AN APPLICATION OF SMALL AREA ESTIMATION TECHNIQUES TO THE
CANADIAN LABOUR FORCE SURVEY

M.A. Hidiroglou\textsuperscript{1} and Z. Patak\textsuperscript{2}

ABSTRACT

The Canadian Labour Force Survey uses a probability sample based on a stratified multi-stage design. Employment estimates include detailed breakdowns by demographic characteristics, industry and occupation, job tenure, and usual and actual hours worked. We illustrate the use of the Structure Preserving Estimation (SPREE) procedure developed by Purcell and Kish (1980), and its incorporation into the Fay-Herriot (1979) estimator to improve the reliability and smoothness of the two digit occupation series.

KEY WORDS: Modeling, Small Area Estimation, Structure Preserving Estimation (SPREE).

RÉSUMÉ

L’Enquête sur la population active du Canada se sert d’un échantillon probabiliste basé sur un plan de sondage stratifié à deux degrés. Les estimations pour l'emploi sont produites par secteur d’activité et par profession, ainsi que selon la durée de chômage, le genre de travail recherché et le nombre d’heures travaillées. Nous illustrons la méthode Structure Préservant l’Estimation (SPREE) élaboré par Purcell et Kish (1980), ainsi que son incorporation dans l’estimateur de Fay et Herriot (1979) afin de rendre la série des occupations codée à deux chiffres d’améliorer plus lisse et d’en améliorer la fiabilité.

MOTS CLÉS : Estimation pour les petites régions; modélisation; structure préservant l’estimation (SPREE);

1. INTRODUCTION

National statistical agencies are facing increasing demand for statistics below the level, for which most large scale surveys have been designed. The survey methodologists are turning toward Small Area Estimation (SAE) techniques to satisfy the need for reliable estimates for small domains. In 2008, Human Resources and Skills Development Canada (HRSDC) requested that a feasibility study be carried out to evaluate available options for SAE of labour market data. These include improving occupational detail to three-digit National Occupational Classification (NOC) level (or four-digit NOC levels) by province, with the possibility of breaking it down further by sub-provincial area or by industry. If such statistics were of sufficient reliability in terms of bias and variance, they would then be used by individuals (e.g. for employment decisions), by businesses (e.g. for investment decisions), and by governments (e.g. for policy analysis and program planning).

Currently, Statistics Canada publishes estimated counts of employed persons for two-digit occupational code by province on a production basis on CANSIM. These are direct estimates from the Labour Force Survey (LFS) that can be requested for a specific month or for a number of months. The three- and four-digit estimated counts by occupation and province are not published on CANSIM, as many of them are not sufficiently reliable for general consumption.

\textsuperscript{1} Michel A. Hidiroglou, 16-D R. H. Coats Building, 100 Tunney’s Pasture Driveway, Ottawa, ON, K1A 0T6, Mike.Hidiroglou@statcan.gc.ca
\textsuperscript{2} Zdenek Patak, 17-J R. H. Coats Building, 100 Tunney’s Pasture Driveway, Ottawa, ON, K1A 0T6, Zdenek.Patak@statcan.gc.ca
A recent pilot project carried out at Statistics Canada using data from Newfoundland and Labrador indicated that small area procedures could result in more accurate estimates for occupational data for employed persons than those obtained using the more traditional direct estimation procedures such as the ones associated with the Labour Force Survey. The Fay-Herriot (1979) procedure was adapted for this study. The auxiliary data that was used to improve the occupation estimation in this study included population counts, income counts, and employment insurance counts (by three-digit occupation) cross-classified by age, gender, and Economic Region as auxiliary data.

This paper is structured as follows. The LFS design is described in Section 2, and its data used for this study in section 4. Small area procedures for estimating occupational counts are briefly explained in Section 4. The results of their application to the LFS occupational employment data are summarized in Section 5.

2. DESIGN OF LABOUR FORCE SURVEY

The Labour Force Survey uses, in most areas, a two-stage stratified clustered design. The first stage consists of clusters (also known as primary sampling units or PSU’s) of approximately 200 households. The second stage is a systematic sample of households within each cluster. All members of a selected household are selected with certainty. The stratification is based on a set of strata that satisfy the requirements of Statistics Canada as well as Human Resources and Social Development Canada. These strata are the intersection of Economic Regions (ER: Statistics Canada needs) and Employment Insurance Economic Regions (EIER: Human Resources and Social Development Canada needs). Table 1 below provides an example of the design for Newfoundland and Labrador.

<table>
<thead>
<tr>
<th>Province</th>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strata</td>
<td>PSUs</td>
</tr>
<tr>
<td>Newfoundland and Labrador</td>
<td>37</td>
<td>967</td>
</tr>
<tr>
<td>EIER</td>
<td>14</td>
<td>310</td>
</tr>
<tr>
<td>02</td>
<td>23</td>
<td>657</td>
</tr>
<tr>
<td>ER</td>
<td>1010</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>1020</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1030</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1040</td>
<td>7</td>
</tr>
</tbody>
</table>

Each sampled dwelling remains in the Labour Force Sample for six consecutive months. Hence, the maximum overlap of the sample between two consecutive months is 5/6 and 1/6 of the sample of dwellings are replaced each month. Given that the total \( Y = \sum_{h=1}^{L} \sum_{c \in U_h} \sum_{1 \in U_{hc}} y_{hc} \) for a characteristic \( y \) is required, where \( L \) is the total number of strata in Canada, \( U_h \) is the universe of clusters (c) within those strata, and \( U_{hc} \) is the set of universe households (1) within a given cluster \( c \). This total is estimated as

\[
\hat{Y} = \sum_{h=1}^{L} \sum_{c \in s_h} \sum_{1 \in s_{hc}} \theta_{hc} y_{hc}
\]

where \( s_h \) is the set of sampled clusters within stratum \( h \), and \( s_{hc} \) is the set of sampled households within a given cluster \( c \). The weight \( \theta_{hc} \) reflects modifications to the original sampling weight that include non-response adjustments, the use of auxiliary counts based on demographic projections, and last month’s values for paired units (composite estimation).

Note that we can write \( \hat{Y} \) as \( \hat{Y}(y_{hc}) \) where \( \hat{Y}(y_{hc}) \) means that \( \hat{Y} \) depends only on the \( y_{hc} \)'s. The estimated variance for \( \hat{Y} \) is obtained using the jackknife procedure. We require the estimator \( \hat{Y}_{(h'c')} \) for each \( (h'c') \) obtained from the sample
after omitting the data from the $c'$-th sampled cluster in the $h'$-th stratum \( (c' = 1, K, n_{h'}; h' = 1, K, H) \). It is obtained from $\hat{Y}$ as follows by redefining the design weight $w_{hc1}$ as $w_{hc1(h'c')}$ where

\[
 w_{hc1(h'c')} = \begin{cases} 
 0 & \text{if } (hc) = (h'c') \\
 \frac{n_{h'}}{n_{h'} - 1}w_{hc1} & \text{if } h = h' \text{ and } c \neq c' \\
 w_{hc1} & \text{otherwise}
\end{cases}
\]

The corresponding $\hat{y}_{hc1(h'c')}$ weight is computed given $w_{hc1(h'c')}$. A new set of $\sum_{h=1}^L n_h$ estimators are computed as

\[
 \hat{Y}_{(h'c')} = \sum_{h=1}^L \sum_{c \in c_h} \hat{y}_{hc1(h'c')} y_{hc1}.
\]

The resulting jackknife estimator of $\hat{Y}$ is

\[
 v_{JACK}(\hat{Y}) = v_{JACK}(\hat{y}_{hc1}) = \sum_{h=1}^L \sum_{c \in c_h} \frac{n_{h'} - 1}{n_{h'}} \sum_{c' \in c_{h'}} \left( \hat{Y}_{(h'c')} - \hat{Y} \right)
\]

Note that for a specific domain $i$, we replace $y_{hc1}$ by $y_{hc1i}$ where $y_{hc1i} = y_{hc1}$ if the unit 1 belongs to domain $i$ and 0 otherwise.

### 3. DESCRIPTION OF LFS DATA

Unemployment rates are produced on a monthly basis in Canada by the Labour Force Survey (LFS). The LFS samples some 53,000 households based on a stratified multi-stage design. The survey reduces response burden by having one-sixth of its sample replaced each month. The published provincial and national estimates unemployment rates are a key indicator of economic performance in Canada. For a detailed description of the LFS design, see Methodology of the Canadian Labour Force Survey (2008).

In 2008, Human Resources and Skills Development Canada (HRSDC) requested that a feasibility study be carried out to evaluate available options for small area estimation (SAE) of labour market data. These include improving occupational detail at the three-digit level (or four-digit levels) by province, with the possibility of breaking it down further by sub-provincial area or by industry. The Census of Population (2001 and 2006) occupational data were used as an “auxiliary” source of data to strengthen the quality of the direct estimates from the Labour Force Survey. The proportions of the three-digit counts within their associated two-digit occupation at the province level between the LFS and the Census (2001) were plotted (see Matrix Plot 3.1). The plots suggest that the associated regression fit is linear with a slope of one and passes through the origin.

### Matrix Plot 3.1: Three-digit occupations proportions within two-digit occupations

<table>
<thead>
<tr>
<th>Reference period</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontario</td>
<td>Prince Edward Island</td>
</tr>
</tbody>
</table>
4. SMALL AREA ESTIMATION FOR THE LFS OCCUPATIONAL DATA

The three-digit occupation counts within a province were obtained by multiplying the corresponding two-digit counts by the proportion of these counts obtained via a small area procedure. The steps are briefly summarized as follows:

- Use the Structure Preserving Estimation procedure proposed by Purcell and Kish (1979) to estimate the three-digit occupation counts within each province. The procedure adjusts the association structure (interior of the original table from the Census) to conform to the current data in the allocation structure (updated margin as reliably estimated by the Labour Force survey), while preserving (as much as possible) the relationships between the variables in the association structure, (as established in a previous census).
- In our application, the three-digit occupational code by province counts within the parent two-digit occupational code are prorated using Census data (2001, or 2006) to the corresponding two-digit occupation by province and national three-digit occupation counts available from the Labour Force Survey at some other time point. This procedure ensures that the three-digit occupation SPREE estimates add up to both the corresponding two-digit occupation counts by province or the corresponding national three-digit occupation counts obtained by the Labour Force Survey. Griffith (1996) used the SPREE to estimate demographic characteristics for congressional districts using Current Population Survey data.
- Combine the resulting SPREE derived proportions with those obtained from the Labour Force Survey.
- Multiply the combined proportions by the two-digit counts for each province.

The procedure can be represented algebraically as follows. Suppose that the doublet \((i,k)\), where \(i=1,\ldots, I\) and \(k=1, K, K\), denotes the cell for which we wish to have three digit occupation counts within province \(k\). The parameter of interest is...
the count of employed persons with a given occupation within province $k$, say $M_{ik}$. It is estimated from the Labour Force Survey as

$$ \hat{M}_{ik} = \sum_{h \in C, c \in c_h} \sum_{l \in l_h} \hat{w}_{hcl} I_{hel,ik} , $$

(1)

where $I_{hel,ik}$ is equal to one if the unit $l$ belongs to the doublet $(i, k)$, and $\hat{w}_{hcl}$ is the final LFS weight. The $h$ and $c$ subscripts indicate to which cluster and stratum of the LFS that the $l$-th individual belongs. Note that $i$ is a domain whereas $k$ is the union of the design strata $h$ that belong to province $k$. We assume that the marginal estimates $\hat{M}_{\cdot k} = \sum_{k=1}^{K} \hat{M}_{ik}$ and $\hat{M}_{\cdot \cdot} = \sum_{i=1}^{I} \hat{M}_{ik}$ are reliable. The SPREE of $M_{ik}$, denoted as $\hat{T}_{ik}$, are obtained by “raking” the corresponding Census counts $N_{ik}$ (2001, 2006) to the reliable marginal totals $\hat{M}_{\cdot k}$ and $\hat{M}_{\cdot \cdot}$. The estimated direct survey and SPREE derived proportions are respectively

$$ \hat{p}_{ik} = \left( \sum_{i=1}^{I} \hat{M}_{ik} \right)^{-1} \hat{M}_{ik} $$

and

$$ \hat{p}_{T,ik} = \left( \sum_{i=1}^{I} \hat{T}_{ik} \right)^{-1} \hat{T}_{ik} . $$

The value $\hat{p}_{T,ik}$ results from the SPREE between the marginal LFS counts, and the available census counts (i.e.: 2001, 2006). The “optimal” linear combination between $\hat{p}_{ik}$ and $\hat{p}_{T,ik}$ is defined as $\hat{p}_{\text{pred},ik}$ where

$$ \hat{p}_{\text{pred},ik} = \lambda_{ik} \hat{p}_{ik} + (1-\lambda_{ik}) \hat{p}_{T,ik} . $$

Fuller (2008) suggested a procedure to estimate $\hat{p}_{\text{pred},ik}$ using the two equations linked to the Fay-Herriot (1979) procedure. The first equation, survey model, links the direct survey estimated proportion $\hat{p}_{ik}$ to the true but unknown population proportion $p_{ik} = \left( \sum_{i=1}^{I} M_{ik} \right)^{-1} M_{ik}$ of the three digit occupation within the corresponding two digit occupation for province $k$. It is given by

$$ \hat{p}_{ik} = p_{ik} + e_{ik} , \quad i = 1, \ldots, I \text{ for } k = 1, \ldots, K . $$

(3)

The second equation, the linking model, relates $p_{ik}$ to the SPREE, but unknown proportion $p_{T,ik}$. The proportion $p_{T,ik}$ results from what would have been the SPREE between the true marginal LFS counts and the corresponding Census counts. It is given by

$$ p_{ik} = p_{T,ik} + v_{ik} ; \quad i = 1, \ldots, I \text{ for } k = 1, \ldots, K . $$

(4)

Combining the above two equations in vector form, for each province $k$, we obtain:

$$ \hat{p}_{k} = p_{T,\cdot k} + v_{\cdot k} + e_{\cdot k} , $$

(5)

where $\hat{p}_{k}^\top = (\hat{p}_{1k}, \hat{p}_{2k})$, $p_{T,\cdot k}^\top = (p_{T,1k}, p_{T,2k})$, $e_{\cdot k}^\top = (0, \Sigma_{\cdot ek})$ and $v_{\cdot k}^\top = (0, \Sigma_{\cdot vk})$.

In our application, the estimated covariance matrix $\hat{\Sigma}_{\cdot ek}$ of the sampling errors $e_{\cdot k}$ was obtained directly from the jackknife variance system of the Labour Force Survey. Smoothing of these estimated variances is useful to get around the assumption that the terms of the covariance matrix $\Sigma_{\cdot ek}$ are known: this was achieved by multiplying the average design
effect and the corresponding variance covariance matrix assuming simple random sampling without replacement. The critical assumption for estimating $\Sigma_{v,k}$, due to Fuller (2008), is given by

$$\Sigma_{v,k} = \left[ \text{diag} \left( p_k \right) - p_k p_k^T \right] \sigma_v^2 = \Gamma_{v,k} \sigma_v^2. \quad (6)$$

The estimated vector of SAE proportions for province $k$ is given by

$$\hat{\mathbf{p}}_{\text{pred},k} = \hat{\mathbf{p}}_{T,k} + \hat{\sigma}_v^2 \hat{\Gamma}_{v,k} \left( \hat{\sigma}_v^2 \hat{\Gamma}_{v,k} + \hat{\Sigma}_{v,k} \right) \left( \hat{\mathbf{p}}_k - \hat{\mathbf{p}}_{T,k} \right) \quad (7)$$

where

$$\hat{\Gamma}_{v,k} = \left[ \text{diag} \left( \hat{\mathbf{p}}_{T,k} - \hat{\mathbf{p}}_{T,k} \hat{\mathbf{p}}_{T,k}^T \right) \right] \quad (8)$$

and $\hat{\sigma}_v^2$ is the smoothed estimate of $\sigma_v^2$. The inversion of the matrix is via the generalized inverse operator “$\cdot$”. If it is further assumed that $\hat{\Sigma}_{v,k}$ is proportional to the multinomial variance, that is

$$\hat{\Sigma}_{v,k} = c_k \hat{\Gamma}_{v,k} \quad (9)$$

for some constant $c_k$, equation (7) becomes

$$\hat{\mathbf{p}}_{\text{pred},k} = \hat{\mathbf{p}}_{T,k} + \hat{\sigma}_v^2 \left( \hat{\sigma}_v^2 + c_k \right)^{-1} \left( \hat{\mathbf{p}}_k - \hat{\mathbf{p}}_{T,k} \right) \quad (10)$$

The $i$-th element of $\hat{\mathbf{p}}_{\text{pred},k}$ can be expressed as

$$\hat{p}_{\text{pred},ik} = \hat{p}_{T,ik} + g_{ik} \left( \hat{p}_{ik} - \hat{p}_{T,ik} \right) \quad (11)$$

where

$$g_{ik} = \left( \hat{\sigma}_v^2 \hat{p}_{T,ik} \left( 1 - \hat{p}_{T,ik} \right) + \hat{\sigma}_v^2 \hat{p}_{T,ik} \right)^{-1} \hat{\sigma}_v^2 \hat{p}_{T,ik} \left( 1 - \hat{p}_{T,ik} \right) \quad (12)$$

The resulting estimated counts are $\hat{y}_{\text{pred},ik} = T_{ik} \hat{p}_{\text{pred},ik}$ where $T_{ik} = \sum_{i=1}^{k} \hat{T}_{ik}$ is the sum of the raked values $\hat{T}_{ik}$ for province $k$.

Berg and Fuller (2009) directly linearise the SPREE estimators by obtaining a quadratic approximation to the objective function that defines the SPREE estimators and expressing the solutions to the resulting approximate optimization problem as linear functions of the direct estimators of the marginal totals. A simpler solution is to Taylor linearize $\hat{T}_{ik} \hat{p}_{\text{pred},ik}$ by assuming that $\hat{p}_{T,ik} = \hat{T}_{ik} / \hat{T}_{ik} \approx \hat{T}_{ik} / \hat{T}_{ik} \approx \hat{T}_{ik} / \hat{T}_{ik} \approx \hat{T}_{ik}$ (i.e.: the proportions of the three digit occupation codes “$i” within a province are similar to those at the Canada level). The estimated variance for $\hat{y}_{\text{pred},ik}$ is then obtained by first re-expressing $\hat{y}_{\text{pred},ik}$ as:

$$\hat{y}_{\text{pred},ik} = T_{ik} p_{T,ik} + (1 - g_{ik}) \left( T_{ik} \hat{p}_{\text{pred},ik} - T_{ik} p_{T,ik} \right) + g_{ik} \left( \hat{M}_{ik} - T_{ik} p_{T,ik} \right) \quad (13)$$

The estimated variance $\hat{\sigma}_{\text{pred},ik}$ is the sum of the estimated variances of $A$ and $B$ where

$$A = (1 - g_{ik}) T_{ik} \left( T_{ik}^{-1} \hat{T}_{ik} + T_{ik}^{-1} \hat{T}_{ik} \right) \quad (14)$$

and

$$B = -g_{ik} T_{ik} v_{ik} \quad (14)$$

where $v_{ik}$ is defined in equation (4). Following Rao (1986), the estimated variance for $A$ is obtained by first defining the variable
\[ z_{hc1k} = (1 - g_{ik}) \hat{T}_k \left( \hat{T}_{i\cdot} a_{i\cdot k} + \hat{T}_{\cdot k} a_{\cdot i k} - \hat{T}_{\cdot \cdot k} \right) + g_{ik} a_{i\cdot k}, \]  

(15)

where \( a_{i\cdot k} = 1 \) if the \( \ell \)-th employed individual belongs to the \( i \)-th three digit occupation and \( k \)-th province, and 0 otherwise: the \( h \) and \( c \) subscripts indicate to which cluster and stratum of the LFS that the \( 1 \)-th individual belongs. The terms \( a_{i\cdot k} \) and \( a_{\cdot i k} \) are similarly defined as \( a_{i\cdot k} \): they refer to presence or absence in the \( k \)-th province or \( i \)-th three digit occupation respectively.

We have that \( \frac{1}{v} \left( \frac{\hat{\mu}_{pred,ik}}{B} \right) = v(A) + v(B) \). The first component \( v(A) \) can now be estimated using the LFS jackknife variance system. That is

\[ v(A) = \sum_{h'=l}^{L} n_{h'} - 1 \sum_{c|c_{h'}} \left( \hat{Z}_{(h',c'),ik} - \hat{Z}_{ik} \right)^2 \]  

(16)

with

\[ \hat{Z}_{ik} = \sum_{h=1}^{L} \sum_{c|c_{h}} \sum_{c|c_{h}} \hat{\mu}_{h,c,k} z_{hc1k} \]

and \( \hat{Z}_{(h',c'),ik} \) is the corresponding jackknifed estimate obtained by deleting the \( c' \)-th cluster within stratum \( h' \). It should be noted that \( i \) is a domain, whereas \( k \) is part of the stratification as a province \( k \) is made up of a subset of the strata \( h \) in Canada. The second component \( v(B) \) is:

\[ v(B) = g_{ik}^2 \hat{\mu}_{i\cdot}^2 \hat{\sigma}^2 \hat{p}_{T,ik} (1 - \hat{p}_{T,ik}) \]  

(17)

5. RESULTS

The computations were applied to the whole LFS data set from May 2001 to May 2006. We illustrate a small subset of the results in matrix plot 3.2 (provincial comparisons in May 2001) and 3.3 (time comparisons) for the three digit occupation Auditors, Accountants and Investment Professionals (B01).

Matrix Plot 3.2: Summary statistics for Proportions and counts for B01 across provinces (May 2001)
Matrix Plot 3.3: Summary statistics for Proportions and counts for B02 within Prince Edward Island across May 2001 to April 2002

Observations from the above plots are:

- Proportions have been smoothed by the SPREE procedure, and corresponding estimated reliabilities are significantly smaller than those obtained from direct estimation: this is particularly true for the smaller provinces.
- There is not much difference between the versions that use the smoothed variances as opposed to the raw variances.
- The SPREE derived counts are not very different from the original direct counts: however, their reliability has been increased. The largest gains occur in the smaller provinces.
- The gain in precision for counts is significantly lower than the corresponding gain for proportions. This is because the gain in precision of the counts is constrained by the precision of their associated marginal counts.
- Similar results hold across the time dimension. There is approximately a 30% gain in reliability in using the SPREED counts as opposed to the direct one for B02 in the small province of Prince Edward Island.

There are some marginal two-digit occupations that have relatively weak reliabilities, and the resulting quality of the three-digit SPREE estimates is weak as well. One way to strengthen the marginal two-digit employed counts of occupation is to cascade the above procedure from one to two-digits, thereby improving the quality of the two-digit marginal employed counts of occupation, and continue with the above procedure. A more detailed version of this section is given by Hidiroglou and Patak (2009).
REFERENCES


