



## Outlier Detection and a Method of Adjustment for the Iranian Manufacturing Establishment Survey Data

Zahra Rezaei Ghahroodi<sup>1\*</sup>, Taban Baghfalaki<sup>2</sup>, and Mojtaba Ganjali<sup>3</sup>

<sup>1</sup>Statistical Research and Training Center, Tehran, Iran

[z\\_rezaei@src.ac.ir](mailto:z_rezaei@src.ac.ir)

<sup>2</sup>Department of Statistics, Tarbiat Modares University, Tehran, Iran

[t.baghfalaki@modares.ac.ir](mailto:t.baghfalaki@modares.ac.ir)

<sup>3</sup>Department of Statistics, Faculty of Mathematical Science

Shahid Beheshti University, Tehran, Iran

[m-ganjali@sbu.ac.ir](mailto:m-ganjali@sbu.ac.ir)

Received: August 03, 2014; Accepted: May 19, 2015

### Abstract

The role and importance of the industrial sector in the economic development necessitate the need to collect and to analyze accurate and timely data for exact planning. As the occurrence of outliers in establishment surveys are common due to the structure of the economy, the evaluation of survey data by identifying and investigating outliers, prior to the release of data, is necessary. In this paper, different robust multivariate outlier detection methods based on the Mahalanobis distance with blocked adaptive computationally efficient outlier nominators algorithm, minimum volume ellipsoid estimator, minimum covariance determinant estimator and Stahel-Donoho estimator are used in the context of a real dataset. Also some univariate outlier detection methods such as Hadi and Simonoff's method, and Hidiroglou-Barthelot's method for periodic manufacturing surveys are applied. The real data set is extracted from the Iranian Manufacturing Establishment Survey. These data are collected each year by the Statistical Center of Iran using sampling weights. In this paper, in addition to comparing different multivariate and univariate robust outlier detection methods, a new empirical method for reducing the effect of outliers based on the value modification method is introduced and applied on some important variables such as input and output. In this paper, a new four-step algorithm is introduced to adjust the input and output values of the manufacturing establishments which are under-reported or over-reported. A simulation study for investigating the performance of our method is also presented.

**Keywords:** Robust multivariate outlier detection; Sampling weight; Winsorization; Mahalanobis distance; Under-reported and Over-reported outliers

MSC 2010 No.: 62P30, 91C99

## 1. Introduction

Manufacturing sector is one of the principal components in the Economic Development Plans. To assess and realize the goals, determined for the manufacturing sector, availability of updated and accurate statistics is essential. Like all statistical surveys, manufacturing establishment survey is subject to measurement errors, including sample and non-sample errors. These measurement errors affect the accuracy of the published statistics.

As outliers are common in every data set in any application such as establishment surveys, identification and correction of outliers are important objectives of survey processing which should be carried out by statistical centers. Many researchers, working with establishment sample surveys, often encounter observations that differ substantially with the bulk of the observations in the sample. This increases the possibility of anomalous data and makes their detection more difficult. Outliers are so unlike or divergent values from the rest of data and ignoring them or considering them in a usual manner can lead to inaccurate survey estimates. Outliers can occur due to errors in the data gathering process or they may be valid measurements. In the former case, data are non-representative (which can be regarded to be unique in the population) and in the latter case, these valid values are referred to representative outliers (which cannot be regarded to be unique in the population, Chambers, 1986).

A common class of such errors is errors in writing out the response, method of choosing samples, misunderstanding of type of unit (e.g. thousands of pounds instead of a single pound) or misunderstanding of the question, which results in an erroneous response. The standard approach for solving these kinds of problems is to use a large number of edits during survey processing. However, sometimes outliers couldn't be identified. Sometimes a correct response can be an outlier. The causes of having outliers in this situation can be related to the method of choosing samples or because of large change in reported values due to a time lag between the time to draw samples drawn and the time these samples are used.

There are different methods for outlier detection. One of the classifications of outlier detection methods is the division of methods to univariate approach (Andrews and Pregibon, 1978) or multivariate approach. Another fundamental classification of outlier detection is the use of a parametric method or a nonparametric method. One of the non-parametric methods is distance-based method. A classical way of identifying multivariate outlier is based on the Mahalanobis distance method. In order to avoid the masking effect, robust estimates of location and scatter parameters are considered. Many methods assume that the data follow some elliptical distribution and they try to estimate the center and the covariance matrix robustly. Then, they use a corresponding Mahalanobis distance to detect outliers.

Many monthly, quarterly and annually manufacturing or business surveys in different countries, such as Monthly and Annual Business Survey (MBS and ABS) in UK and Monthly Survey of Manufacturing (MSM) in Statistics Canada, use different ways of outlier detection and treatment methods. For instance in UK, outliers in MBS are detected automatically and treated by winzorisation or in ABS, the businesses with extreme or a typical value, compared with other businesses in their Standard Industrial Classification (SIC) and employment size, are treated as outliers and post-stratification methods are used for treating them (Office for national statistics,

2011). The Monthly Survey of Manufacturing (MSM) in Statistics Canada performs outlier detection shortly after collecting them by calculating Mahalanobis distance, where the mean vector and covariance matrix are robustly estimated using modified Stahel-Donoho estimates proposed by Patak (1990).

There is a large literature on outlier detection. Many methods for the detection of multiple outliers use very robust methods to split the data into a clean part and the potential outlier part. For example in multivariate data, Rousseeuw and van Zomeren (1990) proposed a method to find the subset of observations within a minimum volume ellipsoid (MVE) as non-outlier data. Rousseeuw and van Driessen (1999) proposed finding the subset of observations with the minimum covariance determinant (MCD). Another option is the forward search method introduced by Hadi (1992a), and Hadi and Simonoff (1993). The basic idea of this method is to identify a clean subset of the data, defined from a robust method, and to include more clean observations until only the outlying units remain out. This method rapidly leads to the detection of multiple outliers. All multiple outlier detection methods suffer from a computational cost that escalates rapidly with the sample size. Billor et al. (2000) proposed a new general approach titled as BACON (Blocked Adaptive Computationally efficient Outlier Nominators) algorithm, based on Hadi (1992b) and Hadi and Simonoff (1993), which can be computed quickly regardless of the sample size. Beguin and Hulliger (2008) proposed the BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Also Hidiroglou and Berthelot (1986) proposed a non-parametric method of outlier detection for periodic manufacturing or business surveys that is a revised version of quartile method.

Since outliers influence the estimates of the population and results of any statistical approach may change greatly depending on how outliers are treated, therefore choosing the best method of treating outliers is necessary. There are several methods of treating outliers that can be classified into three categories “weight modification”, “value modification” and “combination of weight and value modification”, Ishikawa et al. (2010).

Since Iranian Manufacturing Establishment Survey (IMES) data set, the same as all statistical surveys is subject to measurement errors and these measurement errors affect the accuracy of the published statistics, in this paper we concentrate on outlier detection and propose a new empirical method for reducing the effect of outliers. In this empirical method, a new four-step algorithm based on the value modification method is introduced for adjustment of detected outliers in estimating the population parameters of interest in IMES data by identifying the manufacturing establishments which under-report or over-report their input and output variables. In this new empirical algorithm, the over-reported manufacturing establishments are adjusted downwardly and the under-reported manufacturing establishments are adjusted upwardly. Also, some simulation studies are performed for investigating the performance of the proposed adjustment approach, the effect of different sample sizes of manufacturing establishments and the number of outliers in them.

## **2. Description of IEMS**

In order to identify the industrial structure of the country to provide information needed for planning on industrial development, to assess the results of these plans and to formulate the

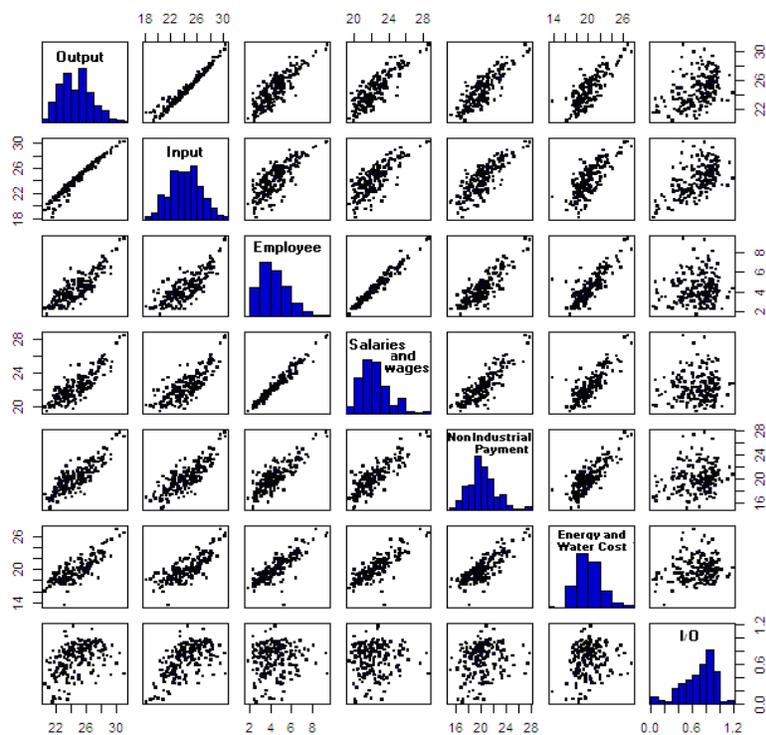
proper economic policies; the Statistical Centre of Iran has implemented the survey on Manufacturing Establishments from 1972. It is obvious that annual data collection is done after finalizing the financial accounts of manufacturing establishments; thus, in this survey, the data of preceding year are collected.

The country's first manufacturing survey was launched in 1963 by the former General Department of Public Statistics. In 1964 to 1972, the Ministry of Economics conducted other manufacturing surveys. It kept on performing the job up to 1973 when the Ministry of Industries and Mines took the duty over. The Statistical Centre of Iran (SCI) launched the first survey of Large Manufacturing Establishments (with over 10 or more workers) in 1972, which has annually repeated the job. In 1997 and 2002, the SCI conducted the General Census of Manufacturing and Mines (GCMM) and the General Census of Establishments, respectively, to collect a frame data set for the nation's economic activities and household's activities.

The target population for this survey includes all manufacturing businesses operating in Iran. The objective of this annually survey is to collect economic data required for compiling National Accounts and, in details, to estimate input value, output value and value added. In this survey which is undertaken by the SCI, the information is collected directly based on face-to-face interview with officer or director of statistical units of the manufacturing establishment. Sampling frame is the list of manufacturing establishments obtained from General Census of Establishments in 2002 and is annually updated. Survey method is used for the complete enumeration of large manufacturing establishments with 50 or more workers and for other establishments with 10-49 workers. It should be mentioned that in the survey, the data on manufacturing establishments with 10-49 workers for some provinces were collected by stratified random sampling method and the data related to the remaining provinces as well as manufacturing establishments with 50 and more workers was collected through a census. The survey population is split into ISIC (Iranian Standard Industry Classification) industries and is stratified according to size. The sampling method is stratified random sampling in which the stratification variables are number of workers and economic activity based on ISIC 4-digit Codes. The stratification method is the Dalinus method and the industry classification utilized in the survey is the ISIC, Revision 3 with some changes. In each stratum, sample establishments are selected using systematic method. The level of estimation is whole country, each province and each ISIC code. In this paper, in order to extend the results to whole population, sampling weights of the survey design are considered in calculating interested variables such as Input and output values. These sampling weights are used in methods for univariate and multivariate outlier detection.

In IEMS, questionnaires are sent out to approximately 12500 manufacturing establishments during June and December every year. The main variables collected are total number of male/female employees, laborers such as average number employed directly (manufacturing) and indirectly (non-manufacturing), average number employed by literacy, production and sales or outputs, purchases of raw materials and components to be used in manufacturing process, salaries and wages, detailed information on energy and water costs or expenses, inventory, fixed assets and other cost and receipt from industry and other services.

In our study, first of all, we apply all of the outlier detection methods which will be introduced in the following on one of the ISIC 4-digit Codes (2710), related to the primary production of iron and steel stick. The number of manufacturing establishments in this code for the data collected in 2010 is 194 units. Some variables of interest for identifying outliers are input, output, total number of employees, salaries and wages, non-industrial payment, energy and water costs or expenses and Input-output (I/O, Input divided by Output). Since the variation of the response variables is high we need to use some transformation. One method of choosing suitable transformation is Box-Cox method (1964). After drawing the Box-Cox plot for response variables of interest, the logarithm transformation is chosen for all above-introduced variables except I/O. Scatter plots and histograms for the logarithm of the variables and for I/O are shown in Figure 1. As it is evident in this Figure, some outlier observations exist for some variables. Of course a multivariate approach can better detect these outliers.



**Figure 1:** Scatter plots of the logarithm of the variables and variable I/O with their marginal histogram in the diagonal

### 3. Methodology

Outlier detection methods have been suggested for several applications such as surveys, clinical trials, voting irregularity analysis, data mining tasks, etc. In this paper, our aim is to detect and to control the impact of outliers on the estimators or statistics of manufacturing establishment survey.

There is different classification for outlier detection; one of them is the univariate and

multivariate classification. In univariate cases, Hadi and Simonoff (1993) propose a forward search approach which is an iterative algorithm for multiple regressions based on robust Mahalanobis distance. This algorithm starts with a clean subset of data set and iterates with a sequence of least squares steps. In the final step, the algorithm uses a  $t$ -distribution based on a threshold value for detecting outlier points. In the multivariate classification, the existing methods can be classified into two major families. Many methods suppose that the data follow some elliptical distributions and try to estimate robustly the center and the covariance matrix. Then the corresponding Mahalanobis distance is used to detect outliers. The second class of methods does not rely on a distributional assumption and uses some measures of data depth (Liu et al., 1999). Unfortunately, the later family is often fails to yield methods computationally feasible for analysig large datasets.

Many robust estimators, such as  $M$ -estimators (Huber, 1981), have advantage of being simple but its breakdown point is at most  $1/(p+1)$  where  $p$  is the dimension of the data. Stahel (1981) and Donoho (1982) were the first to define robust multivariate estimators with high breakdown point of one-half for large data sets such as data from official statistics, regardless of the dimensions of the data. Therefore, some approaches such as Stahel-Donoho (SD) estimator (Stahel, 1981; Donoho, 1982) or the Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1985; Rousseeuw and Leroy, 1987), which will be reviewed in the following section, have high breakdown points, but have the disadvantage of being computationally expensive.

An idea from Wilks and Gnanadesikan (1964) is related to the so-called forward search method which is based on the concept of “growing a clean subset of observations”. The idea is to start with a small subset of the data, “clean subset”, and then add non-outlying observations until no more non-outliers are available. The articles of Hadi (1992) and Atkinson (1993) demonstrate the efficiency of such methods. In this method the “clean subset” grows one point at a time using Mahalanobis distances to rank the observations. This method was developed to a more faster and more sophisticated method by Billor et al. (2000) and Kosinski (1999). Billor et al. (2000) proposed a method which is the most robust and the fastest forward search method with complete multivariate normal data. By comparing this method with other Mahalanobis type methods, the performance of BACON on complete data is very promising (Béguin and Hulliger, 2003).

Now we review some univariate and multivariate outlier detections which will be used for our application.

### **3.1. Univariate outlier detection method**

#### **3.1.1. Hadi and Simonoff’s method**

The basic idea in this method is to start with a relatively clean data set of size  $m$  and include observations until the outlying units remain out. In this method, in order to avoid the masking and swamping problems that can occur when there are multiple outliers in a data set, some outlier detection methods are proposed which is location and scale invariant. These methods identify a clean subset of observations of size  $m < n$  that can be presumed to be free of outliers,

and then perform a forward search. They test the remaining points relative to the clean subset and allow the subset to grow one observation at a time as long as the new subset remains clean of outliers. Fitted values generated by this model are then used to generate  $n$  distances to the actual sample data values. The next step redefines the clean subset to contain those observations corresponding to the  $m+1$  smallest of these distances and the procedure is repeated. The algorithm stops when distances to all sample observations outside the clean subset are all too large or when this subset contains all  $n$  sample units (vide, also Hadi and Simonoff, 1993).

### 3.1.2. Hidiroglou and Berthelot method

A very desirable method for detecting outlier in periodic business or manufacturing establishment survey was created by Hidiroglou and Berthelot (1986). In this method an acceptance boundary that varies according to the size of a unit, is chosen. In this method, outliers will be those observations whose ratio ( $r_i$ ) between the current survey and the previous survey differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population. Let

$$r_i = \frac{x_{it}}{x_{it-1}}$$

be the ratio for value of observation of unit  $i$  at the period  $t$  to that of unit  $i$  at the period  $t-1$ . As the distribution of  $r_i$  is non-symmetric, it is difficult to detect outliers from the left tail of distribution. So, this ratio is transformed to  $S_i$  which is defined as

$$S_i = \begin{cases} 1 - \frac{r^M}{r_i}, & 0 < r_i < r^M, \\ \frac{r_i}{r^M} - 1, & r_i \geq r^M, \end{cases}$$

where  $r^M$  is median of  $r_i$  and the distribution of the  $S_i$  is symmetric. In order to consider the size effect of sample data,  $E_i$  is defined as

$$E_i = S_i \{ \max(x_{it-1}, x_{it}) \}^U,$$

where the value of  $U$  is 0 or 1. If  $U=0$ , the size term goes to 1 and if  $U=1$ , the size term will overpower the size term. In this method,  $E_i$  is judged to be an outlier where it is outside of the range  $(E^M - CD_L, E^M + CD_U)$ , where  $D_L$  and  $D_U$  are defined as

$$D_L = \max(E^M - E^{25}, |AE^M|)$$

and

$$D_U = \max(E^{75} - E^M, |AE^M|).$$

$E^{25}$ ,  $E^M$  and  $E^{75}$  are, respectively, the first quartile, median and the third quartile and  $A=0.05$ .

The  $C$  parameter allows us to narrow or widen the acceptance region. The main challenge in applying this method comes in the selection of appropriate values for  $C$  and  $U$  parameters which is not straightforward. In many papers such as Belcher (2003), the values in the range of 0.3 and 0.5 is recommended as suitable values for  $U$ .

### 3.2. Multivariate outlier detection approach

For a  $p$ -dimensional multivariate sample  $x_1, \dots, x_n$ , let  $X = (x_1, x_2, \dots, x_n)'$  be an  $n \times p$  matrix of multivariate data, where  $x_i = (x_{i1}, \dots, x_{ip})$ . The Mahalanobis distance is defined as

$$MD_i = [(x_i - t)' C^{-1} (x_i - t)]^{\frac{1}{2}}, \quad i = 1, 2, \dots, n, \quad (1)$$

where  $t$  and  $C$  are the estimated multivariate location and covariance matrix, respectively. For a multivariate normally distributed data, in the case of large samples, the values of  $MD_i^2$ 's are approximately distributed chi-square with  $p$  degrees of freedom ( $\chi_p^2$ ).

Since the Mahalanobis distance is very sensitive to the presence of outliers and the sample mean and sample covariance may not be adequate as estimators for the center and scatter of  $X$ , it needs to be estimated by a robust procedure in order to provide reliable measures and better to expose the true outliers in the data. It means that  $t$  and  $C$  in (1) have to be estimated in a robust manner. This leads to the so-called robust distance (RD).

Robust multivariate methods provide an almost complete set of estimators for multivariate location and scatter with high breakdown point. The first such estimator was proposed by Stahel (1981) and Donoho (1982) and it is recommended for small data sets, but the most widely used high breakdown estimator is the minimum covariance determinant (MCD) estimate (Rousseeuw, 1985). In the following definitions of different estimators of location and scatter will be briefly reviewed.

#### 3.2.1. Stahel and Donoho estimator

The first multivariate equivariant estimator of location and scatter with high breakdown point was proposed by Stahel (1981) and Donoho (1982). For a data set  $X = \{x_1, \dots, x_n\}$  which represent a set of  $n$  data points in  $\mathfrak{R}^p$ , the weighted mean and covariance matrix are given, respectively, by

$$T_R = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i},$$

and

$$C_R = \frac{\sum_{i=1}^n w_i^2 (x_i - T_R)(x_i - T_R)^T}{\sum_{i=1}^n w_i^2},$$

where  $T_R$ ,  $C_R$  and  $w_i$  are, respectively, the Stahel-Donoho estimator of location, the estimator of scale covariance matrix and the robust weigh. The Stahel-Donoho estimator looks for a univariate projection that makes an observation an outlier. For more details about this method, see Stahel (1981) and Donoho (1982).

### 3.2.2. MVE and MCD estimator

In multivariate data, one of the RD measures is obtained by the minimum volume ellipsoid (MVE) estimator which searches for the ellipsoid of minimal volume containing at least half of the points in the data set  $X = \{x_1, \dots, x_n\}$  (Rousseeuw and Zomeren, 1990). Then the location estimate is defined as the center of this ellipsoid and the covariance estimate is provided by its shape. Although the MVE method is a robust measure to detect the outlying observation, it is computationally expensive because the implementation of this method via resampling needs a lot of different samples to reach good estimates. Since we need to select the ellipsoid with the minimum volume from all  $\binom{n}{h}$  possible combinations from  $n$  observations, even for moderate sample size, it is computationally expensive.

More recently, the minimum covariance determinant (MCD) estimator (Rousseeuw and Driessen, 1999) has been proposed. This is determined by the subset of  $h$  observations,  $x_{i_1}, x_{i_2}, \dots, x_{i_h}$ , whose covariance matrix has the smallest determinant among all possible subsets of size  $h$ . The location estimator,  $T_{MCD}$ , is the average of these  $h$  points, whereas the scatter estimator,  $C_{MCD}$ , is proportional to their covariance matrix as follows,

$$T_{MCD} = \frac{1}{h} \sum_{j=1}^h x_{i_j},$$

and

$$C_{MCD} \propto \frac{1}{h-1} \sum_{j=1}^h (x_{i_j} - T_{MCD})(x_{i_j} - T_{MCD})^T.$$

A recommendable choice for  $h$  is  $[(n+p+1)/2]$ , but any integer  $h$  within the interval  $[(n+p+1)/2, n]$  can be chosen (vide, Rousseeuw and Leroy, 1987).

Finding the MVE or MCD requires computing the volumes of  $\binom{n}{h}$  ellipsoids and choosing the subset which gives the minimum volume or minimum determinant which are computationally infeasible.

All multiple outlier detection methods, which are described in this section, have suffered in the past from a computational cost. This increases rapidly with the sample size. With complete multivariate data, the BACON (Blocked Adaptive Computationally efficient Outlier Nominators) algorithm (Billor et al., 2000) is a new approach based on the methods of Hadi (1992, 1994) that can be computed quickly and yield a robust estimate of the covariance matrix. For more details about this algorithm, see Billor et al. (2000).

In Section 4, these univariate and multivariate outlier detection methods for cross-sectional and periodic surveys are illustrated using our IEMS data.

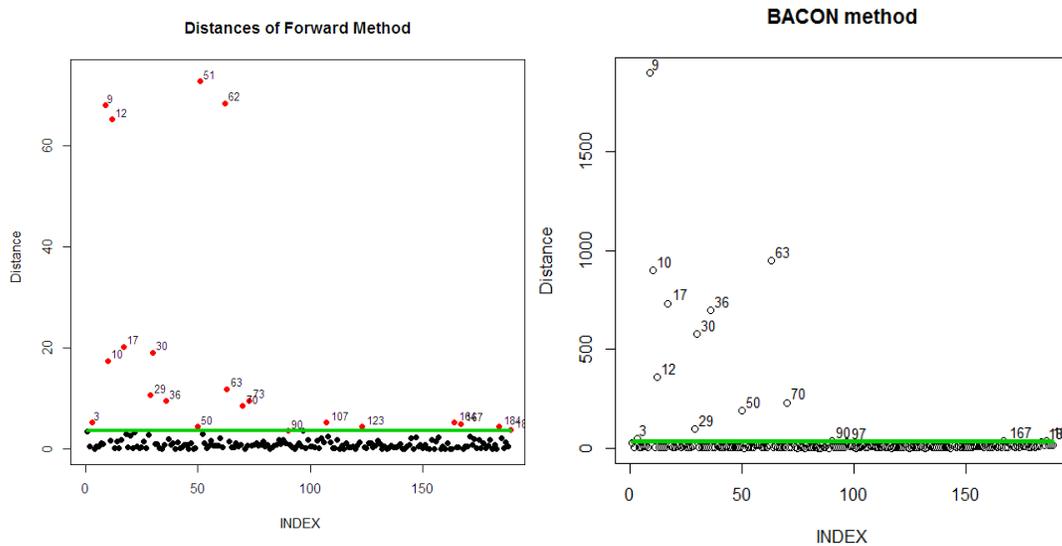
### 4. Results

In our study, first of all, we apply all of the non-periodic presented methods on one of the ISIC 4-digit Codes (2710), which is related to the primary production of iron and steel stick.

Given the linear structure evident in results of Figure 1 between output and other variables of interest, we apply the univariate forward search algorithm described in Section 3. In this method an appropriate model is fitted to the basic subset. We consider the following linear model

$$\log(\text{output}) = \beta_0 \log(RM) + \beta_1 \log(SW) + \beta_2 \log(EW) + \varepsilon,$$

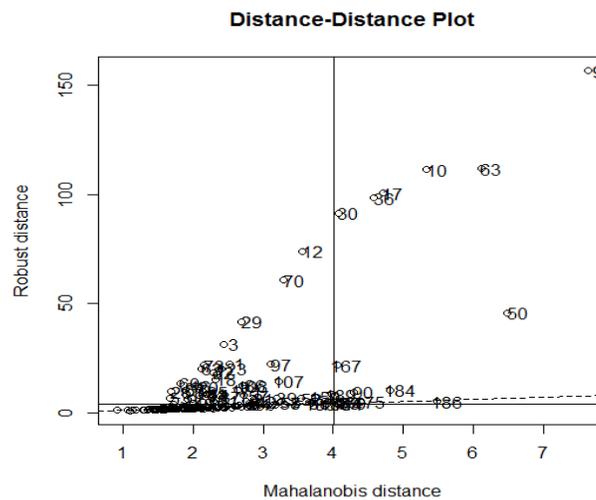
where  $\varepsilon$  has normal distribution,  $\beta = (\beta_0, \beta_1, \beta_2)$  is the vector of regression coefficients for the logarithm of output as response, RM is the amount of raw material, SW is salaries and wages and EW is the energy and water cost. By fitting the above model to these data, 21 observations are detected as outliers. Indices of these outliers are presented in the index plot of distances obtained by forward method in Figure 2. As it is indicated in this Figure, 4 observations indexed by 9, 12, 51 and 62 are far from other observations. These 4 observations are related to the units with the amount of raw material equal to zero.



**Figure 2.** The IMES Data: The index plot of distances obtained by forward and BACON methods

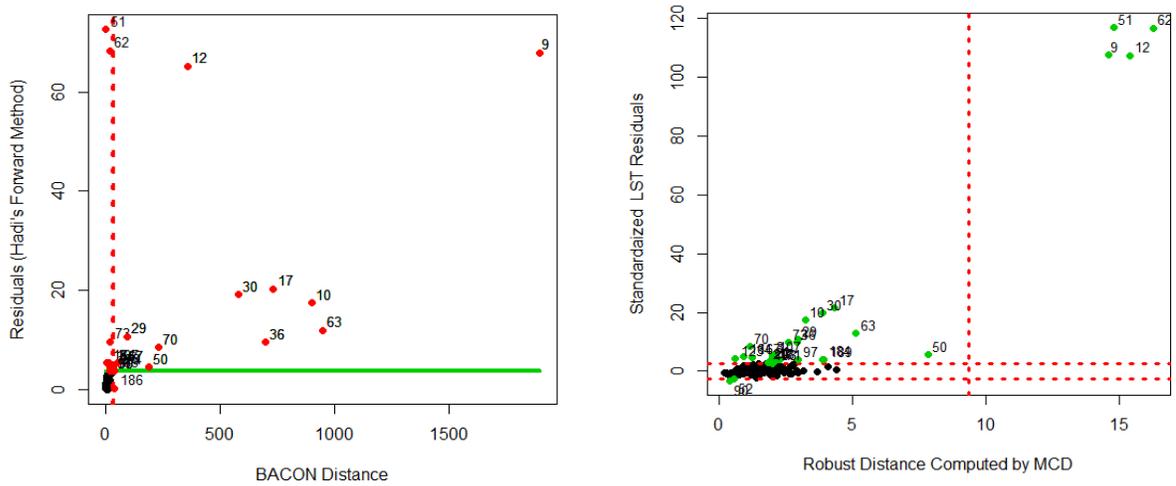
Figure 2 includes index plots of distances obtained by forward and BACON methods. These plots detect values higher than the cut-off point (grey lines) as outliers.

Most of the multivariate statistical methods are based on estimates of multivariate location and covariance; therefore these estimates play a central role in the framework. As mentioned in Section 2, input, output, total number of employees, salaries and wages, non-industrial payment, energy and water costs (or expenses) and input-output (I/O) variables are used in multivariate outlier detection methods. We start with computing the robust minimum covariance determinant (MCD) estimate for the IMES data. Figure 3 shows the Distance-Distance plot introduced by Rousseeuw and Zomeren (1991), which plots the robust distances versus the classical Mahalanobis distances and allows to classify the observations and to identify the potential outliers. Using other robust multivariate methods (MVE and Stahel-Donoho method) gives the same results as MCD. These results are not reported here.



**Figure 3.** The robust distance against classical distance for the IMES data set

Figure 4 plots the robust distances, computed by MCD and BACON methods, against robust residuals from Least Trimmed Squares (LTS) fit and Hadi's forward methods, respectively. This Figure shows the difference between outliers and influential data (left panel). In the left panel of Figure 4, the observations 9, 10, 12, 17, 29, 30, 36, 50, 63, 70 and 72 are chosen as influential data because their distances are more than cut points of BACON distance and their residuals are more than cut points of Hadi's forward method. In the right panel of Figure 4, observations 9, 12, 51 and 62 are identified as influential points since the standardized robust residuals from Least Trimmed Squares (LTS) robust regression are more than 2.5 and their robust distances are more than quantiles of the chi-squared distribution with degrees of freedom equal to the number of parameters estimated in the model ( $p$ ).



**Figure 4.** Distance plot obtained by the BACON method against the residuals obtained by Hadi’s Forward method (left panel) and plot of robust residuals (from LTS fit) against robust distances (right panel)

Table 1 compares various outlier detection methods and identifies the number of detected outliers. According to our investigation using different methods, some of the establishments which are identified as outliers are common in different methods. These are highlighted in Table 1. As it may be seen, all methods have nearly the same results but since the BACON method can be computed more quickly, regardless of the sample size, this method is preferred to be used as a multivariate outlier detection method.

**Table1.** The indices of detected outliers by various methods (bold numbers indicate the common outliers detected by all approaches)

Index of establishment detected as outlier	Methods of identifying outliers
1, <b>3, 9, 10, 12, 17</b> , 18, 22, <b>29, 30, 36, 50</b> , 52, 62, <b>63, 70</b> , 73, 86, 90, 97, 107, 123, 159, 175, 184, 186	<b>MCD</b>
<b>3, 9, 10, 12, 17, 29, 30, 36, 50, 63, 70</b> , 97, 107, 167, 175, 184, 186	<b>Stahel-Donoho</b>
<b>3, 9, 10, 12, 17, 29, 30, 36, 50, 63, 70</b> , 90, 97, 167, 184, 186	<b>BACON</b>
1, <b>3, 9, 10, 12, 17, 29, 30, 36, 50</b> , 60, 62, <b>63, 70</b> , 73, 88, 90, 97, 105, 107, 108, 123, 157, 167, 171, 189	<b>MVE</b>
<b>3, 9, 10, 12, 17, 29, 30, 36, 50</b> , 51, 62, <b>63, 70</b> , 73, 90, 107, 123, 164, 167, 184, 189	<b>Hadi’s Forward Method</b>

Since all above-mentioned outlier detection methods are sensitive to the small sample size, which can occur in each ISIC 4-digit Codes, a very desirable method for detecting outliers in periodic business or manufacturing establishment survey (created by Hidiroglou and Berthelot, 1986) will be revisited in the following. In this univariate method, an acceptance boundary is chosen and outliers are those observations whose growth rate between two consecutive surveys differs significantly from the corresponding overall trend of other observations belonging to the

same subset of the population. In the IMES data, there is a long series of observations belonging to the same manufacturing establishment. Based on the observations of the same manufacturer recorded on both 2009 and 2010 (for 10494 manufacturing establishments), Hidiroglou and Berthelot method has been applied on the input variable for all manufacturing establishments without considering the ISIC codes. By this approach, 606 manufacturing establishments are detected as outliers. This method can be applied for each variable of interest such as output, value added etc. In Hidiroglou and Berthelot's method by increasing  $C$ , the number of detected outliers is decreased. In analyzing our real data set we consider  $C = 20$ ,  $U = 0.4$  and  $A = 0.05$ . In the next section, these 606 detected outliers will be modified based on our proposed algorithm.

## 5. An empirical approach for adjustment of outliers

The next problem is how to treat the detected outliers in estimating the population parameters of interest. In this paper we used a value modification method for adjustment of outliers. In this approach, the value reported by the sample unit will be modified based on our proposed algorithm. The general idea of this method is based on the fact that the value reported by each sample unit could not be more than maximum value of its value reported during the time and the maximum value reported by other establishments in its ISIC 4-digit Codes (since generally all establishments in each ISIC 4-digit Codes are similar in their activities). Also, each value reported by establishment could not be less than minimum value of its value reported during the time and the minimum value reported by other establishments in its ISIC 4-digit Codes. By this idea, we could distinguish over-reported or under-reported establishments. Then, the over-reported manufacturing establishments, based on the 4<sup>th</sup> step in the following algorithm, are adjusted downward and the under-reported manufacturing establishments are adjusted upward.

The four steps of the proposed algorithm and the final results of data analysis are given in the following subsections.

In this section all manufacturing establishments which are investigated as outliers based on Hidiroglou and Barthelot's (1986) method or any other univariate and multivariate detection methods, can be used for adjustment. As mentioned before, since all outlier detection methods, mentioned in section 4, are sensitive to the small sample size, the 606 manufacturing establishments, detected as outliers based on using growth rate of input variable and Hidiroglou and Barthelot's (1986) method are used as outliers in this section. After adjustment of input values based on using the following algorithm, the output of all 606 establishments will be adjusted.

### 5.1. The Algorithm

#### Step 1:

In this step the input value of all manufacturing establishments, detected as outliers based on Hidiroglou and Barthelot's (1986) method in 2010, are used for adjustment and all possible growth rates of input variable during time are calculated. Since, the IMES data are available for 3 consequent surveys from 2008 to 2010, two growth rates are calculated for all detected outliers. These are:

$$r_{i1} = \frac{\text{input}_{i,2009}}{\text{input}_{i,2008}} \quad \text{and} \quad r_{i2} = \frac{\text{input}_{i,2010}}{\text{input}_{i,2009}}, \quad i = 1, 2, \dots, l,$$

where  $l$  is the number of detected outliers. Then, the maximum growth rate,  $\max_{i1}$ , and the minimum growth rate,  $\min_{i1}$ , for  $i = 1, 2, \dots, l$  are computed for each detected outlier as:

$$\max_{i1} = \max(r_{i1}, r_{i2}), \quad \min_{i1} = \min(r_{i1}, r_{i2}) \quad i = 1, 2, \dots, l.$$

For those manufacturing establishments not present in the 2008 survey,  $r_{i2}$  is considered as both the maximum and the minimum growth rates over time. Out of 606 detected outliers in 2010, 240 manufacturing establishments did not present in the first phase of the study in 2008.

### Step 2:

Since generally all establishments in each ISIC 4-digit Codes are the same and homogenous, in this step, maximum and minimum growth rates of all manufacturing establishments, except detected outliers, in each ISIC 4-digit Code at 2010 are calculated. It means that after removing detected outliers for manufacturing establishments available in both years 2009 and 2010, the growth rates,  $r_{i2}$ , are calculated for all remaining establishments. Then, the maximum and minimum growth rates (call them  $\max_2$  and  $\min_2$ ) are calculated for each ISIC 4-digit Code in 2010, by the following formulae:

$$\max_{i2} = \max(r_{i21}, r_{i22}, \dots, r_{i2n_j}), \quad i = 1, 2, \dots, l, \quad j = 1, 2, \dots, m,$$

$$\min_{i2} = \min(r_{i21}, r_{i22}, \dots, r_{i2n_j}), \quad i = 1, 2, \dots, l, \quad j = 1, 2, \dots, m,$$

These are considered to be the same for all detected outlier in each ISIC 4-digit Code. Here,  $n_j$  is the number of establishments in the  $j^{\text{th}}$  ISIC industries (not including outliers detected in the  $j^{\text{th}}$  ISIC industries) and  $m$  is the number of ISIC 4-digit Codes. So, for each ISIC industry, maximum and minimum growth rates are calculated. The attained maximum and minimum growth rates at this stage are allocated to all identified outliers in each ISIC industry.

In these two steps, for manufacturing establishments that have been identified as outlier, two maximum and two minimum growth rates are attained. One, the maximum and minimum growth rates during the time calculated for each establishments identified as outlier data in the first stage ( $\max_{i1}$  and  $\min_{i1}$ ), and the other, the maximum and minimum growth rates calculated for each ISIC industries for all establishments (not including those identified as outlier data,  $\max_{i2}$  and  $\min_{i2}$ ) in the second stage. The point is that the attained maximum and minimum growth rates at the second stage are the same for all identified outliers in each ISIC industry.

### Step 3:

In this stage we should identify the manufacturing establishments which have under-reported or

over-reported input values in 2010.

If declared input by respondents is less than or equal to minimum  $\{\min_{i1}, \min_{i2}\} \times input_{i,2009}$ , this establishment is identified as under-reported establishment. If declared input by respondents is more than or equal to the maximum  $\{\max_{i1}, \max_{i2}\} \times input_{i,2009}$ , this establishment is identified as over-reported establishment. Other establishments are clear of being under-reported or over-reported.

#### Step 4:

In order to adjust or treat the input variable for the manufacturing establishments identified as outlier, the following method is proposed. If the establishment is identified as under-reported, the adjusted input,  $Input_{adj,i,2010}$ , will be calculated based on the following formula:

$$Input_{adj,i,2010} = \max \left[ (\min_{i1} \times input_{i,2009}), (\min_{i2} \times input_{i,2009}), Input_{i,2010} \right], \quad i = 1, 2, \dots, l',$$

where  $l'$  is the number of establishments that are identified as the under-reported establishment and  $Input_{i,2010}$  is the declared input in 2010 by  $i$ th, respondent. If the establishment is identified as over-reported, the adjusted input will be calculated based on the following formula:

$$Input_{adj,i,2010} = \min \left[ (\max_{i1} \times input_{i,2009}), (\max_{i2} \times input_{i,2009}), \dots, input_{i,2010} \right], \quad i = 1, 2, \dots, l'',$$

where  $l''$  is the number of establishments that are identified as over-reported. For establishments that are not over or under-reported, the declared input in 2010 remains unaltered. In this application, out of 606 establishments which are identified as outlier based on Hidiroglou and Barthelot method, 247 establishments are identified as over-reported, 159 establishments are identified as under-reported, and 200 establishments are identified as respondents which are not over or under-reported.

## 5.2. Implementation of algorithm for analyzing IMES data

For all 606 out of 10,494 establishments which are identified as outlier based on input variable, input and output variables, using the above algorithm are adjusted. The results show that the total adjusted input for all establishments in 2010 is  $1.02722 \times 10^{15}$  Rials while the total declared input (without adjustment) for all establishments in 2010 is  $1.077844 \times 10^{15}$ . It shows that the total adjusted input is 0.95 of the total declared input. The results also show that the total adjusted output for all establishments in 2010 is  $1.696392 \times 10^{15}$  Rials while the total declared output for all establishments in 2010 is  $1.482896 \times 10^{15}$ . It shows that the total adjusted output is the total declared output multiplied by 1.15.

Table 2 gives the change and its rate in input and output before and after adjustment in each decile of the variables.

**Table 2.** Distribution of input and output before and after adjustment in each decile (millions of Rials)

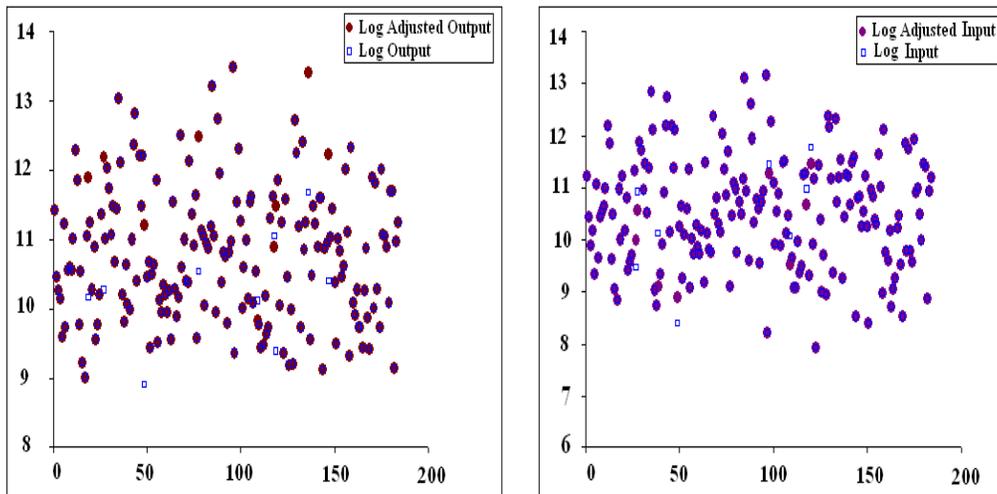
deciles	input	Adjusted input	rate of change	output	Adjusted output	rate of change
1 <sup>th</sup> decile	0.366	0.162	0.443	0.116	0.646	5.569
2 <sup>th</sup> decile	1.306	1.306	1	2.195	2.195	1
3 <sup>th</sup> decile	152.996	151.045	0.987	366.704	374.479	1.021
4 <sup>th</sup> decile	253.555	248.422	0.98	534.067	547.781	1.026
5 <sup>th</sup> decile	411.494	398.867	0.969	809.646	832.336	1.028
6 <sup>th</sup> decile	691.406	661.062	0.956	1259.264	1301.109	1.033
7 <sup>th</sup> decile	1214.074	1157.036	0.953	2053.124	2134.745	1.04
8 <sup>th</sup> decile	2287.316	2176.076	0.951	3735.826	3929.241	1.052
9 <sup>th</sup> decile	5160.055	4931.820	0.956	8243.640	8852.101	1.074
10 <sup>th</sup> decile	97495.900	92765.110	0.951	130966.000	151313.000	1.155

Out of 606 establishments which are identified as outliers based on input variable, by Hidioglou and Berthelot method, 25 establishments are related to 2710 ISIC Codes. Out of 25 establishments which are identified as outliers, 7 establishments are identified as over-reported for input variable and 3 establishments are identified as under-reported. For all 25 establishments which are identified as outlier based on input variable, the adjusted output variable based on the above procedure identified 2 over-reported and 7 under-reported. Table 3 compares the results based on removing outliers, adjustment and unadjustment methods on population parameters of interest such as the sum of input and the sum of output. The results show that omitting the detected outliers can reduce the total amount of output and input comparing with adjusted estimates. The results also show that unadjusted estimates can underestimate output and overestimate input values. So, by applying the adjustment method the over-reported inputs can be reduced and the under-reported outputs can be increased. The results show that the total adjusted output for 2710 ISIC Code is the total declared output multiplied by 1.28 and the total adjusted input for this Code is 0.99 times of total declared input.

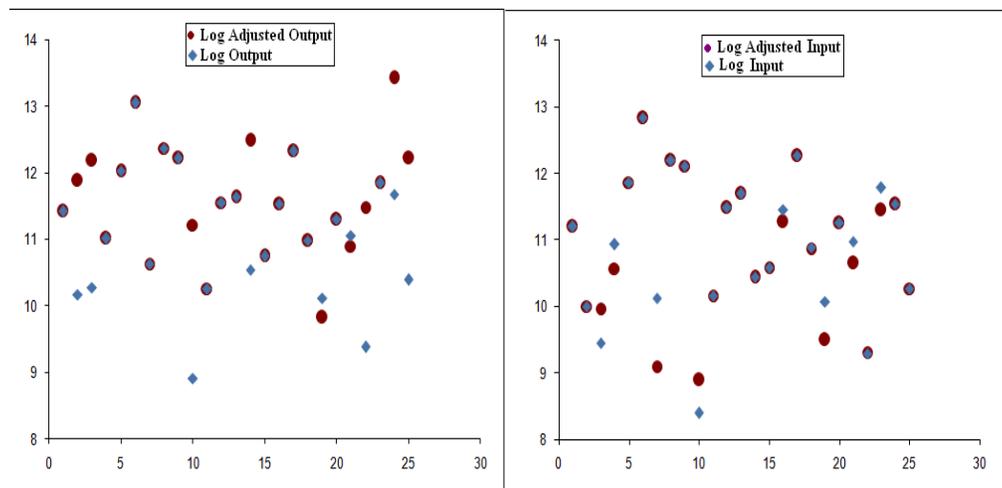
**Table 3.** Total input and output before and after adjustment for 2710 ISIC Code (millions of Rials)

Description	input	output
Adjusted	79155041	152599127
Unadjusted	79684581	119477926
Omitting outliers	64605260	97923934

Figure 5 shows the scatter plot of the logarithm of output and the logarithm of adjusted output (left panel), the logarithm of input and the logarithm of adjusted input (right panel) for all manufacturing establishments in 2710 ISIC Code, recorded on both 2009 and 2010 years. As mentioned before, out of 184 in this code, 25 establishments are identified as outliers based on input variable by Hidioglou and Berthelot method. Figure 6 shows the scatter plot of the logarithm of outputs and adjusted outputs for 25 manufacturing establishments in 2710 ISIC Code (left panel) and the logarithm of inputs and adjusted inputs (right panel) which are identified as outliers in 2710 ISIC Code. These Figures show that by using our adjustment method, how over or under reported outputs or inputs are adjusted.



**Figure 5.** The scatter plot of the logarithm of outputs and adjusted outputs (left panel) and the logarithm of inputs and adjusted inputs (right panel)



**Figure 6.** The scatter plot of the logarithm of outputs and adjusted outputs for 2710 ISIC Code (left panel) and the logarithm of inputs and adjusted inputs for 2710 ISIC Code (right panel)

## 6. Simulation study

In this section, a simulation study is conducted to illustrate the performance of the adjustment approach. In this simulation study, some ISIC 4-digit codes are chosen randomly. We consider surveys from 2008 to 2010 for each ISIC 4-digit code. In order to investigate the effect of sample size, two sample sizes, 859 and 1355 are chosen randomly for three and five ISIC 4-digit codes (2710, 3430, 1711) and codes (2710, 3430, 1711, 1810, 2697), respectively.

In order to simulate data from three selected ISIC 4-digit code from population of IMES, we take the fitted multivariate normal distribution for the subsequence surveys responses for each ISIC 4-digit code. In the next step, we generate  $N=1000$  replications of this three-variate distribution for each ISIC 4-digit code. The sample size is chosen to be equal to the real sample size. We have changed 5 values of randomly selected establishments, for each generated ISIC 4-digit code in 2010, to be outlier points in the generated data set. Since most of the outliers in output values in real data are under-reported and most of the outliers in input values are over-reported, outliers in each ISIC 4-digit code for output are produced in the way that 4 outliers to be under-reported and 1 outlier to be over-reported. And, outlier in input values are produced in the way that 4 outliers to be over-reported and 1 outlier to be under-reported. The proposed adjustment approach is used for obtaining the adjusted sum for the generated data set.

In this simulation study, with 859 observations, 67% of generated over-reported output and 74% of generated over-reported input values of establishments are correctly identified. Also, 52% of generated under-reported output and input values of establishments are correctly identified. Also for a sample size of 1355, 64% of generated over-reported output and 74% of generated over-reported input values of establishments are correctly identified. Also, 35% and 37% of generated under-reported output and input values of establishments are correctly identified. It shows that by increasing sample size, the correct identification of over-reported output and under-reported input is reduced.

We use relative bias, bias and root of mean-squared error for investigating the performance of the approach. These criteria are defined as follows:

$$\begin{aligned}\text{Rel.Bias}(\theta) &= \frac{1}{N} \sum_{i=1}^N \left( \frac{\hat{\theta}_i}{\theta_i} - 1 \right), \\ \text{Bias}(\theta) &= \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i), \\ \text{RMSE}(\theta) &= \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2},\end{aligned}$$

where  $\theta_i$  is population parameter (true value of the parameter) of interest and  $\hat{\theta}_i$  is the adjusted parameter estimate. According to the results of Table 4, the adjusted values of the output and the input are closer to the real values of them. Also, the total adjusted output is more than that of unadjusted one and the total adjusted input is less than that of unadjusted one. The results, given in Table 5, show that the relative biases, biases and root of MSEs of adjusted approach are less than those of the other approach. These are obtained with the presence of outliers in the data set. The results of Table 5 also show that by increasing the sample size, the relative biases, biases and root of MSEs in adjusted and unadjusted approaches increase. In other words, because of the effect of adjustment method on outlier points, the performance of adjustment method for a smaller sample size is better than that of a large sample size.

## 7. Conclusion

This paper examined some methods of detection and adjustments of outliers for the IEMS data set. Performance of different robust multivariate outlier detection methods such as BACON algorithm, minimum volume ellipsoid (MVE) estimator, minimum covariance determinant (MCD) estimator, Stahel-Donoho estimator and two univariate outlier detection methods, based on Hadi and Simonoff (1993) approach in cross-sectional surveys and Hidiroglou-Barthelot method of outlier detection in periodic manufacturing surveys are compared on the real data set of IMES. According to the results, most of the outlier detection methods have found the same most notable observations. However, the BACON method which is also used in Canada is preferred to be used as a multivariate outlier detection method, since this method can be computed quickly regardless of the sample size.

Since IMES data set is subject to measurement errors and these measurement errors affect the accuracy of the published statistics, in addition to outlier detection, we proposed a new empirical method for adjustment of outliers. In this empirical method, a new four-step algorithm is introduced for adjustment of detected outliers in estimating the population parameters of interest. To do this in IMES data, we identify the manufacturing establishments which are under-reported or over-reported. In this new empirical algorithm, the over-reported manufacturing establishments are adjusted downward and the under-reported manufacturing establishments are adjusted upward. A simulation study is also conducted to illustrate the performance of the adjustment approach.

The results of the analysis show that the outlier detection method is effective and the introduced adjustment method is effective in removing the impact of outliers on the main population parameter estimates. However, by increasing the sample size and the number of outlier points, the effect of adjustment method and the performance of our method are found to be weak.

## *Acknowledgements*

*The authors are thankful to Mr. Alireza Zahedian for his practical suggestions which have significantly improved our paper.*

## REFERENCES

- Andrews D. F. and Pregibon D. (1978). Finding the outliers that matter, *Journal of the Royal Statistical Society, Series B (Methodological)*, 40(1), 85-93.
- Atkinson, A. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers in *Data Analysis and Robustness* eds. S. Morgenthaler, E. Ronchetti and W. Stahel, Basel: Birkäuser.

- Béguin, C., and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.
- Béguin C, Hulliger B. (2008). The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data, *Survey Methodology*, 34(1), 91-103.
- Billor, N., Hadi, A.S. and Vellemann, P.F. (2000). BACON: Blocked adaptive computationally - efficient outlier nominators, *Computational Statistics and Data Analysis*, 34(3), 279-298.
- Box, G. E. P., Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, 26 (2): 211-252.
- Chambers, R. (1986). Outlier robust finite population estimation, *Journal of the American Statistical Association*, 81(396), 1063-1069.
- Donoho, D. (1982). Breakdown properties of multivariate location estimators, Ph.D. qualifying paper, Department of Statistics, Harvard University.
- Hadi, A.S. (1992a). Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society Series B*, 54 (3), 761-771.
- Hadi, A.S. (1992b). A new measure of overall potential influence in linear regression, *Computational Statistics and Data Analysis*, 14, 1-27.
- Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society Series B*, 56, 393 – 396.
- Hadi, A.S., Simonof, J.S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, 88, 1264-1272.
- Hidiroglou, M. A. and Berthelot, J. M. (1986). Statistical Editing and Imputation for Periodical Business Surveys, *Survey Methodology*, 12, 1, 73-83.
- Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.
- Ishikawa, A., Endo, S., and Shiratori, T.(2010). Treatment of Outliers in Business Surveys: The Case of Short-term Economic Survey of Enterprises in Japan (Tankan), Bank of Japan Working Paper Series and Review Series.
- Kosinski, A.S. (1999). A procedure for the detection of multivariate outliers, *Computational Statistics and Data Analysis*, 29, 145-161.
- Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, *The Annals of Statistics*, 27(3), 783-858.
- Patak, Z. (1990). Robust principal component analysis via projection pursuit, M. Sc. thesis, University of British Columbia, Canada.
- Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* (Eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz). Reidel. 283-297.
- Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, Inc.
- Rousseeuw PJ, Van Driessen K (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, 41, 212-223.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85 (411), 633–651.
- Stahel, W. (1981), Robuste Schätzungen: Infinitesimale optimalität und Schätzungen von Kovarianzmatrizen, Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.

**Table 4.** Results of parameter estimation (mean and standard deviation) of simulation study for different sample sizes (numbers are in millions of Rials)

Description	Input (n=859)		Input (n=1355)		Output(n=859)		output(n=1355)	
	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.
Adjusted	$1.047891 \times 10^8$	$2.497824 \times 10^6$	$1,065287 \times 10^8$	$2,672775 \times 10^6$	$1.613852 \times 10^8$	$4.232739 \times 10^6$	$1,637803 \times 10^8$	$4,233781 \times 10^6$
Unadjusted	$1.111063 \times 10^8$	$2.52769 \times 10^6$	$1,123475 \times 10^8$	$2,540993 \times 10^6$	$1.602437 \times 10^8$	$4.193357 \times 10^6$	$1,632606 \times 10^8$	$4,045317 \times 10^6$
Real value	$1.028489 \times 10^8$	$1.988296 \times 10^6$	$1,039507 \times 10^8$	$1,984485 \times 10^6$	$1.625768 \times 10^8$	$4.20371 \times 10^6$	$1,655923 \times 10^8$	$4,015431 \times 10^6$

**Table 5.** Results of simulation study (relative bias, bias and root of MSE criteria; millions of Rials) for different sample sizes

Description	Input				output			
	n=859 (ISIC=2710,3430,1711)		n=1355 (ISIC=2710,3430,1711,1810,2697)		n=859 (ISIC=2710,3430,1711)		n=1355 (ISIC=2710,3430,1711,1810,2697)	
	Adjusted method	Unadjusted method	Adjusted method	Unadjusted method	Adjusted method	Unadjusted method	Adjusted method	Unadjusted method
Relative Bias	0.019	0.080	0.0248	0.0807	-0.007	-0.014	-0.0109	-0.0140
Bias	$1.940272 \times 10^6$	$8.257415 \times 10^6$	$2,577977 \times 10^6$	$8,39681 \times 10^6$	$1.19161 \times 10^6$	$2.333121 \times 10^6$	$1.812034 \times 10^6$	$2.331687 \times 10^6$
Root of MSE	$2.413526 \times 10^6$	$8.352909 \times 10^6$	$3,246979 \times 10^6$	$8,497353 \times 10^6$	$1.318933 \times 10^6$	$2.425814 \times 10^6$	$2.146329 \times 10^6$	$2.428006 \times 10^6$