

Representative outliers



Sabine Krieg and Marc Smeets

Statistical Methods (20114)



Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

© Statistics Netherlands, The Hague/Heerlen, 2011.
Reproduction is permitted. 'Statistics Netherlands' must be quoted as source.

Table of contents

1. Introduction to the theme	4
2. One-sided censored estimators	11
3. Two-sided censored estimators.....	22
4. The SBS method	28
5. Consistent estimates of different levels	38
6. Consistent estimate for different target variables	39
7. Estimate of developments for one publication level.....	40
8. Estimate based on registrations.....	41
9. Estimate of developments based on registrations	42
10. More complex situations.....	43
11. Conclusion	44
12. References.....	47
Appendix: Algorithms	49

1. Introduction to the theme

1.1 General description

1.1.1 Problem definition

This document describes estimation methods that are suitable for data files with representative outliers. An outlier is an extreme observation. Only outliers in the sample are discussed in this document. In the introduction, as an example, we keep in mind the estimation for a population total that is based on a simple random sample. The estimators that are described later in this document are also suitable for more complex sampling designs and other population parameters. The total based on a simple random sample y_1, \dots, y_n with n elements is estimated as

$$\hat{Y}_t = \frac{N}{n} \sum_{j=1}^n y_j,$$

where N is the number of elements in the population. Each element in the sample is thus multiplied by N/n . This factor is called the weight.

If an estimator is used that does not take account of outliers (such as the sample mean or the abovementioned total estimation), then an extreme observation has a large influence on the estimation result. A representative outlier is an outlier for which it is assumed that it has been correctly observed and that additional similar elements can be found in the population. For a general description of representative outliers, see also Smeets (2005).

In practice, even after editing, not all errors will have been eliminated from the data. Therefore, this assumption is not fully satisfied.

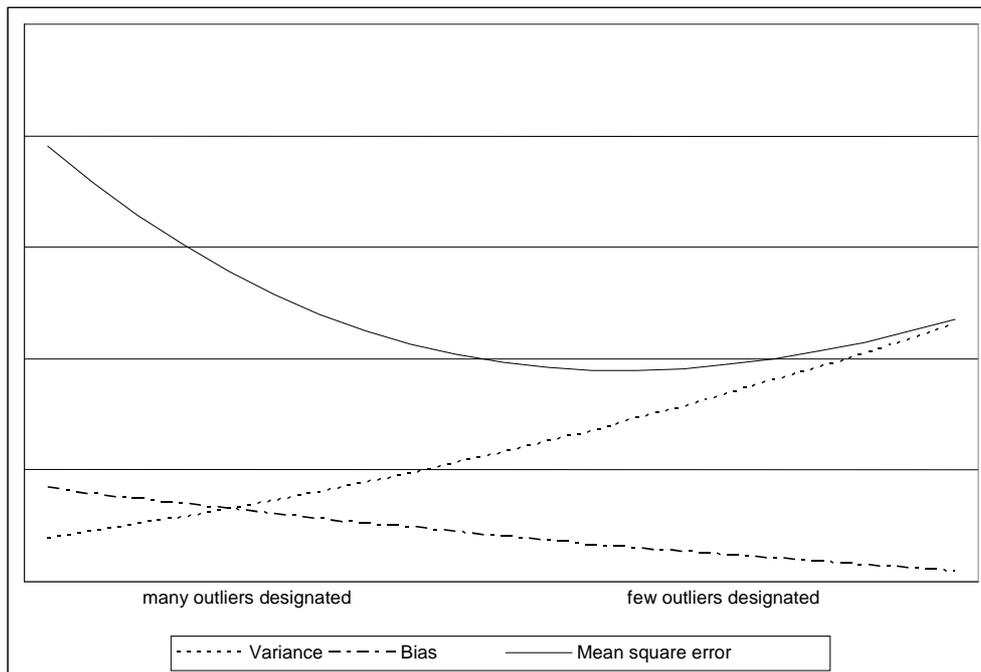
The situation where a single observation has a large influence on the estimation result is often not considered desirable. In this case, the variance of the estimates is often large, which in practice leads to unstable estimates. In other words, the estimates in different periods for the same target variable differ more than expected for the population.

1.1.2 Solution approach

As a solution, the decision is made to reduce the influence of the representative outliers. Consequently, the variance becomes smaller, but bias is introduced in most cases. The treatment of correct outliers therefore boils down to a consideration between the smallest possible variance on the one hand, and the smallest possible bias on the other. The mean square error of an estimator is defined as the expectation of the square of the difference between the estimator and the actual value of the population parameter. This can be calculated as the sum of the variance

and the square of the bias. A small mean square error means that the estimator accurately approximates the actual population parameter. Minimisation of the mean square error is therefore a formalisation of the consideration between a small variance and a small bias. This means that a biased estimator is constructed. The small bias is accepted because the variance decreases. Figure 1 illustrates that the bias is large and the variance is small if many outliers are designated. If only a few or no outliers are designated, the opposite is true. The optimum situation is at the minimum of the mean square error.

Figure 1. Illustration of the effect of outlier treatment on variance, bias and mean square error



In most cases, the more elements designated as outlier (which decreases their influence), the smaller the variance and the larger the bias. For small samples, the loss of accuracy from the bias that arises by designating outliers is relatively small compared with the gain from the smaller variance. For larger samples, the bias weighs heavier than the gain from a smaller variance. For this reason, for larger samples, it is better to designate relatively fewer outliers. This document elaborates various methods to demonstrate how this is specifically put into practice.

If the data are symmetrically distributed, no bias arises from the designation of outliers (at least, if both exceptionally large and exceptionally small values are designated as outliers). In that case, it can be convenient to designate a relatively large number of observations as outlier. This situation occurs in the Structural Business Statistics (SBS) estimator (see Chapter 4).

There are different options to reduce the influence of representative outliers:

- The observation can be ignored.
- The value of the observation can be adapted.
- The weight of the observation can be reduced. Setting the weight to 1 means that the unit only counts for itself and that weighting is done using the other units. The value does not necessarily have to be unique. Even if the value is not unique, or if nothing is known about it, it can be a good solution to set the weight to 1. Note that, based on the sample, it can never be determined whether a value is unique. If the weight is set to 0, then the observation is ignored.

It is often possible to write an estimator in two ways: by adapting values or by adapting weights. The estimation result is then the same in both cases. Chapter 2 contains an elaboration of an estimator that can be used both by adapting the values and by adapting the weights.

Until a number of years ago, the weights or the values were always adapted manually in practice. In recent years, an automatic procedure was introduced for several statistics. In manual outlier detection, the final estimation result depends on the – potentially subjective – considerations of the staff member in question.

1.1.3 Robust estimation

Robust estimators can reduce the influence of outliers. Two simple robust estimators are the α -trimmed mean and the α -winsorised mean. In the α -trimmed mean, the top $\frac{1}{2}\alpha$ percent and the bottom $\frac{1}{2}\alpha$ percent of the observations are ignored. In the α -winsorised mean, the top $\frac{1}{2}\alpha$ percent and the bottom $\frac{1}{2}\alpha$ percent of the observations are adapted. They are given the same values as the largest or smallest non-adapted value respectively. Another well-known but more complex robust estimator is the M-estimator. The three robust estimators mentioned here are not specifically intended for representative outliers and are therefore not further described in this document. For more information about these estimators, see, for example, Huber (1981). In the literature about robust estimation, for which the three abovementioned estimators are intended, it is often assumed that the data are contaminated because of outliers. For example, 90% of the population may be normally distributed with mean μ_1 and variance σ_1^2 , and the other 10% is normally distributed with mean μ_2 and variance σ_2^2 . This 10% is also correct, but is considered as a deviation and not interesting. In robust statistics, we are generally only interested in the non-contaminated part of the data, thus, for example, in μ_1 and σ_1^2 . At Statistics Netherlands, the data are not considered contaminated. In fact, we are interested in

the mean or the total of the entire population, in principle including the exceptionally large or small elements. Another disadvantage of these robust estimators is the dependence of a parameter, for example, α in the α -trimmed mean. The value 5% is often selected for α . However, there is no good way to easily determine the parameter.

These estimators are therefore not suited to representative outliers.

1.1.4 Methods in this document

At Statistics Netherlands, three estimators have been developed that are suitable for representative outliers. These three estimators are described in this document.

- **The one-sided censored estimator**, with the goal of optimally determining the winsorisation in which only the largest (or only the smallest) values are adapted.
- **The two-sided censored estimator**, with the goal of optimally determining the winsorisation in which both the largest and the smallest values are adapted.
- **The SBS method**. In this method, the weights of the observations that deviate the most are set to 1.

Chapter 11 discusses the considerations that play a role when choosing among these methods.

Whether a deviating value has a large influence on the estimation depends not just on the value itself but also on the weight. An exceptional value with a small weight does not have to have a large influence on the estimate.

Whether an observation is deviating and has a large influence on the estimate also depends on what is being estimated. In a sample of all the companies in the Netherlands, a turnover of 10 million euros per year is not exceptional. In the subpopulation of companies with one person employed, this value is indeed exceptional and, depending on the sample size, probably has a strong influence on the estimation result.

Whether an observation is deviating and has a large influence on the estimate also depends on which estimator is used. The company with a turnover of 10 million euros a year has a strong influence on the sample mean of the subpopulation of companies with one staff member. However, if the regression estimator is used with VAT as auxiliary information, and the company also has a large VAT value, then the company's influence is less significant.

The three estimators described in this document are intended for a relatively simple situation: a level estimate must be made for a single population and for a single

target variable, for which a representative sample is selected. The estimators are also suitable (or can be made suitable) for complex sampling designs.

1.1.5 Open problems

However, Statistics Netherlands often has to deal with complex situations. For example, instead of level estimates, often accurate estimates are needed for developments, the goal is to make consistent estimates for different target variables or different target populations, or the estimate is based on a register instead of a representative sample. For these situations (and combinations thereof), no valid methodology has been developed yet. If such a methodology is developed in the future, its description can be added to this document. Chapters 5 to 10 are reserved for this purpose, and additional chapters can be added if necessary.

A provisional and pragmatic solution for these more complex problems is to use one of the methods described in this document. This also occurs in practice, such as in the Structural Business Statistics, where the treatment of outliers focuses on industries and turnover. In this case, the estimates for all of the Netherlands and other aggregation levels and for other target variables are not optimum. The estimates for all of the Netherlands and, for example, the business sectors, are obtained by adding up the estimates of the associated industries. The bias for the estimates per industries is acceptable in relation to the rather large variance. The bias of the estimates for all of the Netherlands and for the business sectors, however, may be unacceptably high in relation to the rather small variance of these estimates. This pragmatic solution therefore does not take sufficient account of different users.

1.2 The concept of the representative outlier

As stated in the previous section, it is assumed for a representative outlier that additional similar elements can be found in the population. The outlier is therefore representative for these elements and that explains the use of the term ‘representative’.

When estimating with representative outliers, the influence of the representative outliers is reduced, for example, by reducing the weight. Reducing the weights has nothing to do with an assumption about how many similar elements are found in the population. It is therefore not assumed that the outlier is not fully representative.

The objective of reducing the influence of representative outliers is to reduce the variance. However, by reducing the influence, bias is introduced. This document elaborates the different ways that the influence of representative outliers can be reduced.

1.3 Scope of the theme and relationship with other themes

As stated above, this document discusses estimation methods for data files with representative outliers. The methods that will be described in this document are intended for situations in which a parameter must be estimated for a single target variable and for a single target population. The methods are suitable if level estimates must be made that are as accurate as possible, and if these estimates are made based on random samples, both for simple and for more complex sampling designs. More research is needed for more complex situations. Space is reserved in Chapters 5 to 10 for this purpose.

Outliers play a role not just in weighting, but also in other phases of the statistical process. During editing, outliers are adapted, where it is determined or assumed that these exceptional observations are incorrect. In imputation, a suitable value is filled in for unit and item non-response. This value is often based on the part of the sample that was observed. This value may have been influenced by outliers, which is not desirable. In addition, outliers can have a disruptive influence in the analysis phase. Finally, outliers can also occur in the auxiliary information and have a disruptive influence.

The methods described in this document are only intended for problems relating to representative outliers in the weighting phase. For techniques that discuss outliers during editing, see Hoogland et al. (2010), and for dealing with outliers during imputation, see Israëls et al. (2007).

1.4 Place in the statistical process

The treatment of representative outliers is part of the weighting process. The process can be interpreted such that the values or weights of outliers are adapted after the data file has been edited and the missing values have been imputed, and just before the sample is weighted. It can also be considered as part of the weighting process. The concept of weighting is discussed briefly in Section 1.1. For more information, see Banning et al. (2010) or Särndal et al. (1992).

1.5 Definitions

Concept	Description
outlier	An outlier is an extreme observation. This document only discusses outliers in the sample.
left outlier	A left outlier is an extremely low observation.
right outlier	A right outlier is an extremely large observation.
representative outlier	A representative outlier is an outlier in the sample, for which it is assumed that it has been correctly observed and that additional similar elements can be found in the population.

non-representative outlier	A non-representative outlier is an outlier in the sample, which has not been correctly observed or is unique in the population.
α -trimmed mean	The α -trimmed mean is the mean of a number of observations, for which the top $\frac{1}{2}\alpha$ percent and the bottom $\frac{1}{2}\alpha$ percent of the observations are ignored.
α -winsorised mean	The α -winsorised mean is the mean of a number of observations, for which the top $\frac{1}{2}\alpha$ percent and the bottom $\frac{1}{2}\alpha$ percent of the observations are adapted. They are given the same value as the largest or smallest non-adapted value respectively.

1.6 General notation

The following general notation is used:

i : an element (such as a company or a person)

n : the sample size

N : the population size

h : index to indicate a certain stratum

L : the number of strata in the population and sample

y : the target variable

x : the auxiliary variable

y_1, \dots, y_n : the observations of the target variable in the sample

w_1, \dots, w_n : the inclusion weights of the observations in the sample

y_{h1}, \dots, y_{hm_h} : the observations of the target variable in stratum h in the sample

\bar{Y} : estimator for the mean of target variable y (with superscript and subscript to specify the estimator and, if desired, the target population).

2. One-sided censored estimators

2.1 Short description

The censored estimator was developed several years ago at Statistics Netherlands; see Renssen et al. (2004) and Krieg et al. (2004). By applying the censored estimator, the influence of outliers is reduced by means of the winsorisation technique (see section 1.1). That means that the influence is reduced by replacing the value of the outlier by a cut-off value. By reducing the influence of outliers, the variance will decrease, but bias will usually be introduced. For the censored estimator, a formula is explicitly derived for the cut-off value that minimises the mean square error (MSE) and, in this way, finds an optimum between a small variance and a small bias. This is the optimum for estimators with such a cut-off value; it is possible that other types of estimators are better. If nothing is known in advance about the optimum cut-off value, this can be estimated using the sample.

If only outliers with an extremely large or an extremely low value are adapted, i.e. only on one side of the distribution, this is referred to as one-sided censoring. One-sided censoring is therefore useful for an asymmetrical distribution with extreme values on only one side. For two-sided censoring, both left outliers (extremely low values) and right outliers (extremely large values) are adapted, and two cut-off values are determined. The left outliers are replaced by the smallest cut-off value and the right outliers by the largest cut-off value. Two-sided censoring is discussed as a separate method in Chapter 3. In this chapter, we elaborate the one-sided censored estimator, for which the cut-off value is estimated using the sample.

If the population were known, then it would be possible to find a better estimator by calculating the cut-off value from the population. However, this is not a realistic situation. At Statistics Netherlands, many surveys are carried out per month, per quarter or per year. In this case, it can be useful to use data from the past to calculate the cut-off value, especially if the population does not change much. However, it has not been studied how much a population may change so that this remains useful. As long as this has not been determined, this idea is not a valid methodology, and it is not described here for this reason.

2.2 Applicability

Using the one-sided censored estimator can be interesting if extreme values in the observations occur on one side of the distribution. In that case, the one-sided censored estimator is expected to produce better results than an estimator that does nothing with outliers.

We elaborate the one-sided censored estimator for simple random samples and stratified samples, where in each stratum simple random sampling without replacement is applied. In other situations, the methods discussed cannot be used unconditionally, and further adaptations are necessary. To illustrate, variations of the one-sided censored estimator are discussed for several specific situations at the end of section 2.3.

At present, the one-sided censored estimator is used in two Statistics Netherlands surveys: the statistics International Services (*Internationale Diensten*, Krieg et al., 2008) and Building Objects in Preparation (*Bouwobjecten in Voorbereiding*, Smeets, 2008).

2.3 Detailed description

2.3.1 One-sided censored estimator for simple random samples

First, we discuss the one-sided censored estimator for a simple random sample. From a population with N elements, a simple random sample of n elements is drawn without replacement. Suppose that, for target variable y , we want to estimate a population mean, such that the influence of outliers is reduced. For this purpose, the target variable y is observed in the sample for all elements i , where $1 \leq i \leq n$. The observations are sorted such that $y_1 \leq y_2 \leq \dots \leq y_n$. A cut-off value t must then be calculated.

We first explain how censoring works if t is given. If an observation y_i is larger than the cut-off value t , the observation is adapted and replaced by this cut-off value. Next, the population mean is estimated by calculating the sample mean of the adapted observations. Suppose that there are r observations smaller than or equal to cut-off value t . The mean is then estimated by

$$\bar{Y}_t^{cens} = \frac{\sum_{j=1}^r y_j + (n-r)t}{n}, \quad (2.3.1)$$

where $y_j \leq t$ for $j = 1, \dots, r$. There are therefore $n-r$ observations designated as outlier. The r observations smaller than or equal to t are also called non-outliers.

Before formula (2.3.1) can be applied, t must be calculated. The objective is to find a cut-off value that minimises the mean square error of (2.3.1). This is the optimum cut-off value. Renssen et al. (2004) provide a formula for the mean square error as sum of the variance and the square of the bias.

The variance, the bias and the MSE can be written as

$$\text{Var}(\bar{Y}_t^{cens}) = (1-f) \frac{p\sigma_m^2 + pq(\mu_m - t)^2}{n}, \quad (2.3.2)$$

$$\text{Bias}(\bar{Y}_t^{cens}) = -q(\mu_r - t), \quad (2.3.3)$$

$$\text{MSE}(\bar{Y}_t^{\text{cens}}) = (1-f) \frac{p\sigma_m^2 + pq(\mu_m - t)^2}{n} + q^2(\mu_r - t)^2. \quad (2.3.4)$$

Here, q is the fraction of adapted values in the population, p the fraction of non-adapted values in the population, σ_m^2 the population variance of the non-adapted values, μ_m the population mean of the non-adapted values, μ_r the population mean of the adapted values and $f = \frac{n}{N}$ the sample fraction.

Note that, for the variance, an approximation formula is used that does not take account of the factor $1/(N-1)$ for simple random samples without replacement. The factor $1/N$ is used instead of this. In practice, N is very large, such that both factors are almost the same.

The cut-off value that minimises equation (2.3.4) is a solution of the following equation:

$$\frac{p}{n}(1-f)(t - \mu_m) - q(\mu_r - t) = 0. \quad (2.3.5)$$

An estimate for the optimum t is then found by using the parameters based on the sample instead of the population parameters in equation (2.3.5).

This means that an estimate for the optimum cut-off value can be found by solving equation (2.3.5), where $q = \frac{n-r}{n}$ is the fraction of outliers in the sample, $p = \frac{r}{n}$ the fraction of non-outliers in the sample, $\mu_m = \frac{1}{r} \sum_{j=1}^r y_j$ the sample mean of the non-outliers, $\mu_r = \frac{1}{n-r} \sum_{j=r+1}^n y_j$ the sample mean of the outliers and $f = \frac{n}{N}$ the sample fraction.

Solving equation (2.3.5) is not straightforward, because the parameters p , q , μ_m and μ_r also depend on t . The value of t determines which observations are designated as outlier. From this, r and the values of the stated parameters follow. Renssen et al. (2004) demonstrate that there is always a unique solution for equation (2.3.5) and that this is smaller than (or equal to) the largest value y_n . There is therefore always at least one element designated as outlier, except in the extreme case of complete enumeration. In this case, it is not useful to designate outliers; after all, all observations have a weight of 1. In that case, $t = \mu_r$ is the solution for (2.3.5). This can be interpreted as a single element being an outlier, for which, however, the value is not adapted. The interpretation that no outliers are designated is also possible in this case. If t is minimally smaller than y_n , then the censored estimate differs minimally from the estimate in which no outliers are designated. Renssen et al. (2002) have algorithms for samples that are drawn with replacement, for simple random sampling and for stratified samples. In the appendix of this document, these algorithms are adapted to the situation of drawing without replacement, and algorithms are provided for additional situations. Algorithm 1 calculates the optimum cut-off value for simple random samples. In this algorithm, t from equation (2.3.5) is calculated for every possible number of outliers, and it is

then checked whether the correct number of outliers is associated with this t . This is repeated until the unique solution of equation (2.3.5) is found.

This algorithm can be used for a sample size of at least two elements. For very small samples, however, calculating estimates is not recommended at all, because they have a very large variance. This is also true for censored estimators.

By inserting the estimated optimum cut-off value t^* in equation (2.3.1), we find the one-sided censored estimate $\bar{Y}_{t^*}^{cens}$ for the population mean.

As stated before, known population parameters are assumed when deriving the formulas for the censored estimator, but ultimately estimates based on the sample are used. If the optimum cut-off value were calculated based on the population, then the estimator based on this would be more accurate. However, even if the optimum cut-off value is estimated based on the sample, then the estimator is usually more accurate than the estimator without outlier treatment. This was demonstrated by various simulations, such as Krieg and Smeets (2005). In Renssen et al. (2004), approximation formulas for variance and bias are derived. It can also be concluded based on these formulas that, in most cases, the censored estimator is more accurate than the estimator without outlier treatment. An exception is a population with multiple outliers with an approximately equal value, in combination with a large inclusion probability. Other exceptions are possible if the approximation formulas in Renssen et al. (2004) are not sufficiently precise. The approximation formulas may not be sufficiently precise for very small samples, small populations or extreme distributions. In the exceptional cases, the accuracy of the censored estimator is comparable with the accuracy of the estimator without outlier treatment. Extreme artificial situations can be constructed where the accuracy of the censored estimator is much lower than the accuracy of the estimator without outlier treatment.

Writing a censored estimator with weights. The censored estimator can be written as a weighted mean of the observations, where the outliers are given a lower weight and the non-outliers a higher weight:

$$\bar{Y}_{t^*}^{cens} = \frac{1}{n} \left[\sum_{j=1}^r g_m y_j + \sum_{j=r+1}^n g_r y_j \right], \text{ where}$$

$$\begin{cases} g_m &= 1 + \frac{q}{p(1+l)} > 1, \text{ for the non-outliers} \\ g_r &= 1 - \frac{1}{1+l} < 1, \text{ for the outliers} \end{cases} \quad \text{and}$$

$$l = \frac{nq}{(1-f)p} > 0. \quad (2.3.6)$$

It is easy to figure out that the weights add up to n ; in other words $rg_m + (n-r)g_r = n$.

Using a simple example, we show how algorithm 1 works.

Example 2.3.1. The sample size is $n = 12$, the population size is $N = 120$ (so $f = 0.1$), the observations are 1, 2, 3, 4, 4, 4, 5, 5, 6, 9, 20, 25. The sample mean is equal to 7.33. In Table 1, for all of the possible values of r , the associated parameters p , μ_m , q and μ_r are calculated. The value of t in the seventh column is determined using equation (B.1). The values of y_r and y_{r+1} are shown in the last two columns.

Table 1. Example of calculating cut-off value t

r	$n-r$	p	q	μ_m	μ_r	t	y_r	y_{r+1}
11	1	0.92	0.08	5.73	25.00	16.29	20	25
10	2	0.83	0.17	4.30	22.50	17.54	9	20
9	3	0.75	0.25	3.78	18.00	15.39	6	9
8	4	0.67	0.33	3.50	15.00	13.50	5	6
7	5	0.58	0.42	3.29	13.00	12.08	5	5
6	6	0.50	0.50	3.00	11.67	11.06	4	5
5	7	0.42	0.58	2.80	10.57	10.18	4	4
4	8	0.33	0.67	2.50	9.75	9.49	4	4
3	9	0.25	0.75	2.00	9.11	8.94	3	4
2	10	0.17	0.83	1.50	8.50	8.40	2	3
1	11	0.08	0.92	1.00	7.91	7.86	1	2

Algorithm 1 goes through all of the rows of Table 1 one by one, and checks whether the t -value found is between the values of y_r and y_{r+1} . In the second row, two elements have been designated as outlier, and $r = 10$. The cut-off value in this row satisfies the condition and the estimated optimum cut-off value of $t^* = 17.54$ has thus been found. The one-sided censored estimator for the population mean is then equal to $\bar{Y}_{t^*}^{\text{cens}} = 6.506$. Note that the algorithm actually stops at $r = 10$, but the table gives all the possibilities for r , to show that the solution for equation (2.3.5) is unique.

Note: In the example, two elements have been designated as outlier. In practice, in most cases, only a small number of outliers are designated, and often only one. This also applies to larger samples. If more elements are designated as outlier, then the bias becomes increasingly larger. This soon plays a greater role than the gains in the variance. Interested readers are invited to calculate the estimated optimum t and the number of outliers in additional examples.

2.3.2 One-sided censored estimator for stratified samples

In a stratified sampling design, the population is divided into L different strata, which consist of N_1, \dots, N_L elements. The total population size is therefore equal to $N = N_1 + \dots + N_L$. A sample of size n is divided among these strata with samples sizes n_1, \dots, n_L . We assume that a simple random sample without replacement is drawn in each stratum. Let y_{h1}, \dots, y_{hn_h} be the observations of the

target variable y in stratum h such that $y_{h1} \leq y_{h2} \leq \dots \leq y_{hn_h}$ is true. A cut-off value is calculated for each stratum. Let $\mathbf{t} = (t_1, \dots, t_h, \dots, t_L)$ be the vector with the cut-off values to be calculated with t_h the cut-off value of stratum h .

If the cut-off values are given, then the censoring per stratum is done in the same way as for a simple random sample. In each stratum h , the values of the target variable are then compared with the given cut-off value t_h . If an observation is larger than the cut-off value, the observation is adapted and replaced by this cut-off value. Next, the stratum mean of the population is estimated by calculating the sample mean of the adapted observations in stratum h . Say that r_h observations in stratum h are smaller than or equal to cut-off value t_h , then the stratum mean is estimated by

$$\bar{Y}_{t_h}^{cens} = \frac{\sum_{j=1}^{r_h} y_{hj} + (n_h - r_h)t_h}{n_h}, \quad (2.3.7)$$

where $y_{hj} \leq t_h$ for $j = 1, \dots, r_h$ and $h = 1, \dots, L$. In stratum h , $n_h - r_h$ observations have been designated as outlier.

The mean of the entire population is subsequently estimated by

$$\bar{Y}_{\mathbf{t}}^{cens} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_{t_h}^{cens}, \text{ where } \mathbf{t} = (t_1, \dots, t_L). \quad (2.3.8)$$

In the stratified situation, the objective is to find a vector of cut-off values $\mathbf{t}^* = (t_1^*, \dots, t_L^*)$, which leads to a minimum mean square error of the estimator (2.3.8). The stratified estimator is therefore optimum for the mean of the entire population and not for the mean per stratum.

Before formula (2.3.7) can be applied, $\mathbf{t}^* = (t_1^*, \dots, t_L^*)$ must be calculated. The optimum cut-off value can be found by solving a system of equations. This system is derived in Renssen et al. (2004), based on the formula for the mean square error. The derivation is based on population parameters. An estimate for the optimum $\mathbf{t}^* = (t_1^*, \dots, t_L^*)$ is found by using the parameters based on the sample instead of the population parameters in the equations. The system is given by:

$$\begin{cases} \frac{N_1(1-f_1)p_1(t_1 - \mu_{1m})}{n_1} - \sum_{h=1}^L N_h q_h (\mu_{hr} - t_h) = 0 \\ \vdots \\ \frac{N_L(1-f_L)p_L(t_L - \mu_{Ln})}{n_L} - \sum_{h=1}^L N_h q_h (\mu_{hr} - t_h) = 0 \end{cases}. \quad (2.3.9)$$

Here, $q_h = \frac{n_h - r_h}{n_h}$ is the fraction of outliers in stratum h in the sample, $p_h = \frac{r_h}{n_h}$ is the fraction of non-outliers in stratum h in the sample, $f_h = \frac{n_h}{N_h}$ the sample

fraction in stratum h , $\mu_{hm} = (y_{h1} + \dots + y_{hr_h})/r_h$ the sample mean of the non-outliers in stratum h and $\mu_{hr} = (y_{h(r_h+1)} + \dots + y_{hn_h})/(n_h - r_h)$ the sample mean of the outliers in stratum h .

As for simple random samples, an approximation formula is also used for the variance in the equation for the mean square error.

Algorithm 2 in the appendix finds the unique solution for system (2.3.9). In principle, the unique solution of (2.3.9) could be found by systematically calculating all possible combinations of numbers of outliers per stratum. However, these calculations would be very time consuming. For this reason, a starting point is first determined using a transformation of the original data. This combination of outliers is then adapted step by step until the solution is found. The starting point is determined by transforming the original data, which simplifies the problem to simple random selection. In practice, based on this starting point, the unique solution of system (2.3.9) is found within several iterations. To determine the interim steps, the strata must be sorted. The selected order does not have any influence on the ultimate solution, only on the interim calculations.

Equation (2.3.8) with the optimum vector of cut-off values $\mathbf{t}^* = (t_1^*, \dots, t_L^*)$ found using algorithm 2 gives the one-sided stratified censored estimator for the population mean $\bar{Y}_{\mathbf{t}^*}^{cens}$.

Writing a stratified censored estimator with weights. The stratified censored estimator can also be written as a weighted mean of the observations, where the outliers are given a lower weight and the non-outliers a higher weight:

$$\bar{Y}_{\mathbf{t}^*}^{cens} = \frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h} \left\{ \sum_{j=1}^{r_h} g_{hm} y_{hj} + \sum_{j=r_h+1}^{n_h} g_{hr} y_{hj} \right\}, \text{ where}$$

$$\begin{cases} g_{hm} &= 1 + \frac{q_h}{p_h(1+\lambda)} \geq 1, \text{ for the non-outliers} \\ g_{hr} &= 1 - \frac{1}{1+\lambda} < 1, \text{ for the outliers} \end{cases}$$

and

$$\lambda = \sum_{h=1}^L \frac{n_h q_h}{(1-f_h)p_h} > 0. \quad (2.3.10)$$

It is easy to see that the weights in each stratum add up to n_h ; in other words $r_h g_{hm} + (n_h - r_h) g_{hr} = n_h$. If no outliers are found in a certain stratum, then $q_h = 0$ and $g_{hm} = 1$ for that stratum and all the elements count the same in this stratum.

In principle, algorithm 2 can be applied for a minimum sample size of two elements per stratum. However, using a minimum sample size of approximately ten elements per stratum is recommended. Outlier detection can become unstable in very small strata. Sampling theory often recommends a minimum sample size of five to ten elements per stratum.

2.3.3 Variants of the one-sided censored estimator

This subsection discusses several variants of the one-sided censored estimator.

One-sided censoring with left outliers

Section 2.3.2 discusses the one-sided censored estimator, in which only the influence of extremely large observations is reduced. If the data are asymmetrically distributed with outliers on the left side (extremely small observations), a variant of the estimator discussed can be used, in which the observations on the left side are reduced by a cut-off value. The objective of this variant is to minimise the mean square error as a function of a cut-off value s of the estimator

$$\bar{Y}_s^{cens} = \frac{(n - r)s + \sum_{j=n-r+1}^n y_j}{n}. \quad (2.3.11)$$

Here, r is the number of observations larger than or equal to cut-off value s , and $n - r$ observations have been designated as outlier. Furthermore, the method works in the same way as the one-sided censored estimators discussed with a right cut-off value t .

Drawing with unequal inclusion weights

In practice, samples are regularly drawn based on a complex sampling design, where the observations are drawn with unequal inclusion weights. By implementing a simple change, the one-sided censored estimators discussed can take unequal inclusion weights into account. Suppose that the observations y_1, y_2, \dots, y_n are drawn with inclusion weights w_1, w_2, \dots, w_n , such that $w_1 + w_2 + \dots + w_n = N$. The censored estimator will then be used based on the values $z_i = w_i y_i$ instead of y_i . The result of incorporating the inclusion weights in the values of $z_i = w_i y_i$ is that the equations (2.3.1), (2.3.7) and (2.3.9) must be adapted.

Equation (2.3.1) is then replaced by

$$\bar{Y}_t^{cens} = \frac{\sum_{j=1}^r z_j + (n - r)t}{N}, \quad (2.3.12)$$

but equation (2.3.5) and algorithm 1 stay the same. Part of the formulas must be adapted because the weight is incorporated in z_i . Note that the parameters p , q , μ_m and μ_r are calculated using z_i instead of y_i .

In the stratified situation, equation (2.3.7) is replaced by

$$\bar{Y}_{t_h}^{cens} = \frac{\sum_{j=1}^{r_h} z_{hj} + (n_h - r_h)t_h}{N_h}. \quad (2.3.13)$$

The system of equations (2.3.9) is replaced by

$$\begin{cases} (1 - f_1)p_1(t_1 - \mu_{1m}) - \sum_{h=1}^L (n_h - r_h)(\mu_{hr} - t_h) = 0 \\ \vdots \\ (1 - f_L)p_L(t_L - \mu_{Lm}) - \sum_{h=1}^L (n_h - r_h)(\mu_{hr} - t_h) = 0 \end{cases}, \quad (2.3.14)$$

where the parameters are calculated as a function of z_{hj} . The solution of (2.3.14) can be found using algorithm 2, the algorithm that is also used to calculate the solution for (2.3.9). Only the starting solution is calculated in another way: using algorithm 3.

Note: If a sample element y_i is close to the stratum mean, but has a very large inclusion weight w_i , this element may be seen as an outlier. That does not seem logical. In that case, the estimator is still good, only the weight was unnecessarily reduced. In practice, this situation is not very likely.

Use of auxiliary information

If, for each observation y_j , a (column) vector of auxiliary information \mathbf{x}_j is available, and this information is also known at population level, the generalised regression estimator can be used. In this case, the censored estimator is applied on the residuals $e_j = y_j - \mathbf{x}_j^t \hat{\boldsymbol{\beta}}$ instead of on the observations themselves. Consider, for example, the generalised regression estimator, represented by

$$\bar{Y}^{\text{reg}} = \bar{Y} + (\bar{\mathbf{X}}_{pop} - \bar{\mathbf{X}}_{sample})^t \hat{\boldsymbol{\beta}} = \bar{\mathbf{X}}_{pop}^t \hat{\boldsymbol{\beta}} + (\bar{Y} - \bar{\mathbf{X}}_{sample}^t \hat{\boldsymbol{\beta}}) =: \bar{\mathbf{X}}_{pop}^t \hat{\boldsymbol{\beta}} + \bar{E}.$$

where \bar{Y} is the sample mean of the observations y_j , $\bar{\mathbf{X}}_{sample}^t$ and $\bar{\mathbf{X}}_{pop}^t$ are the sample mean and the population mean of the auxiliary information \mathbf{x}_j and \bar{E} is the sample mean of the residuals $e_j = y_j - \mathbf{x}_j^t \hat{\boldsymbol{\beta}}$. The generalised regression estimator is described in more detail in Banning et al. (2010) and Särndal et al. (1992).

In this case, the censored regression estimator is given by

$$\bar{Y}_t^{\text{reg,cens}} = \bar{\mathbf{X}}_{pop}^t \hat{\boldsymbol{\beta}} + \bar{E}_t^{\text{reg,cens}}. \quad (2.3.15)$$

2.4 Example

As an example, we discuss the application of the one-sided censored estimator in International Services (*Internationale Diensten*). A report was also published on this application; see Krieg et al. (2008).

International Services estimate quarterly figures for the import and export of companies for several services, including transport, computer and information services. A large part of the response is from companies which do not report imports or exports for any services. These are the so-called zero observations. Therefore, a single large value in the response will have a large influence on the estimates, which means the application of an outlier method is useful.

A stratified sample is drawn, and the strata are classified according to, among others, the standard industry classification (SIC) and the size class. A telephone survey by BES established that the zero observations often involve measurement errors. To correct for this, the inclusion weights are adapted and different inclusion weights occur within a stratum. For this reason, a decision was made to use the one-sided stratified censored estimator with adaptations for the unequal inclusion weights.

Another adaptation concerns not including the zero observations in outlier detection. Theoretically, these observations should be included, because they are part of the sample. For practical reasons, it was decided not to do this. Explicitly, the complete file with the zero observations, which is very large, is not available in the production process when the outliers are detected and would have to be specially retained for outlier detection.

We will use the first quarter of 2007 as an example. The sample frame consists of 39,523 companies and the sample contains 4108 companies (without erroneous zero observations). The censored estimator is used in the manner described above on 11 services, with each service divided into import and export (import and export are called 'flows'). The estimates are therefore optimum for these 22 target variables, but not for aggregates or other breakdowns. The total trade is estimated to be EUR 5.88 billion. If the influence of the outliers is not reduced, then the estimate is EUR 6.14 billion. The mean of the reduction in the estimates over all flow-service combinations by taking outliers into account is therefore 4.22%. In total, 30 outliers were found. For most of the flow-service combinations, one outlier was found; in some cases, there were two to a maximum of four.

2.5 Quality indicators

Given that the objective of the censored estimators is to minimise the mean square error, the obvious choice is to investigate the mean square error of the censored estimator. Renssen et al. (2004) present an analytic approximation for the estimate of the mean square error, in which the variance is estimated using influence functions. The influence function is a measure of robustness, which describes the influence of a specific observation on the estimate. The variance of the estimator can be approximated by taking the expectation of the square of the influence function (Hampel et al., 1986). The accuracy of this approximation formula depends on the sample size. This is comparable with the accuracy of variance estimates for other estimators, such as for the generalised regression estimator. In Renssen et al. (2004), the influence function is derived for several variants of the one-sided censored estimator, and a formula is derived for the estimation of the mean square error for simple random samples.

The mean square error of the censored estimator can be compared with that of estimators that do not take account of outliers, which are also called direct estimators in this context. Besides looking at the mean square error, it is also advisable to examine the bias of the censored estimator. In practice, this can only be approximated as the difference between the estimates that are obtained using the direct estimator and the censored estimator. It is up to the user to assess whether the bias is acceptable. It should be noted here that this is a very inaccurate approximation of the bias (but the only possible one without external information). Nevertheless, a large approximated bias of the censored estimator indicates a large variance for both the direct estimator and the censored estimator.

Another good way to investigate whether the censored estimators have a lower mean square error than the direct estimators is by conducting a simulation study. In this case, it is important to work with realistic population data to obtain reliable results. Section 4.5 offers more information on the requirements for realistic population data.

3. Two-sided censored estimators

3.1 Short description

A two-sided censored estimator can be used to estimate a population mean, such that the influence of both extremely low and extremely large observations is reduced. The two-sided censored estimator was developed at Statistics Netherlands several years ago; see Renssen et al. (2004) and Krieg et al. (2004). The two-sided censored estimator serves to adapt both the very large and the very small extreme observations. This is done in a manner similar to the one-sided censored estimator. However, the formulas and algorithms are more complex.

3.2 Applicability

Using the two-sided censored estimator can be interesting if outliers occur on both sides of the distribution. Like the one-sided censored estimator, the two-sided censored estimator only designates a small number of observations as outlier. This has also been demonstrated in practice. If the distribution is more or less symmetrical, it may be more advantageous to designate more observations as outlier symmetrically on the left and the right, than is done by the two-sided censored estimator.

We elaborate the two-sided censored estimator for simple random samples, which have been drawn without replacement. For stratified sampling designs, the two-sided censored estimator is more complicated, and there is no proof that the algorithms discussed are guaranteed to converge. In practice, the algorithms usually do converge to an estimated optimum cut-off value. More research is needed concerning the use of two-sided censoring on stratified and other complex sampling designs.

At present, two-sided censoring is not used at Statistics Netherlands.

3.3 Detailed description

3.3.1 *Two-sided censored estimator for simple random samples*

In this subsection, we will discuss the two-sided censored estimator for simple random samples. From a population with N elements, a simple random sample of n elements is drawn without replacement. Suppose that we want to estimate the population mean of target variable y , and we want to reduce the influence of both left and right outliers. For all elements i in the sample, where $1 \leq i \leq n$, the target variable y is observed. The sample elements are sorted in such a way that

$y_1 \leq y_2 \leq \dots \leq y_n$. We are going to calculate two cut-off values, a left cut-off value s and a right cut-off value t .

If both cut-off values s and t are given ($s \leq t$), then the two-sided censoring is done in the following way. An observation y_i larger than the cut-off value t is adapted and replaced by t , and the observations smaller than s are replaced by s . Next, the mean of the population is estimated by calculating the sample mean of the adapted observations. Suppose that there are r non-adapted observations, that there are n_l observations smaller than s and n_r observations larger than t . The mean is then estimated by

$$\bar{Y}_{s,t}^{cens} = \frac{\sum_{j=1}^r y_{n_l+j} + n_l s + n_r t}{n}, \quad (3.3.1)$$

where $s \leq y_{n_l+j} \leq t$ for $j = 1, \dots, r$. There are thus $n - r = n_l + n_r$ observations that have been designated as outlier.

Just like for the one-sided censored estimators, we can choose to minimise the mean square error of (3.3.1) as a function of s and t . But if s and t are taken as equal to the population mean, then both the variance and the bias of (3.3.1) are equal to zero. The best estimate is therefore found by taking both s and t as equal to the population mean. If the population mean were known, then this would indeed be optimum. In practice, however, the population mean is not known, and these optimum cut-off values must be estimated using a sample. In this case, the sample mean is taken for s and t , and the censored estimator corresponds with the sample mean. This estimator does not offer any added value.

Instead, Renssen et al. (2004) propose minimising an adapted expression of the mean square error. In this expression, one term from the formula for the mean square error is omitted. An estimate for the cut-off values s^* and t^* , which are optimum for this adapted expression, can be found by solving the following system of equations:

$$\begin{cases} \frac{1-f}{n} [p(\mu_m - s) + q_r(t - s)] - q_l(s - \mu_l) = 0 \\ \frac{1-f}{n} [p(t - \mu_m) + q_l(t - s)] - q_r(\mu_r - t) = 0 \end{cases}, \quad (3.3.2)$$

where $f = \frac{n}{N}$, $p = \frac{r}{n}$, $q_l = \frac{n_l}{n}$, $q_r = \frac{n_r}{n}$, $\mu_m = \frac{1}{r} \sum_{j=1}^r y_{n_l+j}$, $\mu_l = \frac{1}{n_l} \sum_{j=1}^{n_l} y_j$, $\mu_r = \frac{1}{n_r} \sum_{j=1}^{n_r} y_{n-n_r+j}$.

There is always a unique solution for (3.3.2), and in all cases at least one left outlier and one right outlier is designated (see Krieg et al., 2004). Like for the one-sided censored estimator, the application of the two-sided censored estimator is not useful

in the case of complete enumeration. Formula (3.3.2) would then produce cut-off values that are equal to the smallest and largest observation in the sample.

It was noted above that not the mean square error, but another expression, is minimised. At first, this seems like a stopgap measure. The result, however, is an estimator that is similar to the simultaneous use of the one-sided censored estimator on the left and the right. In other words, it is comparable with censoring on the left, while the result of censoring on the right is already known, and, simultaneously, censoring on the right, while the result of censoring on the left is already known. This estimator does not make any assumptions about symmetry and is therefore suitable for situations where the data are not symmetrically distributed. The influence of outliers is not too strongly reduced, and consequently the bias caused by this remains small, comparable to that of the one-sided censored estimator. If it is known that the data are symmetrically distributed, then this choice is not optimum.

The two-sided censored estimator can be used for a sample size of at least two elements. For very small samples, however, calculating estimates is not recommended at all, because they will have a very large variance. This also applies to the censored estimators.

Algorithm 4 in the appendix finds the solution for (3.3.2). This algorithm systematically tries out all possible combinations of numbers of left outliers and right outliers until the unique solution for (3.3.2) is found. In principle, the order of the attempts is arbitrary. However, it is useful to start with a small number of outliers (such as in step 4 in the algorithm) because then the solution will be found rather quickly.

Equation (3.3.1) with the optimum cut-off values s^* and t^* found using algorithm 4 gives the two-sided censored estimator for the population mean $\bar{Y}_{s^*, t^*}^{\text{cens}}$.

Writing the two-sided censored estimator with weights. Like the one-sided censored estimator, the two-sided censored estimator can be written as a weighted mean, in which the outliers are given a lower weight and the non-outliers a higher weight; in other words

$$\bar{Y}_{s^*, t^*}^{\text{cens}} = \frac{1}{n} \left[\sum_{j=1}^{n_l} g_l y_j + \sum_{j=n_l+1}^{n_l+r} g_m y_j + \sum_{j=n-n_r+1}^n g_r y_j \right], \text{ where}$$

$$\begin{cases} g_l < 1, \text{ for the left outliers} \\ g_m = 1, \text{ for the non-outliers and} \\ g_r < 1, \text{ for the right outliers} \end{cases}$$

$$n_l g_l + n_m g_m + n_r g_r = n. \quad (3.3.3)$$

The exact expressions of g_l , g_m and are given in Smeets and Krieg (2004).

3.3.2 Two-sided censored estimator for other sampling designs

If the two-sided censored estimator is used on other sampling designs, the estimator must be adapted for this purpose. Additional methodological research is needed on these adaptations for the two-sided censored estimator. Krieg et al. (2004) elaborate the two-sided censored estimator for stratified samples without replacement. However, there are still open questions for this estimator. For example, it is not clear whether and under which conditions the system of equations, which must be solved to find the estimated optimum cut-off values s and t , has a unique solution. Furthermore, no algorithm is known that is guaranteed to find the optimum solution. However, the algorithm found works in most cases. If one would wish to apply the two-sided censored estimator in practice, then this algorithm could be programmed with a control to see if it works. For the rare case that this does not work, the algorithm for one-sided censoring can be used, first on the left side and then on the right side. This proposal must first be studied in depth before it can be used in practice.

To date, insufficient research has been conducted to use the two-sided censored estimator in other situations.

3.4 Example

At present, the two-sided censored estimator is not yet used at Statistics Netherlands. The simulation study of Krieg and Smeets (2005) investigates an application of the two-sided censored estimator on the Structural Business Statistics (SBS). In this simulation study, the two-sided stratified censored estimator is used, and the cut-off values s and t are estimated using the sample. Here the generalised regression estimator is used and the censored estimator is used on the residuals. This variant of the two-sided censored estimator must still be studied in more depth before it can be described in the Methods Series. This estimator is compared with the one-sided censored estimator (Chapter 2) and the direct estimator. The direct estimator does not adapt any weights or values of outliers. For the one-sided censored estimator, the stratified one-sided censored estimator is studied, in which the cut-off value t is estimated using the sample.

For the simulation study, a population file was created by combining sample data from the years 2000, 2001 and 2002 and three industries in the catering industry (restaurants, snack bars and cafés). In the simulation, this population file is considered as the population for one industry (the restaurants) from which samples are drawn to make estimates for this industry. 10,000 samples were drawn, and these samples were used to determine the variance, the bias and the mean square error of the different estimators.

In the SBS, the generalised regression estimator was used with VAT as auxiliary information. In the first simulation study, the VAT information was ignored and a

sample of $n = 181$ elements was drawn. The table below shows the results of the simulation.

Table 2. Simulation results of one-sided and two-sided censoring without auxiliary information

Estimator	Bias	Variance	MSE
Direct	0	325	325
One-sided censoring	-4.7	249	271
Two-sided censoring	-3.2	258	268

We see that both the one-sided censored estimator and the two-sided censored estimator are more accurate than the direct estimator, and that the two-sided censored estimator is slightly better than the one-sided censored estimator.

In a second simulation, only the elements were drawn for which VAT information was available, and a sample of $n = 351$ elements was drawn. The table below shows the results of this simulation.

Table 3. Simulation results of one-sided and two-sided censoring with auxiliary information

Estimator	Bias	Variance	MSE
Direct	-0.1	27	27
One-sided censoring	-1.3	25	26
Two-sided censoring	-0.3	24	24

In this situation as well, the censored estimators are slightly more effective than the direct estimator, and the two-sided censored estimator is better than the one-sided censored estimator.

3.5 Quality indicators

For two-sided censored estimators as well, the obvious choice is to study the mean square error. The variance can be estimated by using influence functions (Hampel et al., 1986). The accuracy of this approximation formula depends on the sample size. This is comparable to the accuracy of variance estimations for other estimators, such as for the generalised regression estimator. In Renssen et al. (2004), the influence function is derived for the two-sided censored estimator used on a simple random sample. No influence functions have yet been calculated for other versions of the two-sided censored estimator.

The mean square error of the censored estimator can be compared with that of the direct estimators. Besides looking at the mean square error, the bias of the censored estimator can also be examined. In practice, this can only be approximated as the difference between the estimates that are obtained using the direct estimator and the censored estimator. It is up to the user to determine whether the bias is acceptable. It

should be noted here that a large bias in the censored estimator means a large variance in both the direct estimator and the censored estimator.

Another good way of examining whether the censored estimators have a lower mean square error than the direct estimators is to perform a simulation study. In this case, it is important to work with realistic population data to obtain reliable results. Section 4.5 offers more information about the requirements placed on realistic population data.

4. The SBS method

4.1 Short description

In 2001, an outlier method was developed and implemented for the Structural Business Statistics (SBS) (see Vlag et al., 2001, JVG, 2001 and Nieuwenbroek and Vlag, 2001). The method was developed for the weighting process as used for Structural Business Statistics. We call this method here the SBS method; it is also referred to as the Pieter Vlag method. The weighting cell plays a central role in the weighting process of the SBS. For each weighting cell, a residual is calculated as deviation from the standard for each element in the sample. What the standard is depends on the situation. Section 4.3 describes how the standard is defined.

If the absolute value of a residual is too large (in other words, if an element deviates too strongly from the standard), then the element in the sample is designated as outlier and is given the weight of 1. Whether the absolute value is too large is tested using a cut-off value *cut*: if the absolute value is larger than *cut*, then the element is designated as outlier. The cut-off value *cut* is defined as

$$cut = upper_f + c(upper_f - lower_f), \quad (4.1.1)$$

where $upper_f$ and $lower_f$ are the third and first quartile of the absolute values of the residuals. The value of 2.5 was selected for the parameter c .

The calculations described here are performed for the turnover as the main target variable of the Structural Business Statistics. The adapted weights are then also used for weighting the other target variables.

The SBS method was developed in practice, and pragmatic choices were made. It is not possible to provide a methodological foundation for these pragmatic choices. There is also no guarantee that the choices made are optimum. On the contrary, improvements are sometimes possible. Three possible improvements that have been studied are described in this document.

4.2 Applicability

Using this method, the influence of sample elements that strongly deviate from the other elements is reduced by setting the weight to 1.

A simulation study (see Krieg and Smeets, 2005) demonstrated that this method leads to more accurate estimates than an estimator without outlier treatment. The censored estimator (see Chapters 2 and 3) is also less accurate than the SBS method. This result applies in principle only to the range of the simulation study: the catering industry. Because the structure of the data in the other industries probably does not

deviate strongly, the SBS method is probably also a good method for the other industries of the Structural Business Statistics.

Krieg and Smeets (2005) further demonstrated that the SBS method can be improved on various points. The improvements lie in the choice of the parameter c and the way in which the residuals are calculated (no transformation in the remainder cells and taking account of different inclusion weights in the VAT cells). Even without these improvements, the SBS method leads to reasonably accurate estimates. For this reason, section 4.3 describes both the method currently used and the improved method. Please note that an inappropriate choice of parameter c will lead to inaccurate estimates.

The SBS method results in accurate estimates due to the fact that the residuals are more or less symmetrically distributed for a large part of the weighting cells. As a result, the designation of outliers results in very little bias, but it does produce a much smaller variance. This is the reason why the SBS method is a good method for Structural Business Statistics.

The SBS method concentrates on weighting of the Structural Business Statistics and cannot be used unconditionally in other situations. Changes must be made to the method if used for this purpose. Moreover, without further research, it is not known whether the SBS method will also produce good estimates in other situations. Section 4.5 contains several suggestions about how that can be examined.

With a few small changes, the SBS method is also used for the Short Term Statistics (Manual for the Production of Short Term Statistics (*Handboek productie KS-en*), 2004, situation in May 2009). However, it has never been studied whether the method leads to accurate estimates in this context. The objective of the Short Term Statistics is to measure a development, while the Structural Business Statistics are concerned with level figures. Furthermore, for the Short Term Statistics, it has not been studied whether a possible symmetry in the data also leads to more accurate estimates. The use of the SBS method for the Short Term Statistics is therefore not a valid method.

In the 2007-2008 period, research was conducted into whether the Short Term Statistics can be based on VAT information (see De Wolf and Van Bommel, 2007). An adapted version of the SBS method was used for this purpose as well. Due to developments around the rules of tax declaration, it appears that VAT information is no longer available to a sufficient extent, and this new design can probably not be continued (situation in May 2009).

4.3 Detailed description

4.3.1 Sampling design and weighting of the Structural Business Statistics

As stated above, the SBS method was developed specifically for Structural Business Statistics. The description of the method therefore utilises concepts used in the Structural Business Statistics. Moreover, the SBS method is in line with the sample and the weighting method. For proper understanding of this topic, we here provide a short explanation of the Structural Business Statistics. See Resing et al. (2005) for more information.

The objective of the Structural Business Statistics is to publish annual figures about the Dutch business. The publications are made for industries. An industry consists of the enterprises with several Standard Industrial Classifications (SICs). A stratified sample is selected for each industry, with the stratification based on size class. Complete enumeration is applied in size class 6 and higher. Weighting, and therefore also outlier treatment, focuses on size classes 0 to 5. In some industries, the so-called SBS-plus cells, part of the size classes are not observed, but estimated based on fiscal information. The SBS method is applied there in the same way, but reduced to the size classes that are estimated using a sample.

In addition to figures at industry level, figures are also published for higher aggregates and at a more detailed level. The weights established for the estimates at industry level are used for these estimates. The simulation (Krieg and Smeets, 2005) only focused on industries. It is therefore not known whether the SBS method also leads to accurate estimates for other aggregates. A different parameter c would probably be optimum for other aggregates. Neither has a valid methodology been developed at Statistics Netherlands to compute consistent estimates for different aggregation levels (see also Chapter 5).

For outlier detection, each industry is split into weighting cells. For the most part, the same classification is used for this purpose as for the estimation procedure. Companies for which no VAT information is available are classified in the remainder cells; there is one remainder cell for each size class. Companies for which VAT information is available are classified in VAT cells. Here size classes 0 to 3 form one weighting cell and size classes 4 and 5 constitute the other weighting cell. When classifying the industry into weighting cells, minimum cell filling levels are maintained. If these cannot be satisfied, the weighting cells are combined. The exact method of joining is elaborated, for example, in Nieuwenbroek and Vlag (2001). For weighting, the remainder cells for size classes lower than 4 are subdivided further based on the legal form: Legal Entity (*Rechtspersoon* - RP) and Natural Person (*Natuurlijke persoon* - NP).

In the VAT cells, the generalised regression estimator is used with VAT as auxiliary information. Using the generalised regression estimator, the inclusion weights are

adapted in such a way that the estimated VAT total corresponds exactly with the known population totals for VAT. In addition, the estimated number of enterprises corresponds to the known number of enterprises in the population. The use of the generalised regression estimator corrects for a possibly skewed sample, and the precision of the estimate is improved. For a detailed description of the generalised regression estimator, see Banning et al. (2010) or Särndal et al. (1992).

4.3.2 *Outlier detection in the VAT cells, current method*

In the weighting cell, n elements are observed. The observations y_1, \dots, y_n are given for all sample elements. In addition, auxiliary information (in this case, VAT information) x_1, \dots, x_n is given for all sample elements, along with auxiliary information about the population. The regression line is calculated for these elements. Then elements for which the absolute value of the residual is large, i.e. elements which are far away from the regression line, are designated as outlier. Because the regression line is also sensitive to outliers, a robust regression line is calculated first. Elements with a large absolute value of the residual with respect to this robust regression line are not taken into consideration in the calculation of the regression line in the following step. The algorithm is described in the appendix (see algorithm 5). By performing this algorithm, the target variable y_i is adapted for some elements and outliers are designated. Note that the original value must be retained, because this will be needed later.

Medians and quartiles must be calculated in the algorithms in this section. These are initially only defined if the number of elements is odd (for the median), or if the number of elements plus 1 can be divided by 4 (quartiles). In other cases, these values are calculated by interpolation (see JVG, 2001). In addition, for the robust regression line, the data must be divided into three equally large groups. If the number of elements cannot be divided by three, one of the groups will be slightly larger or smaller than the other two groups. For the exact derivation of the groups, see Vlag et al. (2001).

4.3.3 *Outlier method in the VAT cells, improved method*

The inclusion weights w_1, \dots, w_n are also known for the elements in the sample. These inclusion weights are not all equal to each other within a weighting cell, because the inclusion probabilities per size class differ, and because several size classes are combined in a weighting cell. The influence of an element i in the sample is determined by the size of the target variable y_i , but also by the size of the weight w_i . For this reason, the method described in section 4.3.2 can be improved by using z_1, \dots, z_n instead of y_1, \dots, y_n as input, where $z_i = w_i y_i$. The simulation study by Krieg and Smeets (2005) demonstrated that this improved the accuracy of the estimates.

This improvement was not implemented into production.

4.3.4 *Outlier detection in the remainder cells, current method*

In the weighting cell, n elements are observed. The observations y_1, \dots, y_n are given for all sample elements. Residuals are determined for these elements as well. These are defined as the difference between the value of the element itself and the median of all the elements. The cut-off value is not calculated directly for these residuals, but on transformed values. The algorithm can be found in the appendix (Algorithm 6). Performing this algorithm adapts the target variable y_i for some elements. Note that, here too, the original value must be retained, because it will be needed later.

4.3.5 *Outlier method in the remainder cells, improved method*

The outlier method in the remainder cells can be improved in two ways.

First, the inclusion weights w_1, \dots, w_n are also known for the elements in the remainder cells. The method described in section 4.3.4 can therefore be adapted by using z_1, \dots, z_n instead of y_1, \dots, y_n as input, where $z_i = w_i y_i$. This is an improvement only if the weights in a weighting cell are not all the same. This occurs if different size classes are combined in a weighting cell.

The second improvement implies that the logarithmic transformation is not applied and that no absolute values are calculated. That means that $upper_f$ and $lower_f$ are calculated as the third and first quartile of the residuals ε_i . The residuals are then reduced using

$$\varepsilon'_i = \begin{cases} -cut & \text{if } \varepsilon_i < -cut \\ \varepsilon_i & \text{if } -cut \leq \varepsilon_i \leq cut \\ cut & \text{if } \varepsilon_i > cut. \end{cases} \quad (4.3.1)$$

Finally, the y -values are adapted again using $y'_i = m_r + \varepsilon'_i$, and the i -th element is designated as outlier if $y_i \neq y'_i$.

These improvements were not implemented.

4.3.6 *Minimum cell filling level*

The Structural Business Statistics require a minimum cell filling level for the performance of the outlier algorithm. If there are fewer than five elements in a weighting cell, the algorithm is not performed for this weighting cell, and no outliers are detected.

On the one hand, it is a good decision not to base the estimates of medians, quartiles and regression lines on few elements. On the other hand, however, large

observations can have a large influence in very small samples, and designating such a value as outlier can improve the accuracy of the estimate. In the Structural Business Statistics, it is possible to manually designate outliers (see Section 4.3.8), which takes care of this problem in a pragmatic manner.

4.3.7 Selection of the parameter c

Originally, $c = 1.5$ was selected. In the simulation study by Krieg and Smeets (2005) was found that this value was too small and the suggestion was made at that time to increase this parameter to $c = 2.5$. This recommendation was followed.

The simulation study also demonstrated that the SBS method leads to reasonably accurate estimates if c is selected within a rather broad range. It is therefore not essential to accurately determine c . The proper range, however, depends on the situation (sample size and distribution of the residuals). A too small c can result in a strongly biased estimate. That risk is limited by selecting $c = 2.5$ instead of $c = 1.5$, both for the VAT cells and for the remainder cells.

The selection of $c = 2.5$ is a pragmatic choice. More research could show which parameter is optimum in which industry. However, this is very time consuming.

Finally, it must also be noted that $c = 2.5$ would be a poor choice if the transformation were not performed in the remainder cells (this possible improvement is described in Section 4.3.5). In that case, different parameters must be selected for the remainder cells and for the VAT cells.

4.3.8 Manual outlier detection in the Structural Business Statistics

In addition to automatic outlier detection as described in sections 4.3.2 and 4.3.4, outliers are also manually designated in the Structural Business Statistics. Moreover, in some cases outliers which were automatically identified are manually designated as non-outliers.

In general, it can be stated that considerable restraint should be used when making manual changes to automatic outlier detection. After all, the objective of the automatic system is to optimise the accuracy of the estimates. Manual changes can disrupt this optimisation.

Still, there are two reasons why manual outlier detection can be necessary. First, automatic outlier detection is not used for small cells. Second, outlier detection only focuses on total turnover as the main target variable. However, an extreme value for another target variable can also have a large and disruptive influence, which can only be limited by manual outlier detection.

4.3.9 Adapting the weights of both outliers and non-outliers

After the outliers have been detected, the weights of the elements can be adapted. The weights of the outliers are set to 1 to limit their influence. The weights of the non-outliers are therefore raised slightly so that the weights again add up to the population size. Changes are made per stratum and thus per size class. Note that stratification does not take place based on the availability of VAT information. A stratum can therefore contain both elements from a VAT cell and elements from a remainder cell. Suppose that, in a stratum, m elements were observed in the sample. The inclusion weights w_1, \dots, w_m are now adapted. Correction weights g_1, \dots, g_m are calculated. The sample elements that were designated as outlier according to Sections 4.3.2 (or 4.3.3), 4.3.4 (or 4.3.5) and 4.3.8 are given the correction weight $g_i = 1$. The number of outliers in the stratum is r . Next, the weights of the other elements are adapted using the following formula:

$$g_i = \frac{M - r}{m - r} \quad (4.3.2)$$

where M is the population size of the stratum.

4.3.10 Other improvement options

This document describes three improvement options for the SBS method, which were investigated in the simulation study (Krieg and Smeets, 2005):

- Selecting another parameter value;
- No transformation in the remainder cells;
- Taking account of different inclusion weights.

It is possible that the method can also be improved in other ways; for example, by changes to the formula for the cut-off value (4.1.1). However, this has not been studied.

4.4 Example

The SBS method is used every year in the Structural Business Statistics. The simulation study by Krieg and Smeets (2005) studied the characteristics of the method, and the method was used in different situations. The simulation study also used the one-sided and two-sided censored estimators; see Chapters 2 and 3, and Section 3.4 in particular. This section also describes the design of the simulation.

In the first simulation, the VAT information was ignored and the method described in Sections 4.3.4 and 4.3.5 was used. A stratified sample of $n = 181$ elements was selected. The following table presents the results of the simulation:

Table 4. Simulation results for data without auxiliary information

Estimator	c	Transformation	Bias	Variance	MSE
Direct	N/A	N/A	-0.01	325	325
SBS	1.5	Yes	-7.12	240	290
SBS	1.8	Yes	-4.15	258	275
SBS	2.4	Yes	-1.57	285	287
SBS	2	No	-26.07	169	848
SBS	4	No	-11.84	197	337
SBS	8	No	-5.05	213	239
SBS	12	No	-2.79	239	247

Table 4 shows that the direct estimator is not biased, but the SBS estimator is. At the same time, the variance of the direct estimator is larger than the variance of the SBS estimator. For a smaller parameter c , more observations are designated as outlier, as a result of which the bias is larger and the variance smaller. In this situation, the optimum, the smallest mean square error, for the SBS estimator with transformation lies at $c = 1.8$. If the SBS estimator without transformation is used, then $c = 1.8$ would lead to an unacceptably large bias. Also, where $c = 4$, the reduction of the variance cannot compensate for the increase in the bias. The optimum parameter here lies at $c = 8$. A much larger parameter (for example, $c = 12$) produces acceptable estimates, in contrast to a much smaller parameter (for example, $c = 4$).

In a second simulation, only elements in the population for which VAT information is available were included, and the method described in sections 4.3.2 and 4.3.3 is used. A stratified sample of $n = 351$ elements was selected. Table 5 presents the results of the simulation. The Weights column indicates whether corrections took place for the inclusion weights.

This simulation also demonstrates that the direct estimator is not biased. Independent of the parameter, the SBS estimator is much more accurate than the direct estimator. Only if a very large parameter is selected ($c = 20$, or even larger, without taking account of the weights) then the cut-off value cut is so large that the SBS estimator differs very little from the direct estimator. Note that, for a very small parameter, the bias can also become so large that the SBS estimator is less effective than the direct estimator.

Table 5. Simulation results for data with auxiliary information

Estimator	c	Weights	Bias	Variance	MSE
Direct	N/A	N/A	0.058	27.0	27.0
SBS	2	No	-0.17	2.5	2.5
SBS	4	No	-0.619	3.6	4.0
SBS	20	No	-0.938	14.2	15.0
SBS	0.1	Yes	-0.017	0.7	0.7
SBS	1	Yes	0.115	0.9	0.9
SBS	4	Yes	-0.455	3.1	3.3

It is remarkable that, in the improved method where the weights are taken into account, the bias does not increase with a smaller parameter c . The reason for this is that the data are symmetrically distributed here. In this situation, in principle, a very small parameter c can be selected. However, because it is not known in practice whether the population is actually perfectly symmetrically distributed, a very small c is not recommended.

More simulation results can be found in the simulation study (Krieg and Smeets, 2005).

4.5 Quality indicators

The SBS method was developed as an ad-hoc method for the Structural Business Statistics, and no associated quality indicators were developed. For example, no variance formula was derived for the estimator. In that sense, this is not a method with a proper methodological basis. Because the simulation study demonstrated that the estimator is indeed accurate, it was included nevertheless in the Methods Series.

However, some few points of attention deserve to be examined in the application of the SBS method.

1. Symmetry. If the population data are symmetrically distributed, then the SBS estimator leads to rather accurate estimates, compared to the direct estimator. The distribution of the target variable itself is not always important. For example, if the generalised regression estimator is used, then the distribution of the residuals is important. If a stratified sample is used, then the distribution per stratum is important.

To study whether the population data are symmetrically distributed, the sample data must be the basis. Note that an outlier can disrupt a generally symmetrical picture.

The distribution of the sample data can be studied using a diagram, such as a histogram in SPSS. In addition, the asymmetry S can be calculated, which is defined as follows:

$$S = \frac{m_3}{\sqrt{m_2^3}}, \quad (4.5.1)$$

where

$$m_q = \frac{1}{n} \sum_{i=1}^n (y_i - m_1)^q, \quad q = 2,3, \quad (4.5.2)$$

and

$$m_1 = \frac{1}{n} \sum_{i=1}^n y_i. \quad (4.5.3)$$

For symmetrically distributed data, $S = 0$. Small deviations from this are not a problem, first of all because the asymmetry is estimated based on a sample, and second because the estimator does not have to be poor if the data are not perfectly symmetrically distributed.

2. Number of outliers. If the data are symmetrically distributed, then designating a large number of outliers does not lead to bias, but it often does lead to a reduction of the variance. In that case, the designation of a large number of outliers is acceptable.

However, if the data are asymmetrically distributed, then the designation of outliers does lead to bias, which can cancel out the gains in variance. A general rule of thumb here is that in such a case no more than one outlier should be designated per publication cell. For the SBS, a publication cell (industry) consists of VAT cells and remainder cells. Due to symmetry in the VAT cells, a rather large number of outliers can be designated there, and this rule of thumb applies to the remainder cells. So the point is that no more than one outlier should be designated in all the remainder cells together. The simulation study (Krieg and Smeets, 2005) showed that, for a well-chosen parameter c , less than one outlier is designated on average in the remainder cells. The experiences with the censored estimator also show that, in general, only very few outliers should be designated.

3. Simulation. A good way to determine whether the SBS estimator is an improvement compared to the direct estimator is to perform a simulation study. This also allows the optimum parameter c to be determined. However, such a study is time consuming. Another problem is that realistic population data must be available. An important point of attention in this respect is that the population file must contain far more elements than the intended sample. The use of the sample from one period as the population file, where selection with replacement is used, leads to unrealistic simulation results. An option is to combine the data from several periods and several subpopulations. This is realistic population data if the subpopulations are similar and there are no large changes over time. Note that the combination of several periods may be less useful if a panel is applied, where the same sample elements are approached in each period.

Note. If the SBS method is used in a situation other than the Structural Business Statistics, attention must also be paid to the selection of c . One should not assume that $c = 2.5$ is also a good choice in other situations. The abovementioned points of attention can help in the selection of c .

5. Consistent estimates of different levels

Oftentimes, an estimate must be made not just for one level (for example, all of the Netherlands) but also for various subpopulations. In this context, there is a requirement that the different estimates are consistent; for example, the estimates for the totals for all subpopulations must add up to the estimate for the total for all of the Netherlands. Consistency is not automatically achieved if optimum estimates are made independently of one another for different aggregation levels. After all, fewer outliers must be designated at a higher aggregation level for an optimum estimate.

A valid methodology has not yet been developed for this situation.

This situation occurs in the SBS. Here, the decision was made to focus on the industries. The outliers are designated such that the estimates for the industries are optimum (in other words, optimum given the method). The estimates for the other levels are generated automatically by adding up these estimates, but they are not optimum.

6. Consistent estimate for different target variables

Oftentimes, estimates must be made for different target variables that have a certain relationship with each other. For example, the total turnover is equal to the sum of the foreign and domestic turnover.

A valid methodology has not yet been developed for this situation.

This situation occurs in the SBS. A pragmatic method is used that focuses on the turnover as the main target variable. Companies with an extreme turnover value are designated as outlier for all target variables. Because of this pragmatic approach, there are no problems with consistency between the different target variables. Note that such problems could arise if a separate weight were used for each target variable. The simulation study by Krieg and Smeets (2005) studied several situations to assess the quality of the estimates for several other target variables. This study showed that the mean square error is not larger than for an estimator that does not take account of outliers.

This situation also occurs in International Trade in Services (*Internationale Diensten*). In this context, it was decided to focus on the import and export per service. The outliers are designated such that the estimates per service for import and export are optimum (in other words, optimum given the method). The estimates for other target variables, for example, the total import and the total export, are generated automatically by adding up these elements, but they are not optimum.

As described in Chapter 2, the censored estimator can also be written in the form of adapted weights. These adapted weights can be calculated using turnover as the target variable. These adapted weights can also be used on the other target variables. In this case, the same ad-hoc solution is used as for the SBS.

It is likely that improvements can be made by using an outlier method that focuses on more target variables.

7. Estimate of developments for one publication level

Often, the level of the estimate itself is not as important as the development compared to a previous period. The Short Term Statistics (STS) are a well-known example of this. Different ratios are calculated for this statistic; for example, the ratio between the estimates for the turnover of the current month and the same month of the previous year.

In this situation, bias in the level estimates is less of a problem, for if the level estimates for both periods are biased to the same extent, the estimate for the ratio may be (almost) unbiased.

A valid methodology has not yet been developed for this situation at Statistics Netherlands. The method that is currently used for the Short Term Statistics is a slightly adapted version of the SBS method. However, it has not been investigated whether the method also works well for the Short Term Statistics, and how the parameter should be established.

In Great Britain, recent research was performed into outlier treatment when measuring developments (see Lewis, 2008). This paper suggests that the use of an outlier method for level figures leads to better results than the use of an outlier method focused specifically on developments. It must be investigated whether this conclusion applies in general, or only in the example studied. In addition, it must be determined whether the outlier method for developments used in this study can be improved.

8. Estimate based on registrations

Up to now, this document has been based on a randomly selected sample. However, in an increasing number of cases Statistics Netherlands wants to base its published figures on registrations. The definition of representative outliers in the introduction is based on samples, but this can be generalised to registrations: an outlier is an extreme observation in the sample or in the register. A representative outlier is an outlier for which it is assumed it has been correctly observed, and that additional similar elements can be found in the population.

Registrations often cover the publication in question. In this case, there are no representative outliers in the sense that all extreme observations are known. In some cases, the register does not provide full coverage. An example of this is the VAT registration for monthly figures. Because many companies report their VAT quarterly or annually, the companies with monthly VAT information can be seen as a sample from the total population of companies. However, this is not a randomly selected sample. The sampling theory that is developed for random samples therefore does not apply. Similarly, the outlier methods developed for random samples cannot be used unconditionally in this situation.

However, a valid methodology has not yet been developed for this situation.

9. Estimate of developments based on registrations

At Statistics Netherlands, it has been studied how the Short Term Statistics can be produced based on VAT information, and how outliers should be dealt with in this situation. Once a valid method had been developed, this method can be described here.

It is, however, not clear to which extent adequate VAT registration will be available in the future. Therefore, the implementation of this is still very uncertain.

Like in Chapter 8, when estimating developments based on registrations using a register providing full coverage, there are no problems with representative outliers.

10. More complex situations

Chapters 5 to 8 set out the situations that must be researched in the near future. Chapter 9 offers an example for even more complex situations. Other complex situations or combinations of the situations discussed in Chapters 5 to 8 are also possible. A valid methodology has not yet been developed for these situations either.

11. Conclusion

This document has described three different methods for dealing with representative outliers. These methods all focus on a situation in which:

- A random sample has been selected;
- An estimate must be made for the entire population, or for a subpopulation for which other subpopulations are not examined;
- An estimate must be made for a single target variable;
- A level estimate must be made.

In the Structural Business Statistics, the SBS method leads to much more accurate estimates than the censored estimators. This occurs because the data for the Structural Business Statistics is symmetrically distributed for a large part of the weighting cells. This concerns the VAT cells, for which the residuals with respect to the regression line are symmetrically distributed. For this reason, the SBS method is the preferred method for the Structural Business Statistics.

In other situations where the data are – more or less – symmetrically distributed, the SBS method is preferred. However, it is also possible to consider other methods described in the literature, such as an α -trimmed mean or an M-estimator (see Huber, 1981). These methods are not intended for a situation with representative outliers, but for symmetrically distributed data they may result in good estimates in practice. The methods described in the literature have the advantage of having a known variance formula. No comparative research has been conducted for the SBS method versus the M-estimator or the α -trimmed mean. More research is needed before one of these methods can be implemented.

In a situation with extreme values in the observation on one side of the distribution, one of the censored estimators is preferred. Where in the simulation study for the remainder cells, the improved SBS estimator was slightly more effective than the censored estimators, this result probably not true in general since there a well-chosen parameter for the SBS estimator was used. A less well-chosen parameter for the SBS estimator can cause strongly biased estimates. A pragmatic tool for properly selecting the parameter could be the number of outliers that are designated.

In many situations, including situations with only outliers on one side, the two-sided censored estimator leads to slightly better estimates than the one-sided censored estimator (Table 2). At the same time, the two-sided censored estimator is considerably more complex, especially in complex sampling designs. For that reason, in a situation with outliers only on one side, the one-sided censored estimator is preferred. If the data are asymmetrically distributed, but outliers can

occur on both sides, then the two-sided censored estimator is recommended. In this case, one must stay alert and, in the improbable event that the algorithm does not work, find a pragmatic solution.

Based on the censored estimators, the value of an outlier (or the weight) is adapted carefully, while all outliers are given the weight of 1 in the SBS method. Consequently, the SBS estimates are sometimes instable. After all, a small change in the value of a single element can mean that this element changes from “not an outlier” to “outlier”. This can have a large influence on the estimates.

By using one of the methods from this document, the mean square error of the estimators is reduced slightly. Note that this measure, as indicated by its name, is a mean of all possible samples from the population. In some of these possible samples, the estimates based on a method with outlier treatment and the estimates based on a method without outlier treatment differ approximately the same amount from the real population parameter. For other possible samples, the estimates based on a method with outlier treatment are slightly less accurate, but for still other possible samples, the estimates based on a method with outlier treatment are slightly more or even much more accurate than the estimates based on a method without outlier treatment. The potential gains for some samples, with extreme outliers, can be large. The use of a method with outlier treatment can therefore also be viewed as a type of insurance against poor samples. A lower mean square error indicates that the samples where outlier treatment leads to an improvement are dominant.

In any case, whichever estimator is selected, the estimator must be adapted to the sampling design and the weighting method. In addition, changes that take the data structure into account (for example, if many zeroes are observed) can also be useful.

For more complex situations, where more than one target variable is estimated, where optimum estimates must be made for different aggregation levels, where developments must be estimated, or where estimates are made based on registrations, little research has been done to date at Statistics Netherlands. In these situations, a method described in this document could be used as an ad-hoc approach. In this context, pragmatic solutions would have to be thought up to achieve, for example, consistency between different target variables or between different subpopulations. This is then emphatically an ad-hoc solution without methodological basis. If the solution selected has been properly described and research has shown that it leads to reliable estimates, a decision can be made to also describe this method in this document.

In the current applications of the SBS estimator and the censored estimator, it was decided to select a single target variable and a single group of subpopulations (for example, the industries in the SBS), on which the outlier method is used. For other target variables and other subpopulations (for example an entire sector), the estimates follow automatically from the adapted weights, or the adapted variables.

These other estimates, however, are not optimum, because the outlier method is not optimised for this purpose.

Conducting research in the near future is desirable, so that good methods can also be developed for the more complex situations. Space has been reserved in this document to describe these methods at a later date (Chapters 5 to 10).

In practice, it can be useful to also make manual outlier detection possible in addition to automatic outlier detection. This is useful mainly if the automatic method does not take all the objectives of the study into account, such as when using a relatively simple method for a complex situation, such as the situation with more target variables. In addition, manual outlier detection can also be important in the event of low cell filling levels, because some automatic methods become unstable in this situation. However, the manual designation of outliers must take place with considerable restraint, because otherwise the automatically determined optimum solution may become invalid.

Using an outlier method makes the production process more complex; programming and managing it requires capacity. It is obvious that an outlier method should be built into a process only if disruptive outliers are actually a problem.

Finally, we want to emphasise one more time that the use of an estimator focusing on outliers must be properly studied in advance. Negative effects from the bias introduced as a result of this may sometimes only become evident in the longer term. This occurred in the Short Term Statistics, where various effects, including the outlier treatment, led to an increasing bias, the so-called 'runaway effects' (see Van Delden, 2006a and Van Delden, 2006b).

12. References

- Banning, R., Camstra, A. and Knottnerus, P. (2010), *Theme: Sampling Theory, Subthemes: Sample design and Weighting methods*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2011].
- Delden, A. van (2006a), *Wegloopeffecten bij de groeicijfers van de supermarktomzetten: oorzaken en oplossingsrichtingen*. Statistics Netherlands, Voorburg.
- Delden, A. van (2006b), *Herberekening KS bij BES*. Internal report, Statistics Netherlands, Voorburg.
- Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986), *Robust Statistics: the Approach Based on Influence Functions*. Wiley, New York.
- Handboek productie KS-en (2004). Internal report, Statistics Netherlands, Heerlen and Voorburg.
- Hoogland, J., Loo, M.P.J. van der, Pannekoek, J. and Scholtus, S. (2010), *Data editing: detection and correction of errors*. Methods Series document, Statistic Netherlands, The Hague [English translation from Dutch in 2011].
- Huber, P. J. (1981), *Robust Statistics*. Wiley, New York.
- Israëls, A., Pannekoek, P. and Schulte Nordholt, E. (2007), *Imputation*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2011].
- JVGN (2001), *Specificatie Ophooggewichten bepalen*. Documentation for Impect, project UniEdit, Statistics Netherlands, Heerlen.
- Krieg S., R. Renssen and M.J.E. Smeets (2004), *Enkele uitbreidingen voor de gecensureerde schatter*. Internal report, Statistics Netherlands, Heerlen.
- Krieg, S. and M.J.E. Smeets (2005), *De Impect-schatter en mogelijke verbeteringen*. Internal report, Statistics Netherlands, Heerlen.
- Krieg S., M.J.E. Smeets, R.J. Roos, and C.M. Zwaneveld (2008), *Uitbijters bij Internationale Diensten*. Internal report, Statistics Netherlands, Heerlen.
- Lewis, D. (2008), Winsorisation for Estimation of Change. *Bulletin of the Office for National Statistics* 61, 49-61.
- Nieuwenbroek, N.J. and P.A. Vlag (2003), *Evaluatie ophoogmodule binnen Impect*. Internal report, Statistics Netherlands.
- Renssen, R.H., M.J.E. Smeets, and S. Krieg (2002), *Dealing with representative outliers in survey sampling: Algorithms*. Internal report, Statistics Netherlands, Heerlen.

- Renssen, R.H., M.J.E. Smeets, and S. Krieg (2004), *Dealing with representative outliers in survey sampling: Methodology*. Internal report, Statistics Netherlands, Heerlen.
- Resing, B., R. van der Heijden and D. Debie (2005), *Handboek Impact 1 PS-en*. Internal report, Statistics Netherlands, Heerlen and Voorburg.
- Särndal, C-E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.
- Smeets, M.J.E. and S. Krieg (2004), *Gecensureerde schatters bij de Productiestatistieken*. Internal report, Statistics Netherlands, Heerlen.
- Smeets, M.J.E. (2005), *Startcursus Methodologie – Module 11: Representatieve uitbijters*. Internal report, Statistics Netherlands, Heerlen en Voorburg.
- Smeets, M.J.E. (2008), *Een uitbijtermethode voor de statistiek Bouwobjecten in Voorbereiding*. Internal report, Statistics Netherlands, Heerlen
- Vlag, P., G. Heunen, N. Nieuwenbroek, M. Das and H. Pustjens (2001), *Functionaliteit ophoogtraject: technische specificaties detectie uitbijters*. Statistics Netherlands, Heerlen.
- Wolf, P.P. de and K. van Bommel (2007), *Methode- en procesbeschrijving KS en secundaire bronnen*. Internal report, Statistics Netherlands, Voorburg.

Appendix: Algorithms

Algorithm 1. Estimation of the optimum cut-off value for simple random sampling (one-sided)

Step 1. Designate the largest value y_n as outlier and let $r = n - 1$ be the number of non-outliers.

Step 2. Calculate the parameters as follows: $p = \frac{r}{n}$, $q = 1 - p$,

$$\mu_m = \frac{1}{r} \sum_{j=1}^r y_j \text{ and } \mu_r = \frac{1}{n-r} \sum_{j=r+1}^n y_j .$$

Step 3. Solve equation (2.3.5) for t , using the values calculated in Step 2 for the parameters p , q , μ_m and μ_r . The solution can be written as

$$t = \frac{p(1-f)\mu_m + nq\mu_r}{p(1-f) + nq} . \quad (\text{B.1})$$

Step 4. If $y_r < t \leq y_{r+1}$, then t is the optimum cut-off value. Otherwise, we also designate the largest non-outlier as outlier, lower r by 1 and go back to Step 2.

Algorithm 2. Estimation of the optimum cut-off values for stratified samples (one-sided)

The algorithm assumes that $f_h < 1$ applies to all strata h . Strata for which $f_h = 1$ must be deleted from the data beforehand. No outliers are designated there.

First, a starting value is calculated (Steps 1 to 4). Then the optimum cut-off values are calculated using an iterative process (Steps 5 to 8).

Step 1. First, for each stratum h , use algorithm 1 to separately calculate a cut-off value t_h using the observations y_{h1}, \dots, y_{hn_h} . Next, for each stratum, estimate the stratum mean $\bar{Y}_{t_h}^{\text{cens}}$ using equation (2.3.7).

Step 2. Calculate the transformed values $\tilde{y}_{hi} = \frac{N_h}{n_h} (y_{hi} - \bar{Y}_{t_h}^{\text{cens}})$.

Step 3. Let $u_j = \tilde{y}_{hi}$, where $j = 1, \dots, n$, $i = 1, \dots, n_h$ and $h = 1, \dots, L$. Sort the observations so that $u_1 \leq \dots \leq u_n$. Using algorithm 1, calculate a cut-off value \tilde{t} using the observations u_1, \dots, u_n .

Step 4. For each stratum h , determine the number of transformed values \tilde{y}_{hi} that is less than or equal to \tilde{t} , and call this number r_h . The largest $n_h - r_h$ values y_{hi} in a stratum h are thus (provisionally) designated as outlier.

Step 5. Given r_1, \dots, r_L , calculate for each stratum

$$\mu_{hm} = \frac{1}{r_h} \sum_{i=1}^{r_h} y_{hi}, \quad \mu_{hr} = \begin{cases} \frac{1}{n_h - r_h} \sum_{i=r_h+1}^{n_h} y_{hi} & \text{if } n_h - r_h > 0, \\ 0 & \text{if } n_h - r_h = 0 \end{cases},$$

$$q_h = \frac{n_h - r_h}{n_h} \quad \text{and} \quad p_h = \frac{r_h}{n_h}.$$

Step 6. Solve the system of equations represented by (2.3.9) for $\mathbf{t} = (t_1, \dots, t_L)$, using the values calculated in Step 5 for the parameters p_h , q_h , μ_{hm} and μ_{hr} . The vector of cut-off values can be written as

$$\mathbf{t}^t = \mathbf{J}^{-1}[\mathbf{D}_n^{-1} \mathbf{P} \bar{\mathbf{Y}}_m + \mathbf{l} \mathbf{q}_r^t \bar{\mathbf{Y}}_r]. \quad (\text{B.2})$$

Here, $\mathbf{J} = \mathbf{D}_n^{-1} \mathbf{P} + \mathbf{l} \mathbf{q}_r^t$,

where $\mathbf{D}_n = \text{diag}(n_1, \dots, n_L)$, $\mathbf{P} = \text{diag}(N_1(1-f_1)p_1, \dots, N_L(1-f_L)p_L)$, $\mathbf{q}_r = (N_1q_1, \dots, N_Lq_L)^t$, $\bar{\mathbf{Y}}_m = (\mu_{1m}, \dots, \mu_{Lm})^t$, $\bar{\mathbf{Y}}_r = (\mu_{1r}, \dots, \mu_{Lr})^t$ and $\mathbf{l} = (1, 1, \dots, 1)^t$. $\text{diag}(\dots)$ is a diagonal matrix where the diagonal elements are in parentheses. The superscript t means that the transposed matrix must be calculated.

Step 7. For all strata $1 \leq h \leq L$, calculate the number of observations y_{hi} that is smaller than t_h , and call this number \tilde{r}_h .

Step 8. If $r_h = \tilde{r}_h$ for all h , then the estimated optimum cut-off value $\mathbf{t}^* = (t_1^*, \dots, t_L^*)$ has been found. If not, take the first stratum (in other words, take the stratum with the lowest index h) for which $r_h \neq \tilde{r}_h$, define $h_{\min} = \min\{h : r_h \neq \tilde{r}_h\}$ and calculate for that stratum

$$r_{h_{\min}} = \begin{cases} r_{h_{\min}} + 1 & \text{if } r_{h_{\min}} < \tilde{r}_{h_{\min}} \\ r_{h_{\min}} - 1 & \text{if } r_{h_{\min}} > \tilde{r}_{h_{\min}} \end{cases}. \quad \text{This means that you designate one extra}$$

outlier if $r_{h_{\min}} < \tilde{r}_{h_{\min}}$ and one less outlier milder if $r_{h_{\min}} > \tilde{r}_{h_{\min}}$. Go back to Step 5.

Algorithm 3. Determination of the starting solution for algorithm 2 in the case of unequal inclusion weights

Step 1. First, for each stratum h , use algorithm 1 to separately calculate a cut-off value t_h using the observations z_{h1}, \dots, z_{hm_h} . For each stratum,

$$\text{estimate the stratum mean } \bar{Z}_h^{\text{cens}} = \frac{1}{n_h} \left[\sum_{j=1}^{r_h} z_{hj} + (n_h - r_h)t_h \right].$$

Step 2. Calculate the transformed values $\tilde{y}_{hi} = (z_{hi} - \bar{Z}_h^{\text{cens}})$.

Step 3. Let $u_j = \tilde{y}_{hi}$, where $j = 1, \dots, n$, $i = 1, \dots, n_h$ and $h = 1, \dots, L$. Sort the observations so that $u_1 \leq \dots \leq u_n$. Use algorithm 1 to calculate a cut-off value \tilde{t} using the observations u_1, \dots, u_n .

Step 4. For each stratum h , calculate the number of transformed values \tilde{y}_{hi} that is less than or equal to \tilde{t} , and call that number r_h . The largest $n_h - r_h$ values z_{hi} in a stratum h are thus (provisionally) designated as outlier.

Algorithm 4. Estimation of the optimum cut-off value for simple random samples (two-sided)

Step 1. Let $i = 0$ and $j = 0$.

Step 2. Let $n_l = i + 1$ be the number of left outliers and $n_r = j + 1$ the number of right outliers. If $n_l + n_r > n$, then go to Step 4, and otherwise

subsequently calculate the parameters $r = n - n_l - n_r$, $p = \frac{r}{n}$, $q_l = \frac{n_l}{n}$,

$$q_r = \frac{n_r}{n}, \quad m_l = \frac{1}{n_r} \sum_{k=1}^{n_r} y_{n-n_r+k}, \quad m_m = \frac{1}{r} \sum_{k=1}^r y_{n_l+k} \quad \text{and} \quad m_r = \frac{1}{n_l} \sum_{k=1}^{n_l} y_k.$$

Step 3. Solve the system of equations (3.3.2) for s and t , using the values calculated in Step 2 for the parameters p , q_l , q_r , m_l , m_m and m_r . The cut-off values are represented by

$$\begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} (1-f)(p+q_r)+nq_l & -(1-f)q_r \\ -(1-f)q_l & (1-f)(p+q_l)+nq_r \end{pmatrix}^{-1} \begin{pmatrix} (1-f)p m_m + nq_l m_l \\ (1-f)p m_m + nq_r m_r \end{pmatrix} \quad (\text{B.3})$$

Step 4. If $y_{i+1} \leq s \leq y_{i+2}$ and $y_{n-j-1} \leq t \leq y_{n-j}$, then the solution has been found. Otherwise, we make a distinction between three cases:

- If $j = 0$, then $j = i + 1$, $i = 0$ and go back to Step 2;
- If $j > 0$ and $j > i$, then $i = i + 1$, $j = j$ and go back to Step 2;

- If $j > 0$ and $j \leq i$, then $i = i$, $j = j - 1$ and go back to Step 2.

Algorithm 5. SBS method: Designation of outliers in VAT cells, current method

Step 1: Determination of the robust regression line

The data records are divided into three groups of equal size:

1. The group with the smallest x -values;
2. The group with the middle x -values;
3. The group with the largest x -values.

For the first and the third group, the medians $m_{1,x}$ and $m_{3,x}$ of the x -values and the medians $m_{1,y}$ and $m_{3,y}$ of the y -values are calculated. Then, the robust regression line is represented by $y = \alpha x + \beta$, where

$$\beta = \frac{m_{3,y} - m_{1,y}}{m_{3,x} - m_{1,x}} \quad (\text{B.4})$$

and

$$\alpha = m_{1,y} - \beta m_{1,x}. \quad (\text{B.5})$$

This is the straight line that runs through the points $(m_{1,x}, m_{1,y})$ and $(m_{3,x}, m_{3,y})$.

Step 2: Calculation of the residuals and the cut-off value

The residual is calculated for each element i :

$$\varepsilon_i = y_i - \alpha - \beta x_i \quad (\text{B.6})$$

Then the cut-off value cut is calculated as

$$cut = upper_f + c(upper_f - lower_f) \quad (\text{B.7})$$

where $upper_f$ and $lower_f$ are the third and first quartile of the absolute values of the residuals ε_i . For the selection of c , see section 4.3.7.

Step 3: Limiting of residuals and recalculation of y -values

Now the residuals are limited using the following formula

$$\bar{\varepsilon}_i = \begin{cases} -cut & \text{if } \varepsilon_i < -cut \\ \varepsilon_i & \text{if } -cut \leq \varepsilon_i \leq cut \\ cut & \text{if } \varepsilon_i > cut \end{cases} \quad (\text{B.8})$$

Next, the y -values are adapted:

$$y'_i = \alpha + \beta x_i + \bar{\varepsilon}_i. \quad (\text{B.9})$$

Step 4: Recalculation of regression line

A regression line is recalculated using y'_1, \dots, y'_n . The following formula is used for this purpose:

$$\beta' = \frac{\sum_{i=1}^n (x_i - \bar{x})(y'_i - \bar{y}')}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (\text{B.10})$$

$$\alpha' = \bar{y}' - \beta' \bar{x} \quad (\text{B.11})$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y}' = \frac{1}{n} \sum_{i=1}^n y'_i$ are the means of the x -values and the y' -values. So this is the straight line through the point (\bar{x}, \bar{y}') . The result of this calculation corresponds with the least squares method.

Step 5: Recalculation of the residuals, cut-off value, limitation of residuals and recalculation of y-values

Now the residuals ε'_i are calculated as described in Step 2, where α' , β' and y'_1, \dots, y'_n are now used instead of α , β and y_1, \dots, y_n . Next, cut' is calculated for ε'_i .

After this, the residuals are re-limited and the y -values are recalculated, where now cut' , ε'_i , α' , β' and y'_1, \dots, y'_n are used instead of cut , ε_i , α , β and y_1, \dots, y_n . The residuals limited in this way are called $\bar{\varepsilon}_i$, the y -values y''_1, \dots, y''_n .

Step 6: Recalculation of regression line, residuals, cut-off value, limitation of residuals and recalculation of y-values

Using the formulas from Step 4, the regression line is recalculated, where y''_1, \dots, y''_n is now used instead of y'_1, \dots, y'_n . The regression coefficients are now called α'' and β'' . Next, the calculations from Step 5 are repeated, where α'' and β'' and y''_1, \dots, y''_n are now used. The newly calculated y -values are called y'''_1, \dots, y'''_n .

Step 7: Designation of outliers

If $y_i > y'''_i$, then the i -th element is designated as outlier.

Note: The SBS estimator documentation indicates that the element is designated as outlier only if there is a clear difference (more than 1%). This restriction is made to rule out changes due to rounding.

Algorithm 6. SBS method: Designation of outliers in remainder cells, current method

Step 1: Determination of the median

The median m_r of the elements y_1, \dots, y_n is calculated in the remainder cell.

Step 2: Calculation of the residuals and cut-off value

The residual is calculated for each element i :

$$\varepsilon_i = y_i - m_r. \quad (\text{B.12})$$

These residuals are not used for the calculation of the cut-off value cut ; the transformed residuals τ_i are used instead. These are calculated as

$$\tau_i = \begin{cases} \log_{10}(\text{abs}(\varepsilon_i)) & \text{if } \text{abs}(\varepsilon_i) > 0.00005 \\ 0 & \text{otherwise} \end{cases}. \quad (\text{B.13})$$

Then the cut-off value cut is calculated as

$$cut = upper_f + c(upper_f - lower_f) \quad (\text{B.14})$$

where $upper_f$ and $lower_f$ are the third and first quartile of the transformed residuals τ_i . For the selection of c , see section 4.3.7.

Step 3: Limitation of residuals and recalculation of y-values

Now the residuals are limited using the following formula

$$\varepsilon'_i = \begin{cases} -10^{cut} & \text{if } \tau_i > cut, \varepsilon_i < 0 \\ \varepsilon_i & \text{if } -cut \leq \tau_i \leq cut \\ 10^{cut} & \text{if } \tau_i > cut, \varepsilon_i > 0 \end{cases}. \quad (\text{B.15})$$

The y-values are subsequently adapted:

$$y'_i = m_r + \varepsilon'_i. \quad (\text{B.16})$$

Step 4: Designation of outliers

If $y_i \neq y'_i$, then the i -th element is designated as outlier.

Note 1: In the Structural Business Statistics documentation, the procedure is repeated one more time. The median and the quartiles, however, are robust and will not change after adaptation of the outliers. This is why we do not describe this repetition here.

Note 2: For this situation as well, the SBS documentation indicates that the element is designated as outlier only if there is a clear difference (more than 1%). This restriction is made to rule out changes due to rounding.

Note 3: This method was developed for turnover values. These are generally large positive values. The residuals $\varepsilon_i = y_i - m_r$ are also usually greater than 1 or less than -1. In this case, the transformed residuals τ_i are all greater than zero. If the absolute value of ε_i increases, then so does the transformed residual τ_i . The method works reasonably well under this condition, as demonstrated by the simulation results in section 4.4. It is doubtful whether the transformation also works well if the target variable is in another order of magnitude. As also demonstrated by the simulation study, the preference is to not perform the transformation.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Representatieve uitbijters				
1.0	23-11-2009	First Dutch version	Sabine Krieg Marc Smeets	Koert van Bommel Jeroen Pannekoek
English version: Representative outliers				
1.0E	18-07-2011	First English version	Sabine Krieg Marc Smeets	