Working Paper
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and emerging methods

Presentation and development of outlier treatment in HCSO

Prepared by Gergely Horváth, Hungary, HCSO

I. Introduction

1. Detection and treatment of values deviating extremely from other values of the population is an old problem of statistics. These deviating values can result from measurement errors, from mixing more different populations or very skewed distribution. They have a high impact on estimates, especially when data have to be published for small domains or in case of a small sample size.

2. "Hawkins (Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse the suspicion that it was generated by a different mechanism. Barnett and Lewis (Barnett and Lewis, 1994) indicate that an outlying observation, or outlier, is the one that appears to deviate markedly from other members of the sample in which it occurs, similarly, Johnson (Johnson, 1992) defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data". (Ben-Gal 2005)

3. Outlier treatment is especially important for business statistics since they have many variables with strongly skewed distribution. Outliers which are in these cases extremely large values can be handled on the following manners:
   (a) They can be detected by editing
   (b) Their impact on the estimates can be decreased

4. The aim of the outlier detection during the editing phase is to decide whether this value is a real value or an error. Errors – like thousand errors – have to be corrected. Real extreme values cannot be the objects of imputation.

5. Real extreme values can be further classified into two groups: there can be representative and non-representative outliers. Representative outliers are the ones which represent other high values of the population. Non-representative outliers are unique in the population. Outlier treatment aims at improving the estimates and ensuring the consistency of the data coming from different sources or in different time period.

6. There are three main methods of dealing with outliers in a finite population (Cox, 1995):
   (a) Reducing the weights of outliers (trimming weight)
   (b) Changing the values of outliers (winsorization, trimming)
   (c) Using robust estimation techniques such as M-estimation

7. In the following we are going to show the outlier treatment practice used in different business surveys of the HCSO.

II. Outlier detection in the HCSO

8. In some business surveys the outlier treatment is done by the methodology department instead of the field sections, relatively separately from the other steps of the data process. There are many reasons behind this policy. The procedure was developed by methodologists and was implemented in a different environment than the rest of the data process (data process is done by the IT section in SQL, while SAS is used at the methodology department for a long time). The performance of the procedure is also made reasonable to be done by methodologists.

9. The procedure which was developed in the early 2000s can briefly be described as a univariate method which requires experts' decision based on outlier indicators. It takes place in the last part of the data production process right before the estimation. The aim here is not filtering the extreme values but the decrease of the impact on the estimation made by using these extreme values. The result of the procedure is that a new stratum is made for each case marked as outlier, and the weight of the rest of the units are also modified according to the number of outliers in the strata.

10. The investigation of the suspicious values (outliers) was made by the methodologist (previously done by the developer of the method), and the decision was discussed by the field statistician. At the beginning of the use of the procedure the data collection mode was paper based and the process of the data and the editing was made parallel with manual coding.

11. Before presenting the current method of outlier detection, a short description of the data is given. For this experiment we used data of the monthly survey of manufacturing. This survey consists of two parts, a take-all and a sampled part. The sampled part covers the smaller enterprises, with less than 50 employees (and more than 4, because the smallest businesses are not in the scope of the survey). The sampling frame is based on the Business Register, and consists of about 10 thousand units. The sampling ratio is about 15%, because of a very detailed stratification (a lot of NACE categories, also a lot of categories of the number of employees, and two geographical strata: the capital and everything else). (Telegdi 2004.)

A. Description of the method

12. The current method used in HCSO is rather unique. It is similar to the processes used by statistical editing in the sense that it is based on the investigation of the observation units' values driven by the field experts. On the other hand, it is not a typical error detection since there is no correction after the outlier detection phase – or at least the methodologist does not correct any value. Marking the outliers rather belongs to the estimation phase, because its effect will be the decreased (1) weight of the unit. (Csereháti 2004.)

13. The aim of the outlier treatment is improving the estimation. The best known treatments are decreasing the weights and winsorization. Usually, there are more variables involved in the process, but the connection between these variables is not taken into account – of course, the expert is allowed to take into account the values and indicators of the other variables. It is because the correlations between the variables are very small.

14. The process involves the following four steps:
   (a) computing the outlier indicators
   (b) outlier detection by the methodologist/expert
   (c) transfer of the result to the subject matter statistician
   (d) discussion of the result by the field expert and the methodologist (possible modifications), this is similar to the process of selective editing

15. As we have mentioned above the decision whether a particular data is outlier or not is made based on some indicators. We use the distribution properties of the sampling strata (mean, median, sample size, sum of weights) and the standardized values of the observed values per strata to prepare indicators. The auxiliary variables of the indicators are the following:

16. Standardized value of the variables by sampling stratum:

$$STANDARD_{ji} = \frac{Y_{ji} - MEAN_j}{STD_j}$$

17. The starting point is a simple standardization, i.e., we transform the data to a variable with mean of 0 and distribution of 1. It is important to note here, that by this step the strata values under the mean will be transformed to a less than zero value.

18. The next auxiliary indicator is the STAND, which is equal to the ratio of the standardized value and the Grubbs critical value belonging to the stratum. This indicator can only be used for stratum with size greater or equal to 3 because of the Grubbs value. In the monthly sampling there is a huge variety in the sizes of the strata, so it is often the case that this value cannot be computed.

$$STAND_{ji} = \frac{STANDARD_{ji}}{GR_j}$$

19. If we take the $G<G_k$ test of the Grubbs test (Grubbs 1950.), we may try interpreting this indicator as something similar to the Grubbs test. When $STAND_{ji} >1$ than (if the indicator refers to the maximum) then the observed case is an outlier. But the distribution of the variables was not tested (there are strong evidences that the variables are not normally distributed), and the indicator is applied to every values of the variable.

20. The most important indicator of the method is the LNSQRT:

$$LNSQRT_{ji} = LnY_{ji} \cdot \sqrt{STAND_{ji}}$$

21. This indicator has some very interesting characteristics. The most salient is that it gives missing values in many cases:

(a) when the value of Y is equal to 0 (the logarithm is not defined for this value)
(b) when STAND<0 (the square root is not defined for negative values)
(c) when the size of the stratum is 1 or 2 (there is no Grubbs critical value defined for these sample sizes)

22. We get on one hand that LNSQRT cannot be computed for small strata and on the other hand it will only be defined for values greater or equal to the mean of the stratum, which means basically for large values only.

23. The density function of the LNSQRT is a quite strange curve sometimes with two maximums. There is a break at the right end of the histogram which could be used for defining the extreme values. In about 90% of the investigated variables the value of the indicator was less than 13. The indicator identifies the extreme values for each stratum, but in the meantime it takes into account the variable itself as well. This indicator can be used well in practice since we want to filter the extreme big values.

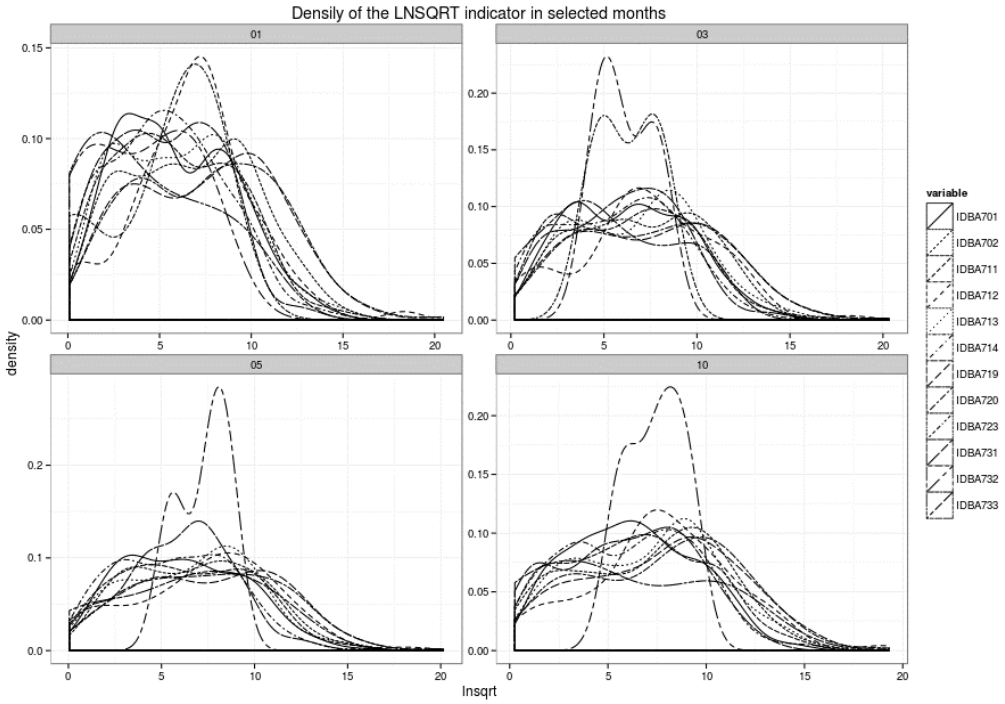

Figure 1: Density of the LNSQRT indicator in selected month and for all variables

24. SQUARED is our second indicator. It multiplies the value of the variables by the square of STAND which is always positive.

$$SQUARED_{ji} = Y_{ji} \cdot STAND_{ji}^2$$

25. For values different from the mean SQUARED can be computed in case of strata with size more than 2, because the STAND auxiliary indicator is also takes part in the computation.

26. MEANX is the ratio of the observed value of the unit and the weighted mean of the stratum without this unit value.

$$MEANX_{ji} = \frac{Y_{ji}}{MEAN_j \cdot N_j - Y_{ji}}\left(N_j - 1\right)$$

27. Because the variable has a very left-skewed distribution, this indicator has skewed distribution as well. The big values of this indicator show the values which are highly different from the other values in their stratum.

28. VALOUT indicator shows the difference between the estimation of the total with and without the given value in a given stratum.

$$VALOUT_{ji} = \frac{N_j - n_j}{n_j - 1}\left(Y_{ji} - \frac{MEAN_j \cdot N_j}{n_j}\right)$$

29. Our last indicator transforms the previous indicator into a ratio (or a percentage).

$$PVALOUT_{ji} = \frac{KIEMER_{ji}}{N_j^2 \cdot MEAN_j} \cdot n_j$$

30. In the process, these indicators are used. Though each has some significance, the most used and examined is the LNSQRT. The potential outliers are presented in tables, so the methodologist expert can examine and compare the original values, and the values of the different indicators.

B. Results of the method

31. In the next section some results of the above presented method will be discussed. The procedure presented above is an important part of the data processing practice for a long time. Great amount of practical experience is gained and this lead to a stable practice. Month by month, about two percent of the cases are identified as outliers.

32. The achieved sample of the small enterprises consists of 1185 enterprises (the total number of the population is bit more than 10 000). From January to October all in all only 45 enterprises were identified as outliers. And only 5 have been selected in every month. A bigger enterprise was 4 times identified as an outlier in average. In the table below the sample size and the number of outliers are presented by month.

Month		Number of Non Outlier Cases	Number of Outliers	All
1	n	1096	22	1118
	Percent	98,0	2,0	100
2	n	1105	20	1125
	Percent	98,2	1,8	100
3	n	1125	21	1146
	Percent	98,2	1,8	100
8	n	1104	17	1121
	Percent	98,5	1,5	100
9	n	1103	19	1122
	Percent	98,3	1,7	100
10	n	1092	21	1113
	Percent	98,1	1,9	100

Table 1: Sample size and selected outliers (using the current method) for some months

33. The second step of the outlier treatment after the detection is the modification of the weights of the outliers. As mentioned earlier, their weights are trimmed to 1. The outlier treatment has serious impact on the estimations. Both the estimated mean and the standard error decrease. Some illustrated values are presented in the table below: the average change of the estimation after the outlier treatment compared to the estimates without outlier treatment.

	Changes in the estimation	Changes in standard error
variables	Avg.	Avg.
Domestic turnover from industrial activities	0.8883	0.4721
Export turnover from industrial activities	0.7342	0.3826
Domestic turnover from non-industrial activities	0.8792	0.7518

Table 2: Average ratios of the estimations before and after the outlier detection

III. The alternative methods.

34. The method presented above effectively identifies the greatest values of some variables. But this method can be criticized at least in two aspects. On the one hand it is not fast, because the expert has to examine big tables or Excel sheets, and compare values of variables and indicators. On the other hand the identifying procedure is not always clear and reproducible. In the use of this method we have gained a lot

of experience and this practical knowledge helps us – as shown above – achieve stable results. So the most obvious way of the modernization is to develop a quicker and automated method.

35. There is always possibility to put in some subjective consideration when we decide to use a specific method. And in my opinion this is even more the case with the outlier treatment, although everybody tries to choose the best suitable method.

36. Our goal is to develop a simple and automated method to eliminate the cumbersome manual process, concentrate more on the effects the outlier treatment has on the estimates, and give better presentation of the whole process for the users to facilitate better understanding.

37. Four alternative methods are presented. We concentrate on univariate methods, except for the Mahalanobis distance based multivariate outlier detection. The alternative methods are the following:

(a) Quantiles
(b) MAD distance
(c) Share in the total
(d) Mahalanobis distance

A. Quantiles

38. The quantile method is a simple and robust technique. An allowable range of values is created based on the empirical data, and values falling outside the range in question are treated as outliers. The method can be applied to a lot of variables, but the connections between the variables cannot be taken take into consideration. In our case, if a variable contains a lot of zeroes, there is a risk, that a non zero value (even a not too big one) is identified as outlier. In a case like this the results are difficult to explain. Applying the method to sampling strata, the size of each stratum also has to be considered.

39. In our experiment we examined the whole sample as one, not taking into account the sampling strata, because there is no significant difference concerning the individual strata for several variables. We have used the same variables as in the original method. Two quantile thresholds were defined: 95% and 99%. At last, we decided that if a case is an outlier in one dimension (variable), then it is treated as an outlier.

40. As to the results the quantile method and the present one have similar outcome. Certainly, the application of two thresholds results in different number of outliers. But in case of the 99% threshold, the number of outliers are almost the same with the current and the quantile method. But the composition of the outliers differs: only the half of the outliers are common with both method. The effects of the two methods on the estimations are also similar. There is a small decrease in the value of the point estimation, and greater reduction in the standard error.

B. MAD distance

41. The MAD based outlier detection is as well a widely used robust method. The MAD is a robust statistic of dispersion, and more resilient of the presence of outliers than the standard deviation. For outlier detection, the base step is the definition of the distance: MEDIAN + k*MAD. A suggested value of the k is 3 or 3.5. (Iglewicz 1993.)

42. It can be stated that quite a big number of the enterprises can be defined as outlier (approximately 20% of the sample). But this great number can be explained by the attributes of the variables, as they are highly skewed. A possible solution can be the increase of the k value. An other option is the modification of the decision procedure: only the cases with two or more outlying variables are taken as outliers. Applying this consideration the number of outliers decreases close to the „usual” 2%.

C. Share in the total

43. Following some classical methods, now a new experimental outlier indicator is presented. It was inspired by the indicators used in the HCSO. In this indicator we tried to take into account the size of the

cases and also the size of the stratum, which contains the case. The first part of the formula shows the share of a case in the total of its stratum (or group). Because if a case shares a considerably great amount of the total, it may differ too much from the others, thus it is an outlier. But in a small group, (i.e two or three cases), the great measure of share has no real meaning. So we have to find a weighting method, to solve this problem. That is why „k" is introduced as a weight, which decrease the value of the indicator proportional to the size of the stratum.

$$psum_i = \frac{x_{ij}}{\sum_i^{n_j} x_{ij}}$$

$$k_j = 1 - \frac{1}{n_j}$$

$$OUT_i = k_j \cdot psum_i$$

44. With this indicator it is possible to consider the share and the size of the stratum in one formula. The advantage of this indicator is that it falls between 0-1, the only disadvantage is that it never reaches 1.

45. Contrary to the previous methods, this indicator is computed for each sampling stratum. Also a threshold has to be defined. Looking at the histogram of the indicator, one can see a skewed distribution. The mean is around 0.15 for many variables, and the value of the indicator is rarely greater than 0.5 (about 10% of the cases). With a threshold value of 0.5, in the case of 2 variables, the method gives a smaller number of outliers (only half of the number according to the usual method.).

D. Mahalanobis distance

46. The Mahalanobis distance can be thought of as a metric of the distance of each case in a multivariate normal space from the centre (Filzmoser 2004.). We can use that distance to detect multivariate outliers. In our experiment, we do not use the usual variables. They are replaced by three computed variables (total net turnover from industrial activities, total turnover, total gross output value). First, we had to log transform the variables to get normally distributed variables. We used the default significance level of the method (p=0.975).

47. The result is quite surprising: we almost have the same cases as outliers as with the usual method.

	Current method	
Robust Mahal.	Non outlier	Outlier
Non outlier	1088	6
Outlier	8	16

Table 3: Cases identified as outliers

48. Considering the results based on the different methods, it is not easy to find a perfect one. Each has some advantage and also disadvantage. And we have to mention other methods which can be interesting: winsorization and the Stahel-Donoho estimation (Franklin et al. 2000.). Based on the results of our experiments it can be stated that the presented methods are able to produce similar outcomes but using simpler computation and methodology. It is also clear that these methods can be easily automated, and put into the production process.

IV. Summary and conclusions

49. Nowdays there is an increasing demand of high quality, accurate and timely statistical products. This is one of the driving forces to develop automated production systems. In the recent years we have also moved from paper to electronic data collection, thus it is the right time to review our methodological tools as well.

50. The fact that outlier treatment, separately from other steps of data process, belongs to the methodology department and plays an important role in the above presented experimental procedures. The department takes and investigates the data and provides the results, the identifiers of the units dedicated to be outliers. With this situation we have more possibilities for experiments because we can change our tools without disturbing the other steps of the workflow. Of course, this does not mean that the procedure is based on absolutely unique or complicated solutions. We have made several steps toward thorough documentation and automatization of the procedure as much as possible during the past years: we have written unified SAS codes for several cases (surveys) and have put the procedure in SAS EG environment.

51. For development we have chosen the R programming language instead of SAS. We had more reasons for this choice: first, the task is completely separated from the other steps of the data process in the Office, we only have to be compatible with the system in receiving the data and providing our output. This only means accessing the databases in practice. Second, we can see the growing importance of R at the Methodology department, and reconstructing-redesigning-investigating outlier treatment seemed to be a good reason to try this new software tool. Last but not least, we can find much more procedures in various R packages for comparing different methods. All the methods presented in this study can be found in different packages of R.

52. One of the main objects of the development is to give comprehensive information to the users (the statisticians) about the results of the process made by us. Various graphic tools of R (like ggplot2 package or R base graphic tools) enable to provide more information as well as the html based reports given by the knitr package. We use Rstudio as the standard environment for R.

A Conclusions, further plans

53. The studies and analyses show that there is no optimal solution. Every procedure has pros and contras. This is the fact which motivates us to get to know the impacts of the data process steps. Monitoring and evaluating the results is crucial, for example, by well-chosen quality indicators. Disseminating the results to the users should be part of the monitoring. This process can be supported by a well-designed and automatic report in which good tools like well designed and informative figures play an important role.

54. The development of the methodology is still in progress.

References

Barnett, V., Toby L. *Outliers in statistical data*. Vol. 3. New York: Wiley, 1994.

Ben-Gal, I. *Outlier detection*. *Data Mining and Knowledge Discovery Handbook*. Springer US, 2005. 131-146. (http://www.eng.tau.ac.il/~bengal/outlier.pdf)

Cox B. G., Binder A., Chinnappa N. B., Christianson A., Colledge M. J., Kott P. S. *Business Survey Methods.* John Wiley & Sons. 1995.

Csereháti, Zoltán. *Az outlierek meghatározása és kezelése gazdaságstatisztikai felvételekben.* Statisztikai szemle, 2004. 728-746.

Franklin, S., Thomas, S., and Brodeur, M. *Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators*. Proceedings of the Second International Conference on Establishment Surveys. 2000.

Filzmoser, P. *A multivariate outlier detection method*. Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling. Vol. 1. No. 1989. Minsk: Belarusian State University, 2004.

Filzmoser, P. Varmuza, K. *chemometrics*: *Multivariate Statistical Analysis in Chemometrics*. R package version 1.3.8. http://CRAN.R-project.org/package=chemometrics. 2012.

Grubbs, F. E. *Sample criteria for testing outlying observations*. The Annals of Mathematical Statistics. 1950. 27-58.

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

Lapsley, M. Ripley B., *RODBC: ODBC Database Access*. R package version 1.3-10. http://CRAN.R-project.org/package=RODBC. 2013.

Lumley, T. *survey: analysis of complex survey samples*. R package version 3.28-2. 2012.

Lumley, T. *Analysis of complex survey samples*. Journal of Statistical Software 9(1): 1-19 2004.

MEMOBUST http://cros-portal.eu/content/memobust Retrieved 25.02.2014.

Murdoch, D. *tables*: *Formula-driven table generation*. R package version 0.7.64. http://CRAN.R-project.org/package=tables. 2013.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.

Starkweather, J. *Multivariate outlier detection with Mahalanobis' distance*. Retrieved: http://www.unt.edu/rss/class/Jon/Benchmarks/Moutlier_JDS_July2013.pdf. 25.02.2014

Telegdi, L. *A kisszervezetek integrált reprezentatív évközi megfigyelése a 2000-es években*. Statisztikai szemle. 2004. 668-690.

Wickham, H. *The Split-Apply-Combine Strategy for Data Analysis*. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/. 2011.

Wickham, H., Francois, R. (2014). *dplyr*: *dplyr: a grammar of data manipulation*. R package version 0.1.1. http://CRAN.R-project.org/package=dplyr

Wickham, H. *Reshaping data with the reshape package*. Journal of Statistical Software, 21(12), 2007.

Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.