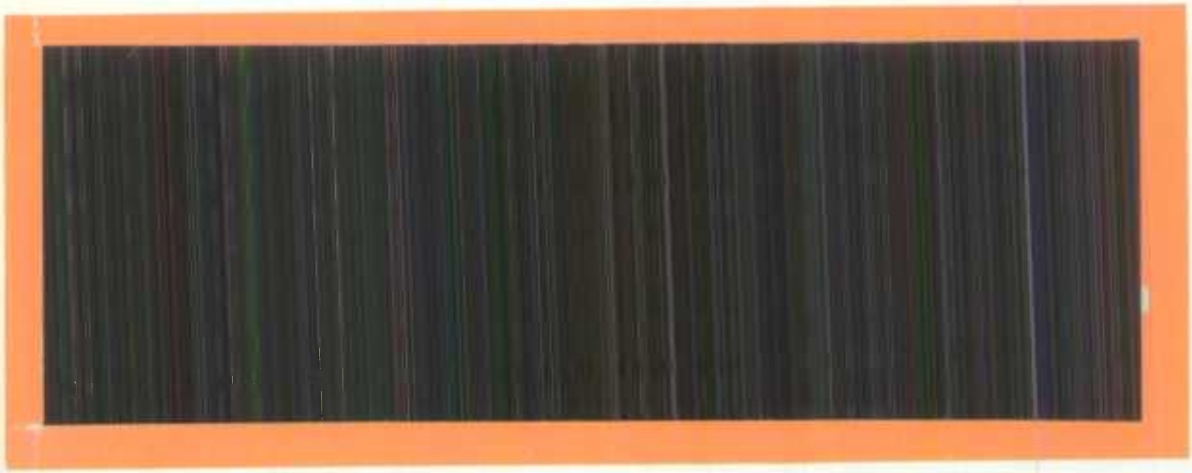
Statistics Canada
Statistique Canada



Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

MODEL BASED UNEMPLOYMENT ESTIMATES FOR SMALL AREAS

I. Trottier and G.H. Choudhry

Census and Household Survey Methods Division

Methodology Branch

CHSM 85-077E

MODEL BASED UNEMPLOYMENT ESTIMATES FOR SMALL AREAS

I. Trottier and G.H. Choudhry

Monthly unemployment estimates for Census Divisions (CD's) were obtained for the period 1981-1983 using synthetic and component methods. The estimates were constructed using Labour Force Survey data, Unemployment Insurance data, and Census data. These estimates were evaluated for their performance relative to existing data sources.

1. Introduction

As part of the modelling work on Labour market data for small areas, synthetic and component methods have been used to estimate total unemployment for Census Divisions (CD's). The monthly estimates of unemployment at the Economic Region (ER) or grouped ER's level with 10% or better c.v. (coefficient of variation) were obtained from Labour Force Survey data. The synthetic estimates for CD's were then obtained from ER (or grouped ER's) level estimates using Unemployment Insurance data and Census population counts. In the component method, two components of unemployment: (i) job losers and job leavers, (ii) new entrants and re-entrants were similarly estimated from ER (or grouped ER's) level data and the estimates of total unemployment for CD's were obtained by adding the two components of unemployment. The job losers and job leavers represent the experienced portion of the unemployed population and the new entrants and re-entrants include people entering the Labour Force or returning to the job market. Only the people in the first category are eligible to receive Unemployment Insurance benefits and the second component consist mainly of student searching for a job during the summer and women returning to work in the fall when children go back to school and women entering the Labour Force after they have raised their children.

The two model based techniques: (a) synthetic, (b) component are described in section 2. The monthly estimates for CD's were obtained for the period 1981-1983 using synthetic and component methods. These series were evaluated by comparing them with the Labour Force post-stratified three-year average estimates and also with sample dependent (Drew, Singh, and Choudhry; 1982) three-year average estimates. The results of this evaluation are presented in Section 3. In section 4, we give the conclusions of the evaluation study and the plans for further work.

2. The Estimators

We have considered two model based methods: (a) synthetic method, (b) component method to obtain monthly series of unemployment for Census Division (CD's). There are a total of 260 CD's across Canada and they vary considerably in size from a population of 2,000 to 3,000,000.

The synthetic series is obtained by disaggregating the monthly Labour Force Survey estimates of unemployed for the economic region ($\hat{Y}^{LFS}_{ER}$) using the number of Unemployment Insurance beneficiaries for the CD (UI_{CD}) and for the economic region (UI_{ER}) and the population for the CD (P_{CD}) and for the ER (P_{ER}) as follows:

$$\hat{Y}^{S}_{CD} = \tfrac{1}{2}\left(\frac{UI_{CD}}{UI_{ER}} + \frac{P_{CD}}{P_{ER}}\right) \times \hat{Y}^{LFS}_{ER}$$

where $\hat{Y}^{S}_{CD}$ is the synthetic estimate of unemployed for a given CD. This in fact is simply the average of two synthetic estimators, one based on Unemployment Insurance beneficiaries counts and the other based on population counts. The simple average of the two synthetic estimates was taken because the previous evaluations showed that the average of the two estimators had smaller bias than either of the two synthetic estimators.

An alternative way of using Unemployment Insurance beneficiaries counts and population is to apply the synthetic estimation method on two components of the unemployed population. In the component approach, the unemployed are subdivided into two categories:

1. Job losers and Job Leavers
2. New entrants and re-entrants

The Labour Force Survey (LFS) provided estimates of the two components of unemployed at the ER level and the component estimates were obtained as follows:

$$\hat{Y}^{C}_{CD} = \left(\frac{UI_{CD}}{UI_{ER}} \times \hat{LoLe}^{LFS}_{ER} \right) + \left(\frac{P'_{CD}}{P'_{ER}} \times \hat{NeRe}^{LFS}_{ER} \right)$$

where $\hat{Y}^{C}_{CD}$: Component estimates of unemployed for a given CD

UI : Unemployment insurance beneficiaries counts

P' : Population comprised of men between the ages of 15 and 24, plus women between the ages of 15 and 44 (based on a study of the distribution of age-sex groups for the new entrants and re-entrants)

$\hat{Lole}^{LFS}_{ER}$: Estimate of job losers and job leavers at the ER level from the LFS

$\hat{NeRe}^{LFS}_{ER}$: Estimates of new entrants and re-entrants at the ER level from the LFS

In the two estimation procedures, previously outlined, ER's were collapsed when the coefficient of variation (CV) of an individual ER estimate was greater than 10%. The collapsed ER's would possess the same UI eligibility criteria (minimal required number of weeks worked) in order to be grouped. A list of the collapsed economic regions for the 36 months (1981 to 1983) is provided in Appendix 1.

Also in some provinces, CD's do no respect ER boundaries. It is therefore possible that a Census Division may be contained in more than one ER. In these cases, the Census Division was split into two parts (if it belonged to two ER's) and estimates were obtained for the two parts and then summed to provide a total Census Division estimate. The CD's were split on the proportion of their 1981 Census admissible population falling into each ER. The auxiliary variables (UI Beneficiaries, monthly population) were also treated accordingly. A list of CD's which are contained in more than one ER is provided in Appendix 2.

3. Evaluation of Experimental Series

The consistency of the experimental series of labour market data (i.e. unemployment), were evaluated against two sources:

a. Labour Force Survey post-stratified estimates for Census Divisions (three year averages)
b. Sample Dependent estimates (see Drew, Singh, and Choudhry; 1982) for groups of Census Divisions (three year average) already produced for the Department of Regional Industrial Expansion (DRIE).

The post-stratified estimator is unbiased except for ratio estimation bias which is negligible for large sample sizes and the sample dependent estimator has been evaluated in a Monte Carlo study (Drew, Singh and Choudhry, 1982) which showed that the estimator had a very negligible bias. Thus the experimental series were evaluated by comparing with the post-stratified estimates and the sample dependent estimates so that any significant differences between the two estimates being compared will be due to bias in the method being assessed. The three year average estimates were compared during this evaluation so that the sampling variance will not be too high and hence relatively small biases can be detected. The experimental series when averaged over three years may not have significant bias, but it could be biased when average is taken over shorter period of time because the bias could be positive for some periods and negative for others and by averaging over longer period of time these biases may cancel out.

Out of the 260 CD's, the three year average post-stratified estimates were obtained for 243 CD's because the remaining 17 CD's did not have sample for all 36 months. The three year average sample dependent estimates are for groups of CD's where 260 CD's are grouped into 183 small areas (domains).

For the purpose of this report, consistency is defined as the absolute percentage difference between the existing data source and the method being assessed, thus representing departure from the existing data source in a percentage form. During evaluation, the experimental series was reformatted for comparability purposes whenever it was necessary e.g. collapsing the CD's.

Although the component method is based on more solid underlying assumptions than the somewhat ad hoc synthetic method, the evaluation of the experimental series showed that these series had similar consistency with the other two data sources. Therefore the results of consistency evaluation are reported for component series only. The test for the significance of the difference between the three year average post-stratified and the experimental series was performed for the synthetic series only, therefore these results are reported for the synthetic series (Table 3) and not for the component series.

The series were not tested for their consistency with LFS monthly estimates of unemployed for Economic Regions since the synthetic estimation method used ensured such a consistency. In some cases, though, the perfect correspondence between the LFS ER estimate and the sum of the CD component estimates included in the ER is not achieved because ER's may have been collapsed. In those cases, the sum of the CD estimates included in the collapsed ER's will correspond to the sum of the LFS ER estimates.

The component series was therefore evaluated against the LFS Post-Stratified series and the Sample Dependent series.

3.1 Post-Stratified Series

The post-Stratified estimates are obtained by using the number of unemployed from the LFS in the small domain and adjusting that estimate by the ratio of the Census Population projections for the small area to the LFS estimated population in the small area.

The monthly, annual averages and three year averages of post-stratified estimates were produced from the LFS data for 260 CD's and were then compared with the component series. Out of these 260 Census Divisions, seventeen were excluded since there was no sample for all the 36 months (9 CD's) or no sample for some of the months (8 CD's). The results for the three year averages are reported here.The monthly and annual post-stratified estimates were not included in the evaluation because of their poor reliability.

Table 1 shows the consistency of the component estimates with the three-year average post-stratified series by population size of the CD's.

Table 1: Distribution of the Absolute Percentage Differences (APD) between the Post-Stratifed three year average and component methods by the CD population size. The column percentages are given in parentheses.

POPULATION SIZE

APD	less than 25000	25000 - 75000	Over 75000	Total
APD ≤ 10%	23 (26%)	39 (38%)	36 (71%)	98 (40%)
10% < APD ≤25%	35 (39%)	40 (39%)	14 (27%)	89 (37%)
APD > 25%	31 (35%)	24 (23%)	1 (2%)	56 (23%)
TOTAL	89 (100%)	103 (100%)	51 (100%)	243 (100%)

The above table indicates that for 77% of the CD's, the absolute percentage differences (APD's) between the component and the post-stratified estimates are less than or equal to 25%. We also notice that most of the cases where the differences are greater than 25% are comprised of smaller CD's.

3.2 Sample Dependent Series

The sample Dependent series is a linear combination of post-stratified method and synthetic method when the sample size for the LFS is not sufficient in the small area and it coincides with the post-stratified method when the sample in the small area exceeds a predetermined critical value (see Drew, Singh, and Choudhry; 1982). This series provides three year averages of unemployed for 183 groups of Census Divisions. The component series was therefore reformatted to the 183 groups and three year average were calculated for the groups. The consistency between the two methods is outlined in the following table, where the absolute percentage differences (APD's) are given by population size of the CD groups.

Table 2: Distribution of the Absolute Percentage Differences between the Sample Dependent and Component Series by the population size (1981 CensusPopulation 15 years and over) of the groups of CD's. The Column percentages are given in parentheses.

POPULATION SIZE

APD	less than 25000	25000 - 75000	Over 75000	Total
APD ≤ 10%	10 (24%)	30 (39%)	42 (65%)	82 (45%)
10% < APD ≤25%	15 (37%)	30 (38%)	19 (30%)	64 (35%)
APD > 25%	16 (39%)	18 (23%)	3 (5%)	37 (20%)
TOTAL	41 (100%)	78 (100%)	64 (100%)	183 (100%)

Table 2 indicates that for 146 groups of CD's out of 183 or 80% of the cases, the absolute percentage difference between the two series is less than or equal to 25%. We also notice that in larger areas, smaller APD's ar observed. The cases for which the APD is greater than 25% is mostly comprised of smaller CD's: 16 out of 41 or 39% of CD's with a population less than 25000 people and 18 out 78 or 23% of CD's (or groups of CD's with a population of 25000 to 75000 people.

The comparison of the three year average component estimates with the post-stratified estimates and with the sample dependent estimates shows that for most cases (over 75%) the absolute percentage differences are less than 25% and the larger relative differences are in smaller CD's. The consistency assessment is based on the reliability of the method used, in this instance, the absolute percentage difference. A significance test of the APD was conducted, in order to determine the reliability of the consistency measure. The test was based on the assumption that the difference between the two methods (three year average) follows a Student distribution for which the test-statistics is defined as:

$$t = \frac{\bar{y}_s - \bar{y}_p}{\sqrt{Var(\bar{y}_s - \bar{y}_p)}} \quad ,$$

where $\bar{y}_s$ is the average of the experimental series and $\bar{y}_p$ is the average of post-stratified series and var $(\bar{y}_s - \bar{y}_p)$ is the variance of the difference between the two estimates.

Significance values (t - values) were obtained for each Census Division for the difference between the Post-Stratified three year average and synthetic three year average estimates. Non-significant differences between the two series are determined by t values comprised between -2 and 2. The statistical test for the significance of difference

between three year average post-stratified and the experimental series was performed for the synthetic series only as the results for the component series should be similar due to similarity between the two experimental series. The following table contains the results of the significance evaluation of the difference between the two series.

Table 3: Significance of the difference between the three year average post-stratified and synthetic method by the CD Population Size

POPULATION SIZE

	less than 25000	25000 - 75000	Over 75000	Total
Significant	42	41	19	102
Non-Significant	47	62	32	141
TOTAL	89	103	51	243

As outlined in the above table, the difference between the synthetic series and the post-stratified three year average series is significant for 102 CD's or 42% of all CD's (243). It is also observed from the same table that the proportion of CD's with significant differences is higher for smaller CD's. But the proportion of larger CD's with significant differences is also quite high (37% for CD's with population over 75,000). One reason why bias is present in many CD's may be that the estimate which is disaggregated, in this case the Economic Region unemployment estimate, was confined to a too stringent reliability criterion, i.e., the economic regions which had an estimated number of unemployed with a CV greater than 10% were to be collapsed with another economic region. From a small case study, it was apparent that collapsing of some ER's were to yeild bias since the industry composition of the grouped ER's were different although contiguous. The synthetic estimates without collapsing the ER's were also obtained and the tests of significance were carried out. The number

of CD's with significant differences between the synthetic and the post-stratified estimates decreased from 102 with collapasing to 88 without collapasing the ER's. Thus we are confronted by a problem that the increased reliability of the synthetic estimates can be achieved at the risk of higher potential bias. This is also demonstrated by Choudhry and Belanger (1985) in a monte carlo study.

As pointed in tables 1 and 2, most CD's for which large differences were observed have a population less than 25,000 and these differences turned out to be significant. This would suggest that for smaller CD's another approach should be envisaged because these cases have appeared to be the major source of inconsistency throughout the evaluation. These census divisions could be collapsed to other CD's in order to improve the small area estimates. The remaining cases where the difference proved to be significant require more attention and investigation.

4. Conclusions

The evaluation of the experimental series proved that for over 40% of the CD's, the estimates of unemployment are subject to model bias. The proportion of CD's with bias is larger for CD's with smaller populations. One solution would be to group the smaller CD's with others in order to improve the quality of the estimates. However, some Census Divisions with larger populations have an estimated number of unemployed which is significantly different from what external sources seem to indicate. We suggest that further investigation be conducted for these cases from an economic point of view. It seems that the unemployment insurance data and the census population counts at the small area level may not be sufficient to model the unemployment at the small area level. We are currently testing the logit models for unemployment at the CD level using unemployment insurance counts, Labour Force by industry from Census, and the composition of the CD by types of area. If the model gives an adequate fit, then more reliable smoothed estimates based on the model can be produced for the Census Divisions.

References

1. Choudhry, G.H., and Belanger, Y. (1985), "Small Area Estimates from Sample Surveys," Presented at the International Symposium on Small Area Statistics, May 1985; Ottawa, Canada.

2. 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 Journal, Vol. 8, pp 17-47.

APPENDIX 1

COLLAPSED ECONOMIC REGIONS

Province	Economic Region	
Newfoundland	3,4 (for 36 months)	
Nova Scotia	1,2 (for 36 months)	
	4,5 (for 36 months)	
New Brunswick	1,2 (for 36 months)	
	4,5 (for 36 months)	
Quebec	1,3 (for 36 months)	
	2,9 (for 36 months)	
	4,7 (for 4 months)	6,7 (for 32 months)
	5,6 (for 4 months)	4,5 (for 32 months)
	0,8 (for 36 months)	
Ontario	0,9 (for 36 months)	
	2,8 (for 25 months)	2,3 (for 11 months)
		7,8 (for 11 months)
	5,6 (for 36 months)	
Manitoba	2,3 (for 36 months)	
	4,5 (for 36 months)	
	6,8 (for 36 months)	
Saskatchewan	2,3 (for 12 months)	1,2 (for 24 months)
	3,4 (for 24 months)	1,4 (for 12 months)
	5,6 (for 36 months)	
Alberta	1,2 (for 27 months)	1,2 (for 9 months)
	2,4 (for 27 months)	4,5 (for 9 months)
	7,8 (for 36 months)	
British Columbia	1,2 (for 36 months)	
	3,4 (for 12 months)	3,5 (for 24 months)
	6,9 (for 36 months)	
	7,8 (for 36 months)	

APPENDIX 2

CENSUS DIVISIONS CONTAINED IN MORE THAN ONE ECONOMIC REGION

Province	Census Division	Economic Region
New Brunswick	9	1,2
Quebec	7	1,3
	20	3,4
	24	3,5
	26	3,5
	27	3,4
	28	3,4
	29	3,4
	34	4,5
	42	4,6
	47	4,6
	75	6,7
	76	6,7
	84	4,8
	97	3,9
	98	0,2,8
Ontario	18	2,3
	20	3,9
	23	3,7
	26	4,5
	28	4,5
	30	4,7
	43	3,8
Manitoba	17	5,8
	19	1,5,8
Saskatchewan	2	1,2
	6	1,4
	9	4,5
	11	3,4,5
	15	5,6
	16	5,6
	17	5,6
	18	5,6
Alberta	7	5,6
	8	4,5
	10	5,6
	11	3,6,8,
	12	7,8
	13	4,8
	15	4,7,8