Survey Reweighting for Tax Microsimulation Modelling

John Creedy

NEW ZEALAND TREASURY

WORKING PAPER 03/17

SEPTEMBER 2003

THE TREASURY

Kaitohutohu Kaupapa Rawa

NZ TREASURY WORKING PAPER 03/17

Survey Reweighting for Tax Microsimulation Modelling

MONTH/YEAR

September 2003

AUTHOR

John Creedy
New Zealand Treasury
PO Box 3724
Wellington 6000
NEW ZEALAND

Email john.creedy@treasury.govt.nz
Telephone 64-4-471-5009
Fax [64 4 473 1151]

ACKNOWLEDGEMENTS

This paper arose from Treasury modelling of hypothetical reforms to the New Zealand tax and transfer system. I should like to thank Ivan Tuckwell for helpful discussions and for providing HES data in the required form. I have also benefited from discussions with Guyonne Kalb and Nathan McClellan, and comments by Mike Doherty and Dean Hyslop on an earlier version.

NZ TREASURY

New Zealand Treasury
PO Box 3724
Wellington 6000
NEW ZEALAND

Email information@treasury.govt.nz
Telephone 64-4-472 2733
Website www.treasury.govt.nz

Abstract

This paper describes a range of 'minimum distance' methods used to compute new weights for large cross-sectional surveys used in microsimulation modelling. Extraneous information about a range of population variables is used for calibration purposes. An iterative solution procedure is described and numerical examples are given, involving comparisons among alternative distance functions. An application to the New Zealand Household Economic Survey (HES) is reported.

JEL CLASSIFICATION	C30 - Econometric Methods: - Multiple/Simultaneous Equation Models - General
	C42 - Survey Methods
	C61 - Optimization Techniques; Programming Models; Dynamic Analysis
KEYWORDS	Household surveys; calibration; survey weights

Table of Contents

List of Tables

List of Figures

Survey Reweighting for Tax Microsimulation Modelling

1 Introduction

Tax microsimulation models are based on large-scale cross-sectional survey data. Each individual or household has a sample weight provided by the statistical agency responsible for collecting the data. The typical starting point is to use weights that are inversely related to the probability of selecting the individual in a random sample, with some adjustment for non-response. It has become common for agencies, using 'minimal' adjustments, to produce revised weights to ensure that, for example, the estimated population age/gender distributions match population totals obtained from other sources, in particular census data. Such calibration methods appear to be well known among survey statisticians, a highly influential paper being that by Deville and Särndal (1992).[1]

Users of official data usually take the weights as given, when 'grossing up' from the sample in order to obtain estimates of population values. This applies not only to simple aggregates, such as income taxation, or the number of recipients of a particular social transfer, or the number of people in a particular age group, but the weights are also used in the estimation of measures of population inequality or poverty. However, there is no guarantee that weights calibrated on demographic variables produce appropriate revenue, expenditure and income distribution results.

One aim of this paper is therefore to describe the basic calibration approach to economic modellers who are not familiar with the survey literature but need to reweight their samples. This may arise, for example, if population aggregates, not used for official calibrations, are not sufficiently close to population values obtained from other data sources, such as tax and benefit administration data. A further important reason for wanting to reweight the data arises when a survey from one year is used to examine the likely implications of, say, a tax and transfer policy in a later year. This need can arise if cross-sectional surveys are not carried out every year or if there are long delays in releasing data. Nevertheless, other administrative data may be available at more frequent intervals. It is also useful to be able to allow for changes in, say, the age distribution of the population or in aggregate unemployment rates over time.

[1] A detailed description of calibration and Generalised Regression (GREG) methods used in Belgium is given in Vanderhoeft (2001), which also describes the SPSS based program g-CALIB-S. Bell (2000) describes methods used in the Australian Bureau of Statistics household surveys, involving the SAS software GREGWT. Statistics Sweden uses the SAS software CLAN, described by Andersson and Nordberg (1998) and also used by the Finnish Labour Force Survey. All results in the present paper were obtained using Fortran programs written by the author.

The basic problem of obtaining 'minimum distance' weights is described more formally in section 2. The chi-squared distance function has an explicit solution and this is derived in section 3.[2] A more general class of distance measures is discussed in section 4, where iterative solutions are needed. These sections provide a simplified exposition, with derivations, of some of the results stated by Deville and Särndal (1992), whose more sophisticated and comprehensive treatment concentrated on statistical inference issues.[3] The use of Newton's method for the solution of the nonlinear equations is explored. Numerical examples are used to compare alternative distance functions, based on a small hypothetical sample. Finally, in section 5 the methods are applied to New Zealand Household Economic Survey (HES) data. Brief conclusions are in section 6.

2 The problem

For each of K individuals in a sample survey, information is available about J variables; these are placed in the vector:

$$x_k = \begin{bmatrix} x_{k,1} \\ . \\ . \\ . \\ x_{k,J} \end{bmatrix} \tag{1}$$

For present purposes these vectors contain only the variables of interest for the calibration exercise (rather than all measured variables). Many of the elements of x_k are likely to be $0/1$ variables. For example $x_{k,j} = 1$ if the k th individual is in a particular age group (or receives a particular type of social transfer), and zero otherwise. The sum $\sum_{k=1}^{K} x_{k,j}$ therefore gives the number of individuals in the sample who are in the age group (or who receive the transfer payment).

Let the sample design weights (provided by the statistical agency responsible for data collection) be denoted s_k for $k = 1,...,K$. These weights can be used to produce estimated population totals, $\hat{t}_{x|s}$ based on the sample, given by the J -element vector:

$$\hat{t}_{x|s} = \sum_{k=1}^{K} s_k x_k \tag{2}$$

The problem examined in this paper can be stated as follows. Suppose that other data sources, for example census or social security administrative data, provide information about 'true' population totals, t_x. The problem is to compute new weights, w_k, for $k = 1,...,K$ which are as close as possible to the design weights, s_k, while satisfying the set of J calibration equations:

$$t_x = \sum_{k=1}^{K} w_k x_k \tag{3}$$

[2] The link between this method and Generalised Regression estimators of population totals is discusssed briefly at the end of the section. See especially Särndal *et al.* (1992).

[3] Deville and Särndal (1992) used fewer than two pages to state the results discussed here.

It is thus necessary to specify a criterion by which to judge the closeness of the two sets of weights.

In general, denote the distance between w_k and s_k as $G(w_k, s_k)$. The aggregate distance between the design and calibrated weights is thus:[4]

$$D = \sum_{k=1}^{K} G(w_k, s_k) \tag{4}$$

The problem is therefore to minimise (4) subject to (3). The Lagrangean for this problem is:

$$L = \sum_{k=1}^{K} G(w_k, s_k) + \sum_{j=1}^{J} \lambda_j \left(t_{x,j} - \sum_{k=1}^{K} w_k x_{k,j} \right) \tag{5}$$

where λ_j for $j = 1, ..., J$ are the Lagrange multipliers. The following two sections consider methods of obtaining values of w that minimise (5).

3 An explicit solution

The constrained minimisation problem stated above has an explicit solution for a distance function based on the chi-squared measure. This is discussed in subsection 1. A numerical example is examined in subsection 2.

3.1 The Chi-squared distance measure

Consider the chi-squared type of distance measure, where the aggregate distance is given by:

$$D = \frac{1}{2} \sum_{k=1}^{K} \frac{(w_k - s_k)^2}{s_k} \tag{6}$$

The Lagrangean in (5) can be written as:

$$L = \frac{1}{2} \sum_{k=1}^{K} \frac{(w_k - s_k)^2}{s_k} + \sum_{j=1}^{J} \lambda_j \left(t_{x,j} - \sum_{k=1}^{K} w_k x_{k,j} \right) \tag{7}$$

where the λ_j, for $j = 1, ..., J$, are the Lagrange multipliers, and $t_{x,j}$ represents the jth element of the vector of known population aggregates, t_x.

[4] Some authors, such as Folson and Singh (2000) write the distance to be minimised as $\sum_{k=1}^{K} s_k G(w_k, s_k)$, but the present paper follows Deville and Särndal (1992).

Differentiation of (7) gives the set of K first-order conditions:

$$\frac{\partial L}{\partial w_k} = \left(\frac{w_k}{s_k} - 1 \right) - \sum_{j=1}^{J} \lambda_j x_{k,j} = 0 \tag{8}$$

for $k = 1, ..., K$, along with the J conditions in (3). Rewriting $\sum_{j=1}^{J} \lambda_j x_{k,j}$ as $x_k' \lambda$, where the prime indicates transposition, and multiplication of each equation in (8) by s_k gives, after rearrangement:

$$w_k = s_k \left(1 + x_k' \lambda \right) \tag{9}$$

for $k = 1, ..., K$.

To solve for the Lagrange multipliers, pre-multiply (9) by x_k and rearrange, so that:

$$w_k x_k - s_k x_k = s_k x_k x_k' \lambda \tag{10}$$

Summing (10) over all K, and making use of the calibration equations, gives:

$$t_x - \hat{t}_{x|s} = \left[\sum_{k=1}^{K} s_k x_k x_k' \right] \lambda \tag{11}$$

where the term in brackets on the right hand side of (11) is a J by J square matrix. Hence, if this matrix can be inverted, the vector of Lagrange multipliers is given by:

$$\lambda = \left[\sum_{k=1}^{K} s_k x_k x_k' \right]^{-1} \left(t_x - \hat{t}_{x|s} \right) \tag{12}$$

The resulting values of λ are substituted into (9) to obtain the new weights.[5]

3.2 A small example

The above procedure may be illustrated using a simple example. Suppose there are four variables, $x_{k,j}$ (for $j = 1, ..., 4$), of concern, for which population values t_x are available. The hypothetical data, for a sample of 20 individuals, are shown in Table 1. Suppose variable 1 refers to age, so that $x_{k,1} = 1$ for those who are 'young' and is zero otherwise, while $x_{k,2} = 1$ for those who are unemployed, and zero otherwise. Variable 3 measures

---

[5] Write (9) as $w_k = s_k \left(1 + \lambda' x_k \right)$ and (12) as $\lambda' = \left(t_x - \hat{t}_{x|s} \right)' T^{-1}$ with T as the symmetric matrix $\sum_{k=1}^{K} s_k x_k x_k'$. Given sample observations on the variable y_k, an estimate of the population total, $\hat{t}_y$, can be obtained as $\sum_{k=1}^{K} w_k y_k$. Substituting for w_k gives the result in Deville and Särndal (1992, p.377) that $\hat{t}_y = \sum_{k=1}^{K} s_k y_k + \left(t_x - \hat{t}_{x|s} \right)' B$, where $B = T^{-1} \sum_{k=1}^{K} s_k x_k y_k$. This provides the link between reweighting and the Generalised Regression (GREG) estimator. The production of asymptotic standard errors is often based on this estimator, in view of the result that other distance functions are asymptotically equivalent; see Deville and Särndal (1992, p.378). The present discussion concentrates only on reweighting.

earnings from employment, while variable 4 is another categorical variable referring to location ($x_{k,4}=1$ if the individual lives in a city, and is zero otherwise).[6] Given the sample design weights shown in the penultimate column of Table 1, the estimated population totals are equal to $\hat{t}_{x|s}=\left[44,24,213,32\right]^{'}$.

The symmetric matrix $\sum_{k=1}^{K}s_k x_k x_k^{'}$ and its inverse are given in Table 2. The zero elements reflect the property of the basic data, that only individuals who work (for whom $x_{k,2}=0$) are assumed to receive positive earnings, $x_{k,3}$. Suppose that the known population totals are $t_x=\left[50,20,230,35\right]^{'}$, reflecting a younger population than in the sample weights and a lower unemployment rate. The resulting calibrated weights are shown in the final column of Table 1.

Table 1 – Sample values and calibrated weights

k	$x_{k,1}$	$x_{k,2}$	$x_{k,3}$	$x_{k,4}$	s_k	w_k
1	1	1	0	0	3	2.753
2	0	1	0	0	3	2.109
3	1	0	2	0	5	5.945
4	0	0	6	1	4	4.005
5	1	0	4	1	2	2.484
6	1	1	0	0	5	4.589
7	1	0	5	0	5	5.752
8	0	0	6	1	4	4.005
9	0	1	0	0	3	2.109
10	0	0	3	1	3	3.120
11	1	0	2	0	5	5.945
12	1	1	0	1	4	3.985
13	1	0	3	1	4	5.019
14	1	0	4	0	3	3.490
15	0	0	5	0	5	4.678
16	0	1	0	1	3	2.345
17	1	0	2	1	4	5.070
18	0	0	6	0	5	4.614
19	1	0	4	1	4	4.967
20	0	1	0	0	3	2.109

[6]The number of variables needed is of course one less than the number of categories of each type, otherwise singularity problems arise.

Table 2 – Matrix $\sum_{k=1}^{K} s_k x_k x_k'$ and its inverse

$\sum_{k=1}^{K} s_k x_k x_k'$			
44.000	12.000	101.000	18.000
12.000	24.000	0.000	7.000
101.000	0.000	981.000	101.000
18.000	7.000	101.000	32.000

$\left[\sum_{k=1}^{K} s_k x_k x_k'\right]^{-1}$			
0.037	-0.016	-0.003	-0.008
-0.016	0.053	0.003	-0.011
-0.003	0.003	0.002	-0.005
-0.008	-0.011	-0.005	0.053

The required adjustments to the weights can clearly be seen to be consistent with expectations, given the calibration requirements and the characteristics of the individuals. For example, the weights for individuals 2, 9 and 20 fall by a relatively large amount (from 3 to 2.109), since these individuals are all unemployed, old and living in rural locations, for all of which the aggregates are required to fall. The weights for individuals 1 and 6 do not drop so far because, although these are unemployed and in a rural location, they are young. The weight for person 12 falls by a small amount because, although unemployed, this person is young and in a city. The weights for individuals 13, 17 and 19 increase by relatively large amounts as they are young, employed and living in a city.

4 Alternative Distance Functions

The chi-squared distance function is convenient because it enables an explicit solution for the calibrated weights to be obtained, requiring only matrix inversion. However, a modified form of the same approach can be applied to a range of alternative distance functions, as shown in this section. These functions belong to a class of functions having two features: the first derivative with respect to w can be expressed as a function of w/s and its inverse can be obtained explicitly. An interactive solution procedure is required for the calculation of the Lagrange multipliers. The general case of this class is presented in subsection 1. An iterative approach based on Newton's method is described in subsection 2. Several weighting functions are described in subsection 3 and illustrated in subsection 4.

4.1 The general case

The Lagrangean for the general case, stated in section 2, was written as:

$$L = \sum_{k=1}^{K} G(w_k, s_k) + \sum_{j=1}^{J} \lambda_j \left(t_{x,j} - \sum_{k=1}^{K} w_k x_{k,j} \right) \tag{13}$$

Suppose that $G(w_k, s_k)$ has the property, shared with the chi-square distance function, that the differential with respect to w_k can be expressed as a function of the ratio w_k/s_k, so that:

$$\frac{\partial G(w_k, s_k)}{\partial w_k} = g\left(\frac{w_k}{s_k}\right) \tag{14}$$

The K first-order conditions for minimisation can therefore be written as:

$$g\left(\frac{w_k}{s_k}\right) = x_k'\lambda \tag{15}$$

Write the inverse function of g as g^{-1}, so that if $g(w_k/s_k) = u$, say, then $w_k/s_k = g^{-1}(u)$. In the case of the chi-square distance function used above, $g(w_k/s_k) = w_k/s_k - 1$, and the inverse takes a simple linear form. In general, from (15) the k values of w_k are expressed as:

$$w_k = s_k g^{-1}\left(x_k'\lambda\right) \tag{16}$$

If the inverse function, g^{-1}, can be obtained explicitly, equation (16) can be used to compute the calibrated weights, given a solution for the vector, λ.

As before, the Lagrange multipliers can be obtained by post-multiplying (16) by the vector x_k, summing over all $k = 1, ..., K$ and using the calibration equations, so that:

$$t_x = \sum_{k=1}^{K} w_k x_k = \sum_{k=1}^{K} s_k g^{-1}\left(x_k'\lambda\right) x_k \tag{17}$$

Finally, subtracting $\hat{t}_{x|s} = \sum_{k=1}^{K} s_k x_k$ from both sides of (17) gives:

$$t_x - \hat{t}_{x|s} = \sum_{k=1}^{K} s_k \left\{g^{-1}\left(x_k'\lambda\right) - 1\right\} x_k \tag{18}$$

The term $s_k\left\{g^{-1}\left(x_k'\lambda\right) - 1\right\}$ is of course a scalar, and the left hand side is a known vector. In general, (18) is nonlinear in the vector λ and so must be solved using an iterative procedure, as described in the following subsection.

4.2 An iterative procedure

Writing $t_x - \hat{t}_{x|s} = a$, the equations in (18) can be written as:

$$f_i(\lambda) = a_i - \sum_{k=1}^{K} s_k x_{k,i} \left\{g^{-1}\left(x_k'\lambda\right) - 1\right\} = 0 \tag{19}$$

for $i = 1,...,J$. The roots can be obtained using Newton's method, described in the Appendix. This involves the following iterative sequence, where $\lambda^{[I]}$ denotes the value of λ in the I th iteration:[7]

$$\lambda^{[I+1]} = \lambda^{[I]} - \left[\frac{\partial f_i(\lambda)}{\partial \lambda_\ell}\right]^{-1}_{\lambda^{[I]}} \left[f(\lambda)\right]_{\lambda^{[I]}} \tag{20}$$

The Hessian matrix $\left[\partial f_i(\lambda)/\partial \lambda_\ell\right]$ and the vector $f(\lambda)$ on the right hand side of (20) are evaluated using $\lambda^{[I]}$.

The elements $\partial f_i(\lambda)/\partial \lambda_\ell$ are given by:

$$\frac{\partial f_i(\lambda)}{\partial \lambda_\ell} = -\sum_{k=1}^{K} s_k x_{k,i} \frac{\partial g^{-1}\left(x_k'\lambda\right)}{\partial \lambda_\ell} \tag{21}$$

which can be written as:

$$\frac{\partial f_i(\lambda)}{\partial \lambda_\ell} = -\sum_{k=1}^{K} s_k x_{k,i} x_{k,\ell} \frac{\partial g^{-1}\left(x_k'\lambda\right)}{\partial \left(x_k'\lambda\right)} \tag{22}$$

Starting from arbitrary initial values, the matrix equation in (20) is used repeatedly to adjust the values until convergence is reached, where possible.

As mentioned earlier, the application of the approach requires that it is limited to distance functions for which the form of the inverse function, $g^{-1}(u)$, can be obtained explicitly, given the specification for $G(w,s)$. Hence, the Hessian can easily be evaluated at each step using an explicit expression for $dg_k^{-1}\left(x_k'\lambda\right)/d\left(x_k'\lambda\right)$. As these expressions avoid the need for the numerical evaluation of $g^{-1}\left(x_k'\lambda\right)$ and $dg_k^{-1}\left(x_k'\lambda\right)/d\left(x_k'\lambda\right)$ for each individual at each step, the calculation of the new weights can be expected to be relatively quick, even for large samples.[8] However, it must be borne in mind that a solution does not necessarily exist, depending on the distance function used and the adjustment required to the vector $t_x - \hat{t}_{x|s}$.

4.3 Some distance functions

One reason why the chi-squared distance function produces a solution is that no constraints are placed on the size of the adjustment to each of the survey weights. It is therefore also possible for the calibrated weights to become negative. However, Deville and Särndal (1992) suggested the following simple modification to the chi-squared function, although the explicit solution for the chi-squared case is no longer available and the iterative method must be used.

---

[7] The approach described here differs somewhat from other routines described in the literature, for example in Singh and Mohl (1996) and Vanderhoeft (2001). However, it provides extremely rapid convergence.

[8] Using numerical methods to solve for each $g^{-1}(u)$ and $dg^{-1}(u)/du$, for $u = x_k'\lambda$, for every individual in each iteration, would increase the computational burden substantially.

Suppose it is required to constrain the proportionate changes to certain limits, different for increases compared with decreases in the weights. Define r_L and r_U such that $r_L < 1 < r_U$. The objective is to ensure that, for increases, the proportionate change, $w/s - 1$, is less than $r_U - 1$, or that $r_U > w/s$. For decreases, the aim is to ensure that $1 - w/s$ (or the negative of the proportional change) is less than $1 - r_L$, so that $r_L < w/s$.

For the chi-squared distance function, it has been seen that $g^{-1}(u) = 1 + u$, where $u = x'\lambda$ and $g^{-1}(u)$ solves for w/s. Hence if $g^{-1}(u) = w/s$ is outside the specified range, it is necessary to set it to the relevant limit, either r_U or r_L, rather than allow it to take the value generated. Since $g^{-1}(u) - 1 = w/s - 1 = u$, it is clear that the limits are exceeded if $u < r_L - 1$ and if $u > r_U - 1$. In each case where the value of $g^{-1}(u)$ has to be set to the relevant limit, the corresponding value of $dg^{-1}(u)/du$ is zero. This approach ensures that weights are kept within the range, $r_L s_k < w_k < r_U s_k$. Hence, negative values of w are avoided simply by setting r_L to be positive.[9]

It has been seen above that the solution procedure requires only an explicit form for the inverse function $g^{-1}(u)$, from which its derivative can be obtained. It is not necessary to start from a specification of $G(w, s)$. Deville and Särndal (1992) suggest the simple form:

$$g^{-1}(u) = \left(1 - \frac{u}{2}\right)^{-2} \tag{23}$$

The gradient function, $g(w/s)$, is given by solving (23) for u, so that:

$$g\left(\frac{w}{s}\right) = u = 2\left(1 - \left(\frac{w}{s}\right)^{-1/2}\right) \tag{24}$$

and the form of the distance function can be obtained by integrating (24).[10] This is referred to as Case A, and its properties are given in the first row of Table 3. The second row of the table provides details of Case B, where $g^{-1}(u) = (1-u)^{-1}$, and the final row gives the corresponding properties of the basic chi-squared function.[11] A feature of these functions is that they do not require any parameters to be set.

[9] This is much more convenient than imposing inequality constraints and applying the more complex Kuhn-Tucker conditions. Also, it is desirable to restrict the extent of proportional changes even where they produce positive weights.

[10] Hence it is required to obtain $2\int\left\{1 - \left(\frac{w}{s}\right)^{-1/2}\right\}dw = 2\left(w - 2\sqrt{s}\sqrt{w}\right)$, which can be written as $2\left(w + s - 2\sqrt{s}\sqrt{w}\right) - 2s$, and dropping the last term, which is a constant, this is equal to $2\left(\sqrt{w} - \sqrt{s}\right)^2$.

[11] Deville and Särndal (1992) discuss the use of a normalisation whereby $g^{-1'}(0)$ is set to some specified value, but this is not necessary for the approach.

Table 3 – Alternative distance functions

Case	$G(w,s)$	$g(w/s)$	$g^{-1}(u)$	$dg^{-1}(u)/du$
A	$2\left(\sqrt{w}-\sqrt{s}\right)^2$	$2\left(1-\left(\frac{w}{s}\right)^{-1/2}\right)$	$\left(1-\frac{u}{2}\right)^{-2}$	$\left(1-\frac{u}{2}\right)^{-3}$
B	$-s\log\left(\frac{w}{s}\right)+w-s$	$1-\left(\frac{w}{s}\right)^{-1}$	$(1-u)^{-1}$	$(1-u)^{-2}$
Chi-squared	$(w-s)^2/2s$	$\frac{w}{s}-1$	$1+u$	1

Deville and Särndal (1992) also suggest the use of an inverse function $g^{-1}(u)$ of the form:[12]

$$g^{-1}(u)=\frac{r_L(r_U-1)+r_U(1-r_L)\exp\alpha u}{(r_U-1)+(1-r_L)\exp\alpha u} \tag{25}$$

where r_L and r_U are as defined above and:

$$\alpha=\frac{r_U-r_L}{(1-r_L)(r_U-1)} \tag{26}$$

Thus $g^{-1}(-\infty)=r_L$ and $g^{-1}(\infty)=r_U$, so that the limits of w/s are r_L and r_U. This function therefore has the property that adjustments to the weights are kept within the range, $r_L s_k < w_k < r_U s_k$, although, unlike the chi-squared modification, no checks have to be made during computation.

The derivative required in the computation of the Hessian is therefore:

$$\frac{dg^{-1}(u)}{du}=g^{-1}(u)\left\{r_U-g^{-1}(u)\right\}\frac{(1-r_L)\alpha\exp\alpha u}{(r_U-1)+(1-r_L)\exp\alpha u} \tag{27}$$

Since $g^{-1}(u)$ solves for w/s, (25) can be rearranged, by collecting terms in $\exp\alpha u$, to give:

$$\frac{\frac{w}{s}-r_L}{1-r_L}=\frac{r_U-\frac{w}{s}}{r_U-1}\exp\alpha u \tag{28}$$

so that the gradient of the distance function is:

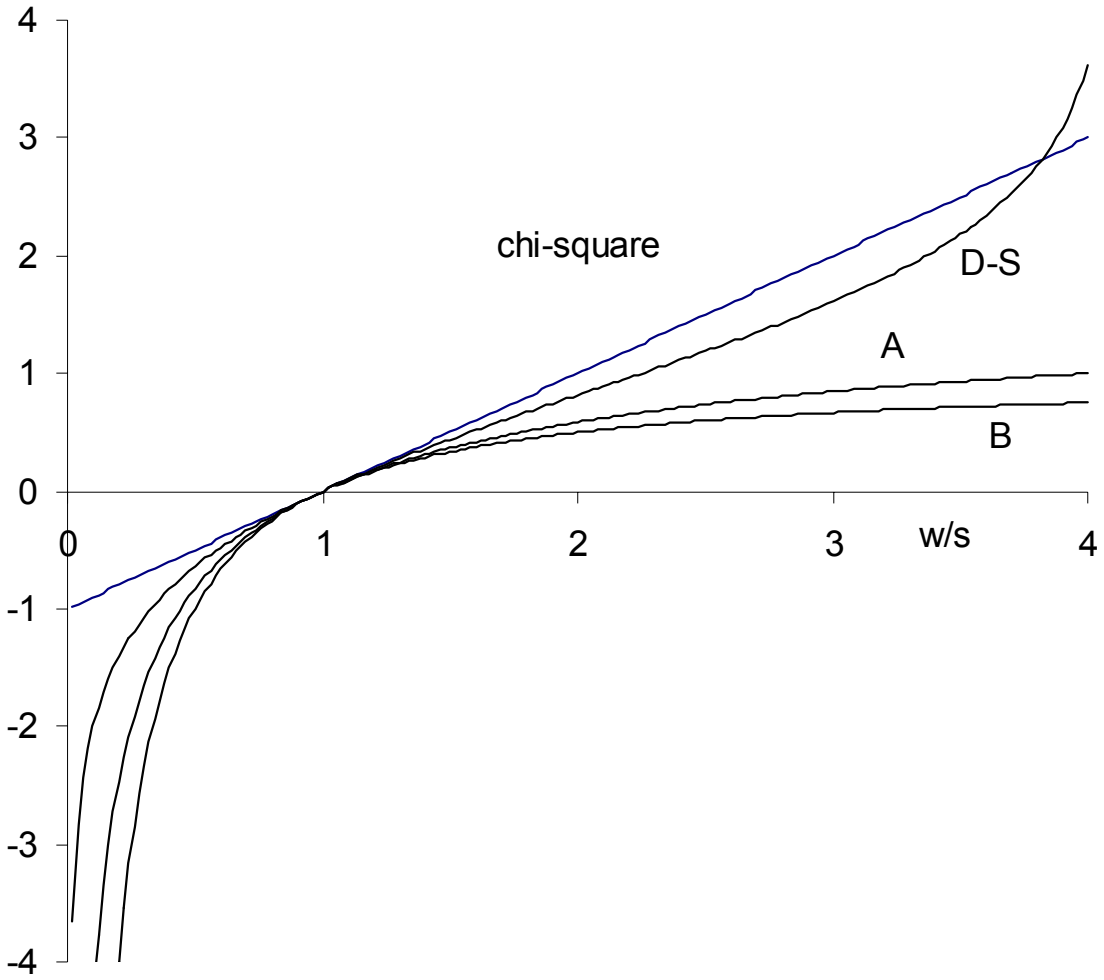
$$g\left(\frac{w}{s}\right)=u=\frac{1}{\alpha}\left[\log\left(\frac{\frac{w}{s}-r_L}{1-r_L}\right)-\log\left(\frac{r_U-\frac{w}{s}}{r_U-1}\right)\right] \tag{29}$$

The special nature of this gradient function is illustrated by the line D-S in Figure 1, which shows the profile of (29) for the wide range where $r_U=4.1$ and $r_U=0.01$. The first characteristic of the S-D function that is evident is the restriction of w/s to the range

[12] Singh and Mohl (1996), in reviewing alternative calibration estimators, refer to this 'inverse logit-type transformation' as a Generalised Modified Discrimination Information method.

specified. Figure 1 also shows the function $g(w/s)$ for the other cases discussed above. In all cases, the slope is zero (corresponding to a turning point of the distance function) when $w/s = 1$. Given the quadratic U-shaped nature of the chi-squared distance function, the gradient increases at a constant rate, being negative in the range $w/s < 1$. Cases A and B also imply U-shaped distance functions, but with the gradient increasing more sharply for $w/s < 1$ and more slowly than the chi-square function in the range $w/s > 1$.

Figure 1 – Alternative gradient functions



The distance function is given by integrating (29) with respect to w. It is most convenient to apply the variate transformation $x = w/s$, so that $dw = sdx$, and it is required to obtain:

$$\frac{s}{\alpha}\int\left[\log\left(\frac{x-r_L}{1-r_L}\right)-\log\left(\frac{r_U-x}{r_U-1}\right)\right]dx \qquad (30)$$

Using the result that:

$$\int \log\left(\frac{x-a}{b}\right)dx = (x-a)\left[\log\left(\frac{x-a}{b}\right)-1\right] \tag{31}$$

and:

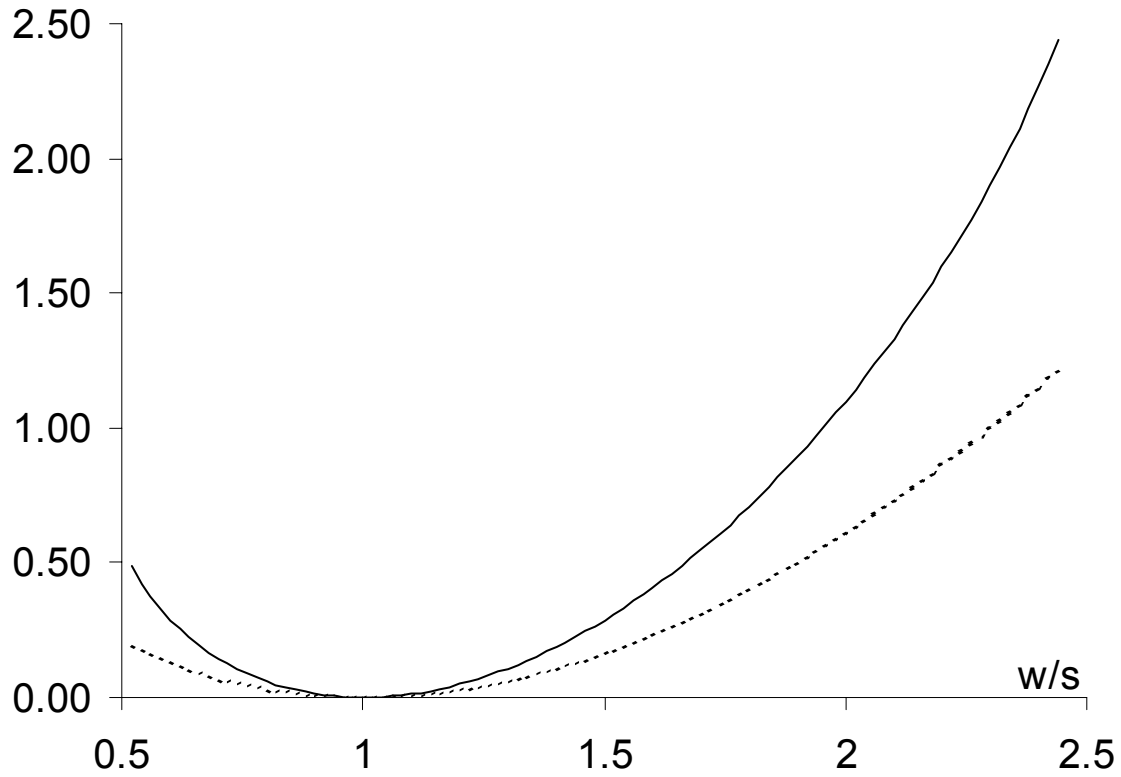
$$\int \log\left(\frac{a-x}{b}\right)dx = -(a-x)\left[\log\left(\frac{a-x}{b}\right)-1\right] \tag{32}$$

substitution and rearrangement gives s/α multiplied by:

$$G(w,s) = \left(r_U - \frac{w}{s}\right)\log\left(\frac{r_U - \frac{w}{s}}{r_U - 1}\right) + \left(\frac{w}{s} - r_L\right)\log\left(\frac{\frac{w}{s} - r_L}{1 - r_L}\right) \tag{33}$$

plus a term $(r_U - r_L)s/\alpha$, which, since it is a constant, may be dropped without loss.[13] Examples of this distance function are shown in Figure 2.[14]

Figure 2 – Deville-Särndal distance functions



[13] Equation (33) is the result stated without proof by Deville and Särndal (1992, p. 378).

[14] Folsom and Singh (2000) propose a variation on this, which they call a 'generalised exponential model', in which the limits are allowed to be unit-specific. In practice they suggest the use of three sets of bounds for low, medium and high initial weights.

4.4 Further Numerical Examples

The application of the distance functions presented in the previous subsection to the hypothetical sample used earlier gives the results shown in Table 4, where the simple (unrestricted) chi-squared results are added for comparison. In cases where limits are imposed on the degree of adjustment of the weights, it cannot be expected that a solution will always be available. For this reason, care is needed in the choice of r_L and r_U, as discussed below.

The values of $r_U = 3$ and $r_L = 0.2$ were initially selected as being well outside the range of ratios obtained using the other distance functions. When the range was reduced to the potentially restrictive values of $r_U = 1.3$ and $r_L = 0.8$, none of the ratios obtained was actually at the limits specified. Nevertheless, the change to the weighting function produces a different set of weights, as shown by comparisons in Table 4: some actually move further away from their initial, or survey, weights.

Table 4 – Revised weights using alternative distance functions

k	s_k	A	B	$r_U = 3$ $r_L = 0.2$	$r_U = 1.3$ $r_L = 0.8$	$r_U = 1.25$ $r_L = 0.8$	Chi-squared
1	3.000	2.674	2.654	2.706	2.513	2.483	2.753
2	3.000	2.228	2.260	2.178	2.408	2.400	2.109
3	5.000	5.998	6.012	5.976	6.162	6.187	5.945
4	4.000	3.944	3.926	3.974	3.951	4.019	4.005
5	2.000	2.514	2.521	2.501	2.534	2.493	2.484
6	5.000	4.456	4.423	4.510	4.189	4.138	4.589
7	5.000	5.729	5.717	5.747	5.911	6.094	5.752
8	4.000	3.944	3.926	3.974	3.951	4.019	4.005
9	3.000	2.228	2.260	2.178	2.408	2.400	2.109
10	3.000	3.086	3.074	3.106	3.213	3.325	3.120
11	5.000	5.998	6.012	5.976	6.162	6.187	5.945
12	4.000	3.814	3.762	3.897	3.645	3.769	3.985
13	4.000	5.108	5.136	5.065	5.094	4.990	5.019
14	3.000	3.490	3.487	3.494	3.604	3.680	3.490
15	5.000	4.665	4.666	4.665	4.442	4.314	4.678
16	3.000	2.370	2.380	2.355	2.428	2.408	2.345
17	4.000	5.191	5.232	5.128	5.115	4.993	5.070
18	5.000	4.603	4.604	4.600	4.366	4.237	4.614
19	4.000	5.028	5.043	5.001	5.069	4.986	4.967
20	3.000	2.228	2.260	2.178	2.408	2.400	2.109

The choice of $r_U = 1.25$ and $r_L = 0.8,$ shown in the penultimate column of the table, actually places some adjustments to the weights at the lower limit of the range: for individuals 2, 9 and 20, the value of w/s is equal to 0.8. However, no adjustments are at the upper range specified. If r_L is raised to 0.82 (with r_U unchanged), unreported results show that individual 16 is placed at the lower limit, along with 2, 9 and 20 as before; in addition individuals 5, 13, 17, 19 are pushed to the upper limit of 1.25. The attempt to raise r_L to 0.83 means that no solution is possible. However, if r_U is set to the higher value of $3.0,$ then $r_L = 0.83$ is found to be the highest value (where the range of variation is limited to the second decimal point) of r_L for which a solution is possible. The two highest ratios needed in this case are for persons 13 and 17, who have w/s values of 1.328 and 1.376 respectively. If r_L is kept at this value of $0.83,$ the lowest value of r_U for which a solution exists is $r_U = 1.26$. In this case individuals 1, 2, 6, 9, 16 and 20 are placed at the lower limit and individuals 5, 13, 17 and 19 are placed at the upper limit. Clearly, some care needs to be exercised in the choice of upper and lower limits.

While these examples help to explore the characteristics of the different approaches, it is necessary to examine the practical implementation of the method. This is carried out in the following section.

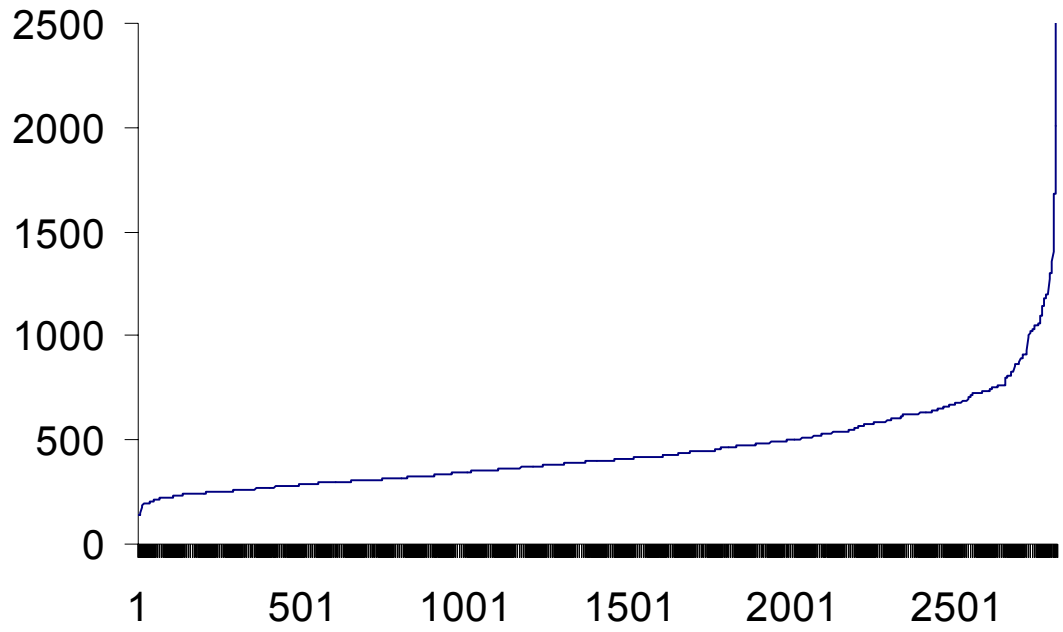
5 The NZ Household Economic Survey

This section applies the above approaches to the New Zealand Household Economic Survey 2000/01, which is the latest survey available. The aim is to illustrate the application of the approach in a practical context, and to compare the performance of the alternative distance functions. At this point it may be useful to stress that reweighting may cause non-calibrated variables to change in undesirable ways, so that various other checks need to be made.[15]

The variation in the survey weights provided by Statistics NZ for the period 2000/01 is illustrated in Figure 3, where the weights are arranged in ascending order for a sample of 2808 households.[16] It can be seen that the majority of these weights are within a fairly narrow range, although some are substantially higher, suggesting a considerable degree of under-representation of these household types in the sample.

For present purposes, new weights were obtained using calibration values for 2003/4, therefore allowing for population changes. A total of 36 calibration equations were used, covering the total numbers in the following categories for: 11 family composition types; 16 age/sex types; 2 unemployment benefits; 2 Domestic Purpose Benefits; 2 invalidity benefits; 2 sickness benefits; and 1 widow's benefits.[17]

Figure 3 – Survey weights



[15]The precision of some survey estimates may also be lowered, particularly where many calibration constraints are used. Examples are given in Skinner (1999); see also Kalton and Flores-Cervantes (2003).

[16]These are integrated weights, not the original weights. For a discussion of the use of integrated weighting, as described by Lemaître and Dufour (1987), by Statistics New Zealand, see StatsNZ (2001).

[17]For each of these types, there is of course one additional category not used. The motivation for selecting these variables involved the use of the data for projecting taxes and benefit expenditures. For a general discussion of variable selection, see Nascimento Silva and Skinner (1997).

Figure 4 – Calibrated weights: Chi-Square

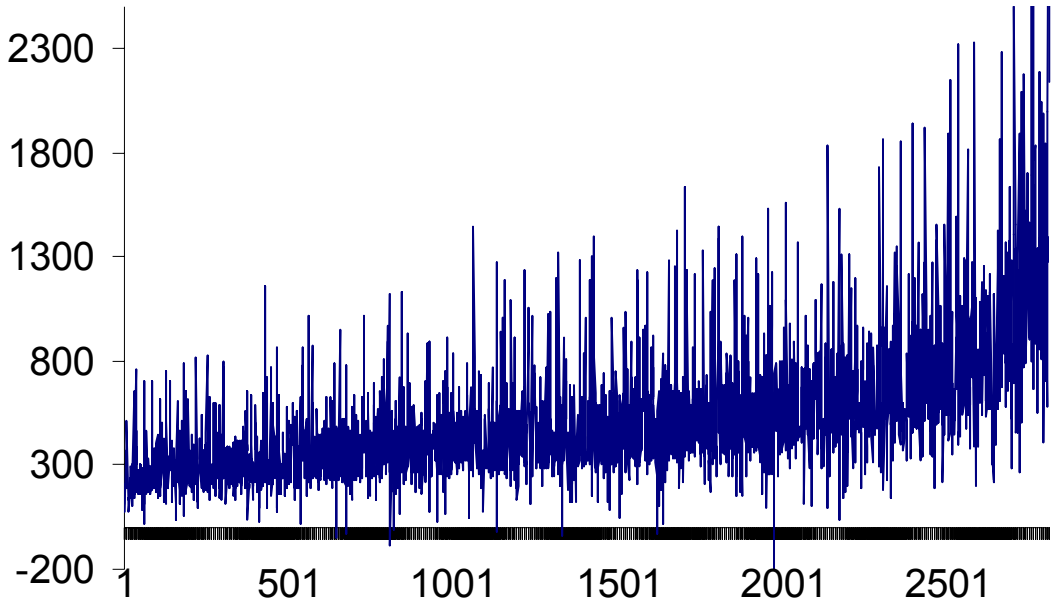
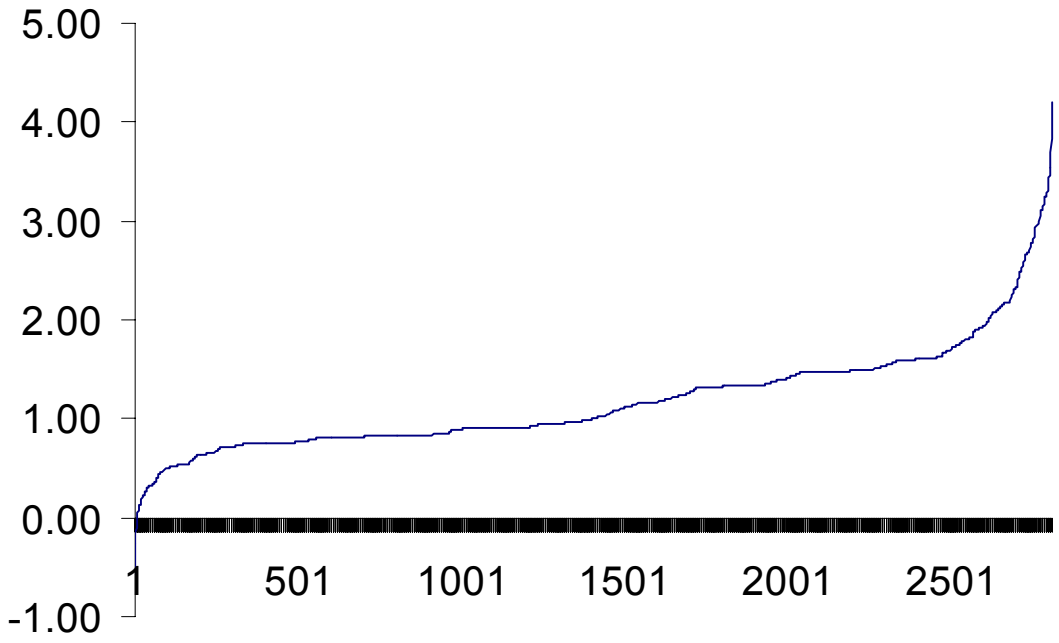


Figure 5 – Ratio of calibrated to survey Weights: Chi-Square

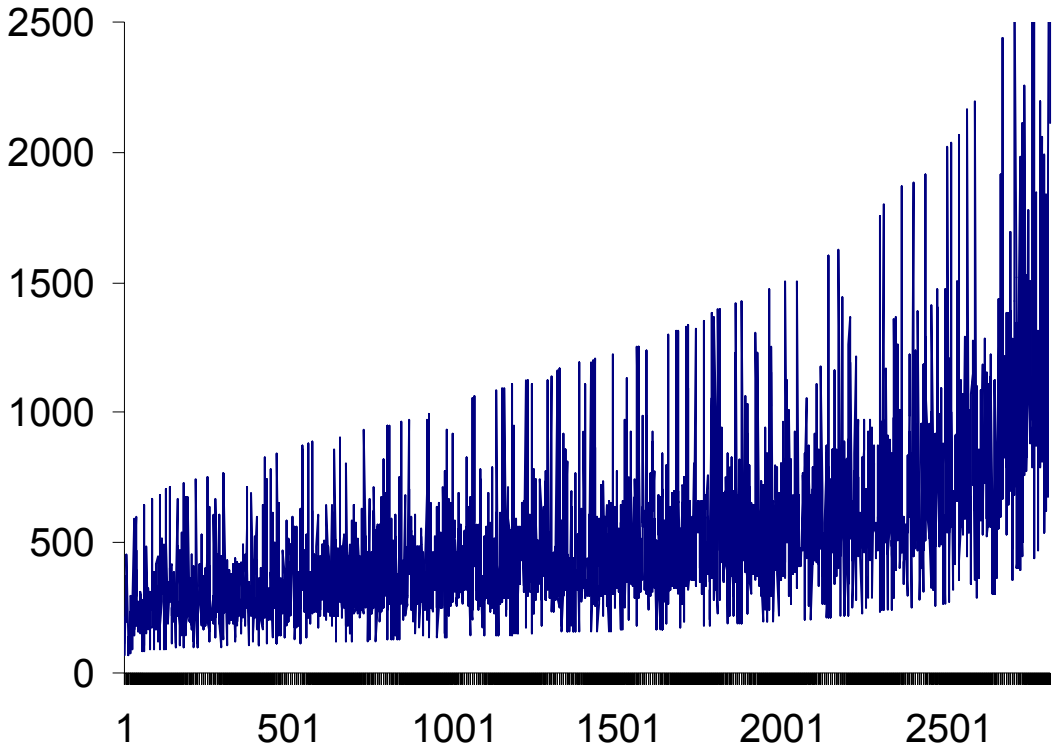


The calibrated weights obtained using the basic chi-square distance function are shown in Figure 4, where households are arranged in the same order as in Figure 3 (although the vertical axis has been truncated at 2,500). The corresponding ratios of calibrated to survey weights, w/s, are displayed in increasing order in Figure 5. This clearly shows considerable variability in the weights, with some negative weights resulting. Using the modified chi-square distance function, allowing the ratio of weights, w/s, to be restricted with the limits $r_U = 3.0$ and $r_L = 0.4$, resulted in values displayed in Figure 6. The effects of the adjustment can clearly be seen in the extent to which the new weights are restricted

to a range of variation around the initial profile. The top and bottom of the profile in Figure 6 are substantially 'smoothed'; clearly a significant number are placed at the limits, particularly at the lower limit.

The calibrated weights obtained using the distance function in (33), allowing for upper and lower limits to w/s of $r_U = 3$ and $r_L = 0.4$ are shown in Figure 7, and the ratios are shown in ascending order in Figure 8. In both cases where limits were imposed, the range shown is the narrowest for which a solution was obtained (that is, for which the iterative method used to obtain the Lagrange multipliers converged).[18] The main difference between the modified chi-square case and the distance function in (33) is that the former appears to push more values to the lower edge of the profile.

Figure 6 – Calibrated weights: Modified Chi-Squared function



[18]Where a solution was not available, the procedure 'exploded' relatively quickly, after just a few iterations. Otherwise convergence was achieved rapidly.

Figure 7 – Calibrated weights: Deville-Särndal function

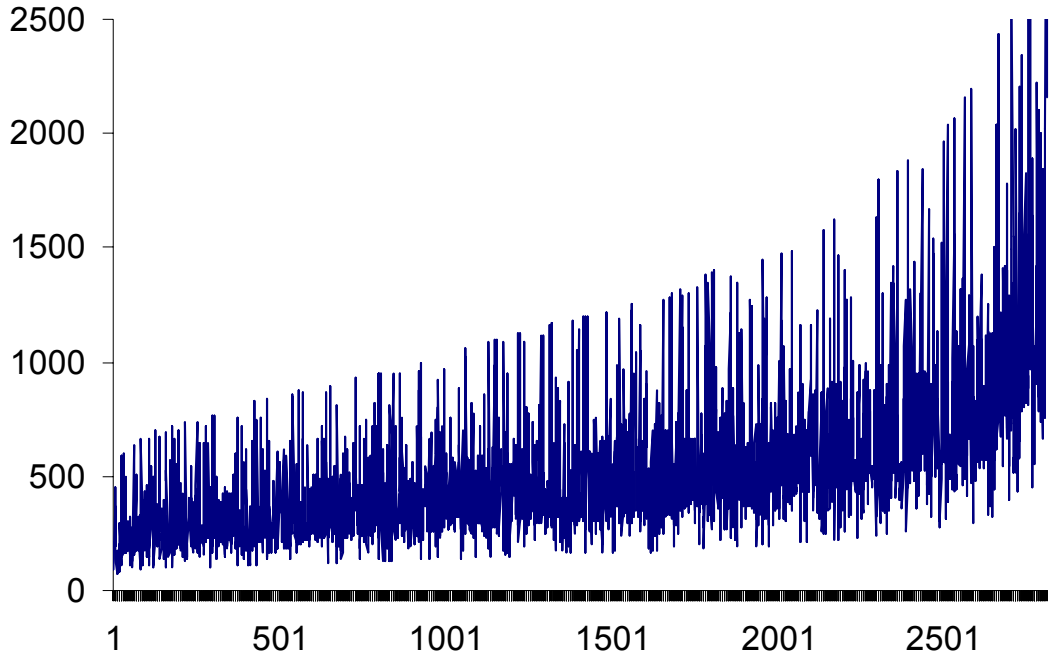
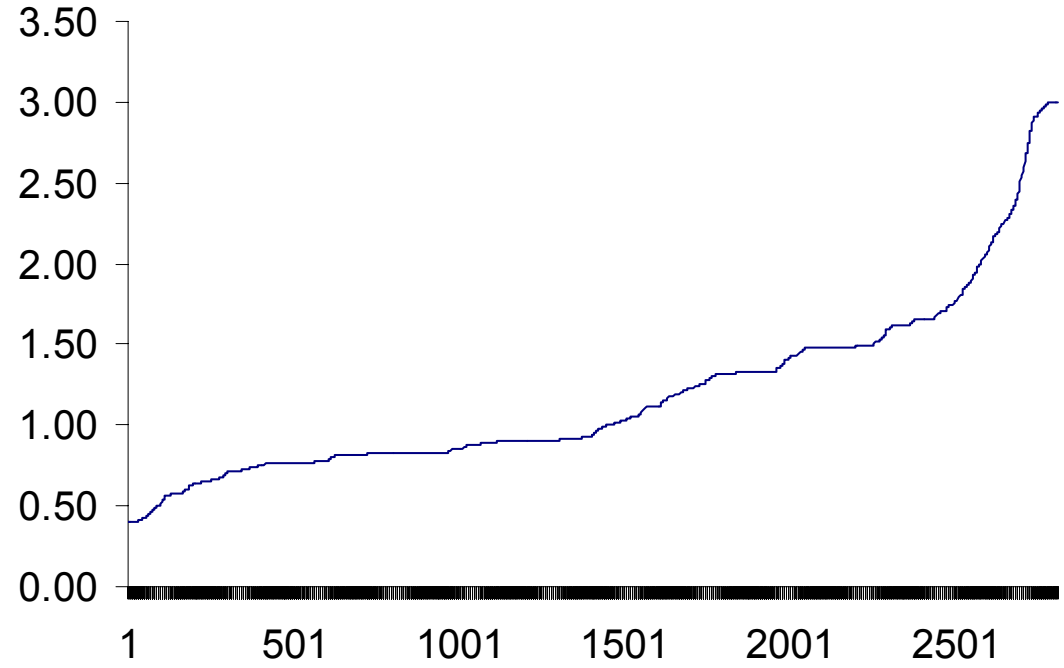


Figure 8 – Ratio of calibrated to survey weights: Deville-Sarndal function



The use of the other two distance functions failed to produce solutions. Comparing the results for the distance functions producing solutions, it seems that the only serious contenders are the two cases imposing constraints on the proportionate changes in weights. The numerical values of the limits which can be imposed, while still obtaining a solution, appear to be the same for the adjusted chi-squared function and the function in equation (33). Where solutions are available, there seems little to choose between those

two cases. However, in further experiments using a larger number of calibration equations, it was found that no solution was available using the distance function in (33), however wide the range of variation allowed. Nevertheless a solution could be obtained using the modified chi-squared distance function. The standard chi-squared function also gave a solution, as expected, but this produced a number of negative weights.

6 Conclusions

This paper has examined a range of minimum distance methods used to compute new weights for large cross-sectional surveys used in microsimulation modelling. The methods involve the use of extraneous information about a range of population variables, for calibration purposes. The distance functions were restricted to those for which the first derivative can be expressed as a function of the ratio of the new weights to the survey weights, and for which that function can be inverted explicitly. In general, an iterative solution procedure is required. An approach based on Newton's method was described and numerical examples were given for several distance functions. Finally, the performance of the method was examined using the New Zealand Household Economic Survey. Rapid convergence of the iterations was obtained, although care needs to be taken when imposing limits on the proportional adjustments to sample weights. Since the same basic approach (and computer program) can easily examine a range of distance functions, and Newton's method converges extremely quickly, it is relatively costless to consider the full range of distance measures. However, in practice, convergence cannot be expected using all measures.

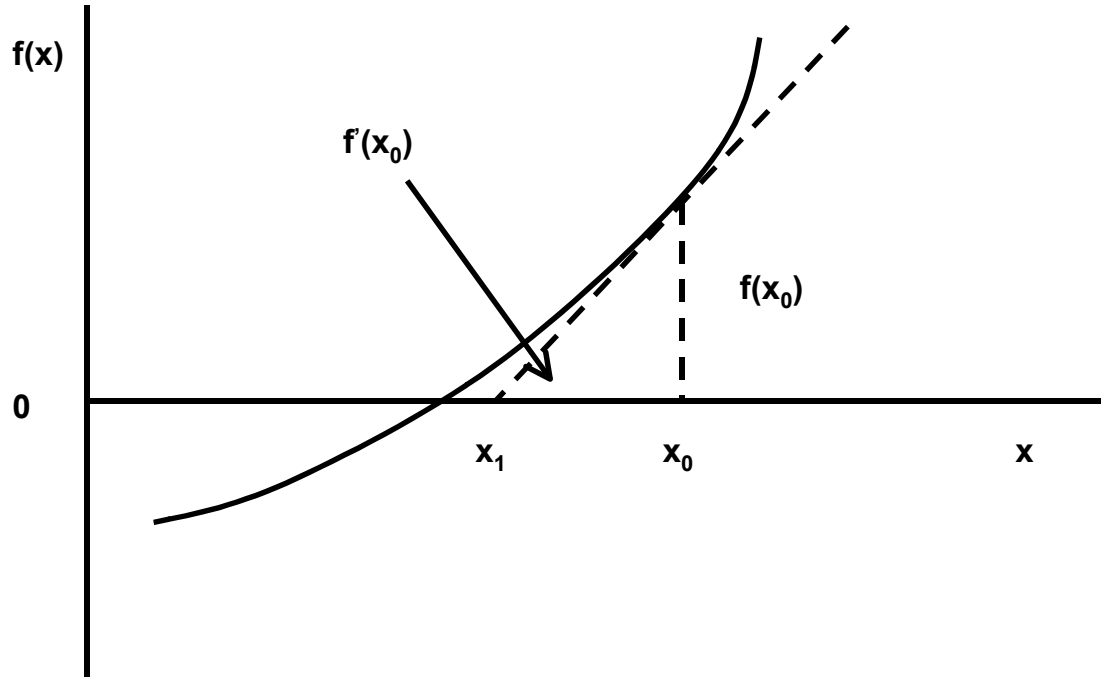
Finally it is worth remembering that reweighting may cause the distribution of important variables, in particular alternative sources of income, to change.[19] Checks on changes in a range of distributions are therefore recommended.

[19] This point is also made by Klevmarken (1998).

Appendix: Newton's method

Consider finding the root of the single equation in one variable, $f(x)=0$, where $f(x)$ takes the form shown in Figure 9. Newton's method involves taking an arbitrary starting point, x_0 and drawing the tangent, with slope $f'(x_0)$. By approximating the function by the tangent, the new value is given by the point of intersection of this tangent with the x axis, at x_1. Selecting x_1 as the next starting point leads quickly to the required root.

Figure 9 – Newton's method



From the triangle in Figure 9:

$$f'(x_0)=\frac{f(x_0)}{x_0-x_1} \tag{34}$$

Hence, starting from $I=0$, the sequence of iterations is:

$$x_{I+1}=x_I-\left\{f'(x_I)\right\}^{-1}f(x_I) \tag{35}$$

Convergence is reached when $x_{I+1}-x_I<\varepsilon$ and ε depends on the accuracy required. Newton's method is easily adapted to deal with a set of equations, $f_i(x)$, where x is a vector. The method involves repeatedly solving the following matrix equation, where $x^{[I]}$ now denotes the vector in the I th iteration and $f(x)$ is a vector containing the $f_i(x)$ values.

$$x^{[I+1]}=x^{[I]}-\left[\frac{\partial f_i(x)}{\partial x_\ell}\right]^{-1}_{x^{[I]}}\left[f(x)\right]_{x^{[I]}} \tag{36}$$

References

[1] Andersson, C. and Nordberg, L. (1998) A User's Guide to CLAN 97. *Statistics Sweden.*

[2] Bell, P. (2000) Weighting and standard error estimation for ABS household surveys. *Paper prepared for ABS Methodology Advisory Committee: Australian Bureau of Statistics.*

[3] Deville, J.-F. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, pp. 376-382.

[4] Statistics New Zealand (2001) Information Paper: The Introduction of Integrated Weighting to the 2000/2001 Household Economic Survey. *Statistics New Zealand.*

[5] Folson, R.E. Jnr. and Singh, A.C. (2000) The generalized exponential model for sampling weight calibration for extreme values, non-response and post-stratification. *Proceedings of the Survey Research Methods Section: American Statistical Association*. http://www.amstat.org/sections/srms/proceedings/papers/2000_099.pdf.

[6] Kalton, G. and Flores-Cervantes, I. (2003) Weighting methods. *Journal of Official Statistics*, 19, pp. 81-98.

[7] Klevmarken, N.A. (1998) Statistical inference in micro-simulation models: incoporating external information. *Uppsala University Department of Economics Working Paper.* http://www.nek.uu.se/Pdf/1998wp20.pdf.

[8] Lemaître, G. and Dufour, J. (1987) An integrated method for weighting persons and families. *Survey Methodology*, 13, pp. 199-207.

[9] Nascimento Silva, P.L.D. and Skinner, C. (1997) Variable selection for regression estimation in finite populations. *Survey Methodololgy*, 23, pp. 23-32.

[10] Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

[11] Singh, A.C. and Mohl, C.A. (1996) Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, pp. 107-115.

[12] Skinner, C. (1999) Calibration weighting and non-sampling errors. *Research in Official Statistics*, 2, pp. 33-43.

[13] Vanderhoeft, C. (2001) Generalised calibration at Statistics Belgium. *Statistics Belgium Working Paper,* no. 3.