Model-based Approach to Small Area Estimation of Disability counts and rates using Data from the 2006 Participation and Activity Limitation Survey

Valérie Bizier[1] Yong You[2] Lucie Veilleux[3], Chantal Grondin[4]

[1]Statistics Canada, 100 Tunney's Pasture Driveway RHC 17P, Ottawa ON K1A 0T6
[2]Statistics Canada, 100 Tunney's Pasture Driveway RHC 16D Ottawa ON K1A 0T6
[3]Statistics Canada, 100 Tunney's Pasture Driveway RHC 15R, Ottawa ON K1A 0T6
[4]Statistics Canada, 100 Tunney's Pasture Driveway RHC 15O, Ottawa ON K1A 0T6

Abstract

This paper presents the small area model considered to produce estimates of the number of persons with disabilities and disability rates for health regions and selected municipalities using the 2006 Participation and activity limitation survey data. The paper describes the transformations applied to direct estimators and to the variances associated with these estimators in order to meet certain fundamental criteria. The log linear unmatched model to which the hierarchical Bayes (HB) approach was applied by relying on the Gibbs sampling method and the results from the latter model are presented. Lastly, the paper presents the data sampling methodology used to ensure that the final statistics take into account province level results.

Key Words: Small area estimation, Disability counts and rates, Log linear unmatched model, Gibbs sampling method

1. Introduction

The main source of information on adults and children with disabilities, that is to say those whose day-to-day activities are limited because of a condition or health problem, is the Participation and Activity Limitation Survey (PALS). This nation-wide survey financed by Human Resources and Skills Development Canada and conducted by Statistics Canada, provides key information on the prevalence of different types of disabilities, on support provided to people with disabilities, on their labour force profile, their income and their participation in society. However the number of respondents to the survey, approximately 29,000 adults and 7,000 children, does not allow for accurate direct estimates at the sub-provincial level. Following the demands to that effect which were expressed by many provincial governments as well as municipalities, Statistics Canada has put in place a model-based approach to small area estimation for the disability count and rate. This document describes this approach and presents the results.

Note to readers

This paper is based on the Participation and Activity Limitation Survey (PALS). PALS is a post-censal survey that collected information about persons with disabilities whose

everyday activities are limited because of a health-related condition or problem. The survey took place between November 2006 and February 2007. PALS is funded by Human Resources and Skills Development Canada (HRSDC).

PALS is a post-censal survey which used the 2006 Census as a sampling frame to identify its population. The 2006 Census questionnaire included two general questions on activity limitations. The PALS respondents were selected through the use of the census information on age, geography and the responses to these two general questions. The PALS interview began with the census activity limitation filter questions identical to the Census questions followed by a series of detailed screening questions on disability. If respondents answered NO to all of the filter questions and screening questions, the interview ended and the respondent was not considered to be a person with a disability according to PALS. If respondents answered YES to any of the filter questions or screening questions, they were considered disabled. The interview went on to collect information on the impact of that disability on their everyday activities and other aspects of their life, such as education, employment, leisure, transportation and accommodation.

The PALS sample was 48,000, consisting of approximately 39,000 adults and 9,000 children. The sample was selected using a two-phase stratified design where at the first phase, a Census questionnaire was distributed to approximately one out of five persons, and at the second phase, a stratified sample was selected based on characteristics from the first phase. Interviews were conducted by telephone with the interviewers using a computer assisted collection methodology. Two questionnaires were used, one for adults aged 15 and over and one for children under the age of 15. The interviews for the children's questionnaire were conducted with the parent or guardian of the child. The overall response rate was 75.0%.

The population covered by the survey was persons residing in private and some collective households in the ten provinces and three territories. Persons living in institutions and on First Nations reserves were excluded from the survey. PALS 2006 followed the groundwork laid by the Health and Activity Limitation Survey (HALS) in 1991 and the Participation and Activity Limitation Survey of 2001. The data for HALS 1991 and PALS 2001 could not be compared because of significant differences in their sampling designs, the operational definition of their target population and the content of their questionnaires. However, the PALS 2006 results can be compared with the 2001 survey to identify trends in the previous five years. For more information about PALS 2006, see the *Participation and Activity Limitation Survey 2006: Technical and Methodological Report* (89-628-XWE2007001, free), published in December 2007.

2. Basic Model for Small Area

In order to get a basic model for small areas, let's assume that the parameter of interest θ_i for the small region i is related to region-specific auxiliary data $x_i = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)'$ through a linear model

$$\theta_i = x_i'\beta + \nu_i, \; i = 1,..., m, \qquad (1)$$

where m is the number of small areas, $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the px1 vector of regression coefficients, and ν_i are region-specific random effects which are assumed to be independent and identically distributed (iid) with E(ν_i) = 0 and V(ν_i) = σ_ν^2. The normality hypothesis for ν_i is often included. The model is called a linked model for θ_i.

The basic model for small area also assumes that, given the sample size of a specific region $n_i > 1$, there exists a direct survey estimate $\hat{\theta}_i$ (usually design-unbiased) for the regional parameter of interest θ_i such as

$$\hat{\theta}_i = \theta_i + e_i, \; i = 1,..., m, \qquad (2)$$

where e_i are the sampling errors associated with the direct estimators $\hat{\theta}_i$. We also assume that the e_i are normal independent random variables of mean E($e_i \mid \theta_i$) = 0 and of sampling variance V($e_i \mid \theta_i$) = σ_i^2. This model (2) is called the sampling model for the direct survey estimator $\hat{\theta}_i$.

The combination of linking model (1) with sampling model (2) yields a regional mixed linear model, called Fay-Herriot model (Fay and Herriot, 1979)

$$\hat{\theta}_i = x_i'\beta + \nu_i + e_i, \; i = 1,..., m, \qquad (3)$$

which includes the design-based random variables e_i and model-based random variables ν_i. Standard methods, such as the empirical best linear unbiased predictor (EBLUP) method and the hierarchical Bayes (HB) approach using the Gibbs sampling method, can be applied to this model (3) to obtain estimates based on a small area model. These methods assume that σ_i^2 is known, then usually, a smoothed estimator of σ_i^2 is used and treated as known. To do so, one can use a generalized variance function. Among others, this was used to estimate provincial census undercoverage rates in Canada (Dick and You, 1998).

However, the assumption that E($e_i \mid \theta_i$) = 0 may not be valid if the sample size n_i is small and the relationship between θ_i and the auxiliary information available is not necessarily linear, even if the direct estimator $\hat{\theta}_i$ is design-unbiased for θ_i.

Then, it is possible to consider a more realistic linking model

$$g(\theta_i) = x_i'\beta + \nu_i, \; i = 1,..., m \qquad (4)$$

based on a function *g(.)* of θ_i and some regional random effects ν_i. Model (4) and model (1) are called unmatched, since they cannot be combined directly to produce a linear mixed model.

The two models described here were applied to PALS data to obtain small area level estimates. The log-linear unmatched model produced the best results. We will come back to this model in the section on "Small area model used by PALS".

3. Ratio Adjustment of Direct Estimators

Direct estimators for the desired statistics, meaning number of people with disabilities Y_i, and disability rate, p_{yi}, within a region i , can be written as follows:

$$\hat{Y}_i = \sum_{j \in S} y_j w_j \delta_{ij} \qquad (5) \quad \text{and} \quad \hat{p}_{yi} = \frac{\hat{Y}_i}{N_i} = \frac{\sum_{j \in S} y_j w_j \delta_{ij}}{N_i} \qquad (6)$$

where w_j is the survey weight, y_j is a binary variable taking on the value 1 or 0 whether person j has a disability or not, δ_{ij} is a binary variable taking on the value 1 or 0 whether person j in the sample S belongs to region i or not and N_i is the total number of persons living in region i according to the 2006 Census.

With regards to the total number of people living in area i, it is often preferable to take the weighted estimate of this number to preserve the same trend in the numerator and denominator. However in the case of PALS, the survey data did not allow the computation of a direct estimate of the population size; they could only allow the computation of the number of people who reported an activity limitation in the Census. Hence to obtain an estimate of the total population, one must add to this number people who were not part of our target population, that is to say people who did not report an activity limitation in the Census. This number is known and adding it to the sum of the weights in our sample would not have been difficult. However the variance computation using the bootstrap weights for the ratio would have been more complicated. For the survey publications, a sample of people without an activity limitation was selected and from this sample, we computed 1000 series of bootstrap weights to allow the computation of disability rates. This sample however doesn't guarantee a sufficient number of people in each small area. We would have had to select another sample of people without disabilities for which we would have needed to compute 1000 bootstrap samples to estimate the variance, which would have further delayed the project. As well, if we had used the estimated population size, we would have needed two models for small areas: one to predict the disability rates and one to predict the number of disabled people. It would not have been possible to simplify only by a constant the variances of the disability rates direct estimators and their estimates produced by the model to obtain these quantities for the totals.

Since PALS is a post-censal survey where the sample is selected from people who responded "yes" to at least one of the general activity limitation questions in the Census, we expect that for each region i, the number of people with a disability Y_i estimated from the survey be at most equal to the number of people having responded "yes" to at least one of the Census filter questions, M_i. For certain small areas however, the very small available sample sizes and high survey weights made it impossible for this condition to hold. The direct estimators were thus adjusted using a ratio to ensure that the desired

statistics would be consistent with the already known census totals. Hence, direct estimators used for small areas are:

$$\hat{Y}_i^R = \frac{M_i}{\sum_{j\in S} w_j \delta_{ij}} * \sum_{j\in S} y_j w_j \delta_{ij} \quad (7) \qquad \text{and} \qquad \hat{p}_{yi}^R = \frac{\hat{Y}_i^R}{N_i} = \frac{M_i}{\sum_{j\in S} w_j \delta_{ij}} * \frac{\sum_{j\in S} y_j w_j \delta_{ij}}{N_i} \quad (8)$$

These ratio adjusted direct estimators, although they are somewhat biased like any ratio estimator, are clearly much more stable in terms of variance compared to the unadjusted direct estimators. They are also much closer to the true values we are aiming for, since they use known totals M_i as their possible maximum value.

4. Modeling sampling variance

The variance of the ratio-adjusted direct estimators can be estimated using 1,000 Bootstrap samples provided with survey data. The formula for this variance is given as:

$$\hat{V}(\hat{Y}_i^R) = {M_i}^2 * \hat{V}_{BOOT}\left(\frac{\sum_{j\in S} y_j w_j \delta_{ij}}{\sum_{j\in S} w_j \delta_{ij}}\right) = {M_i}^2 * \frac{1}{1000}\sum_{b=1}^{1000}\left(\frac{\sum_{j\in S} y_j w_j^b \delta_{ij}}{\sum_{j\in S} w_j^b \delta_{ij}} - \frac{\sum_{j\in S} y_j w_j \delta_{ij}}{\sum_{j\in S} w_j \delta_{ij}}\right)^2 \quad (9)$$

$$\hat{V}(\hat{p}_{yi}^R) = \left(\frac{M_i}{N_i}\right)^2 \hat{V}_{BOOT}\left(\frac{\sum_{j\in S} y_j w_j \delta_{ij}}{\sum_{j\in S} w_j \delta_{ij}}\right) = \left(\frac{M_i}{N_i}\right)^2 \frac{1}{1000}\sum_{b=1}^{1000}\left(\frac{\sum_{j\in S} y_j w_j^b \delta_{ij}}{\sum_{j\in S} w_j^b \delta_{ij}} - \frac{\sum_{j\in S} y_j w_j \delta_{ij}}{\sum_{j\in S} w_j \delta_{ij}}\right)^2 \quad (10)$$

where w_j^b is the weight associated with person j in the b^{th} Bootstrap sample.

For most regions, the variance estimate obtained is good. However in the case of regions with very small sample sizes, the estimated variance tends to be unstable. It also happens that the variance can be null in regions where all respondents reported a disability. This is due to the fact that since all y_j take on a value of 1 within a region, the ratio will always be 1 no matter which Bootstrap sample is used, so we observe no variability in this ratio.
In order to stabilize the variance estimation and prevent the problems that null variances would create, a generalized variance model based on the non zero $\hat{V}(\hat{Y}_i^R)$ was found. This model takes the following form:

$$\tilde{V}(\hat{Y}_i^R) = \exp\left\{\beta_0 + \beta_1 \log(M_i) + \beta_2 \log(N_{P_j})\right\} \quad (11)$$

where N_{P_j} represents the total number of people in province P_j to which region i belongs and M_i represents the number of people having responded "yes" to at least one of the Census filter questions.

This model was used to obtain smoothed estimates of the sampling variance associated with direct estimators $\hat{Y}_i^R$ and to impute the sampling variances for null

variances $\widehat{V}(\hat{Y}_i^R)$. Sampling variances were then divided by N_i^2 in order to obtain the ones associated with the direct estimators $\hat{p}_{yi}^R$. Subsequently, these smoothed and imputed variances were considered to be known variances for direct estimators $\hat{Y}_i^R$ and $\hat{p}_{yi}^R$ in the small area models implementation.

5. Selection of auxiliary variables

To identify the explanatory variables of disability among adults and children at the small area level, we conducted a weighted least squares regression analysis using direct estimators. The weight used in this analysis was the ratio of the sample size in the small area to the total sample size, multiplied by the number of areas. Thus, more importance was given to the existing links between the auxiliary variables and the variable of interest in the larger areas than in the smaller areas when determining the most significant variables to explain disability rates. Analysis of the choice of explanatory variables did not rely solely on the R-square result, but also on the diagnostics of heteroscedasticity, normality of residuals and multicolinearity of the explanatory variables. Note that the intercept was included in all linking models to avoid having to center the auxiliary variables.

The auxiliary variables considered for the linking models are based on totals from the 2006 Census long forms and including only people who reported at least one “yes” to the Census filter questions. These totals were transformed to better predict the parameter of interest. As a result, in order to model disability rates, the totals were transformed into proportions to produce a more linear relationship between the parameter of interest and the auxiliary variables.

For adults, the auxiliary variables studied were age, severity of limitations according to the Census, employment status, main source of income, language spoken at home, immigration status, Aboriginal identity, whether they are living or not below the poverty line, as well as average income, average number of hours worked and average value of the residences of persons who answered “yes” to the screening questions in the area.

For children, the auxiliary variables studied were age, severity of limitations on the Census, Aboriginal identity and language spoken at home, as well as characteristics of the adults living with these children in terms of average income, average number of hours worked, average value of their residence, immigration status, employment status, number of unpaid hours providing childcare and main source of income.

6. Small area model used by PALS

As mentioned, both the Fay-Herriot model (3) and the log-linear unmatched model (12 and 13) were assessed for the purpose of estimating the number of persons with disabilities and disability rates. However, the log-linear unmatched model was preferred over the Fay-Herriot model because it appeared to perform better in predicting the desired statistics and produced more stable results when estimating variance.

Since the number of persons with disabilities is deduced directly from the disability rate and vice versa, it was decided to develop the models from the disability rates. The rates obtained from the model were multiplied by N_i to obtain estimates of the number of persons with disabilities.

The log-linear unmatched model for disability rates p_{yi} relies on the sampling model for the direct estimator:

$$\hat{p}_{yi}^{R} = p_{yi} + e_i,\ i = 1,\ldots, m \qquad (12)$$

and the linking model:

$$\log(p_{yi}) = x_i'\beta + \nu_i,\ i = 1,\ldots, m, \qquad (13)$$

where the e_i are sampling errors associated with the direct estimators $\hat{p}_{yi}^{R}$ and the ν_i are random effects related to the linking model for $log(\hat{p}_{yi}^{R})$. We assume that the e_i are independent normal random variables of mean $E(e_i \mid p_{yi}) = 0$ and sampling variance $Var(e_i \mid p_{yi}) = \sigma_i^2$ treated as known and corresponding to the smoothed and imputed variance previously defined.

To obtain the posterior mean $E(p_{yi} \mid \hat{p}_y^R)$ and posterior variance $Var(p_{yi} \mid \hat{p}_y^R)$, where $\hat{p}_y^R = (\hat{p}_{y1}^R, \ldots, \hat{p}_{ym}^R)$, we followed the hierarchical Bayes approach using the Gibbs sampling method. The Gibbs sampling method is a Markov Chain Monte Carlo iterative method that generates samples from the posterior distribution, and then uses these samples to estimate the desired posterior quantities (Gelfand and Smith, 1990). To implement this method, the following conditions must be satisfied:

$$1)\ \pi(p_{yi} \mid \hat{p}_y^R, \beta, \sigma_\nu^2) \propto h(p_{yi})\, f(p_{yi} \mid \beta, \sigma_\upsilon^2),\ i = 1,\ldots, m \quad (14)$$

with $h(p_{yi}) = \dfrac{1}{p_{yi}} \exp\left[-\dfrac{(\hat{p}_{yi}^R - p_{yi})^2}{2\sigma_i^2} \right]$ (14a) and $f(p_{yi} \mid \beta, \sigma_\upsilon^2)$ is the log-normal density function $\dfrac{1}{p_{yi}} \exp\left[-\dfrac{\{\log(p_{yi}) - x'\beta_i\}^2}{2\sigma_\upsilon^2} \right]$ (14b)

$$2)\ \left[\beta \mid \hat{p}_y^R, p_y, \sigma_\nu^2\right] \sim N\left(\frac{\sum_{i=1}^{m} x_i \log(p_{yi})}{\sum_{i=1}^{m} x_i x_i'}, \frac{\sigma_\nu^2}{\sum_{i=1}^{m} x_i x_i'}\right),\ i = 1,\ldots, m \qquad (15)$$

$$3)\ \left[\sigma_\nu^2 \mid \hat{p}_y^R, p_y, \beta\right] \sim IG\left(a_0 + \frac{m}{2}, b_0 + \frac{1}{2}\sum_{i=1}^{m} \{\log(p_{yi}) - x_i'\beta\}^2\right),\ i = 1,\ldots, m \quad (16)$$

where *IG* denotes an inverse gamma distribution and a_0, b_0 are known positive constants and usually set to be very small to reflect our limited knowledge about σ_ν^2.

Sampling using conditions 2 and 3 is straightforward. However, condition 1 does not have a closed form. To update p_{yi}, a rejection sampling algorithm, such as the Metropolis Hastings algorithm within the Gibbs sampler, can be used (Chib and Greenberg, 1995).

The Gibbs sampling algorithm for drawing samples from the posterior distribution is as follows:

a) Using starting values $\beta^{(0)}$, $\sigma_{\upsilon}^{2(0)}$, draw $p_{yi}^{(1)}$, *i = 1, …, m,* from the log-normal density function given by (14a)

b) From (15), draw $\beta^{(1)}$ using $p_{yi}^{(1)}$, *i = 1, …, m,* and $\sigma_{\upsilon}^{2(0)}$

c) From (16), draw $\sigma_{\upsilon}^{2(1)}$ using $p_{yi}^{(1)}$, *i = 1, …, m,* and $\beta^{(1)}$.

These steps correspond to the first cycle of the algorithm. For the following cycles, we incorporate the Metropolis Hastings algorithm into step (a). Therefore, for cycle *k+1,* we draw the candidate $p_{yi}^{(k+1)}$ using $\beta^{(k)}$, $\sigma_{\upsilon}^{2(k)}$ and the log-normal density function (14a) which will be accepted with probability

$$\alpha(p_{yi}^{(k)}, p_{yi}^{(k+1)}) = \min\{h(p_{yi}^{(k)}) / h(p_{yi}^{(k+1)}),1\} \qquad (17).$$

If the candidate is rejected, set $p_{yi}^{(k+1)} = p_{yi}^{(k)}$.

We perform a large number of cycles, say *B*, which we call the "burn-in" period, until convergence, and then we can treat

$$\{p_{yi}^{(B+k)}, \beta^{(B+k)}, \sigma_{\upsilon}^{2(B+k)}\}, \quad k = 1, \ldots, G$$

as *G* samples from the joint posterior distribution. Estimations of $E(p_{yi} \mid \hat{p}_y^R)$ and $Var(p_{yi} \mid \hat{p}_y^R)$ are then based on the marginal sample $\{p_{yi}^{(B+k)}\}, k = 1, \ldots, G$ from the Gibbs sampler.

7. Evaluation of the models

In order to assess the overall validity of the proposed model, the posterior predictive p-value model (Meng, 1994) was used based on the deviation measurement $T(\hat{p}_{yi}^R, p_{yi}) = \sum (p_{yi} - \hat{p}_{yi}^R)^2 / \sigma_i^2$. These statistics revealed that the log-linear unmatched model produced better results than the Fay-Herriot model. The p-values associated with the log-linear unmatched models were more satisfactory, i.e. they were significantly closer to 0.5 than those of the Fay-Herriot model.

8. Benchmarking estimates

The estimates obtained from the log-linear unmatched model were then benchmarked to provincial direct estimates, partly because these estimates are reliable and unbiased with respect to the sample design, but also because of the need for consistency with previous releases based on the PALS.

Mathematically, the purpose of benchmarking is to ensure that the benchmarked estimates p_{yi}^{FINAL} meet the constraint $\sum_{i\in P_j} p_{yi}^{FINAL} * N_i = \sum_{i\in P_j} \hat{Y}_i = \sum_{i\in P_j} \hat{p}_{y_i} * N_i$ where P_j, j=1,…, J, are groups (provinces in our case) of the disjoint m_j small areas such that $\sum_{j=1}^{J} m_j = m$, the total number of small areas. As a result, we get benchmarked estimates using the following formula:

$$p_{yi}^{FINAL} = p_{yi}\left(\sum_{k\in P_j} \hat{Y}_k \Big/ \sum_{k\in P_j} p_{yk} * N_k \right)$$

and the following posterior mean square error (You, Rao and Dick, 2004)

$$PMSE(p_{yi}^{FINAL}) = \left[p_{yi}\left(\sum_{k\in P_j} \hat{Y}_k \Big/ \sum_{k\in P_j} p_{yk} * N_k \right) - p_{yi} \right]^2 + v(p_{yi} \mid \hat{p}_{yi}^{R})$$

9. Results

After consulting with municipal and provincial governments regarding disability data, small area models were developed for two sets of small areas. The first small area set consisted of census metropolitan areas (CMAs) and census agglomerations (CAs) for which we had respondents in the survey. Individuals not covered by these areas were combined to produce estimates at the infraprovincial urban and rural area level. This geographic subdivision consisted of a total of 114 areas.

The second small area set consists of health regions as defined by the provincial departments of health in 2007. In Nova Scotia and Ontario, we had to choose between two possible provincial groupings. We chose to produce estimates by zones in Nova Scotia and by health units in Ontario. Some areas had to be combined because of the very small number of respondents available in those areas. This was the case in the North Shore, Northern Quebec and Nunavik health regions in Quebec; the Sudbury and District and Temiscamingue health units in Ontario; the Burntwood and Churchill regional health authorities in Manitoba; and the Mamawetan Churchill River, Keewatin Yatthé and Athabasca regional health authorities in Saskatchewan. As a result, estimates were obtained for 119 health regions.

In addition, since the availability of auxiliary variables for children was reduced compared to adults, and because the concepts associated with disability differ for the two

groups, they were modelled separately. As a result, four models were developed. For the CAs and CMAs, the sample sizes available for adults ranged between 12 and 1,751, and for children, between 3 and 448. In the case of the health regions, the sample sizes available for adults ranged between 16 and 1,653 and for children, between 3 and 358.

For adults, the following explanatory variables were selected for the linking models:

Health regions:

- the proportion of adults aged between 25 and 34 years who answered "yes" to one of the Census screening questions in the area;
- the proportion of adults aged 75 years and older who answered "yes" to one of the Census screening questions in the area;
- the proportion of non-immigrant adults who answered "yes" to one of the Census screening questions in the area;
- the proportion of adults whose main source of income was the government, and who answered "yes" to one of the Census screening questions in the area;
- the logarithm of the average value of private residences belonging to owners who answered "yes" to one of the screening questions in the area.

Census agglomerations and census metropolitan areas:

- the proportion of adults aged 35 to 44 years who answered "yes" to one of the Census screening questions in the area;
- the proportion of adults aged 65 to 74 years who answered "yes" to one of the Census screening questions in the area;
- the proportion of adults who speak one of the official languages at home and who answered "yes" to one of the Census screening questions in the area;
- the logarithm of the average value of private residences belonging to owners who answered "yes" to one of the screening questions in the area.

These variables converge with the results of studies of false positives. As mentioned in the note to readers textbox, the PALS deems a person to be disabled only if he or she answers "yes" to at least one of the Census screening questions for activity limitation asked on the PALS, AND to at least one of the more detailed screening questions on the PALS. Persons who do not respond positively to at least one of these questions are deemed to be false positives. Based on false positive studies, age, immigration status, language spoken at home, and a person's economic profile could have a clear impact on the chances of being a false positive. For this reason, it is worthwhile to see that these links also exist at the area level.

For children, the following explanatory variables were selected for the linking models:

Health regions:

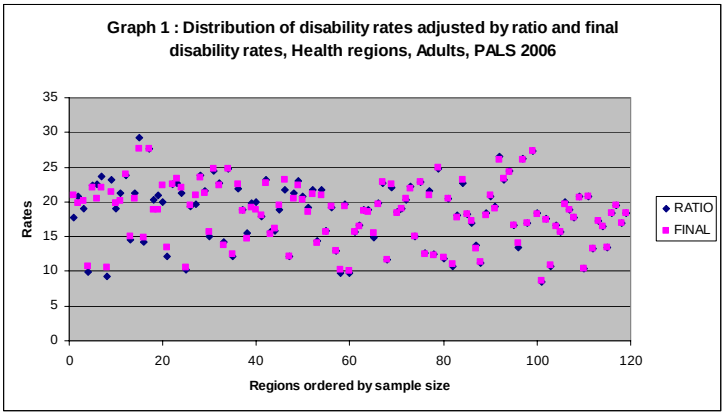
- the proportion of children aged 0 to 1 year for whom "yes" was answered to one of the Census screening questions in the area;
- the proportion of children who speak one of the official languages at home and for whom "yes" was answered to one of the Census screening questions in the area;
- the proportion of non-immigrant persons living with at least one child for whom "yes" was answered to one of the Census screening questions in the area;
- the logarithm of the average value of residences of children for whom "yes" was answered to one of the Census screening questions in the area.

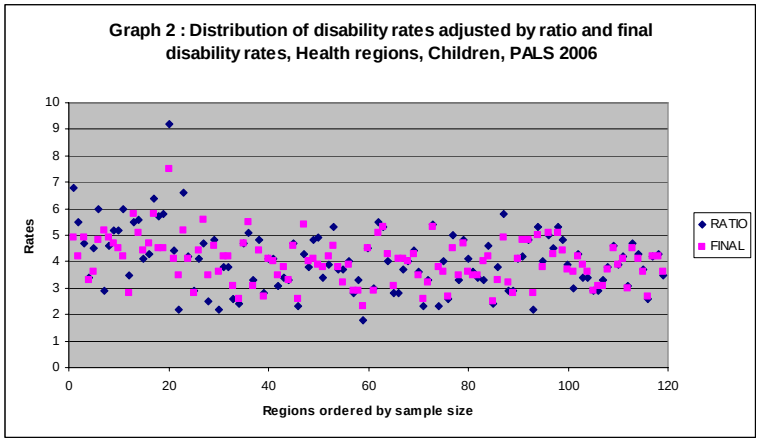
Census agglomerations and census metropolitan areas:

- the proportion of children aged 0 to 1 year for whom "yes" was answered to one of the Census screening questions in the area;
- the proportion of children aged 5 to 9 years for whom "yes" was answered to one of the Census screening questions in the area;
- the proportion of children who speak one of the official languages at home, for whom "yes" was answered to one of the Census screening questions in the area;
- the proportion of non-immigrant persons with at least one child for whom "yes" was answered to one of the Census screening questions in the area.

Again here, it is not surprising to find that these variables are significant in the models for small areas. The false positive analyses show that a very large proportion of children aged 0 to 1 year become false positives. Other factors linked to false positive children are the age of the child, the immigration status of the parents, the language spoken at home, and the economic profile of the child's family. However, as in the false positive analysis, it was found that it is much more difficult to explain disability among children than among adults. Two reasons can be put forward to justify this phenomenon. First, right from the start, disability among children is a lot more difficult to explain by explanatory variables than disability among adults. For adults, age has a very strong correlation with the presence of disability. For children, there isn't a variable that has this strong relationship with the presence of disability. Second, the number of auxiliary variables available for children used for these studies is really limited. Indeed, the Canadian Census collects very little information for children less than 15 years of age. To compensate for this lack of information at the children level, more indirect information about the adults living with children was used. However these variables were less significant in the models.

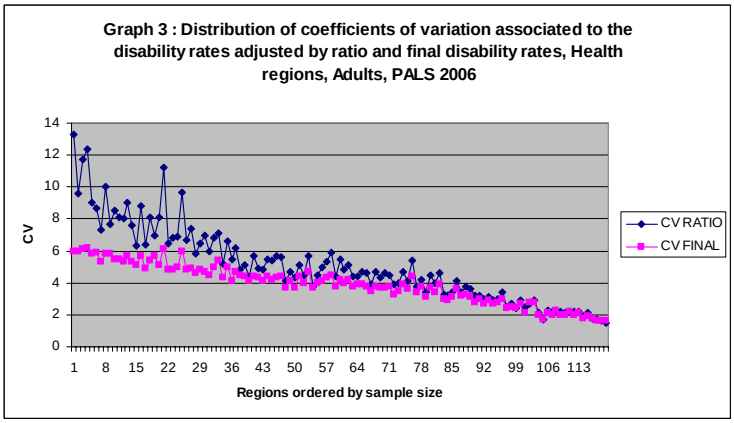
The following graphs show the distribution of disability rates adjusted by the ratio $\hat{p}_{yi}^{R}$ and final disability rates p_{yi}^{FINAL} obtained by the log-linear unmatched model and benchmarking of the estimates. The rates are in order of the sample size of the region. Thus, regions with a smaller sample size are shown on the left and those with a larger sample size are on the right. For ease of reading, only the results for the health regions are presented here, but the results for the census agglomerations and census metropolitan areas are provided in the appendix.

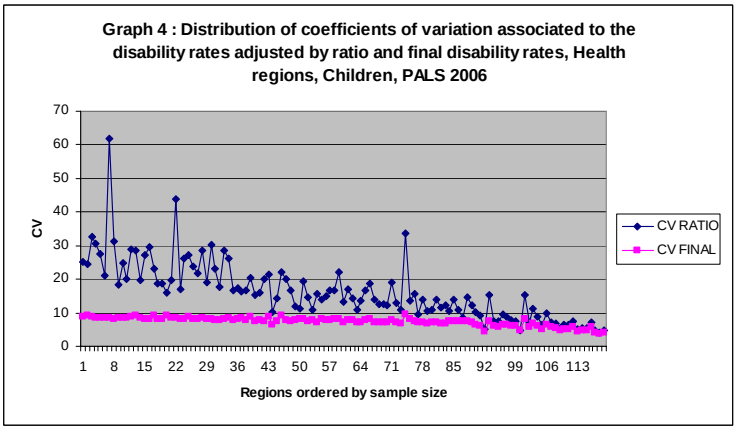




Note that for adults, final estimates obtained by the small area model are very close to the ratio-adjusted direct estimates, especially in the case of large sample sizes. For children, given the very small sample sizes available and consequently the lack of accuracy in the direct estimates, there is a larger variation between direct estimates and final estimates which fades as the sample size becomes larger. However, even for the larger regions which have fairly good estimates to start with (estimates on the right side of the graph), it can be seen that the model doesn't fit the data perfectly. Thus, we can suppose that the lack of explanatory variables has an effect on the validity of the model.

The following graphs show the coefficients of variation (CV) associated with these rates.





There is a sharp decrease in the CVs with the application of the small area model. In the case of both adults and children, applying the model has the greatest impact on the CVs in areas with the smallest sample sizes. Indeed, the more accurate the direct estimator, the more importance the final estimate gives to it, which means less gain with respect to the variance of the final estimates. For children where the CVs associated with the direct estimates are very high, we find that the distribution of final CVs is much more uniform than for adults. This result indicates that the linking model is more important in calculating the final estimate for children than it is for adults.

10. Conclusion and comments

The small area estimation project for the number of persons with disabilities and disability rates comes from a need for data expressed by the provincial governments and some municipalities. However, the real driving force behind this project was the fact that PALS offers ideal conditions for using small area estimation methods. Since PALS is a post-censal survey, estimates produced from the Census could be used directly as auxiliary variables. These variables are available in large numbers for the desired small areas and many are relevant for explaining disability. In addition, the PALS sample sizes, some 29,000 adult respondents and 7,000 child respondents, also make it possible to obtain the direct estimators needed to apply the models for most of the desired areas.

The estimates for PALS small areas were produced by applying the log-linear unmatched model to which the hierarchical Bayes approach was applied using the Gibbs sampling method. This approach uses a sampling model based on ratio-adjusted direct estimators of the disability rate and the number of persons with disabilities. It is also based on a log-linear linking model, which determines the link between these parameters of interest and auxiliary variables from the known totals of the Census long form.

The estimates obtained for adults are very close to the ratio-adjusted direct estimates. This can be attributed to the fact that these estimates were of high quality at the outset and that the linking model performed well for predicting parameters of interest. We are therefore very confident that the adult estimates are accurate and robust.

The small area estimates obtained for children differ more from their direct estimates. Since the accuracy of the direct estimates for children were much poorer than that of adults, greater importance to the model was given when calculating the final estimates. But the linking model to predict the parameters of interest for children were also not performing as good for children compared to adults (there were also less available auxiliary variables). As well, as for any result produced almost exclusively from a model, it is more difficult to judge the validity of the results. Consequently, only the adults' estimates will be officially released based on the 2006 PALS data. The children model will be re-evaluated in 2011 as this small area project will probably be carried out using the 2011 PALS data.

The 2011 instance of PALS will also give us the opportunity to look at several other small area models which could be considered. For example, we might consider using models at the unit level. This would require incorporating the survey's sample design into the models, which would complicate the estimation method given the complexity of the survey design. We might also assess semi-parametric models, such as the "penalized

spline" models recently used in a survey of lakes in the North-eastern United States (Opsomer and al., 2008).

Finally, another project that could be realized using the 2011 PALS data would be to extend the small area methods to other important statistics produced by the survey. It would definitely be valuable to obtain small area estimates by type of disability and by severity of disability.

Acknowledgments

The authors would like to thank Dave Dolson, Jean-Pierre Morin, Julie Bernier and Mike Hidiroglou for their suggestions and constructive comments when reviewing this paper. The authors would also like to thank Professor John N.K. Rao from Carleton University, whose expertise in the small area estimation field greatly influenced the direction of this project. Finally, since this work was done as part of the 2006 Participation and Activity Limitation Survey, the authors would like to thank the PALS manager, Mrs Susan Stobert, as well as Human Resources and Skills Development Canada, official sponsor of this survey, who have made this project possible with their financing support.

References

Chip, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, vol. 49, pp 327-335.

Dick, P. (1995). Modeling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, vol. 21, pp 45-54.

Dick, P. and You, Y. (1998). A Hierarchical Bayes analysis of census undercoverage**.** *Symposium 97, New directions in surveys and censuses: proceedings*. Statistics Canada, Ottawa, pp 101-105.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income for small places: An application of James-Stain procedures to census data. *Journal of the American Statistical Association*, vol. 74, pp 269-277.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, vol. 85, pp. 398-409.

Meng X. L. (1994). Posterior predictive *p*-value. *The Annals of Statistics*, vol. 22, pp 1142-1460.

Opsomer, J. D. and al. (2008). Non-parametric small area estimation using penalized spline regression. *JRSSB*, vol. 70, part 1, pp 265-286.

Singh, M.P., Gambino, J. and Mantel, H.J (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, June 1994, vol. 20, no 1, pp 3-22.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, June 2008, vol. 34, no 1, 19-27.

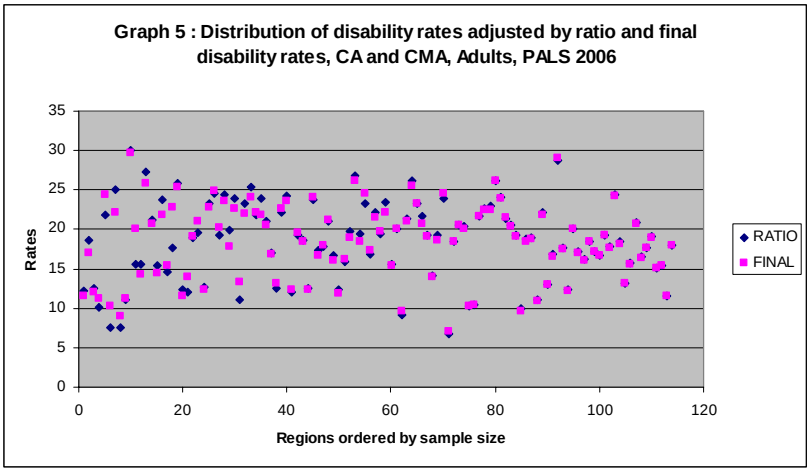
You, Y. and Chapman, B. (2006). Small Area Estimation Using Area Level Models and Estimated Sampling Variances. *Survey Methodology*, June 2006, vol. 32, no1, 97-103.

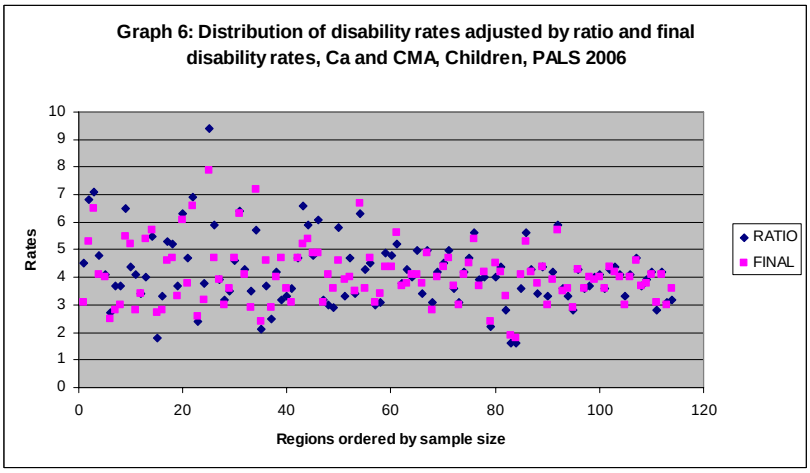
You, Y. and Rao, J.N.K. (2002) Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

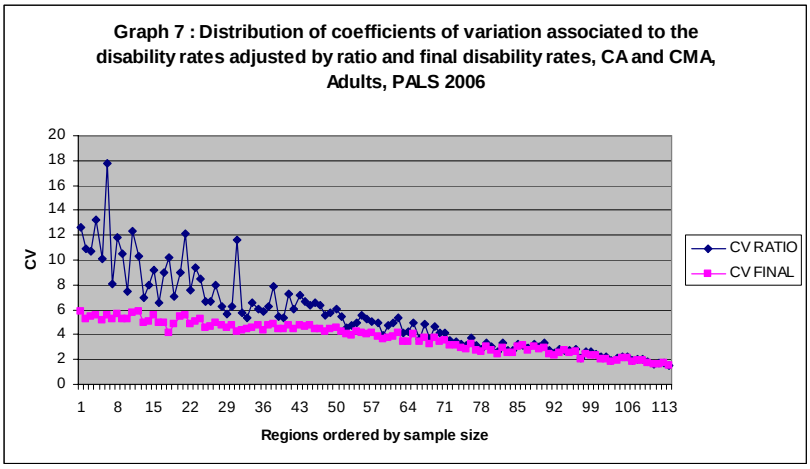
You, Y, Rao, J.N.K and Peter Dick (2004). Benchmarking Hierarchical Bayes Small Area Estimators in the Canadian Census Undercoverage Estimation. Statistics in Transition, April 2004, Vol. 6, No. 5, pp. 631-640.

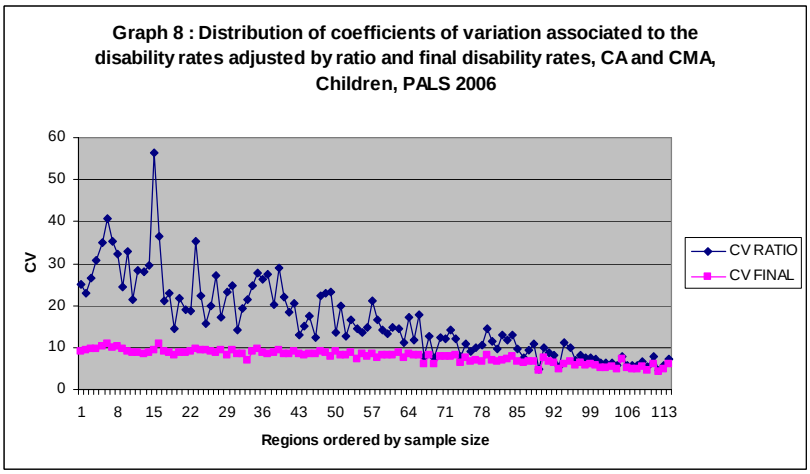
Zhou, Q.M. and You, Y. (2007). Hierarchical Bayes small area estimation for the Canadian Community Health Survey. Methodology Branch Working Paper, HSMD-2007-007E, Statistics Canada.

Appendix






Graph 7 : Distribution of coefficients of variation associated to the disability rates adjusted by ratio and final disability rates, CA and CMA, Adults, PALS 2006
CV
20
18
16
14
12
10
8
6
4
2
0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99 106 113
Regions ordered by sample size
CV RATIO
CV FINAL


Graph 8 : Distribution of coefficients of variation associated to the disability rates adjusted by ratio and final disability rates, CA and CMA, Children, PALS 2006
CV
60
50
40
30
20
10
0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99 106 113
Regions ordered by sample size
CV RATIO
CV FINAL