Outlier Detection in Weight Time Series of Connected Scales

Saeed Mehrang*, Elina Helander*, Misha Pavel†, Angela Chieh‡ and Ilkka Korhonen*
*Signal Processing Department, Tampere University of Technology, Tampere, Finland
Email: {saeed.mehrang, elina.helander, ilkka.korhonen}@tut.fi
†College of Computer and Information Science and Bouve College of Health Sciences, Northeastern University
Boston, Massachusetts, USA
Email: m.pavel@neu.edu
‡Data and Studies Department at Withings, Paris, France
Email: angela.chieh@withings.com

***Abstract*—In principle, connected sensors allow effortless long-term self-monitoring of health and wellness that can help maintain health and quality of life. However, data collected in the "wild" may be noisy and contain outliers, e.g., due to uncontrolled sources or data from different persons using the same device. The removal of the "outliers" is therefore critical for accurate interpretation of the data. In this paper we study the detection and elimination of outliers in self-weighing time series data obtained from connected weight scales. We examined three techniques: (1) a method based on autoregressive integrated moving average (ARIMA) time series modelling, (2) median absolute deviation (MAD) scale estimate, and (3) a method based on Rosner statistics. We applied these methods to both a data set with real outliers and a clean data set corrupted with simulated outliers. The results suggest that the simple MAD algorithm and ARIMA performed well with both test sets while the Rosner statistics was significantly less effective. In addition, the ARIMA approach appeared to be significantly less sensitive to long periods of missing data than MAD and Rosner statistics.**

***Keywords*-weight time series analysis, outlier detection, ARIMA modeling, MAD scale estimate, Rosner statistics**

I. Introduction

Throughout the past hundred years people's lifestyle has changed significantly due to highly automated machines evolving day by day and as a result, forced us to be less active than our ancestors. This lack of physical activity combined with unhealthy eating habits have led to one of the most severe today's health problems; obesity. Overweight and obese individuals are at increased risk of having heart problems, diabetes, musculoskeletal disorders and high blood pressure [1].

Different methods have been studied and developed to help people change their behavior for example by modifying eating habits as well as physical activity patterns [2]–[4]. However, according to recent studies, adherence to such behavioral therapies represents a significant challenge. In this respect, one of the most effective approaches to mitigate this challenge suggested by behavioral therapists was self-monitoring combined with automatic feedback as a central principle for weight management [5], [6]. The working hypothesis is that the insights of how everyday behavior affects their weight helps people either lose or maintain their weight. In particular they will get rough notions about the long-term changes in their weight occurring over time. Hence, regular self-weighing can play an important role in long-term weigh-loss maintenance.

A variety of sensors and mobile applications are available for consumers to support behavior change by means of self-monitoring. Those sensors collect and display a wealth of personal health and wellness data. For self-weighing, modern connected weight scales automatically transmit weight data to network servers so that users can access their weight on-line or through a mobile phone application. The availability of long-term weight time series may also help coaches to get insight about individuals' behaviors through weekly and annual weight variations [7] or e.g. the correlation between self-weighing frequency and weight change [8]. Equipped with this information it would be possible to optimize just-in-time adaptive interventions.

Self-monitoring with connected sensors is done during daily life in uncontrolled conditions, which may result in data that is contaminated by outliers. They may arise from conditions where different users are using the same devices, or due to some external uncontrolled influences; in self-weighing, exceptionally heavy clothing or carrying some objects during measurement may lead to wide deviations from the real body weight. Scales may also be used to weigh other things than persons, such as pets, suitcases or individuals with (and without) suitcases. In order to support accurate analysis of subtle changes and patterns in body weight, removal of these outliers is necessary.

A challenge in detecting outliers arises from the fact that people's adherence to self-weighing varies temporally, and hence missing data with long gaps between consecutive measurements can occur. These long gaps of time without measurements may lead to changes in weight levels [8]. On the other hand, even 2-3 percent variation of body weight can occur within a day [8]. These fluctuations alone can easily mask subtle changes and delay detection of significant trends. Therefore, weight dynamics need to be taken into account.

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

In this paper, we examined three different outlier detection methods based on: (1) autoregressive integrated moving average (ARIMA) specification of time series [9], (2) median absolute deviation (MAD) scale estimate [10], [11], and (3) Rosner statistics anomaly detection [12].

These three techniques were applied to two different types of test sets; the first one comprised simulated outliers added to 20 clean real weight time series and the second one included 20 visually annotated real weight time series that contained outliers.

II. Methods

A. Materials

1) Simulated data: We randomly selected 20 clean (i.e. no visually observable outliers in the time series) weight time series, whose length varied between 300 and 350 measurements taken from Withings (Withings, Paris, France) weight scale users. Any time series with possible anomalies were excluded from the data set and replaced with another randomly selected time series. The total number of weight measurements included in this data set is 6494.

On average 4.5 percent of the data points in each of the 20 clean weight time series were randomly selected and intentionally corrupted with normally distributed noise. That is, the original data points were replaced with simulated outliers. Mean value of half of the outliers was equal to mean value of original weight time series increased by 5 kg. The mean value of the other half was equal to mean value of the original weight time series minus 10 kg. The standard deviation of the outliers was defined equal to the median standard deviation of the time series included in the 20 clean time series. The goal was to simulate outliers due to occasional interference by weighing two individuals different from the target person. The total number of outliers simulated in this test set is 294.

2) Real data: A subset of 20 time series including outliers were randomly selected among a set of 10,000 self-weighing time series database. A time series was included if it was visually assessed to contain at least one outlier. Outliers were visually annotated by one researcher by visual inspection. The number of weight measurements included in this data set altogether is 14112 in which 68 points were visually identified to be outliers.

B. Data analysis

The data analysis and implementation of the algorithms were done by R version 3.2.1. The ARIMA-based outlier detection algorithm used in this study was deployed from the package called tsoutliers [13]. The Rosner statistics anomaly detection test deployed EnvStats package [14].

C. Outlier detection methods: ARIMA approach

A non-seasonal time series named as X_t follows an autoregressive integrated moving average (ARIMA) process of order (p,d,q) if the dth difference of the X_t can be considered as an autoregressive moving average (ARMA)(p,q) process. Autoregressive (AR) term refers to the fact that any value of a variable X at time point t can be explained by p previous values of X at time points $t-p, t-p+1,..., t-1$. Moving average (MA) part of the model denotes the forecast error at time instant t can be explained by q past forecast errors at time points $t-q, t-q+1,..., t-1$. A general representation of ARIMA models called seasonal autoregressive integrated moving average model used for the outlier detection was described in [15], [16].

The ARIMA-based approach can detect four types of outliers described in [9], namely (1) a level shift outlier (LS), (2) an innovational outlier (IO), (3) an additive outlier (AO), and (4) a temporary change (TC). The detection procedure based on ARIMA specification can be divided into three iterative steps as follows [13].

I. Locate outliers: First the algorithm computes the initial ARIMA model parameters based on the maximum likelihood (ML) or minimum conditional sum of squares (CSS) as specified by the ARIMA choice of parameters. Subsequently, it checks every data point of the series to get four different τ-statistics corresponding to the four above-mentioned types of outliers. The algorithm then chooses the largest absolute value of each time point τ-statistics as a dominant outlying effect. The dominant τ-statistics is compared to a critical value C which was specified for the function in advance. This threshold variable is used to decide whether the time point can be considered as an anomaly or a valid data point that must be kept unchanged.

II. Iterate: After finding a set of m potential outlying points, the algorithm computes new τ-statistics for these data points based on outlier effects and estimated residuals obtained from the fitted ARIMA model. In order to make sure valid data points are not included in the set of outliers the algorithm considers a condition by which every outlying point with τ-statistics smaller than C in absolute value is removed from the set of outliers. Then, again new τ-statistics will be computed based on this new set of outliers and the above-mentioned condition is tested iteratively until the point no τ-statistics smaller than C is found within the set of outlying data points.

III. Remove outliers: The estimated effects of identified outliers are removed from the model and new ARIMA parameters will be calculated based on ML-CSS criteria.

D. Outlier detection methods: MAD

Detection of outliers using median absolute deviation implements a moving window of length k centered at each sample point and estimates two variables: The first one is the median of the window and the second one is a scale estimate, namely the median absolute deviation (MAD). The values of $MAD_i = median_i[|X_i - median_j(X_j)|]$ are scale estimate of X_i within the window. Each X_i is then compared to the

corresponding MAD_i . If the absolute deviation of the data point is greater than the threshold, then the data point is considered as an anomaly [10], [11]. A threshold value t_0 controls sensitivity of the algorithm to local fluctuations. The greater the t_0, the less sensitive is the outlier detector.

E. Outlier detection methods: Rosner statistics

Rosner statistics is an anomaly detection procedure applicable for time series of normally distributed samples. The computation of Rosner statistics is based on the assumption that both the the clean data and the outliers are normally distributed which means after removing k outlying data points, the series should be distributed normally.

Basically in Rosner statistics the highest k value that can be chosen for a time series of length n is the largest integer smaller than $\frac{n}{10}$. The detail description of Rosner's anomaly detection algorithm can be found in [12].

I. Locate outliers: At the outset, the algorithm finds the k extreme studentized deviates (ESD statistic) values described in detail in [12]. Then the most deviant outlier, is removed and the algorithm recomputes the ESD values for the remaining $k-1$ points. The algorithm iterates to compute the ESD values for the remaining until all k points are evaluated. At the end of this step a series of ESDs are recalculated based on the sample sizes of $n, n-1, n-2,..., n-k+1$ consecutively.

II. Remove outliers: In order to determine which one of the k potential anomalies are likely outliers, they need to be successively compared with critical values of ESD statistics corresponding to each sample size of $n, n-1,..., n-k+1$. If ESD of the kth outlying point (the least extreme studentized statistic) is larger than corresponding kth critical value then all of the suspected k data points are outliers. Otherwise, this point will be removed from the set of outliers. The algorithm compares the subsequent extreme outlying points to their equivalent critical value.

This procedure continues until all outlying data point with ESD bigger than their equivalent critical value are removed or all of the k potential outlying points were tested [12].

III. RESULTS AND DISCUSSION

Table I depicts the settings used for corresponding variables. These values were obtained based on receiver operating characteristic curves of each technique. The threshold values that revealed the best diagnostic performance were chosen to be used as settings of corresponding algorithms.

The results of outlier detection in real annotated time series are given in Table II. It can be observed that MAD scale estimate has the best sensitivity with more than 98 percent true detection of outliers. After that, sensitivity of ARIMA approach and Rosner statistics were 93 and 81 percent respectively. The specificities of all three algorithms were very high, all beyond 98 percent true identification of correct data points. Fig. 1 clearly illustrates a case where

Table I
SETTINGS OF THE IMPLEMENTED ALGORITHMS.

Variable	Value
ARIMA critical value (C)	2.5
MAD threshold value (t_0)	2
MAD window length (k)	30
Rosner error rate (α)	0.5

Table II
THE RESULTS OF OUTLIER DETECTION IN REAL ANNOTATED TEST SET.

Methods	Sensitivity	Specificity
ARIMA	0.93	0.98
MAD	0.98	0.99
Rosner	0.81	0.99

implemented methods detected outlying points in a real annotated weight time series.

Table III shows outlier detection performance in simulated test set. The results suggest that ARIMA outlier detection performs slightly better than MAD scale estimate and significantly better than Rosner statistics in terms of sensitivity. The corresponding specificity of all three methods is again quite high and approximately equal which means almost all of the correct data points were classified correctly. Fig. 2 and Fig. 3 reveal two examples of how these three methods removed outliers from simulated test set. There are big differences between the sensitivity of Rosner statistics and two other methods in Fig. 2 while in Fig. 3 all three methods performed excellent. As explained deeper in the following paragraph, the reason for such an enormous performance degradation in Rosner statistics in Fig. 2 can be due to masking effect of the level shift occurred at the beginning of time series.

By analyzing simulated time series individually, it was observed that there were a few cases where the sensitivity of MAD and Rosner statistics dropped by approximately 10 and 50 percent respectively while the sensitivity of ARIMA approach was still reasonably high. The main reason for these outcomes is the effect of sudden level shifts in vicinity of undetected outliers. These level shifts occur usually when there is either sudden weight gain or sudden weight loss after a long period of time of missing measurements. To explain the effect of level shifts resulted by gaps of measurements, one of the simulated cases corrupted by artificial outliers is shown in Fig. 4. There are two major level shifts occurring at time instants 100 and 200 approximately. These level shifts are the main causes of performance degradation in MAD and Rosner statistics. As can be seen in Fig. 4(c) ARIMA approach detected almost all of the outliers. However, neither MAD (Fig. 4(d)) nor Rosner statistics (Fig. 4(e)) was able to identify outlying points occurred in neighborhood of level shifts. ARIMA appears to be the best choice for cases

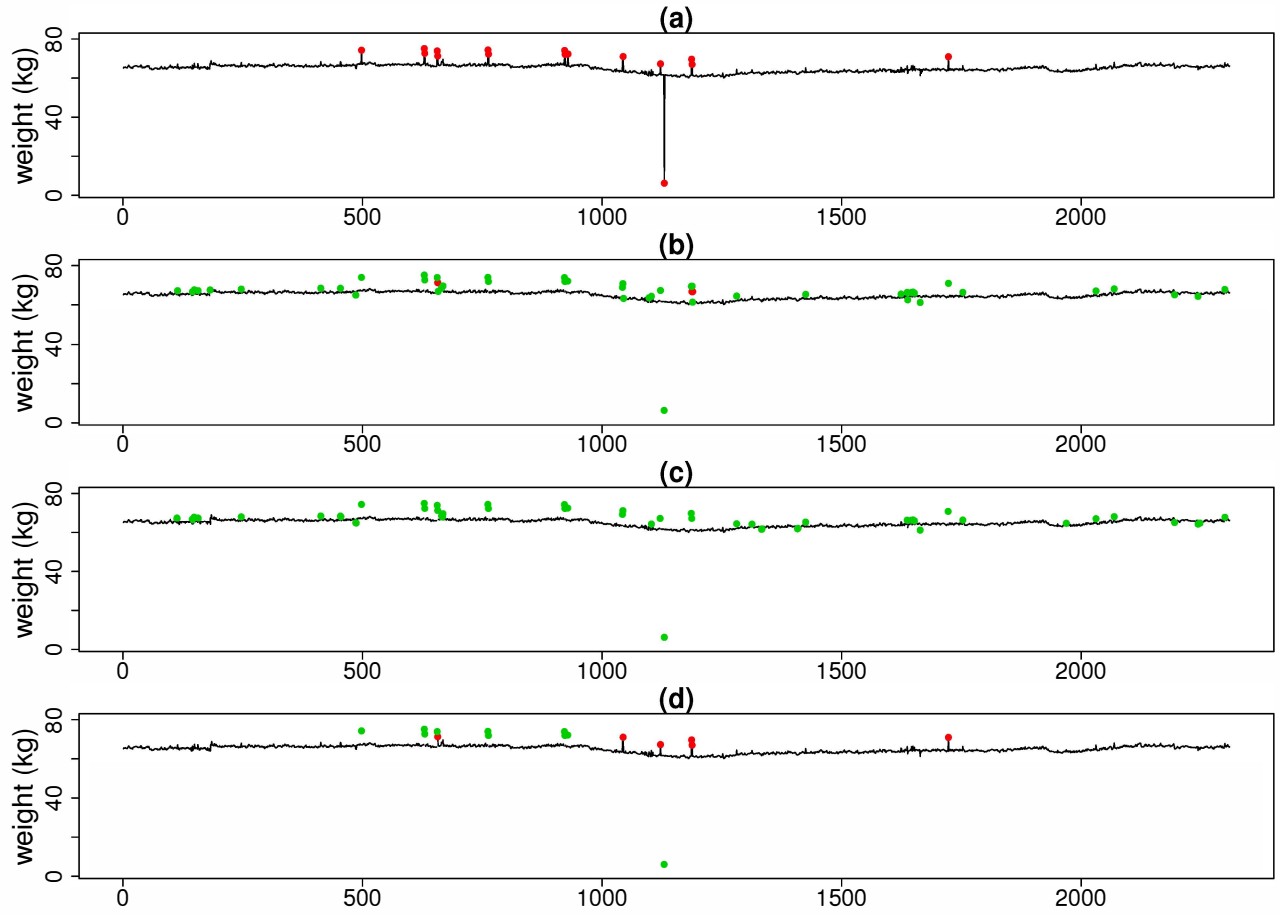


Figure 1. A real example of outlier detection using implemented algorithms: (a) original time series, results of (b) ARIMA approach, (c) MAD scale estimate, and (d) Rosner statistics. Red dots denote visually annotated outliers while green dots are the ones detected by algorithm.

Table III
THE RESULTS OF OUTLIER DETECTION IN SIMULATED TEST SET.

Methods	Sensitivity	Specificity
ARIMA	0.93	0.96
MAD	0.94	0.96
Rosner	0.65	0.99

Table IV
PEARSON CORRELATION COEFFICIENTS AND CORRESPONDING p-VALUES BETWEEN STANDARD DEVIATION OF TIME SERIES AND DIAGNOSTIC PERFORMANCE OF ALGORITHMS.

Method	r	p-value
ARIMA	-0.16	0.50
MAD	-0.44	0.049
Rosner	-0.98	< 0.001

where level shifts can mask anomalies in their vicinity. This similar masking effect took place in four of the simulated time series included in simulation test set. Therefore, the average sensitivity of Rosner statistics dropped significantly there.

Owing to occurrence of level shifts the standard deviation of time series often increases. Consequently the increase in the standard deviation of the time series, reduces the sensitivity of Rosner statistics and MAD scale estimate because these techniques ignore sequential aspects of the time series. To clarify this issue, Table IV depicts Pearson correlation coefficients between standard deviation of time series and diagnostic performance of implemented algorithms for simulated data set. It is obvious that ARIMA approach performed quite independent from standard deviation of time series. In contrast, there are correlations between standard deviation and performance of MAD and Rosner statistics with r coefficients equal -0.44 and -0.98 respectively.

Receiver operating characteristic (ROC) curves depicted in Fig. 5 were estimated based on statistical performance of simulated test set. Based on Fig. 5 it is conceivable that by decreasing critical value (C) from 5 to 2.25 there was a gradual increase in sensitivity in contrast to the decline of specificity. Similarly, corresponding sensitivity

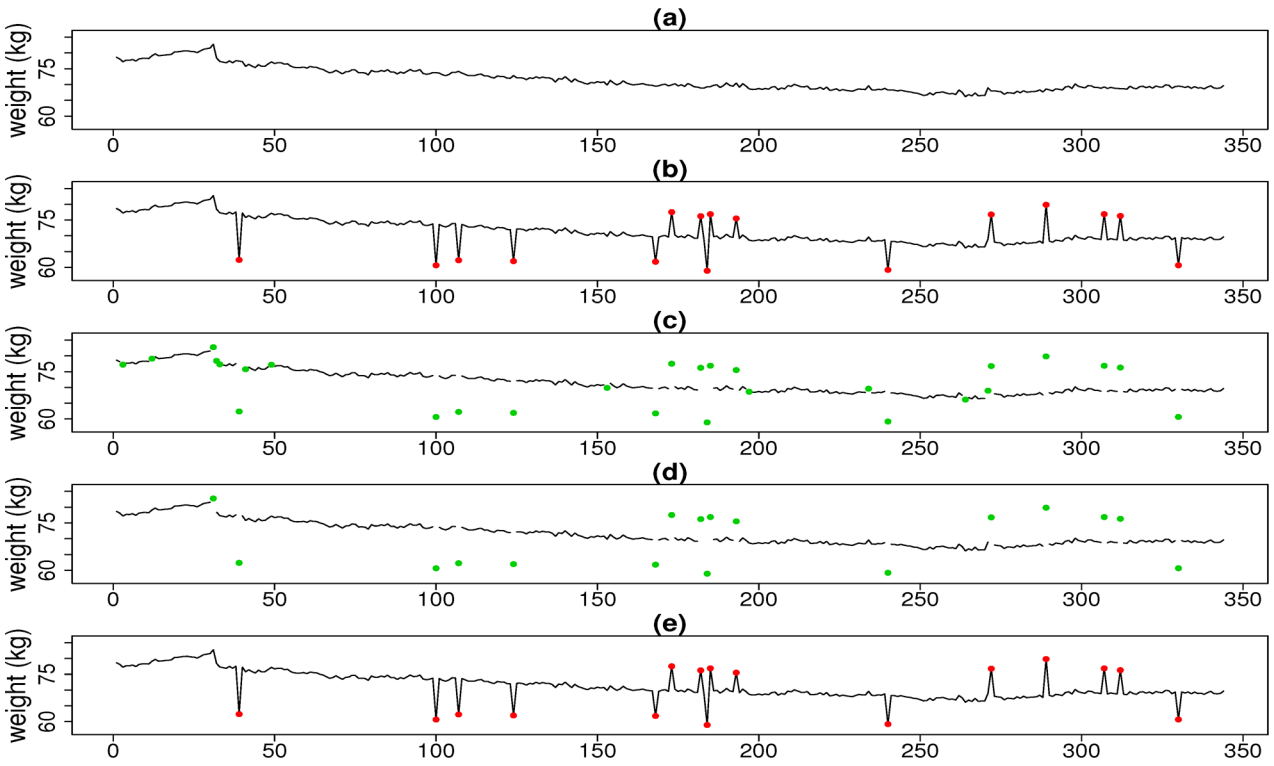


Figure 2. An example of a time series from simulated test set: (a) original time series, (b) corrupted time series, results of (c) ARIMA, (d) MAD scale estimate, and (e) Rosner statistics. Red dots describe simulated outliers while green dots represent detected outliers.

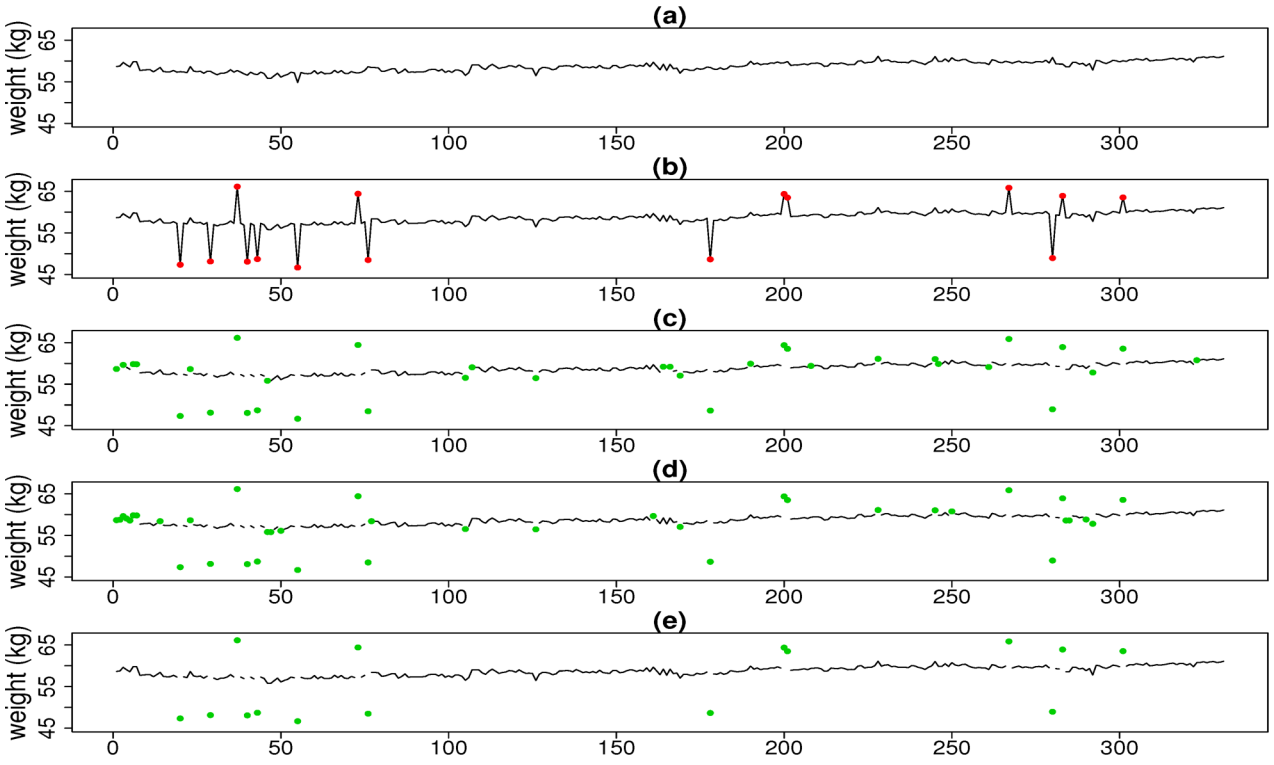


Figure 3. An example of a time series from simulated test set: (a) original time series, (b) corrupted time series, results of (c) ARIMA approach, (d) MAD scale estimate, and (e) Rosner statistics. Red dots represent simulated outliers while green dots denote detected outliers.

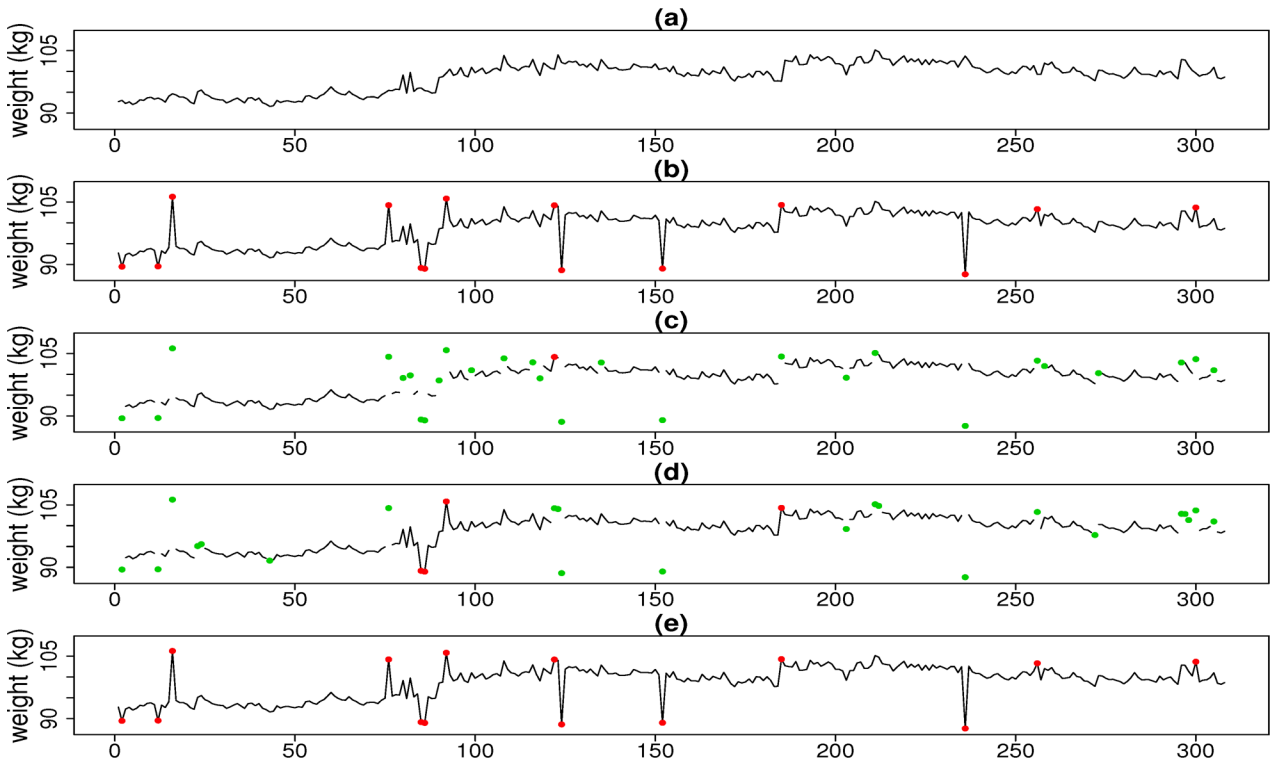


Figure 4. An example of a time series from simulated test set: (a) original time series, (b) corrupted time series, results of (c) ARIMA approach, (d) MAD scale estimate, and (e) Rosner statistics. Red dots represent simulated outliers while green dots denote detected outliers.

and specificity of MAD scale estimate changed by reducing the threshold value (t_0) from 6 to 2. Concerning ROC curve of Rosner statistics, it can be observed that by slowly increasing the error rate (α), there were slight improvements in the sensitivity; however, the specificity remained quite high without any comprehensible declines. Therefore, here the best critical value (C) for ARIMA approach is 2.5, the most efficient threshold value (t_0) for MAD scale estimate is 2, and ultimately the best error rate (α) for Rosner statistics can be 0.99. Although the optimum value of Rosner statistics error rate is 0.99, there were no significant changes in the results of outlier detection between error rate value 0.5 and 0.99. The value of α was chosen to be 0.5 for the whole parts of this study. The area under curve (AUC) is 0.97, 0.96, and 0.86 respectively for ARIMA, MAD, and Rosner statistics.

The advantage in performance of ARIMA comes at a cost in computational complexity. The average processing time of ARIMA approach is $\sim$100 times longer than the other two methods as illustrated in Table V. This can be considered as the most significant drawback of using such a complex iterative algorithm.

Table V
AVERAGE PROCESSING TIME OF ALGORITHMS USED IN THIS STUDY.

Methods	Average Processing Time
ARIMA	$87.7(s)$
MAD	$0.017(s)$
Rosner	$0.0025(s)$

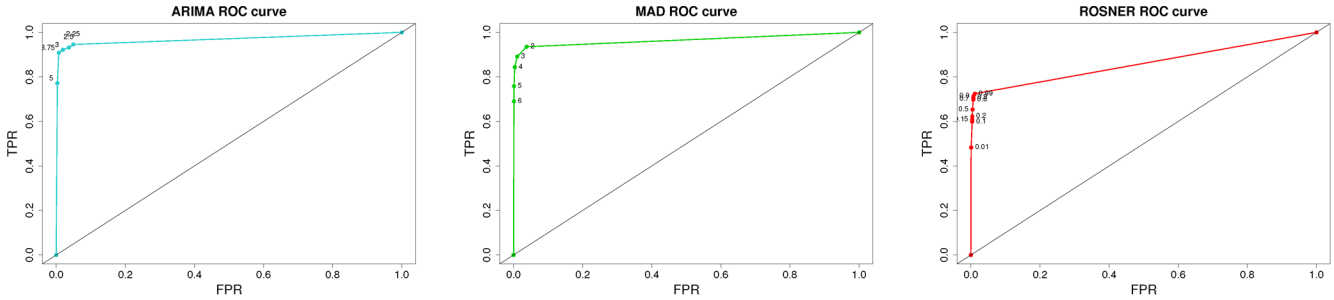
IV. CONCLUSION AND FUTURE WORK

Based on our preliminary investigation, the median absolute deviation scale estimate can be a good candidate for analysis of anomalies in the context of weight time series. It is computationally light and simple to implement. However, its performance was slightly degraded in the presence of level shifts and gaps in the data.

The results of our study suggest that ARIMA outlier detection approach is superior to other methods but the algorithm is computationally expensive compared to other techniques. An additional shortcoming of the ARIMA-based approach not mentioned earlier is its low power in detection of outliers at the very beginning of time series.

A method that constitutes the strengths of both MAD scale estimate and ARIMA anomaly detection could be useful. For instance a serial architecture of MAD followed by ARIMA approach may compensate the deficiencies of both methods. However computational problems of ARIMA approach must be solved in advance.

Rosner statistics anomaly detection method used in this study performed less well than the other techniques. We



(a) ARIMA (AUC = 0.97) (b) MAD (AUC = 0.96) (c) Rosner statistics (AUC = 0.86)

Figure 5. ROC curves corresponding to (a) ARIMA, (b) MAD, and (c) Rosner statistics. Points in the graphs represent threshold values. TPR and FPR denote true positive rate and false positive rate accordingly.

suspect that some of the assumptions underlying this method may be violated in the real life weight sequences. It is possible, however that implementation of a windowed Rosner statistics may result in better outcomes accordingly.

One limitation of this study is the fact that in real weight time series it is challenging to decide whether any data point can be considered as an outlier or a normal point. Therefore defining a ground truth based on visual inspection may sometimes lead to faulty decisions. By the same token, the results using the simulated outliers may not correspond to all real life situations. Adding contextual information in the future may alleviate these concerns.

The specific characteristics of weight dynamics should be considered in order to assess whether weight change between neighboring observations given their temporal distance can be realistic. As a weak point, the temporal weight dynamics were not taken into account in this study. Therefore, one of the future challenges is to incorporate the information of time intervals between measurements into the outlier detection process.

We acknowledge that the study of the outlier detection techniques presented in this paper are not comprehensive. Rather, this paper serves as a preliminary investigation of the topic. That is, in spite of obtaining acceptable results related to both data sets tested in this study, there were still some real cases where none of the implemented methods performed well. At last, combination of real time automatic assignment of measurements with techniques used in this study can be worth to examine in future.

REFERENCES

[1] A. E. Field, E. H. Coakley, A. Must, J. L. Spadano, N. Laird, W. H. Dietz, E. Rimm, and G. A. Colditz, "Impact of overweight on the risk of developing common chronic diseases during a 10-year period," *Archives of internal medicine*, vol. 161, no. 13, pp. 1581–1586, 2001.

[2] D. B. Sarwer, A. von Sydow Green, M. L. Vetter, and T. A. Wadden, "Behavior therapy for obesity: where are we now?" *Current Opinion in Endocrinology, Diabetes and Obesity*, vol. 16, no. 5, pp. 347–352, 2009.

[3] M. Espeland, "Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look ahead trial," *Diabetes care*, 2007.

[4] W. C. Knowler, E. Barrett-Connor, S. E. Fowler, R. F. Hamman, J. M. Lachin, E. A. Walker, and D. M. Nathan, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," *The New England Journal of Medicine*, vol. 346, no. 6, pp. 393–403, 2002.

[5] J. A. Linde, R. W. Jeffery, S. A. French, N. P. Pronk, and R. G. Boyle, "Self-weighing in weight gain prevention and weight loss trials," *Annals of Behavioral Medicine*, vol. 30, no. 3, pp. 210–216, 2005.

[6] D. M. Steinberg, D. F. Tate, G. G. Bennett, S. Ennett, C. Samuel-Hodge, and D. S. Ward, "The efficacy of a daily self-weighing weight loss intervention using smart scales and e-mail," *Obesity*, vol. 21, no. 9, pp. 1789–1797, 2013.

[7] E. Helander, M. Pavel, H. Jimison, and I. Korhonen, "Time-series modeling of long-term weight self-monitoring data," in *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, Aug 2015.

[8] E. Helander, A.-L. Vuorinen, B. Wansink, and I. Korhonen, "Are breaks in daily self-weighing associated with weight gain?" *PLoS ONE*, 2014.

[9] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, 1993.

[10] R. K. Pearson, *Exploring Data in Engineering, the Science and Medicine*, 1st ed. New York: Oxford University Press, 2011.

[11] H. W. Borchers, *pracma: Practical Numerical Math Functions*, 2015, r package version 1.8.6. [Online]. Available: http://CRAN.R-project.org/package=pracma

[12] B. Rosner, *Fundamentals of biostatistics*, 7th ed. Cengage Learning, 2010.

[13] J. L. de Lacalle, *tsoutliers: Detection of Outliers in Time Series*, 2015, r package version 0.6. [Online]. Available: http://CRAN.R-project.org/package=tsoutliers

[14] S. P. Millard, *EnvStats: An R Package for Environmental Statistics*. New York: Springer, 2013. [Online]. Available: http://www.springer.com

[15] G. E. Box and G. M. Jenkins, "Some recent advances in forecasting and control," *Applied Statistics*, pp. 91–109, 1968.

[16] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2011, vol. 734.