MODEL-BASED SMALL AREA ESTIMATION OF UNEMPLOYMENT: A CANADIAN EXPERIENCE

S. Earwaker, N. Brien, J. F. Gosselin, Statistics Canada

1. INTRODUCTION

During 1983, Statistics Canada received funding to initiate a new program called the Small Area Data Program whose objective is to expand the bureau's capacity to produce and disseminate geographically detailed statistical data. One of the tasks given high priority is the development of labour market data for small areas and in particular Census Divisions (CDs).

This paper presents some results of an ongoing study of model-based small area estimation of unemployment using administrative, survey and census data. The primary sources of data and the need for small area labour market data will first be discussed. The estimators considered will then be presented together with an assessment of their performance. The paper will then conclude with an indication of future directions of work.

PRIMARY DATA SOURCES

The Canadian Census of Population and Housing provides the most comprehensive source of statistical data on the labour market. Since 1971 every census has included some questions on labour force activity, although previous censuses did not always do so. It appears that the censuses will remain an important data source since the plans for the 1986 Census are to essentially replicate the 1981 Census content.

Although censuses provide the richest source of small area data in terms of geographic detail and range of variables available, they are conducted infrequently because of their high cost and, therefore, they cannot be used to monitor ongoing labour market conditions for local areas.

The Canadian Labour Force Survey (LFS) is a monthly survey of approximately 55,000 households which provides the basis for the production of monthly unemployment rates. Given its sample size, the survey provides an excellent vehicle for estimating levels and trends for provinces.

Sub-provincial estimates can also be produced from the LFS. For example, monthly and yearly averages are currently published for some 53 Economic Regions and 23 Census Metropolitan Areas. Under the new design which will take effect in 1985 (see Singh, Drew and Choudhry, 1984) there will be an increase in the number of sub-provincial areas for which estimates are released.

Administrative data on the Federal Government Unemployment Insurance (UI) Program are available to Statistics Canada. Since 1978, the geographic identifier on the UI records has been the postal code. The files are currently being used to produce counts of beneficiaries cross-classified by variables such as age and sex for small areas together with a derived measure, namely the UI to population ratio (Leyes, Bobet and Radley (1982)). The main advantages of using these data are the good small area potential, the flexibility of area systems which the postal code could provide, and the ease of producing tables. These counts and ratios of UI beneficiaries differ from the unemployment estimates produced from the LFS primarily because of conceptual differences between beneficiaries and unemployed. However, comparisons between beneficiary counts and Census estimates of unemployed at the CD level indicate that the beneficiary data are a good basis for developing model-based estimates.

SMALL AREA LABOUR MARKET DATA

There are 260 CDs in Canada (excluding the Yukon and Northwest Territories) averaging approximately 90,000 population. The population sizes of CDs vary significantly, ranging from 2,000 to about 2,000,000 people.

The LFS, with its large sample size, has a fairly good small area potential. However, LFS estimates of unemployment in CDs tend to be subject to high levels of sampling error, particularly since some CDs are very small, and since CDs generally cut across survey stratum boundaries introducing an additional source of variability. Techniques such as collapsing of smaller CDs and averaging over months are used in order to stabilize the estimates. For example, 3-year averages have been produced for approximately 145 CDs or groups of CDs with coefficients of variation less than or equal to 25%.

More recently, a sample dependent estimator (Drew, Singh and Choudhry (1982)) has been developed and evaluated. For small areas where the sample yield is sufficient to produce estimates of acceptable precision, the procedure relies on a post-stratified design-based estimator. When the sample yield is insufficient, the procedure then switches to a linear combination of the post-stratified estimator and a synthetic estimator using population as the auxiliary variable.

The estimators discussed in this paper attempt to make use of other sources of unemployment data including the administrative UI file. Such estimators could eventually be applied directly or could be used as input to the sample dependent procedure if the new estimators prove to be more efficient than the model-based component (i.e. population-based synthetic) currently being used.

2. THE ESTIMATORS FOR UNEMPLOYMENT

In constructing estimation models, the objective has been to make the best possible use of the available data sources. The models typically rely on sample survey estimates to provide the relevant control totals for the calibration of the closely related UI data which provide the small area detail.

Efforts to date have focused primarily on minimizing and estimating model error, that is, the error which may be attributed to the departure from assumptions underlying the model. Accordingly, most of the models tested to date were developed and evaluated using 1981 Census data.

As a measure of model error, the percent Absolute Relative Difference (ARD) was calculated as the relative difference between the model-based estimate and the count of unemployed as estimated in the 1981 Census at the CD level. To date, three estimators have been tested: synthetic, SPREE, and regression. These will now be discussed.

SYNTHETIC ESTIMATION

The synthetic estimator uses survey estimates at the province level for age-sex classes of unemployed persons. These estimates are then distributed over Census Divisions within the province, according to the distribution of UI beneficiaries in each age-sex class.

The "survey estimates" were taken from the Census for this study, although in practice the method would use estimates from the monthly Labour Force Survey, which are subject to much greater sampling variability. The evaluations here seek to study only the model error. Subsequent work will address the combined error from both the model and the sample.

The average ARD for all 260 Census Divisions was 19.6%. A weighted average ARD which is weighted by the number of unemployed persons in each CD was estimated to be 16.0%. This is somewhat lower than the average ARD since the estimator performs better in larger CDs.

A similar synthetic estimator was constructed using population data in place of UI beneficiary data as the auxiliary source. The average ARD for all 260 Census Divisions was 23.5% and the weighted average ARD was 18.8%. These estimates do not compare favourably with the UI-based synthetic estimator described above, suggesting that the UI data are better suited for this and similar modelling applications.

SPREE ESTIMATION

One of the difficulties in modelling with UI data is that the UI eligibility criteria (which govern the processing of claims from unemployed persons) differ from region to region depending upon the regional labour market characteristics. To overcome this problem, a more sophisticated model was constructed to incorporate information at the regional level. Following the Structure Preserving Estimation (SPREE) procedure suggested by Purcell and Kish (1979), a second set of survey estimates was added to those of the earlier synthetic model, namely, total unemployed at the the Economic Region level within provinces.

This produced estimates of unemployed which were consistent (1) with the survey estimates both for age-sex groups at the province level and Economic Regions within provinces, while otherwise preserving (as much as possible) the structure of the UI data. The average ARD for all 260 CDs was 14.6%, and the weighted average ARD was 8.2%.

Another concern with respect to the model error is the incidence of extreme errors. For the SPREE model it was found that 23% of the CDs had ARDs in excess of 20%. The larger or more extreme errors were generally found in the smaller CDs. It is not surprising that the method did not perform well for every CD, especially considering the great range in sizes of CDs. Nevertheless, these extreme errors should be examined more closely.

At first glance, one might suspect that a SPREE model, based on more detailed survey estimates might perform better than those tested to date. However, these more detailed survey estimates might be subject to large sampling errors. One would suspect, therefore, that the possible reductions in model error to be gained by such a model would be offset by increases in sampling error. More work will be necessary in order to assess the relative contributions of these two sources of error to the total error. In particular, a Monte Carlo study is planned which will address this issue.

REGRESSION ESTIMATION

Several variations of the linear regression model were tested, but the description here will focus on the one which was considered to be superior. A regression equation was obtained using 1981 Census estimates of unemployed at the Census Division level as the dependent variable. UI beneficiaries and total population (from 1981 Census) were used as independent variables.

To address the problem of differing UI eligibility criteria, CDs were stratified into four strata according to the prevailing regional requirements for "number of insured weeks". An analysis of these data indicated that there were four distinctly different linear trends, one for each of the four strata. A weighted least squares (2) regression procedure was used to fit four separate regressions.

Model error was estimated using ARD. Overall the estimated average ARD was 12.3%, and the weighted ARD was 7.8%.

The distribution of extreme errors was remarkably similar to that of SPREE. For the regression method, 18% of the CDs had ARDs in excess of 20%.

An important insight was gained in the development of regression models. When the independent variable on population was not present in the regression, the model error was almost doubled. This suggests that there is some part of the unemployed, which was not "explained" or represented by UI beneficiaries. In reality, we know this is true. The "uninsured unemployed" include all such persons who are unemployed but do not qualify for UI benefits. Based upon our experiences we observe that a linear combination of the population variable and the UI variable is a better predictor for some subgroups of the unemployed than UI variable alone, and that this subgroup probably constitutes the uninsured unemployed.

3. DISCUSSION AND GENERAL COMMENTS

The ARD's for the three methods tested are shown in Table 1 (see Appendix). We can make a few observations on these results at this time. First, it is apparent that when additional cross classification structures are available, the SPREE method can be used to advantage. In our experience, the SPREE method had lower model error than the synthetic method.

A second observation is that the SPREE and regression models gave similar results, although they can not be compared fairly on the basis of the ARD's shown here. The SPREE model included age/sex information which was not used in the regression approach. Similarly the regression made use of an additional symptomatic variable, namely population. In attempting to develop the best version of each method examined, the authors concede that they may have produced results which to not permit a direct comparison of methods.

A final observation is that all methods tended to yield higher relative errors for the smaller CD's. This is reflected in the lower weighted average ARD's as compared to the unweighted averages. This observation suggests that there may be some lower size limit to the small area estimation capability of these methods.

One of the limitations which exists when relating UI data to LFS or Census data pertains to the geographic identifier on UI data, namely the postal code. Firstly, the areas serviced by a single postal code can intersect more than one Census Division. This is particularly a problem in rural areas where the entire area serviced by each local post office is covered by a single postal code. A second problem is that many people receive their mail in locations other than their usual place of residence. Household surveys on the other hand would enumerate such individuals at their usual place of residence. An initial study of these problems has estimated average geographic conversion errors of 3.3% over all CD's. Although the average is relatively small, the conversion error varies significantly across areas and has been estimated to be as high as 40% for certain CDs. We recognize that this is a limitation of the postal code which will probably exist for all models using the UI data source.

The coefficients of the regression model have been estimated using the 1981 Census results. If we use the regression equation to produce estimates in post-censal time periods, the model may suffer from time lag bias - a bias due to changes in the modelled relationships. Such changes might result from changes in the administration and coverage of the UI Program or from changes in labour market conditions.

One alternative to this approach is to estimate regression coefficients monthly using Labour Force Survey data. This alternative to symptomatic regression was suggested by Ericksen (1974) and is known as "sample regression". As noted in his work, the sampling variability in observations on the unemployed is an additional source of error in the estimates produced from this regression approach.

The SPREE methods would also employ monthly LFS data and would be subject to additional sampling variability. The models investigated to date utilize survey data at large area levels. In practice, we know that sampling variability is acceptable at these levels, (although we have not yet evaluated the sampling error present in the SPREE estimates). SPREE models constructed at lower levels would probably be less biased, but in general would suffer from higher levels of sampling error. In future work we shall investigate the tradeoffs between bias and variance, and try to find some level of construction which is optimal in the sense of mean squared error.

4. CONCLUSIONS AND FUTURE WORK PLANS

The results of initial investigations of model-based small area estimation of unemployment have led to the following conclusions:

(i) in spite of basic definitional differences and other problems, the
UI data are highly correlated with the statistical concept of
unemployed;

(ii) modelling techniques such as SPREE and regression can be
used and appear to have good potential.

Our experience with this application has suggested several areas for future work:

a) The models tested essentially rely on the distribution of UI data across Census Divisions. Postal code conversion problems which are known to exist introduce distortions in these distributions which lead to biased estimates. Improved conversion files which are becoming available are almost certain to lead to improvements in the estimates. Tests should be run with such files. Such files would also permit the use of other, more homogeneous areal units. Such units could be clustered according to labour market characteristics like the labour force participation rate. This should help to minimize model error due to variation in labour market conditions.

b) To date, the only cross-classification variables tested are age and sex. Another variable available on the UI file which should have good potential for modelling purposes is 'occupation' (which has obvious relevance to labour market statistics). Preliminary investigations have indicated that UI occupation is conceptually compatible with LFS occupation at major group levels. This cross-classification variable may improve estimation.

c) As mentioned earlier, there is a certain subgroup of the unemployed not covered by the UI data, namely the "uninsured unemployed". This suggests the use of a component method which would involve the use of UI data as a predictor of unemployment for the insured population (or some proxy of it) and some other variable(s) (e.g., population, population in certain age groups, etc.) for the residual category.

d) A Monte Carlo study is being planned which will provide a means of assessing the combined effects of sampling variance and model error on the estimators presented in this paper (and others). It will also offer the opportunity of fine-tuning the methods. The question of optimal level of constructing SPREE models may be investigated in this context.

FOOTNOTES

(1) Under the SPREE procedure, consistency is obtained by adjusting the structure (here, UI Beneficiaries) iteratively using the Iterative Proportional Fitting algorithm.

(2) Weights of 1/UI were selected from a number of alternatives.

REFERENCES

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**, pp.17-47.

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1984). "Post 1981 Censal Redesign of the Canadian Labour Force Survey". Unpublished manuscript to be presented at the 1984 American Statistical Association Meetings, Philadelphia.

Ericksen, E.P. (1974). "A Regression Method for Estimating Population Changes of Local Areas". **JASA 69**, pp.867-875.

Leyes, J., Bobet, E. and Radley, L. (1982). "The Use of Unemployment Insurance Records to Derive an Unemployment Indicator". **American Statistican Association Proceedings 1982**, Section on Survey Research Methods.

Purcell, N.J. (1979). "Efficient Small Domain Estimation: A Categorical Data Analysis Approach". Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

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

APPENDIX

TABLE 1
Estimated Average Model Errors

Model	Average ARD (%)	Weighted Average ARD (%)
Synthetic (Pop-based)	23.5	18.8
Synthetic (UI-based)	19.6	16.0
SPREE	14.6	8.2
Regression	12.3	7.8