

A Comparison of the Outlier Detecting Methods: An Application on Turkish Foreign Trade Data

Dr. Aylin KOLBAŞI¹, Prof. Dr. Aydın ÜNSAL²

Abstract

In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps in obtaining a consistent analysis is to identify observations that are far from the center. Outlier values often carry important information, even if they are considered to be errors or problems. However detected outliers lead to incorrect modeling, biased parameter estimates, and incorrect results. For this reason, it is very important to identify them before modeling and analyzing. In this study it has been tried to determine the method of detecting outliers which can detect the outliers in foreign trade data in the most accurate way. For this, z-score, median z-score, box-plot, adjusted box-plot, B-H method, k-means clustering method and robust regression methods were applied to the foreign trade data set and the application results of outlier detection methods were compared by considering some criteria. In the light of these criteria, the most appropriate outlier detection methods that can be applied to the foreign trade data set have been determined.

Keywords: Outlier, foreign trade, outlier detection methods.

1. Introduction

Foreign trade is expressed as a commodity trading activity of one country to other countries and it has a dynamic structure that changes continuously. Foreign trade is very important in the development process of the countries. Therefore, when foreign trade and economic policies are determined, basic determinations should be made such as which sector imports or exports, protectionist or free foreign trade policy should be followed, and developments in foreign trade should be observed. The most important tools used for this are foreign trade data and foreign trade statistics and indicators calculated using these data. Foreign trade statistics and indicators are calculated and published by Turkey Statistical Institute (TURKSTAT) officially. The data source of foreign trade statistics is the records of customs declarations received from the Ministry of Trade. It's possible that make mistakes in the values of the quantity and value (dollar) variables in the monthly export and import during the data entry. These possible errors lead to erroneous calculation of figures such as total exports, total imports, foreign trade volume, foreign trade deficit, export and import change rates compared to the same month of the previous year calculated as a result of data analysis. For this reason, it is necessary to identify and correct these erroneous records before calculating foreign trade statistics. These erroneous records are considered as outliers so outlier detection methods are used to detect these possible errors in the foreign trade data set.

¹ aylinkarabacak@tuik.gov.tr, Turkish Statistical Office, Ankara/Türkiye

² aunsal@gazi.edu.tr, Hacı Bayram Veli University, Ankara/Türkiye

There are many definitions of outliers. One of them is Hawkins' definition. Hawkins defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. (Hawkins,1980).

In this paper, it is aimed to determine the method of detect the outliers in the foreign trade data set which can detect the most accurate values.

2. Literature

There are many methods in the literature that detect outliers. Determination of the appropriate outlier detection method applied to the data set to be analyzed requires a research on its own. From this point of view, it is aimed in this study to determine the outlier detection method which make the most accurate detection of outlier in foreign trade transactions. There are many studies in the literature to compare outlier detection methods. In their study of Kuppusamy and Kaliyaperumal in 2013, they compared some methods such as the Grubbs test, the Dixons Test, and the Hampel method, which were used to detect outliers (Kuppusamy ve Kaliyaperumal, 2013, p.713). In their study in 2001, Penny and Jolliffe compared the use of multivariate outlier detection methods using clinical laboratory safety data (Penny ve Jolliffe, 2001, p.306). The European Statistical Office has compared the methods using foreign trade data in a study they have done in their own field (Aribbas, 2012, p.49). It is possible to come across many studies aimed at comparing the outlier detection methods. While some of these studies only compare methods among themselves, some compare the results of methods using specific data sets. In his 2006 study, Seo compared the methods with specific criteria (Seo, 2006, p.6). However, Seo only made this comparison taking into account the univariate outlier detection methods. In this study, the advantages and disadvantages of the methods are summarized and the results are compared with each other. It is revealed within the framework of the criteria that determine how and under which conditions the result is weak and strong. In this respect, this study aims to give a different perspective to the methods of outlier evaluation.

3. Review of Methods Used

The purpose of data analysis is to evaluate and summarize the compiled information about a large number of variables. One of the first steps to achieving a consistent analysis is the detection of observations far from the center. Although outliers are generally considered as errors or problems, they are considered as variables that carry important information in some studies. However, outliers are generally the values that lead to incorrect models, incorrect parameter estimations, and inaccurate results. Therefore, it is great importance to identify them before modeling and analysis.

Outlier detecting methods are classified in different ways. One of them is to classify these methods in two categories as formal tests and non-formal tests. Many formal tests require test statistics for hypothesis testing, and often a distribution assumption. The goal is to test that the outliers are deviating from the default distribution. Although formal tests are quite powerful under statistical assumptions, the distribution of real world data sets is not known or does not follow a certain distribution, such as normal, gamma or exponential distribution. Another limitation of formal testing is that it is sensitive to masking and swamping problems. These problems will be described in the next section.

On the other hand, many non-formal tests or in other words outlier labeling methods create a range or criteria for outliers instead of hypothesis testing and any observation outside the range or criterion is considered outliers. Many location and scale parameters are used to create a reasonable criterion or range for the detection outliers. There are 2 reasons to use the outlier labeling method;

First; formal tests generally require a hypothesis test and test statistic based on the distribution assumption, in order to detect whether the target outlier is the actual outlier of the distribution, while most of the outlier labeling methods use an interval created by the scale and location parameters of the data. Although the method of labeling is easy to use, some observations outside the range can be mistakenly defined as outliers. However, the primary objective of the detection of outliers is not to detect the outliers that violate the distribution assumptions of basic statistical analyzes such as t-test, ANOVA and regression but rather to detect the outlier values that are located away from the majority of the data set regardless of the distribution, labeling methods are used. At the same time, it is difficult to determine the distribution of the data for large data sets or to convert it into a suitable distribution such as normal distribution.

In this study, the outlier labeling methods will be used since the purpose of the detection of the outliers is to detect the values that are far away from the majority of the data set. In addition to outlier labeling methods, the clustering method and the robust regression method will be used to detect the outliers. It is also possible to evaluate the outlier labeling methods as univariate outlier detection methods, clustering and robust regression method as a multivariate outlier detection methods.

The outliers detection methods used in this study are as follows;

- 1-Z-score
- 2-Median z-score
- 3-Box-Plot Method
- 4-Adjusted Box-Plot Method
- 5-Hidroglou - Berthelot Method
- Multivariate outlier detection methods;
- 1-k-means clustering
- 2-Robust regression method

3.1 Z-score

The z-score is a method that uses mean and standard deviation values to detect outliers in the data set.

$$Z_i = \frac{x_i - \bar{x}}{sd} \quad 3.1$$

$X_i \sim N(\mu, \sigma^2)$ ve $sd = \text{standart deviation}$

The main basis of this rule is that X has a normal distribution and therefore z has a standard normal distribution. When the integer values of the z-scores exceed 3, the corresponding observation value is detect as an outlier (Aribbas, 2012). Although this method is simple, it is convenient to use when the data set distribution is normal.

According to the Shiffler, the maximum possible z-score value depends on the number of observations and is calculated by the formula $\frac{(n-1)}{\sqrt{n}}$ (Shiffler, 1988). Proof of this formula is as follows;

Let X_1, \dots, X_n , be a random sample size of n-1 population with unknown mean and variance and number of observation n , assume the X_i are already ordered and $\bar{X}_{n-1} = 0$ Suppose another observation X_n added to

the sample such that $X_n > X_{n-1}$. The mean of the n values is X_n/n , and the variance estimate $S_n^2 = \frac{\sum(X_i - \bar{X}_n)^2}{(n-1)}$ can be written as;

$$S_n^2 = \left[\frac{(n-2)}{(n-1)} \right] S_{n-1}^2 + \frac{X_n^2}{n} \quad 3.2$$

The largest positive z score for the n values;

$$Z_{(n)} = (X_n - \frac{X_n}{n})/S_n \quad 3.3$$

is maximized whenever S_n minimized. From 3.2 this occurs when $S_{n-1}^2=0$; hence equation 3.3 reduces to;

$$Z_{(n)} = \frac{(n-1)}{\sqrt{n}} \quad 3.4$$

Equation 3.4 is the maximum achievable value for the largest positive z score based on n values. Table 1 lists this maximum z-scores for several sample sizes.

Table 1. Maximum absolute z score per sample size

n	Z-score
3	1.155
4	1.5
5	1.789
10	2.846
11	3.015
17	3.881
18	4.007

As can be seen from Table 1, the z-score method is not particularly suitable for small data sets, since no z-score value exceeds 3 when the number of observations is less than 10 and equal to 10. Another limitation for this method is that the standard deviation value is affected by one or more outliers. This leads to masking problems. Due to the masking, it is not possible to detect significant outliers.

3.2 Medyan z-score

The mean and standard deviation estimators used in the z-score method are influenced by a single outlier as previously mentioned. In order to avoid this problem, median and median absolute deviation values are used instead of mean and standard deviation in the median z-score method. The median absolute deviation value is calculated according to the following formula.

MAD=median ($|X_i - \tilde{X}|$), \tilde{X} , sample median, Median z-scores are calculated as follows;

$$Mi = \frac{0,6745(X_i - \tilde{X})}{MAD} \quad \text{large for a normal data set } E(MAD) = 0,6745$$

Iglewicz and Hoaglin (1913) stated that the observation values should be labeled as outliers in the case of $|Mi| > 3.5$ As in the z-score method, Mi scores are effective for data sets from normal distribution (Seo, 2006).

3.3 Box-Plot Method

The Tukey method, which forms a box chart, is a well-known simple graphical tool, which shows information such as the upper and lower outliers of the univariate data set, lower and upper quartiles, and the median. According to the methods used the mean and the standard deviation, it is less affected by outliers because it uses quartiles. The rule for this method is given below.

- 1- Firstly, the interquartile range called IQR is calculated. $IQR = Q3 - Q1$
- 2- The internal boundaries are calculated by the formulas $Q1 - 1.5IQR$, $Q3 + 1.5IQR$.
- 3- The external boundaries are calculated with $Q1 - 3IQR$, $Q3 + 3IQR$ formulas.
- 4- Observation values between internal and external boundaries are defined as possible outliers.

Observation values outside the external boundaries are determined as outliers. There is no statistical basis for the 3 and 1.5 coefficients that Tukey uses to calculate the boundaries (Seo, 2006). While the normal distribution hypothesis is required in the methods described previously, Box-Plot method can be used in data sets with skewed distribution since it does not use mean and standard deviation values and does not require any distribution assumption. However, the Tukey method is not suitable for data sets for very small number of observations (Seo, 2006).

3.4 Adjusted Box-Plot Method

Although Tukey's Box-Plot method is applicable to both symmetric data and skewed data, it causes a large number of observations to be determined as outliers in the data with high skewness. This is due to the use of the lower and upper quadrants and interquartile distances measured without considering the skewness of the data set. Vanderviere and Huber have defined a new rule taking into account the medcouple (MC), which is a robust skewness measure for skewed data (Vanderviere and Huber, 2004). The calculation of MC coefficient is described below;

Let $X_n = \{X_1, \dots, X_n\}$ be an independent data set from a continuous univariate distribution. Let the data set be sorted from small to large with $X_1 \leq \dots \leq X_n$. In this case the MC coefficient is calculated according to the following formula;

$$MC(X_1, \dots, X_n) = \text{med} \frac{(X_j - \text{med}_k) - (\text{med}_k - X_i)}{X_j - X_i}$$

med_k = Median of the X_n data set

$X_i \leq \text{med}_k \leq X_j$ ve $X_i \neq X_j$ Ranges are calculated as follows;

$$[L, U] = [Q_1 - 1.5 \exp(-3.5MC)IQR, Q_3 + 1.5 \exp(4MC)IQR] \quad MC \geq 0$$

$$= [Q_1 - 1.5 \exp(-4MC)IQR, Q_3 + 1.5 \exp(3.5MC)IQR] \quad MC \leq 0$$

L: Lower boundary

U: Upper boundary

Observation values outside the boundaries are determined as outliers. The MC value ranges from -1 to 1. When the data is symmetrical, the MC value is 0. If $MC > 0$, the data is skewed to the right and $MC < 0$, the

data is skewed to the left. Vanderviere and Huber (2004) calculated the average percentages of outliers from the lower and upper limits for adjusted Box-Plot and original Box-Plot methods, taking into account the different numbers and distributions of observations. In the simulation study, it was concluded that the less observation value was found to be an outlier compared to the Tukey's Box-Plot method for a right skewed dataset. Compared to the Tukey method, in the case of a slightly right skewed distribution, the lower boundary moves to the right and more observation on the left side is determined as an outlier. This difference is due to the decrease in the lower limit and the increase in the upper limit. An example is given below.

Example:

Let take the data sets consisting of values of 3,2,3,4,3,7,3,7,3,8,3,9,4,4,4,1,4,2,4,7,4,8,14 ve 15. For this data set, Tukey's Box-Plot method and the adjusted Box-Plot method were applied and result are given in the figure as follows (Seo, 2006).

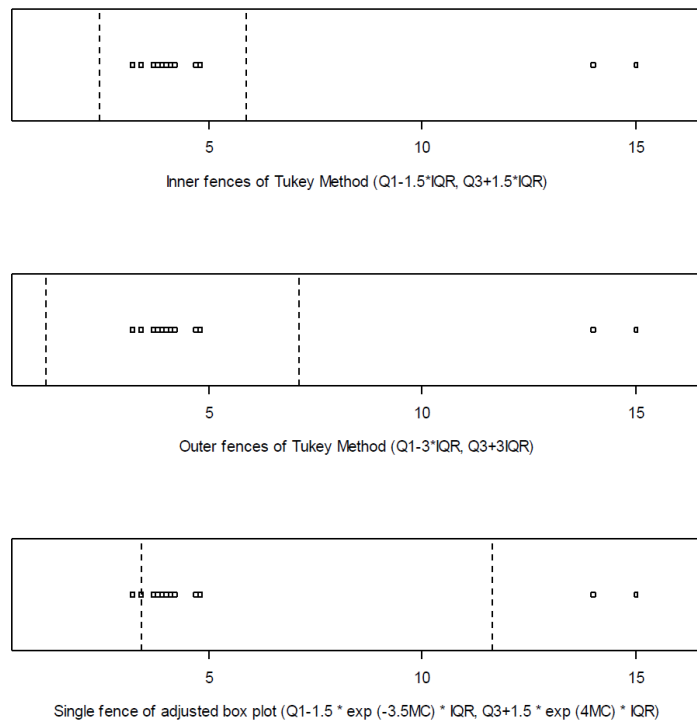


Figure 1. Change of the intervals of two different box-plot methods

The first figure in Figure 1 represents the internal boundaries of the Tukey's Box-Plot method calculated for the data set in the sample, the second figure shows outer boundaries of Tukey's Box-Plot and the final figure shows the boundaries calculated using the adjusted Box-Plot method. According to the Tukey's Box-Plot method, the internal borders are calculated as [2.45; 5.85] and the external boundaries are calculated as [1,18; 7,13]. When we look the inner and lower limits are calculated according to the Box-Plot method, we can see that only 14 and 15 values in the data set are detected as outliers. Boundaries are calculated according to the adjusted Box-Plot method as [3,41; 11,62]. Therefore, when the results of the adjusted Box-Plot method were examined, the lower boundary was moved to the right and 2 of the observation values to the left of the data set were detected as outliers. According to the results of the first method, no observation

value to the left of the data set can be detected as an outlier, while the second method can also detect the observation values on the left side. This is because the second method takes into account a boundary calculation that is free of the effect of skewed the data set. Therefore, a healthier result is obtained than the first method.

3.5 Berthelot-Hidiroglou Method

The BH method is a micro-correction method defined by Berthelot and Hidiroglou (1986). This method is based on the rates at which the boundaries are calculated automatically from the data set. In order to prevent the boundaries from being influenced by the outliers, parameters such as median, quartiles and interquartile distance are used, which are robust parameters instead of mean and standard deviation. The lower and upper limits are calculated according to the following formula (Belcher, 2003).

$$L = R_{\text{median}} - k * D_{rQ1}$$

$$U = R_{\text{median}} + k * D_{rQ2}$$

However, there are two disadvantages of using the rate method in this way. These;

1- It is difficult to determine the remaining values in the left tail of the data set.

2- It does not take into account that the variability in the ratios of small business is greater than in large business.

The BH method solves these problems with a size transformation following symmetric transformation. Transformations are given below.

Symmetric transformation;

$$S_i = \begin{cases} 1 - \frac{R_{\text{median}}}{R_i}, & 0 < R_i < R_{\text{median}} \\ \frac{R_i}{R_{\text{median}}} - 1, & R_i \geq R_{\text{median}} \end{cases}$$

Size transformation;

$$E_i = S_i * (\max(X_i(t), X_i(t+1))) \cup 0 \text{ Xiu max} 1$$

EQ1 and EQ3, 1st and 3rd quartiles of transformation;

$$DQ1 = \max(E_{\text{median}} - EQ1, |A * E_{\text{median}}|)$$

$$DQ3 = \max(EQ3 - E_{\text{median}}, |A * E_{\text{median}}|)$$

$$L = E_{\text{median}} - C * DQ1$$

$$U = E_{\text{median}} + C * DQ3$$

L: lower limit U: upper limit

A is fixed at 0.05. Therefore, U and C, it is possible to say that there are two parameters used in this method (Belcher, 2003). Conventionally, the ratio arrangement compares the ratios of two substances at pre-defined boundaries. This approach leads to the problem in the economic data set with highly skewed distribution. First, large cell ratios tend to dominate the distribution, and smaller outliers in the left tail of the

distribution cannot be detected. Second, in many economic programs, the variability of smaller cell estimates from period to period is greater than that of larger ratios. Berthelot and Hidirolou call it the width masking effect. Therefore, the traditional rate method leads to the detection of very small cell estimates as outliers. The BH method solves these two problems for positive proportions of distribution, using the E_i transformation, which takes into account the size of the data and S_i , which makes it possible to detect outliers in both tails and use the median to obtain a symmetric distribution.

For controlling importance of the magnitude of the data, U parameter is used in transformation E_i s are distributed around zero. If $U = 0$, there is no emphasis on the magnitude of the data and $E_i = S_i$. However, as the U parameter approaches 1, the outliers are composed of only large units.

When Emedyan-EQ1 and EQ3-Emedyan are too small The term $|A * Emedyan|$ prevents the detection of incorrect outliers. Hidirolou and Berthelot recommend that it be taken as $A = 0.05$.

In summary, the BH method works on the transformation series applied on ratios. One of the disadvantages of this application is the necessity to define the appropriate values for the U , C and A parameters. It is also possible to calculate the E_i values which are quite far from the data set for the R_i values near the data set. This situation reveals the fact that the observation values which are not outliers are detected as outliers (Ishikawa, Endo and Shiratori, 2010). Similarly, for the R_i values which are quite far from the data set, the E_i values which are very close to the data set can be calculated. In this case, in fact, the value of the observation value that should be detected as an outlier value can not be detected as outliers not be detected (Ishikawa, Endo and Shiratori, 2010). This means that the outliers cannot be detected correctly in the data set.

This method is adapted for the foreign trade data set as follows. First, P_i values are calculated and then S_i values are calculated by taking these values into account;

$$P_i = \text{Dollars/quantity}$$

$$Q = \text{Quantity}$$

$$S_i = \begin{cases} p_i/p_{Q2} - 1 & \text{if } p_i \geq p_{Q2} \\ 1 - p_{Q2}/p_i & \text{if } p_i < p_{Q2} \end{cases}$$

Then the e_i values are calculated according to the following formula:

$$e_i = s_i \max \{V_i, p_{Q2} Q_i\}^u$$

After calculating $d_{Q1} = \max\{e_{Q2} - e_{Q1}, |a e_{Q2}|\}$, $d_{Q3} = \max\{e_{Q3} - e_{Q2}, |a e_{Q2}|\}$, BH-scores are calculated as follows;

$$BH - Score_{e_i} = \begin{cases} (e_i - e_{Q2})/d_{Q3} & \text{if } e_i \geq e_{Q2} \\ (e_{Q2} - e_i)/d_{Q1} & \text{if } e_i < e_{Q2} \end{cases}$$

After the calculated scores are compared with the threshold value, it is decided whether the relevant observation value is an outlier. In this method, it is recommended to use values of 0.05, 0.5 and 4 in the parameters a and u and for the threshold value (Arribas, 2012). P_i values were calculated as $P_i = \text{Dollar / quantity 1, Dollar / quantity 2 and Quantity1 / Quantity2}$.

Multivariate outlier detection methods are discussed after univariate outlier detection methods.

3.6 k-means Clustering

k-clustering is an algorithm that separates the observation values into a number of groups according to certain characteristics. Grouping is done by minimizing the sum of squares of data between the data and the related cluster center. Therefore, the k-clustering method classifies the data set in this way. In the method of k-clustering outliers are expected to be collected in the same cluster. So, it is possible to say that clusters with low number of points are clusters consisting of outliers. The k-mean cluster method does not rely on any statistical assumption. The performance of the method depends on the capacity of the clustering variables that will distinguish between the observation values. In this method, two criteria are taken into account when detecting outliers.

1- R^2 value should be meaningful and the distance between the cluster thought to be composed of outliers and cluster center is high.

2- The number of observations within the cluster which is thought to be composed of outliers is small.

When R^2 is less than 50% and the number of observations is less than 15, this method does not work effectively (Arribas, 2012). Before applying the clustering method, the data are standardized in order to remove the measurement effect. Another issue to be considered in the clustering method is to determine the appropriate number of clusters. There are many criteria for determining the number of k. Because no method solves this problem in full. One of the criteria used to determine the appropriate number of sets is the distortion rule (Arribas, 2012). The application stages of this rule are as follows;

1.k-average clustering algorithm $k = 1, \dots, n$ is applied individually. For each application, $\min = (\text{sum of squares between classes}) / (n.p)$ is calculated.

n =number of observations in sample

p =number of values

2. Then $J_k = d_k^{-p/2} - d_{k-1}^{p/2}$ is calculated.

3. When the J_k value reaches the maximum, k is the appropriate number of clusters

Another criterion for determining the appropriate number of clusters is the categorized cubic clustering criterion (CCC) in the k-mean clustering analysis of the SAS Enterprise program. When the CCC value calculated for different k numbers is maximum, the suitable number of cluster is determined (SAS, 1998).

The following steps are followed in the method of detection outliers by using k-mean clustering method.

1- The data is primarily standardized.

2- The suitable number of cluster is determined.

3- After determining the appropriate number of clusters, clusters are determined by k-mean clustering method. After the clusters are determined, clusters with the least number of observations are determined and the distance of these clusters to the center is examined. Observation values in the smallest clusters with the highest distances to the center are detected as outliers. Before the clustering method was applied in the foreign trade data set, BRF1, BRF2 and M1 / M2 were standardized for unit measurement effect and the clusters were determined using these 3 variables.

3.7 Robust Regression Method

The ordinary least squares method (OLS) with ease of calculation is a method commonly used in regression analysis. However, the results of the OLS method for regression analysis are misleading unless some statistical assumptions are provided. The ordinary least-squares method is highly resistant to outliers. Such observations lead to not to be provided the assumption that the errors are identical and independent, and that the estimates are biased and ineffective. In these cases, it is more accurate to use robust regression methods. At the same time, robust regression methods serve as an important tool for analyzing the data set that is full of outliers. It is used to detect outlier and to obtain more resistant results against outliers (Chen, 2002).

In statistical studies, it is assumed that the observational values drawn randomly are independent and identical in order to create a sample. However, in actual data sets, there may be situations in which some observation values are too large or too small compared to other observation values, and it is not possible to expect such observations to have the same distribution as most of the data set. The summarized information (average, variance, etc.) of the estimators calculated by using the data sets with such observation values is misleading. Such estimators are sensitive to outliers. If an estimator is not affected by outliers, it is a robust estimator. Similarly, the estimation method can be classified as a robust and non-robust method. Robustness characteristics of the estimator and the degree of this robustness are expressed in various sizes such as breakdown point, effect function and sensitivity curve.

The breakdown point is defined as the supremum of the ratio of the number of outliers to the total number of observations that cause the estimator to move away from the parameter value (Yorulmaz, 2009). If the estimator breakdown point is greater than 0, then the estimator is said to be resistant estimator. The breakdown point of the arithmetic mean is 0. Because the existence of a single outlier affects the results of the arithmetic mean. The breaking point of the median is 0.5, which is the highest limit for a estimator. The breakdown point does not correspond to more than 50% of the total number of observations. Because if the number of outliers corresponds to more than half of the total number of observations, it is not possible to draw meaningful results about the distribution representing the data.

The most important aim of the robust regression method is to produce resistant results against outliers. For this reason, it limits the effect of outliers. The robust regression method overcomes the problems in 3 categories. These;

1. The outlier problems in y-direction
2. Multiple outlier problems in the covariance field (defined as lever point, outliers in the x-area)
3. Problems related to outliers in both the y-direction and x-area

Many methods have been developed to solve these problems. One of these methods is the least trimmed squares method with a large breakdown point defined by Rousseeuw.

3.7.1 Least Trimmed Squares Method

The least trimmed sum of squares (LTS) method is one of the methods used for robust regression (Kılınç B.K. and Alpu Ö., 2015). There are many algorithms used to calculate the LTS estimate in the literature. Ruppert & Carol (1980), Neykov & Neytchev (1991), Tichavsky (1991), Atkinson & Weisberg

(1991), Ruppert (1992), Stramberg (1993), Hawkins (1994), Rouseeuw & Driessen (2006) generate these algorithms (Kılınç and Alpu, 2015).

The objective function for the LTS method is defined as follows;

$$\text{Min}_{\beta} \sum_{i=1}^h (e^2)_{i:n}$$

h is the number of good data values. The LTS estimator trims $(n-h)$ data with the largest residuals from the data set and provides a robust estimate (Kılınç ve Alpu, 2015). If $h = n$, the LTS estimator is equal to the ordinary least squares estimator. If $h = n / 2$ is selected, an estimator with high (50%) breakdown point can be obtained. However, under normal conditions, sometimes a high breakdown point can lead to poor results (Al-Atari and Al-Amleh, 2016). On the other hand, selecting $h = (n + p + 1) / 2$ increases the efficiency of the estimator. After determining the number of h , all possible subsets are determined. The regression coefficients are then calculated. With these coefficients, the squares of the remains are listed as $e_{(1)}^2 \leq \dots \leq e_n^2$ and the largest residual squares are excluded from the total. This ensures that outliers are completely excluded. This method can be very effective depending on the value h and the construction of outliers. If the all outliers are excluded correctly, this method is equivalent to the least squares method. However, this method is not effective if there are more outliers than trimmed ones. Conversely, if the data is trimmed other than the outliers, this leads to some good data being excluded from the calculation (Alma, 2011).

The regression equation for foreign trade data is established as follows;

$$\widehat{BRF1} = \beta_0 + \beta_1 \text{quantity1} + \beta_2 \text{dolar} + e$$

$$\widehat{BRF2} = \beta_0 + \beta_1 \text{quantity2} + \beta_2 \text{dolar} + e$$

These equations will be used in order to detect outlier in foreign trade data for variables BRF1 and BRF2.

3.8 Criteria for Comparison of Outlier Detection Methods

- 1-Masking-swamping effect
- 2- Skewness of data set (data with high skewness coefficient)
- 3- Large, small number of records

Masking-swamping effects: Data sets with more than one outlier are exposed to masking and swamping effects. For these effects, although not mathematically qualified, Acuna and Rodriguez provided intuitive understanding (Acuna and Rodriguez, 2004) . According to the definitions;

Masking effect is that one outlier masks the other one. Therefore, in case of deletion of the first outlier value, the 2nd observation value is detected as an outlier. Masking occurs when a group of outlying points skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

Swamping effect is the detection of a regular observation value as an outlier due to an outlier. In other words, when the first outlier is deleted, the regular observation is detected as a normal observation. The swamping effect occurs when a group of outliers skewed the mean and covariance estimates towards it. As a result, regular observations are moving away from the mean. The distance of good observations to the mean is large enough to cause them to appear as outliers.

Methods of detecting outliers shall be compared to whether they are sensitive to these two effects. It is of great importance that methods are not exposed to these effects. Because the outlier detection methods exposed to these effects cannot accurately detect the outliers. Consequently this criterion is the most important and before applying the outlier detection method, it should be tested whether the method is exposed to these effects.

Skewness of data: In the case of high skewness, outlier detection methods cannot detect outliers that are in the tail of data at the opposite of skewness direction. Therefore, in the case of high skewness, it is desired to examine the results obtained by using different outlier detection methods. So the high degree of skewness is another criterion to be considered when comparing the methods. After the masking and swamping effects, the most important criterion to be considered in determining the appropriate outlier detection methods is this criterion so the methods should not be affected by the skewness of the data. The method which is not affected by the skewness of the data can detect outliers in both tail of the data.

Large,small number of data: One of the criteria for comparing outlier detection methods is the large or small number of observations. There are outlier detection methods that cannot detect an outlier in data with small number of observations so it is desired to investigate whether there are any methods which can detect outliers in cases where the number of observations is small. Similarly, it should also be checked whether the detection methods are working effectively in data sets where the number of observations is too large.

3.9 Sample Selection

In this work, outlier detection methods are compared by an implementation. For the implementation, import data for the last 6 months of 2015 that were taken from The Ministry of Commerce is used. The following features were taken into consideration when selecting outlier detection methods to be used in the implementation;

1- Foreign trade data is a very large data set that does not provide statistical assumptions such as normal distribution.

2- The primary objective of the detection of outliers in foreign trade data is not to detect the outliers that violate the distribution assumptions of basic statistical analyzes such as t-test, ANOVA and regression but rather to detect the outlier values that are located away from the majority of the data set regardless of the distribution.

Considering these features, the methods described above are selected for the implementation. A representative example of the data set to be implemented is given as follows.

A Comparison of the Outlier Detecting Methods: An Application on Turkish Foreign Trade Data

	GTIP Code of Product	GTIP NAME Definition of product	COUNTRY	MODE OF TRANSPORT	BRF1 Unite price1	BRF2 Unite price2	M1/M2 quantity ratio
Group 1	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
	X	CAR	GERMANY	ROAD	.	.	.
Group 2	X	CAR	FRANCE	SEA	.	.	.
	X	CAR	FRANCE	SEA	.	.	.
	X	CAR	FRANCE	SEA	.	.	.
	X	CAR	FRANCE	SEA	.	.	.
	X	CAR	FRANCE	SEA	.	.	.
	X	CAR	FRANCE	SEA	.	.	.
.	
.	
.	
Group n	T	APPLE	ITALY	AIR	.	.	.
	T	APPLE	ITALY	AIR	.	.	.
	T	APPLE	ITALY	AIR	.	.	.
	T	APPLE	ITALY	AIR	.	.	.
	T	APPLE	ITALY	AIR	.	.	.

Figure 2. Foreign trade data set

As shown in Figure 2, data are grouped by taking product, country, mode of transport variables into consideration. The lines in each group represent the foreign trade register for the product. The variable BRF1 denotes unit price in KG. The BRF2 variable shows the unit price calculated in supplementary unite. The M1 / M2 variable shows the ratio of the quantity value of KG to supplementary unite like square meters, cubic meters, pieces etc. For example, the car is registered both in KG and in number. Accordingly, BRF1, 1 KG car is how many dollars, BRF2 1 car is how many dollars, M1 / M2 means how many KG is a car. As a result, the methods of detecting outliers were applied to each group separately for BRF1, BRF2 and M1 / M2 variables.

Later, since the foreign trade data set is a very large set of data, it has been simplified to allow comparison of the methods in accordance with the criteria. The data set is simplified by arbitrary sampling. In Figure 2.2 below, the sampling steps and sample data set are given in a summary form.

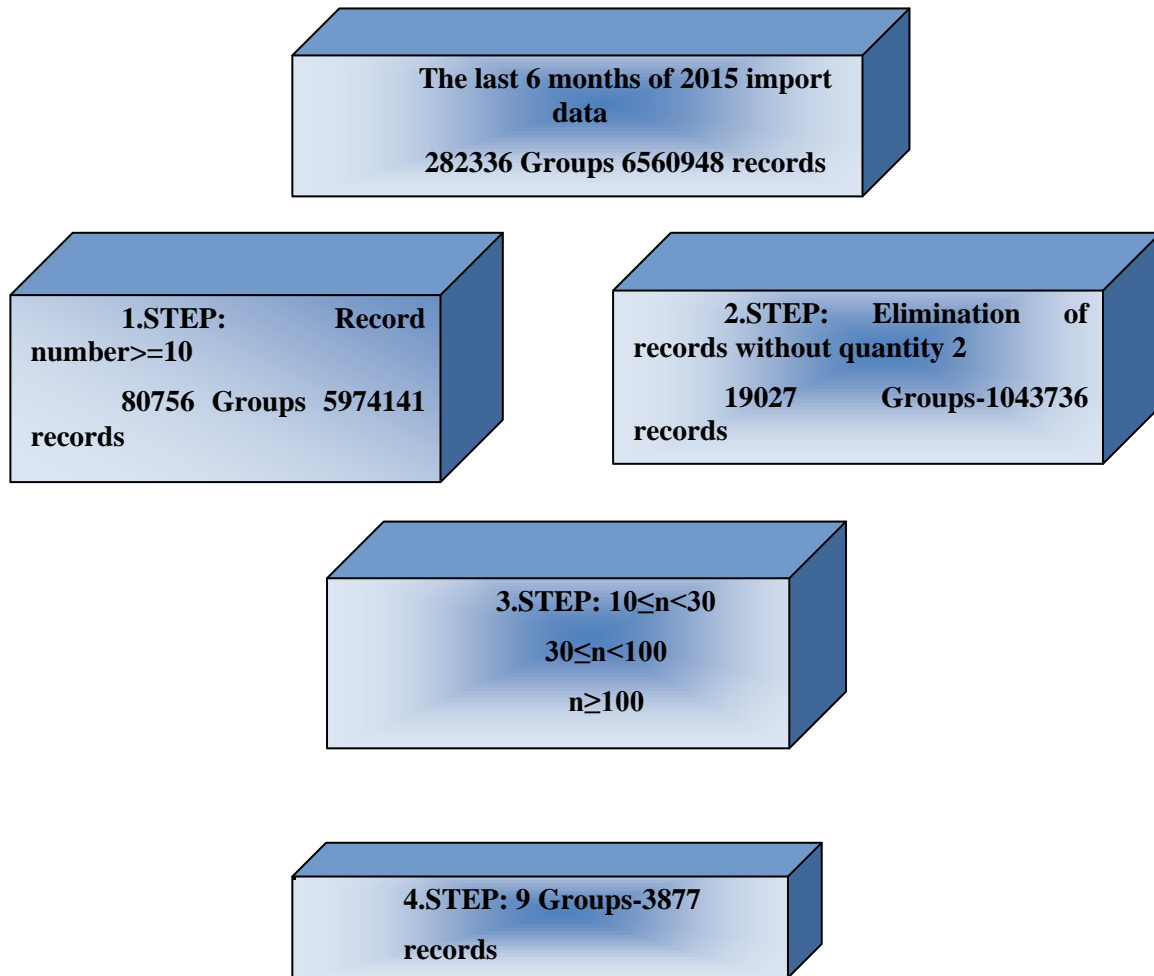


Figure 3. Sampling steps

The first cube in the figure 3 shows the number of groups and records of the data set before it was simplified. In the first step of the sample selection, groups with more than 10 and 10 records were identified. According to this, there are 80756 groups and 5974141 records. In the second step, groups without knowledge of quantity 2 were removed from the data set. After this operation, the group and records numbers are 19027 and 1043736 respectively. In step 3, the data is grouped by considering the interval defined in the cube. $10 \leq n < 30$, $30 \leq n < 100$ and $n \geq 100$ denote small, medium and big sample data respectively. Then, considering each interval, the coefficients of skewness were calculated for each of the BRF1, BRF2 and M1 / M2 variables in each group. Calculated skewness coefficients are sorted from large to small. Then, the groups that have the largest and the smallest and equal to zero coefficients were determined for each interval. At the end of this, 9 groups and 3877 records were selected for implementation.

4. Results

The implementation was made using SAS Enterprise Guide and SQL programming included in this program. In this study some criteria were taken into consideration while comparing the methods of detecting outliers. With these criteria, answers to the following questions were searched;

1- Is the outlier detection method affected by masking and swamping?

2- Is the outlier detection method affected by the data with a high degree of skewness?

3- Does the outlier detection method not work or is ineffective in the case of data with a large number of records?

4- Does the outlier detection method not work or is ineffective in the case of data with a small number of records?

5-Does the outlier detection method have the application difficulty for the foreign trade dataset?

In order to find out the answer to these questions, the above-mentioned methods of outlier detection were applied individually to the foreign trade data set. Then the outliers that each method detected from the data set in the application result were examined. By taking the criteria into consideration, it was determined whether the detected outliers were correctly or incorrectly detected, and the outlier value percentages were calculated accordingly. It has been determined that the method of detecting outliers is subject to masking and swamping effects by looking at the existence of outliers that are detected incorrectly. It was then checked whether each method could detect outliers on the left and right tail of the data set used in the application. As a result of this check, it was also revealed whether the methods of outlier detection are affected by the skewness of the data. In the next step, the number of records of the data groups used in the application were taken into consideration to determine whether the methods of outlier detection are working effectively. According to this, the outlier detection methods are discussed one by one and the results of some applications are given below.

Table 2. Results of z-score method

GTP	COUNTRY	MODE OF TRANSPORT	BRF1	ZBRF1	OUTLIER	GTP	COUNTRY	MODE OF TRANSPORT	BRF1	ZBRF1	OUTLIER
220870100000	1	30	1619.00	9.79	1	220870100000	1	30	50.50	5.97	1
220870100000	1	30	50.50	0.14	0	220870100000	1	30	28.43	2.60	0
220870100000	1	30	28.43	0.00	0	220870100000	1	30	27.50	2.46	0
220870100000	1	30	27.50	0.00	0	220870100000	1	30	27.33	2.44	0
220870100000	1	30	27.33	0.00	0
.
.
.
220870100000	1	30	4.25	0.14	0	220870100000	1	30	4.25	0.14	0
220870100000	1	30	4.17	0.15	0	220870100000	1	30	4.17	0.15	0
220870100000	1	30	3.82	0.15	0	220870100000	1	30	3.82	0.15	0
220870100000	1	30	3.63	0.15	0	220870100000	1	30	3.63	0.15	0

A Comparison of the Outlier Detecting Methods: An Application on Turkish Foreign Trade Data

In Table 2, one of the results obtained after applying the z-score method is given. This example contains information on the liquor product imported from France via the code of 220870100000 GTIP (Customs Tariff Statistics Position). According to this result, the first table shows the outliers for the BRF1 variable. According to the first table, the highest value of 1619 was detected as an outlier, and the next 50.5 value could not be detected as an outlier. After the largest BRF1 value was deleted, the z-score method was applied to the rest of the data set then the value of 50.5 that could not be detected as an outlier in the first application was detected as an outlier in the second table. Therefore, it is possible to say that the z-score method is exposed to masking effect.

Table 3. Results of Box-Plot and Adjusted Box-Plot Method

Box-Plot Method							D-Box-Plot Method				
GTIP	COUNTRY	MODE OF TRANSPORT	MIN2	UPPER BOUNDARY	LOWER BOUNDARY	OUTLIER	GTIP	COUNTRY	MODE OF TRANSPORT	MIN2	OUTLIER
401120900011	18	17	75.00	77.99	35.05	0	401120900011	18	17	75.00	1
401120900011	18	17	69.28	77.99	35.05	0	401120900011	18	17	69.28	1
401120900011	18	17	68.45	77.99	35.05	0	401120900011	18	17	68.45	1
401120900011	18	17	63.12	77.99	35.05	0	401120900011	18	17	63.12	0
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
401120900011	18	17	49.55	77.99	35.05	0	-	-	-	-	-
401120900011	18	17	49.12	77.99	35.05	0	401120900011	18	17	49.12	0
401120900011	18	17	47.82	77.99	35.05	0	401120900011	18	17	47.82	0
401120900011	18	17	47.43	77.99	35.05	0	401120900011	18	17	47.43	0
401120900011	18	17	34.10	77.99	35.05	1	401120900011	18	17	34.10	1

In Table 3, one of the results were obtained after the box-plot method was applied This example contains information on the tire for truck and bus imported from Luxembourg via the GTIP code 401120900011. The data set in the table 3 is is skewed to the left. After the box-plot method was applied to this data set, according to the table on the left in table 3 , it is seen that for the M1 / M2 variable, only outliers in the left tail of the data set were detected. The table on the right shows the results obtained after the adjusted box-plot method was applied to the same data set. Accordingly, when the adjusted box-plot was applied, it was observed that the outliers in both tails of the data set could be detected. According to these results, it is possible to say that the Box-plot method is affected by the skewness of the data set.

Table 4. Results of Adjusted Box-Plot

GTIP	COUNTRY	MODE OF TRANSPORT	M1M2	OUTLIER	GTIP	COUNTRY	MODE OF TRANSPORT	M1M2	OUTLIER
902219000000	720	10	34997	1	902219000000	720	10	34997	1
902219000000	720	10	1095	1	902219000000	720	10	1095	1
902219000000	720	10	738.75	1	902219000000	720	10	738.75	0
902219000000	720	10	578	0	902219000000	720	10	578	0
902219000000	720	10	514.333	0	902219000000	720	10	514.333	0
902219000000	720	10	462	0	902219000000	720	10	501	0
902219000000	720	10	450	0	902219000000	720	10	500	0
902219000000	720	10	434.091	0	902219000000	720	10	485	0
902219000000	720	10	425	0	902219000000	720	10	471	0
902219000000	720	10	408.333	0	902219000000	720	10	462	0
902219000000	720	10	400	0	902219000000	720	10	461	0
902219000000	720	10	400	0	902219000000	720	10	450	0
902219000000	720	10	400	0	902219000000	720	10	450	0
902219000000	720	10	400	0	902219000000	720	10	434.0909	0
902219000000	720	10	150	0	902219000000	720	10	425	0
902219000000	720	10	150	0	902219000000	720	10	410	0
902219000000	720	10	6.5	0	902219000000	720	10	408.3333	0
902219000000	720	10	0.42343	0	902219000000	720	10	405	0
902219000000	720	10	0.02825	0	902219000000	720	10	402	0
					902219000000	720	10	401	0
					902219000000	720	10	401	0
					902219000000	720	10	400	0
					902219000000	720	10	400	0
					902219000000	720	10	400	0
					902219000000	720	10	400	0
					902219000000	720	10	376	0
					902219000000	720	10	375	0
					902219000000	720	10	371	0
					902219000000	720	10	370	0
					902219000000	720	10	365	0
					902219000000	720	10	360	0
					902219000000	720	10	310	0
					902219000000	720	10	255	1
					902219000000	720	10	155	1
					902219000000	720	10	150	1
					902219000000	720	10	150	1
					902219000000	720	10	140	1
					902219000000	720	10	125	1
					902219000000	720	10	6.5	1
					902219000000	720	10	0.423427	1
					902219000000	720	10	0.02825	1

In Table 4, one of the results obtained after adjusted box-plot method was applied. This example contains information on the product of other X-ray devices imported by sea from China via the GTIP code 902219000000. For the M1 / M2 variable, outliers are detected. The number of records in the left table is 18 in table 4. According to the results in this table, it was observed that the adjusted box-plot method was not effective. Because only outliers in the right tail of the data could be detected. In order to show that this situation is due to the small number of records in the data set, the data set was expanded by using regular observations and the method was re-applied. The outlier obtained by the extended data set are shown in the table on the right in table 4. When the number of records was increased, it was observed that the remaining outliers in both tails of the data set were detected. Accordingly, it is possible to say that the adjusted box-plot method does not work effectively when the number of records is small.

The other methods were treated in a similar way and the results of the application were evaluated as above. As a result of all these operations, the following table is formed and the cases of outliers are presented in a summary by comparison criteria.

Table 5. The status of the outlier detection methods according to the criteria

Methods	Masking-Swamping	Skewness	Large number of record	Small number of record	Application difficult	Percentage of success
Adjusted Box-plot	NO	NO	NO	YES	NO	80%
Median z-score	NO	YES	NO	YES	NO	60%
Box-Plot	NO	YES	NO	YES	NO	60%
Z-score	YES	YES	NO	YES	NO	40%
B-H Methods	YES	NO	NO	YES	NO	40%
K-means	YES	NO	YES	NO	YES	40%
Robust regression	YES	NO	NO	YES	NO	60%

The criteria for masking and swamping effect are the most important benchmarks. Because the method exposed to this effect can not accurately detect outliers. Therefore, the methods that is exposed to this effect are shown in red in the table and eliminated in the first place. Then the methods that are affected by the skewness of the data were determined and these methods eliminated because if the methods exposed to skewness, they can not detect outliers in both of tails of data. Later, it was found that the methods were working effectively when there was a large number of records or a small number of records. Then methods evaluated whether has the ease of implementation in the foreign trade data set. At the end of, success percentages were calculated according to criteria. The evaluations of each method and similar findings in the literature are given below.

Z-score Method

1- Because of using mean and standart deviations that are sensitive to outliers, z-score method exposes to masking effect. In his 2006 study, Seo gave examples of how the z-score method was exposed to masking (Seo,2006).

2- When the number of data is less than 10, z-score method cannot detect the outliers

Consequently z-score method is not proper method for foreign trade data set

B-H Methods

1- As previously mentioned B-H method exposes to masking and swamping effect. Ishikawa, Endo and Shiratori have shown that the B-H method is subjected to masking and swamping effects in their studies (Ishikawa, Endo, Shiratori, 2010, p.17)

2- When the number of data is equal to or less than 10, B-H method cannot detect the outliers

Consequently B-H method is not proper method for foreign trade data set

k-Mean Clustering

1- The performance of the method depends on the capacity of the clustering variables that will distinguish between the observation values. Furthermore, this method does not work effectively when R^2 is less than 50% and the number of observations is less than 15.

2- The k-mean algorithm is sensitive to outliers and outliers have a disproportionate effect on the final cluster configuration. This causes the observation values, which are actually masked by clustering, not to be detected as outliers or some regular values are detected as outliers (Bokare and Thakare, 2016). Consequently k-mean clustering exposes to masking and swamping effect

3- The k-mean clustering method does not work properly in data sets with a large number of observations (Marghny and Taloba, 2011).

4- In the literature, it is discussed whether it is appropriate to use clustering algorithms for the detection of outliers. For example, Zhang and Whang reported that clustering algorithms should not be taken into account as outlier detection methods (Bokare and Thakare, 2016).

5- A further disadvantage of the k-average clustering method is the determination of the appropriate number of clusters. Since the appropriate number of cluster for different observation number varies, the standart number of clusters cannot be determined. Determining the number of clusters for each data requires another analysis on its own. Therefore, it is not appropriate to use k-mean clustering method for the foreign trade data set which has very different number of observations.

Due to the above mentioned features k-mean clustering method is not proper method for foreign trade data set.

Robust Regression Method

1- One of the robust regression methods, the LTS method, needs the number of h which will provide the maximum reduction in the sum of the residual squares after to be deleted from the data set. h denotes the number of the largest outliers numbers. Unfortunately, the number of h is not generally known (Zioutas, Avramidis and Pitsoulis, 2007). If the number h cannot be determined correctly, there are two possibilities. Accordingly, this method is ineffective if there are more outliers than those that are trimmed in the data. Conversely, if the data is truncated other than the outliers, this also leads to some good data being excluded from the calculation. Therefore, this method may be subject to masking and swamping effect (Zioutas, Avramidis and Pitsoulis, 2007).

2- The regression model established with the variable to be detected as an outlier should be meaningful. A system where the most appropriate regression model is not installed is not working properly.

3- In addition, it was observed that the LTS method did not detect an outlier when the median absolute deviation (MAD) of any of the variables used in the regression model was zero.

As a result robust regression method is not proper method.

Box-Plot Method

1-Box-Plot method is a method that does not need any distribution assumption. It also uses quartiles so it is not affected by the presence of outliers. Therefore, it is not exposed to masking or swamping effect.

2-In cases where the number of observations is small, it is not able to detect an outlier. Seo gave an example for this in his work (Seo,2006).

3- Box-Plot is affected by the skewness of the data. In skewed data, it is not able to detect outliers in the opposite direction of skewness while detecting a large number of outliers in the direction of skewness.

Due to the above mentioned features Box-Plot method is not proper method for foreign trade data set.

Median Z-Score Method

1- Median z-score method is not affected by outlier values compared to z-score method. Therefore, it is not exposed to masking and swamping effect. However, as in the z-score method, it works more efficiently in a data with normal distribution.

2- In addition, it is not able to detect an outlier in the data with a small number of observations. The efficiency of the method increases as the number of observations increases.

3- In the non-symmetrical datasets, it is not possible to detect the outlier in the opposite direction of skewness and it can detect the outlier in the direction of skewness effectively. Therefore, it can detect outliers in a data set with symmetric and normal distribution more effectively.

Adjusted Box-Plot

1- The adjusted Box-Plot method is an outlier detection method that does not require any assumption, is not exposed to masking or swamping effect, has ease of application and is not affected by the skewness of data.

2- However, it has been observed that it cannot effectively detect outliers in data with small number of records.

3- Hubert and Vandervieren warn that the adjusted Box-Plot method does not yet take into account tail weights when considering the skewness (M. Hubert, E.Vandervieren, 2007). For this reason, in some of the data with high skewness due to the weight effect, the adjusted Box-Plot method is not able to effectively detect the outliers in the direction of skewness while detecting outliers in the opposite direction of the skewness direction.

As a result, it is concluded that the most appropriate and most accurate methods of detection are the adjusted Box-Plot and median Z-score methods for the foreign trade data set, considering the benchmarking criteria and the percentage of undetected outliers. Since these two methods have been observed to complement each other's weaknesses, it has been determined that these two methods should be applied to the data set together in order to detect the correct outlier in the data set.

5. Conclusions

As previously stated, outliers are observations that are not well-defined in normal behavior in the data set. Outlier can occur almost in every real data set. One of these data sets is the foreign trade data set used to calculate the foreign trade statistics. In order to calculate the foreign trade statistics in a way that reflects the truth, it is of great importance that outliers are detected in the data set effectively.

In this study, seven different outlier detection methods were considered and some criteria were used to determine the most proper outlier detection methods in the foreign trade data set. The outlier detection methods in the study do not represent all outliers in the literature. The methods that are suitable for the structure of foreign trade data set and serve the purpose of detection outliers were used. Each method is effective under certain conditions, but the foreign trade data set does not provide these conditions, it was observed that they are not working effectively in this data set. The choice of the best method varies according to the type of data to be applied and the purpose of detection outliers. For this reason, it is recommended to follow the following steps when determining the outlier detection method.

1- First, characteristics of data should be determined. Questions should be investigated, such as whether the data is symmetrical or skewed, the number of records is small or large or has the normal distribution or not.

2- The purpose of detection outliers must be determined. What is the primary objective of the detection of outliers? Is it to detect the outliers that violate the distribution assumptions of basic statistical analyzes such as t-test, ANOVA and regression or to detect the outlier values that are located away from the majority of the data set regardless of the distribution.

After determining the structure of the data set and the purpose of detection the outliers, the appropriate methods in the literature should be investigated and selected.

The views expressed are of the authors and do not reflect those of TURKSTAT.

References

- [1] Acuna E., Rodriguez C. A. (2004), "Meta analysis study of outlier detection methods in classification", Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrived from academic.uprm.edu/eacuna/paperout.pdf. ,In proceedings IPSI 2004, Venice, 2004.
- [2] Arribas I. (2012). Analysis of Three Methods of Detection of Important Influential Data Errors in ITG. Luxembourg: European Commission EUROSTAT.
- [3] Alma, Ö. (2011). "Comparison of Robust Regression Methods in Linear Regression". *Int. J. Contemp. Math. Sciences*, 6(9), 409 - 421.
- [4] Belcher R. (2003). "Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys". SSC Annual Meeting, June 2003, Proceedings of the Survey Methods Section.
- [5] Bokare M., Thakare V.M., (2016). "Outlier Discernment Using K-Means". *GE- International Journal of Engineering Research*, 4(12), ISSN(O): 2321-1717, ISSN(P): 2394-420.
- [6] Chen C. (2002). "Robust Regression and Outlier Detection with The ROBUSTREG Procedure". SAS Institute Inc., Cary, NC, Paper 265-27.
- [7] Ishikawa A., Endo S., Shiratori T. (2010). " Treatment of Outliers in Business Surveys: The Case of Short-term Economic Survey of Enterprises in Japan (Tankan)", Bank of Japan 2-1-1 Nihonbashi-Hongokucho, Chuo-ku, Tokyo 103-8660.

- [8] Kılınç B.K., Alpu Ö. (2015). "Combining Some Biased Estimation Methods with Least Trimmed Squares Regression and its Application". *Revista Colombiana de Estadística*, 38(2), 485-502.
- [9] Kuppasamy M., Kaliyaperumal K. (2013). "Comparison of Methods for Detecting Outliers". *International Journal of Scientific and Engineering Research*, Volume 4, Issue 9, ISSN 2229-5518.
- [10] Marghny M.H., Taloba A.I. (2011). "Outlier Detection Using Improved Genetic K-means". *International Journal of Computer Applications (0975-8887)*, 28(11), 33-36 NC.
- [11] Penny K.I., Jolliffe I.T. (2001). "A Comparison of Multivariate Detection Methods for Clinical Laboratory Safety Data". *Journal of The Royal Statistical Society: Series D (The Statistician)*, Volume 50, Issue 3, 295-307.
- [12] Seo S. (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*, Master Thesis, Pittsburgh University, Pensilvanya, 1-39.
- [13] Vanderviere E., Huber M. (2007), *An Adjusted Boxplot for Skewed Distributions*.
- [14] Yorulmaz Ö., (2009). "Robust Regression Method and Diagnose Of Outliers on Several Social Data". *Balıkesir University Social Science Institute Journal*, 12(21), 76-88.
- [15] Zioutas G., Avramidis A., Pitsoulis L., (2007). "Penalized Trimmed Squares and A Modification of Support Vectors for Unmasking Outliers in Linear Regression". *REVSTAT – Statistical Journal*, 5(1), 115-136.