Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada | Statistique Canada

Canada

Proceedings of Statistics Canada's Symposium 2005
Methodological Challenges for Future Information Needs

GETTING AROUND A CONCEPTUAL DIFFERENCE BETWEEN TWO DATA SOURCES WHEN PRODUCING ESTIMATES

Serge Godbout and Chantal Grondin[1]

ABSTRACT

The Survey of Employment, Payrolls and Hours is a monthly survey using two data sources: a census of administrative records and an establishment survey. The administrative source, derived from the Canada Revenue Agency's payroll deduction accounts, contains employment and gross monthly payroll variables. The survey source also provides employment and gross monthly payroll data, along with a set of related variables that are not available from the administrative source. The survey data are used to build models in order to mass impute several derived variables on the administrative source. The survey design relies on the fact that the concepts for number of employees and gross monthly payroll are the same on the two data sources. However this is not always the case in practice, which causes instability in the estimates. In this article, we describe different solutions that were brought to the survey design and to the mass imputation model to allow us to get around this conceptual difference, hence producing estimates that are more stable over time. The article concludes with some results from different estimation scenarios for average weekly earnings.

KEY WORDS: Data sources; sampling plan; regression estimation

1. INTRODUCTION

The Survey of Employment, Payrolls and Hours (SEPH) is a monthly survey using two data sources: a census of administrative records and a survey of a sample of establishments, the Business Payrolls Survey (BPS). The SEPH produces level and trend estimates for employment, earnings, hours and related variables by province and by industry (Grondin and Lavallée, 2001).

The administrative source consists of payroll deduction accounts forms supplied by the Canada Revenue Agency (CRA). Since the forms cannot be linked to a particular province or industry, the data have to be aggregated at the enterprise level and then disaggregated at the establishment level, as establishments are associated with one province and one industry. Hence we have, for each establishment j in the establishment universe U, the number of employees $E_{A,j}$ and the gross monthly payroll $P_{A,j}$. The survey portion consists of a stratified sample S of about 11,000 establishments selected from a list frame, Statistics Canada's Business Register (BR). Those establishments can be linked to the administrative source. Among the variables collected for each unit $i \in S$ are number of employees $E_{E,i}$, gross monthly payroll $P_{E,i}$ and weekly earnings $G_{E,i}$. The sampling weight of the units in stratum h is $w_i = 1 / p_i$, where p_i is the selection probability of unit i. The survey data are used to build models for the mass imputation of a range of derived variables on the administrative source, including weekly earnings. The final estimates are produced with data generated through mass imputation of the administrative source.

This survey design assumes that the employment and gross monthly payroll concepts are the same on both sources. Experience has shown, however, that that is not always the case in practice, which causes instability in the estimates.

In this article, we will describe various solutions brought to get around this difference in concepts between the two sources. First, we will examine some of the problems that underlie or contribute to the difference between the two sources. We will propose various ways of improving the survey design so that there will be greater consistency between the two sources. Then we will compare various models and scenarios for estimating the derived variables and look at their impact on the stability of the estimates. In this article, we will confine our

[1] Serge Godbout, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, R. H. Coats Building, 11th Floor, Ottawa, Ontario, Canada, K1A 0T6, Serge.Godbout@statcan.ca.
Chantal Grondin, Special Surveys Division, Statistics Canada, 150 Tunney's Pasture Driveway, Main Building, Room 2500-G, Ottawa, Ontario, Canada, K1A 0T6, Chantal.Grondin@statcan.ca.

analysis to the estimation of average weekly earnings, even though the SEPH is interested in other similar estimates, such as average weekly paid hours.

2. PROBLEMS AND IMPROVEMENTS IN THE SAMPLING PLAN

The SEPH's survey design is based on the assumption that the concepts that are common to the two sources are equivalent. Yet there are a number of differences. First, the employment and payroll of establishments in the administrative source do not always match the employment and payroll of establishments selected for the BPS. This occurs when data are missing and have to be imputed, for example. Another reason is that the establishment employment and payroll figures in the administrative source are not reported directly by respondents but are obtained through a process of aggregation (at the enterprise level) and disaggregation (at the establishment level). The disaggregation is based on an approximate proportion of employees in each establishment from the BR. While this disaggregation generally produces quite reliable employment figures, the same cannot be said for monthly payroll. Using only a proportion of employees to disaggregate the payroll is tantamount to assuming that all employees of the various establishments in an enterprise earn the same average monthly salary, which may not be true. At this time, there is no other information source that would allow us to disaggregate the payroll in some other way. However, to make the data from the two sources more consistent, we plan to collect BPS data at the enterprise level and use the same disaggregation process as is used for the administrative source to generate the information at the establishment level in the survey. This will result in greater consistency between the employment and payroll figures for the two sources.

The BPS currently collects data at the establishment level. What happens in some instances, though, is that, without informing us, units report data for a different level (the enterprise or a group of establishments, for example). As a result, the relative weight of the data reported by those units is too high when we build the regression models. This problem can be remedied by collecting information at the enterprise level, as long as respondents actually report at that level. However, the problem also arises in the case of units that change classifications or provinces over time. After weighting, their figures, particularly the number of employees, become too influential. The solution we propose is to reduce the weight of these overly influential units. This weighting adjustment is performed by attempting to fit the BPS distribution of employment by size group to the distribution found in the administrative source.

There is also a problem with the payroll variable. In the administrative source, gross monthly payroll is equal to the salaries paid out during the reference month, which generally includes amounts for days worked in the previous month, depending on how the establishment's pay periods are defined. In the BPS, however, the gross monthly payroll, though defined in the same way as it is in the administrative source, seems to be reported differently. Some respondents report the amounts payable for the days worked between the first and last days of the month, while others report two biweekly pay periods or four weekly periods, ignoring the months that have an extra pay period. As a result, gross monthly payroll in the survey does not show the same trend between two months as gross monthly payroll in the administrative source. The solution we propose is to redefine the regression models so that they no longer use the survey's gross monthly payroll variable as an explanatory variable.

3. ESTIMATION MODELS

In this study, we focus on estimating average weekly earnings (AWE). We will start with the estimation of AWE using BPS data alone. Then we will look at two ways of producing AWE estimates with combined data from the two sources. We will see that the problem of inconsistency between the two sources can be remedied with a change in the way the data are modelled.

3.1 Estimation with survey data alone (method E)

AWE can be estimated from survey data alone. We use the Horvitz-Thompson estimator, with the sampling weight set to the inverse of the selection probability. For a domain *d*, average weekly earnings ($A\hat{W}E$) are estimated using a ratio of weighted sums: $A\hat{W}E_{\text{E}} = \sum_{i \in d \cap S} w_{E,i} G_{E,i} \Big/ \sum_{i \in d \cap S} w_{E,i} E_{E,i}$, where $w_{E,i}$ is the sampling weight, $G_{E,i}$ is the weekly earnings and $E_{E,i}$ is the number of employees of each establishment *i* sampled by the survey.

AWE are estimated by a known method, but the sample size does not provide precise enough estimates for a number of domains (province by industry).

3.2 Estimation using both sources without microdata matching (method A1)

To estimate AWE with both sources without microdata matching, we use synthetic estimators. That is, we model weekly earnings per employee as a function of monthly payroll per employee from the survey data, and we apply the estimated coefficients to the monthly payroll per employee on the administrative source. To simplify the calculation of these synthetic estimators, we first mass-impute the weekly earnings variable $\text{G}_{A,j}$ for each establishment *j* in the administrative source, as follows. We divide the BPS establishments into homogeneous groups (referred to as model groups) to model weekly earnings per employee as a function of monthly payroll per employee. The regression model for a given model group *g* is as follows: $(G/E)_{E,i(g)} = \beta_{1(g)} + \beta_{2(g)} (P/E)_{E,i(g)} + \varepsilon_{i(g)}$, where $(G/E)_{E,i} = \text{G}_{E,i} / \text{E}_{E,i}$ is the weekly earnings per employee and $(P/E)_{E,i} = \text{P}_{E,i} / \text{E}_{E,i}$ is the monthly payroll per employee. Parameters $\beta_{1(g)}$ and $\beta_{2(g)}$ are estimated by the least squares method (Lohr, 1999) using survey data. Then they are applied to all *j* establishments in the administrative source (by model group) to mass-impute weekly earnings per employee, $(G\hat{/}E)_{A,j}$. From there, we derive the weekly earnings for establishment *j* by finding $\hat{G}_{A,j} = (G\hat{/}E)_{A,j} \times E_{A,j}$. Lastly, the estimate of $A\hat{W}E$ for domain *d* is computed as follows: $A\hat{W}E_{\text{A1}} = \sum_{j \in d \cap U} \hat{G}_{A,j} \Big/ \sum_{j \in d \cap U} E_{A,j}$.

When the model groups are very homogeneous, this method considerably reduces the mean square error of estimates for very small domains, since the variance is much lower and the bias is small. In addition, if the model is a good fit, we will have very high coefficients of determination R^2. However, this method is particularly sensitive to the differences between the employment and payroll concepts in the two sources.

3.3 Estimation using the two sources with microdata matching (method A2)

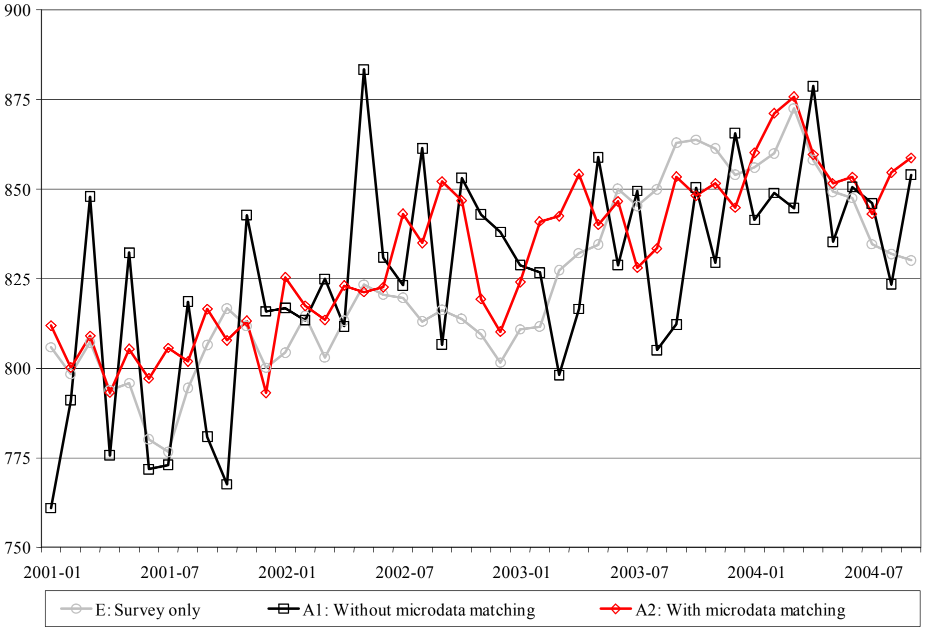
AWE can also be estimated by matching survey microdata with administrative microdata. For each BPS sample unit *i*, we estimate the parameters of the regression model using the weekly earnings per employee from the survey, $(G/E)_{E,i} = \text{G}_{E,i} / \text{E}_{E,i}$, and the monthly payroll per employee from the administrative source, $(P/E)_{A,i} = \text{P}_{A,i} / \text{E}_{A,i}$. Hence, the regression model for a given model group *g* is $(G/E)_{E,i(g)} = \beta_{1(g)} + \beta_{2(g)} (P/E)_{A,i(g)} + \varepsilon_{i(g)}$. Again here, coefficients $\beta_{1(g)}$ and $\beta_{2(g)}$ are estimated using the least squares method, and mass imputation is performed as shown in section 3.2. The final estimate of $A\hat{W}E$ for domain *d* is computed by finding $A\hat{W}E_{\text{A2}} = \sum_{j \in d \cap U} \hat{G}_{A,j} \Big/ \sum_{j \in d \cap U} E_{A,j}$, where $\hat{G}_{A,j} = (G\hat{/}E)_{A,j} \times E_{A,j}$.

The advantage of this model is that it is not affected by the problems associated with the difference in concepts between the two sources. The monthly payroll per employee used to estimate the model's parameters is derived from the administrative source, $(P/E)_{A,j}$, and in mass imputation, we apply the estimated parameters to that same variable.

3.4 Results

Using SEPH data, we estimated the AWE for January 2001 through September 2004 by the three methods described above. We examine here the results for a domain in which we observed a substantial discrepancy between the two sources. For this particular domain *d*, the sample size of the survey portion is about 1,300 units, and the administrative source covers more than 60,000 units. We used only one model group (*g* = *d*). When we estimated the parameters $\beta_{1(g)}$ and $\beta_{2(g)}$ over 45 months, model A1 had better coefficients of determination R^2 (between 0.96 and 0.99) than model A2 (between 0.62 and 0.94), but this statistic is misleading because the problem with model A1 stems from the difference in concepts between the variables in the two sources. When the three series are plotted on a graph (Figure 1), model A1 shows greater variability than the other two methods.

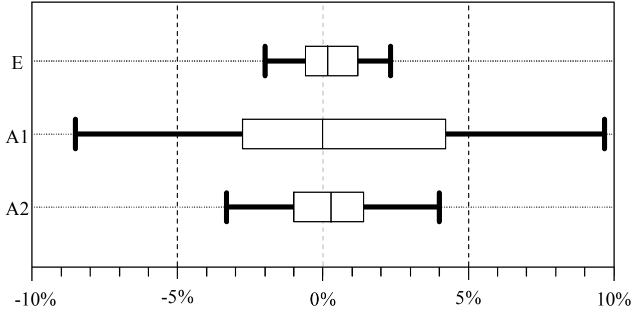
Figure 1
Estimates of average weekly earnings in dollars using three different methods
(Domain *d*, January 2001 - September 2004)



In the case of method A1 (without microdata matching), the observed variations in the AWE seem to be caused by sharp variations in monthly payroll in the administrative source. The result is a sawtooth curve. This model is too rigid over time, since it is a function of monthly payroll per employee from the survey, whereas the estimated coefficients are applied to the monthly payroll per employee from the administrative source, and that variable tends to be much more stable in the survey than in the administrative source. For example, we can see that the spikes in the method A1 AWE series correspond to months with five Thursdays or five Fridays. We saw the same spikes when we produced the monthly-payroll-per-employee series from the administrative source.

Visual inspection shows that the curves for methods E and A2 are much more stable than the curve for A1 The distribution of monthly percentage changes (Figure 2) has a standard deviation of 4.35 percentage points for A1, compared with 1.53 for A2 and 1.08 for E. If we assume that the trend over the 45 months is linear, the A2 series is better than the E and A1 series, since its coefficient of determination R^2 is 0.76, compared with 0.69 and 0.33 for the other two series.

Figure 2
Distribution of monthly percentage changes for the three methods
(Domain *d*, January 2001 - September 2004)



Finally, analysis of the A1 and A2 models has shown that the residuals generally have a normal distribution and are free of heteroscedasticity.

CONCLUSION

We have observed certain problems associated with the difference in concepts between the SEPH's data sources. Some of the inconsistencies can be mitigated by making changes in the BPS. In the future, we plan to collect data at the enterprise level rather than at the establishment level. The information will then be disaggregated at the establishment level using the same disaggregation process as is used for the administrative data. We will also use a method that reduces the impact of influential values by fitting the distribution of employment in the survey data to the distribution of employment in the administrative data.

Of the three methods of estimating average weekly earnings that we studied, the one that involved using two sources with microdata matching (method A2) yields much more stable results than the other two. It reveals trends over time and, unlike the method based on survey data alone, it produces more precise province-by-industry estimates. In addition, the administrative source's "monthly payroll per employee" variable can be regarded as complementary to the survey data in the modelling process. This new method has also been used in the estimation of average weekly paid hours and special payments. The results were similar to those presented in this paper.

ACKNOWLEDGEMENTS

We wish to thank the editors and the team of SEPH methodologists (Pierre Lavallée, Anne-Marie Houle, Richard Belcher, Édith Hovington and Sharon Wirth) for their assistance, comments and suggestions. We are also grateful to Don Royce, John Kovar and Éric Rancourt for their suggestions concerning models.

REFERENCES

Grondin, C. and Lavallée, P. (2001), "Current Methodology of the Survey of Employment, Payrolls and Hours", Ottawa, Canada: Statistics Canada.

Kroese, A.H. and Renssen, R.H. (2000), "New Applications of Old Weighting Techniques: Constructing a Consistent Set of Estimates Based on Data from Different Sources", *Proceedings of the Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms, and Institutions*, Buffalo, New York: American Statistical Association, p. 831-840.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Pacific Groove: Duxbury Press.