Statistics Canada   Statistique Canada

Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes-ménages

Canada

APPROACHES TO SMALL AREA ESTIMATION AT STATISTICS CANADA

Paper submitted by Statistics Canada
for Conference of European Statisticians
Meeting on Statistical Methodology, February 1986.

This paper was prepared by J. Douglas Drew,
drawing upon the work of many others at Statistics Canada.

Methodology Branch Working Paper
Number: CHSM 85-078EF

APPROACHES TO SMALL AREA

ESTIMATION AT STATISTICS CANADA

1. In recent years there has been an increased demand for small area data in Canada. There has been increasing governmental concern with issues of distribution, equity and disparity, with targeting of programs to assist the disadvantaged, be they geographic regions or sub-groups of the population, and with subsequent monitoring of the impact of those programs. Also since many decisions for businesses are made in the context of local social, economic and environmental conditions, private sector demand for small area statistics is strong. Finally, elected representatives want statistics on how well their constituents are faring relative to others.

2. Together with this increased demand there has been progress on both technological and statistical fronts conducive to the development of small area data. These include the automation of administrative data systems, the increased use of the postal code as a geographical identifier on these files, reduction in the cost of computing due to more powerful mainframe computers, and progress on development of statistical modelling methods that allow small area estimates, based jointly on survey data and other census or administrative records, to be produced below the level directly supportable by the survey sample size.

3. In 1982 Statistics Canada began developing a proposal for systematic and integrated development and dissemination of small area data, under which small area data was viewed from a geographic perspective, with the small area being the unit to be described and the different subject matters representing the detail of that description.

4. The Small Area Data Program (SADP) was initiated in 1983, with a 3 year developmental period, after which the program was to become largely self-sustaining through revenues generated by sale of small area data products. In addition to its development and co-ordination role within Statistics Canada, the SADP was also seen as a means of co-ordinating other foci of small area statistical work that exist in other federal departments and in the provinces. Policy issues and a general overview of the program are described by Brackstone (1985).

5. Three components of the program were identified: data development, data systems; and infra-structure. The later sections of this paper are concerned with four principal areas of data development: labour market data, business data, family income data, and post-censal population estimates. In the remainder of this section, the other program components are briefly discussed.

a) Data Systems

6. A key element in the SADP's data system is an inventory of existing small area data sets, including both their characteristics and how they can be accessed. The data sets themselves can be classified either as source data bases or summary data bases. The source bases usually contain micro-level data, which usually reside with the collectors of the data. Access to them

is generally restricted for reasons of confidentiality. Summary data bases, on the other hand, represent consolidations of data from a variety of sources at a single geographic level. Feeding off the summary data bases will be a series of products which can be customised to a user's needs, such as fact sheets on particular characteristics, data on diskettes, etc.

b) Infrastructure

7. The most critical element of the infrastructure is geography. The SADP seeks to provide data for other than just standard areas. Non-standard areas can be defined in terms of standard geography for which data are available by means of geocoding, or by use of postal codes. Geocoding involves assigning longitude and latitude co-ordinates, and has been in use since the 1971 Census of Population. The postal code method uses a conversion file to translate postal codes into standard geography. Both methods are accurate to the block face level in urban areas, but are less accurate in rural areas. An important element of the SADP has been and will continue to be the further improvement of these methods. For example, in the 1986 Census the postal code is being captured on a 20% sample basis to strengthen the conversion process.

II. LABOUR MARKET DATA DEVELOPMENT

a) Objectives

8. The objectives of the labour market data development project have been to examine possible extensions to the array of labour market measures for small areas, and to make recommendations on which data should be produced by Statistics Canada. Initial efforts have been focussed mostly on estimates for Census Divisions, these being the standard geographic areas immediately below the level at which data are currently published.

b) Sources

9. This section deals with labour market measures based on a place of residence concept (collected from households or individuals), while measures based on place of work (collected from businesses) are dealt with in the next section.

10. The major sources of small area labour market data are the Labour Force Survey, the Censuses of Population, and administrative data. Small area strengths and weaknesses of each are discussed below.

11. The Labour Force Survey is a monthly survey of 51,000 households covering the civilian, non-institutional population of Canada's 10 provinces. A redesigned sample was introduced in January 1985 which has resulted in an increase in both the quantity and quality of subprovincial data available from the survey. Prior to its redesign, the survey was designed primarily with a view to providing the most reliable national and provincial data. With the principal focus of the redesign being on improved subprovincial data, data are now published monthly for 66 subprovincial regions and 24 Census Metropolitan

Areas (CMA's) with average CV's for unemployed of 15%. This compares with 53 regions and 23 CMA's under the old design. A major weakness of the LFS in terms of its small area potential is the limited ability to produce monthly estimates for smaller geographic areas due to the small sample sizes involved. This problem can be combatted to some extent by accumulating monthly samples to produce quarterly, annual or multi-year average data. An example of this is that under the redesigned sample, quarterly data are being produced for 42 non CMA cities with an average CV of 20% for unemployed.

12. In Canada the census of population and housing is conducted every 5 years. In recent censuses, information on employment and unemployment have been collected on a 20% sample basis. This will also be the case for the 1986 Census.The principal strengths of census data from a small area perspective are their flexibility to tabulate not only for standard sub-provincial areas such as Census Divisions and Federal Electoral Districts, but also for custom user defined areas, and the relatively small sampling errors (for example, an average 5% CV for estimates of unemployed at the Census Division level). Weaknesses are their infrequency, and the relatively long lead time for release of the data (approximately one year).

13. The most pertinent administrative information for estimation of unemployment comes from the Unemployment Insurance (UI) records. The survey concept of unemployment distinguishes two components: job losers/leavers, and new entrants or re-entrants into the labour force. Counts of regular UI beneficiaries without earnings conceptually correspond quite closely to the job loser/leaver component, which typically accounts for 70% of the unemployed. Statistics Canada produces direct counts of the regular beneficiaries without earnings at the Census Division level. In addition Statistics Canada also produces the UI/P ratio - the ratio of regular beneficiaries without earnings to the working age population (15-64 years of age). The strengths of the administrative data are their frequency and timeliness. There is also considerable geographic flexibility, although a limiting factor in this respect is that for smaller areas the error arising from geographic conversion of postal codes can be problematic. Weaknesses of administrative data include the lack of consistency with survey concepts and coverage, and the potential for breaks in the time series due to changes in the administrative program.

c) Small Area Estimators

14. The objective of labour market data developmental work has been to extend the range of small area data while ensuring: (i) consistency with LFS concepts, and (ii) aggregation to already published LFS estimates at higher geographic levels. The small area estimation techniques considered can be broadly classified as either design based estimators, model based estimators or mixture estimators (combinations of the other two). These are described below, with brief reference to some earlier work pre-dating the Small Area Data Program.

15. Design based estimators include the direct (or domain or simple expansion) estimator, and the post-stratified domain estimator. The direct estimator is obtained by applying the standard estimation procedure appropriate to the design of the survey to the sample units falling into each small area or

domain. A problem with the direct estimator stems from the general lack of co-incidence between boundaries of survey design strata and those of the small areas of interest, and this is aggravated by the clustered nature of sample designs encountered in practice. These conditions may result in the small area being over- or under-sampled, so that estimates of totals, conditionally on the given sample, may be severely biased. When benchmark information on a related variable (such as population of the small area) can be obtained from external sources, post-stratification can overcome this problem.

16. Model based estimators include synthetic estimators and structure preserving estimators (SPREE) due to Purcell and Kish (1979). Synthetic estimators were first investigated at Statistics Canada by Ghangurde and Singh (1972), in the context of deriving small area labour market estimates. Their estimator proportioned survey estimates (by age/sex group) for larger areas to small areas on the basis of the small area's share of the larger area's population (by age/sex group) in a recent census. While synthetic estimators yielded significant reductions in the level of sampling error, there is an implicit assumption of homogeneity within the population sub-group between the small area and the larger area which if untrue may lead to serious bias.

17. Mixture estimators seek to combine design based and model based estimators in such a way to take advantage of the strengths and overcome the weaknesses of each. Drew, Singh and Choudhry (1982) proposed and evaluated an estimator they termed the "sample dependent estimator" which relies solely on the post-stratified estimator when the sample in the small area equals or exceeds that expected under the survey design, but otherwise switches to a linear combination of post-stratified and synthethic, with increasing weight on the synthetic component as the small area becomes increasingly under-sampled. In a Monte Carlo study, the sample dependent estimator was found to have comparable m.s.e. but smaller bias than the synthetic estimator. A strategy of using the sample dependent estimator in combination with averaging monthly data to produce reliable annual or multi year average estimates for Census Divisions was recommended.

d) Labour Market Data Development under SADP

18. Two streams of data development were identified under the SADP. The first of these was the implementation of a sample dependent estimation capacity, initially geared to production of employment and unemployment estimates for Census Divisions, but with a view to later generalization to handle flexible area systems. The second stream was further research into model based estimators, which might at some future point improve upon and replace the sample dependent estimator.

19. This strategy was necessary to respond to immediate user demands while continuing research efforts with good potential for future payoffs. The most important of the demands came from the Department of Regional and Industrial Expansion, which required 3 year average estimates of unemployment rates and employment-to-population ratios for individual or combined Census Divisions. These data are now being produced, and used as input into an annual Development Index, in which Census Divisions are ranked into 4 tiers qualifying for successively higher maximum levels of assistance for approved industrial

development programs.

20. In terms of further research, Choudhry and Bélanger (1985) undertook a Monte Carlo study in which the LFS sample design was simulated for one province of Canada using Census Data. Direct, post-stratified, sample dependent, synthetic and SPREE estimators were evaluated. Their formulation of the synthetic estimator was the same as that of Ghangurde and Singh (1972), while the SPREE estimator started with LF estimates by age/sex groups for the small areas from the previous Census as an association structure, which was raked to the following marginals: survey estimates by age/sex for the Economic Region, and total population for the Census Division.

21. Their findings were that the synthetic and SPREE methods had fairly large relative biases for unemployed (7.7 and 11.3% respectively), but on the other hand their sampling error (coefficient of variation) was 50% lower than that for the sample dependent estimator.

22. Based on these findings the sample dependent estimator has lower m.s.e. than the model based estimators for Census Division estimates when a 3 year averaging period is taken, as is the case for the Development Index. This is because sampling errors are reduced by the long averaging period to the point where the high bias of the model based estimators dominates. For averaging periods of less than one year, however, variances dominate the m.s.e., and hence model based estimators perform better.

23. Since bias plays such a key role in the choice between different small area estimators, one direction of further research will be an in depth study of bias. The geographic base of the study will be expanded to encompass all provinces, and the effects of different choices of auxiliary information will be investigated. These include the use of UI data; the use of Census age/sex counts by type of area (rural/urban); and the use of UI data for estimation of the job loser/leaver component of unemployed, and Census data for estimation of the entrant/re-entrant component. Those estimators performing best in terms of bias may then be included in a further Monte Carlo study to evaluate m.s.e.

24. Another possible direction for further research is the investigation of regression estimators. Earlier investigations into logistic regression models to estimate unemployment at the small area level using UI and Census population counts as independent variables did not yield promising results. It is planned to study a more refined regression estimator incorporating as independent variables the log-ratio of UI beneficiaries to population, the industry and degree of urbanization from the Census, and dummy variables for month and year, with standard adjustments to take account of auto-correlation of sampling errors from month to month.

III. SMALL AREA BUSINESS DATA DEVELOPMENT

25. To parallel developmental activity on social statistics, a project was initiated to develop small area business data. Specific objectives were data for Census Divisions, with breakdowns by Industry Division (of which there are

18) and by size of establishment (2-4 classes), for the following variables: Gross Business Income (GBI), Wages and Salaries (WS), Employment, and number of establishments.

26. The general setting for development of small area business data is not unlike that for small area labour market data described in the previous section. Survey data are available for a sample of the population with detailed information based on established concepts, and typically the small areas of interest are domains, that is, they are disregarded by the survey design strata. Administrative data are available which can be geocoded to small areas using postal codes or addresses, and their coverage and concepts differ to some extent from those desired.

27. Given the similarity of setting, it is not surprising that many of the issues to be addressed are similar to those discussed in the previous section. The major added difficulty is that for business statistics, the unit providing the information may be quite different from the unit for which the statistics are desired. For example for incorporated businesses, administrative data may be furnished at the corporation level, whereas the small area data are more relevant when they relate to the local sites of business activity.

28. A summary of the data development activity follows, as reported in more detail by Sande (1985). The estimation of GBI and WS is considered first for unincorporated businesses and then for incorporated businesses, including a discussion of the disaggregation of corporations into employment reporting units (ERU's). Finally the estimation of employment is considered.

a) Financial Data on Unincorporated Business

29. Revenue Canada records, to which Statistics Canada has access under the Statistics Act of 1971, constitute the source of financial information on both incorporated and unincorporated businesses. For unincorporated businesses, every individual with a business must file a financial statement with his tax return. A sample of these returns is sent to Statistics Canada, where financial data are transcribed and an industry code is added, for use in a variety of surveys. The sample design is simple with the only stratification variable being GBI. A 10% sample of filers is chosen with GBI from $10,000 to $25,000, a 25% sample for those from $25,000 to $500,000, and a 100% sample for those over $500,000. In 1982, the sample size was 150,000 from an inscope population of 500,000.

30. In addition to the sample data, an auxiliary source of information exists for filers with GBI of $25,000 to $500,000, from a file maintained by Revenue Canada for auditing purposes. In 1981 the audit file contained information on total sales which was quite similar in concept to GBI, however the file did not contain anything comparable to WS. A problem with use of this file has been frequent changes in its structure in the last few years. It contains less information now than it used to.

31. The initial emphasis in data development has been on deriving direct estimates of GBI and WS from the sample data. Two major issues have been:

i) Automated geo-coding of the files, relating business addresses and/or postal codes to Census Divisions by means of address to postal code and postal code to Census Division conversion files. This has been very successful, with only a small number of cases requiring manual intervention.

ii) Dealing with missing data. In the past, in transcribing Revenue Canada data on wages and salaries no distinctions were made between zero entries and non-response. Consideration is being given to how to deal effectively with this at the imputation stage.

32. Studies have also been carried out examining ways in which the audit file under certain simplifying assumptions could be used to improve upon the sample based estimate for businesses with GBI in the $25,000 to $500,000 range (Hidiroglou 1985). Phases of these studies consisted of examining: i) comparability of data on the audit file with those on the sample file, and ii) performance of alternative small area estimators. Highlights are discussed briefly below.

33. Findings from the data comparisons were:

i) Geographic Data. The filers' address on the audit file was found to be a good proxy for the business address (available only on the sample file), with 95% agreement when using the different addresses to assign Census Divisions.

ii) Industry Data.There was a relatively good agreement (78%) between the industrial codes assigned by Statistics Canada for the sample and by Revenue Canada for the audit file, for the 18 major industry divisions. Procedures were suggested for reconciling the differences.

iii) Financial Data. Sales data on the audit file were found to agree closely with those on sample file, with discrepancies being within a 10% range for over 90% cases.

34. The second phase considered estimation of WS present on the sample file only, using the audit file as a source of auxiliary information on the number of businesses and GBI. Of variables present on both the sample and audit files, GBI was found to be the best predictor of WS. A Monte Carlo study of design based, model based and mixture estimators was conducted. Findings were that the mixture estimators were best in terms of lowest bias and root mean square error, the model based (synthetic) estimators may be very biased, the direct estimator displayed the highest root mean square error, and finally the performance of the post-stratified estimator was only marginally worse than that of the mixture estimators.

b) Financial Data on Incorporated Businesses

35. The primary data sources are:

i) Revenue Canada Data: A detailed sample of 25,000 records exists, from which WS and GBI are imputed with a ratio type model based on assets, sales and other information for the universe of 500,000 businesses.

ii) T4 Supplementary File: The T4 slip in Canada is the annual summary of pay and deductions that every tax filer receives from his employer. Each T4 record contains a Payroll Deduction (PD) number, which can be used to identify the employer.

iii) The Business Register (BR): Statistics Canada maintains a register of businesses, built around the PD system. It links PD numbers to establishments to corporations, and is also linked to the frames of major economic surveys. It is coded both for geography and industry.

36. The major difficulty with Revenue Canada corporate data is that they pertain to corporate headquarters and not the actual location of the economic activity. Alternative means of disaggregating the corporate data are under investigation. One approach is use of T4 data to obtain WS by PD number, and then linking the PD number to establishments or ERU's via the BR and the Survey of Employment, Payroll and Hours. PD numbers linked to more than one ERU on SEPH will have WS allocated proportionately to the reported employment. Another approach for WS is the linking of the imputed Revenue Canada file directly to the BR. For GBI disaggregation, use of economic censuses is being considered, since this approach will permit use of the Census GBI data as a basis for allocating Total Revenue from the T2 file among the component establishments.

c) Employment Estimates

37. The Survey of Employment, Payrolls and Hours is a monthly survey which collects data from 70,000 Employment Reporting Units. The sample is stratified by province, size of business, and major industry division.

38. Initial efforts are being devoted to producing direct and post-stratified estimates of employment by industry division for Census Divisions and estimates of their variances. Strategies for collapsing industries and CD's are also being investigated. Pending the results, consideration may later be given to more sophisticated model based or mixture estimators such as those discussed in earlier sections, and to producing annual averages.

IV. STATISTICS ON FAMILIES AND FAMILY INCOME

39. Current sources of sub-provincial family income statistics are the Census of Population and the Annual Survey of Consumer Finances (SCF). Income data are collected on a sample basis for each decennial Census, and also for some quinquennial Censuses - a case in point being the upcoming 1986 Census. In 1981, with a 20% sample, the average CV for annual income for Census Divisions was 1.5%. From the SCF currently not a great deal of subprovincial data is produced. Estimates are published for selected CMA's. For the 16 CMA's published for 1982, the CV for average family income was 3.6%.

40. Research efforts have concentrated on examining the potential of Revenue Canada's personal income tax file for statistics on families and family income. The Revenue Canada files are files of individuals. They contain Social Insurance Number (SIN), marital status, spousal SIN, sources of income,

and amounts of tax deductions and exemptions.

41. In deriving families from the Revenue Canada files, the attempt was to come as close as possible to the Census definition of family, which corresponds to a husband and wife or a lone parent together with any never married children living in the same dwelling. Spouses who both filed tax returns were linked using the spousal SIN where present, or on the basis of name, address, age and marital status otherwise. It was also attempted to link single filers under 30 years of age to a family on the basis of name and address. Finally for each family, the number of non-filing children was estimated from the amounts of the different tax benefits claimed.

42. Experimental family estimates were produced for the smallest Canadian province and evaluated, using Census data (Auger 1985). Comparisons showed that the administrative estimates of the number of families and unattached individuals were not too far off those from the Census, with a 6% under-estimate and a 9% over-estimate respectively. The coverage of children under 18 years old was found to be close to 100% while for children 18 years old and over, the coverage was closer to 80%. The last result was probably due to the limitations in the computer methods used in matching children who filed a tax return. However, these procedures have been improved. One of the major problems associated with this administrative file was found to be the low coverage of individuals 65 years of age and over. Also the estimated number of unattached individuals under 35 years of age was much larger than its Census counterpart. This was mostly due to common-law couples who were not considered husband-wife families in the study, and to the presence of unmatched children.

43. Based on the demonstrated potential for estimating families and family income (principally for families with members under 65 years of age), efforts are currently underway to derive family estimates for all provinces, incorporating matching of common-law couples. Finally some separate work has been initiated with the old age security pension file, which will be instrumental in improving the coverage of the elderly and their families.

V. POST-CENSAL POPULATION ESTIMATION

44. Prior to 1982, Statistics Canada produced annual population estimates for Census Divisions and CMA's based on a component method of estimation. These data appeared 15-18 months after the reference period.

45. Research was carried out under the SADP to evaluate alternative estimation procedures with a view to improving both timeliness and accuracy. This research lead to development and implementation of a procedure under which, for Census Divisions and CMA's, two sets of estimates will be published yearly, appearing 3-4 months and 11-15 months after the reference date respectively. Research findings are reported by Verma and Basavarajappa (1985). Some salient points are noted below.

46. Under the approach adopted, estimates for the current year are based on the component estimates from the previous year, to which is added an estimate

of the population change, based on the difference between the regression estimates for the two points in time. The estimates obtained in this fashion have been termed the "regression-nested estimates". For the same reference period, estimates are produced later using a component approach which provides information on components of change, namely births, deaths and net migration.

47. Symptomatic variables used in the regression estimation differed from province to province depending on the availability of different administrative records, and included family allowance data, counts from provincial health care files, and electricity connections. The component method uses births and deaths from vital registers, and estimated migration from the Revenue Canada personal tax files.

48. Evaluation studies found the regression-nested estimates to be better than either the component or standard regression method for estimation of total population, with a mean absolute error of 1.7% for Census Divisions. Hence the estimation strategy adopted yields timely and accurate estimates of total population, with estimates of components of population change being available later with the component estimates.

49. Further research work is planned to evaluate the effect of a log-linear transformation of the model, and to test the model more fully with different sets of symptomatic data files (such as health care files, and drivers licences).

VI. CONCLUSIONS

50. As it nears the end of its third year, the Small Area Data Program has successfully accomplished much of its initial developmental mission. Despite increased emphasis in the second phase of the program on production and cost recovery, important developmental opportunities still exist and a portion of the program's efforts will continue to be channelled into these.

51. These opportunities include the greater exploitation of data from surveys other than the Labour Force Survey and the Survey of Employment, Payrolls and Hours, further research into techniques to preserve confidentiality particularly for small area business data, and the acquisition and development of new administrative data sources such as old age security records, files of telephone and electricity connections, motor vehicle registrations and municipal tax assessments.

VII. REFERENCES

Auger, E. (1985): Family Income Statistics from Revenue Canada's Personal Income Tax File: A Case Study. Paper presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.

Brackstone, G.J. (1985): Small Area Data: Policy Issues and Technical Challenges. Paper presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.

Choudhry, G.H., and Bélanger, Y. (1985): Small Area Estimates from Sample Surveys. Paper presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.

Drew, J.D., Singh, M.P., and Choudhry, G.H. (1982): Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, Survey Methodology, Vol. 8, No. 1, pp. 17-47.

Ghangurde, P.D. and Singh, M.P. (1973): Synthetic Estimation in Periodic Household Surveys, Survey Methodology, Vol. 3, No. 2, pp. 152-181.

Hidiroglou, M.A. (1985): Problems Associated with the Estimation of Small Area Business Data. Paper presented at Statistical Society of Canada meetings, Winnipeg, Canada. June 1985.

Purcell, N.J. and Kish, L. (1979): Estimation for Small Domains, Biometrics, Vol. 35, pp. 365-384.

Sande, I. (1985): The Small Area Business Data Development Project at Statistics Canada. Paper presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.

Verma, R.B.P. and Basavarajappa, K.G. (1985): Recent Developments in the Estimation of Population for Small Areas in Canada by Regression Method. Paper presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.