INVERSION PROCEDURES FOR SYSTEMATICALLY RANDOMLY ROUNDED INCOME DATA

Émile Allie[1]

ABSTRACT

Natural rounding of data by respondents creates bias in the estimation of an average. Systematic rounding of income data worsens the problem. This paper examines alternative inversion procedures and concludes that the use of fiscal information solves the bias problem in the estimation of an average. The inversion procedures are applied to systematically rounded data and their outcome is evaluated for wages and salaries, self-employment income non-farm, and two truncated distributions: Old Age Security and Guaranteed Income Supplement, and Canada/Quebec Pension Plan.

KEY WORDS: Income; Inversion; Rounding; Systematic

RÉSUMÉ

Les arrondissement naturels des données par les répondants créent des biais dans l'estimation des moyennes. Les procédures d'arrondissement systématiques empirent le problème. Ce texte examine des procédures d'inversion alternatives et conclue que l'utilisation de données fiscales résout le problème du biais dans l'estimation de la moyenne. Les procédures d'inversion sont appliquées à des données systématiquement arrondies et les résultats sont évalués pour les salaires et traitements, les revenus de travail autonome non-ferme et deux distributions tronquées : la sécurité de la vieillesse et le supplément de revenu garanti, et les prestation du Régime des pensions du Canada et du Régime des rentes du Québec.

MOTS CLÉS : Arrondissement, inversion, systématique, revenu

1. INTRODUCTION

1.1 Description of the Problem

In most surveys on income, rounding from respondents occurs. To insure confidentiality of published data, some statistical agencies in U.S.A. and the public release of the Survey of Labour and Income Dynamics (SLID) 1998 from Statistics Canada are systematically randomly rounding income data. Data analysts are now facing a new set of problems.

For example, an analysis of income distribution may provide large changes in the outcome, given the income interval selected. Any 'threshold measure' like the low-income level is now highly sensitive to the rounding interval and threshold values. In a microsimulation model, it could also generate false conclusions on the number of gainers or losers from changes in policies because people are grouped at rounded income values. The worse case scenario is when all the action involving changes in policies fall between rounded values. The estimated impact of the policy is null while it is supposed to generate some changes in disposable income.

1.2 Organization of the Paper

Section 2 reviews the literature. Sections 3 will show that systematically rounded data generate bias in an average estimate, and examines the impact of two inversion procedures on this bias. Section 4 evaluates the two inversion procedures on two continuous distributions (wages and salaries and self-employment income (non-farm)) and two truncated distributions (OAS/GIS, CPP/QPP) from SLID Public Release 1998, Statistics Canada. Section 5 concludes.

2. LITERATURE REVIEW

Tricker (1984) shows that rounding can disturb the estimated moments of the distribution. Rowe and

[1]Émile Allie, Personal Income Tax Division, Department of Finance, 16th Floor, 140 O'Connor Street, Ottawo, Ontario, K1A 0G5, Allie.Emile@fin.gc.ca.

Gribble (1994) conclude that rounding can cause spurious trends and/or discontinuities. Brachmann et al. (1996) argue that rounding can be problematic when exact data are required in specific statistical measures.

To solve the respondent rounding problems, Qian (1996) evaluates different approaches including a theoretical frequency distribution applied to data. Building on this idea, we suggest to use fiscal distribution to solve the rounding problem.

3. ESTIMATION OF AVERAGE UNDER ROUNDING AND INVERSION PROCEDURES

3.1 The rounding procedure

A given income, e.g. 24,431, is assigned to a rounding class, e.g. 200. Then a probability of selection is assigned to rounded values in the rounded interval $24{,}000 \pm 200$ which is the set of values {24,200, 24,300, 24,400, 24,500, 24600} . A random number help selecting the final random rounded value (FRRV), e.g. 24,600. The final rounded value 24,600 could be generated by many values. Let us define the rounding interval as the set of all values generating a given rounded value.

If we define w_x as the probability of observing the original value x, and $\Sigma w_x = 1$, then $\mathrm{E(X)}$ is the weighted average of the ^{x}x, where x is any positive integer value.

$$\mathrm{E(X)} = \sum_x w_x \quad x\,, \tag{1}$$

and x could be seen as a random variable.

Let us define:

p^*_{xy} the probability of x being rounded to y, where p^* is a symmetric distribution and $\Sigma_y p^*_{xy} = 1$ for $\forall i$,

v_{xy} is the error generated by the rounding process as the difference between y and x,

then, the expected value $\mathrm{E(Y)}$ is a biased measure of $\mathrm{E(X)}$ because $E(vxy) \neq 0$. This is because the distribution of x in a rounding interval is not symmetric around the rounded value y but follow a decreasing distribution like most income distributions.

$$\mathrm{E(Y)} = \mathrm{E(X)} + \sum_x w_i \sum_y p^*_{xy} v_{xy}\,. \tag{2}$$

3.2 The Inversion procedures

3.2.1 Bias in Inversion Procedures

An inversion procedure applies a density function to rounded values to spread the data on the rounding interval.

Let us define:

g_{xyz} be the probability that an original value x is rounded to y and generates an inverted value z on the rounding interval such that $\Sigma_z g_{xyz} = 1 \quad \forall x, \quad \forall y$;

ζ_{xyz}, is the error associated with the difference between the inverted value z and the original data x when rounded to y.

The expected value of z, $\mathrm{E(Z)}$, is biased only if the last term in the right hand side of (3) is different from zero.

$$\mathrm{E(Z)} = \mathrm{E(X)} + \sum_x w_x \sum_y p^*_{xy} \sum_z g_{xyz}\, \zeta_{xyz}\,. \tag{3}$$

3.2.2 Simple Inversion Procedure

In a simple inversion procedure, the value of y is symmetrically distributed so that the expected value of z, given y is y.

$$E(z \mid y) = y\,. \tag{4}$$

This mean the noise added by the simple inversion procedure has an expected value of zero, and the simple inversion procedure still produces a biased expected value of x.

3.2.3 Fiscal Inversion Procedure

Building on Qian's idea of a theoretical distribution to inverse the respondent rounded data, we used fiscal distribution to represent the original survey distribution to inverse the systematically rounded data.

Then the probability g_{xyz} is the product of the symmetric distribution used to round data and the fiscal distribution underlying the original distribution. The last right hand side of (3) can be written as:

$$\sum_x w_x \sum_y p^*_{xy} \sum_z g_{xyz}\, \zeta_{xyz} = \sum_x w_x \sum_y p^*_{xy} \sum_z g_{xyz}\,(z - i) = \sum_x w_x (i - i) = 0 \tag{5}$$

This is because, by construction, $\Sigma_y p^*_{xy} \Sigma_z g_{xyz}\, z = i$.
So, the use of fiscal information on systematically rounded data produces an unbiased estimate of the mean.

4. EVALUATION OF THE INVERSION PROCEDURES

4.1 The algorithm

We used rounded data from SLID, Public Release 1998, Statistics Canada. The original data are in SLID Internal

Database and are confidential[2]. The simple inversion procedure is easy to implement. The problem starts with the use of fiscal data because the distribution of values in a rounding interval must be calculated for each rounded value in the initial database.

The algorithm we developed is the following:

- sort rounded data by income rounded values;
- produce income distribution from fiscal data;
- select a rounded income value and generate the distribution of values associated with the rounded value; for each observation with the selected rounded value, randomly assign a value from the fiscal distribution;
- the fiscal distribution is corrected for the weight of the observation because in selecting one value we are selecting as many times the observation weight relative to the total weight of all the observations having the same rounded value.

Each observation in the database is assigned to its province and the algorithm was applied separately to each province.

4.2 Evaluation

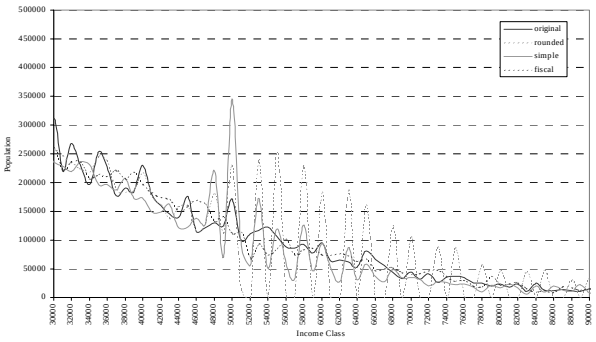
The evaluation of the inversion procedures is based on four income variables: wages and salaries, self-employment income non-farm, Old Age Security and Guaranteed Income Supplement (OAS/GIS), and Canada and Quebec Pension Plan (CPP/QPP). The two last variables, which are truncated distributions because there is a maximum to the level of benefits, will be discussed together.

4.2.1 Wages and Salaries

The following graph shows the Wages and Salaries distribution on the segment of $30,000 to $90,000. The original data in SLID is the continuous dark line. We can see that even if a large part of the income information in SLID is coming from the respondent tax file, a significant number of the other respondents may be rounding their responses as illustrated by peaks of observations around "natural" rounding values. There might be also normal rounding associated with a tendency to offer annual pay in rounded values like $50,000 per year.

Figure 1 shows the wages and salaries on the 30K to 90K interval. The distribution of the original data is the

Figure 1
Wages and Salaries $30,000-$90,000
Canada
1998



continuous dark line. Rounded values are in dash gray. The simple inversion procedure is the continuous gray line and the fiscal inversion is the dark dash line. We can see that rounding creates fixed intervals in the income distribution. When the rounded data have a high frequency, there is a high probability that the simple inversion procedure keeps the same peak, sometimes a little bit out of phase with the rounded date, as shown by the continuous gray line. The fiscal inversion produces a distribution slightly closer to the original SLID data.

If we consider an inversion procedure as a process similar to the minimization of the quadratic error between the imputed value (rounded value, simple inversion, fiscal inversion) and the original data, we can use the R^2 as a goodness of fit measure. Table 1 shows the value for wages and salaries. We observe that fiscal inversion procedure produces a better fit than the simple inversion procedure. We also know from the previous section that fiscal inversion produces unbiased estimate of the mean while even original data might have produced a biased estimate because of "natural" rounding of respondents.

4.2.2 Self-Employment Income Non-Farm

The results for the self-employment income non-farm are similar to those observed for wages and salaries, as illustrated in Figure 2, for income between $50,000 and $90,000. The size of the problem here is roughly 1/20 of the problem related to wages and salaries.

Table 1 Wages and Salaries Goodness of Fit

Rounded/Original	0.982
Rounded/Original	0.980
Rounded/Original	0.973

2 To implement Statistics Canada procedures on confidential data, all the data processing was done at Statistics Canada and only final aggregated distributions were sent to the author.

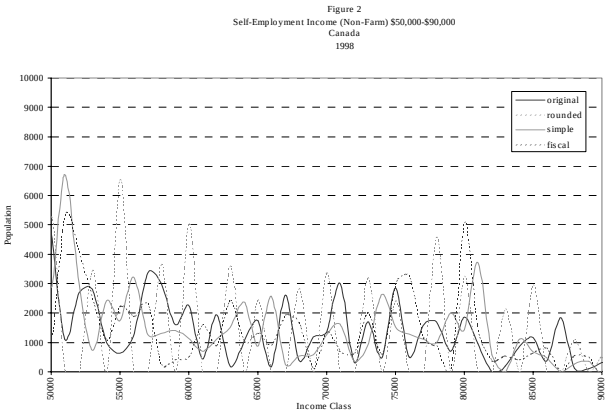


Figure 2 shows the self-employment income non-farm on the 60K to 80K interval. The distribution of the original data is the continuous dark line. Rounded values are in dash gray. The simple inversion procedure is the continuous gray line and the fiscal inversion is the dark dash line. This is similar to what we observe in Figure 1 for wages and salaries.

From Table 2, it seems that fiscal inversion is marginally less efficient than the simple inversion, but we should remember that simple inversion produces a biased average.

We may have to review the fiscal imputation procedure because it seems to produce a high peak of observation around $80,000. This may be because in some provinces the fiscal file may not contain enough observations in the imputation interval.

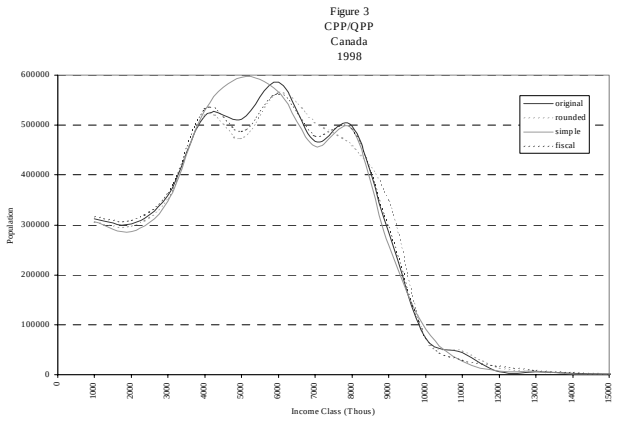
4.2.3 Two Truncated Distributions

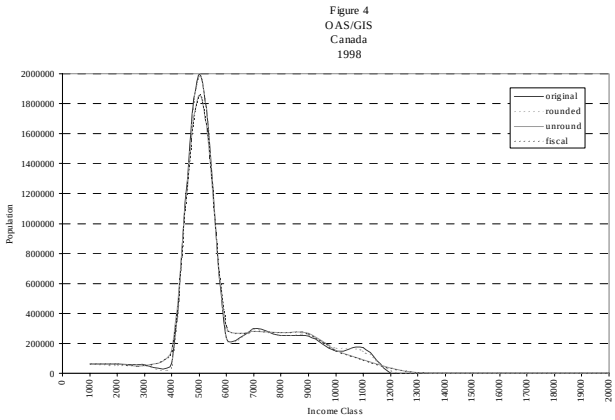
The number of observations imputed for the two truncated distributions, OAS/GIS and CPP/QPP, may be relatively large but they are concentrated in very few values. In these cases, it seems that simple inversion or fiscal inversion produce the same kind of results as illustrated in Figures 3 and 4.

One reason for the relatively good performance of even the rounding procedure is the small size of the rounding interval.

Table 2 Self-Employment Income Goodness of Fit

Rounded/Original	0.953
Rounded/Original	0.951
Rounded/Original	0.953





5. CONCLUSIONS

One reason for rounding data in SLID was to ensure the confidentiality of the fiscal information provided by respondents in SLID who gave access to their fiscal information instead of answering the income part of the survey.

It is well known that “natural” rounding is creating problems in the estimated moments of the distribution, systematic rounding worsens the situation. Following Qian (1996) who suggested using a theoretical distribution to solve the rounding problem, we were able to show that fiscal information as the representation of the actual distribution solves the bias problem generated by systematically rounded data.

If the fiscal inversion were applied to the original data it might have solved the bias generated by the “natural” rounding of respondents and at the same time preserved the confidentiality of the respondents.

In some provinces, the use of a Canadian distribution instead of a provincial distribution might help to solve some problems associated with rare densities in the rounding interval.

ACKNOWLEGEMENT

This research started at the Socioeconomic Modeling Group, Statistics Canada. Special thanks go to Shawna Brown who helped to complete it after the author left to the Department of Finance, and to J-F Beaumont for comments. The author remains responsible for any mistakes remaining.

REFERENCES

Bachman, Klaus, and Andreas Stich and Mark Trede, «Rounding Errors in Income Data», *Seminar für Wirtshafts- und Sozialstatistik, University zu Köln*, 1996.

Qian, Jiahe, «Restoration of Data with Rounding and Bounding Errors», *Proceedings – Section on Survey Research Methods American Statistical Association*, 1996, vol 1, p. 446-451.

Rowe, Geoff and Steve Gribble, «Income Statistics from Survey Data: Effects of Respondent Rounding», *Proceedings – American Statistical Association Business and Economic Statistics Section*, 1994, p. 77-82.

Tricker, A.R. «Effects of Rounding on the Moment of a Probability Distribution», *The Statistician*, vol 33, 1984, p. 381-390.