Selective Editing Using Contamination Model

Ieva Burakauskaitė (ieva.burakauskaite@stat.gov.lt)
Statistics Lithuania,

Vilma Nekrašaitė-Liegė (vilma.nekrasaite-liege@stat.gov.lt, vilma.nekrasaite-liege@vilniustech.lt)
Statistics Lithuania, Vilnius Gediminas Technical University

ABSTRACT

Results of an outlier detection study with a focus on selective editing are presented in the paper. The aim of selective editing is to identify observations affected by errors that have a major impact on the quality of sample estimates. This way the data editing process can be focused on the corresponding observations therefore allocating excess human resources and reducing time costs though maintaining the quality of sample estimates. These objectives are especially important for national statistical institutions such as Statistics Lithuania seeking to optimize the data editing process.

A few different versions of selective editing were applied to the data editing process of the quarterly statistical survey on service enterprises (turnover indicator) of Statistics Lithuania. Predictions of the target variable were obtained using the contamination model. An impact of a potential error on a sample estimate was evaluated using a score function with a standard structure – a difference between the observed value of the target variable and its prediction multiplied by a sample weight and a suspicion component. Two types of the suspicion component (discrete and continuous) were used and an impact of the suspicion component on the effectiveness of selective editing was investigated. Efficiency of the continuous suspicion component supported its advantage over the discrete suspicion component, and therefore turned out to be a major factor in optimizing the data editing process.

Keywords: *selective editing; contamination model; data validation; statistical survey; official statistics.*

1. Introduction

An appropriate accuracy of sample estimates is one of the most important objectives to be achieved using sampling methods in official statistics. The quality of statistical data together with the sampling strategy (a sampling plan and an estimator) have a major impact on the accuracy of sample estimates. Commonly, an erroneous part of statistical data is unknown and therefore may only be detected by either logical or mathematical methods using some kind of known additional information. Previously the data editing process was usually focused on editing all of the detected errors. However, according to various studies, in order to achieve the desired accuracy of sample estimates, it is unnecessary to edit all errors. The main idea of selective editing is to identify and sort errors according to the influence they have on sample estimates (Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000). For this purpose, a score function is constructed that portrays the influence possibly erroneous observation has on sample estimates (Latouche and Berthelot, 1992; Hedlin, 2003). As the error detection procedure is usually carried out before the calculation of sample estimates, the best versatile selective editing model for identifying only the most influential part of erroneous data has to be chosen in advance.

Although selective editing is already implemented as one of the methods for outlier detection and data validation procedure in Statistics Lithuania, it still remains an important, uncommon topic for research in Lithuania. The currently used selective editing model identifies a high number of outliers as a result of its similarity to the default selective editing model accessible through the package "SeleMix" of the programming language R. In order to adopt a more suitable selective editing model for a specific statistical survey of Statistics Lithuania, an outlier detection study was carried out. Results of the latter study have been briefly presented at Summer School on Survey Statistics 2021 (Burakauskaitė and Nekrašaitė-Liegė, 2021) and The Use of R in Official Statistics (uRos2021) conferences. This paper provides a closer look into two cases of the carried out study, and indicates the main features that have to be taken into consideration while choosing the most suitable selective editing model.

Section 2 of the paper introduces the contamination model and the selective editing method that form a base for the practical study of the outlier detection. Section 3 presents the study that was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. During the study some randomly selected values of statistical data were replaced with errors. The detection of randomly introduced errors was then carried out using a few cases of selective editing. The comparison of results as well as its summary are presented in Section 4. Calculations were carried out using the statistical programming language R and its package "SeleMix" that has been designed to execute the selective editing method (Guarnera and Buglielli, 2013).

2. Methodology on Selective Editing

2.1 Contamination Model

Suppose that true (unobserved) data are independent realizations of p-variate random vectors $\mathbf{Y}_i^* = (\mathbf{Y}_{i1}^*, \dots, \mathbf{Y}_{ip}^*)'$, $i = 1, \dots, n$, with a Gaussian distribution with mean vectors $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})'$ and a common covariance matrix $\boldsymbol{\Sigma}$. Also, a set of q covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})'$ exists for every sampled unit i and $\boldsymbol{\mu}_i = \mathbf{B}'\boldsymbol{x}_i$ where $\mathbf{B}$ is a $q \times p$ matrix of unknown coefficients (Di Zio and Guarnera, 2013). The corresponding true data model can be expressed as

$$\mathbf{Y}^* = \mathbf{XB} + \mathbf{U} \qquad [1]$$

where $\mathbf{Y}^*$ is the $n \times p$ true data matrix, $\mathbf{X}$ – $n \times q$ covariate matrix and $\mathbf{U}$ – $n \times p$ matrix of normal residuals. Rows of the matrix $\mathbf{U}$ are independent realizations of Gaussian random vectors with mean equal to $\mathbf{0}$ and a covariance matrix $\mathbf{\Sigma}$.
Generic marginal probability distributions of the ith sampled unit of matrices $\mathbf{Y}^*$ (true data) and $\mathbf{U}$ (residuals) are denoted as

$$f(\boldsymbol{y}_i^*) = N(\boldsymbol{y}_i^*;\ \boldsymbol{\mu}_i, \mathbf{\Sigma}),\ \ f(\boldsymbol{u}_i) = N(\boldsymbol{u}_i; \mathbf{0}, \mathbf{\Sigma}),\ \ i = 1, \dots, n. \tag{2}$$

In general, form $N(\boldsymbol{y};\ \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a marginal probability distribution of the p-variate random vector $\mathbf{Y}$ with mean equal to $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.
It is assumed that the presence of errors in data are described by independent Bernoulli random variables. Therefore, the observed (erroneous) data can be expressed as

$$\mathbf{Y} = \mathbf{Y}^* + \mathbf{I}\boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{I}$ is a diagonal $n \times n$ matrix with its diagonal elements equal to Bernoullian variables $I_1, \dots, I_n$ ($I_i = 1$ if the corresponding sampled unit is erroneous, and $I_i = 0$ otherwise, $i = 1, \dots, n$). A marginal probability distribution of the p-variate random vector $\boldsymbol{\epsilon}_i$ (random noise) can be expressed as

$$f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \mathbf{0}, \mathbf{\Sigma}_\epsilon),\ \ \mathbf{\Sigma}_\epsilon = (\alpha - 1)\mathbf{\Sigma}, \tag{4}$$

with a numeric constant $\alpha > 1$.
$f(\boldsymbol{y}|\boldsymbol{y}^*)$ denotes a conditional marginal probability distribution of random variables $\mathbf{Y}$ and $\mathbf{Y}^*$. Therefore, [3] can be expressed equivalently:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1 - \pi)\delta(\boldsymbol{y} - \boldsymbol{y}^*) + \pi N(\boldsymbol{y};\ \boldsymbol{y}^*, \mathbf{\Sigma}_\epsilon) \tag{5}$$

where π is "a priori" probability of contamination and $\delta(\boldsymbol{y} - \boldsymbol{y}^*)$ is the Dirac delta function with mass at $\boldsymbol{y}^*$.
Furthermore, a marginal probability distribution of the observed data can be expressed as

$$\begin{aligned} f(\boldsymbol{y}_i) &= \int_0^\infty f(\boldsymbol{y}_i, \boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^* \\ &= \int_0^\infty f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^* \\ &= (1 - \pi) N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \mathbf{\Sigma}) + \pi N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \alpha\mathbf{\Sigma}). \end{aligned} \tag{6}$$

Coefficients of the latter observed data model can be obtained by the maximum likelihood estimation.

2.2 Selective Editing

Selective editing is based on the comparison between the observed data and predictions of the true (unobserved) data. The latter can be obtained from a conditional marginal probability distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ (Di Zio and Guarnera, 2013). An application of the Bayes formula provides:

$$\begin{aligned} f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) &= \frac{f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)}{\int_0^\infty f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^*} \\ &= \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i) N(\boldsymbol{y}_i^*;\ \tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{\Sigma}}) \end{aligned} \tag{7}$$

where

$$\tilde{\boldsymbol{\mu}}_i = \frac{y_i + (\alpha - 1)\boldsymbol{\mu}_i}{\alpha},$$

$$\widetilde{\boldsymbol{\Sigma}} = \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Sigma},$$

$\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i)$ is the Dirac delta function with mass at $\boldsymbol{y}_i$, $\tau_1(\boldsymbol{y}_i)$ and $\tau_2(\boldsymbol{y}_i)$ are posterior probabilities that the ith sampled unit with observed values $\boldsymbol{y}_i$, $i = 1, \dots, n$, is not erroneous and that it is contaminated respectively:

$$\tau_1(\boldsymbol{y}_i) = P(\boldsymbol{y}_i = \boldsymbol{y}_i^* | \boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})},$$
$$\tau_2(\boldsymbol{y}_i) = P(\boldsymbol{y}_i \neq \boldsymbol{y}_i^* | \boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i). \quad [8]$$

Posterior probabilities [8] are defined in terms of the conditional expected value $\widetilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$, $i = 1, \dots, n$. Therefore, the expected error can be defined as

$$\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_i = \tau_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \widetilde{\boldsymbol{\mu}}_i). \quad [9]$$

In practice, [9] is usually applied by using maximum likelihood estimates instead of the corresponding true data values.

2.2.1 Definition of the Score Function

Hereinafter $\hat{p}$ denotes a maximum likelihood estimate of some parameter p.

Suppose one seeks to estimate a sum of the variable Y_j^*, $j = 1, \dots, p$, with a sampling weight w_i of the ith sampled unit, – $T_j^* = \sum_{i=1}^n w_i y_{ij}^*$. A ratio between the expected error [9] with a sampling weight w_i multiplied by a suspicion component s_{ij} (probability that the ith sampled unit is erroneous) and the target parameter estimate $\hat{T}_j = \sum_{i=1}^n w_i \hat{y}_{ij}$ represents a conditional error of the ith sampled unit:

$$r_{ij} = \frac{s_{ij} w_i (y_{ij} - \hat{y}_{ij})}{\hat{T}_j}. \quad [10]$$

A local score function for the variable Y_j is denoted as $S_{ij} = |r_{ij}|$. Separate local scores can be combined into one global score GS_i in a few different ways:

$$GS_i = \max_j S_{ij} \quad \text{or}$$
$$GS_i = \sum_j S_{ij}.$$

In order to identify an optimal number of observations to be edited, the corresponding sampled units are sorted in descending order according to the GS_i. First $\tilde{k}$ observations are then chosen for the editing procedure:

$$\tilde{k} = \min\left\{k^* \in 1, \dots, n \,\middle|\, \max_j R_{kj} < \eta,\ \ \forall\, k > k^*\right\} \quad [11]$$

where $R_{ij} = \left|\sum_{k \geq i}^n r_{kj}\right|$ with an accuracy level η.

The suspicion component s_{ij} might have a discrete form (e.g., $s_{ij} \in \{0, 1\}$) or a continuous form ($s_{ij} \in [0, 1]$). In the paper the latter continuous suspicion component is defined according to Norberg et al. (2010). An additional test variable should be defined prior to defining the suspicion component:

Definition 1 (Test variable) *Test variable* $\boldsymbol{t}$ can be a combination of variables from a statistical survey and (or) some additional information. Statistical errors might then be identified by

checking whether a value of the test variable $\boldsymbol{t}_{j'}$, $j' = 1, \dots, p'$, for the ith sampled unit falls into a chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$.

The above-mentioned statistical error might be an observation that stands out compared to the rest of observations in the corresponding data set, to observations of a previous round of the same statistical survey or to some other additional information (e.g., administrative data). A few examples of a non-statistical error might be inconsistent answers the same respondent provides to the same question over different periods of time (e.g., a variable does not equal to the sum of its summands), disallowed values (e.g., observations that do not fall into a previously defined quantitative interval), item non-response, etc.

Definition 2 (Discrete suspicion component) *Discrete suspicion component*

1. $s_{ij} = 1$ if a value of the jth ($j = 1, \dots, p$) survey variable of the ith ($i = 1, \dots, n$) sampled unit y_{ij} is a non-statistical error;
2. $s_{ij'} = 1$ if a value of the j'th ($j' = 1, \dots, p'$) test variable of the ith sampled unit $t_{ij'}$ is a statistical error, i.e., $t_{ij'} \notin \left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$. In this case, $s_{ij} = 1$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$;
3. $s_{ij} = 0$ otherwise.

Nonetheless, it is important to take into consideration a different distance between observations that do not fall into the chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$ and the corresponding bounds of the region. The continuous suspicion component might convey the information on the latter distance more effectively.

Definition 3 (Continuous suspicion component) *Continuous suspicion component*

1. $s_{ij} = 1$ if a value of the jth ($j = 1, \dots, p$) survey variable of the ith ($i = 1, \dots, n$) sampled unit y_{ij} is a non-statistical error;
2. $\tilde{s}_{ij'} = \frac{\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) - t_{ij'}}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} < \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right)$;
3. $\tilde{s}_{ij'} = \frac{t_{ij'} - \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} > \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$;
4. $\tilde{s}_{ij'} = 0$ if $\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) < t_{ij'} < \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$.

The continuous suspicion component then equals to $s_{ij'} = \tilde{s}_{ij'} / \left(\tau + \tilde{s}_{ij'}\right)$ with parameters $\kappa \geq 0, \alpha > 0$ and $\tau > 0$ that regulate the size of the acceptance region. $s_{ij} = \max_{j'} s_{ij'}$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$.

3. Results of the Outlier Detection Study

The outlier detection study was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. The purpose of the statistical survey is to prepare and publish statistical information on the sales income (turnover) and their indices of service enterprises and provide the data for users on short term statistics. Enterprise turnover[1] of the accounting period was the target variable of the study.

In order to obtain the most suitable yet flexible enough selective editing model, four different predictor variables were chosen for comparison purposes – turnover from value-added tax (hereinafter referred to as VAT) declarations, turnover from the quarterly F-01 questionnaire[2], average number of employees and total hours worked. It is known that the first two predictor variables (turnover indicators) tend to have a high correlation with the target variable. On the contrary, the last two predictor variables (labour statistics indicators) usually have a lower

correlation with the target variable of the study. The latter characteristic will be explored more in the following part of the paper. Every chosen predictor variable was used separately providing four data sets and therefore four different cases of selective editing application.
For simplicity, the outlier detection study was carried out using only units with non-missing values that are greater than 0 for both the target variable and the corresponding predictor variable. In practice, an acceptance region, e.g., $(Q_1 - 3IQR, Q_3 + 3IQR)$ with the first quartile Q_1, the third quartile Q_3 and an interquartile range IQR, is usually constructed for the outlier detection in other units. As four data sets contain different number of non-missing values that are greater than 0, the number of observations left for the further study varies. The corresponding number of observations in data sets (primary populations) according to the predictor variable is given in Table 1 below.

Table 1: Number of observations in statistical data sets

Predictor variable	**Number of observations**
Turnover from VAT declarations	4085
Turnover from the quarterly F-01 questionnaire	574
Average number of employees	4867
Total hours worked	4931

The data contamination process was an important part of the outlier detection study as it contributed to selecting a level of accuracy for the further study (see
Table ***2***), and finding the most suitable outlier detection procedure given the chosen accuracy. In order to control the data contamination process, influential outliers in primary populations had to be replaced with plausible values. Therefore, default outlier detection procedure was performed using the statistical programming language R and its package "SeleMix", and the detected outliers in primary populations were replaced with contamination model predictions. The following algorithm was then applied to every modified primary population:

1. The target variable was contaminated in 3 different ways:
    a. 1.5 percent of observations were multiplied by 100,
    b. 2 percent of observations were trimmed leaving only the first and the last digits,
    c. 20000000 was added to 1.5 percent of observations;
2. Estimation of model coefficients and outlier (potential error) detection were carried out using the statistical programming language R and its package "SeleMix" (function "ml.est");
3. Values of the target variable were sorted in descending order according to estimates of the global score function obtained using the statistical programming language R and its package "SeleMix" (function "sel.edit"). An estimate of the global score function is close to 0 when a value of the target variable is not identified as an outlier and therefore has no major impact on the quality of sample estimates, and greater than 0 when a value of the target variable is identified as an outlier;
4. The part of outliers that have a major impact on the quality of sample estimates (influential errors) was chosen for the editing procedure.

The latter influential error detection procedure was repeated in two different ways – by calculating estimates of the score function (1) with a discrete suspicion component that stays the same among all observations ($s_i = 1$), and (2) with a continuous suspicion component. The latter suspicion component was designed using an acceptance region between the first and the third quartiles $\left(\hat{t}^{(L)}, \hat{t}^{(U)}\right)$ where $\hat{t}_i = \hat{y}_i \ \ (i = 1, \dots, n)$. Parameters κ and τ varies ($\kappa \in \{0, 0.5, 1, 1.5\}$, $\tau \in \{0.1, 0.5, 1, 1.5, 2\}$) and specific values are chosen depending on the lowest number of identified influential errors, $\alpha = 0.05$.

Selective editing with different levels of accuracy gives a different number of influential errors. If all of the detected influential errors were introduced by the above-mentioned data contamination procedure, the corresponding level of accuracy was chosen for the further study (see
Table ***2*** below).

Table 2: Levels of accuracy (threshold values) for statistical data sets

Predictor variable	**Level of accuracy**
Turnover from VAT declarations	0.011
Turnover from the quarterly F-01 questionnaire	0.004
Average number of employees	0.027
Total hours worked	0.026

As it can be observed from the above-given
Table ***2*****:** Levels of accuracy (threshold values) for statistical data sets, levels of accuracy fall into two groups – below 0.02 and above 0.02. It gives an insight that selective editing models with either turnover indicator as a predictor variable might perform better and result in higher quality sample estimates. Whereas labour statistics indicators chosen as predictors might provide lower quality selective editing models. The same two groups of predictor variables were mentioned earlier in the paper while comparing the correlation with the target variable of the study. Having the four specific statistical data sets, correlation coefficients between predictor and target variables may be calculated easily (see Table 3 below). Correlation coefficients separate predictor variables into the same two groups – turnover indicators have a higher correlation with the target variable (both above 0.9) while labour statistics indicators tend to have a lower correlation with the target variable (both below 0.6).

Table 3: Correlation coefficients between target and predictor variables

Predictor variable	**Correlation coefficient**
Turnover from VAT declarations	0.96
Turnover from the quarterly F-01 questionnaire	0.97
Average number of employees	0.52
Total hours worked	0.52

Keeping in mind the chosen levels of accuracy (see
Table ***2***), the results of different selective editing approaches were then compared by estimating the relative absolute bias after every sequential edit of each influential error. The latter procedure helped determining a number of influential errors that have to be edited in order to achieve desired levels of accuracy. The combined results are provided in Table 4 below.

Table 4: Number of influential errors in statistical data sets

Predictor variable	**Total number of influential errors**	**Number of influential errors to be edited**
(1) Selective editing with the discrete suspicion component		
Turnover from VAT declarations	134	92
Turnover from the quarterly F-01 questionnaire	23	14
Average number of employees	90	>90
Total hours worked	111	>111
(2) Selective editing with the continuous suspicion component		

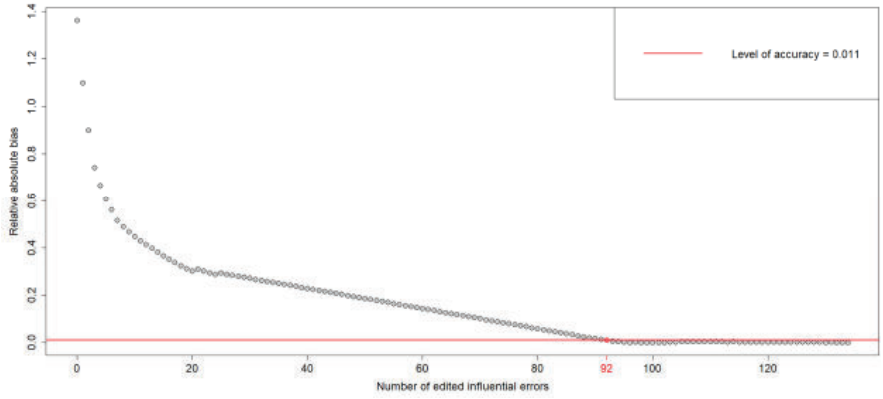
Turnover from VAT declarations	93	92
Turnover from the quarterly F-01 questionnaire	15	14
Average number of employees	136	121
Total hours worked	124	123

Results observed applying selective editing with different approaches may be compared in two different ways – the impact of the discrete / continuous suspicion component (Subsection 3.1) and the impact of the predictor variable (Subsection 3.2).

3.1 The Impact of the Suspicion Component

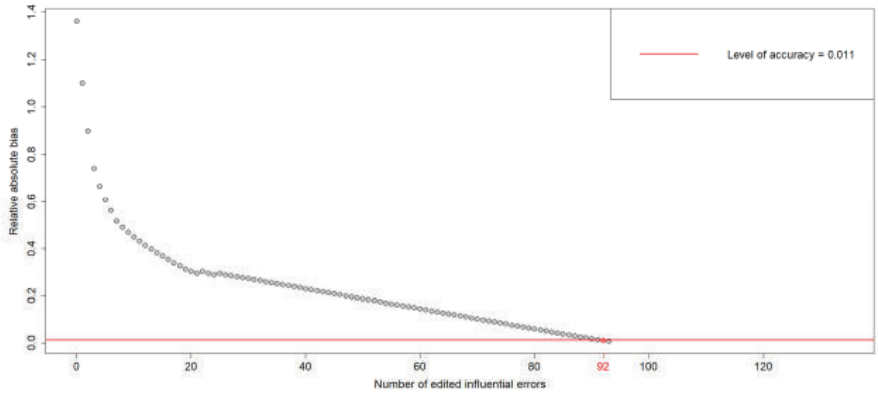
Consider the case when a predictor variable is turnover from VAT declarations. As observed in Table 4, a total number of influential errors differ significantly depending on a type of the suspicion component.
When the discrete suspicion component is used in the selective editing model, the same value of suspicion is used for every observation ($s_i = 1$). Therefore, no additional impact is put on determining whether the corresponding observation is an influential error or not. Although a total of 134 errors are identified as influential, the relative absolute bias calculation shows that only 92 of the identified influential errors have to be edited in order to achieve the desired level of accuracy (0.011), see Figure 1 below.

Figure 1: Relative absolute bias dependency on the number of edited influential errors using the discrete suspicion component with turnover from VAT declarations as the predictor variable



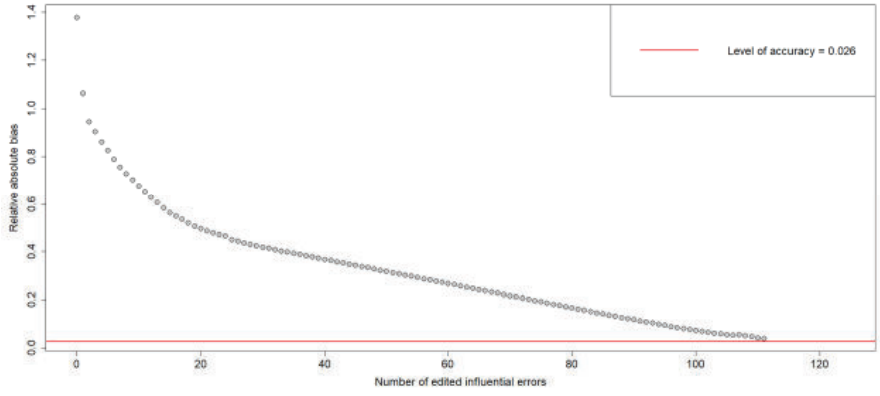
On the contrary, the use of the continuous suspicion component lets to take into consideration distances between observations that do not fall into the chosen acceptance region and the corresponding bounds of the region. With this additional impact on the selective editing model, the calculation of relative absolute bias shows that almost every identified influential error (92 out of 93) has to be edited in order to achieve the desired level of accuracy, see Figure 2 below. The latter example illustrates an advantage of the continuous suspicion component over the discrete suspicion component. In practice, the use of the continuous suspicion component could prevent statisticians from the overediting of data while preserving the preferable quality of sample estimates.

Figure 2: Relative absolute bias dependency on the number of edited influential errors using the continuous suspicion component with turnover from VAT declarations as the predictor variable



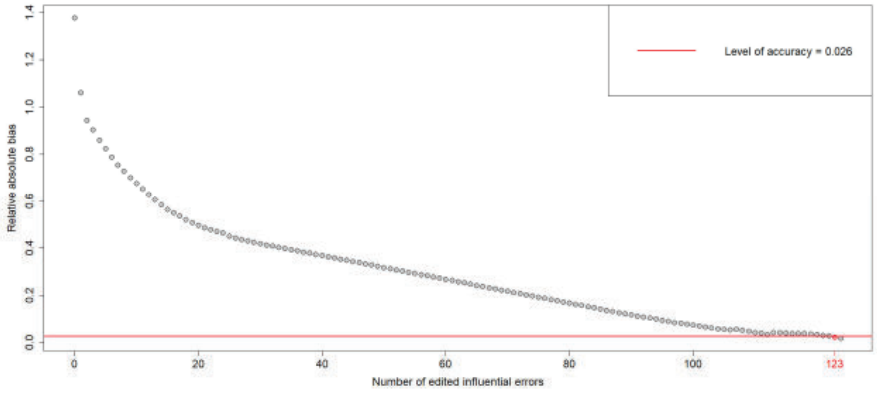
Now consider the case when a predictor variable is one of labour statistics indicators – total hours worked. Table 4 indicates a significant difference between a total number of influential errors while applying selective editing with different types of the suspicion component.
In the case when the discrete suspicion component is used, relative absolute bias calculation demonstrates that editing all the identified 111 influential errors does not result in sample estimates with the preferable level of accuracy (0.026), see Figure 3 below.

Figure 3: Relative absolute bias dependency on the number of edited influential errors using the discrete suspicion component with total hours worked as the predictor variable



Similar to the previously seen example with turnover from VAT declarations as the predictor variable, the use of the continuous suspicion component increases the chance of achieving the preferable level of accuracy. The calculation of relative absolute bias shows the need to edit almost every identified influential error (123 out of 124), see Figure 4 below.

Figure 4: Relative absolute bias dependency on the number of edited influential errors using the continuous suspicion component with total hours worked as the predictor variable



As it was illustrated in Table 4 and Figure 1 through Figure 4, applications of selective editing with different predictor variables have shown an effectiveness of the continuous suspicion component on the outlier detection procedure. This approach to selective editing lets to identify only the most important influential errors, and therefore prevents from the overediting or, in other cases, insufficient data editing.

3.2 The Impact of the Predictor Variable

Another important factor in choosing the most suitable selective editing model besides the suspicion component is the predictor variable. As it was previously noted in the paper, a weak dependency between the predictor variable and the target variable of the study might be considered as a drawback of a selective editing model. The latter model would result in a lower level of accuracy and therefore lower quality of sample estimates. Moreover, the use of the discrete suspicion component might be inconsistent with the preference to achieve the previously chosen level of accuracy. As it was shown in Table 4 and Figure 3, a previously chosen, i.e., "desired", level of accuracy was not achieved using the discrete suspicion component when one of labour statistics indicators was the predictor variable.
In contrast, a strong dependency between the predictor variable and the target variable of the study results in a better selective editing model. A previously chosen level of accuracy may be achieved using both the discrete and the continuous suspicion components. However, as it was illustrated in Subsection 3.1, the continuous suspicion component might be a more efficient choice.

4. Conclusions

After calculations of the relative absolute bias dependency on the number of edited influential errors, selective editing with the continuous suspicion component was determined to be an optimal method for the outlier detection procedure. The latter version of selective editing prevents from the overediting or, in some cases, insufficient statistical data editing.
Turnover from VAT declarations and turnover from the quarterly F-01 questionnaire were identified as the most suitable predictor variables for the outlier detection procedure. The main property of a suitable predictor variable turned out to be a high correlation between the latter predictor variable and the target variable of the study.

Main findings of the outlier detection study contributed to the outlier detection procedure currently implemented at Statistics Lithuania. The construct of the continuous suspicion component lets to improve the existing selective editing model and focus statistical data editing on the most important influential errors while preserving the quality of sample estimates.
In the paper, all four cases of selective editing were modelled using undivided data sets. A possible extension of the carried out study could focus on the search of the most suitable selective editing model when data sets are separated into groups according to some factor, e.g., the economic activity of enterprises.

5. References

Burakauskaitė, I. and Nekrašaitė-Liegė, V. (2021), "Selective Editing Using Contamination Model", SUMMER SCHOOL ON SURVEY STATISTICS 2021, BNU Network on Survey Statistics, Statistics Lithuania, Vilnius, Lithuania, 40–45.
Di Zio, M. and Guarnera, U. (2013), "A Contamination Model for Selective Editing", Journal of Official Statistics, 29, 4, 539–555.
Guarnera, U. and Buglielli, M. T. (2013), "SeleMix: an R Package for Selective Editing", Italian National Institute of Statistics, Rome, Italy, 2013-12-12.
Hedlin, D. (2003), "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics", Journal of Official Statistics, 19, 177–199.
Latouche, M. and Berthelot, J. M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys", Journal of Official Statistics, 8, 389–400.
Lawrence, D. and McDavitt, C. (1994), "Significance Editing in the Australian Survey of Average Weekly Earnings", Journal of Official Statistics, 10, 437–447.
Lawrence, D. and McKenzie, R. (2000), "The General Application of Significance Editing", Journal of Official Statistics, 16, 243–253.
Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P. and Nordberg, L. (2010), "A General Methodology for Selective Data Editing", Statistics Sweden, Stockholm, Sweden, 2010-02-04.
Official Statistics Portal (2021), "The sales income (turnover) and indices of servise enterprises", Metadata, Statistics Lithuania, Vilnius, Lithuania, 2021-11-26, retrieved from https://osp.stat.gov.lt/documents/10180/5118910/Paslaug%C5%B3+%C4%AFmoni%C5%B3+rodikliai+%5BEN%5D+636.html/62ebe741-2e13-4398-8630-34f666003eb0, accessed 2022-01-14.
Official Statistics Portal (2022), "Financial indicators of enterprises", Metadata, Statistics Lithuania, Vilnius, Lithuania, 2022-01-05, retrieved from https://www.stat.gov.lt/documents/10180/5118910/%C4%AEmoni%C5%B3+finansiniai+rodikliai+%28AB%2C+UAB%2C+V%C4%AE%2C+S%C4%AE+ir+kiti%2C+i%C5%A1skyrus+I%C4%AE+ir+fizinius+asmenis%29+%5BEN%5D+5101.html, accessed 2022-01-14.

[1] *Enterprise sales income (turnover)* refers to income from selling goods and / or providing services received by an economic entity in the reporting period (VAT excluded). Income from selling long-term fixed assets, financial and investment activity, dividend, etc., as well as funding from the budget, are excluded. The services provided do not include the acquisition value of customer's materials, products, spare and component parts (Official Statistics Portal, 2021).
[2] The quarterly F-01 questionnaire was a tool for collecting statistical data for the quarterly statistical survey on financial indicators of enterprises. The objective of statistical information was to provide for users quarterly statistical information on the performance and financial indicators (employees, income, expenses, profit, equity, liabilities, assets, etc.) and efficiency ratios (profitability, liquidity, turnover, etc.) of non-financial corporations (Official Statistics Portal, 2022).