Survey Methodology, June 2001
Vol. 27, No. 1, pp. 65-74
Statistics Canada, Catalogue No. 12-001

Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation

Jack Gambino, Brian Kennedy and Mangala P. Singh [1]

Abstract

The Canadian Labour Force Survey (LFS) is a monthly survey with a complex rotating panel design. After extensive studies, including the investigation of a number of alternative methods for exploiting the sample overlap to improve the quality of estimates, the LFS has chosen a composite estimation method which achieves this goal while satisfying practical constraints. In addition, for variables where there is a substantial gain in efficiency, the new time series tend to make more sense from a subject-matter perspective. This makes it easier to explain LFS estimates to users and the media. Because of the reduced variance under composite estimation, for some variables it is now possible to publish monthly estimates where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.

Key Words: Rotating panel survey; Estimation system; Weighting; Change estimate; Level estimate.

1. Introduction

1.1 Why Composite Estimation?

The Canadian Labour Force Survey (LFS) is a monthly survey of 54,000 households selected using a stratified multistage design. Households stay in the sample for six consecutive months, thus five-sixths of the sample is common between consecutive months. Each month, the members of a selected household are asked questions about their labour force status, earnings, and so on. In the LFS estimation system used prior to 2000, initial design weights were modified using regression to produce final weights that respect age-sex and geographical (subprovincial region) *population control totals*. Each record then had a *unique final weight* that is used for all tabulations.

The estimation system used data from the current month only. No attempt was made to exploit the fact that the common sample can be used to improve estimates. However, characteristics such as employment by industry are highly correlated over time and unemployment is moderately correlated over time, thus there is potential for efficiency gains. Because of these gains, surveys similar to the LFS, such as the United States Current Population Survey (CPS), have used composite estimation to improve their estimates for many years. However, the LFS did not introduce composite estimation until January 2000.

In the early 1980s (see Kumar and Lee 1983), the CPS approach to composite estimation was studied for possible implementation in the LFS. Although the results showed that there were efficiency gains for Employed and, to a lesser extent, for Unemployed, it was felt that these gains were outweighed by the negative aspects of the method. These include the fact that the optimal parameters for Employed and Unemployed are quite different, which would have forced a trade-off between, on the one hand, using a compromise set of parameters, thereby diluting the efficiency gains, and, on the other hand, having variables that do not *add up to totals* (*e.g.*, Employed plus Unemployed would not equal Labour Force, unless one of the three is obtained as a residual). Another factor that worked against this form of composite estimation was that it was not *compatible with the existing weighting, estimation and dissemination systems* used by the LFS – the introduction of composite estimation would have required a complete overhaul of these systems.

Traditionally, the key estimates produced by the Labour Force Survey were monthly unemployment rates. However, with the increasing emphasis on estimates of employment level and on estimates of change in recent years, the need to find ways to make use of the common sample also increased since these estimates would benefit significantly. In the mid-1990s, therefore, interest in composite estimation was revived at Statistics Canada, and a regression-based method that fit in well with the existing LFS estimation system was developed. This method is described in Singh, Kennedy, Wu and Brisebois (1997) with a more up to date version included in Singh, Kennedy and Wu (2001). The new methodology allows for a choice of methods, depending on one's objectives. If the primary interest is in estimates of level, then one can use level-driven predictors in the procedure. If change is most important, then change-driven predictors can be used. One can go one step further and include both types of predictor in the procedure. However, in the latter case, the number of independent variables in the regression becomes large, which can lead to distortion of the final sample weights.

Preliminary results based on the new method using change-driven predictors and others using level-driven predictors were discussed at meetings of Statistics Canada's Advisory Committee on Statistical Methods. The method

1. Jack Gambino, Brian Kennedy and Mangala P. Singh, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

addressed the problems with traditional composite estimators and showed substantial gains in efficiency. It was noted, however, that the estimator using change-driven predictors may lead to a drift in level estimates over time in some extreme situations. Also, it was decided, based on the committee's recommendation, that both estimates of level and of change should be given importance in the choice of predictors. After the exchange of technical notes between Wayne Fuller, J.N.K. Rao and Statistics Canada staff, a method suggested by Fuller, that combines the change-driven and level-driven approaches without the constraints associated with including both sets of predictors in the regression was adopted (see Fuller and Rao 2001). The solution is remarkably straightforward: take a linear combination of the level and change predictors: $X = (1-\alpha)X_L + \alpha X_C$, and use it as the predictor. The change- and level-driven predictors are now special cases. Furthermore, one can choose α to reflect the relative importance one wishes to give to level versus change.

The present paper describes the new composite estimator in section 2. An extensive evaluation of this estimator was carried out using actual LFS data for a number of characteristics over a long period of time. The results of these studies are summarized in section 3. Unlike traditional composite estimators, the regression based composite estimator requires that the matching of the sample between two consecutive months be done at the individual record level. This creates some interesting situations where one has to deal with nonrespondents and in scope and out of scope individuals between two consecutive months in such away that the quality of estimates of change is not compromised. Section 4 discusses the imputation procedure developed to deal with various situations that arise when dealing with incomplete data for two consecutive months. Finally, the success of this new composite estimator is judged not only on its statistical efficiency but its stability over time and its cost effectiveness, while achieving the following objectives: (i) minimizing changes to the old estimation system, (ii) producing a unique weight for each sample unit (iii) respecting age-sex and geography control totals and (iv) producing consistent estimates (in the sense that, *e.g.,* Employed + Unemployed = Labour Force and Labour Force + Not In Labour Force = Population 15+). These objectives are discussed at various points in the paper, but especially in section 3.

2. The Regression Composite Estimator

Surveys such as the United States Current Population Survey have exploited their sample overlap by using *K*-composite or *AK*-composite estimators. Initially, the CPS used the *K*-composite estimator

$$y'_t = (1-K)y_t + K(y'_{t-1} + \text{change}_{t-1,t})$$

with $K = 1/2$ for time *t*, where $\text{change}_{t-1,t}$ denotes an estimate of change based on the common, or matched, sample. This was later replaced by the *AK*-composite estimator

$$y'_t = (1-K)y_t + K(y'_{t-1} + \text{change}_{t-1,t}) + A(\text{unmatched} - \text{matched})$$

with $A = 0.2$ and $K = 0.4$ (see Cantwell and Ernst 1992). The optimal values of A and K depend on the variable of interest, and using different values for different variables poses problems of consistency (in the sense that parts do not add up to totals) in this approach. This prompted us to look for alternative approaches that satisfy the objectives mentioned at the end of the previous section.

It should be noted that we describe the new approach here at the person level, but in practice, person-level information is aggregated to the household level, and household-level records are then used by the estimation system.

To use regression for weighting in the old LFS estimation system, a regression matrix X is formed. Each person in the sample corresponds to a row of X. Each column of X corresponds to a control total; *e.g.*, column *c* may be Male 20-24, and the value in row *i*, column *c* will equal 1 if person *i* is a male between the ages of 20 and 24, and 0 otherwise (similarly for columns corresponding to geographical areas). For further details on the estimation methods used by the Labour Force Survey, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998).

To exploit the sample that is common between months, the X matrix is augmented by columns whose elements are defined in such a way that when this month's final weights are applied to the elements of each new column, the total is a composite estimate from the previous month, *i.e.*, last month's composite estimate is used as a control total (strictly speaking, the control total is based on weights that reflect the current month's population). As we noted in the introduction, there are several ways to define the new columns, depending on one's objectives. We present below only the alternatives that were proposed for implementation.

A typical new column will correspond to employment in some industry, say agriculture. If one is primarily interested in estimates of level, the following way of forming columns produces good results. Let M and U denote the matched (common) and unmatched (birth) sample, respectively. For person *i*, and times $t-1$ and *t*, let $y_{i,t-1}$ and $y_{i,t}$ be indicator variables which equal 1 whenever the person was employed in agriculture. Then let

$$x_i^{(L)} = \begin{cases} \bar{y}'_{t-1} & \text{if} \quad i \in U \\ y_{i,t-1} & \text{if} \quad i \in M, \end{cases}$$

where $\bar{y}'_{t-1}$ is last month's composite estimate of the proportion of people employed in agriculture; in practice, we use $\bar{y}'_{t-1} = \hat{Y}'_{t-1} / P_{15+}$, where P_{15+}, denotes the population aged 15 and over. The corresponding control total is last month's estimate of the number of people employed in

agriculture, *i.e.*, $\hat{Y}'_{t-1}$. Thus the end result is that the final weighted sum of the elements of the new column will equal last month's estimate. This is almost the same as forcing this month's weights, applied to last month's values for the common sample, to reproduce last month's estimate of employment in agriculture (after adjusting by 5/6). We have used the superscript L as a reminder that the goal here is to improve estimates of level.

If interest lies primarily in estimates of change, the following way of forming new columns of X produces good results:

$$x_i^{(C)} = \begin{cases} y_{i,t} & \text{if} \quad i \in U \\ y_{i,t} + R(y_{i,t-1} - y_{i,t}) & \text{if} \quad i \in M, \end{cases}$$

where R is a ratio that adjusts for the fact that five-sixths of the sample between months is common. The value $R = \sum_{\text{all}} w_i / \sum_{\text{matched}} w_i$ is used in the production system. For convenience, we used R = 6/5 during development since, in practice, the difference between the two is small because procedures to balance the weights by rotation group are used (*e.g.*, nonresponse adjustment is done separately by rotation group). As before, the corresponding control total is last month's estimate of the number of people employed in agriculture. Applying the final weights to the elements of this column of the X matrix and summing produces the equality

$$\hat{Y}'_{t-1} = \hat{Y}'_t - \hat{\Delta}^{M,f}_{t-1,t},$$

or, in words, last month's estimate equals this month's estimate minus an estimate $\hat{\Delta}$ of $Y_t - Y_{t-1}$ based on the common sample. We use the superscript f in $\hat{\Delta}$ as a reminder that the estimate of change is based on the final weights following composite estimation. In terms of the "pre-composite" weights, it is easy to show in the univariate case that

$$\hat{Y}'_t = (1-b)\hat{Y}_t + b(\hat{Y}'_{t-1} + \hat{\Delta}^{M}_{t-1,t,}),$$

where b is the regression coefficient and $\hat{\Delta}$ is the estimate of change based on the original weights. The more general case where auxiliary variables are present is given by Fuller and Rao (2001, equation 2.3).

Earlier results have shown that using the L controls produces better estimates of level for the variables added to the X matrix as controls. Similarly, adding C controls produces good estimates of change for the variables that are added. Singh *et al.* (1997, 2001) present efficiency gains for C-based estimates of level and change and refer to earlier results on L-based estimates.

Early in the development, an estimation system that used only the C-based controls was considered. However, there was some concern expressed about an estimation system based solely on change-driven controls since estimates of level are also very important (for example, they play a key role in the federal government's Employment Insurance program). These concerns are summarized in Fuller and Rao (2001).

In principle, we can add both L and C controls to the regression, but this would result in a large number of columns in the X matrix, which has undesirable consequences such as an increased number of extreme final weights, including negative weights. To avoid this, we would have to limit the number of industries included in the estimator. Wayne Fuller (see Fuller and Rao 2001) proposed an alternative which allows us to include the industries of greatest interest while allowing us to compromise between improving estimates of level and improving estimates of change. Fuller's alternative is to take a linear combination of the L column and the C column for an industry and use it as the new column in the X matrix, *i.e.*, use

$$x_i = (1-\alpha)\, x_i^{(L)} + \alpha\, x_i^{(C)}.$$

The original level- and change-driven variables are special cases of Fuller's compromise.

Choice of α: Fuller and Rao (2001) showed that, based on some reasonable assumptions, values of α such as 0.65 and 0.75 produce reasonable estimates of both level and change. The actual choice of α depends on the variable of interest (specifically, its correlation over time) and on the relative importance of level versus change.

Our studies (see Appendix 1) showed that for the two most important variables, employed and unemployed, the best choices of α for estimates of level are 0.39 and 0.24, respectively. The corresponding values for estimates of change are 0.99 and 0.81, respectively. Clearly, there is a need to compromise between the goals of improving estimates of level and estimates of change.

To decide which values of α to study, we obtained compromise values of α by averaging the level-driven and change-driven values for each variable, *i.e.*, we obtained approximately 0.7 and 0.52 for employed and unemployed, respectively. Results based on the values α = 1 and α = 0.75 had already been produced, so we added results for α = 0.67 and α = 0.6. Based on the results discussed below, which show that there are no substantial differences in the results for the three values 0.6, 0.67, and 0.75, we chose to implement the value α = 2/3 in the production system.

3. Features, Properties and Results

We present a summary of some of the features and properties of the regression composite estimator. Some graphical and numerical results are presented in section 3.1 below.

Systems implementation. An important advantage of the estimator is that it can be implemented within the old LFS estimation system in a straightforward manner since, essentially, one needs to augment the regression matrix, as described above. This was an important factor in our initiative to study and finally introduce composite estimation as

otherwise it would have cost a great deal more to change the system.

Weighting. Unlike the *A-K* estimator, where weighting to satisfy population control totals and composite estimation are separate steps, weighting for the regression composite estimator is done in one step, *i.e.*, simultaneously with weighting to satisfy the age-sex and geographical controls. For illustration, the way the regression matrix would be augmented when elements $x_i^{(C)}$ defined in section 2 are added is shown in Appendix 3. Adding the elements $x_i = (1-\alpha)x_i^{(L)} + \alpha\, x_i^{(C)}$ is similar. This not only preserves the consistency mentioned next but also retains the benefits of the controls applied to the usual regression estimator, *i.e.*, the age-sex and geographic controls in our case.

Consistency. Because weighting for age-sex and geographical controls is done at the same time as weighting for the composite estimate controls, consistencies are preserved. In particular, parts add up to totals; *e.g.*, Employed + Unemployed = Labour Force. In other approaches to composite estimation, consistency is achieved by other means which require either a separate step or a compromise of some kind.

Efficiency gains. For the variables that are added as control totals, there are substantial gains in efficiency for both estimates of level and of change. For $\alpha = 1$, the gains for estimates of change can be dramatic; by choosing a smaller value of α we gain more for estimates of level while reducing the magnitude of the gains for estimates of change. Some results for the case $\alpha = 2/3$ are given in section 3.1.

Seasonal adjustment. The time series of employment by various industries are scrutinized by both internal and external users of the Labour Force Survey. One important consequence of the gain in efficiency is that several of these series which could not be seasonally adjusted in the past can now be seasonally adjusted. In other words, composite estimation increases the signal-to-noise ratio sufficiently that seasonal adjustment becomes effective. A related consequence of composite estimation that is popular with users is that several estimates that were published as three-month moving averages are now published as monthly estimates.

Systematic differences between composite and usual level estimates. In theory, the expectations, taken over all possible samples, for both the usual and composite estimators should be the same, making them both unbiased or almost unbiased. One would therefore expect that the estimates of level obtained using the two estimators would criss-cross each other over time. In practice, however, this does not happen. This is due to the fact that, when actual survey conditions are taken into account, the composite estimator and the usual estimator do not have the same expected value; for example, see Bailar (1975) and Kumar and Lee (1983) for results on the *K*- and *AK*-composite estimator, respectively. Kumar and Lee show this by deriving explicit expressions for the expected value of the usual estimator and the *AK*-composite estimator. The matched and unmatched samples differ because of differences in nonresponse rates and the mode of data collection (*e.g.*, personal versus telephone interviewing, centralized versus decentralized interviewing). In practice, the units in the "birth" sample have a higher nonresponse rate, and the missing households tend to be smaller and have higher employment rates than the responding ones. Since the usual estimator and the composite estimator give different weights to the matched and unmatched sample, they will have different expected values. Thus time series for the two estimators can display systematic differences. In practice, these differences are usually swamped by sampling variation, but they become evident for more precise series such as Employed for big provinces like Ontario and for Canada. Our results for Employed are consistent with those described by Bailar (1975) for the U.S. Current Population Survey, *i.e.*, the composite estimates for Employed tend to be smaller than the usual estimates. For Unemployed in Ontario, the difference between the two types of estimates tends to be much smaller.

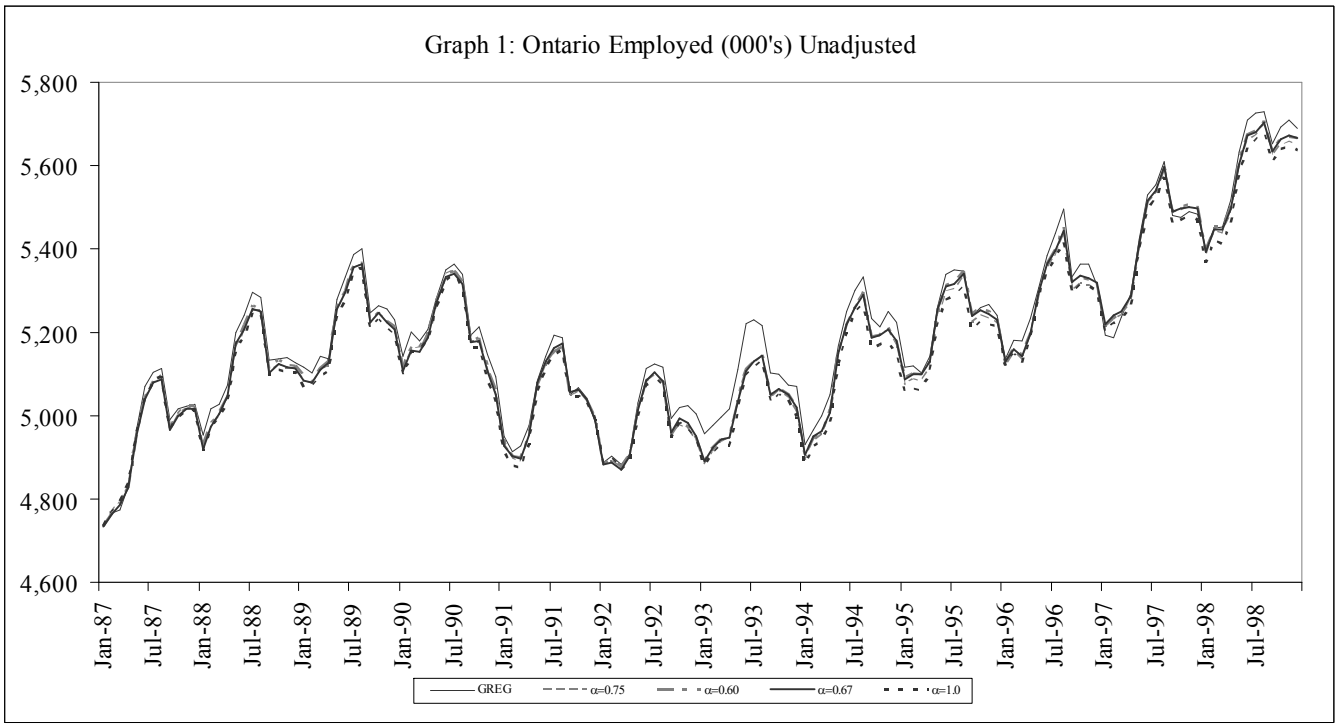
Ways of reducing systematic differences between estimates from different rotation groups are currently being investigated. In particular, the possibility of introducing a weight adjustment for the number of households of different sizes by rotation group is being studied as a way of adjusting for the fact that small households are under-represented in the birth rotation. This would benefit both the composite estimators and the usual regression estimator, and would probably reduce the gap between them.

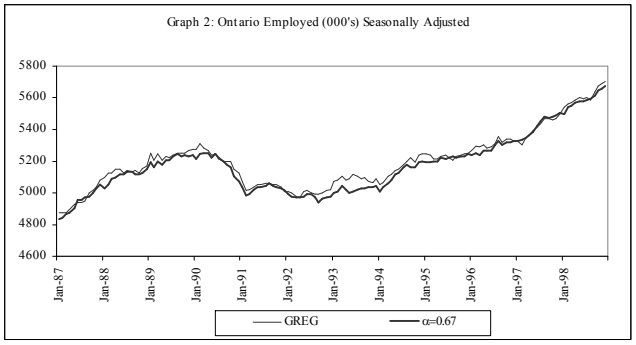
3.1 Empirical Results

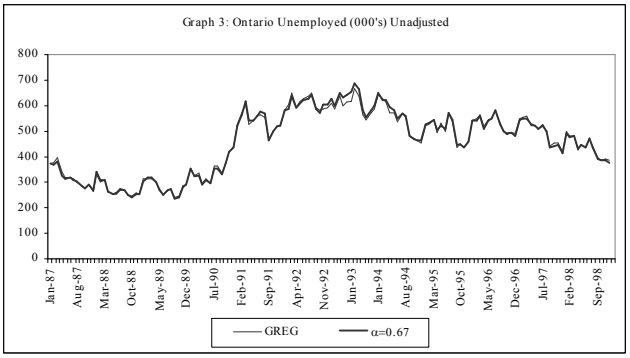
Employment and unemployment at the provincial level. Graph 1 shows total employment at the province level from 1987 to 1998 for Ontario. The time series for the composite estimation series for the four values of α, *i.e.*, for 0.6, 0.67, 0.75 and 1 behave similarly. In these graphs, it is clear that there is a change in level for this series under composite estimation – the estimated number of employed persons is lower. The seasonally adjusted versions of the Ontario employment series based on the usual estimator and on the composite estimator for $\alpha = 0.67$ are shown in Graph 2.

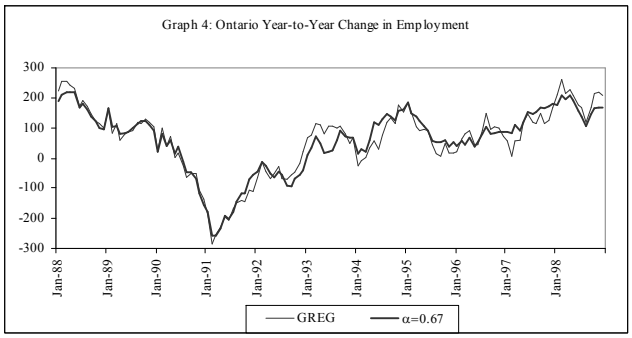
Graph 3 compares the usual estimates of Ontario unemployed to the regression composite estimate for $\alpha = 0.67$. The effect of composite estimation on this variable is clearly less pronounced than on employment-related variables.

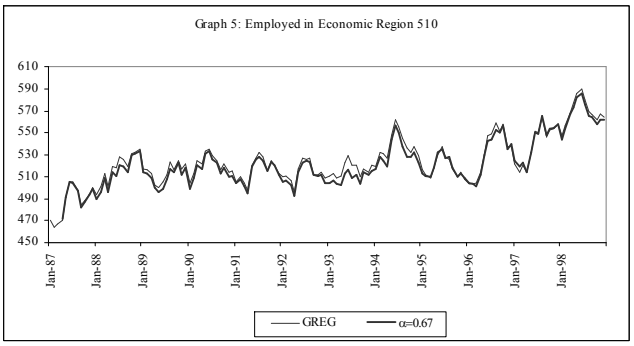
Graph 4 compares year-to-year changes in Ontario employment for the two estimators. Each point in the series is the difference between employment in year y, month m and year $y-1$, month m. For example, the first point is January 1988 employment minus January 1987 employment. The composite estimation series is clearly smoother, especially in the second half of the twelve-year period.











Employment by subprovincial region. Graph 5 compares the usual estimate of employment with the composite estimate with $\alpha = 0.67$ for an economic region in Ontario. The results for other subprovincial regions are similar. The behaviour of the usual and composite estimate series are very similar, thus, the effect of composite estimation is neither beneficial nor harmful. For special tabulations, the LFS estimation system has the flexibility to allow the user to add controls at the economic region level if needed. There is already a control for the total population in each economic region.

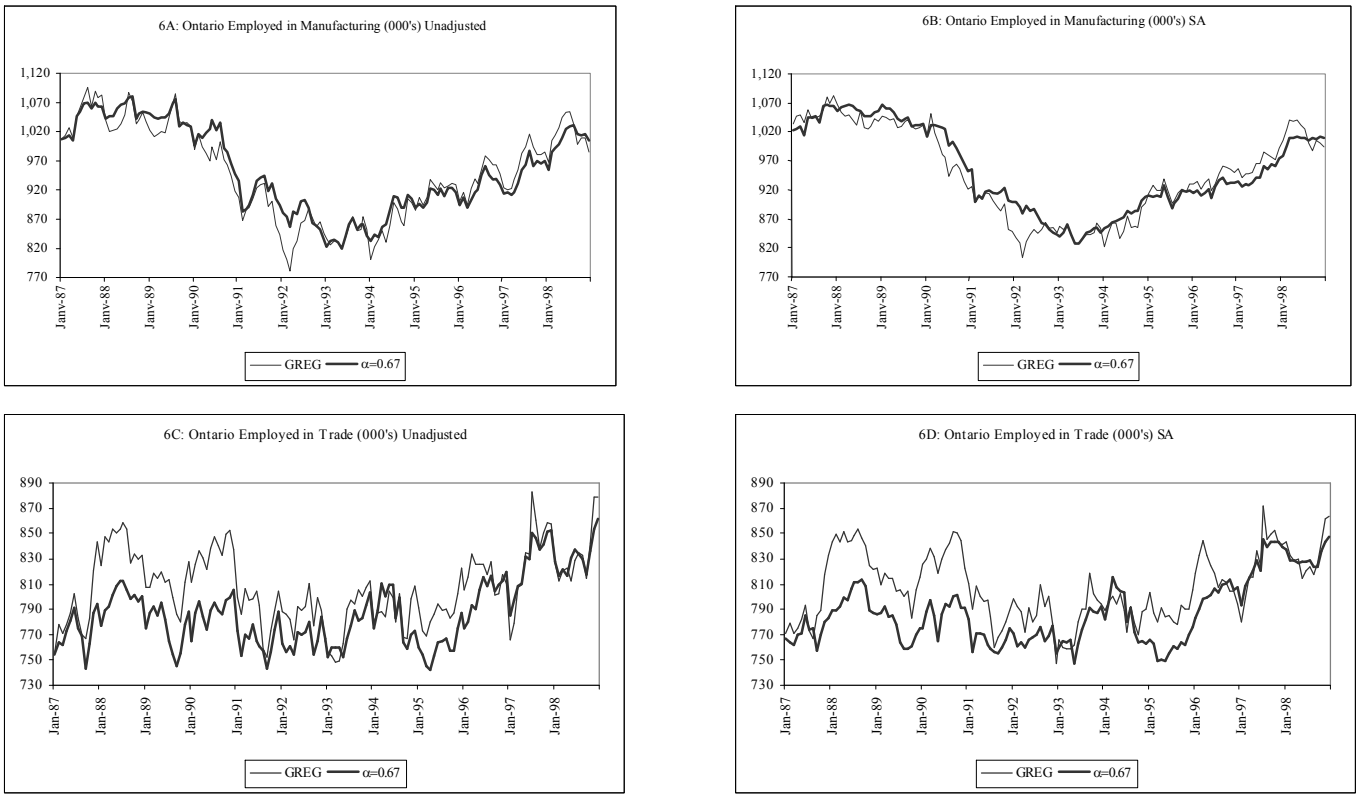
Employment by industry, and seasonal adjustment. The composite estimates were compared to the usual regression estimate for sixteen industries. Graph 6A-6D show the results for two of them in Ontario. Though not included in these graphs, once again, the four values of α result in composite estimation series that generally behave similarly, although sometimes the series for $\alpha = 1$ departs from the others. The composite estimation series tend to be less volatile than the regression series. This is particularly noticeable for the seasonally adjusted Trade series which we have included here because it illustrates the most extreme case. For this series, the behaviour of the original regression estimates in the first few years, in both the seasonally adjusted and unadjusted series, is difficult to explain from a subject-matter viewpoint. The behaviour of the Manufacturing series is more typical of the remaining fourteen industries.

Comparing the seasonally adjusted (Graph 6D) and unadjusted (Graph 6C) series for Trade, we see that seasonal adjustment has had relatively little effect on the regression series, but has changed the composite series significantly. This is a manifestation of the ability of composite estimation to increase the signal-to-noise ratio sufficiently to make seasonal adjustment effective.

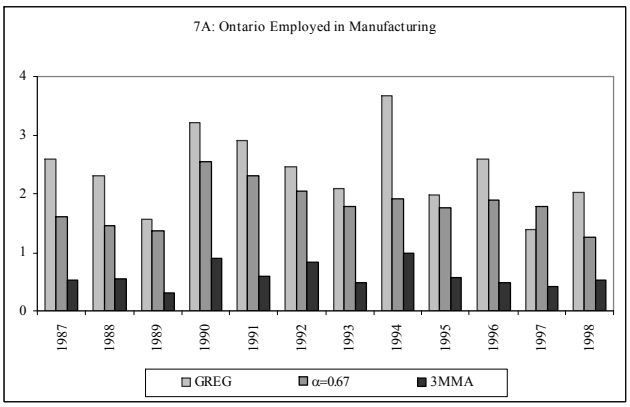
The seasonal adjustment program used by the Labour Force Survey computes a variety of measures that are used as indicators of the effectiveness of seasonal adjustment. Some of these measures are presented in Appendix 2. These show that, for Ontario employment in the twelve-year period 1987-1998, composite estimation increases the number of industries that can be successfully seasonally adjusted. Results for other provinces and for Canada as a whole are similar.

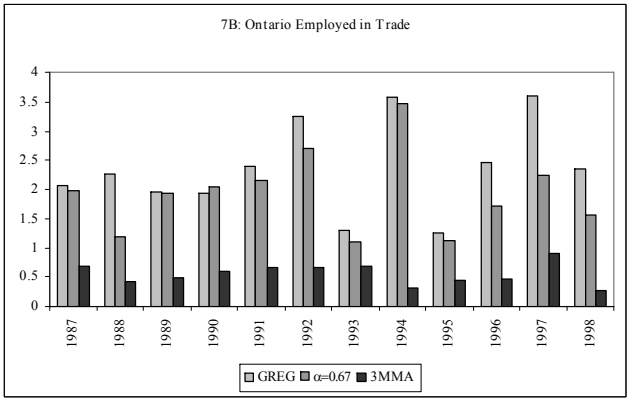
A measure of stability. For several important data series, instead of monthly estimates, three-month moving averages were published in the past. This was due to the high sampling variability associated with these series, leading to unacceptable volatility in the monthly series. Of particular interest are province-level estimates by industry and by class of worker. It had been anticipated that the composite estimates for these series would demonstrate more stability, allowing the publication of monthly estimates instead of three-month averages. A measure of stability, the index of volatility, is computed as follows. For each industry, the month-to-month change in employment is calculated from seasonally adjusted estimates. Then the difference between consecutive change estimates is computed. The absolute value of this "change in the change" is expressed as a percentage of the corresponding monthly total estimate. These percentages are then averaged over the entire year. Large values of this measure occur when a series has many consecutive movements in opposite directions, indicating volatility.

The volatility index was computed for sixteen industries. Graphs 7A and 7B for two of these industries, Ontario Manufacturing and Trade, are included here, comparing the usual estimator, the three-month moving average of the usual estimator and the montly composite estimator with



Graph 6. Selected Employment by Industry.





$\alpha = 0.67$. For Manufacturing, the average indeces for the usual, composite and moving average estimates are 2.4, 1.8 and 0.60, respectively. For Trade, the corresponding values are 2.4, 1.9 and 0.55. For all industries, the volatility of the composite estimates typically falls between that of the usual monthly and three-month average estimates. Occasionally, for isolated years, the composite estimates are less volatile than the three-month averages or more volatile than the usual monthly estimates, but generally the volatility of the composite estimates is between that of the usual monthly estimates and that of the three-month moving averages. We also note that when the usual monthly estimates exhibit extreme volatility, the composite series tend to be more stable. The monthly regression estimates compete with the composite estimates only when the volatility index is low for both of them.

With the introduction of composite estimation, three-month moving averages were dropped in favour of the more desirable monthly estimates for industry series.

Variance estimates. For variables that are added as control totals, such as employment by industry, there can be substantial gains in efficiency at the province level, where efficiency is defined as Var(greg)/Var(composite). For most industries, gains of 10 to 20 percent are typical, but they can be as a high as 40 percent. A 40 percent efficiency gain corresponds, for example, to reducing a 15 percent coefficient of variation to 12.7 percent and a 10 percent coefficient of variation to 8.5 percent. For province-level employment and unemployment estimates, the efficiency gains are more modest, typically in the five to ten percent range. For estimates of month-to-month change, the efficiency gains for controlled variables are bigger, usually more than double the gains for estimates of level.

For variables that are not controlled, there is little or no effect of composite estimation on efficiency unless the variable is highly correlated with a controlled variable. For example, at the province level, Employed Males shows a gain in efficiency because it is correlated with total employed, which is controlled. On the other hand, employment by subprovincial economic region shows neither gains nor losses.

4. Treatment of Missing Data

By definition, the x_i variables involve data from the current and previous month. This leads to complications when, for a given person in the common sample, data is available only for one month. This may occur due to non-response in either month or when a move or change in scope has taken place between the two months. The different cases that may occur are represented in the following diagram, where R denotes a response, X denotes a nonresponse and O denotes a unit that is out of scope.

	A	B	–	C	D
Month t	XXX…	RRR…	RRR…	RRR…	OOO…
Month $t-1$	RRR…	XXX…	RRR…	OOO…	RRR…

In all these cases, namely A, B, C, and D, the objective is to find a solution such that $\sum_{i \in S} w_{it} x_{it}$ is still an estimator of Y_{t-1}. We set the following two objectives for handling the situation of missing data from either month of the common sample:

i) retain as many valid responses as possible, *i.e.*, the option of removing a unit from the estimation process is rejected

ii) develop an imputation method that does not understate the estimate of change in any significant way.

In the case of nonresponse, there are two situations: Case A, where a household responded last month but not this month, and Case B which is the reverse situation. In the following, i denotes a person in an affected household.

Case A: Replace y_{it} by $\hat{y}_{it}$. This can be achieved in a number of ways. A simple approach is to replace y_{it} by the corresponding response from the previous month, *i.e.*, $y_{i,t-1}$. During the early stages of the study, this approach was used but rejected later as it can bias (understate) the estimate of change significantly. For the LFS estimation system, it was decided to use the previous month's known demographic and employment characteristics of persons to

form imputation classes and then use hot deck imputation (*i.e.*, current month's data) to obtain $\hat{y}_{it}$. An alternative would be to use a mean of some sort.

Case B: The procedure is analogous, *i.e.*, when last month's value is missing, then imputation classes are formed using data from month *t* and the donor is found using data from responding units in month $t-1$.

In the case where unit *i* has moved or changed scope, the following situations may arise.

Case C: Suppose that unit *i* was out of scope at time $t-1$ but is in scope at time *t* (*e.g.*, a person who just turned 15, or a newly arrived immigrant). Then unit *i* should contribute 0 at time $t-1$ and y_{it} at time *t*. Hence we let $x_{it}=0$ since $\sum w_{it}x_{it}$ should estimate Y_{t-1}.

Case D: Conversely, suppose that unit *i* was in scope and is now out of scope. This includes, *e.g.*, people who left the country, joined the military or died. Such units should be dropped since the target population is the in-scope population at time *t* (and the ultimate goal is to estimate Y_t). Since we sample dwellings but collect data for individuals within those dwellings, two other situations arise due to movement of persons in and out of the sampled dwellings.

Case i): Suppose that unit *i* was in the population at both times but in a sampled dwelling only at time *t* (*i.e.*, a person who moved from a non-sampled dwelling to a sampled dwelling). Then his/her status at time $t-1$ is unknown, *i.e.*, $y_{i,t-1}$ is unknown. For all such cases, as in the nonresponse case, we can impute a value $\hat{y}_{i,t-1}$ for $y_{i,t-1}$ either from a donor in the sample or by a sample mean. The LFS uses hot deck imputation.

Case ii): Finally, consider the case where unit *i* was in the sample at time $t-1$ but moved to a non-sampled dwelling at time *t*. Since the LFS sample is a sample of dwellings and not a sample of people, this unit should simply be dropped when computing estimates of Y_t.

5. Conclusion

The composite estimator described in this document meets all the objectives that were set at the beginning of this project and summarized in the introduction. It produces estimates of level and change that are more efficient than the estimates produced by the usual regression estimator while satisfying all operational and consistency constraints. The impact of the composite estimator with the value α = 2/3 on the many time series produced by the Labour Force Survey is generally moderate. When the impact is substantial, as in the Ontario Trade series, for example, the new series tend to make more sense from a subject-matter expert's perspective. This type of improvement in the series makes it easier to explain LFS estimates to users and the media.

The composite estimates have other features that users find very desirable. Because of the reduced variance under composite estimation, it is possible to publish monthly estimates in many cases where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.

Implementation of composite estimation for the LFS is an important first step. Studies to improve the treatment of nonsampling errors are ongoing, and their results can be incorporated into the weighting and estimation system at any time. The system has the great advantage that it is very flexible. For example, the value of α can be changed easily, hence a comparison of a broad range of α values for a number of important variables is planned. This may lead to a system in which different α values are used for different control variables, while still having a unique final weight per record.

Acknowledgements

We would like to thank Avi Singh and Statistics Canada's Advisory Committee on Statistical Methods for their contributions to this project. We are also grateful to the many people whose comments on earlier versions of this paper improved it greatly.

Appendix 1

Relationship between α, ρ and (*A*, *K*). Kumar and Lee (1983) found optimal values of *A* and *K* in *AK*-composite estimation for estimates of level and change as a function of the correlation coefficient ρ. We derived an approximate relationship between the *A* and *K* values, ρ and α. This result was then used to find good values of α for several variables. These are presented in Tables 1 and 2 for estimates of level and change, respectively. The *A* and *K* values in the tables are the optimal ones for the corresponding value of ρ. The values of α in the tables are consistent with those obtained by Wayne Fuller based on an AR(1) model (personal communication). The value of α for Labour Force in Table 2 exceeds one because of the approximation.

Table 1
α Values for Several Variables – Level

Variable	ρ	*A*	*K*	α
Employed	0.852	0.49	0.8	0.385
Unemployed	0.580	0.38	0.5	0.242
Labour Force	0.843	0.48	0.8	0.403
E.P. Agriculture	0.955	0.38	0.8	0.448

Table 2
α Values for Several Variables – Change

Variable	ρ	*A*	*K*	α
Employed	0.852	0.1	0.9	0.995
Unemployed	0.580	0.2	0.6	0.806
Labour Force	0.843	0.1	0.9	1.009
E.P. Agriculture	0.955	0.0	0.9	0.959

Appendix 2

Seasonal adjustment measures for Ontario employment by industry

	F Value			M7			SMOOTH	
Industry	greg	$\alpha = 0.60$	$\alpha = 0.75$	greg	$\alpha = 0.60$	$\alpha = 0.75$	greg	$\alpha = 0.60$
Agriculture	87.76	120.18	112.70	0.27	0.23	0.24	37.94	45.36
Forestry	21.34	24.58	23.22	0.50	0.52	0.57	21.76	26.78
Utilities	4.29	3.48	6.80	1.10	1.25	0.82	15.39	15.52
Construction	128.3	275.06	246.93	0.26	0.16	0.17	41.68	57.50
Manufacturing	38.22	55.60	69.21	0.37	0.30	0.30	29.02	31.94
Trade	9.93	15.12	20.35	0.80	0.68	0.53	25.13	34.92
Transportation	9.16	8.64	9.69	0.94	0.75	0.70	15.36	23.33
Finance	6.49	8.94	8.84	1.22	0.76	0.77	13.45	19.67
Professional	5.30	12.91	9.81	1.03	0.72	0.76	12.45	19.52
Management	14.72	24.98	20.35	0.67	0.52	0.52	16.20	22.17
Education	67.37	219.62	214.37	0.33	0.16	0.19	53.25	66.47
Health Care	8.78	10.73	8.48	0.80	0.66	0.75	16.09	19.92
Information	21.13	52.31	62.94	0.66	0.36	0.35	24.29	33.46
Accommodations	44.85	75.37	78.03	0.36	0.34	0.30	31.89	44.29
Other Services	2.61	13.17	12.00	1.41	0.75	0.81	18.58	26.27

Description of Measures

F-value: F-value for the test performed within the X11-ARIMA program to check for the presence of stable seasonality. The higher the F, the more significant is the presence of stable seasonality.

M7: Measure that combines the test for stable and moving seasonality. Generally, when M7 is greater than 1, there is no identifiable seasonality present in the series; therefore, the series should not be adjusted.

SMOOTH: Percentage difference between the standard deviation of the month-to-month changes in the original series and the standard deviation of the month-to-month changes in the seasonally adjusted series. The larger this value, the more smoothing was obtained through the seasonal adjustment process.

Appendix 3

Implementing Regression Composite Estimation within the LFS Estimation Framework: Illustrated Using the Change-driven Approach

Original *X* matrix

Age-sex indicators								Region indicators						
0	0	1	0	.	.	.	0	0	1	0	.	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	.	0
.	.						.	.	.	.				
.	.						.	.	.	.				
.	.						.	.	.	.				
X_1	X_2	.	.	.	.	.	X_k	X_{k+1}	.	.	.	.	.	X_p

← Population control totals

Modified *X* matrix for composite estimation when $x_i^{(C)}$ are added

Age-sex indicators								Region indicators							E	U	Ag	mining				services
0	0	1	0	.	.	.	0	0	1	0	.	.	.	0	a	0	0	b	.	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	.	0	c	0	d	0	.	.	.	0
.	.						.	.	.	.												
.	.						.	.	.	.												
.	.						.	.	.	.												
X_1	X_2	.	.	.	.	.	X_k	X_{k+1}	.	.	.	.	.	X_p	E′	U′	Ag′	.		.	.	.	S′

E′ is last month's employment estimate

For *birth* units, set a, b, c, ... to indicate this month's status (*e.g.*, a = 1 if employed, 0 otherwise). For *matched* units, do the following:

$a = e_t + (e_{t-1} - e_t) \times 6/5$ where e = 1 if person is employed, e = 0 otherwise

$d = ag_t + (ag_{t-1} - ag_t) \times 6/5$ where ag = 1 if person is employed in agriculture, ag = 0 otherwise

Examples:

(i) Suppose Person 2 was employed in agriculture both last month and this month. Then $e_{t-1} = e_t = 1$ and $ag_{t-1} = ag_t = 1$, hence $c = 1 - 0 = 1$ and $d = 1 - 0 = 1$.

(ii) Suppose Person 2 was employed in agriculture last month and in mining this month. Then $e_{t-1} = e_t = 1$, $ag_{t-1} = 1$ and $ag_t = 0$ hence $c = 1 - 0 = 1$ and $d = 0 + (1 - 0) * 6/5 = 1.2$.

(iii) Suppose Person 2 was employed in mining last month and in agriculture this month. Then $e_{t-1} = e_t = 1$, $ag_{t-1} = 0$ and $ag_t = 1$ hence $c = 1 - 0 = 1$ and $d = 1 + (0 - 1) * 6/5 = -0.2$.

References

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Cantwell, P.J., and Ernst, L.R. (1992). New developments in composite estimation for the Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 121-130.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.

Kumar, S., and Lee, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, 27, 33-44.

Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 300-305.