Southampton Statistical Sciences Research Institute
University of Southampton

Outlier Robust Imputation of Survey Data via Reverse Calibration

R. Ren, R. Chambers

Abstract

Outlier robust methods of survey estimation, e.g. trimming, winsorization, are well known (Chambers and Kokic, 1993). However, such methods do not address the important practical problem of creating an “outlier free” data set for general and public use. In particular, what is required in this situation is a data set from which the outlier robust survey estimate can be recovered by the application of standard methods of survey estimation. In this paper we describe an imputation procedure for outlying survey values, called reverse calibration, that achieves this aim. This method can also be used to correct gross errors in survey data, as well as to impute missing values. The paper concludes with an evaluation of the method based on a realistic survey data set.

S^{3}RI Methodology Working Paper M03/19

Outlier Robust Imputation of Survey Data via Reverse Calibration

R. Ren
Survey Statistics Laboratory
CREST/ENSAI, Campus de Ker Lann
Rue Blaise Pascal. 35170 Bruz. France

R. L. Chambers
Department of Social Statistics
University of Southampton, Highfield
Southampton, SO17 1BJ. United Kingdom

August 2002

Abstract: Outlier robust methods of survey estimation, e.g. trimming, winsorization, are well known (Chambers and Kokic, 1993). However, such methods do not address the important practical problem of creating an "outlier free" data set for general and public use. In particular, what is required in this situation is a data set from which the outlier robust survey estimate can be recovered by the application of standard methods of survey estimation. In this paper we describe an imputation procedure for outlying survey values, called reverse calibration, that achieves this aim. This method can also be used to correct gross errors in survey data, as well as to impute missing values. The paper concludes with an evaluation of the method based on a realistic survey data set.

Key words: Auxiliary information, finite population, sample survey, outliers, gross errors, missing data, imputation, robust estimation, winsorization, calibration.

1. Introduction

Outlying data values are frequently encountered in sample surveys, particularly surveys measuring economic and financial phenomena. Chambers (1986) classifies these values into two groups. The first are representative outlier values. These are correctly measured sample values that are outlying relative to the rest of the sample data and for which there is no reason to believe that similar values do not exist in the non-sampled part of the survey population. The second group consists of non-representative outlier values. These are gross errors in the sample data, caused by deficiencies in survey processing (e.g. miscoding). Such errors have nothing to do with the values in the non-sampled part of the survey population. Either type of outlier can have a substantial impact on the eventual survey estimate if ignored. Typically, non-representative outliers are detected and corrected during the survey editing process, while representative outliers are handled in the survey estimation process, generally by the use of outlier robust or resistant estimation procedures.

Design-based approaches to dealing with outliers in survey estimation are described by Kish (1965), Searl (1966) and Hidiroglou and Srinath (1981). Chambers (1982, 1986) developed model-based outlier robust estimation techniques for sample surveys. Recent work in this area is described in Chambers and Kokic (1993), Lee (1991, 1995), Hulliger (1995), Welsh and Ronchetti (1998) and Duchesne (1999). The research described in this paper has been carried out within the Euredit Project (2000), which is aimed at the development and evaluation of new methods for editing and imputation, and in particular the development of imputation methods that can be used with outliers in survey data.

After carrying out survey estimation, the statistician often has to deliver a data set for general and public use. It is hard to imagine that a non-expert user of this data set will employ the same sophisticated robust techniques that the statistician has applied to those parts of the data set containing outliers. Consequently the survey statistician must deliver a "clean" data set, with outlier values appropriately modified, such that the data set is suitable for general use with standard statistical software. Ideally, this is where one can recover the results obtained from the robust estimation method using this standard software. This can be achieved by using an outlier imputation procedure that we call *reverse calibration*. In this paper we describe this method and compare it with more standard imputation methods that are typically used for imputation of missing data.

The structure of this paper is as follows: in the next section we describe the reverse calibration approach to outlier imputation. Since this procedure depends on the actual method used for outlier robust estimation used in the survey, we then describe some outlier robust estimation methods in section 3. In section 4 these methods are then used to determine outlier imputations, via reverse calibration, which are applied to a realistic survey data set. This data set has been created within the Euredit project and is based on the *Annual Business Inquiry* (hereafter abbreviated as the *ABI*) survey carried out by the UK *Office for National Statistics* (hereafter abbreviated as the *ONS*). We discuss the results of this evaluation in section five.

2. Outlier imputation by reverse calibration

Imputation methods have traditionally been used for missing data. The basic idea in this case is that, by "filling in" the missing values in a data set, standard methods of inference, which typically assume "complete" data, are applicable. In this section we take this idea and apply it to another common survey data problem. This is the presence of outliers in these data. As noted in the previous section, such outliers can be representative or non-representative. Once the outliers in the survey data have been identified and classified in this way, we can treat them appropriately. Non-representative outliers are very similar in concept to missing data. By definition these values are, for one reason or another, wrong. Consequently, they need to be changed back to their correct values. This can be done by re-interview of the survey respondents that provided these values, in the same way that one can carry out follow-up interviews of survey non-respondents. Alternatively these values can be replaced by imputed values derived from the non-outliers or "inliers" in the survey data set, similar to the way imputed values based on respondent data are used to replace missing data. Note that this approach makes the assumption that, conditional on known (and correct) values for covariates, the error creation process leading to non-representative outliers is independent of the process underpinning generation of the true values for these outliers.

Representative outliers, on the other hand, are more difficult to handle. By definition, there is nothing to be gained by re-interview of the respondents that provided them (beyond the knowledge that these values are in fact correct). Imputation of these values based on relationships in the inlier data values is also inappropriate, since these outlier values clearly do not have the same relationships. Modern outlier resistant methods of estimation allow for this difference, but control the impact of the corresponding outlier contribution to the overall survey estimate. What is required in this case is a method of outlier imputation that mimics this behaviour.

2.1 Reverse calibration imputation

A basic assumption is that all representative sample outliers are identifiable. To minimise notation, we initially assume that application of survey editing and follow-up procedures implies that there are no missing values or non-representative outliers in the sample data. That is, all outliers in these data are representative. Let s denote the sample of n units and let $\{w_i ; i \in s\}$ denote a <u>target</u> set of estimation weights that we wish to apply to <u>all</u> the sample values, outliers as well as inliers, in order to estimate the population total of interest. Often these weights will be the inverses of inclusion probabilities or regression (e.g. GREG or BLUP) weights. Their main characteristic is that they are known for each sample unit and are fixed. Our problem is then one of imputing sample data values such that when these imputed values are multiplied by the $\{w_i ; i \in s\}$ and summed over the sample, they then lead to an "acceptable" estimate of the population total.

By "acceptable" we mean here that this estimate equals one that we obtain when we apply an appropriate outlier resistant technique to the sample data. For example, suppose that

$$\hat{t}_y = \sum_{i \in s} w_i^* Y_i$$

is such an estimate, where the $\{w_i^* ; i \in s\}$ are outlier resistant weights. Then this condition is satisfied when

$$\hat{t}_y = \sum_{i \in s} w_i^* Y_i = \sum_{i \in s} w_i Y_i^*$$

where the $\{Y_i^* ; i \in s\}$ denote the imputed sample values. Let s_2 be the sub-sample of size n_2 consisting of the representative sample outliers and let s_1 be the sub-sample of size $n_1 = n - n_2$ that consists of the sample inliers. A natural restriction is $Y_i^* = Y_i ; i \in s_1$ and $w_i^* = w_i ; i \in s_1$ in which case the problem can be re-expressed as one of defining a set of imputed values $\{Y_i^* ; i \in s_2\}$ that satisfies

$$\hat{t}_y - \hat{t}_{1y} = \hat{t}_y - \sum_{i \in s_1} w_i Y_i = \sum_{i \in s_2} w_i Y_i^* = \hat{t}_{2y}. \quad (1)$$

A natural way of choosing the Y_i^*, $i \in s_2$ is so that they remain as close as possible to the true values Y_i, $i \in s_2$ subject to the constraint (1). In turn, this requires that we specify a distance measure $d(Y^*, Y)$ between the imputed values and the true values that must be then minimised subject to this constraint. It is easy to see that this is equivalent to a calibration problem where the survey variable Y plays the role of sample weight and the sample weight variable w plays the role of the survey variable. It is well known (Deville and Särndal, 1992) that

$$Y_i^* = Y_i F_i(w_i \lambda) \quad (2)$$

where $F_i(\cdot)$ is a calibration function that satisfies $F_i(0) = 1$, $F_i'(0) = q_i$ and λ is a constant determined by $\sum_{i \in s_2} w_i Y_i F_i(w_i \lambda) = \hat{t}_{2y}$.

Suppose that $Y > 0$. A simple distance measure is

$$d\left(Y^*, Y\right) = \sum_{i \in s_2} \left(Y_i^* - Y_i\right)^2 / 2 q_i Y_i \quad (3)$$

where $q_i > 0$, $i \in s_2$ are constants that can be chosen by the statistician. Using this distance measure, we have $F_i(t) = 1 + q_i t$ (Deville and Särndal, 1992). From (2) it follows

$$Y_i^* = Y_i \left[1 + q_i w_i \frac{\hat{t}_{2y} - \sum_{j \in s_2} w_j Y_j}{\sum_{j \in s_2} q_j w_j^2 Y_j} \right]. \quad (4)$$

The second term on the right-hand-side of (4) is negative if the outliers are mainly 'big' outliers, i.e. take values much larger than the values associated with the inliers in the sample. Consequently the observed true value Y_i associated with a representative outlier is decreased. In contrast this term is positive if the outliers are mainly 'small' outliers, i.e. take values much smaller than the values associated with the inliers in the sample. In this case the true value Y_i associated with a representative sample outlier is increased. This is consistent with the general idea of outlier modification or winsorization.

A potential advantage of reverse calibration imputation is that a calibration program *CALMAR* (Sautory, 1993) is available, containing several different distance functions $d\left(Y^*, Y\right)$. Standard choices of q_i are $q_i = 1$ or $q_i = d_i^{-1}$. In the latter case (4) simplifies to a ratio-type imputation

$$Y_i^* = Y_i \frac{\hat{t}_{2y}}{\sum_{j \in s_2} w_j Y_j}. \quad (5)$$

Note that neither (4) nor (5) guarantee that the imputed values satisfy editing rules (e.g. are positive). To prevent negative values, we can use one of the alternative distance measure proposed in Deville and Särndal (1992) or use the distance measure (4) with $q_i = d_i^{-1}$, which leads to ratio-type imputation (5). Alternatively, we can integrate the editing rules into the calibration procedure.

2.2 The general case

The reverse calibration method described above treats all outliers similarly. In particular they are all either decreased or increased in value. This is sensible if these values are all of one type, i.e. all big or all small. However, in practice outliers relative to a regression model for Y tend to be a mix of these two types, and these two different types of outliers need to be treated differently in imputation (the small outliers need to be increased and the big outliers need to be decreased). Furthermore, there are typically also missing values for Y in the sample data, and these need to be imputed at the same time as these outliers are imputed.

Suppose that a sample s is subject to both outlier and missing values. Let s_1 be the sub-sample of inliers and respondents, and let s_2 be the sub-sample consisting of outliers and missing values. Suppose also that a reliable population total estimate $\hat{t}_y$ is obtained by some outlier resistant procedure that takes non-response into account. Let $\hat{t}_{1y}$ be an estimate of the population total of the inliers and respondents. Then an estimate of the population of the outliers and non-respondents can be obtained as $\hat{t}_{2y} = \hat{t}_y - \hat{t}_{1y}$.

What we mean by a population here is open to interpretation. In fact, we have four populations (or, to be more precise, domains). These are the respondent inlier population, the nonrespondent population, the respondent “small outlier” population and the respondent “big outlier” population. We assume that our overall target population estimate can be broken down into four components that effectively represent our best estimates for the totals of each of these domains. Similarly we assume that the sample units can be divided among these four domains. The reverse calibration process is then straightforward. We adjust the observed sample values in each domain (including the respondent inlier domain if necessary) so that when multiplied by their target weights w_i they recover the respondent inlier + outlier components of the overall estimate. Finally, we impute sample values for the missing cases in order to recover the last component of the estimate.

To be more precise, let $s_2^{(+)}$ denote the responding sample units corresponding to large outliers, $s_2^{(-)}$ the responding sample units corresponding to small outliers, and $s_2^{(m)}$ the nonresponding sample units. The corresponding decomposition of the estimated population total is $\hat{t}_y = \hat{t}_{1y} + \hat{t}_{2y}^{(-)} + \hat{t}_{2y}^{(+)} + \hat{t}_{2y}^{(m)}$. The reverse calibrated imputed values are then given by

$$Y_i^* = \begin{cases} Y_i\left[1 + q_i w_i\left(\sum_{j\in s_1} q_j w_j^2 Y_j\right)^{-1}\left(\hat{t}_{1y} - \sum_{j\in s_1} w_j Y_j\right)\right], i \in s_1 \\ Y_i\left[1 + q_i w_i\left(\sum_{j\in s_2^{(-)}} q_j w_j^2 Y_j\right)^{-1}\left(\hat{t}_{2y}^{(-)} - \sum_{j\in s_2^{(-)}} w_j Y_j\right)\right], i \in s_2^{(-)} \\ Y_i\left[1 + q_i w_i\left(\sum_{j\in s_2^{(+)}} q_j w_j^2 Y_j\right)^{-1}\left(\hat{t}_{2y}^{(+)} - \sum_{j\in s_2^{(+)}} w_j Y_j\right)\right], i \in s_2^{(+)} \\ \tilde{Y}_i\left[1 + q_i w_i\left(\sum_{j\in s_2^{(m)}} q_j w_j^2 \tilde{Y}_j\right)^{-1}\left(\hat{t}_{2y}^{(m)} - \sum_{j\in s_2^{(m)}} w_j \tilde{Y}_j\right)\right], i \in s_2^{(m)} \end{cases} \quad (11)$$

where the values $\tilde{Y}_i$ represent initial (uncalibrated) imputed values for the missing data cases. An obvious choice for $\tilde{Y}_i$ is the fitted value for this case generated by the observed sample inliers, which corresponds to assuming that all nonrespondents are inliers. Observe that these imputed values lead to ratio type imputations when $q_i = w_i^{-1}$, while if w_i equals the inverse of the sample inclusion probability then we generally need to change the values of all the observed sample units (inliers as well as outliers) in order to achieve calibration.

A sufficient condition for the imputed values for the inliers to be the same as their observed values is when $\hat{t}_{1y} = \sum_{j\in s_1} w_j Y_j$. Consequently, if it is a requirement that inlier values remain unchanged, then we can define our estimate of the observed inlier contribution to the overall population total using this identity. This immediately leads to the restriction

$$\hat{t}_y - \sum_{j\in s_1} w_j Y_j = \hat{t}_{2y}^{(-)} + \hat{t}_{2y}^{(+)} + \hat{t}_{2y}^{(m)}.$$

Since it is unlikely that the domain estimates for the two outlier contributions and the missing inlier contribution will satisfy this restriction a priori, we need to modify these estimates so that they do. The easiest way to do this is by apportioning out the difference $\hat{t}_y - \sum_{j\in s_1} w_j Y_j$ among these estimates. This leads to modified domain estimates that need to be substituted for $\hat{t}_{2y}^{(-)}$, $\hat{t}_{2y}^{(+)}$ and $\hat{t}_{2y}^{(m)}$ in the reverse calibration formula above, given by

$$\hat{\hat{t}}_{2y}^{(-)} = \hat{t}_{2y}^{(-)} \left(\frac{\hat{t}_y - \sum_{j\in s_1} w_j Y_j}{\hat{t}_{2y}^{(-)} + \hat{t}_{2y}^{(+)} + \hat{t}_{2y}^{(m)}} \right)$$

$$\hat{\hat{t}}_{2y}^{(+)} = \hat{t}_{2y}^{(+)} \left(\frac{\hat{t}_y - \sum_{j\in s_1} w_j Y_j}{\hat{t}_{2y}^{(-)} + \hat{t}_{2y}^{(+)} + \hat{t}_{2y}^{(m)}} \right)$$

$$\hat{\hat{t}}_{2y}^{(m)} = \hat{t}_{2y}^{(m)} \left(\frac{\hat{t}_y - \sum_{j\in s_1} w_j Y_j}{\hat{t}_{2y}^{(-)} + \hat{t}_{2y}^{(+)} + \hat{t}_{2y}^{(m)}} \right)$$

respectively.

3. Outlier resistant estimation of population totals

In this section we briefly describe some outlier resistant estimators of population totals that can be used with the reverse calibration imputation method introduced in section 2. All the methods we consider assume that the sample outliers are representative. In addition, we assume that these sample outliers have been identified, so that the sample can be decomposed into inliers and outliers (i.e. $s = s_1 \cup s_2$ as in the previous section). We also assume non-response is ignorable given auxiliary information, and so estimation can be based on the responding sample data.

3.1 The Hidiroglou-Srinath estimator

Hidiroglou and Srinath (1981) assumed prior identification of sample outliers and introduced a class of robust estimators of a finite population total based on the idea of down-weighting these sample outliers relative to the sample inliers. For the simple random sampling situation, these authors proposed an estimator of the form

$$\hat{t}_{HS} = \lambda \sum_{i\in s_2} Y_i + q(\lambda) \sum_{i\in s_1} Y_i \tag{12}$$

where s_1 is the inlier sub-sample, of size n_1, s_2 is the outlier sub-sample, of size n_2, with $n_1 + n_2 = n$; $\lambda < N/n$ is a strictly positive down-weighting factor and $q(\lambda) = n_1^{-1}(N - n_2\lambda)$. Following the approach of Chambers (1982), we can obtain an optimal value for λ by minimising the mean squared error of (12) under the assumption that the population Y-values are randomly drawn from a mixture of inlier and outlier values. This optimal value is

$$\lambda_{opt} = \frac{(N - n_1)\sigma_1^2 + n_1\sigma_2^2 + n_1 N_2 (\mu_2 - \mu_1)^2}{n_2\sigma_1^2 + n_1\sigma_2^2 + n_1 n_2 (\mu_2 - \mu_1)^2}$$

where μ_i, σ_i^2 denote the mean and variance of the inliers (i = 1) and outliers (i = 2) in the population. These parameters can be estimated from the sample inlier/outlier values of *Y*. Assuming that the sampling fraction for outliers is the same as the overall sampling fraction f leads to the approximations

$$\lambda_{opt} \approx f^{-1}\gamma_{opt}$$

and

$$q(\lambda_{opt}) \approx f^{-1}n_1^{-1}\left(n - n_2\gamma_{opt}\right)$$

where

$$\gamma_{opt} = \frac{\left(n - n_1 f\right)\sigma_1^2 + n_1 f\sigma_2^2 + n_1 n_2\left(\mu_2 - \mu_1\right)^2}{n_2\sigma_1^2 + n_1\sigma_2^2 + n_1 n_2\left(\mu_2 - \mu_1\right)^2}. \tag{13}$$

The Hidiroglou-Srinath (HS) estimator (12) can be generalised to the case of non-uniform weights $\{d_i ; i \in s\}$ by writing it in the form

$$\hat{t}_{HS,d} = \gamma_{opt}\sum_{i\in s_2} d_i Y_i + q(\gamma_{opt})\sum_{i\in s_1} d_i Y_i . \tag{14}$$

where γ_{opt} can be calculated using (13), with all parameter estimates replaced by appropriately weighted alternatives.

Chambers (1982) observed that this estimator can be extended to incorporate auxiliary information in a straightforward way. To start, observe that (12) can equivalently be written

$$\hat{t}_{HS} = \sum_{i\in s} Y_i + (N - n)\bar{y}_1 + n_2(\lambda - 1)(\bar{y}_2 - \bar{y}_1)$$

where $\bar{y}_i$ denotes the mean of the *Y*-values in s_i, $i = 1, 2$. Assuming that the outlier and inlier sub-populations follow different regression models defined in terms of an auxiliary variable *X* and specified by

$$\begin{aligned} E_\xi(Y_i \mid i \in s_j) &= \beta_j X_i \\ V_\xi(Y_i \mid i \in s_j) &= \sigma_j^2 X_i \end{aligned} \qquad j = 1, 2,$$

one can then define a generalised HS estimator of the form

$$\hat{t}_{GHS} = \sum_{i\in s} Y_i + \left((1-\gamma)\frac{\bar{y}_1}{\bar{x}_1} + \gamma\frac{\bar{y}_2}{\bar{x}_2}\right)\sum_{j\notin s} X_j \tag{15}$$

where $\bar{x}_i$ denotes the mean of the *X*-values in s_i, $i = 1, 2$, and $\gamma \geq 0$ is a down-weighting parameter to be determined by minimising the model-based mean squared error of (15) under this model. This optimum value of γ is

$$\gamma_{opt} = \frac{n_2(\beta_2 - \beta_1)^2\bar{x}_2 / \sum_{j\notin s} X_j + \sigma_1^2 / n_1\bar{x}_1}{(\beta_2 - \beta_1)^2 + \sigma_1^2 / n_1\bar{x}_1 + \sigma_2^2 / n_2\bar{x}_2}. \tag{16}$$

Provided there are sufficient outliers in the sample data the parameters in this expression can be estimated from the s_1 and s_2 sub-samples as appropriate using standard least squares formulae.

Again, we note that the extension of (15) to the case of variable sample weights is straightforward. Basically all that is necessary is that the sample means $\bar{y}_i, \bar{x}_i$, i = 1, 2 in (15) and (16) be replaced by corresponding weighted means. Similarly, the parameter estimates in (16) are replaced by weighted least squares estimates.

3.2 Winsorized estimation

The *d*-weighted winsorized estimator of the population total of a positive-valued survey variable *Y* is

$$\hat{t}_{WR} = \sum_{i\in s} d_i Y_i^* \tag{17}$$

where

$$Y_i^* = \begin{cases} Y_i, & \text{if } Y_i \le K \\ \dfrac{Y_i + (d_i - 1)K}{d_i}, & \text{otherwise.} \end{cases}$$

Here *K* is a predefined cut-off that needs to be chosen. Kokic and Bell (1993) describe a procedure for doing this in the case of stratified random sampling that is also valid for simple random sampling.

Chambers and Kokic (1993) proposed an extension of winsorization to the linear regression context. For the situation where the regression is through the origin (i.e. a ratio model), their extension leads a ratio type predictor of the form

$$\hat{t}_{CK} = \sum_{i\in s} Y_i + \frac{\bar{y}_1}{\bar{x}_1}\sum_{j\notin s} X_j + \left(\frac{N\bar{X}}{n\bar{x}} - 1\right)\sum_{i\in s}\sqrt{X_i}\begin{cases} \dfrac{Y_i - (\bar{y}_1/\bar{x}_1)X_i}{\sqrt{X_i}} & \text{if } \dfrac{Y_i - (\bar{y}_1/\bar{x}_1)X_i}{\sqrt{X_i}} \le K(X_i) \\ K(X_i) & \text{otherwise} \end{cases}$$

where $\bar{X}$ and $\bar{x}$ are the population and sample means of *X*-values respectively, $\bar{y}_1$ and $\bar{x}_1$ are the means of *Y* and *X* values in s_1 respectively and $K(x) = c\hat{\sigma}_s\sqrt{x} + x\bar{y}_1/\bar{x}_1$. Here $\hat{\sigma}_s$ is a scale estimate based on the regression residuals in s_1 and *c* is a tuning constant. In the application described in section 4 we took *c* = 4. This estimator can be written as

$$\hat{t}_{CK} = \frac{\sum_{i\in s} Y_i^*}{\sum_{i\in s} X_i}\sum_{i\in U} X_i \tag{18}$$

where

$$Y_i^* = \begin{cases} Y_i & \text{if } Y_i \le K(X_i) \\ \alpha Y_i + (1-\alpha)K(X_i) & \text{otherwise.} \end{cases}$$

Here $\alpha = n\bar{x}/N\bar{X}$. The estimator (18) can be adapted to unequal weights, leading to

$$\hat{t}_{CK,d} = \frac{\sum_{i\in s} d_i Y_i^{**}}{\sum_{i\in s} d_i X_i}\sum_{i\in U} X_i \tag{19}$$

where

$$Y_i^{**} = \begin{cases} Y_i & \text{if } Y_i \le K(X_i) \\ \alpha_i Y_i + (1-\alpha_i)K(X_i) & \text{otherwise.} \end{cases}$$

In this case $\alpha_i = \bar{x}_d / d_i\bar{X}$, with $\bar{x}_d = \sum_{i\in s} d_i X_i / \sum_{i\in s} d_i$.

3.3 A model-based robust regression estimator

Suppose that the finite population values $\{Y_i, i \in U\}$ satisfy $Y_i = \beta X_i + \varepsilon_i$, where the ε_i are *iid* with mean zero and variance $\sigma^2 v^2(X_i)$, with $v(x) > 0$ a known function. Then the Best Linear Unbiased Predictor of the finite population total of *Y* is (Royall, 1970)

$$\hat{t}_{LS} = \sum_{i\in s} Y_i + \hat{\beta}_{LS}\sum_{j\notin s} X_j \tag{20}$$

where $\hat{\beta}_{LS}$ is the generalised least squares estimator of β

$$\hat{\beta}_{LS} = \left(\sum_{i \in s} X_i^2 / v^2(X_i) \right)^{-1} \left(\sum_{i \in s} Y_i X_i / v^2(X_i) \right).$$

It is well known that $\hat{t}_{LS}$ is sensitive to outliers. Chambers (1986) noted that this estimator can be decomposed as

$$\hat{t}_{LS} = \sum_{i \in s} Y_i + \beta \sum_{j \notin s} X_j + \sum_{i \in s} \sigma w_i \left(\frac{Y_i - \beta X_i}{\sigma v(X_i)} \right).$$

Here β is the true regression coefficient and the w_i satisfy

$$w_i = X_i v^{-1}(X_i) \left(\sum_{j \notin s} X_j \right) \left(\sum_{i \in s} X_i^2 / v^2(X_i) \right)^{-1}.$$

Based on this decomposition, Chambers (1986) proposed a class of robust alternatives to (20)

$$\hat{t}_{rob} = \sum_{i \in s} Y_i + \hat{\beta}_s \sum_{j \notin s} X_j + \sum_{i \in s} \hat{\sigma}_s w_i \psi \left(\frac{Y_i - \hat{\beta}_s X_i}{\hat{\sigma}_s v(X_i)} \right)$$

where $\hat{\beta}_s$ and $\hat{\sigma}_s$ are outlier-robust slope and scale estimators respectively and ψ is a real valued influence function that controls the contribution of the outliers to the estimate. In the case described in section 4, where an outlier identification exercise is first carried out, these estimators are based on the values in s_1. If we denote these estimators by a subscript of "1", then the analogue of the Chambers (1986) estimator for this case is

$$\hat{t}_{LSC} = \sum_{i \in s} Y_i + \hat{\beta}_1 \sum_{j \notin s} X_j + \sum_{i \in s_2} \hat{\sigma}_1 w_i \psi \left(\frac{Y_i - \hat{\beta}_1 X_i}{\hat{\sigma}_1 v(X_i)} \right). \tag{21}$$

A common situation is where the outliers are mainly big outliers, in which case the distribution of the residuals is skewed. Chambers and Kokic (1993) suggest that in this case (21) could be based on the asymmetric influence function

$$\psi(t) = \begin{cases} t & \text{if } t \le c \\ c & \text{otherwise} \end{cases}$$

where c is a tuning constant. Typically this is taken to be rather large, in order to allow the (representative) sample outliers to contribute to the total population estimate. Again, in the application reported in section 4 we used this asymmetric influence function, with $c = 4$.

3.4 An outlier resistant and approximately unbiased estimator

All the outlier resistant estimators discussed so far are biased, sometimes substantially, when the outlier distribution is skewed. This is necessary since they achieve efficiency by trading increased bias for decreased variance. However, Ren and Chambers (2002) propose an outlier resistant estimator that is less prone to bias in this situation. This works by rescaling both the population inliers and the outliers in order to reduce the variability caused by outliers, while at the same time maintaining the population total. This allows approximately unbiased estimation of this total based on these rescaled values.

Let U_1 and U_2 denote the sub-populations containing inliers and outliers, respectively. The aim is to find an optimum rescaling constant λ, $\lambda \ge 0$, such that when the Y-values are rescaled to

$$Y_i^* = \begin{cases} f(\lambda) Y_i & i \in U_1 \\ \lambda Y_i & i \in U_2 \end{cases}$$

the population total remains unchanged:

$$t_y = \sum_{i\in U} Y_i = \sum_{i\in U} Y_i^* = f(\lambda)\sum_{i\in U_1} Y_i + \lambda \sum_{i\in U_2} Y_i\,. \tag{22}$$

The optimum value of λ is the value that minimises the rescaled population variance

$$S_y^{*2} = \frac{1}{N-1}\sum_{i\in U}\left(Y_i^* - \bar{Y}\right)^2 = \frac{1}{N-1}\left\{\sum_{i\in U_1}\left(f(\lambda)Y_i - \bar{Y}\right)^2 + \sum_{i\in U_2}\left(\lambda Y_i - \bar{Y}\right)^2\right\} \tag{23}$$

subject to (22). It is easy to see that (22) holds provided

$$f(\lambda) = 1 + \delta(1-\lambda) \tag{24}$$

where $\delta = \sum_{i\in U_2} Y_i / \sum_{i\in U_1} Y_i$. Minimisation of (23) subject to (24) leads to

$$\lambda_{opt} = \frac{\delta(\delta+1)\sum_{i\in U_1} Y_i^2}{\delta^2 \sum_{i\in U_1} Y_i^2 + \sum_{i\in U_2} Y_i^2}\,. \tag{25}$$

Estimation of the population total of the rescaled values (and therefore, by definition, an estimator of the original population total) then proceeds in the usual way, using the sample *d*-weights provided.

In practice, both δ and λ_{opt} are unknown, so sample estimates must be substituted. Let $\hat{\delta}$ and $\hat{\lambda}_{opt}$ be these estimators. The resulting outlier resistant estimator is

$$\hat{t}_{Rob} = f(\hat{\lambda}_{opt})\sum_{i\in s_1} d_i Y_i + \hat{\lambda}_{opt}\sum_{i\in s_2} d_i Y_i\,. \tag{26}$$

The main problem in (26) is estimation of δ. If we denote this estimate by $\hat{\delta}$, then λ_{opt} can be estimated by the simple sample-weighted expression

$$\hat{\lambda}_{opt} = \frac{\hat{\delta}\left(\hat{\delta}+1\right)\sum_{i\in s_1} d_i Y_i^2}{\hat{\delta}^2 \sum_{i\in s_1} d_i Y_i^2 + \sum_{i\in s_2} d_i Y_i^2}\,.$$

Ren and Chambers (2002) suggest that in the absence of any prior or external information about the proportion of outliers in the population, δ be estimated by

$$\hat{\delta} = \left(\frac{\hat{M}_2 \sum_{i\in s_2} d_i \sum_{i\in s_2} d_i Y_i}{\hat{M}_1 \sum_{i\in s_1} d_i \sum_{i\in s_1} d_i Y_i}\right)^{1/2}$$

where $\hat{M}_1$ and $\hat{M}_2$ are sample medians of the *Y*-values in s_1 and s_2, respectively.

4. Numerical evaluation of robust estimation and imputation

We evaluate the reverse calibration imputation method using the 1997 sector one *ABI* data, as prepared for the *Euredit* (2000) project. In particular we focus on one auxiliary (*X*) variable *turnreg* corresponding to the register value of estimated turnover for a business and four analysis variables (*Y*). These are total turnover (*turnover*), total tax paid (*taxtot*), total purchases (*purtot*) and total employment costs (*emptotc*). Since *turnreg* is a register variable we know its overall total as well as its stratum totals. The strata themselves correspond to size strata defined in terms of the register measure of the number of employees of a business and the *turnreg* value for the business. Sample weights (*d*-weights) are also available.

The *ABI* dataset has 6099 cases and comes in two versions. The first, which we call the *true data*, has no errors and no missing data, but still has many representative outliers. The second, which we refer to as the *perturbed data* contains a mix of representative (i.e. true) outliers, introduced errors (many leading to non-representative outliers) and introduced missing values. Since we have access to the *true data* we can construct two approaches to imputation based on these data. The first, which leads to the *unverified data*, treats all detected outliers in the *perturbed data* as representative, estimates the underlying population total on this basis and then imputes values for all detected outliers and missing values. The second approach, which leads to the *verified data*, is more realistic, in the sense that all identified outliers in the *perturbed data* are first checked to see whether they are errors or not. Any detected errors are then set to their true values and estimation/imputation proceeds as before.

Table 1 gives the *d*-weighted estimates of the population total based on the *true data* and the *perturbed data*, respectively. Note the huge impact that untreated errors in the data have on these estimates.

Table 1. Weighted estimates of population totals

	turnover	*taxtot*	*purtot*	*emptotc*
true data	269088777	4631853	189689033	29419325
perturbed data	24116695453	436375032	20739928268	2357859187

4.1 Outlier detection for the *ABI* data

For estimators using auxiliary information, outliers were detected using an across-stratum forward search procedure (Chambers, Hentges and Zhao, 2002) based on a linear model in the log scale of the data. For estimators that do not use auxiliary information, outliers were detected within strata using the *MAD* (*Median Absolute Deviation*) procedure. This declares a sample value in a stratum to be an outlier when it lies outside the interval [*Med* – 4*MAD*, *Med* + 4*MAD*] where *Med* is the median of the stratum sample data and *MAD* is the median of the absolute deviations from this median. In what follows we refer to true outliers as outliers detected in the *true data*, and perturbed outliers as those detected in the *perturbed data*. The latter can be split further, into detected outliers and detected errors. Tables 2 and 3 give the total number of missing values, errors, true outliers (detected in the true data), detected outliers and detected errors, and therefore the number of undetected outliers and undetected errors in the *verified data*. Finally, we note that since the data do not follow a linear model for *taxtot* and *emptotc* in the log scale, the forward search procedure failed to detect most of the errors for these variables. Consequently, we carried out outlier detection for these variables by combining the soft edit rules used in the *ABI* data with the forward search method.

Table 2. Outlier and error detection performance: forward search method

Variable	Missing values	Actual errors	True outliers	Detected outliers	Detected errors	Undetected outliers	Undetected errors
turnover	42	241	106	71	224	35	17
taxtot	45	482	23	23	247	0	235
purtot	28	629	111	64	275	47	354
emptotc	41	332	39	26	237	13	95

Table 3. Outlier and error detection performance: *MAD* method

Variable	Missing values	Actual errors	True outliers	Detected outliers	Detected errors	Undetected outliers	Undetected errors
turnover	42	241	165	136	205	29	36
taxtot	45	482	258	145	339	113	143
purtot	28	629	265	183	257	82	372
emptotc	41	332	194	128	252	66	80

4.2 Outlier resistant estimation of population totals

We consider four estimators that assume an across-stratum linear relationship between the survey variable and the auxiliary variable *turnreg* (and hence ignore the size stratification), and four estimators that do not. The estimators in the former class are the generalised HS estimator (15), coded as *Estghsrc*; the generalised Kokic and Bell estimator (19), coded as *Estgkbrc*; the robust regression estimator (21), coded as *Estrerc*; and the classical model-based predictor $\hat{t}_{LS}$, see (20), coded as *Clprdreg*, which serves as the reference estimator for this class. The estimators in the latter class are all stratified by size. They are the weighted HS estimator (14), coded as *EstimHS*; the weighted scale transform estimator (26), coded as *EstimRob*; the weighted winsorized estimator (17), coded as *EstWins*; and the Horvitz-Thompson estimator (in this case the stratified expansion estimator), coded as *EstimHT*, which serves as the reference estimator for this class.

For each data configuration and each study variable, estimated population totals and their estimated coefficients of variation (based on Jackknife estimates of standard errors) were calculated for all the estimators listed above. These are shown in Tables 4 and 5 for the estimators that use auxiliary information, and in Tables 6 and 7 for the estimators that do not. The coefficient of variation is the ratio of the estimated standard error to the corresponding estimate of total, and is only useful for comparing estimators that are unbiased. For example, a positively biased estimator with the same standard error as an unbiased estimator will have a lower coefficient of variation. Consequently, in Tables 5 and 7 we also show the relative difference between an estimator and its reference estimator. Note that outliers tend to right skew most economic populations. Consequently we expect negative relative changes for the resistant estimators.

Inspection of the results for the linear model-based estimators set out in Tables 4 and 5 shows that the three resistant estimators are superior to the non-resistant reference estimator for the *true data*, with smaller coefficients of variation and moderate negative relative differences between these estimators and the reference estimator. For the *verified data*, all four estimators overestimate by a small amount. This is probably due to the presence of undetected errors in the sample data. In contrast, the results for the *unverified data*, where the detected errors are treated as representative outliers, are clearly unacceptable. In this case all four methods of estimation severely overestimate the population totals of interest. Focussing on the *true data* and the *verified data*, we see that *Estgkbrc* and *Estregrc* are slightly superior to *Estghsrc*.

Similar results can be seen in Tables 6 and 7 for the stratified estimators. Here again the three resistant estimators are generally superior to the non-resistant reference Horvitz-Thompson estimator. The weighted winsorized estimator *EstWins* appears to be biased lower than the weighted HS estimator *EstimHS* and the weighted scale transform estimator *EstimRob*. Other numerical results not reported here show that *EstimHS* generally performs well when there are few outliers in the sample data, while *EstimRob* performs well when there are many outliers in the sample data.

Comparing the results for the linear model-based estimators in Tables 4 and 5 with those for the stratified estimators in Tables 6 and 7 we see that the former group of estimators generally have smaller coefficients of variation, but slightly larger negative relative changes. Overall, it is clear that using the auxiliary information in *turnreg* in estimation is beneficial for the *ABI* data. It is also very clear that, irrespective of the particular resistant method of estimation used, verification of identified outliers before estimation is crucial. Allowing errors in the survey data to be treated as representative outliers in estimation is a recipe for disaster.

Table 4. Estimation of totals: linear model-based estimators

Variable	Estimator			
	Clprdreg	*Estghsrc*	*Estgkbrc*	*Estregrc*
	true data			
turnover	269545407	253270677	247464410	249342600
taxtot	4655782	4258105	4091774	4238522
purtot	192575028	180852297	174049744	177459170
emptotc	27526483	26895756	26398197	26724563
	verified data			
turnover	280461470	269675223	260157507	264729267
taxtot	5211889	4911781	4621905	4782497
purtot	199913301	188594856	181051536	184992156
emptotc	27636629	27105079	26606150	26864103
	unverified data			
turnover	27140509644	29581299196	11820399165	14924616179
taxtot	459377641	308095114	216914827	163573025
purtot	22686419578	25637777042	10167753954	12910204562
emptotc	2517666251	2764776476	1174848045	1397232943

Table 5. Coefficients of variation and relative changes in total estimation: linear model-based estimators

Variable	Estimator						
	Clprdreg	*Estghsrc*	*Estgkbrc*	*Estregrc*	*Estghsrc*	*Estgkbrc*	*Estregrc*
	Coefficient of variation (%)				Relative change (%)		
	true data						
turnover	4.5	3.8	2.8	2.9	-6.0	-8.2	-7.5
taxtot	10.4	8.1	7.6	7.9	-8.5	-12.1	-8.9
purtot	4.9	4.4	3.2	3.3	-6.1	-9.6	-7.8
emptotc	5.3	4.7	6.0	4.6	-2.3	-4.1	-2.9
	verified data						
turnover	5.3	5.0	4.3	4.4	-3.8	-7.2	-5.6
taxtot	11.6	9.6	8.1	8.6	-5.8	-11.3	-8.2
purtot	5.7	5.3	4.7	4.5	-5.7	-9.4	-7.5
emptotc	5.3	4.7	6.0	4.6	-1.9	-3.7	-2.8
	unverified data						
turnover	81	87	80	85	9	-56	-45
taxtot	44	61	40	58	-34	-52	-64
purtot	67	71	68	71	13	-55	-43
emptotc	74	79	70	78	10	-53	-44

Table 6. Estimation of totals: stratified estimators

Variable	Estimator			
	EstimHT	*EstimHS*	*EstimRob*	*EstWins*
	true data			
turnover	269088777	262919117	266872511	253448998
taxtot	4631853	4462770	4478555	4279186
purtot	189689033	186878783	183314050	181321657
emptotc	29419325	27961081	28147126	26417602
	verified data			
turnover	268880011	263806971	264411474	253222992
taxtot	4857658	4693038	5144381	4506820
purtot	188855602	186090830	189547911	180483346
emptotc	29431502	27984426	28200362	26424474
	unverified data			
turnover	23954027905	20373805300	9536712579	19464659853
taxtot	434439983	336263777	193135103	274003138
purtot	20628515242	18297388571	8267416072	17618818717
emptotc	2342642556	2012001740	1013491684	1891822323

Table 7. Coefficients of variation and relative changes in total estimation: stratified estimators

Variable	Estimator						
	EstimHT	*EstimHS*	*EstimRob*	*EstWins*	*EstimHS*	*EstimRob*	*EstWins*
	Coefficient of variation (%)				Relative change (%)		
	true data						
turnover	13.8	14.0	13.6	13.8	-2.3	-0.8	-5.8
taxtot	12.0	11.3	9.6	10.5	-3.7	-3.3	-7.6
purtot	13.9	14.0	13.3	13.7	-1.5	-3.4	-4.4
emptotc	15.8	14.4	11.5	12.9	-5.0	-4.3	-10.2
	verified data						
turnover	14.8	14.8	13.7	14.2	-1.9	-1.7	-5.8
taxtot	10.8	10.4	9.1	9.9	-3.4	5.9	-7.2
purtot	13.6	13.7	12.7	13.5	-1.5	0.4	-4.4
emptotc	14.3	13.1	10.8	12.2	-4.9	-4.2	-10.2
	unverified data						
turnover	80.6	78.9	18.3	80.0	-14.9	-60.2	-18.7
taxtot	37.7	37.9	11.0	42.7	-22.6	-55.5	-36.9
purtot	65.6	65.0	21.0	63.6	-11.3	-59.9	-14.6
emptotc	70.3	68.3	17.9	70.4	-14.1	-56.7	-19.2

4.3 Outlier imputation

To illustrate the performance of outlier imputation based on the reverse calibration method we calibrated to the robust population total estimate *Estrerc* calculated using (21). The reverse calibration imputations themselves were computed using (11). They were then compared with values obtained using standard imputation methods for these type of data: regression imputation under a linear model in *turnreg* (in both the raw and log scale of the data) and nearest neighbour imputation, based on distances between sample values of *turnre*g. Since the observed differences in imputations between these standard three methods were small, only the results of regression imputation under a linear model in *turnreg* are reported below.

Imputations were carried out using three different types of data. To start, all detected outliers were imputed in the *true data*. Since it is clear that calibrating to estimates based on the *unverified data* is a waste of time, we explored calibration to two versions of the *verified data*. The first, which we call the *100% verified data* does not depend on an outlier identification process for error detection. All sample records in the *perturbed data* are verified and all errors corrected by replacement of their true values. Thus the *100% verified data* contains only true outliers and missing values, with some of the true outliers detected via the outlier detection process. The second version of the *verified data* we refer to as the *outlier verified data*. This data set is defined by only verifying the error status of detected outliers in the *perturbed data*, with all errors within this group then being set to missing.

Note that zeros in the survey data (either *Y* or *X*) define another type of outlier. In this paper, however, we do not consider this special kind of outlier, assuming that cases with zero values are error-free. This leads to outlying observations parallel to the *x*-axis and *y*-axis in the scatter plots presented in Figure 2.

In order to assess the quality of the imputations, we present results for four evaluation criteria in Table 8. The first three are evaluation criteria recommended by the *Euredit Project* (Chambers, 2001), while the fourth is a measure of the proportion of imputed values in the population that pass the *ABI* soft editing rules. The evaluation criteria are:

i. The weighted mean absolute difference between the true values Y_i and the imputed values Y_i^*:

$$MADI = \sum\nolimits_{i \in imp} d_i \left| Y_i - Y_i^* \right| / \sum\nolimits_{i \in imp} d_i$$

where imp is the imputed sub-sample.

ii. The weighted mean absolute relative difference between the true values Y_i and the imputed values Y_i^*:

$$MARD = \sum\nolimits_{i \in imp^+} d_i \left| \frac{Y_i - Y_i^*}{Y_i} \right| / \sum\nolimits_{i \in imp^+} d_i$$

where imp^+ is the same as in **i**, but restricted to $Y_i > 0, i \in imp$.

iii. The weighted Pearson moment correlation coefficient between the true values Y_i and the imputed values Y_i^*:

$$PECC = \sum\nolimits_{i \in imp} d_i \left(Y_i - \overline{Y}_{s_2} \right)\left(Y_i^* - \overline{Y}_{imp}^* \right) / \sqrt{\sum\nolimits_{i \in imp} d_i \left(Y_i - \overline{Y}_{imp} \right)^2 \sum\nolimits_{i \in imp} d_i \left(Y_i^* - \overline{Y}_{imp}^* \right)^2}$$

where $\overline{Y}_{imp} = \sum\nolimits_{i \in imp} d_i Y_i / \sum\nolimits_{i \in imp} d_i$ and $\overline{Y}_{imp}^* = \sum\nolimits_{i \in imp} d_i Y_i^* / \sum\nolimits_{i \in imp} d_i$.

iv. The weighted proportion of valid imputations:

$$MSVI = \sum\nolimits_{i \in imp} d_i \delta_i / \sum\nolimits_{i \in imp} d_i$$

where $\delta_i = 1$ if Y_i^* is a valid value (i.e. Y_i^* passes the soft edit rules of the *ABI*); $\delta_i = 0$ otherwise.

Table 8. Evaluation statistics for imputed values

	Weighted mean of imputed values			*MADI*		*MARD*		*PECC*		*MSVI*	
	(a)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)
	true data										
turnover	1456	405	214	1681	1544	8.8	7.8	0.42	0.07	0.57	1.00
taxtot	910	382	16	528	922	0.6	14.9	1.00	-0.15	0.65	1.00
purtot	763	245	172	938	877	10.3	13.7	0.08	0.04	0.96	1.00
emptotc	135	494	28	590	158	268	16	0.16	0.09	0.62	1.00
	100% verified data										
turnover	1166	331	273	1182	1144	4.6	5.7	0.30	0.15	0.77	1.00
taxtot	159	17	7	168	160	13.4	3.5	0.05	0.01	1.00	1.00
purtot	659	189	183	723	717	6.0	10.9	0.08	0.04	0.97	1.00
emptotc	75	120	31	162	79	63	8.6	0.21	0.16	1.00	1.00
	outlier verified data										
turnover	2493	2201	2153	593	538	2.4	2.5	0.999	0.999	0.92	1.00
taxtot	56	46	45	55	56	1.2	1.2	0.940	0.940	0.96	0.96
purtot	1833	1714	1654	415	374	5.9	4.3	0.999	0.999	0.96	1.00
emptotc	238	282	276	90	98	1.7	1.8	0.999	0.999	0.96	1.00

Note: (a) true value, (b) reverse calibration, (c) regression imputation.

When considering the results set out in Table 8, it is important to remember that there are relatively few imputations in the *true data* and the *100% verified data* (essentially only missing values and true outliers) while there are relatively many imputations in the *outlier verified data* (since these include all the detected errors, which are treated as missing). Consequently, comparisons between these different data sets should be avoided. This is particularly true for *PECC* where we see low values for the *true data* and the *100% verified data* (due to the fact that the imputations for the true outliers are not designed to recover their values, and these dominate the measure) while in the case of the *outlier verified data* the robust regression model underpinning the imputations for the detected errors actually fits their true values rather well, and these "swamp" the imputations for the true outliers.

The results in Table 8 show that reverse calibration imputation and standard regression imputation seem to be not too different in terms of pure imputation performance, with perhaps the reverse calibration approach scoring better in terms of the correlation between true and imputed values. To an extent this similarity is driven by the fact that the true outliers in the ABI data tend to have small weights and hence are discounted by the above criteria.

The performance of reverse calibration imputation can be better appreciated from an inspection of Figures 1 and 2. In Figure 1 we plot the imputed values of *turnover* against the true values of this variable, while in Figure 2 we show the values of *turnover* plotted against those of the covariate *turnreg* both before and after imputation. All plots are on the log scale of the data. Here it can clearly be seen that there is a strong linear relationship between the imputed values and the true values for true outliers. There is also no significant difference between the imputed values for missing data and perturbed errors generated by the two methods.

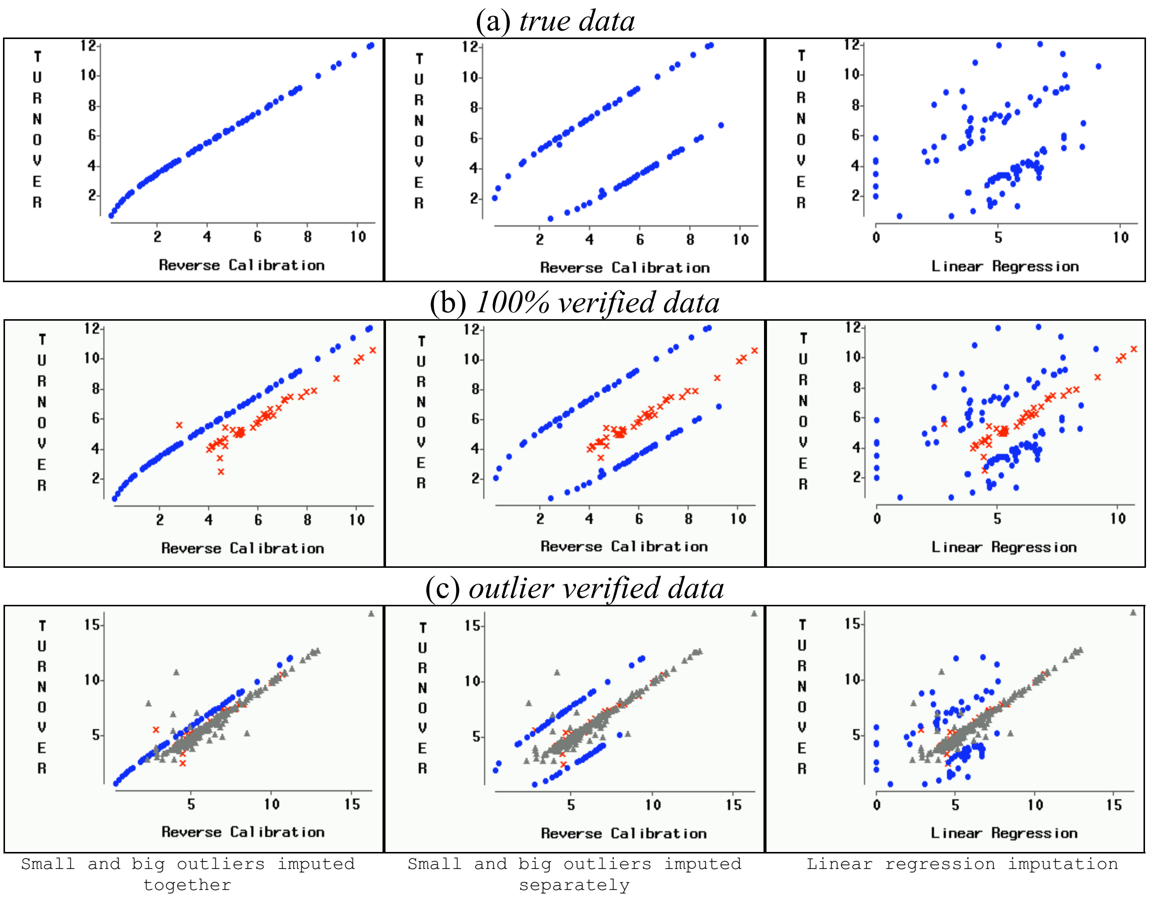
Finally, in Table 9 we show the stratified expansion estimates and the regression estimates of the population totals before imputation (in brackets) and after imputation (first value is based on reverse calibration imputation, second value is based on robust regression imputation). For the regression estimation results, pre-imputation estimation is robust regression estimation, while after imputation it is simple regression estimation. For the expansion estimation results, both of pre-imputation estimation and post-imputation estimation are stratified expansion estimation. Since there are actually relatively few true outliers in the ABI data (compared to the number of errors), there is little difference between the overall estimates based on reverse calibration imputation and those based on regression imputation. However, it is also clear that the estimates based on the reverse calibration imputations are systematically slightly larger than those based on regression imputation, indicating that the former imputation method tends to allow outliers to have more "say" in estimation.

Table 9. Estimates of population total before and after imputation

	Variable			
	turnover	*taxtot*	*purtot*	*emptotc*
	Stratified expansion estimation			
true data	(269088777) 265054049 264320456	(4631853) 4455522 4333276	(189689083) 187334925 187003002	(29419325) 29824257 29298019
100% verified data	(270468949) 264665106 264361038	(4654708) 4349873 4329947	(190425133) 187007851 186970993	(29576816) 29536876 29308487
outlier verified data	(264654385) 277552404 276756021	(5219167) 5107820 5168510	(184932869) 195346631 194553015	(28650608) 30146782 30070948
	Regression estimation			
true data	(249342600) 253085063 252225059	(4238522) 4445173 4204309	(177459170) 180939121 180483772	(26724563) 27146206 26745912
100% verified data	(249011055) 249402861 249099514	(4226175) 4133132 4121348	(177325785) 176197519 176158917	(26689600) 26981405 26841312
outlier verified data	(243517947) 251443928 250679084	(4431328) 4906132 4898020	(170611125) 176725974 175963690	(25222498) 27671226 27570233

As conclusion, the reverse calibration imputation can be a competitive alternative to the conventional imputation methods, especially for imputation of outlier values.

Figure 1. Plots of true values (*y*) against imputed values (*x*) for *turnover*, log scale. Blue indicates true outliers, red indicates missing and grey indicates errors.



(a) *true data*

(b) *100% verified data*

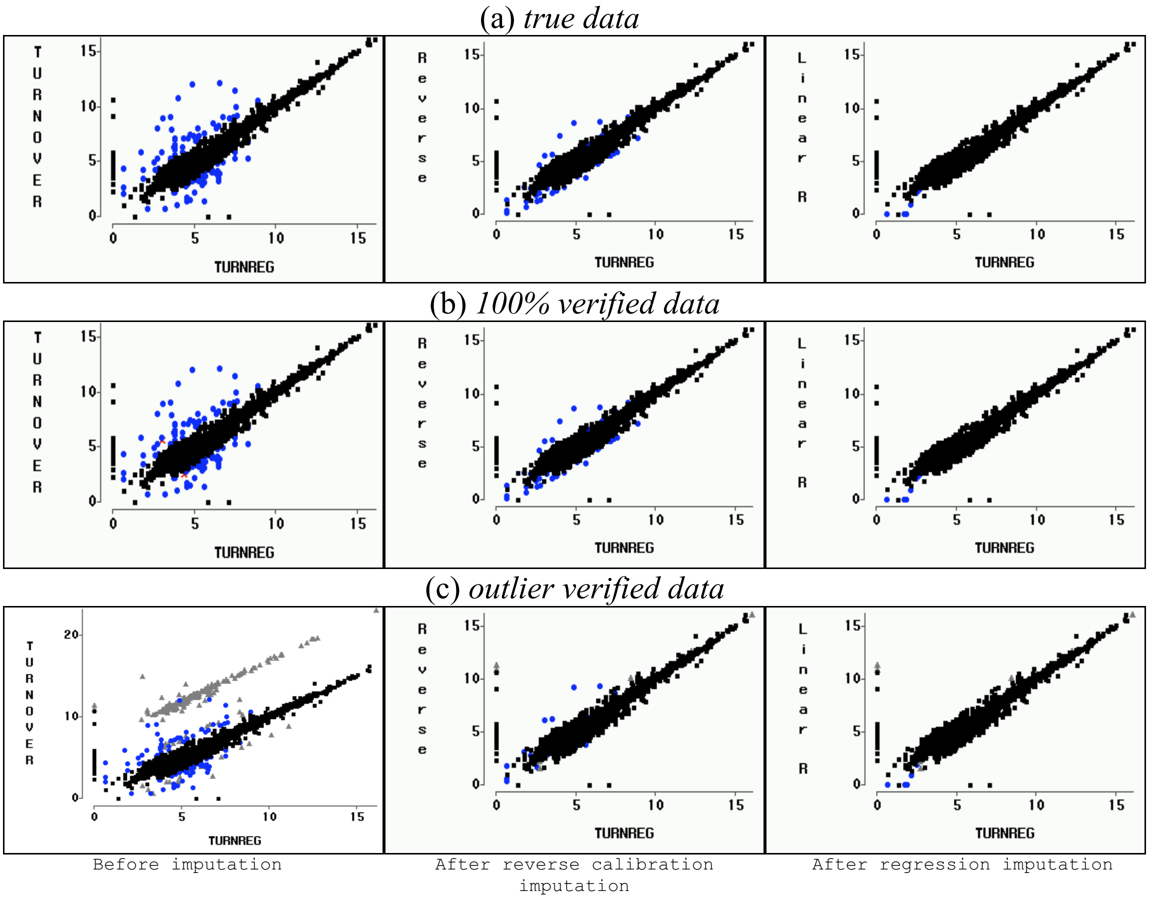
(c) *outlier verified data*

Small and big outliers imputed together

Small and big outliers imputed separately

Linear regression imputation

Figure 2. Plots of *turnover* (*y*) against *turnreg* (*x*) for *turnover*, log scale. Blue indicates true outliers, red indicates missing and grey indicates errors.



(a) *true data*

(b) *100% verified data*

(c) *outlier verified data*

Before imputation

After reverse calibration imputation

After regression imputation

References

Chambers, R. L. (1982). *Robust Finite Population Estimation*. PhD. Thesis. *The Johns Hopkins University*, Baltimore.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association,* 81, 1063-1069.

Chambers, R. L. and Kokic, P. N. (1993). Outlier robust sample survey inference. Invited Paper, *Proceedings of the 49th Session of the International Statistical Institute*, Firenze.

Chambers, R. L. (2001). Evaluation criteria for statistical editing and imputation. *Euredit Project* Report.

Chambers, R., Henteges, A. and Zhao, X. (2002). Using robust tree-based methods for outlier and error detection. *Manuscript submitted for publication*.

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

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology,* 25, 43-56.

Euredit Project (2000). *Euredit Project* document. *ONS*.

Hidiroglou, M. H. and Srinath, K. P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association,* 76, 690-695.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimation. *Survey Methodology*, 21, 79-87.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.

Lee , H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the 1991 Annual Research Conference.* U.S. Bureau of the Census.

Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods*, (Eds. B.G. Box, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), John Wiley & Sons, New York.

Ren, R. and Chambers, R. L. (2002). Unbiased outlier resistant estimation for finite populations. *Manuscript submitted for publication*.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika* 57, 377-387.

Sautory, O. (1993). La macro CALMAR: Redressement d'un échantillon par calage sur marges. *Technical Report* F9310: INSEE.

Searl, D. T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.

Welsh, A. H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, B, 60, 413-428.