**STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

Working paper No. 17
English only

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UNECE Work Session on Statistical Data Editing**
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

# PERFORMANCE OF JACKKNIFE VARIANCE ESTIMATION USING SEVERAL IMPUTATION METHODS

**Contributed paper**

Submitted by Instituto Nacional de Estadistica, Spain[1]

**Abstract**

In this paper we study the performance of jackknife variance estimation based on adjusted imputed values. We apply several imputation methods to random samples from a population of business survey data.

Key words: jackknife variance estimation, ratio imputation, hot deck imputation , mean imputation.

## I. INTRODUCTION

1. The main objective of this study is to analyze the performance of jackknife variance estimation under several imputation methods. This performance is measured through the monte carlo relative bias and mean square error of the jackknife variance estimation and its coverage rate. Missing values are generated in samples from a population of industrial businesses. The imputation methods used are ratio imputation with and without residuals, mean imputation and hot deck imputation. We usually need more than one imputation method in business surveys.

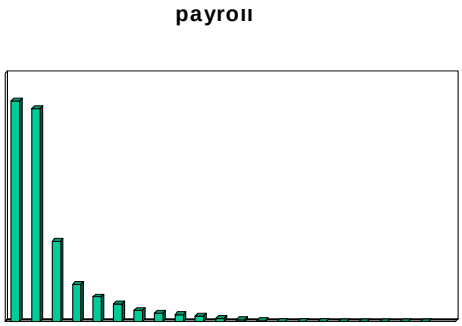
2. In the following section we describe the data that will be used. In section III we present the imputation methods and the jackknife estimators. Section IV presents the results of the montecarlo experiments. Finally, some conclusions are provided.

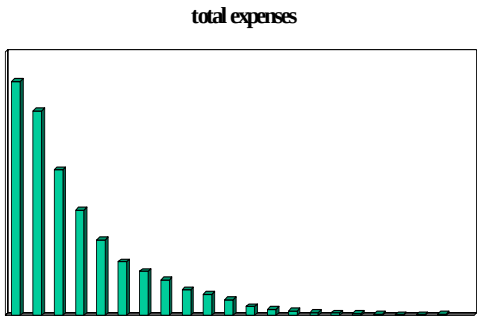
## II. INDUSTRIAL BUSINESS SURVEY DATA

3. In Spain, the Industrial Business Survey completely enumerates businesses with 20 or more employees. We take these as our population. We simulate 200,000 samples from this population to test the performance of the jackknife variance estimation.

4. The variables considered in this paper are turnover, payroll expense and total expense. Their distributions are skewed as the histograms show. Table 1 shows a high and very high correlation between them.

[1] Prepared by F. Aparicio (fapape@ine.es) and D. Lorca (mdlorca@ine.es).



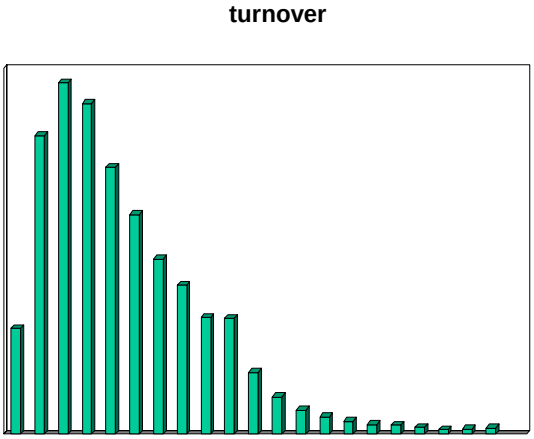




**Table 1**:

| Correlation: | Turnover |
|---|---|
| Payroll expenses | 0.77 |
| Total expenses | 0.99 |

In this study we impute the turnover from each of the other two variables.

## III. BACKGROUND

### A. Jackknife variance estimation

5. Let U={1...,k,...N} be the index set of the finite population and s a simple random sample without replacememt of size n drawn from U. The set of respondents , r, is of size m and the set of non-respondents , o, is the size l=n-m. Let y be the variable of interest to be imputed using more than one imputation method and x the auxiliary variable.

6. We consider the jackknife variance estimation based on adjusted imputed values proposed by Rao and Shao (1992). This takes into account the fact that some data are imputed values. If a respondent is deleted, a re-imputation is done using the response set reduced by one unit.
Let $\hat{y}_k$ be the imputed value and $\hat{y}_k(j)$ the value obtained by computing the imputed values using the reduced response set after unit j has been deleted.
The data after imputation are given by $\{y_{.k} : k \in s\}$ where

$$y_{.k} = \begin{cases} y_k & If \quad k \in r \\ \hat{y}_k(j) & if \quad k \in o \ \ and \ \ j \in r \\ \hat{y}_k & if \quad k \in o \ \ and \ \ j \in o \end{cases}$$

The jackknife variance estimator of $\bar{y}_{.s}$ is:

$$\hat{V}_J = (1-f)\frac{n-1}{n}\sum_{j \in s}\left(\bar{y}^r_{.s}(j) - \bar{y}_{.s}\right)^2$$

with $s(j) = s - \{j\}$

$$\bar{y}^r_{.s}(j) = \frac{1}{n-1}\sum_{s(j)} y^r_{.k}(j)$$

$$f = \frac{n}{N}$$

**B. Imputation methods**

7. Here we present the formulae used for ratio imputation with residuals and hot-deck imputation when they are both applied to the same data set. For ratio and mean imputation the formulae are provided by Rancourt, Lee and Särndal (1994).

8. The response set is divided into two parts: $r_1$ and $r_2$ of sizes $m_1$ and $m_2$ respectively and the non-response set is accordingly divided into two parts: $o_1$ and $o_2$ of sizes $l_1$ and $l_2$ .
We suppose that in $r_1$ and $o_1$ we have the values of the x variable avalaible. Therefore we use ratio imputation in $o_1$. In $o_2$ we apply with replacement hot deck imputation.
The imputed values are given by:

$$y_{.k} = \begin{cases} y_k & if \quad k \in r \\ \left(\bar{y}_{r_1}/\bar{x}_{r_1}\right)x_k + \varepsilon_k^* & if \quad k \in o_1 \\ \delta_k^* & if \quad k \in o_2 \end{cases}$$

where $\delta_k^* \in \{y_i : i \in r\}$, is a with replacement and equal probabilities realization from the observed values of y,

$\varepsilon_k^* \in \left(y_i - \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}}x_i : i \in r_1\right)$, is a with replacement and equal probabilities realization from the residuals of the complete observations.
The point estimator for $\bar{Y}$ after imputation is:

$$\bar{y}_{.s} = \frac{1}{n}\left(m\bar{y}_r + l_1\bar{x}_{o_1}\frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} + \sum_{k \in o_1}\varepsilon_k^* + \sum_{k \in o_2}\delta_k^*\right)$$

The jackknife mean after deletion of unit is:

$$\bar{y}_{.s}^{r}(j)=\begin{cases}\dfrac{1}{n-1}\left(\left(m+\dfrac{l_2}{m-1}\right)\bar{y}_r-\left(1+\dfrac{l_2}{m-1}\right)y_j+\dfrac{\bar{y}_{r_1}(j)}{\bar{x}_{r_1}(j)}l_1\bar{x}_{o_1}+\sum_{k\in o_1}\varepsilon_k^*+\sum_{k\in o_2}\delta_k^*\right) & \text{if} \quad j\in r_1\\ \dfrac{1}{n-1}\left(\left(m+\dfrac{l_2}{m-1}\right)\bar{y}_r-\left(1+\dfrac{l_2}{m-1}\right)y_j+\dfrac{\bar{y}_{r_1}}{\bar{x}_{r_1}}l_1\bar{x}_{o_1}+\sum_{k\in o_1}\varepsilon_k^*+\sum_{k\in o_2}\delta_k^*\right) & \text{if} \quad j\in r_2\\ \dfrac{1}{n-1}\left(m\bar{y}_r+\dfrac{\bar{y}_{r_1}}{\bar{x}_{r_1}}\left(l_1\bar{x}_{o_1}-x_j\right)+\sum_{k\in o_1-\{j\}}\varepsilon_k^*+\sum_{k\in o_2}\delta_k^*\right) & \text{if} \quad j\in o_1\\ \dfrac{1}{n-1}\left(m\bar{y}_r+\dfrac{\bar{y}_{r_1}}{\bar{x}_{r_1}}l_1\bar{x}_{o_1}+\sum_{k\in o_1}\varepsilon_k^*+\sum_{k\in o_2-\{j\}}\delta_k^*\right) & \text{if} \quad j\in o_2\end{cases}$$

where

$$\sum_{k\in o_1}\varepsilon_k^*=l_1\bar{\varepsilon}_{o_1}^*$$

$$\sum_{k\in o_2}\delta_k^*=l_2\bar{\delta}_{o_2}^*$$

$$\sum_{k\in o_1-(j)}\varepsilon_k^*=l_1\bar{\varepsilon}_{o_1}^*-\varepsilon_j^*$$

$$\sum_{k\in o_2-(j)}\delta_k^*=l_2\bar{\delta}_{o_2}^*-\delta_j^*.$$

## IV. MONTECARLO STUDY AND RESULTS

### A. Montecarlo study

9. From our population of industrial businesses of size N=16438, simple random samples without replacement of sizes n=100, 500, 1000 and 5000 are drawn.
Non-response in the y variable is randomly generated, assuming that the response mechanism is uniform. A loss of about 30 per cent is simulated.
The number of replications is 200000 for each x variable, imputation method and sample size.

10. Fortran programs were developed to make the simulations. They can be obtained from the authors upon request.

11. Within each replication, we compute the percentage relative bias, the relative mean square error and the coverage rate of the 95 percent confidence interval based on the normal approximation.

### B. Results

12. Tables 2 to 5 show the simulation results for each sample size and imputation method.

**Table 2**. Imputed variable: turnover. Auxiliary variable: payroll expense
Imputation methods: ratio and mean imputation

| Sample size | Ratio: | 100% | | 90% | | | 70% | | | 0% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean: | 0% | | 10% | | | 30% | | | 100% | | |
| | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) |
| 100 | 2.54 | 12334 | 58.5 | 2.30 | 13355 | 58.4 | 1.26 | 14856 | 58.4 | 0.23 | 15821 | 57.1 |
| 500 | -0,30 | 1135 | 68.0 | -0.12 | 1158 | 67.9 | 0.086 | 1226 | 67.3 | -1.22 | 1343 | 65.8 |
| 1000 | -0.89 | 381 | 74.3 | -0.60 | 392 | 74.0 | -0.78 | 411 | 73.7 | -2.14 | 454 | 72.6 |
| 5000 | -6.40 | 22.6 | 88.4 | -7.0 | 23.4 | 88.1 | -8.5 | 24.8 | 87.8 | -11.4 | 27.9 | 87.3 |

**Table 3**. Imputed variable: turnover. Auxiliary variable: total expense
Imputation methods: ratio and mean imputation

| Sample size | Ratio: | 100% | | 90% | | | 70% | | | 0% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean: | 0% | | 10% | | | 30% | | | 100% | | |
| | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) |
| 100 | 0.72 | 12671 | 59.2 | 0.68 | 12574 | 58.8 | 0.14 | 13107 | 58.6 | 0.29 | 15675 | 56.9 |
| 500 | 1.22 | 970 | 69.9 | 0.32 | 986 | 69.4 | 0.38 | 1036 | 68.9 | -0.90 | 1339 | 65.8 |
| 1000 | 0.36 | 317 | 76.7 | -0.07 | 321 | 76.5 | -0.50 | 338 | 75.9 | -2.12 | 455 | 72.6 |
| 5000 | -2.22 | 17.5 | 90.3 | -2.41 | 17.7 | 90.3 | -4.25 | 18.8 | 89.8 | -12.0 | 28.0 | 87.3 |

**Table 4**. Imputed variable: turnover. Auxiliary variable: payroll expense
Imputation methods: ratio with residuals and hot deck imputation

| Sample size | Ratio: | 100% | | 90% | | | 70% | | | 0% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hot deck | 0% | | 10% | | | 30% | | | 100% | | |
| | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) |
| 100 | 2.47 | 12773 | 59.5 | 3.17 | 13932 | 59.2 | 1.79 | 14958 | 59.3 | -1.44 | 19413 | 58.7 |
| 500 | -0.41 | 1262 | 68.7 | 0.32 | 1314 | 68.7 | -0.36 | 1396 | 68.3 | -2.01 | 1678 | 67.6 |
| 1000 | -1.62 | 436 | 75.0 | -1.35 | 451 | 74.9 | -1.54 | 485 | 74.6 | -2.45 | 569 | 73.8 |
| 5000 | -9.70 | 26.6 | 88.1 | -10.7 | 28.0 | 88.0 | -11.7 | 30.2 | 87.6 | -15.8 | 36.0 | 87.0 |

**Table 5**. Imputed variable: turnover. Auxiliary variable: total expense
Imputation methods: ratio with residuals and hot deck imputation

| Sample size | Ratio: | 100% | | 90% | | | 70% | | | 0% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hot deck | 0% | | 10% | | | 30% | | | 100% | | |
| | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) | RB(%) | MSE | COVR(%) |
| 100 | 0.61 | 12521 | 59.4 | 0.52 | 13058 | 59.1 | 0.64 | 14010 | 59.2 | -0.02 | 19635 | 58.8 |
| 500 | 0.54 | 986 | 70.0 | -0.05 | 1015 | 69.8 | 0.64 | 1126 | 69.6 | -1.21 | 1681 | 65.5 |
| 1000 | 0.34 | 319 | 76.9 | -0.90 | 332 | 76.6 | -0.94 | 370 | 76.3 | -2.27 | 569 | 73.9 |
| 5000 | -2.85 | 17.7 | 90.4 | -4.35 | 18.6 | 90.1 | -7.16 | 21.1 | 89.6 | -15.5 | 36.1 | 87.1 |

13. As expected, the performance of the jackknife variance estimator is better for larger sample sizes, for ratio imputation and for total expense as auxiliary variable (this variable has a higher correlation with the y-variable, as shown in table 1) .
The relative bias is large for small sample sizes, then decreases and increases again when the sampling fraction becomes non-negligible. This effect has been reported by other authors ( Lee, Rancourt, and

Särndal, 1995, Full, S, 1999 ). The coverage rate is not close to the true one even for large samples. This is probably due to the skewed and heavy-tailed distributions of the variables.

## V. CONCLUSIONS

14. From the point of view of jackknife performance ratio, imputation should be used instead of mean and hot deck imputation whenever auxiliary information is available. The use of residuals and the hot deck imputation is not recommended over plain ratio and mean imputation from the point of view of the jackknife performance.

15. The definition of imputation classes should improve the quality of both the imputation and the jackknife performance, as also should the stratification of the sample. This is actually under research and will be presented in the future.

**References**

Full, S. ((1999). Estimating variance due to imputation in ONS business surveys. International Conference on Survey Nonresponse. Portland, Oregon.

Kover, J.G and Whitridge, P.J (1995). Imputation of business survey data. Business survey methods. Eds B.G. Cox, D.A. Binder,B.N. Chinnappa, A. Christianson, M.J.Colledge, P.S. Kott), pp 403-423. New Yoerk . Wiley.

Lee, H, Rancourt, E and Särndal, C. (1995). Variance estimation in the presence of imputed data for the Generalizad estimation System, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 384-389.

Rancourt, E., Lee, H. and Särndal, C. (1994). Variance estimation under more than one imputation method. International Conference on establishment Surveys, pp 374-379.

Rao, J.N.K and Shao, J (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika,79,4, pp 811-822.