



SCHOOL OF INFORMATION SCIENCES AND
TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**Detection and Treatment of Outliers in Survey
Data**

by

Eleni Th. Mavropoulou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
October 2015





Acknowledgements

Ioannis Doukas





ΠΕΡΙΛΗΨΗ

Σε αυτή τη διπλωματική εργασία γίνεται μία ανασκόπηση των πιο γνωστών μεθόδων που χρησιμοποιούνται για την ανίχνευση και αντιμετώπιση των ακραίων τιμών σε δεδομένα που προέρχονται από δειγματοληπτικές έρευνες. Ο όρος «ακραία τιμή», στις δειγματοληπτικές έρευνες, αναφέρεται είτε σε παρατηρήσεις που έχουν μεγάλη τιμή, είτε σε παρατηρήσεις που έχουν μεγάλο δειγματικό βάρος, με αποτέλεσμα να έχουν μεγάλη επίδραση στις εκτιμήσεις. Υπάρχει πληθώρα μεθοδολογιών που μπορούν να εφαρμοστούν σε δεδομένα δειγματοληπτικών ερευνών, οι οποίες παρουσιάζονται στα κεφάλαια που ακολουθούν. Ωστόσο, λίγες από αυτές τις μεθόδους είναι κατάλληλες για την ανίχνευση και αντιμετώπιση των ακραίων τιμών όταν ο σχεδιασμός της δειγματοληψίας είναι περίπλοκος. Σε αυτή την εργασία εφαρμόστηκαν κάποιες από αυτές τις μεθόδους σε μια εμπειρική μελέτη στα δεδομένα της Έρευνας για το Εισόδημα και τις Συνθήκες Διαβίωσης της Ελλάδας, που διενεργείται από την ΕΛΣΤΑΤ, για το έτος 2013.





Abstract

In this thesis, we review the most common methods that are used for the detection and treatment of outliers in survey data. The term outliers, in survey sampling, refers to sample units that have either a high value or are associated with a large sampling weight, thus having a high impact on the estimates. There is a variety of methodologies that can be applied to survey data and which are presented in the following chapters of this thesis. However, only few of those methods are suitable for detecting and treating outliers when the survey design is complex. In this thesis, some of the methods are applied in an empirical study of outliers in the data of the Survey on Income and Living Conditions of Greece, for the year 2013.





Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Outliers in Survey Data | 3 |
| 3 | Detection of Outliers | 7 |
| 3.1 | The relative distance | 8 |
| 3.2 | The MAD | 9 |
| 3.3 | The quartile method | 9 |
| 3.4 | The Hidioglou-Berthelot method | 10 |
| 3.5 | Outlier detection in the HCSO | 11 |
| 3.6 | Cook's D | 13 |
| 3.7 | Mahalanobis distance | 14 |
| 4 | Treatment of Outliers | 17 |
| 4.1 | Winsorized estimators | 17 |
| 4.1.1 | Searls' Winsorized mean | 18 |
| 4.1.2 | The Once-Winsorized mean | 19 |
| 4.1.3 | Fuller's mean | 20 |
| 4.2 | Re-weighted estimators | 21 |



| | | |
|----------|---|-----------|
| 4.2.1 | The post-stratified estimator | 21 |
| 4.3 | The variance-inflation model | 22 |
| 4.4 | Dalén and Tambay’s method | 23 |
| 4.5 | ψ -type M-estimators | 24 |
| 4.5.1 | The Huber M-Estimator | 25 |
| 4.5.2 | Bruce’s period-to-period change estimator | 25 |
| 4.5.3 | Regression Estimators | 26 |
| 4.5.4 | The Conditional Bias | 32 |
| 5 | Empirical Study | 37 |
| 5.1 | Detection | 42 |
| 5.2 | Treatment | 46 |
| A | Detection | 57 |
| B | Treatment | 59 |



List of Tables

| | | |
|-----|--|----|
| 5.1 | Descriptive Statistics | 40 |
| 5.2 | Outlier Detection for HY010 | 43 |
| 5.3 | Robust H-T and Huber M-estimator for HY010 | 46 |
| 5.4 | D-T mean of HY010 | 49 |
| 5.5 | RV for HY010 | 52 |
| | | |
| A.1 | Outlier Detection for HH070 | 57 |
| A.2 | Outlier Detection for HY020 | 58 |
| A.3 | Outlier Detection for HY022 | 58 |
| | | |
| B.1 | Robust H-T and Huber M-estimator for HH070 | 59 |
| B.2 | Robust H-T and Huber M-estimator for HY020 | 60 |
| B.3 | Robust H-T and Huber M-estimator for HY022 | 60 |
| B.4 | D-T mean of HH070 | 61 |
| B.5 | D-T mean of HY020 | 61 |
| B.6 | D-T mean of HY022 | 62 |
| B.7 | RV for HH070 | 62 |
| B.8 | RV for HY020 | 63 |
| B.9 | RV for HY022 | 63 |





List of Figures

| | | |
|-----|--|----|
| 5.1 | Matrix Plot | 41 |
| 5.2 | Histogram of HY010 | 42 |
| 5.3 | RHT and Huber's M-estimator of HY010 | 48 |



Chapter 1

Introduction

In the first chapter of this thesis, we describe what outliers are in classical statistics and furthermore, why they play such an important role in the estimation of means or totals in data obtained from a design-based survey.

In surveys, the data are associated with design weights, which indicate the number of similar units in the target population. Let's say we are interested in measuring variable Y and we obtain a sample unit i , that has a value $y_i = a$ and a design weight $w_i = b$. This would mean that there are b units in the population that have the value a for this variable. So, because of the fact that the estimators of means and totals in surveys are weighted, one could easily imagine what would happen to our estimator when either the value of a or b increases drastically.

The second chapter of this paper is about the detection of outliers in survey data. There are many methods that can be used to identify if a particular observation is an outlier. Most of these methods, e.g. the relative distance, identify as outliers



the units that have a y -value located far away from the center of the data, without however taking into consideration large design weights, if any. Also, many of the proposed methods include a certain threshold, above which a unit is detected to be an outlier, and which is not unambiguously determined.

Moving on to the third chapter, we discuss methods that treat outliers in survey data. There is a significant amount of literature about estimators that are insensitive to such observations, which can be roughly classified into three main categories: winsorized estimators, re-weighted estimators and M-estimators. Unfortunately, the majority of those methods require either the assumption of simple random sampling (srs) or a known empirical distribution, which makes them unsuitable for complex samples, where the distribution of the data is usually unknown. There are methods, such as the estimator proposed by Dalén and Tambay (1988), that perform detection and treatment of outliers simultaneously.

The last chapter of this thesis involves an empirical study using the dataset of the Greek SILC for 2013. The data were obtained by the Greek Statistical Authority, under a two-stage stratified cluster sampling design. We chose four variables of interest, where we applied the methodologies we thought were most suitable to be applied on this type of data, regarding both the detection and treatment of outliers. The treated estimators are evaluated by their relative variance compared to that of the Horvitz-Thompson estimator.



Chapter 2

Outliers in Survey Data

Chambers (1986) described outliers as a “perennial problem for applied survey statisticians”. In classical statistics, a sample is usually assumed to have been generated from a population or a model with a certain parametric distribution. Outliers are then defined as the observations that have extremely high or low values, with respect to that distribution.

However, in the framework of design-based surveys, the concept is quite different. Each unit of the sample is assigned a sampling weight, which indicates how many units of the finite population this unit represents. So, an outlier in a design-based survey may be an observation that is not extreme-valued, but can greatly influence the estimates because of the large weight that has been assigned to it. Some authors make a distinction between representative and non-representative outliers: non-representative are the outliers that are unique in the population, or values that have been incorrectly recorded, whereas representative outliers are the ones that represent other large values of the population.



In business or economic surveys, the populations from which we sample are usually skewed. Therefore it is very common that the sample we obtain contains outliers. Lee (1995), in order to give an indication of how skewed those populations can be, wrote that “20 percent of units in the population account 80 percent of the population value”. Inclusion or exclusion of these values in the estimation would affect the Horvitz-Thompson estimator so drastically, that it would become unreliable. This is the reason why special attention should be paid to those extreme population values while designing the survey.

In business surveys, where the target population is usually stratified by enterprise size, it is common that enterprises which by the time of stratification were allocated to a stratum that contains small-sized businesses, are later found to belong to another stratum. Those sample units are called stratum jumpers. It may also be the case that the time lag between the survey design and the data collection causes a change in the size of some units. The stratum jumpers carry their original sampling weights. Therefore, stratum jumpers with large weights are likely to be identified as outliers at the editing stage.

Sampling with probability proportional to size (pps) or stratification by size are two of the most common methods used when we sample from a skewed population. By employing these methods, large values are selected almost certainly, so that very small sampling weights are assigned to them. However, this means that outliers have to be identified at the design stage and without error, which is barely the case, because of the volatility in economic surveys. Inevitably, outliers will



occur, no matter how thoroughly the sample is designed.

All the above leads us to the subject of this thesis: how we deal with outliers in survey data. There are two stages, namely, the identification or detection and the treatment. The detection is done during the editing phase, where one has to decide whether a certain value is real, or an error. Errors have to be corrected or imputed; on the other hand, real extreme values should be the object of treatment.





Chapter 3

Detection of Outliers

Many methods have been proposed for the detection of outliers. In order to choose among them, one should first consider the following factors:

1. In most cases, we cannot make a realistic assumption about the underlying model or distribution for the sampled data.
2. Often, the population units are included in the sample with unequal probabilities and thus they have unequal sampling weights.
3. When conducting either business or household surveys, it is very common that the population from which the sample is drawn is skewed.

Below we present some of the most common methods that can be used to detect outliers in survey data.



3.1 The relative distance

The relative distance is the most basic method for identifying extreme observations. For each sample unit, we calculate its relative distance from the centre of the data. Let $y_{(1)}, \dots, y_{(n)}$ be the ordered data obtained from a sample of size n , m the estimate of location, and s the estimate of scale. The relative distance of y_i is given by:

$$d_i = \frac{|y_i - m|}{s} \quad (3.1)$$

Unit i is detected as an outlier if it lies outside the tolerance interval $(m - C_L s, m + C_U s)$, where C_L, C_U are predetermined constants.

The relative distance can be efficient in detecting a single outlier, it is however ineffective when trying to detect multiple outliers. If the sample mean and standard deviation are used as location and scale measures in computing d_i , it can be affected by outliers, because both m and s are sensitive to outliers. Especially when we are sampling from a skewed population, the relative distances of the outliers may appear smaller, and therefore may not be detected by (3.1). This is called the “masking effect”.

In order to avoid the “masking effect”, one should consider using a robust alternative to location and scale estimates when computing d_i .



3.2 The MAD

The Median Absolute Deviation (MAD), often used as a robust scale estimate, see Andrews et al. (1972), is given by

$$MAD = \text{median}_i[|y_i - \text{median}_j(y_j)|] \quad (3.2)$$

The MAD is more resistant to outliers than the sample standard deviation, but was not widely used for survey sampling until mid 90's. Lower and upper interquartile ranges presented below were used instead.

When detecting outliers, we first need to define the distance $\text{median} + kMAD$, see Rousseeuw et al (1993). The term $kMAD$ is a consistent estimator of the population standard deviation σ , and k is a constant scale factor, for which Iglewicz and Hoaglin (1993) proposed a value of 3 or 3.5. For highly skewed samples, one may need to increase the value of k , otherwise almost 1/5 of the sample may be detected as outliers.

3.3 The quartile method

Let q_1, q_2, q_3 denote the quartiles of the sample, with $q_2 = m$ the sample median. Also, let $d_L = q_2 - q_1$ and $d_U = q_3 - q_2$, i.e., the lower and upper interquartile ranges, be the scale measures. Then the tolerance interval is given by $(q_2 - C_L d_L, q_2 + C_U d_U)$, for some predetermined constants C_L, C_U .

The quartile method is a robust method against outliers. Usually, the constants



C_L , C_U are obtained from past data. Because of the fact that in most of the cases the population is skewed, it occurs that the lower and upper bounds of the tolerance interval are unequal. Sometimes the sample may be bounded on one side. In such cases it would be preferable to use a one-sided tolerance interval, obtained by replacing one of the limits by the lowest or highest value.

3.4 The Hidioglou-Berthelot method

Hidioglou and Berthelot (1986) developed a modified quartile method for detecting outliers in trend data.

Let $r_{i,t} = y_i(t)/y_i(t-1)$ be the ratio of the value of y for unit i , obtained at time t to the value for the same unit at time $t-1$, i.e., the change ratio. For instance, consider a business survey where $y_i(t)$ denotes the size of company i at period t . If the company increases in size between two periods, then the change ratio will increase as well, and the present value will be considered an outlier.

The change ratio fluctuates more likely for units with small values and is frequently skewed. So, when comparing small to large companies, the method tends to detect the small ones as outliers (the “size masking effect”). For this reason, Hidioglou and Berthelot proposed a two-step transformation, which reduces both the skewness of the data and the “size masking effect”.



Let s_i be the transformed values of r_i , which are calculated as follows:

$$s_i = \begin{cases} 1 - \frac{m}{r_i}, & \text{if } 0 < r_i < m, \\ \frac{r_i}{m} - 1, & \text{if } r_i \geq m, \end{cases} \quad (3.3)$$

where m is the median of the sample. For the second step the effect of unit i , E_i , is computed as follows:

$$E_i = s_i [\max(y_i(t-1), y_i(t))]^V, V \in [0, 1] \quad (3.4)$$

To increase the value of the exponent V , means also increasing the significance of large values and thus decreasing the “size masking effect”, while a value of zero to the exponent means that we do nothing about it. Hidiroglou and Berthelot applied the quartile method to the E_i 's to detect trend outliers.

3.5 Outlier detection in the HCSO

A rather unique method was used in 2014 for detecting outliers in the Hungarian Central Statistical Office. The method aims at decreasing the impact of outlying observations on the estimates. Therefore, a new stratum is created for every outlier, and the rest of the observations are re-weighted according to the amount of outliers in the strata.

Some auxiliary indicators were computed in order to decide whether or not a particular observation is an outlier. The steps for creating the most important indicator, $lnsqr$, are the following:



1. Compute the standardised value of each variable y_{ji} in each sampling stratum j

$$standard_{ji} = (y_{ji} - mean_j) / std_j,$$

where $mean_j, std_j$ are the mean and standard deviation of stratum j , respectively, so that the new variable has mean 0 and variance 1.

2. Compute the next indicator, $stand_{ji}$, only for strata with size $n_j \geq 3$.

$$stand_{ji} = \frac{standard_{ji}}{gr_j} \quad (3.5)$$

The denominator gr_j is the Grubbs¹ critical value for stratum j . If $stand_{ji} > 1$, then the unit i is detected as an outlier.

3. Finally, calculate $lnsqrt_{ji}$ for every observation, which is:

$$lnsqrt_{ji} = \ln y_{ji} \sqrt{stand_{ji}} \quad (3.6)$$

What is interesting about $lnsqrt$ is that it gives missing values when $y_{ji} = 0$, $stand_{ji} < 0$ and when the stratum size n_j is smaller than 3. So, we can only calculate $lnsqrt$ for strata with large sizes and for those observations whose values are greater than the mean of the stratum they belong to. Another interesting characteristic about this index is its density function, which is a curve, usually with two maxima and a break at the right end, used for selecting outliers.

¹Grubbs' test for outliers (1950), is based on the assumption that the data are normally distributed and it is used for detecting one outlier at a time. The null hypothesis is: the sample contains no outliers and the one-sided test statistic for checking whether the maximum is an outlier is

$$G = \frac{y_{max} - \bar{y}}{s}. H_0 \text{ is rejected at significance level } \alpha, \text{ if } G > \frac{n-1}{n} \sqrt{\frac{t_{\alpha/n, (n-2)}^2}{n-2+t_{\alpha/n, (n-2)}^2}}.$$

3.6 Cook's D

Cook's distance, named after R. Dennis Cook (1977), is used to measure the influence of a data point in the LS estimate of the regression coefficient. It is defined as:

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}, \quad (3.7)$$

where \hat{y}_j is the predicted value of y_j for the full model, $\hat{y}_{j(i)}$ the predicted value of y_j if we omit observation i , p the number of fitted parameters in the model, MSE the mean square error of the model, e_i the regression residuals and h_{ii} the leverage points².

Cook's distance measures the effect on the regression line when we exclude a certain observation from the full model. Units that have either a high residual and/or high leverage need to be further examined because including them in the regression would possibly alter its outcome.

Cook (1986) proposed a cut-off for values with $D_i > 1$, while Bollen (1990) suggested using data points with $D_i > 4/n$. However, this method suffers from the "masking effect", as it is based on an LS estimator which is not robust.

²Leverage are the observations that are outlying in the value of the independent variable, and are defined as $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$, $h_{ii} \in [0, 1]$.



3.7 Mahalanobis distance

In 1936, P. C. Mahalanobis introduced the Mahalanobis distance, as a measure of the distance between a point P and a distribution D . It is widely used in cluster analysis, as it actually measures how many standard deviations P lies away from the mean of D in a multi-dimensional space. The distance is also scale-invariant and takes into account the correlations of the data.

The Mahalanobis distance is defined as $D_M(x) = \sqrt{(x - \mu)^\top S^{-1}(x - \mu)}$, where $x = (x_1, \dots, x_n)^\top$ the vector of the observations, $\mu = (\mu_1, \dots, \mu_n)^\top$ the vector of the means and S the covariance matrix. So, it measures the distance of a multi-dimensional vector x from its center.

In survey sampling, the Mahalanobis distance can be used for detecting multivariate regression outliers. First, one has to compute the distance using robust estimates for the location and scale parameters, and then use the normalized residuals of the robust regression to detect ordinary least squares outliers and leverage points, see Rousseauw and Van Zomeren (1990). The method however is not easy to apply on large datasets, as it demands a lot of computation.

The Mahalanobis distance was used on normally distributed variables (a log-transformation of the original variables) for identifying multivariate outliers in the HCSO, and the results were satisfactory.

Apart from the methods presented above, other ad hoc methodologies have also



been occasionally proposed for the detection of outliers in survey data. The lack of distributional assumption makes it difficult to develop a detection method, and the problem becomes even bigger when a complex sample design is used.





Chapter 4

Treatment of Outliers

The treatment of outliers is performed at the estimation stage, when the data are already clean from errors and the remaining outliers in the dataset are representative of other units in the population. When the outliers are identified, the objective is to reduce the variance of a nearly unbiased estimator, but it is possible that this reduction may introduce bias. The most common treatment procedures can be divided into three main categories: winsorization of the largest values, reduction of the weights assigned to outliers and robust estimation techniques, such as M-estimation.

4.1 Winsorized estimators

Many variables of interest in most household and business surveys are non-negative and skewed to the right. Winsorization, i.e., replacement of the largest values of the sample with a large order statistic, is applied to the expansion estimator



$\hat{Y} = (N/n) \sum_{i=1}^n y_i$, for simple random sampling without replacement.

Let $y_{(1)}, \dots, y_{(n)}$ be the ordered observations of the sample. The k -times Winsorized estimator of the total, and the corresponding estimator of the mean are respectively:

$$\hat{Y}_w = \frac{N}{n} \left(\sum_{i=1}^{n-k} y_{(i)} + ky_{(n-k)} \right), \quad (4.1)$$

$$\hat{Y}_w = \frac{1}{n} \left(\sum_{i=1}^{n-k} y_{(i)} + ky_{(n-k)} \right) \quad (4.2)$$

An alternative winsorized estimator can be calculated if we replace the k largest values of the variable y with a threshold $R \in (y_{(n-k)}, y_{(n-k+1)}]$, above which a sample unit is considered an outlier.

$$\hat{Y}_{w*} = \frac{N}{n} \left(\sum_{i=1}^{n-k} y_{(i)} + kR \right), \quad (4.3)$$

$$\hat{Y}_{w*} = \frac{1}{n} \left(\sum_{i=1}^{n-k} y_{(i)} + kR \right). \quad (4.4)$$

4.1.1 Searls' Winsorized mean

Searls (1966) proposed a non-parametric alternative estimator for the population mean μ that is obtained as follows:

$$\hat{Y}_S = \sum_{i=1}^n \min(y_i, R) / n \quad (4.5)$$



where R is the cut-off point. Rivest and Hurtubise (1993) proposed that, when the population distribution $F(y)$ is known or can be estimated from past data, the optimal cut-off \hat{R} for fixed n is

$$R = (n - 1)E_F[\max(Y - R), 0] + \mu \quad (4.6)$$

where E_F is the expectation taken with respect to the distribution F . For given R , the estimated MSE of \hat{Y}_S is:

$$MSE(\hat{Y}_S) = \frac{\sum_{i=1}^n [\min(y_i, R) - \bar{y}_R]^2}{n(n-1)} + \left[\frac{\sum_{i=1}^n \max(y_i - R, 0)}{n} \right]^2 \quad (4.7)$$

When we obtain R from (4.6), the expected number of Winsorization points, $n[1 - F(R)]$, decreases as the skewness of F increases.

4.1.2 The Once-Winsorized mean

Assume now that we use a threshold $R = y_{(n-1)}$. Then the estimator we obtain for the population mean is the once-Winsorized mean:

$$\hat{Y}_{once} = \bar{y} - \frac{y_{(n)} - y_{(n-1)}}{n}, \quad (4.8)$$

where $\bar{y} = 1/n \sum_{i \in S} y_i$. Rivest (1994) suggested the following MSE estimator :

$$MSE(\hat{Y}_{once}) = \frac{s^2}{n} - \frac{(y_{(n)} + y_{(n-1)} - 2\bar{y}_{once})(y_{(n)} - 3y_{(n-1)} + 2y_{(n-2)})}{n^2}. \quad (4.9)$$

The once-Winsorized mean is the most efficient Winsorized mean when R is an extreme-order statistic, and is consistent even when the distribution has infinite



variance. For distributions that are more skewed than the exponential, winsorizing improves the efficiency of the sample mean. It is also interesting that \hat{Y}_{once} and \hat{Y}_{searls} have the same asymptotic distribution as \bar{y} .

4.1.3 Fuller's mean

Fuller (1991) proposed a model-based estimator for the sample mean. The first step of the method is to find out whether the assumed distribution F has heavier tails than the exponential. If the null hypothesis that the tails are exponential is rejected, then the high values are winsorized.

For the first step the test statistic is:

$$F_{T_j} = \frac{[\sum_{i=n-j+1}^n (n-i+1)(y_{(i)} - y_{(i-1)})]/j}{[\sum_{i=n-T_j}^{n-j} (n-i+1)(y_{(i)} - y_{(i-1)})]/(T_j - j + 1)} = \frac{N_j}{D_j}, \quad (4.10)$$

where j and T_j are tuning parameters that have to be defined. Under the null hypothesis $F_{T_j} \sim F_{2j, 2(T_j - j)}$, where $F_{2j, 2(T_j - j)}$ the F distribution with $2j, 2(T_j - j)$ degrees of freedom.

Now we calculate Fuller's Winsorized mean as:

$$\hat{Y}_F = \begin{cases} \bar{y}, & \text{if } F_{T_j} < K_j, \\ \frac{\sum_{i=1}^{n-j} y_{(i)} + j(y_{(n-j)} + K_j D_j)}{n}, & \text{if } F_{T_j} \geq K_j, \end{cases} \quad (4.11)$$

Fuller suggested the values $j = 3, T_j = 4\sqrt{n} - 10, K_j = 3.5$ for the three tuning parameters. Unfortunately, there is no estimator for the $MSE(\hat{Y}_F)$.

4.2 Re-weighted estimators

Instead of replacing the highest values of a variable y with a large-order statistic, one can alternatively modify the weights that are assigned to them. The outliers are usually down-weighted and the remaining sample units have their weights increased, so that the new weights sum up to N , as before. The method can be applied both in the case of s.r.s. and stratified s.r.s.

Bershad (1960) proposed an estimator for the total in srs, which reduced the weight of the k -outliers to $r < N/n$:

$$\hat{Y}_{rew} = \frac{N}{n} \sum_{i=1}^{n-k} y_i + r \sum_{i=n-k+1}^n y_i, \quad (4.12)$$

while a different version of the above under the restriction that the sum of weights equals N is:

$$\hat{Y}_{rew2} = \frac{N - rk}{n - k} \sum_{i=1}^{n-k} y_i + r \sum_{i=n-k+1}^n y_i. \quad (4.13)$$

To calculate (4.12) and (4.13), one has first to decide the reduced weight r for the k -outliers. Rao (1971) and Chinnappa (1976) proposed the estimator (4.13) with $r = 1$ for outliers that are unique and under the assumption that all the outliers of the population are sampled. If we divide the above estimators by N we obtain estimators of the mean.

4.2.1 The post-stratified estimator

Assume that we have sampled k out of K outliers in the population, where K is known i.e. the outliers are representative of other units in the population. Then,



we can calculate the post-stratified estimator of the total, which is:

$$\hat{Y}_{post} = \frac{N-K}{n-k} \sum_{i=1}^{n-k} y_i + \frac{K}{k} \sum_{i=n-k+1}^n y_i. \quad (4.14)$$

In order to find the optimal weight for estimator (4.13), Hidiroglou and Srinath (1981) minimized its MSE, conditional and unconditional on k . They found that the optimal weight is a function of K , \bar{Y} (the population mean), and the variances of outliers and non-outliers.

They also computed another re-weighted estimator, in which the weight reduction depends on the number of outliers in the sample, k :

$$\hat{Y}_{rew3} = \frac{N}{n} \left[1 + \frac{k}{2n} \right] \sum_{i=1}^{n-k} y_i + \frac{N}{n} \left[1 - \frac{n+k}{2n} \right] \sum_{i=n-k+1}^n y_i \quad (4.15)$$

4.3 The variance-inflation model

Ghangurde (1989a, 1989b) used the variance-inflation model to create a model-based estimator that is robust to outliers. The model is:

$$y_i = \beta x_i + \varepsilon_i, \quad (4.16)$$

$\varepsilon_i \sim (0, \sigma_1^2 x_i)$, $i = 1, \dots, n-k$ and $\varepsilon_i \sim (0, \sigma_2^2 x_i)$, $i = n-k+1, \dots, n$. The parameters β , σ_1^2 , σ_2^2 are unknown. We consider the first $n-k$ observations to be non-outliers, and the rest k to be outliers, and let $w = \sigma_1^2 / \sigma_2^2$, $w \in (0, 1]$. The BLUE estimator



for β , with respect to the ratio estimator $\hat{Y} = \bar{Y}X/\bar{X}$ is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n-k} y_i + w \sum_{i=n-k+1}^n y_i}{\sum_{i=1}^{n-k} x_i + w \sum_{i=n-k+1}^n x_i} \quad (4.17)$$

The above expression shows that w is the weight reduction factor i.e., the weights of the outliers are optimally reduced to w .

4.4 Dalén and Tambay's method

Dalén (1987) and Tambay (1988) developed a method that detects and treats influential observations in unequal probability sampling. Their method can be considered as both Winsorization and weight-reduction, as the expression presented below involves both a cut-off point and a modified weight for outliers.

Let

$$z_i = \begin{cases} w_i y_i, & \text{if } w_i y_i < T \\ T + (y_i - \frac{T}{w_i}), & \text{otherwise} \end{cases} \quad (4.18)$$

where w_i is the design weight of observation i , T is a threshold and $\hat{Y}_{DT} = \sum_i z_i$ the estimation for the total. This estimator gives a reduced weight of 1 to the observations that exceed the cut-off point $w_i y_i - T$. Hidiroglou (1991) proved that if we have simple random sampling and the location parameter is estimated by the winsorized mean at a cut-off point fT , where $f = n/N$, the above estimator reduces to that of Fuller.



4.5 ψ -type M-estimators

M-estimators, introduced by Huber (1964), are a broad class of maximum-likelihood-type estimators, widely used in robust statistics. In particular, a ψ -type M-estimator is defined as follows:

Let \mathcal{Y} and Θ be measure spaces with $\theta \in \Theta$ the vector of parameters, for a positive integer r . The ψ -type M-estimator is defined through a measurable function $\psi: \mathcal{Y} \times \Theta \rightarrow \mathbb{R}^r$, and is the solution of

$$\int_{\mathcal{Y}} \psi(y, \theta) dF(y) = 0, \quad (4.19)$$

for some distribution F . Expression (4.19) can also be used when we want to estimate the location parameter θ in a sample of size n as

$$\sum_{i=1}^n \psi(y_i - \theta) = 0. \quad (4.20)$$

If we take $\psi(r) = r$, the estimator obtained is the Ordinary Least Squares estimator i.e, the sample mean.



4.5.1 The Huber M-Estimator

Huber (1964) used the following ψ -function, in order to obtain an M-estimator:

$$\psi(t) = \begin{cases} c, & \text{if } t > c, \\ t, & \text{if } |t| \leq c, \\ -c, & \text{if } t < -c, \end{cases} \quad (4.21)$$

where c is the tuning constant that shows how many outliers are treated. As $c \rightarrow 0$, the above estimator becomes the sample median, while when c is very big, then no outliers are treated and the estimator for the location parameter is the same as the least squares estimator.

4.5.2 Bruce's period-to-period change estimator

Bruce (1991), in order to estimate period-to-period change for finite populations, used Huber's M-estimator, under the constraint that the weight of outliers should not be less than 1. This gives us a slightly different version of (4.21), which is:

$$\psi(t) = \begin{cases} (1-f)c + ft, & \text{if } t > c, \\ t, & \text{if } |t| \leq c, \\ -(1-f)c - ft, & \text{if } t < -c, \end{cases} \quad (4.22)$$

with f being the sampling fraction. The ψ -type functions are widely used in the context of robust regression estimators; see next section.



4.5.3 Regression Estimators

Assume that for each sample unit i we can obtain a vector \mathbf{x}_i of auxiliary variables, whose totals in the population, $\sum_{i \in U} \mathbf{x}_i$, are known. The sampling weights can be calibrated to the known totals. So, we can write the calibration equation $\sum_{i \in U} \mathbf{x}_i = \sum_{i \in S} \mathbf{x}_i c_i$, where c_i stands for the calibrated weight for unit i . Consequently, the calibrated estimator for the total of Y can be written as $\hat{Y} = \sum_{i \in S} y_i c_i$.

For the linear model we have that $E(y_i | \mathbf{X}) = \mathbf{x}_i' \boldsymbol{\beta}$ and $\text{Var}(y_i | \mathbf{X}) \propto v_i = \mathbf{x}_i' \boldsymbol{\lambda}$, $i \in U$, with respect to the model, where $\boldsymbol{\beta}$ is the vector of the unknown parameters of the model and $\boldsymbol{\lambda}$ is known.

Chambers' M-estimator

Chambers (1986) developed a robust model-based estimator for the population total Y . As the Least Squares estimator for $\boldsymbol{\beta}$ can be affected from outliers, he used an M-estimator to make $\hat{\boldsymbol{\beta}}$ robust, and therefore introduced bias. A bias correction had to be incorporated.

Beaumont and Rivest (2008) used a formula similar to that of Chambers, in order to show that calibrated estimators can be affected by outlying observations:

$$\hat{Y}^C = \sum_{i \in S} y_i + \sum_{i \in U-s} \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R + \sum_{i \in S} u_i \frac{y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R}{\sqrt{v_i}}, \quad (4.23)$$

where $u_i = (c_i - 1)\sqrt{v_i}$ are the new weights, $y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R / \sqrt{v_i}$ are the standardized residuals for the units included in the sample, and $\hat{\boldsymbol{\beta}}^R$ the robust estimator of $\boldsymbol{\beta}$.



Two interesting conclusions can be drawn from (4.23): The first is that we may obtain extreme-valued residuals for the non-sample units, which can affect the sampling error very much, but there is nothing that can be done to prevent this from happening at the estimation stage. The second conclusion is that observations which bring a large standardized residual, when multiplied with a large weight u_i will also affect the sampling error. Therefore, such observations should be down-weighted.

To limit the impact of outliers, Beaumont and Rivest (2008) used a generalized Schweppe-type M-estimator, see also Hampel et al. (1986). The robust estimator for Y is:

$$\hat{Y}_{rob} = \sum_{i \in s} y_i + \sum_{i \in U-s} \mathbf{x}_i' \hat{\beta}^R + \sum_{i \in s} \frac{u_i}{h_i} \psi \left[h_i \frac{y_i - \mathbf{x}_i' \hat{\beta}^R}{\sqrt{v_i}} \right], \quad (4.24)$$

where $\psi(t)$ is Huber's ψ -function and h_i is a weight that can both depend on the value of the i -th observation x_i and on the original sampling weight $w_i = 1/\pi_i$. In the case of a census where $s = U$ and $c_i = 1, i \in U$, the above estimator becomes $Y = \sum_{i \in U} y_i, \forall \psi, h_i$.

Under the assumption that $\mathbb{E}[\hat{Y}_{rob}] \approx Y$, the MSE of \hat{Y}_{rob} can be written as follows:

$$MSE[\hat{Y}_{rob}] = \mathbb{E}[\hat{Y}_{rob} - Y]^2 \approx Var(\hat{Y}_{rob}) + Bias(\hat{Y}_{rob})^2 \quad (4.25)$$

$$\approx Var(\hat{Y}_{rob}) + [\mathbb{E}(\hat{Y}_{rob} - \hat{Y}^C)^2 - Var(\hat{Y}_{rob} - \hat{Y}^C)] \quad (4.26)$$



However, Gwet and Rivest (1992) proposed another estimator for the mean squared error of \hat{Y}_{rob} , in the direction of finding a tuning constant c for the Huber ψ -function that would minimize the $MSE(\hat{Y}_{rob})$. The estimator is:

$$MSE_2[\hat{Y}_{rob}] = v(\hat{Y}_{rob}) + \max\{0, (\hat{Y}_{rob} - \hat{Y}^C)^2 - v(\hat{Y}_{rob} - \hat{Y}^C)\}, \quad (4.27)$$

where $v(\cdot)$ is a design-consistent estimator of $Var(\cdot)$.

Lee's M-estimator

Lee (1991a) studied another version of (4.23) in order to propose a robust multiple regression estimator. He used the generalized M-estimation technique to derive the form of a Mallows-type estimator. Such estimators tend to be biased and inconsistent, see Isaki and Fuller (1982). So, to make the estimator consistent, Lee (1991a) multiplied the sum of residuals by $\theta \in [0, 1]$:

$$\hat{Y}_{lee} = \sum_{i \in s} y_i + \sum_{i \in U-s} \mathbf{x}'_i \hat{\beta}^R + \theta \sum_{i \in s} w_i (y_i - \mathbf{x}'_i \hat{\beta}^R), \quad (4.28)$$

where $w_i = 1/\pi_i$. For this estimator to be consistent, θ needs to be close to 1, as n approaches infinity.

Hulliger's H-T estimator

Basu's elephant example (1971) indicates that the Horvitz-Thompson estimator is sensitive to outliers. So, Hulliger (1995), used the M-estimation method to create a robust Horvitz-Thompson estimator.



Hulliger used the model $y_i = \beta z_i + \varepsilon_i$, where the variable of interest y is linearly associated with the design variable z , which is a size measure, and has the form $z_i = \sum_{i \in s} w_i / (n w_i)$. Also, for the error term of the model, we have that $\sigma_\varepsilon^2 = z_i \sigma^2$.

In this case, we are interested in estimating the population mean for the variable y , $\bar{Y} = \sum_{i \in U} y_i / N$. The Horvitz-Thompson estimator for the mean is $\hat{Y}_{HT} = \bar{z} \hat{\beta}$, where $\bar{z} = \sum_{i \in s} w_i z_i / \sum_{i \in s} w_i = 1$ the Hajek estimator.

Hulliger (1995) assumed that there are no outlying weights, therefore robustification should be applied only on the residuals $(y_i - \beta x_i) / \sqrt{z_i}$. The estimating equation for $\hat{\beta}$ is as follows:

$$\sum_{i=1}^n w_i \psi \left[\frac{y_i - \beta z_i}{\sqrt{z_i}} \right] \sqrt{z_i} = 0, \quad (4.29)$$

where $\psi(\cdot)$ is the Huber psi-function. The M-estimator for $\hat{\beta}$ is obtained through an Iteratively Reweighted Least Squares (IRLS) algorithm, for which a starting value $\beta^{(0)} = \text{med}_i(y_i, w_i) / \text{med}_i(z_i, d_i)$ was proposed, see Hulliger et al (2011), and $\text{med}_i(y_i, w_i)$ is the weighted median of y_i . Then, the standardized residuals are calculated as $r_i = (y_i - \beta^{(0)} z_i) / \sqrt{z_i}$, in order to compute the robust estimate for the scale parameter $\hat{\sigma}_\varepsilon = \text{MAD}(r_i, w_i)$.

So, the robust weight for unit i is calculated as:

$$u_i = \frac{\psi(r_i / \hat{\sigma}_\varepsilon)}{|r_i / \hat{\sigma}_\varepsilon|}. \quad (4.30)$$



For the new robust weights we have that $u_i = 1$ for observations that are not outliers, and $u_i < 1$ for outliers. Moreover, $r_i \rightarrow 0$ for extreme outliers. Finally, from the IRLS algorithm we have that the robust estimate for the regression coefficient at the $(t + 1)^{th}$ iteration is:

$$\beta^{(t+1)} = \frac{\sum_{i=1}^n w_i u_i y_i}{\sum_{i=1}^n w_i u_i z_i}. \quad (4.31)$$

In equation (4.29), if we replace z_i by a covariate x_i which is related to y_i , and for which the population total \bar{x} is known, then we can obtain a robust ratio estimator, see Gwen and Rivest (1992) and Hulliger (1995).

Weight Smoothing

Beaumont (2008) introduced a weight smoothing approach to the estimation of the population total $Y = \sum_{i \in U} y_i$, under the assumption that the design weights w_i are random quantities.

Let us consider the vector of design variables $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)'$, for example a size measure. Let us also define the inclusion indicator of population unit i in the sample, $\mathbf{I} = (I_1, \dots, I_N)'$ i.e.,

$$I_i = \begin{cases} 1, & \text{if unit } i \text{ is included in the sample,} \\ 0, & \text{otherwise.} \end{cases} \quad (4.32)$$



The design weights are a function of \mathbf{Z} only. Under the assumption of randomness, the following model ξ is constructed:

$$E_{\xi}(w_i|\mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{x}_i, y_i; \alpha_s), i \in s, \quad (4.33)$$

where g_s is some function, α_s is a vector of unknown model parameters that have to be estimated from the sample and \mathbf{X} is the matrix of explanatory variables. The variable of interest y can either be univariate or multivariate. The purpose of using the model ξ is to eliminate the redundant variability in the design weights.

Let us consider $\hat{\alpha}_s$, a consistent estimator of α_s with respect to the model ξ , in order to obtain $\hat{w}_i = g_s(\mathbf{x}_i, y_i; \hat{\alpha}_s)$. Then, the smoothed estimator for the population total (if we apply the method on the Horvitz-Thompson estimator) becomes

$$\hat{Y}_{SDB} = \sum_{i \in s} \hat{w}_i y_i. \quad (4.34)$$

For the mean squared error of the smoothed estimator we have

$$MSE(\hat{Y}_{SDB}) = v(\hat{Y}_{SDB}) + \max\{0, [\hat{Y}_{SDB} - \hat{Y}_{HT}]^2 - v[\hat{Y}_{SDB} - \hat{Y}_{HT}]\}, \quad (4.35)$$

where $\hat{Y}_{HT} = \sum_{i \in s} (1/\pi_i)y_i$ and $v(\hat{Y}_{SDB})$, $v[\hat{Y}_{SDB} - \hat{Y}_{HT}]$ are design-consistent estimators of $Var(\hat{Y}_{SDB})$ and $Var(\hat{Y}_{SDB} - \hat{Y}_{HT})$, respectively.

The weight smoothing approach can as well be applied to calibration estimators. Especially for the case of the H-T estimator, Beaumont (2008) showed that, if the linear model is true, \hat{Y}_{SDB} is unbiased and equally or more efficient than the



Horvitz-Thompson estimator, under the model ξ and the sampling design.

4.5.4 The Conditional Bias

Beaumont, Haziza and Gazen (2013) introduced a new approach for deriving robust estimates when sampling from a finite population, by using the conditional bias that is associated with a sample unit. They used the conditional bias in order to construct both a model based estimator and a robust Horvitz-Thompson estimator.

The Model-Based Approach

Assume that the variable of interest Y has been generated by some model. Let also X be the matrix that contains the auxiliary variable x_i^\top in its i^{th} row, and the x_i^\top 's are assumed to be independent and generated from the same distribution F . For the linear model, we have that $Z|X \sim (0, 1)$, where $Z_i = (Y_i - x_i^\top \beta) / \sigma_i$, $i = 1, \dots, N$ and are mutually independent.

Now assume we are interested in estimating the population total $Y = \sum_{i \in U} Y_i$ and we select a random sample s from the population U . The best linear unbiased predictor (BLUP) of Y , as proposed by Royall (1976), is $\hat{Y} = \sum_{i \in s} r_i y_i$, where r_i are weights

$$r_i = 1 + \frac{x_i^\top}{\sigma_i^2} \left[\sum_{i \in s} \frac{x_i x_i^\top}{\sigma_i^2} \right]^{-1} \left(\sum_{i \in U-s} x_i \right). \quad (4.36)$$



The conditional bias of unit i is therefore $B_i(y_i; \beta) = \mathbb{E}_F(\hat{Y} - Y | s, Y_i = y_i)$ and the expectation is evaluated with respect to the model F . Also, if we note that $\sum_{i \in s} r_i x_i = \sum_{i \in U} x_i$, we can see that the conditional bias takes a different form for sample and non-sample units. So, we have that

$$B_i(y_i; \beta) = \begin{cases} (r_i - 1)(y_i - x_i^\top \beta), & \text{if } i \in s, \\ -(y_i - x_i^\top \beta), & \text{if } i \in U - s \end{cases} \quad (4.37)$$

The prediction error of \hat{Y} is therefore, $\hat{Y} - Y = \sum_{i \in U} B_i(Y_i; \beta)$. In order to make the best linear unbiased predictor robust, one has to downweight the contribution of influential units in the term on the right-hand side of the above expression.

Therefore, the robust BLUP of Y will be

$$\hat{Y}_R(\beta) = \sum_{i \in s} Y_i + \sum_{i \in U-s} x_i^\top \beta + \sum_{i \in s} \psi[(r_i - 1)(Y_i - x_i^\top \beta)], \quad (4.38)$$

where ψ is a bounded function, for example Huber's psi. The vector of unknown parameters β should be estimated either by a robust estimator such that of Chambers (1986) or by an independent source of data.

The Robust Horvitz-Thompson Approach

Assume now we are interested in estimating the finite population total, $Y = \sum_{i \in U} y_i$. The Horvitz-Thompson estimator of Y , $\hat{Y}_{HT} = \sum_{i \in s} w_i y_i$, is design-unbiased, which means $\mathbb{E}_{\mathcal{P}}(\hat{Y}_{HT}) = Y$ with respect to the sampling design \mathcal{P} . The probability of including unit i in the sample is $\pi_i = P(I_i = 1)$, where I_i the inclusion indicator for



unit i .

The conditional bias of the HorvitzThompson estimator for a unit i is

$$B_{1i}^{HT} = \mathbb{E}_{\mathcal{D}}(\hat{Y}_{HT}|I_i = 1) - Y = \sum_{j \in U} \left[\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right] y_j, \quad (4.39)$$

where $\pi_{ij} = P(I_i = 1, I_j = 1)$ the second-order inclusion probability of units i and j . For a unit that is not included in the sample, we have $B_{0i}^{HT} = \mathbb{E}_{\mathcal{D}}(\hat{Y}_{HT}|I_i = 0) - Y = -(w_i - 1)^{-1} B_{1i}^{HT}$.

For Poisson sampling, stratified srs and for fixed-size high-entropy sampling with varying first-order inclusion probabilities, we have that

$$\hat{Y}_{HT} - Y \simeq \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in U-s} B_{0i}^{HT} \quad (4.40)$$

If we use a decomposition analogous to (4.38), it can be shown that the robust Horvitz-Thompson estimator can be expressed as follows:

$$\hat{Y}_{RHT} = \hat{Y}_{HT} - \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in s} \psi[B_{1i}^{HT}] \quad (4.41)$$

Because the conditional bias B_{1i}^{HT} depends on unknown population parameters, it should be estimated by either an independent source of data or by current data.



In this chapter, various methods were presented for treating outliers at the estimation stage of a survey. Of course, some treatment of outliers can be done at the editing stage, but that will not concern us within the context of this thesis. The lack of a parametric assumptions about finite survey populations make dealing with outliers a difficult task.





Chapter 5

Empirical Study

An empirical study was conducted in order to evaluate the methodologies presented in chapters 2 and 3, regarding the detection and treatment of outliers in survey data. The data that were used for the study were those of the Greek component of the European Survey on Income and Living Conditions (EU-SILC) for year 2013, provided by the Greek Statistical Authority (ELSTAT).

EU-SILC is the EU reference source for comparative statistics on income distribution and social exclusion at European level, particularly in the context of the “Programme of Community action to encourage cooperation between Member States to combat social exclusion” and for producing structural indicators on social cohesion for the annual spring report to the European Council. It provides two types of annual data:

- Cross-sectional data pertaining to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions, and



- Longitudinal data pertaining to individual-level changes over time, observed periodically over a four year period

The reference population of EU-SILC is all private households and their current members residing in the territory of the Member States at the time of data collection. Persons living in collective households and in institutions are generally excluded from the target population.

An integrated design (rotational design) has been proposed by Eurostat. Rotational design refers to the sample selection based on a number of sub-samples or rotations, each of them similar in size and design and representative of the whole population. From one year to the next, one of the rotations (the “older”) is replaced by a newly selected rotation.

The fundamental characteristic of the integrated design is that the cross-sectional and longitudinal statistics are produced from essentially the same set of sample observations, thus avoiding unnecessary duplications which entirely separate cross-sectional and longitudinal surveys would involve.

Now we turn to the Greek EU-SILC for year 2013. A two-stage stratified cluster sampling design was implemented, and a sample of 7439 respondent households and 15318 respondent persons were obtained. The population was stratified by geographic area (13 regions plus the metropolitan areas of Athens and Salonica) and by population density (rural, semi-urban and urban areas). In the first stage of sampling, a number of geographic clusters (small communities or city blocks) were selected independently within strata with probability proportional to



size (number of households). In the second stage of sampling, a fixed number of dwellings were selected in each cluster with systematic sampling.

For the empirical study, we chose the following four variables of interest from the household file:

1. HH070: Total Housing Cost
2. HY010: Total Household Gross Income
3. HY020: Total Disposable Household Income
4. HY022: Total Disposable Household Income Before Social Transfers Other Than Old-Age and Survivor's Benefits.

The first variable, HH070, refers to the total housing cost of a household on a monthly basis, while the other three variables refer to yearly quantities.

Table 5.1 shows the sample measures of tendency, dispersion and asymmetry for each variable. In particular, the term “mean” refers to the sample mean, $1/n \sum_{i=1}^n y_i$. It can be seen that the distributions of all variables are extremely skewed; in a symmetric distribution, the sample skewness $\gamma_1 = \mu_3/\mu_2^{3/2}$ would equal 0 whereas in our data it is greater than 7 for every variable. Also, the Horvitz-Thompson weighted mean for each of the four variables is respectively 398.6485, 21054.45, 15324.45 and 14653.36 euros.

Figure 5.1 shows the correlation between the four variables and the design weights (variable DB080). Looking at the scatter plots, we can see that large

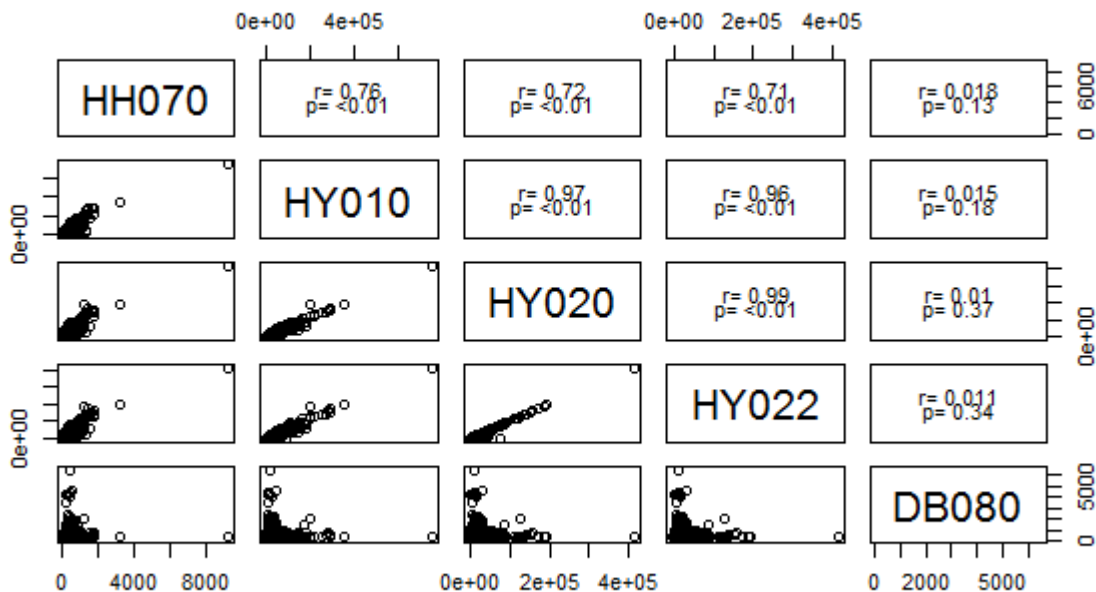


Table 5.1: Descriptive Statistics

| Variable | Descriptives | | | |
|----------|--------------|----------|----------|-----------|
| HH070 | Min | 82.55 | Max | 9189.72 |
| | Q1 | 294.64 | Skewness | 17.98 |
| | Median | 375.73 | St. Dev. | 176.63 |
| | Mean | 397.03 | M.A.D. | 127.76 |
| | Q3 | 469.07 | | |
| HY010 | Min | 0 | Max | 748822.13 |
| | Q1 | 9597.9 | Skewness | 8.91 |
| | Median | 15579.82 | St. Dev. | 21314.96 |
| | Mean | 20884.46 | M.A.D. | 10661.76 |
| | Q3 | 25876.99 | | |
| HY020 | Min | -2600 | Max | 415950 |
| | Q1 | 8000 | Skewness | 7.09 |
| | Median | 12420 | St. Dev. | 12855.08 |
| | Mean | 15255.2 | M.A.D. | 7715.45 |
| | Q3 | 19151.25 | | |
| HY022 | Min | -5140 | Max | 415950 |
| | Q1 | 7200 | Skewness | 7.01 |
| | Median | 11810 | St. Dev. | 12930.92 |
| | Mean | 14579.65 | M.A.D. | 7857.78 |
| | Q3 | 18500 | | |

values of the y -variables do not seem to be associated with large design weights. Also, we can assume that the design weights are linearly uncorrelated to the variables of interest, as the correlation coefficient $r = \hat{\rho}$ is less than 2 percent in each case (p stands for the p -value when we test the null hypothesis $H_0 : \rho = 0$).

Figure 5.1: Matrix Plot



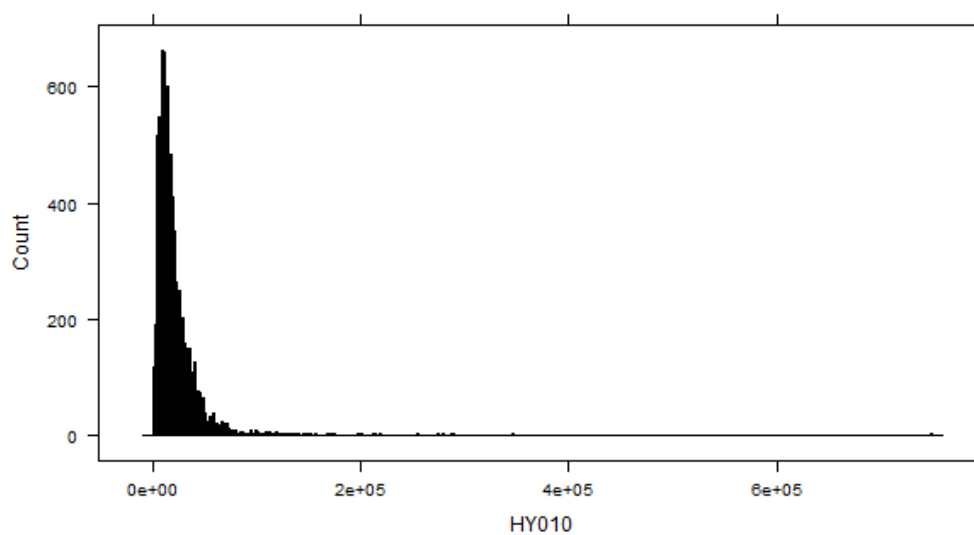
In the next two sections, we illustrate the application of the methods which are most appropriate for complex data on the Total Household Gross Income,¹ HY010. The results for the rest of the variables are included in Appendices A and B.

¹The total household gross income is computed as the sum for all household members of gross personal income components plus gross income components at household level.

5.1 Detection

Figure 5.2 shows the histogram of the total household gross income. The bulk of the data is located on the left hand side of the diagram; 75% of the households have a total gross income that is less than 25877 euros per year, but there are also observations that have very large y-values and appear with low frequency. Such observations are most likely to be identified as outliers when we perform detection.

Figure 5.2: Histogram of HY010



The methods that were used for the detection of outliers are the relative distance,

the quartile method and the M.A.D. The use of these methods requires the specification of certain parameters, therefore we tested each method with 10 different values for each parameter. The results are shown in Table 5.2.

Table 5.2: Outlier Detection for HY010

| Relative Distance | | | Quartile Method | | | M.A.D. | | |
|-------------------|-----------|-----------------|-----------------|-----------|-----------------|--------|-----------|-----------------|
| c | n_{out} | $y_{(1)}^{out}$ | c_u | n_{out} | $y_{(1)}^{out}$ | k | n_{out} | $y_{(1)}^{out}$ |
| 5.50 | 127.00 | 74276.00 | 5.00 | 188.00 | 67203.10 | 6.00 | 104.00 | 81380.50 |
| 6.00 | 104.00 | 81380.50 | 6.00 | 114.00 | 77573.12 | 7.00 | 86.00 | 90239.64 |
| 6.50 | 100.00 | 85011.94 | 7.00 | 89.00 | 88622.99 | 8.00 | 60.00 | 101454.42 |
| 7.00 | 86.00 | 90239.64 | 8.00 | 70.00 | 98027.08 | 9.00 | 47.00 | 111687.61 |
| 7.50 | 75.00 | 95710.67 | 9.00 | 51.00 | 108809.33 | 10.00 | 35.00 | 122761.06 |
| 8.00 | 60.00 | 101454.42 | 10.00 | 38.00 | 118943.65 | 11.00 | 29.00 | 133957.05 |
| 8.50 | 55.00 | 106266.00 | 11.00 | 31.00 | 131721.70 | 12.00 | 23.00 | 144303.75 |
| 9.00 | 47.00 | 111687.61 | 12.00 | 24.00 | 141589.92 | 13.00 | 18.00 | 156994.50 |
| 9.50 | 41.00 | 117287.78 | 13.00 | 19.00 | 151944.45 | 14.00 | 17.00 | 168609.98 |
| 10.00 | 35.00 | 122761.06 | 14.00 | 17.00 | 168609.98 | 15.00 | 13.00 | 196112.34 |

For each one of the detection methodologies we have calculated n_{out} , the number of outliers detected in each case, and $y_{(1)}^{out}$, the smallest outlier with respect to its y-value. The relative distance of each sample unit was calculated using robust estimates for location and scale i.e., the sample median and the M.A.D., in order to avoid the masking effect.

In all three methods, the outliers fall on the right hand side of the histogram, which means that as outliers are identified the sample units that have the largest



y-values. Of course, the number of the detected outliers, n_{out} , depends on the constant used for each method, so a choice of a smaller value for the parameter would probably identify as outliers units that have small y-values. The choice of a value for the constant in each method is not obvious. In order to decide about it, one should either have a thorough knowledge of the present data, or use relevant past experience.

Because of the fact that each method depends on a different parameter, the comparison between those methods cannot be straightforward. An interesting conclusion is, though, that because we computed the relative distances using the M.A.D. as a scale estimate and the sample median as a location estimate, the methods Relative Distance and M.A.D. became identical, since both identify as outliers the observations for which $y_i \geq median + c * M.A.D., i \in s$; see Table (5.2) where $c = k$.

One of the issues that come up regarding outlier detection is that none of the methodologies that are currently used take into account the sampling weights. In unequal probability sampling, the distribution of the target variable may be a lot different in the population than in the sample. So, a detected outlier in our sample may not be an outlier in the population. In our case, we tried to get a more accurate picture of the population by inflating the data with the design weights, but the results did not differ much.

Another problem that arises with the methodologies that ignore the sampling weights is that those methods cannot detect influential observations, and thus



are restricted in detecting only sample units that have large y -values. For instance, in our data in Table 5.2, if we perform detection using the M.A.D. method with $k = 15$, the first outlier we encounter is the unit that has a total income of 196112.34 euros. Looking back at our dataset, this sample unit has a design weight of 328.48. Another sample unit that had a total income of 172429.29 euros was, however assigned a weight of 1485.71. Although the second unit is much more influential to the Horvitz-Thompson estimator than the first one, the studied methods fail to detect it as an outlier.

But even if we accept that the sample contains a certain number of outliers which have been detected by any of the methods we described above, the next issue that comes up is what to do with those observations. In the case of srs without replacement, it would be theoretically justified to either winsorize the values of the outliers, or employ a re-weighted estimator. In our case, we could not perform any of those because (a) the sampling design of the SILC is complex and, (b) we had no past knowledge about the empirical distribution function of the data. It might be the case that detection of outliers in complex surveys should be performed as a first step, in the hope that it will lead to clean data without measurement errors. As a second step then treatment of outliers may be performed i.e., the influence of the outliers on the estimate gets reduced, by an estimator that is a little biased, but has smaller variance than the H-T estimator. The results for the rest of the variables were similar and are presented in Appendix A.



5.2 Treatment

Moving on to the treatment of outliers, we show the application of the three methods that are most appropriate to use in a survey with a complex design, such that of the EU-SILC: the robust Horvitz-Thompson estimator proposed by Hultiger (1995), the hybrid of winsorization and weight reduction proposed by Dalen and Tambay (1988) and the robust location M-estimator of Huber (1964). The methodologies we used achieve detection and treatment of outliers at the same time. The results for the robust H-T estimator and Huber's M-estimator for the mean of the variable HY010 are in Table 5.3.

Table 5.3: Robust H-T and Huber M-estimator for HY010

| c | \hat{Y}_{Huber} | n_{out} | \hat{Y}_{RHT} | n_{out} |
|-----|-------------------|-----------|-----------------|-----------|
| 1 | 16849.97 | 4922.00 | 17778.65 | 2144.00 |
| 2 | 18887.93 | 1927.00 | 18213.81 | 843.00 |
| 3 | 19816.72 | 619.00 | 19111.03 | 422.00 |
| 4 | 20207.33 | 224.00 | 19664.77 | 237.00 |
| 5 | 20426.80 | 101.00 | 19990.14 | 147.00 |
| 6 | 20547.80 | 56.00 | 20198.14 | 98.00 |
| 7 | 20623.53 | 32.00 | 20345.72 | 72.00 |
| 8 | 20674.11 | 20.00 | 20455.08 | 52.00 |
| 9 | 20714.64 | 17.00 | 20538.07 | 37.00 |
| 10 | 20746.59 | 13.00 | 20599.60 | 28.00 |

where \hat{Y}_{RHT} is the robust Horvitz-Thompson, \hat{Y}_{Huber} is the Huber M-estimator, $c = 1, \dots, 10$ are the values of the tuning constant and n_{out} are the number of outliers that were detected. The ψ -function used for the computation of \hat{Y}_{RHT} is that of Huber. The standard H-T estimator of the mean \bar{Y} is $\sum_{i \in S} w_i y_i / \sum_{i \in S} w_i$.



The Horvitz-Thompson estimate for the mean of HY010 was 21054.5 euros. \hat{Y}_{RHT} can be expressed as a re-weighted estimator, as it actually changes the weights of outliers by multiplying those weights by a robust weight $u_i < 1$. The number of outliers n_{out} depends on the choice of the tuning constant c which is used for the Huber ψ -function. The results for the other variables can be found in Appendix B.

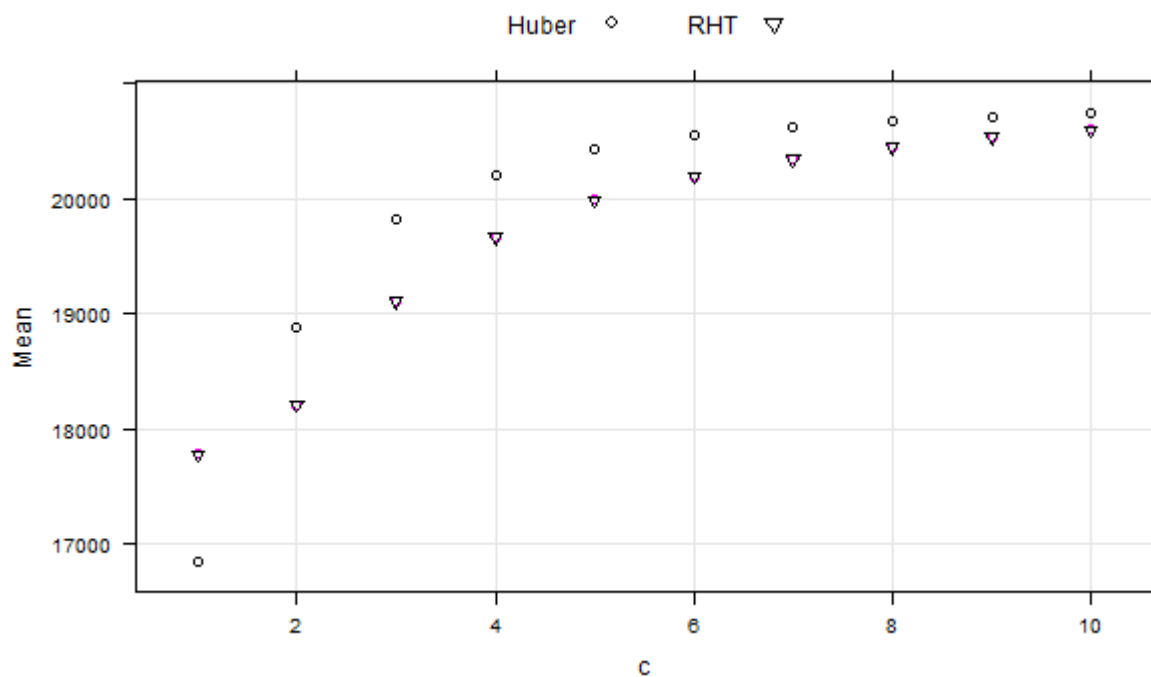
As the value of the tuning constant c decreases, \hat{Y}_{RHT} becomes more robust and more outliers are downweighted. This estimator reduces the weight of the observations that have the largest residuals in the model $y_i = \beta z_i + \varepsilon_i$. In our study, the outliers that were detected in each case were the observations that had also the largest y -values, which seems reasonable given the skewness of the data. As we increase the value of c , \hat{Y}_{RHT} approaches the value of the H-T estimator. Regarding Huber's location estimator, \hat{Y}_{Huber} , we can see that as the value of c decreases, the estimator becomes less sensitive to outliers and we get a more robust estimate for the mean of HY010. The limiting case of \hat{Y}_{Huber} as $c \rightarrow 0$ is the sample median, which is 15579.82, whereas when $c \rightarrow \infty$ the estimate tends towards the unweighted sample mean, which is 20884.46 euros.

In Figure 5.3 we see the increase of \hat{Y}_{Huber} and \hat{Y}_{RHT} as the value of the tuning constant c increases. It is interesting that for most values of c , \hat{Y}_{Huber} is slightly greater than \hat{Y}_{RHT} , that is, \hat{Y}_{RHT} seems to be more resistant to outliers than \hat{Y}_{Huber} . The robust H-T applies the ψ -function on the standardized residuals of the model that is described above, so the difference between the estimates could be explained by the poor fit of the model, since $\hat{\rho}(y, z) = -0.026$.

When we treat multivariate data using weight modification, an important issue is



Figure 5.3: RHT and Huber's M-estimator of HY010



whether we should use a single set of weights for all variables of interest or use a different set of weights for each variable. We computed \hat{Y}_{RHT} for the variables HH070, HY020 and HY022 twice; the first time we used HY010 as a key variable and carried out the estimation for the other variables using the weights for HY010, while the second time each variable had its own set of weights. The results are presented in tables B.1, B.2 and B.3 in appendix B, where the notation \hat{Y}_{RHT}^* and n_{out}^* refers to the estimation using the robust weights for HY010.

When we calculated the robust H-T estimator for the mean, the variable that had in general the most outliers was HY010. So, when we carried out the estimation for

the other variables using the robust weights of HY010, the estimates of \hat{Y}_{RHT} decreased. But despite the large difference between n_{out} and n_{out}^* , $\forall c$, the differences between \hat{Y}_{RHT} and \hat{Y}_{RHT}^* appear to be quite small for medium to large values of c . From another point of view, the two estimators may take approximately equal values for very different number of detected outliers, see Table 5.3.

Moving on to Dalén and Tambay's estimator, we present the results for the mean of HY010 in Table 5.4, where T is the threshold for detecting outliers, \hat{Y}_{DT} is the estimator for the mean and n_{out} the number of detected outliers.

Table 5.4: D-T mean of HY010

| T | \hat{Y}_{DT} | n_{out} |
|----------------|----------------|-----------|
| 10^7 | 17145.23 | 2049.00 |
| $2 \cdot 10^7$ | 19221.46 | 557.00 |
| $3 \cdot 10^7$ | 19925.34 | 218.00 |
| $4 \cdot 10^7$ | 20264.66 | 115.00 |
| $5 \cdot 10^7$ | 20451.50 | 65.00 |
| $6 \cdot 10^7$ | 20563.86 | 41.00 |
| $7 \cdot 10^7$ | 20639.95 | 27.00 |
| $8 \cdot 10^7$ | 20692.12 | 15.00 |
| $9 \cdot 10^7$ | 20730.92 | 13.00 |
| 10^8 | 20765.81 | 12.00 |

Dalén and Tambay's estimator identifies as influential the sample units for which $w_i y_i \geq T$. As we have noted above, this estimator can be considered both a re-weighted and a winsorized estimator. The robust weights w'_i are obtained when we multiply the sampling weights of the outliers by $h_i = (w_i T + w_i y_i - T) / w_i^2 y_i$, with $h_i < 1$. Expressing \hat{Y}_{DT} as a winsorized estimator, the y -values of the outliers are replaced by $y'_i = (w_i T + w_i y_i - T) / w_i^2$. The weights and the y -values of non-



outliers are kept intact.

As the value of the cut-off T decreases, \hat{Y}_{DT} detects more outliers and therefore becomes more robust, whereas when $T \rightarrow \infty$ the estimator approaches the H-T mean. A choice for an appropriate value of T requires past knowledge of the data, so we tested the method using ten different values for the parameter. Since the issue of using one *vs* multiple sets of weights remains open, we calculated \hat{Y}_{DT} twice: in the first try each variable had its own set of robust weights while in the second try we used the robust weights we obtained from HY010 to estimate the mean of the remaining variables. The results can be found in Appendix B, where the notation T^* , \hat{Y}_{DT}^* and n_{out}^* indicate the use of the robust weights of HY010.

The robust weights w_i^* used in the estimation for HY010, which were also used to estimate the mean of the other variables in the second try, were calculated as follows:

$$w_i^* = \begin{cases} w_i, & \text{if } w_i y_i^* < T^* \\ \frac{T^*}{y_i^*} + 1 - \frac{T^*}{w_i y_i^*}, & \text{if } w_i y_i^* \geq T^* \end{cases} \quad (5.1)$$

where T^* is the cut-off that is used in Table 5.4, y_i^* are the y -values of HY010 and w_i the design weights. The number of outliers n_{out}^* is the one that resulted from the application of the method for HY010, and therefore is the same across tables 5.4, B.4, B.5 and B.6. For the variables HY020 and HY022 the estimate \hat{Y}_{DT}^* seems to be slightly greater than \hat{Y}_{DT} and increases as the value of T^* increases. The differences between \hat{Y}_{DT}^* and \hat{Y}_{DT} cannot be explained directly because a different number of observations is downweighted in each try. Moreover, some sample units that are not detected as outliers for variable HY010 (therefore are not down-



weighted) may have large y -values for the rest of the variables. However, for cases where $n_{out} = n_{out}^*$, eg. Table B.4 where $n_{out} = n_{out}^* = 27$ and Table B.6 where $n_{out} = n_{out}^* = 115$, the difference between \hat{Y}_{DT}^* and \hat{Y}_{DT} is almost negligible and that is reassuring. We expect Dalén and Tambay's estimator to perform better than \hat{Y}_{RHT} in terms of variance, because of the poor fit of the model that is used to compute \hat{Y}_{RHT} .

In order to compare the efficiency of each of \hat{Y}_{RHT} and \hat{Y}_{DT} to the efficiency of the H-T estimator \hat{Y} , we used the general relation $V_{\mathcal{P}}(\hat{Y}) = \text{deff } V_{srs}(\hat{Y})$, where $V_{\mathcal{P}}$ is the variance of \hat{Y} under the design \mathcal{P} and V_{srs} is the variance of \hat{Y} under srs, and deff is the design effect. An estimator of V_{srs} based on data from a complex design and with design weights $\{w_i\}$ is

$$\hat{V}_{srs} = \left(1 - \frac{n}{N}\right) \frac{1}{n} \left[\frac{1}{N-1} \sum_{i \in s} w_i \left(y_i - \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} \right)^2 \right], \quad (5.2)$$

see Gambino (2009), and with treated weights $\{w_i^t\}$ is

$$\hat{V}_{srs}^t = \left(1 - \frac{n}{N}\right) \frac{1}{n} \left[\frac{1}{N-1} \sum_{i \in s} w_i^t \left(y_i - \frac{\sum_{i \in s} w_i^t y_i}{\sum_{i \in s} w_i^t} \right)^2 \right]. \quad (5.3)$$

Then,

$$\frac{\hat{V}(\hat{Y}^t)}{\hat{V}(\hat{Y})} = \frac{\text{deff } \hat{V}_{srs}^t}{\text{deff } \hat{V}_{srs}} = \frac{\sum_{i \in s} w_i^t \left(y_i - \frac{\sum_{i \in s} w_i^t y_i}{\sum_{i \in s} w_i^t} \right)^2}{\sum_{i \in s} w_i \left(y_i - \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} \right)^2}, \quad (5.4)$$

where \hat{Y}^t is the estimator with treated weights. As an efficiency measure we define the relative variance, $RV = (\hat{V}[\hat{Y}^t]/\hat{V}[\hat{Y}] - 1) \cdot 100$. A value of $RV < 0$ indicates the gains in efficiency when using a treated estimator relative to the efficiency of

the H-T estimator. The robust weights used in the calculation of \hat{Y}_{RHT} and \hat{Y}_{DT} are placed on the numerator of (5.4), in order to compare the efficiency of those two estimators to the efficiency of \hat{Y} . The results for the variable HY010 are in Table 5.5, where RV_{RHT} and RV_{DT} denote the relative variances of \hat{Y}_{RHT} and \hat{Y}_{DT} respectively.

Table 5.5: RV for HY010

| c | RV_{RHT} | T | RV_{DT} |
|-----|------------|----------------|-----------|
| 1 | -12.42 | 10^7 | -69.05 |
| 2 | -7.80 | $2 \cdot 10^7$ | -51.86 |
| 3 | -0.55 | $3 \cdot 10^7$ | -41.83 |
| 4 | -1.25 | $4 \cdot 10^7$ | -35.06 |
| 5 | -6.89 | $5 \cdot 10^7$ | -30.22 |
| 6 | -1.96 | $6 \cdot 10^7$ | -26.66 |
| 7 | 1.75 | $7 \cdot 10^7$ | -23.77 |
| 8 | -1.85 | $8 \cdot 10^7$ | -21.38 |
| 9 | -2.05 | $9 \cdot 10^7$ | -19.35 |
| 10 | -8.72 | 10^8 | -17.50 |

Both treated estimators appear to be more efficient than the H-T estimator, as the relative variances are less than zero in all but one case. We would expect RV_{RHT} to decrease as the value of the tuning constant c increases, however this was not the case here. On the other hand, the relative variances of \hat{Y}_{DT} decrease as T increases, which makes sense, because the estimator becomes less resistant to outliers. Dalén and Tambay's estimator seems to be more efficient than the robust H-T, probably because of the lack of fit of the model that is used to derive \hat{Y}_{RHT} .

The results for the remaining variables are in tables B.7, B.8 and B.9 in Appendix B, where the notation RV_{RHT}^* and RV_{DT}^* stands for the relative variance of the esti-



mators \hat{Y}_{RHT}^* and \hat{Y}_{DT}^* . There seems to be no specific pattern regarding RV_{RHT} and RV_{RHT}^* ; however \hat{Y}_{RHT} attains a maximum in gains of variance for $c = 1$ for every variable, regardless of the set of weights used. \hat{Y}_{DT} performed better than the other estimators. The relative variance of \hat{Y}_{DT}^* is slightly smaller than that of \hat{Y}_{DT} yet negative in all of the cases, which indicates that the use of one set of robust weights could still result in estimates that are more accurate than those of the H-T estimator.

The methodologies we examined in this section require the specification of certain parameters, the tuning constant c of \hat{Y}_{RHT} and \hat{Y}_{Huber} , and the cut-off T of \hat{Y}_{DT} . There is always a bias-variance trade-off. For example, a small value for c might produce too many outliers, and hence, an estimator that would have a large bias, but a large value would lead to an unrobust estimator. In our opinion, one should be conservative when choosing the value for the parameter, in order not to end up with an estimator that is too biased. That risk of bias becomes even bigger for variables for which the treated estimates are produced by the robust weights determined by the treatment of outliers for the key variable. Like in our study, a small value of c would down-weight too many observations of the key variable and moreover, the estimations for the rest of the variables would drift along. There are methods that can be used to find an optimal value for c by minimizing the *MSE* estimator of the desired estimator but they require additional information, to which we had no access.

EU SILC also produces regression (calibration) estimators of totals of interest.



There exist methods that may produce robust regression estimators, which, however, require knowledge of population totals (obtained from the census), and to which we had no access.

Concluding Remarks

In this thesis, we have reviewed the methods that are used to detect and treat outliers in survey data. The majority of those methods deal with outliers in a univariate setting, although survey data are usually multivariate.

There is a variety of techniques that can be used for the detection of outliers. The relative distance, the M.A.D. and the quartile method are frequently used at the editing stage of a survey, in order to identify measurement errors, the correction of which will lead to a clean dataset. The quartile method is usually preferred over the others, because of its non-parametric nature. For trend outliers, the method used, (e.g., in Eurostat) is the one proposed by Hidioglou and Berthelot (1986). The disadvantage of the existing methodologies is that they do not take into account the sampling weights, let alone the possible correlation between the weights and the variables of interest.

Regarding outlier treatment, the methods used can be mainly split into three categories: winsorization, weight modification and robust estimation. The first two approaches are more appropriate for infinite populations or for srs without replacement. There are winsorized estimators that can be used for finite populations, but they usually require a distributional assumption, which is often not



easy to substantiate. The re-weighted estimators that are presented in Section 4.2 are only applicable when the sampling design is srs. In multivariate surveys and when treatment involves weight modification, one has to decide whether one set of weights should be used for the estimation, or if each variable will have its own vector of robust weights. Robust estimation techniques such M-estimation are often employed because they implicitly identify and treat outliers at the same time. M-estimators are generally biased, but the gains in variance are often substantial.

In our empirical study, we examined the relative merits of some of the methodologies that can be used for detecting and treating outliers in multistage surveys, using data from the EU SILC. The findings are useful in choosing the most appropriate procedure for such survey data.





Appendix A

Detection

Table A.1: Outlier Detection for HH070

| Relative Distance | | | Quartile Method | | | M.A.D. | | |
|-------------------|-----------|-----------------|-----------------|-----------|-----------------|--------|-----------|-----------------|
| c | n_{out} | $y_{(1)}^{out}$ | c_u | n_{out} | $y_{(1)}^{out}$ | k | n_{out} | $y_{(1)}^{out}$ |
| 5.50 | 23.00 | 1115.43 | 5.00 | 75.00 | 843.84 | 6.00 | 18.00 | 1155.60 |
| 6.00 | 18.00 | 1155.60 | 6.00 | 43.00 | 935.99 | 7.00 | 12.00 | 1281.43 |
| 6.50 | 17.00 | 1209.88 | 7.00 | 26.00 | 1036.32 | 8.00 | 8.00 | 1431.29 |
| 7.00 | 12.00 | 1281.43 | 8.00 | 22.00 | 1125.22 | 9.00 | 6.00 | 1703.07 |
| 7.50 | 9.00 | 1377.95 | 9.00 | 16.00 | 1221.74 | 10.00 | 6.00 | 1703.07 |
| 8.00 | 8.00 | 1431.29 | 10.00 | 11.00 | 1323.85 | 11.00 | 3.00 | 1791.97 |
| 8.50 | 7.00 | 1503.68 | 11.00 | 8.00 | 1431.29 | 12.00 | 2.00 | 3221.74 |
| 9.00 | 6.00 | 1703.07 | 12.00 | 7.00 | 1503.68 | 13.00 | 2.00 | 3221.74 |
| 9.50 | 6.00 | 1703.07 | 13.00 | 6.00 | 1703.07 | 14.00 | 2.00 | 3221.74 |
| 10.00 | 6.00 | 1703.07 | 14.00 | 6.00 | 1703.07 | 15.00 | 2.00 | 3221.74 |



Table A.2: Outlier Detection for HY020

| Relative Distance | | | Quartile Method | | | M.A.D. | | |
|-------------------|-----------|-----------------|-----------------|-----------|-----------------|--------|-----------|-----------------|
| c | n_{out} | $y_{(1)}^{out}$ | c_u | n_{out} | $y_{(1)}^{out}$ | k | n_{out} | $y_{(1)}^{out}$ |
| 5.50 | 84.00 | 55228.00 | 5.00 | 147.00 | 46080.00 | 6.00 | 69.00 | 58734.35 |
| 6.00 | 69.00 | 58734.35 | 6.00 | 95.00 | 53100.00 | 7.00 | 48.00 | 66890.00 |
| 6.50 | 55.00 | 63300.00 | 7.00 | 68.00 | 59745.96 | 8.00 | 33.00 | 74200.00 |
| 7.00 | 48.00 | 66890.00 | 8.00 | 49.00 | 66400.00 | 9.00 | 23.00 | 83000.00 |
| 7.50 | 41.00 | 70400.00 | 9.00 | 36.00 | 73500.00 | 10.00 | 15.00 | 90773.00 |
| 8.00 | 33.00 | 74200.00 | 10.00 | 24.00 | 80210.00 | 11.00 | 14.00 | 110034.00 |
| 8.50 | 27.00 | 78500.00 | 11.00 | 16.00 | 86705.00 | 12.00 | 14.00 | 110034.00 |
| 9.00 | 23.00 | 83000.00 | 12.00 | 14.00 | 110034.00 | 13.00 | 13.00 | 115630.00 |
| 9.50 | 18.00 | 85860.00 | 13.00 | 14.00 | 110034.00 | 14.00 | 11.00 | 122898.00 |
| 10.00 | 15.00 | 90773.00 | 14.00 | 14.00 | 110034.00 | 15.00 | 8.00 | 137750.00 |

Table A.3: Outlier Detection for HY022

| Relative Distance | | | Quartile Method | | | M.A.D. | | |
|-------------------|-----------|-----------------|-----------------|-----------|-----------------|--------|-----------|-----------------|
| c | n_{out} | $y_{(1)}^{out}$ | c_u | n_{out} | $y_{(1)}^{out}$ | k | n_{out} | $y_{(1)}^{out}$ |
| 5.50 | 79.00 | 55370.00 | 5.00 | 151.00 | 45400.00 | 6.00 | 66.00 | 59745.96 |
| 6.00 | 66.00 | 59745.96 | 6.00 | 94.00 | 52190.00 | 7.00 | 47.00 | 66890.00 |
| 6.50 | 54.00 | 63300.00 | 7.00 | 66.00 | 59745.96 | 8.00 | 30.00 | 75050.00 |
| 7.00 | 47.00 | 66890.00 | 8.00 | 49.00 | 65700.00 | 9.00 | 22.00 | 83000.00 |
| 7.50 | 38.00 | 71450.00 | 9.00 | 35.00 | 73500.00 | 10.00 | 15.00 | 90773.00 |
| 8.00 | 30.00 | 75050.00 | 10.00 | 24.00 | 78800.00 | 11.00 | 14.00 | 110034.00 |
| 8.50 | 25.00 | 78650.00 | 11.00 | 18.00 | 85860.00 | 12.00 | 14.00 | 110034.00 |
| 9.00 | 22.00 | 83000.00 | 12.00 | 14.00 | 110034.00 | 13.00 | 13.00 | 115630.00 |
| 9.50 | 16.00 | 86705.00 | 13.00 | 14.00 | 110034.00 | 14.00 | 11.00 | 122898.00 |
| 10.00 | 15.00 | 90773.00 | 14.00 | 14.00 | 110034.00 | 15.00 | 8.00 | 137750.00 |



Appendix B

Treatment

Table B.1: Robust H-T and Huber M-estimator for HH070

| c | \hat{Y}_{Huber} | n_{out} | \hat{Y}_{RHT} | n_{out} | \hat{Y}_{RHT}^* | n_{out}^* |
|-----|-------------------|-----------|-----------------|-----------|-------------------|-------------|
| 1 | 380.19 | 7409.00 | 405.34 | 2074.00 | 381.73 | 2144.00 |
| 2 | 389.00 | 6688.00 | 393.08 | 625.00 | 381.05 | 843.00 |
| 3 | 392.83 | 3519.00 | 391.31 | 229.00 | 386.22 | 422.00 |
| 4 | 394.28 | 1014.00 | 392.81 | 107.00 | 390.09 | 237.00 |
| 5 | 394.90 | 243.00 | 393.56 | 49.00 | 392.20 | 147.00 |
| 6 | 395.28 | 84.00 | 394.87 | 29.00 | 393.37 | 98.00 |
| 7 | 395.51 | 33.00 | 395.72 | 18.00 | 394.16 | 72.00 |
| 8 | 395.66 | 22.00 | 396.36 | 15.00 | 394.76 | 52.00 |
| 9 | 395.78 | 12.00 | 396.84 | 10.00 | 395.23 | 37.00 |
| 10 | 395.85 | 8.00 | 397.14 | 8.00 | 395.60 | 28.00 |

Table B.2: Robust H-T and Huber M-estimator for HY020

| c | \hat{Y}_{Huber} | n_{out} | \hat{Y}_{RHT} | n_{out} | \hat{Y}_{RHT}^* | n_{out}^* |
|-----|-------------------|-----------|-----------------|-----------|-------------------|-------------|
| 1 | 13059.78 | 5566.00 | 14197.79 | 2126.00 | 13591.97 | 2144.00 |
| 2 | 14222.55 | 2210.00 | 13891.52 | 703.00 | 13680.30 | 843.00 |
| 3 | 14743.49 | 712.00 | 14406.03 | 314.00 | 14218.59 | 422.00 |
| 4 | 14945.70 | 230.00 | 14711.56 | 161.00 | 14548.01 | 237.00 |
| 5 | 15044.74 | 91.00 | 14870.87 | 90.00 | 14736.18 | 147.00 |
| 6 | 15098.86 | 49.00 | 14967.09 | 56.00 | 14852.73 | 98.00 |
| 7 | 15126.54 | 27.00 | 15033.34 | 41.00 | 14934.05 | 72.00 |
| 8 | 15150.38 | 14.00 | 15078.53 | 25.00 | 14994.52 | 52.00 |
| 9 | 15173.17 | 14.00 | 15112.45 | 21.00 | 15041.04 | 37.00 |
| 10 | 15189.11 | 12.00 | 15140.64 | 15.00 | 15074.87 | 28.00 |

Table B.3: Robust H-T and Huber M-estimator for HY022

| c | \hat{Y}_{Huber} | n_{out} | \hat{Y}_{RHT} | n_{out} | \hat{Y}_{RHT}^* | n_{out}^* |
|-----|-------------------|-----------|-----------------|-----------|-------------------|-------------|
| 1 | 12444.74 | 5253.00 | 13867.91 | 2190.00 | 12893.45 | 2144.00 |
| 2 | 13568.51 | 2024.00 | 13219.43 | 694.00 | 13011.94 | 843.00 |
| 3 | 14071.94 | 653.00 | 13739.86 | 319.00 | 13550.43 | 422.00 |
| 4 | 14271.66 | 209.00 | 14042.62 | 160.00 | 13876.45 | 237.00 |
| 5 | 14369.56 | 83.00 | 14197.01 | 89.00 | 14063.39 | 147.00 |
| 6 | 14422.91 | 47.00 | 14292.39 | 56.00 | 14179.38 | 98.00 |
| 7 | 14450.74 | 23.00 | 14358.12 | 42.00 | 14261.40 | 72.00 |
| 8 | 14474.74 | 14.00 | 14404.69 | 26.00 | 14322.30 | 52.00 |
| 9 | 14497.69 | 14.00 | 14439.06 | 21.00 | 14369.08 | 37.00 |
| 10 | 14513.70 | 11.00 | 14467.84 | 16.00 | 14403.05 | 28.00 |



Table B.4: D-T mean of HH070

| T | \hat{Y}_{DT} | n_{out} | T^* | \hat{Y}_{DT}^* | n_{out}^* |
|----------------|----------------|-----------|----------------|------------------|-------------|
| 10^5 | 360.69 | 6299.00 | 10^7 | 376.96 | 2049.00 |
| $2 \cdot 10^5$ | 383.59 | 1713.00 | $2 \cdot 10^7$ | 388.71 | 557.00 |
| $3 \cdot 10^5$ | 391.61 | 399.00 | $3 \cdot 10^7$ | 392.43 | 218.00 |
| $4 \cdot 10^5$ | 393.97 | 172.00 | $4 \cdot 10^7$ | 394.18 | 115.00 |
| $5 \cdot 10^5$ | 395.16 | 94.00 | $5 \cdot 10^7$ | 395.11 | 65.00 |
| $6 \cdot 10^5$ | 395.83 | 64.00 | $6 \cdot 10^7$ | 395.67 | 41.00 |
| $7 \cdot 10^5$ | 396.33 | 50.00 | $7 \cdot 10^7$ | 396.07 | 27.00 |
| $8 \cdot 10^5$ | 396.69 | 34.00 | $8 \cdot 10^7$ | 396.36 | 15.00 |
| $9 \cdot 10^5$ | 396.94 | 27.00 | $9 \cdot 10^7$ | 396.60 | 13.00 |
| 10^6 | 397.14 | 20.00 | 10^8 | 396.81 | 12.00 |

Table B.5: D-T mean of HY020

| T | \hat{Y}_{DT} | n_{out} | T^* | \hat{Y}_{DT}^* | n_{out}^* |
|-----------------|----------------|-----------|----------------|------------------|-------------|
| $5 \cdot 10^6$ | 12131.16 | 3659.00 | 10^7 | 13018.08 | 2049.00 |
| 10^7 | 13801.08 | 1199.00 | $2 \cdot 10^7$ | 14300.19 | 557.00 |
| $15 \cdot 10^6$ | 14458.14 | 455.00 | $3 \cdot 10^7$ | 14705.79 | 218.00 |
| $2 \cdot 10^7$ | 14748.74 | 206.00 | $4 \cdot 10^7$ | 14894.77 | 115.00 |
| $25 \cdot 10^6$ | 14900.73 | 122.00 | $5 \cdot 10^7$ | 14998.24 | 65.00 |
| $3 \cdot 10^7$ | 14991.92 | 79.00 | $6 \cdot 10^7$ | 15060.06 | 41.00 |
| $35 \cdot 10^6$ | 15052.47 | 48.00 | $7 \cdot 10^7$ | 15102.20 | 27.00 |
| $4 \cdot 10^7$ | 15090.84 | 35.00 | $8 \cdot 10^7$ | 15130.31 | 15.00 |
| $45 \cdot 10^6$ | 15122.41 | 28.00 | $9 \cdot 10^7$ | 15150.08 | 13.00 |
| $5 \cdot 10^7$ | 15147.07 | 21.00 | 10^8 | 15168.02 | 12.00 |



Table B.6: D-T mean of HY022

| T | \hat{Y}_{DT} | n_{out} | T^* | \hat{Y}_{DT}^* | n_{out}^* |
|-----------------|----------------|-----------|----------------|------------------|-------------|
| $5 \cdot 10^6$ | 11487.87 | 3451.00 | 10^7 | 12338.66 | 2049.00 |
| 10^7 | 13151.67 | 1132.00 | $2 \cdot 10^7$ | 13623.70 | 557.00 |
| $15 \cdot 10^6$ | 13796.07 | 433.00 | $3 \cdot 10^7$ | 14029.08 | 218.00 |
| $2 \cdot 10^7$ | 14080.92 | 194.00 | $4 \cdot 10^7$ | 14218.76 | 115.00 |
| $25 \cdot 10^6$ | 14228.34 | 115.00 | $5 \cdot 10^7$ | 14323.61 | 65.00 |
| $3 \cdot 10^7$ | 14318.14 | 77.00 | $6 \cdot 10^7$ | 14386.26 | 41.00 |
| $35 \cdot 10^6$ | 14379.42 | 45.00 | $7 \cdot 10^7$ | 14429.02 | 27.00 |
| $4 \cdot 10^7$ | 14417.56 | 34.00 | $8 \cdot 10^7$ | 14457.53 | 15.00 |
| $45 \cdot 10^6$ | 14449.40 | 28.00 | $9 \cdot 10^7$ | 14477.53 | 13.00 |
| $5 \cdot 10^7$ | 14474.00 | 19.00 | 10^8 | 14495.62 | 12.00 |

Table B.7: RV for HH070

| c | RV_{RHT} | RV_{RHT}^* | T | RV_{DT} | T^* | RV_{DT}^* |
|-----|------------|--------------|----------------|-----------|----------------|-------------|
| 1 | -10.66 | -10.85 | 10^5 | -68.87 | 10^7 | -57.59 |
| 2 | -8.86 | -3.88 | $2 \cdot 10^5$ | -50.46 | $2 \cdot 10^7$ | -45.55 |
| 3 | -3.24 | 0.93 | $3 \cdot 10^5$ | -40.96 | $3 \cdot 10^7$ | -39.69 |
| 4 | -1.88 | 0.30 | $4 \cdot 10^5$ | -36.14 | $4 \cdot 10^7$ | -35.85 |
| 5 | 0.11 | -2.77 | $5 \cdot 10^5$ | -32.94 | $5 \cdot 10^7$ | -32.85 |
| 6 | -0.67 | 1.51 | $6 \cdot 10^5$ | -30.51 | $6 \cdot 10^7$ | -30.41 |
| 7 | -2.27 | 0.18 | $7 \cdot 10^5$ | -28.26 | $7 \cdot 10^7$ | -28.25 |
| 8 | -5.35 | -0.69 | $8 \cdot 10^5$ | -26.30 | $8 \cdot 10^7$ | -26.25 |
| 9 | -4.94 | -2.37 | $9 \cdot 10^5$ | -24.52 | $9 \cdot 10^7$ | -24.39 |
| 10 | -11.56 | -3.06 | 10^6 | -22.87 | 10^8 | -22.61 |



Table B.8: RV for HY020

| c | RV_{RHT} | RV_{RHT}^* | T | RV_{DT} | T^* | RV_{DT}^* |
|-----|------------|--------------|-----------------|-----------|----------------|-------------|
| 1 | -17.65 | -14.60 | $5 \cdot 10^6$ | -70.21 | 10^7 | -62.82 |
| 2 | -8.00 | -7.07 | 10^7 | -52.53 | $2 \cdot 10^7$ | -44.92 |
| 3 | 11.61 | -1.05 | $15 \cdot 10^6$ | -41.23 | $3 \cdot 10^7$ | -35.59 |
| 4 | -6.27 | -1.49 | $2 \cdot 10^7$ | -34.09 | $4 \cdot 10^7$ | -29.63 |
| 5 | -6.31 | -5.85 | $25 \cdot 10^6$ | -29.24 | $5 \cdot 10^7$ | -25.39 |
| 6 | 1.95 | -2.31 | $3 \cdot 10^7$ | -25.61 | $6 \cdot 10^7$ | -22.27 |
| 7 | -5.57 | 0.34 | $35 \cdot 10^6$ | -22.76 | $7 \cdot 10^7$ | -19.72 |
| 8 | -5.87 | -2.67 | $4 \cdot 10^7$ | -20.53 | $8 \cdot 10^7$ | -17.64 |
| 9 | -5.38 | -3.32 | $45 \cdot 10^6$ | -18.52 | $9 \cdot 10^7$ | -16.05 |
| 10 | -3.96 | -7.74 | $5 \cdot 10^7$ | -16.74 | 10^8 | -14.59 |

Table B.9: RV for HY022

| c | RV_{RHT} | RV_{RHT}^* | T | RV_{DT} | T^* | RV_{DT}^* |
|-----|------------|--------------|-----------------|-----------|----------------|-------------|
| 1 | -22.33 | -14.78 | $5 \cdot 10^6$ | -69.44 | 10^7 | -62.46 |
| 2 | -7.32 | -7.22 | 10^7 | -51.85 | $2 \cdot 10^7$ | -44.71 |
| 3 | 6.67 | -1.24 | $15 \cdot 10^6$ | -40.79 | $3 \cdot 10^7$ | -35.48 |
| 4 | -5.75 | -1.58 | $2 \cdot 10^7$ | -33.80 | $4 \cdot 10^7$ | -29.55 |
| 5 | -6.57 | -5.97 | $25 \cdot 10^6$ | -29.07 | $5 \cdot 10^7$ | -25.30 |
| 6 | 2.06 | -2.23 | $3 \cdot 10^7$ | -25.51 | $6 \cdot 10^7$ | -22.18 |
| 7 | -5.10 | 0.25 | $35 \cdot 10^6$ | -22.66 | $7 \cdot 10^7$ | -19.64 |
| 8 | -6.34 | -2.58 | $4 \cdot 10^7$ | -20.44 | $8 \cdot 10^7$ | -17.56 |
| 9 | -5.32 | -3.30 | $45 \cdot 10^6$ | -18.43 | $9 \cdot 10^7$ | -15.97 |
| 10 | -4.04 | -7.72 | $5 \cdot 10^7$ | -16.65 | 10^8 | -14.52 |





Bibliography

Andrews, D. F., P. 1. Bickel, F. R. Hampel, P. 1. Huber, W. H. Rogers. and W. Tukey. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, NJ: Princeton University Press.

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edition. John Wiley and Sons Inc., New-York.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. *Foundations of statistical inference*, V.P. Godambe and D.A. Sprott (Editors), Holt, Rinehart, and Winston, Toronto, 203-233.

Beaumont, J.-F. (2008). A New Approach to Weighting and Inference in Sample Surveys, *Biometrika*, 95, 3, 539-553.

Beaumont, J.-F., Rivest, L. P. (2009). Dealing with Outliers in Survey Data, *Sample Surveys: Design, Methods and Applications*, Vol. 29A.

Belcher, R. (2003). Application of The Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys, *SSC Annual Meeting, June 2003*

Bershad, M.A. (1960). Some observations on outliers, unpublished report, Washington, DC: US Bureau of the Census.



Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Bollen, K. A., and Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases, *Modern Methods of Data Analysis* (pp. 257-91).

Bruce, A. G. (1991). Robust Estimation and Diagnostics for Repeated Sample Surveys, *Mathematical Statistics Working Paper 1991/1*, Wellington, Statistics New Zealand.

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

Chinnappa. B. N. (1976). A Preliminary Note on Methods of Dealing with Unusually Large Units in Sampling from Skewed Populations. unpublished report, Ottawa: Statistics Canada.

Cook, R. D., Weisberg, S. (1982). Residuals and influence in regression, New York, NY: Chapman & Hall

Dalén, J. (1987). Practical Estimators of a Population Total Which Reduce the Impact of Large Observations. *R & D Report*. Stockholm: Statistics Sweden.

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

Eurostat. (2012). Working Group meeting Statistics on Living Conditions, Doc 65-operation 2013, May 2012.



Filzmoser P., Reimann C., Garrett R.G. (2003). Multivariate outlier detection in exploration geochemistry. Technical report TS 03-5, Department of Statistics, Vienna University of Technology, Austria. Dec. 2003.

Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

Garrett R.G. (1989). The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, Vol. 32, pp. 319-341.

Ghangurde, P. D. (1989a). Outlier Robust Estimation in Finite Population Sampling, unpublished report, Ottawa: Statistics Canada.

Ghangurde, P. D. (1989b). Outliers in Sample Surveys, *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 736-739.

Grubbs, F. (1950). Sample criteria for testing outlying observations, *The Annals of Mathematical Statistics* 21(1), p.27-58

Gwet, J.-P., and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. John Wiley and Sons Inc., New-York.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12, 73-83.



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

Horváth, G. (2014). Presentation and development of outlier treatment in HCSO, United Nations Economic Commission for Europe, Conference of European Statisticians, Work Session on Statistical Data Editing, Paris, France, 28-30 April 2014.

Horvitz, D. G., and D. J. Thompson. (1952). A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, 47, pp. 663-685.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

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

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia, 54-63.

Hulliger, B., Alfons, A., Filzmoser, P., et. al. (2011), Robust Methodology for Laeken Indicators, *Advanced Methodology for European Laeken Indicators*, Deliverable 4.2.

Iglewicz, B., Hoaglin, D. (1993). Volume 16: How to Detect and Handle Outliers, *The ASQC Basic References in Quality Control: Statistical Techniques*



Isaki, C. T., and W. A. Fuller. (1982). Survey Design Under the Regression Superpopulation Model, *Journal of the American Statistical Association*, 77, pp. 89–96.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419-435.

Lee, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 178-202.

Lee, H. (1995). Outliers in business surveys. *Business Survey Methods*, Chapter 26, John Wiley & Sons, Inc., New-York.

Lee, H., and Patak, Z. (1998). Outlier robust generalized regression estimator. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 231-235.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, Series B, 61, 166-186.

Rao, C. R. (1971). Some aspects of statistical inference in problems of sampling for finite populations. *Foundations of Statistical Inference*, eds V. P. Godambe and D. A. Sprott, Toronto, Rinehart & Winston.

Raymond, J. (2007). Survey of Electronic Commerce and Technology: Past, Present and Future Challenges. *Papers presented at the ICES-III*, June 18-21, Montreal, Quebec, Canada.



Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81, 373-383.

Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. *Survey Methodology*, 21, 119-129

Rousseeuw, P. J., Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88 (424): 1273-1283.

Searls, D.T. (1966). An estimator for a population mean which reduces the effect of large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, Virginia, 229-234.

