TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

SAEED MEHRANG

OUTLIER DETECTION IN WEIGHT TIME SERIES OF CONNECTED SCALES: A COMPARATIVE STUDY

Master of Science thesis

Examiners: Prof. Ilkka Korhonen
PhD Elina Helander
Examiner and topic approved by the
Faculty Council of the Faculty of
Computing and Electrical Engineering
on December 9, 2015

ABSTRACT

SAEED MEHRANG: Outlier Detection in Weight Time Series of Connected Scales: A Comparative Study
Tampere University of Technology
Master of Science thesis, 64 pages, 1 Appendix page
September 2016
Master's Degree Programme in Information Technology
Major: Medical Informatics
Examiners: Prof. Ilkka Korhonen
PhD Elina Helander
Keywords: big data mining, body mass analysis, time series analysis, outlier detection, self-monitoring of weight, smart and connected health, Withings

Smart and connected health technologies as part of the digitally supporting health and heathcare plans can play an explicitly important role in improving preventive healthcare and patient outcomes, decreasing costs, and speeding up the scientific discoveries. Rigorous information processing approaches, such as outlier detection and data cleaning, are therefore needed to enhance the reliability of the acquired data. A "smart electronic weight scale" is a connected sensor that regularly measures and stores time series of body mass values. The long-term self-weighing time series data, like any other time series data, may occasionally contain abnormal values which are called "outliers". The existence of these outlying values can distort or mislead the data analysis. In this thesis, detection of outliers in time series of weight measurements of 10,000 anonymous Withings weight scale users is investigated. Four point-wise outlier detection approaches are studied and compared from different aspects. These techniques are: (1) a method based on Autoregressive Integrated Moving Average (ARIMA) time series modelling, (2) moving Median Absolute Deviation (MAD) scale estimate, (3) conventional Rosner statistic, and (4) windowed Rosner statistic. The results suggest that ARIMA approach, moving MAD and windowed Rosner statistic can properly find the outliers; however, in case of facing missing data the only method which was able to ideally identify the outliers was ARIMA approach. In contrast, conventional Rosner statistic did not show acceptable outlier detection power. The computational complexity of the ARIMA approach was unsatisfactorily costly, whilst the rest of the tested techniques were quite fast in terms of computation time.

PREFACE

This thesis was done as part of the research projects in Personal Health Informatics (PHI) group in Department of Signal Processing, Tampere University of Technology, Finland between July 2015 and February 2016 under the supervision of PhD Elina Helander and Prof. Ilkka Korhonen.

Here I would like to express my deep gratitude to PhD Elina Helander, Prof. Ilkka Korhonen, Dr. Hannu Nieminen, and Prof. Misha Pavel for all of the invaluable support and recommendations they gave me during my thesis work.

Many thanks also to Withings as the provider of the data set used in this study. They enabled a new era of research by gathering a wealth of physiological data collected in everyday life.

Last, but not the least, I wish to thank my family for all their priceless encouragements and faith, specially my mother, Maryam, with her unconditional support.

Tampere, September 2016

Saeed Mehrang

Table of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS AND SYMBOLS

ADHR	Acute Decompensated Heart Failure
AM	Before Noon
AO	Additive Outlier
AUC	Area Under Curve
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BMI	Body Mass Index
CSS	Conditional Some of Squares
ESD	Extreme Studentized Deviate
FPR	False Positive Rate
HF	Heart Failure
IO	Innovational Outlier
LS	Level Shift
MA	Moving Average
MAD	Median Absolute Deviation
ML	Maximum Likelihood
ROC	Receiver Operating Characteristic
TC	Temporary Change
TPR	True Positive Rate

a_t	forecast error at time point t
B	backward shift operator
c	number of outliers in conventional Rosner statistic
C	critical value of ARIMA algorithm
d	differencing order
D	seasonal differencing order
k	moving MAD window length
L	window length in conventional Rosner statistic
L_j	outlier dynamic of type j
m	number of outliers in ARIMA algorithm
$maxit$	maximum outer loop iteration in ARIMA algorithm
$maxit.iloop$	maximum inner loop iteration in ARIMA algorithm
MED_j	median of j_{th} window

n	time series length
p	autoregressive polynomial order
P	seasonal autoregressive polynomial order
q	moving average polynomial order
Q	seasonal moving average polynomial order
s	seasonality period
t	time
$\bar{x}$	arithmetic mean of x
α	Type-I error rate
β	Type-II error rate
δ	dampening effect
θ_m	moving MAD threshold value
θ_p	autoregressive polynomial of order p
Θ_P	seasonal autoregressive polynomial of order P
ϕ_q	moving average polynomial of order q
Φ_Q	seasonal moving average polynomial of order Q
ω	magnitude of outlier

1. INTRODUCTION

According to World Health Organization (WHO) prevalence of obesity almost doubled between 1980 and 2008. Based on the estimation of countries in WHO European Region, more than 50 percent of men and women were overweight in 2008 [1]. In the same year, near 23 percent of women and 20 percent of men were obese. WHO reported also that in 2014 every third 11-year-old child is overweight or obese. Such a huge rising prevalence of obesity can be observed in United States of America as well. As reported by Obesity Rate and Trends, in the whole country more than 68 percent of adults were overweight of whom half were obese as of 2011 to 2012 [2]. In September 2015, the minimum rate of adult obesity was as big as 20 percent among all the states while, in as many as 25 states, close to one third of the citizens were obese. Moreover, according to the 2015 Youth Risk Behavior Surveillance System (YRBSS) in USA, almost 30 percent of high school students were overweight or obese.

One of the solutions that has been proved to be effective in weight-loss and weight-maintenance interventions, is self-weighing. That is, self-monitoring of weight can help those who wish to either lose weight or keep their weight at a constant level [3, 4, 5, 6, 7]. Currently self-weighing can be easily done via "smart connected weight scales" or simply "smart weight scales". These weight scales provide the capability of storing a digital version of body mass along with corresponding measurement time. These digitally available series of weight measurements are then called "weight time series" in this thesis. The analysis of these weight time series plays an important role in understanding the changes in body mass and the overall health status of individuals throughout the time [8, 9, 10]. Multivariate analysis of weight time series along with other physiological and psychological variables using connected sensors can be counted as the next frontier of healthcare informatics. At population level, the weight time series analysis can also increase the insight about the effect of climate, public holidays, and cultural factors on body mass variations in different time intervals and geographical areas [11, 12, 13]. Investigation of differences between the behavioral models of weight gainers and weight losers is another application of weight time series analysis at population level.

Since self-monitoring of weight is done during daily life and mainly in uncontrolled conditions, it can be contaminated by **outliers**. Here outliers are the measurements that are distinctly separate from the rest of the adjacent measurements or in other words, the measurements which are physiologically unlikely to occur. These outlying values can arise from conditions where different persons are using the weight scale, carrying some objects during measurement such as suitcase, or due to some external influences such as exceptionally heavy clothing. Weighing pets can be another source of producing outlying values in the case of weight time series [14]. The presence of these noticeably deviated values usually lead to remarkable distractions in the outcome of analysis. For instance, properties of a mathematical model fitted to a weight time series could dramatically change in presence of outliers. It can be even more critical if heart failure patients' weight are being screened. In this case, a deviation from baseline weight values may notify body overfluid that if neglected, may lead to mortal outcomes. Discriminating these natural deviations from outliers is exceptionally crucial and challenging. Subsequently, the problem of outliers needs to be addressed before designing any study, any intervention, or drawing any conclusion.

Basically, outlier detection procedure can be dramatically challenging depending on the nature of the data. Presence of missing data and diversity of the range of normal data usually complicate the process of identifying outliers. In the context of weight time series, a challenge may arise from temporal variation of people's adherence to self-weighing. Stopping self-weighing for a while leads to generation of missing data. The long periods of time without measurements then might cause larger than average fluctuations in weight levels [15]. The average daily variations of body mass can rise even up to 3 percent [16]. In longer periods the normal fluctuations can follow even more disparate models depending on the people's behavior. Therefore distinguishing physiologically unreasonable measurements from normal measurements would be an extremely difficult process.

In this thesis the main objective is to tackle the problem of detecting outliers in weight time series of connected scales. After a comprehensive literature analysis, a few of the most suitable statistical outlier detection methods were selected to be investigated. The chosen methods were regarded as the most fitting techniques considering the nature of the problem. The examined techniques are: (1) a method based on Autoregressive Integrated Moving Average (ARIMA) time series modelling, (2) moving Median Absolute Deviation (MAD) scale estimate, (3) conventional Rosner statistic, and (4) windowed Rosner statistic. Detailed explanation about each technique can be found in chapter 4. It should be noted that, there has not been any study of similar kind addressing detection of outliers in time series of weight measurements recorded in uncontrolled daily life conditions.

Withings as a consumer electronics company manufactures smart weight scales that can be connected with mobile phones and personal computers. This thesis work is based on a data set that includes weight time series of 10,000 randomly selected anonymous Withings weight scale users from all over the world.

This document is structured as follows. Chapter 2 discusses briefly the essence of self-weighing, variability of body mass, and weight time series properties. In Chapter 3, a short review of literature about different outlier detection techniques along with a concise explanation of context-wise categorization of univariate time series are described. Different types of outliers and categorization of outlier detection methods based on data labels are discussed in Chapter 3 as well. Chapter 4 accounts for the implemented outlier detection methods. Chapter 5 describes the used data set and some of the statistical properties of the included population. Chapter 6 includes the results and discussion corresponding to comparison of statistical performance of the used methods. Finally in Chapter 7 conclusions of the acquired results are presented.

2. THEORETICAL BACKGROUND

2.1 Essence of Self-Weighing

According to [17], the rate of mortality in population of young and middle-aged white inhabitants of North America and Europe increases when the body mass index (BMI) value exceeds $30(kg/m^2)$. Having a BMI value over $30(kg/m^2)$ means being exceptionally vulnerable to heart problems, diabetes, musculoskeletal disorders and high blood pressure [18, 14]. In addition, based on what has been presented in [19], obesity and inactivity together are the second cause of death and disability in the United States after smoking. The two mentioned factors contribute in forming "behavioral determinants of health". Therefore, refining the individuals' behavior is considered to be the key element in not only improving the health status but also decreasing the costs of healthcare. Increasing the people's awareness about the effect of their behavior on their health and well-being using what is called "self-monitoring" can be a fundamental solution. In this regard, a variety of equipment and services are available for consumers to support behavior change by means of self-monitoring that provides the capability of collecting and displaying a wealth of personal health and wellness data. In the case of self-monitoring of weight, smart connected weight scales have been recently designed and are now widely used. In fact, they help increasing the understanding about the variations of individuals' body mass as a function of time. Recently, self-monitoring of weight combined with automatic feedback have shown helpful outcomes in weight management since it increases the people's insight about both short and long-term changes occurring in their weight [4, 20, 21, 22]. Therefore, regular self-weighing can play an important role in long-term weight-loss and weight-maintenance interventions.

A series of weight measurements recorded by connected weight scales is called "weight time series". The analysis of these time series enables us to investigate the individual and population level variations of body mass. This can further lead to profound understanding about the individuals' behavior [23, 24]. Weekly and monthly fluctuations, correlation of self-monitoring frequency and weight change, as well as behavioral models are a few of the things that can be studied using weight time series data. The effect of self-weighing adherence is one of the factors that has lately been studied and claimed to be impactful in weight-loss and weight-maintenance interventions. Frequent self-weighing may have considerable positive influence on the procedure of losing or maintaining weight [20, 16]. Nevertheless, a number of studies have argued that the self-weighing by itself cannot affect the weight-loss interventions if it is added to traditional methods of weight-loss. However, it was claimed that "daily weighing" either combined with or without electronic feedback can produce a small, but significant weight-loss [6, 25]. Daily weighing also appears to be beneficial in inhibiting weight-regain after weight-loss [26]. In opposition to supporters of frequent self-weighing, cognitive behavioral interventions advised at most weekly self-weighing to prevent the discouragement caused by negligible weight losses [27, 28].

2.2 Variability of Body Mass

Weight varies due to various factors such as body fluid status, digestion, and diet. Detection of the changes in above-mentioned factors can be utilized by analysis of body mass changes. On the other hand, the current digital weight scales are equipped with a new technology that enables body fat measurements at the same time with scaling the weight. utilizing both fat mass and body weight screening, subtle variations in fat mass, lean mass, and even body fluid level can be monitored. As mentioned earlier, up to 3 percent of daily weight variations can be considered normal [15]. In addition to daily variations, body mass also varies weekly, that is typically leading to weight-gain during weekends as opposed to weight-loss during weekdays [8].

By looking into short-term variations of body mass we can obtain invaluable information about the body fluid status. The clinical importance of body fluid via body mass screening is unveiled regarding heart failure (HF) patients whose fluid balance is of great importance. Fluid retention can sometimes lead to severe impacts such as Acute Decompensated Heart Failure, therefore assessment of fluid status is crucial in managing hypervolemia (or fluid overload) in both hospitalization and after discharge of HF patients [29, 30].

The imbalance of electrolytes can also be tracked by body mass screening. The imbalance of electrolytes might be caused by excessive sweating, diuretic medicines, heart failure, or kidney disease. The mentioned imbalance is usually followed by the variations in either ionic or water levels of the body [31]. Detection of such variations can be managed by careful tracking of the body mass changes.

The effect of ambient temperature and humidity on body fluid changes can also be studied via weight time series analysis. As the temperature and humidity rises the rate of sweating also increases by up to approximately 1 liter per hour. In case of ultra endurance sports, the rate of sweating may even grow up to 3 liter per hour. Healthy individuals are able to compensate the mentioned water outtake by increasing the water intake but in case any fluid imbalance happens, there will be a noticeable change in the body mass [32].

2.3 Properties of Weight Time Series

Every time series can be decomposed into a combination of trend, seasonal, and irregular components [33]. These components are basically used for describing the properties of a time series. In this section the weight time series properties are discussed in terms of the trend, seasonal, and irregular components.

The trend component is the first and sometimes the most informative feature of the weight time series in the context of weight-loss/maintenance interventions. Here, the trend component represents the long-term variations of weight indicating if the person is either losing, gaining, or maintaining his/her weight. The second element which is called seasonal component parametrizes the periodic variations of weight such as diurnal [16], weekly [8], quarterly, and menstrual [34]. The irregular component describes the random variations of body mass. These random variations might be related to irregular changes in diet [35], physical activity [12, 36], fluid balance [37, 38], and ambient temperature [32].

In time series analysis the "stationarity" of time series is of great importance. For example, in order to model the time series using ARIMA, the staionarity condition must be met. By definition, a time series is stationary if its statistical properties such as mean, variance, and autocorrelation are constant over time [33]. On the other hand, presence of stationarity is necessary in forecasting time series values in future. If the time series properties are not meeting the stationarity criteria, then a set of actions like detrending, differencing, and log-transforming are used to stationarize the time series. According to the aforementioned principles, weight time series can be easily "non-stationary" since most of them generally contain non-zero trend, non-zero seasonal component, or varying autocorrelation through time.

It is worth mentioning that, like all other digital sensors, connected weight scales rounds the measurand (weight) to the nearest tenths value. Such a discretization inevitably leads to a discrete probability density function.

3. WEIGHT TIME SERIES AND OUTLIERS

3.1 Problem of Outliers

The problem of outlier detection has been investigated extensively since the very beginning stages of digital signal processing. Basically, an outlier (or anomaly) is defined as the abnormalities that do not follow the well-defined expected normal behaviors. The anomalies can be induced by several factors such as malicious activity, breakdown of the system, fraud, user interferences, and etc. It should be noted that detection of anomalies is distinct but related to "noise removal" [39]. Noise denotes the unwanted data which is not of interest of the data analyst; however, there is an "interestingness" in the detection and observation of the anomaly values and their occurrence in the data [40].

The outlier detection process sometimes becomes a challenging procedure owing to the following items [40, 41]:

1. Identifying and defining a normal region that comprises all the possible normal behavior is always very complex.
2. Current normal behavior may evolve throughout the time, and in case of biomedical signals the normal behavior probably varies person-wise.
3. Definition of anomaly in one application domain can dramatically be different than other domains. That means, normal behavior in one domain may cover even significantly large fluctuations; however, in another domain very tiny deviations are counted as anomaly.
4. It might be sometimes a challenging issue to find sufficient amount of training and validation data.
5. The process of outlier detection becomes notably complicated if the data contains noise and outliers at the same time.

6. Presence of missing data in the time series complicates the outlier detection process as well.

Based on the formerly published review articles and books, nature-wise the following major categories of outlier detection techniques can be considered. Classification based [42, 43], clustering based [42, 44, 45], statistical based [46, 47, 48, 49, 50, 51], and information theoretic [52, 53, 54]. Each one of the above-mentioned categories has broad applicability in different domains. Further description of the each technique can be found in cited references.

The studied methods in this thesis fall into the category of statistical based outlier detection techniques. The underlying concept of outlier detection techniques under this category is: "An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed" [55]. There is a key assumption in all the statistical based outlier detection techniques and that is: "Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model" [40]. These techniques work based on a statistical inference test that detects the data points that have a low probability to be generated by the model that represents normal behavior.

Dimensions of the data in process is a factor that restricts the number of alternatives of outlier detection techniques. In general, each time series is a collection of data instants. These data instants can nature-wise be described by either one variable (univariate) or multiple variables (multivariate) at a time [40, 42]. The weight time series in this study are univariate because for each time point there is only a one-dimensional variable which is body mass in kilograms. Consequently, the outlier detection approaches investigated in this study are restricted to the category of univariate statistical based techniques. In following sections the detailed description of the underlying principles concerning univariate time series categorization, types of outliers in univariate time series, as well as types of outlier detection methods based on data labels are presented.

3.2 Types of Univariate Time Series

The problem of outlier detection in univariate time series data is considerably diverse and highly dependent on the context and nature of the data. Context wise, a set of univariate time series within a data set can be divided into four different categories as follows [41]:

1. Periodic and synchronous that stands for constant periodicity as well as being temporally aligned.
2. Aperiodic and synchronous that implies on temporally aligned time series that are not periodic.
3. Periodic and asynchronous suggesting a category in which the time series follow a constant time period but not a same starting time point.
4. Aperiodic and asynchronous time series that neither have periodicity, nor are temporally aligned.

Identification of outliers become more complex in aperiodic and asynchronous category in comparison with the other three categories. The main reason for this complexity stems in lack of similarity criteria for selection of outlying values. In other words, detection of anomalous behavior is much easier when all of the time series follow a similar periodic and synchronous pattern. In contrast, if one of the time series is allowed to follow a distinct unknown pattern, the detection procedure cannot be established based on similarity tests using the rest of the time series available in the data set. Basically, weight time series fall into the category of aperiodic and asynchronous time series.

3.3 Types of Outliers in Univariate Time Series

The outlying values (or anomalies) can be divided into three groups, (1) point anomalies, (2) contextual anomalies, and (3) collective anomalies [40]. Point anomalies are the values that abnormally deviate from the rest of the data points, whereas contextual anomalies stand for the cases where an identical datum can be considered as normal in one context in spite of being anomaly in another context. The collective anomalies represent a category in which a set of consecutive instances of data are considered anomalous although none of them are abnormal individually. The majority of the outliers in this study fall into the category of point anomalies. There are a few cases where collective anomalies were observed in the data set but they are not considered in this study since investigation of collective anomalies fall outside the scope of this thesis.

3.4 Types of Outlier Detection Methods Based on Data Labels

Depending on the availability and the extent of data labels, the anomaly (outlier) detection methods usually operate in one of the three modes, namely supervised, semi-supervised, and unsupervised [40]. In supervised mode there is usually a set of training data where all of the normal and anomalous data points are labeled by an expert. The goal of supervised methods is to find a predictive model using the training data that classifies the unseen data points in the test data. In this regard, one of the challenges is the process of labeling in which there might be human faults. A solution for this challenge can be employment of different experts to reduce the false labeling. Other than that, power of the model in predicting test data labels collapses if the training data set does not cover all the possible combinations of anomaly occurrences. However, this can be somehow addressed by artificial injection of anomalies in training phase. The semi-supervised mode similarly constitutes training and test phases whereas in training phase only the normal data points will be labeled so that in test phase everything that differs from the normal class will be considered as anomalies. In detail, the training phase in semi-supervised mode tries to find a model that describes the normal behavior which will be used as ground truth in test phase. Lastly, the unsupervised mode stands for the cases where there is no training data. In this case the assumption is that the anomalous values occur rarely in the test data. The weakness of unsupervised methods is therefore the cases where the mentioned assumption is violated that often leads to high false alarm rate [40].

4. OUTLIER DETECTION METHODS

In this section the four studied outlier detection techniques are described in detail. The four techniques are namely (1) ARIMA based approach, (2) moving MAD, (3) conventional Rosner statistic, and (4) windowed Rosner statistic. Please note that in this thesis "ARIMA approach" and "ARIMA technique" both refer to "ARIMA based approach".

It should be emphasized here that none of the studied methods take the time domain features of weight measurements into account. In other words, they do not differentiate if two consecutive points are apart with hours of time difference or months of time difference. Therefore, measurement time does not play any role in outlier detection process.

4.1 ARIMA Approach

A non-seasonal time series named X_t follows an ARIMA process of order (p, d, q) if the d_{th} difference of the X_t can be considered as an ARMA (p, q) process. The whole ARIMA model is represented in equation 4.1 where autoregressive part θ_p and moving average ϕ_q are polynomials of order p and q, respectively. B represents the backward shift operator, and d shows the order of differencing as well. Here autoregressive (AR) term suggests 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 that the forecast error at time instant t can be explained by q past forecast errors at time points $t-q, t-q+1, ..., t-1$ [56, 57, 14].

$$\theta_p(B)(1-B)^d X_t = \phi_q(B)\omega_t \tag{4.1}$$

Now a general representation of ARIMA models called Seasonal Autoregressive Integrated Moving Average model is shown in equation 4.2 which is the basis for detection of outliers based on ARIMA approach in this study.

$$\Theta_P(B^s)\theta_p(1-B^s)^D(1-B)^d X_t = \Phi_Q(B^s)\phi_q(B)\omega_t \tag{4.2}$$

Θ_P, θ_p, Φ_Q , and ϕ_q are polynomials of order P, p, Q, and q, respectively. Here seasonal autoregressive polynomial Θ_P, seasonal moving average polynomial Φ_Q as well as D times seasonal differencing were added to the model to take seasonal components of the time series into account. In addition, the variable s defines the number of time points until the pattern repeats again (or seasonality period). In general, seasonality refers to repeating patterns of time series such as diurnal, weekly, monthly, and annual periods. A seasonal ARIMA model employs differencing at lags equal to the periodicity of the time series in order to eliminate the additive effects of seasonal components. The above-mentioned model will be considered stationary if $D = d = 0$ and roots of the polynomials on the left hand side of equation 4.2 are all out of unit circle [58]. Stationarity means having constant statistical properties such as mean, variance, and autocorrelation over time.

There can be four types of outliers depending on the definition introduced in [58]. A level shift outlier (LS), an innovational outlier (IO), an additive outlier (AO), and a temporary change (TC) are of those four types of outliers which can be detected within the time series. Now, suppose that the series X_t is an n point time series in which there are m outlier points. For the sake of simplicity it was assumed that X_t is nonseasonal, hence the corresponding ARIMA model representing it is as equation 4.3.

$$X_t = \frac{\theta(B)}{\alpha(B)\phi(B)} a_t \ , \ t = 1, ..., n \tag{4.3}$$

$\theta(B)$, and $\phi(B)$ are polynomials with all roots out of the unit circle expressing the autoregressive and moving average components of the model, respectively. In contrast, $\alpha(B)$ is a polynomial with all roots on the unit circle denoting d times differencing in case the time series is non-stationary. Moreover, the term a_t represents the forecast error.

To consider the influence of outlying values in the model it is needed to expand the definition introduced in equation 4.3 in such a way that for every time point where a potential outlying point might have occurred the effect of that outlier will be added to model. If we assume that only one of the four types of outliers occurs at time point t_1 then the expanded model looks like equation 4.4.

$$X_t^* = X_t + \omega \frac{A(B)}{G(B)H(B)} I_t(t_1) \tag{4.4}$$

Where X_t is the ARMA process described in equation 4.3, $I_t(t_1) = 1$ if $t = t_1$, and $I_t(t_1) = 0$ otherwise. This $I_t(t_1)$ is an indicator flag that turns to 1 for occurrence of the outlier, ω stands for the magnitude of outlier, and $\frac{A(B)}{G(B)H(B)}$ denotes corresponding dynamic of each outlier type [58]. These dynamics are illustrated in equation 4.5.

$$\frac{A(B)}{G(B)H(B)} = \begin{cases} \frac{\theta(B)}{\alpha(B)\phi(B)} & \text{for innovative outlier (IO),} \\ \frac{1}{(1-\delta B)} & \text{for temporary change (TC),} \\ 1 & \text{for additive outlier (AO),} \\ \frac{1}{(1-B)} & \text{for level shift (LS).} \end{cases} \tag{4.5}$$

As can be seen AO and LS are two boundary state of TC where the parameter δ equals 1 for the LS and 0 for the AO, respectively. A TC produces an effect at time point t with magnitude ω and it slowly decays by a pace specified by δ. The default value for the parameter δ that controls the dampening effect is 0.7 although analyst can define it accordingly. To explain more about the equivalent effects of outliers on the time points, if the outlier is an AO then the effect is immediate like an impulse however for LS there will be a step change that raises the level of succeeding time points. It should be noted that the effect of AO, LS, and TC are all independent of the model while in case of an IO it depends on the stationarity, seasonality, and parameters of the time series model [58].

Now, a generalized version of the model introduced in equation 4.4 is expressed in equation 4.6 for a time series of length n in which there are m outliers. The $I_t(t_j)$ flag turns to 1 in case there is an outlying effect at time point t_j with the dynamic L_j corresponding to one of the four outlier types declared in equation 4.5.

$$X_t^* = \sum_{j=1}^{m} \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\alpha(B)\phi(B)} a_t \tag{4.6}$$

After explaining the mathematical concepts of ARIMA outlier detection technique, it is time to describe its iterative outlier detection process. Detection procedure can be divided into three repeating steps as follows [14, 58, 59],

Step I

First the algorithm computes the initial model parameters based on the maximum likelihood (ML) or minimize conditional some of squares (CSS) defined within the ARIMA parameter selection. Afterward, it calculates four different τ-statistics corresponding to four types of outliers for every point of the time series. Now the algorithm chooses the biggest absolute value of each time point τ-statistic as a dominant outlying effect. Then it compares the dominant τ-statistic with the critical value (C) which was defined for the function in advance in order to decide whether the time point can be considered as an anomaly or it is a valid data point that must be kept unchanged.

Step II

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 ascertain that 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. This repeating procedure is done as many as the *maxit.iloop* value which is defined inside the options of the algorithm and is called the "maximum inner loop iteration". The default value of *maxit.iloop* equals 4 which means the algorithm repeatedly computes the new τ-statistics four times. At this point, the identified outliers will be removed from the time series.

Step III

The output of the algorithm is ready unless the user wants to repeat the two above-mentioned steps for the time series obtained after removing the detected outliers in previous step. This is allowed by increasing the value of an option named *maxit* to higher than the default value which is 1. If the value of *maxit*, that is called the "maximum outer loop iteration", is set to values bigger than 1, then a new ARIMA model will be fitted to the time series based on Maximum Likelihood or Conditional Some of Squares criteria. In this case the algorithm repeats the two above-mentioned steps until reaching the *maxit* value.

4.2 Moving Median Absolute Deviation Scale Estimate

Detection of outliers using median absolute deviation (MAD) is much simpler from both computation and implementation point of views compared to ARIMA approach. This method implements a moving window of length k centered at each time point in which two principle values are calculated. First value is median of the data points and second one is the MAD scale estimate. Based on equations 4.7 and 4.8 the MAD scale estimate is a measure of deviation of data points from median of the corresponding window. Every data point on which the window is centered will be compared to the corresponding MAD scale estimate. If there is an absolute deviation of bigger than MAD scale estimate, then the data point is served as an outlier and should be treated accordingly [60, 61, 62]. Equations 4.7 and 4.8 explicitly illustrates how MAD scale estimate is measured [14, 63].

$$MED_j = median(X_j) \tag{4.7}$$

$$MAD\ scale\ estimate = \theta_m \times median(abs(X_j - MED_j)) \tag{4.8}$$

Here X_j is a window of length k centered at the data point j which is tested for outlier decision. MED_j represents the median of the window while variable θ_m controls the sensitivity of the algorithm. The bigger the θ_m the less sensitive the algorithm will be. It should be noted that for the $\frac{k-1}{2}$ time points at the beginning and the end of the time series the MAD scale estimate cannot be computed using a centered window. In this case there are few alternatives such as leaving these data points unchanged, appending the series from both ends as long as $\frac{k-1}{2}$ data points, or using right-sided and left-sided windows for the beginning and the end of time series, respectively. In this study the time series are appended from both ends to facilitate centered window implementation.

The two variables, window length k and threshold value θ_m, are the user defined controlling parameters of the moving MAD algorithm. By changing these two variables, sensitivity and specificity of the moving MAD can be adjusted accordingly.

4.3 Conventional Rosner Statistic

Conventional Rosner statistic is an outlier detection procedure applicable for those time series that are normally distributed after exclusion of c outlying points. This is the reason that allows no more than $c = [\frac{n}{10}]$ outlying data points for a time series of length n. Following steps must be taken to achieve the result of Rosner statistic [50].

Step I

To initialize the algorithm, first the c extreme studentized deviate (ESD statistics) values must be calculated according to equation 4.9,

$$ESD\ statistic = max_{i=1,\ldots,n} \frac{|X_i - \bar{X}|}{s}. \tag{4.9}$$

where $\bar{X}$ and s are the mean and standard deviation of the series, respectively. The main idea of conventional Rosner statistic is that the first ESD statistics (or the most extreme deviation) corresponding to X_n (that is the most outlying data point) is computed based on the full sample size. Then, for estimation of the second ESD, first the most outlying point (X_n) must be removed from the time series and this procedure needs to be continued until reaching the c_{th} outlying data point with its equivalent ESD. Therefore, at the end of this step a series of ESDs are calculated based on the sample size of $n, n-1, n-2, ..., n-c+1$, consecutively [14, 63, 64].

Step II

In order to determine which one of the c potential anomalies are detected correctly, their equivalent ESD statistics are successively compared with corresponding critical values of ESD statistics (as shown in Appendix A) in each sample size of $n, n-1, ..., n-c+1$ [65]. In detail, the decision is made based on the following procedure, if ESD of the c_{th} outlying point (the least extreme studentized statistic) is bigger than corresponding c_{th} critical value then all of the suspected c data points are outliers. Otherwise, this point is eliminated from the set of outliers and similarly the second least extreme outlying point is tested if it is greater than its equivalent critical value. This procedure continues until all outlying data points with ESDs bigger than their equivalent critical values are removed or all of the c potential outlying points were tested [50].

The diagnostic performance of conventional Rosner statistic can be controlled by setting the value of α (probability of Type-I error) or the significance level of the test which is a value between 0 and 1. Accordingly, critical values of ESD statistics change when different values of α are set. In other words, by increasing the probability of Type-I error, the critical values of ESD statistics decrease allowing more freedom to classify a point into the category of outliers. In fact, the role of α is controlling the probability of wrong rejections of the null hypothesis. Here the null hypothesis (H_0) proposes that there are no outliers in the time series whilst the alternative hypothesis (H_1) suggests existence of up to c outlying points. By setting bigger values of α, probability of wrong rejections of null hypothesis rises and correspondingly causes reduction of specificity. On the other hand, increasing the probability of Type-I error (α) coincides with decreasing the probability of Type-II error (β) that is defined as not rejecting the null hypothesis when in fact the alternative hypothesis is true. The more the Type-II error probability is shrunk, the more sensitive the algorithm becomes.

4.4 Windowed Rosner Statistic

Windowed Rosner statistic was implemented after observing the weak points of the conventional Rosner statistic. Windowed Rosner statistic operates in such a way that the time series is first divided into 50 percent overlapping windows of length L. Here L is a user defined variable and it has to be an even number. Then the ESD statistics are computed within each window in a similar manner described in the two-step procedure of conventional Rosner statistic (Section 4.3). A point is labeled as an outlier if and only if it is found to be outlier in two adjacent overlapping windows. Note that for the first and last $0.5L$ points at the beginning and ending of the time series there are not overlapping windows. In this case the classification decision is done by looking into the only window comprising the aforementioned points.

Diagnostic performance of windowed Rosner statistic can be controlled using both window length (L) and Type-I error rate (α).

5. MATERIALS

This study is based on a data set that includes weight time series of 10,000 randomly selected anonymous Withings (Withings, Paris, France) weight scale users from all over the world. All of the subjects whose data were used in this study gave their consent to allow use of their anonymous data for research purposes at the time of setting up their user accounts as part of approving the Terms and Conditions (see [66]). All the data processed in this study were anonymized and identification of an individual user was not possible. No experimental procedures or intervention of any kind was provided for the users. In following section the detailed description of the underlying data set can be found.

5.1 Description of Underlying Data Set

The data set contains 5,534,898 measurements altogether where anonymous Withings weight scale users from 109 countries are included. A few of the basic demographics of the data set are shown in Table 5.1.

The distribution of the measurement times within the day is the first statistical parameter that is sketched in Figure 5.1. The daily measurement times are important in the sense that people usually have lower weight values in the morning compared to the evening [16]. In other words, for comparing subsequent weight

Table 5.1 *Demographics of the data set. The values are expressed as mean and standard deviation (in parenthesis) or percentage.*

Population size	**Age**	**BMI**	**Males (%)**	**Number of measurements**	**Measurement period in days**
10, 000	44.0 (11.1)	26.8 (5.3)	65.9	553.5 (474.1)	1094.4 (434.9)

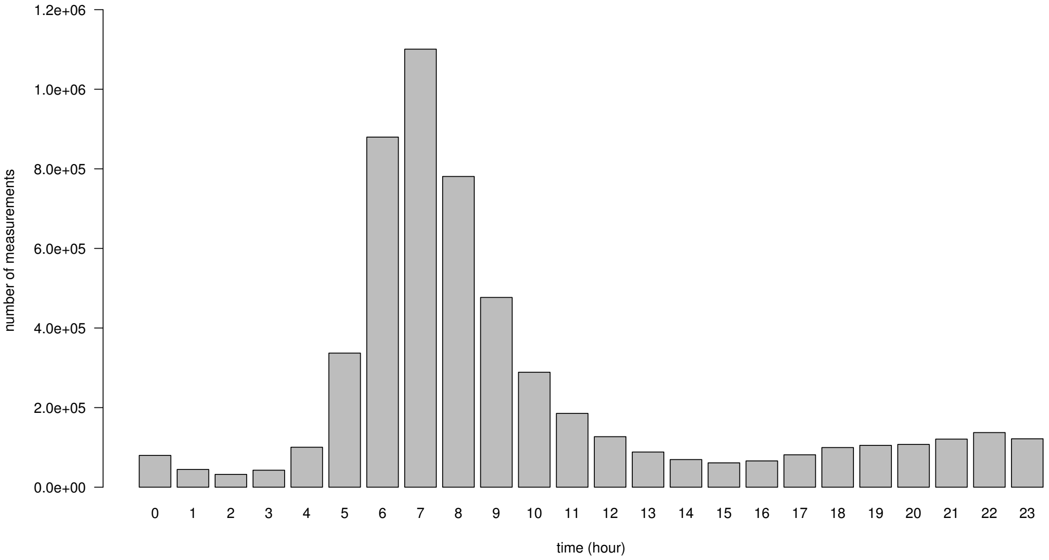


Figure 5.1 *Distribution of measurement times within day.*

values, self-weighers and health coaches should take the time of each measurement into account, as well. Here, the studied population tended to weigh themselves more in the morning between 5 to 10 AM as clearly shown in Figure 5.1. As mentioned earlier the highest bars belong to morning hours in comparison with afternoon and evening hours.

Yet another interesting parameter in terms of self-weighing is the frequency of self-monitoring. Figure 5.2 shows how often the weight scale users tended to record their weights. The most frequent time intervals between consecutive measurements are 12 to 24 and 24 to 36 hours. That means, most of the users repeated their recordings between 12 to 36 hours after their previous recordings.

Figure 5.3 shows the average weekly measurement activity of all of the 10,000 subjects in the data set. The following procedure was done to obtain the weekly measurement activities of each subject. Firstly, for each subject the number of weekdays and weekend days in which there were at least one weight measurement were counted. Next, all of the weekdays and weekend days between the first and the last measurement were counted. Then, the percentage of each measured weekdays and weekend days with respect to the total number of the equivalent weekdays and weekend days in the measurement period of each subject were calculated. Finally, the average percentage of measurement activity in each day of the week was computed for all of

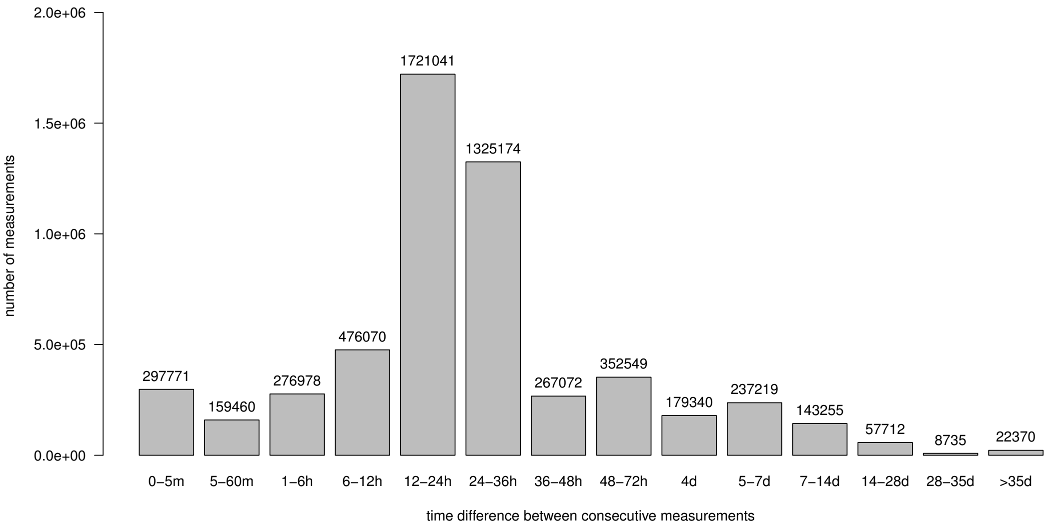


Figure 5.2 *Frequency of self-weighing within different time intervals. In the figure the letters m, h, and d are abbreviations of minutes, hours, and days, respectively. The bars are labeled with the total number of weight measurements corresponding to each time difference within the whole data set.*

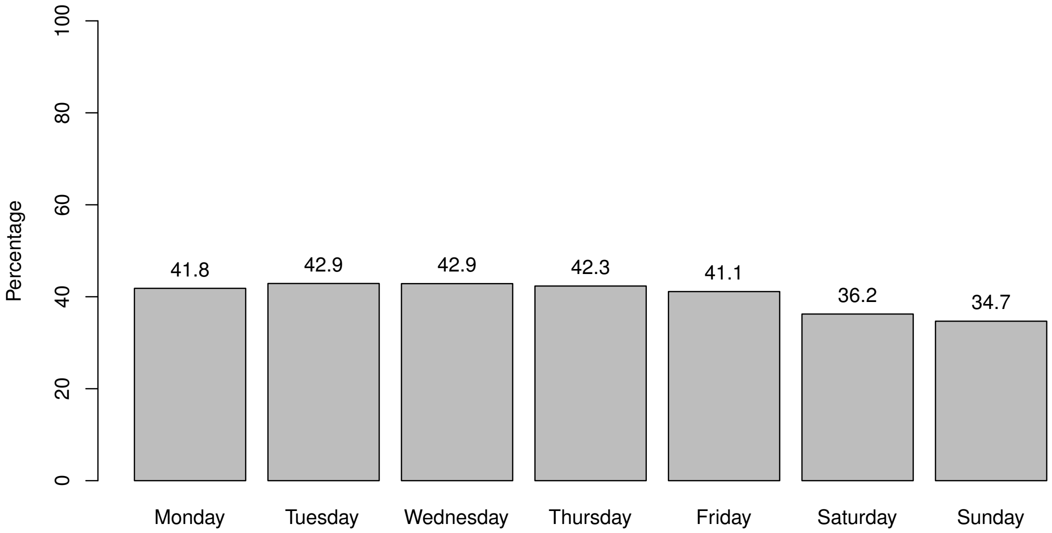


Figure 5.3 *Average percentages of weekly measurement activity over the whole data set.*

the 10,000 subjects. According to Figure 5.3, users weighed themselves more often during weekdays than weekend days.

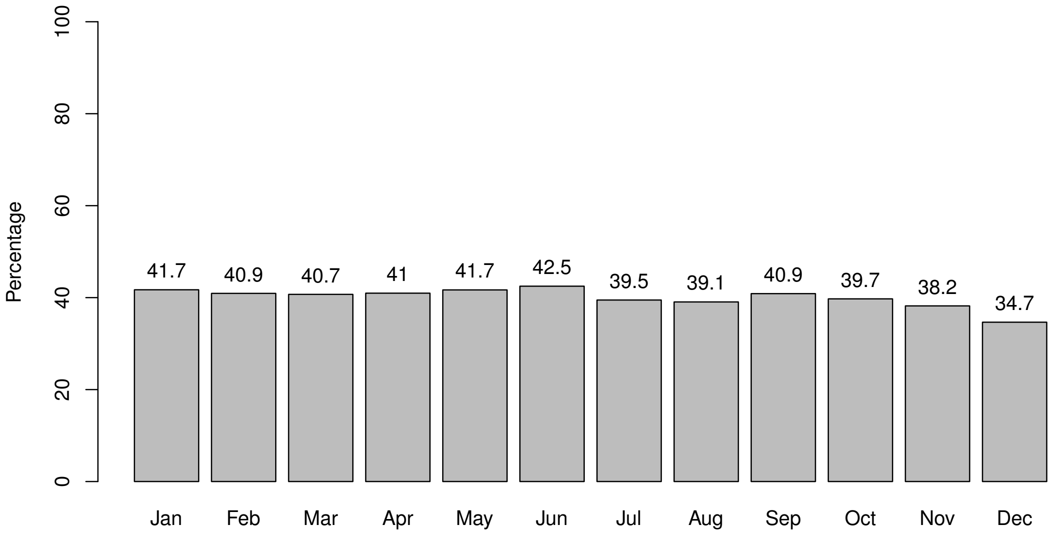


Figure 5.4 *Average percentages of monthly measurement activity over the whole data set.*

With a similar approach to weekly measurement activity, the monthly measurement activities of the studied population can be inspected. They are calculated by counting the number of unique days in each month, in which there were weight measurements, divided by the total number of days in the corresponding month. Averaging the subject-wise monthly measurement activities over the whole data set, Figure 5.4 was resulted.

The subjects' adherence to self-weighing clearly decreased during July and August as well as November and December. The reason for such a decrement may originate in the occurrence of summer vacations in northern hemisphere countries and Christmas holidays in the aforementioned months, respectively.

5.2 Test Sets

Examination of the four outlier detection methods was done using two different 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 originally contained outliers. The reason for using simulated outliers was to test the performance of outlier detection techniques under different circumstances. Testing different controlling parameters of each method was another aim of employment of simulated outliers. After finding the strength, weaknesses, and best controlling parameters of each method, the performance of the algorithms were explored separately over both simulated and real test sets.

5.2.1 Simulated Test Set

A subset of 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, was chosen as simulated test set. Any time series containing possible anomalies were excluded from this subset and replaced with another randomly selected time series. The total number of weight measurements included in this subset is 6494.

20 clean weight time series were randomly selected and intentionally corrupted with normally distributed outliers. 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 selected 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 that corresponds to 4.5 percent of the data points.

5.2.2 Real Test Set

A subset of 20 time series originally containing outliers were randomly selected among a set of time series suspected to contain outliers. In order to identify the subjects whose weight time series were probably contaminated by outliers, the whole 10,000 weight time series were fed into moving MAD technique. Here moving MAD was chosen as the benchmark of outlier detection since it showed quite acceptable performance in previous studies [14]. Then among the whole data set, those time series that at least one of their measurements were detected as outlier were preselected and filtered out. These preselected time series formed a group of subjects whose measurements were expected to contain outlying values. In next step, for final selection of 20 time series out of the aforementioned preselected group, random selection and visual inspection were employed. That is, a time series was first randomly selected and then visually assessed if it contained at least one outlier. If the randomly selected time series was confirmed of having outliers, then it was qualified to be placed in the real test set. Outliers were then annotated by the author of the thesis.

For the above-mentioned preselection step, moving MAD controlling parameters were chosen in such as way that the algorithm became highly sensitive. Since there was a visual inspection phase after the preselcetion, it was preferred to decrease the chance of missing contaminated time series. Altogether 169 time series were preselected before random picking and visual inspection phases. The number of weight measurements included in the real test set altogether was 14112 in which 68 points were visually identified to be outliers.

5.3 Data Analysis

The data analysis and implementation of the algorithms were done using R version 3.2.1, on a 64-bit Intel Core i7 3.60 GHz processor, with 16 GB RAM.

The ARIMA algorithm was deployed from "tsoutliers" and "forecast" packages [67, 68, 69]. Besides, conventional Rosner statistic algorithm was exploited from the "EnvStats" package [70]. In addition, moving MAD was a modified versions of Hampel filter that has been extensively used in digital filtering [60, 61, 62]. Windowed Rosner statistic was also developed based on the conventional Rosner statistic. Moving MAD and windowed Rosner statistic algorithms were implemented by the author since they were not available in any R packages.

6. RESULTS AND DISCUSSION

After implementation of the four outlier detection approaches explained in previous chapter the results of the outlier detection processes are shown and discussed in current chapter. There are three main sections in this chapter discussing the performance of the implemented techniques firstly over simulated test set and secondly over real test set. The last section comprises comparison of the observed weaknesses and strengths of the studied techniques.

Diagnostic performance of each technique with respect to different cut-off (threshold) values of a controlling parameter is investigated by looking into corresponding receiver operating characteristic (ROC) curves at the beginning of the two first sections of this chapter. Basically, the accuracy of a method in discrimination of diseased cases (here outliers) from normal cases is evaluated using ROC curve analysis [71, 72]. In a ROC curve the true positive rate (TPR, i.e. the rate of correct detection of diseased cases) is plotted as a function of the false positive rate (FPR, i.e. the proportion of normal cases that are wrongly detected as outliers) for different cut-off values of a controlling parameter. Each point on the ROC curve denotes a pair of TPR and FPR values corresponding to a particular decision threshold. It is worth mentioning that TPR and sensitivity are equivalent by definition and can be used interchangeably. While, FPR is equivalent to 1−specificity (1 minus specificity), meaning that the lower the FPR, the more powerful a method is in terms of finding normal cases.

Principally the points on ROC curves that are closer to the point (FPR=0,TPR=1) represent the best cut-off values in terms of diagnostic performance [73]. The best cut-off values basically provide the highest sensitivity and specificity rates regarding a controlling parameter.

Furthermore, in ROC analysis, the variable named area under the ROC curve (AUC) is estimated in order to compare the diagnostic performance of two or more methods in a test. AUC is a measure of how well a technique distinguishes between diseased and normal cases with respect to different values of a controlling parameter [74].

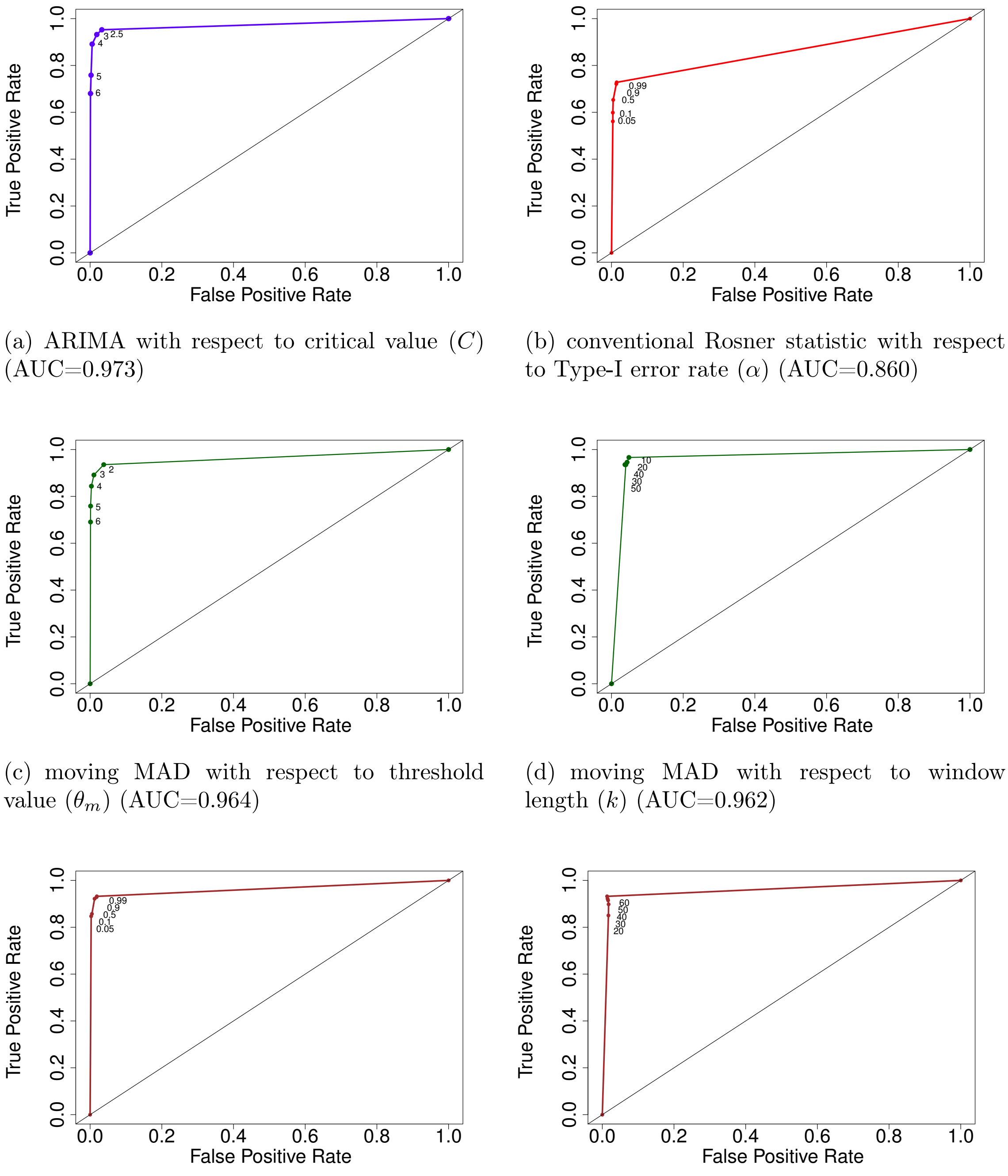
6.1 Results of Simulated Test Set

ROC curves of each of the examined techniques corresponding to simulated test set can be observed in Figure 6.1. The best parameters showing the highest average statistical sensitivity and specificity of each algorithm were identified based on ROC curves. By definition the points on ROC curves that are closer to the point (FPR=0,TPR=1) represent the best controlling parameters in terms of classification [73]. These parameters are depicted in Table 6.1. The best average statistical performance of each technique in terms of sensitivity and specificity using the identified parameters can be seen in Table 6.2.

Average processing time of the studied techniques for each time series in simulated test set can be explored in Table 6.3. Accordingly, the differences in terms of computational complexity are comprehended.

The ROC curves in Figure 6.1(a) and Figure 6.1(b) demonstrate the evaluation of different critical values (C) and Type-I error rate values (α) regarding ARIMA approach and conventional Rosner statistic, correspondingly. The curves depicted in Figure 6.1(c) and Figure 6.1(d) reveal the performance of moving MAD regarding the two controlling variables, threshold value (θ_m) and window length (k), sequentially. Similarly, Figure 6.1(e) and Figure 6.1(f) illustrates the performance of windowed Rosner statistic concerning Type-I error rate value (α) and window length (L), respectively.

In order to recognize the best choices of θ_m and k for moving MAD, first the window length (k) was arbitrarily chosen to be 30, while different threshold values (θ_m) were tested (Figure 6.1(c)). Thereafter, by choosing and keeping the best recognized threshold value unchanged and varying the window length, the best value of k was obtained (Figure 6.1(d)). In the same way, the best choices of Type-I error rate values (α) and window length (L) for windowed Rosner statistic were determined. First the window length was arbitrarily chosen to be 50 while different values of Type-I error rate were examined (Figure 6.1(e)). Then by choosing the best value of Type-I error rate and keeping it unchanged, different window length values were evaluated (Figure 6.1(f)).



(a) ARIMA with respect to critical value (C) (AUC=0.973)

(b) conventional Rosner statistic with respect to Type-I error rate (α) (AUC=0.860)

(c) moving MAD with respect to threshold value (θ_m) (AUC=0.964)

(d) moving MAD with respect to window length (k) (AUC=0.962)

(e) windowed Rosner statistic with respect to Type-I error rate value (α) (AUC=0.963)

(f) windowed Rosner statistic with respect to window length (L) (AUC=0.957)

Figure 6.1 *ROC curves of the simulated test set corresponding to (a) ARIMA approach with respect to critical value (C), (b) conventional Rosner statistic with respect to Type-I error rate (α), (c) moving MAD with respect to threshold value (θ_m), (d) moving MAD with respect to window length (k), (e) windowed Rosner statistic with respect to Type-I error rate (α), and (f) windowed Rosner statistic with respect to window length (L). For ARIMA, five critical values $C = \{6, 5, 4, 3, 2.5\}$ were tested. For moving MAD five threshold values $\theta_m = \{6, 5, 4, 3, 2\}$ as well as five window length $k = \{10, 20, 30, 40, 50\}$ were examined. Conventional Rosner statistic was assessed by five Type-I error rate values $\alpha = \{0.05, 0.1, 0.5, 0.9, 0.99\}$. Windowed Rosner statistic was also tested by five Type-I error rate values $\alpha = \{0.05, 0.1, 0.5, 0.9, 0.99\}$ as well as five window length $L = \{20, 30, 40, 50, 60\}$.*

Table 6.1 *The best controlling parameters of the implemented algorithms regarding simulated test set.*

Variable	Value
ARIMA critical value (C)	2.5
moving MAD threshold value (θ_m)	2
moving MAD window length (k)	10
conventional Rosner Type-I error rate (α)	0.99
windowed Rosner Type-I error rate (α)	0.99
windowed Rosner window length (L)	60

Table 6.2 *The best statistical performance of implemented outlier detection techniques in simulated test set.*

Methods	Average Sensitivity	Average Specificity
ARIMA	0.952	0.967
moving MAD	0.966	0.951
conventional Rosner statistic	0.728	0.986
windowed Rosner statistic	0.932	0.987

Table 6.3 *Average processing time of implemented outlier detection techniques for each time series in simulated test set.*

Methods	time (s)
ARIMA	65.411
moving MAD	0.014
conventional Rosner statistic	0.003
windowed Rosner statistic	0.019

6.1.1 Results of Simulated Test Set: ARIMA Technique

The ARIMA outlier detection method was examined only by Critical value (C) as a controlling parameter. The ARIMA technique can also be controlled by maximum number of inner loop and outer loop iterations explained in Chapter 4. However, increasing the number of iterations dramatically increases the computation time. Indeed, in all stages of the current study the two above-mentioned variables were set to their default values as in "tsoutliers" package [67]. The default values are as follows, inner loop iteration ($maxit.iloop = 4$) and outer loop iteration ($maxit = 1$).

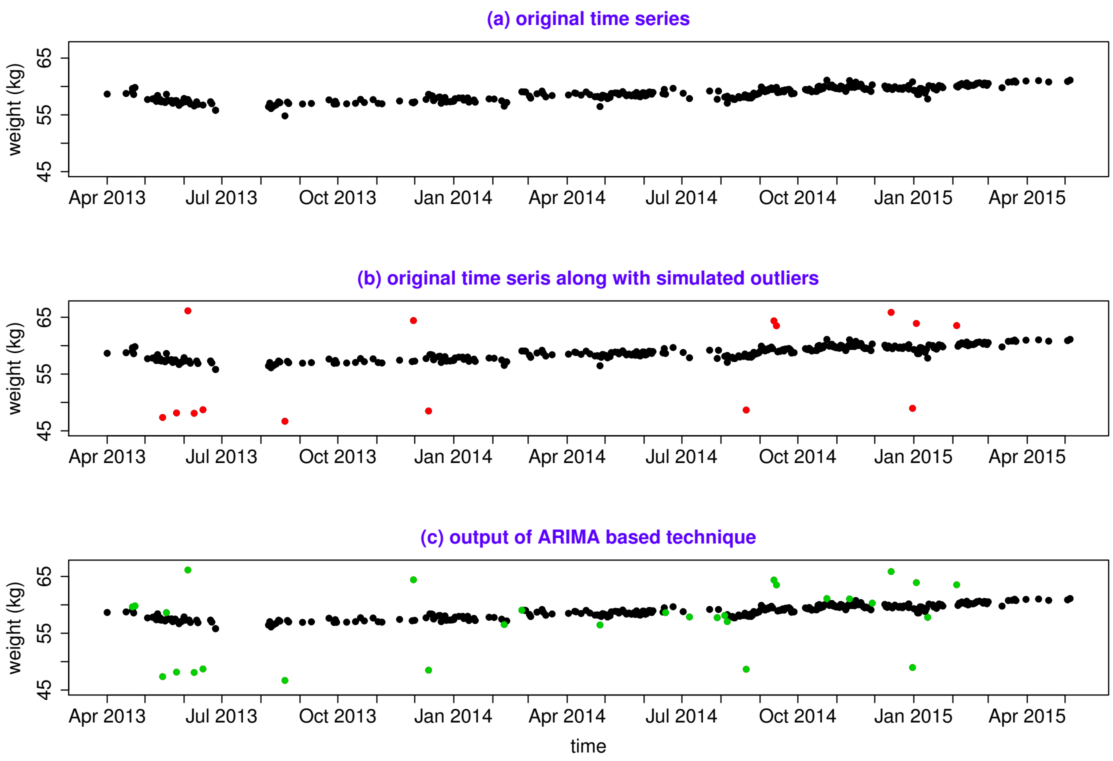


Figure 6.2 *(a) original time series before adding simulated outliers, (b) original time series along with simulated outliers, represented by red dots, and (c) output of ARIMA technique ($C = 2.5$) where green dots represent the points that the algorithm identified as outliers.*

The best classification performance was obtained with $C = 2.5$ as it led to closest point to (FPR=0,TPR=1) coordinate on ROC graph (Figure 6.1(a)). The average sensitivity and specificity values corresponding to the selected critical value were equal to 0.952 and 0.967, respectively.

Strengths

The ARIMA outlier detection technique performed quite well over simulated test set. The reason for such a strength is that the detection of outliers is done by taking the sequential aspects of weight time series into account. In other words, by iteratively fitting Autoregressive Integrated Moving Average models to each time series, most of the artificially added outliers were spotted. Figure 6.2 depicts one of the sample time series where artificial (simulated) outliers were added to a clean time series. The result of ARIMA technique can be observed in Figure 6.2(c) where almost all of the outliers were detected correctly.

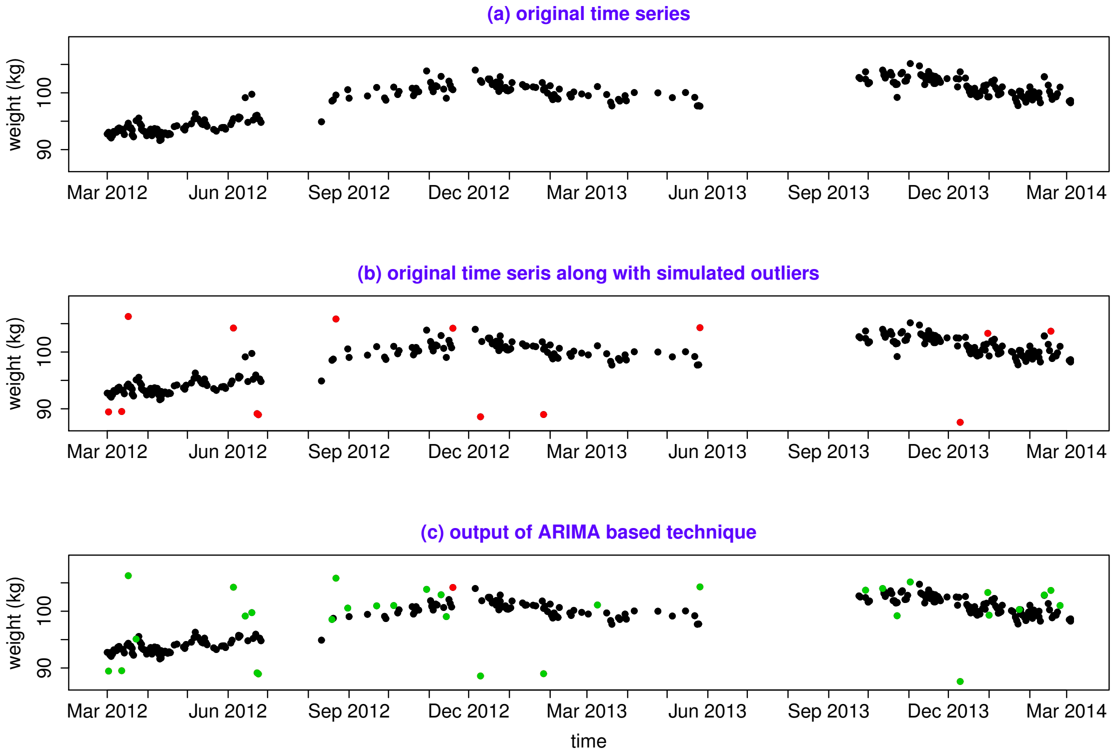


Figure 6.3 *(a) original time series before adding simulated outliers, (b) original time series along with simulated outliers, represented by red dots, and (c) output of ARIMA technique ($C = 2.5$) where green dots represent the points that the algorithm identified as outliers.*

One of the most important assets of ARIMA outlier detection technique is the ability to spot the outlying values occurring immediately before and after long gaps of measurements. Figure 6.3 clearly shows how the points lying on the edges of measurement gaps were accurately identified. The simulated outlier which was placed at the beginning of the second measurement gap (about June 2013 on time-axis) looks almost in the same weight-axis level as the data points after the gap. This is one of the most challenging cases where detection of the outlying value can only be done by taking the dynamics of weight time series into consideration.

Weaknesses

Considering the weight time series shown in Figure 6.2(c) and Figure 6.3(c), there are a few points which were wrongly detected as outliers. The average specificity of the algorithm is roughly 97 percent, which means false positive rate of the algorithm should be improved (Table 6.2).

Another weakness of the ARIMA technique is being computationally expensive. For each weight time series in simulated test set the average processing time of the ARIMA technique was approximately 69.411 seconds using $C = 2.5$ (Table 6.3). It is suspected that, the amount of time this algorithm needs for identification of outliers may go beyond hours in case the time series length falls in the range of thousand points. Complexity of the dynamics of time series may also affect the computation time of the algorithm. That means, the more fluctuation in weight values the heavier the computation of the algorithm might be.

6.1.2 Results of Simulated Test Set: Moving MAD

Moving Median Absolute Deviation (MAD) scale estimate is the second method implemented in this study. It involves two controlling parameters named θ_m and k that allow finding the optimal statistical performance. Based on Figure 6.1(c) and Figure 6.1(d) the best threshold value (θ_m) was equal to 2 and the best window length (k) was equal to 10. Accordingly, the average values of sensitivity and specificity were 0.966 and 0.951, sequentially (Table 6.2).

Strengths

The results of this technique for the simulated test set showed quite reasonable average sensitivity. Figure 6.4 shows how the moving MAD algorithm performed in one of the simulated cases. All of the red dots in Figure 6.4(b) were replaced by green dots in Figure 6.4(c) meaning that they were all accurately detected.

The low computation time of moving MAD is its another asset. The average computation time of each weight time series using the best identified controlling parameters was estimated to be 0.014 seconds (Table 6.3).

Weaknesses

According to Figure 6.4, moving MAD false positive rate has to be improved since there are quite a number of true weight measurements wrongly detected as outliers. According to Figure 6.1(c), by decreasing the sensitivity of the algorithm the false

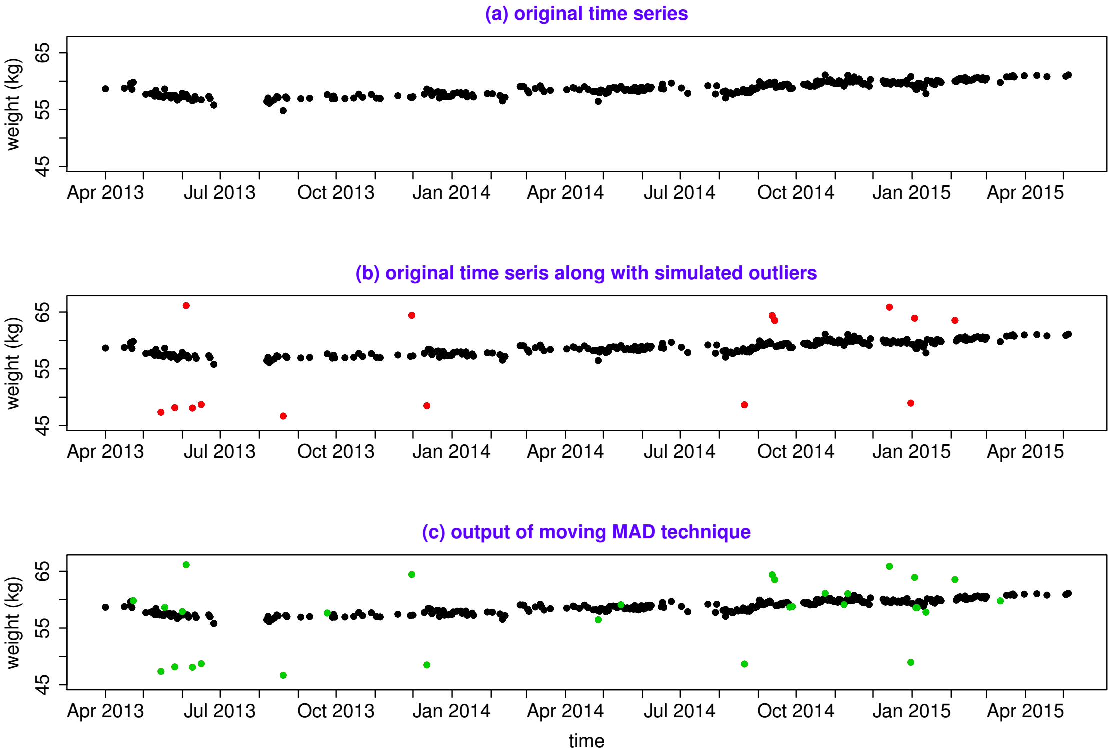


Figure 6.4 *(a) original time series before inserting simulated outliers, (b) original time series after inserting simulated outliers, represented by red dots, and (c) output of moving MAD ($\theta_m = 2$ and $k = 10$) where green dots represent the points that the algorithm detected as outliers.*

positive rate can be decreased equivalently although the power of detecting true outliers drops.

Another imperfection of the moving MAD can be noted in Figure 6.5(c) where the outliers in the neighborhood of measurement gaps were remained undetected. This was caused by occurrence of a "level shift" immediately after the measurement gap started around June 2013. This has eventually led to reduction of sensitivity of the algorithm. Except the points on the edges of measurement gaps, the rest of outliers were reasonably detected. That means, the presence of measurement gaps may become less destructive if the algorithm can be equipped with a new feature by which the temporal variations of the weight are also considered in the process of outlier detection.

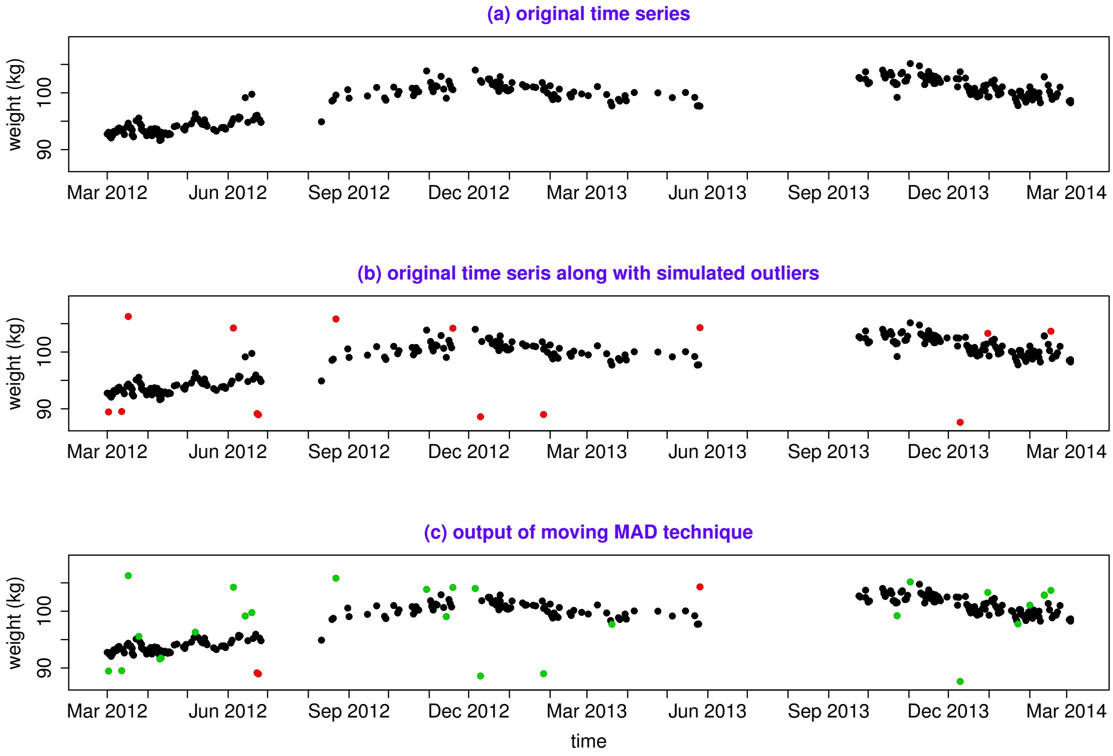


Figure 6.5 *(a) original time series before inserting simulated outliers, (b) original time series after inserting simulated outliers, represented by red dots, and (c) output of moving MAD ($\theta_m = 2$ and $k = 10$) where green dots represent the points that the algorithm detected as outliers.*

6.1.3 Results of Simulated Test Set: Conventional Rosner Statistic

Conventional Rosner statistic is the third method implemented in this study. This method involves only one controlling variable named Type-I error rate (α). It works in such a way that by increasing the Type-I error rate the algorithm has more freedom to label a point as an outlier. In other words, increasing the Type-I error rate leads to higher sensitivity whereas decreasing it causes higher specificity. Consequently, the selected Type-I error rate value was equal 0.99 that resulted in 0.728 and 0.986 average sensitivity and average specificity levels.

Strengths

Conventional Rosner statistic showed significantly high average specificity that denotes considerably low number of wrongly detected outliers. According to the ROC

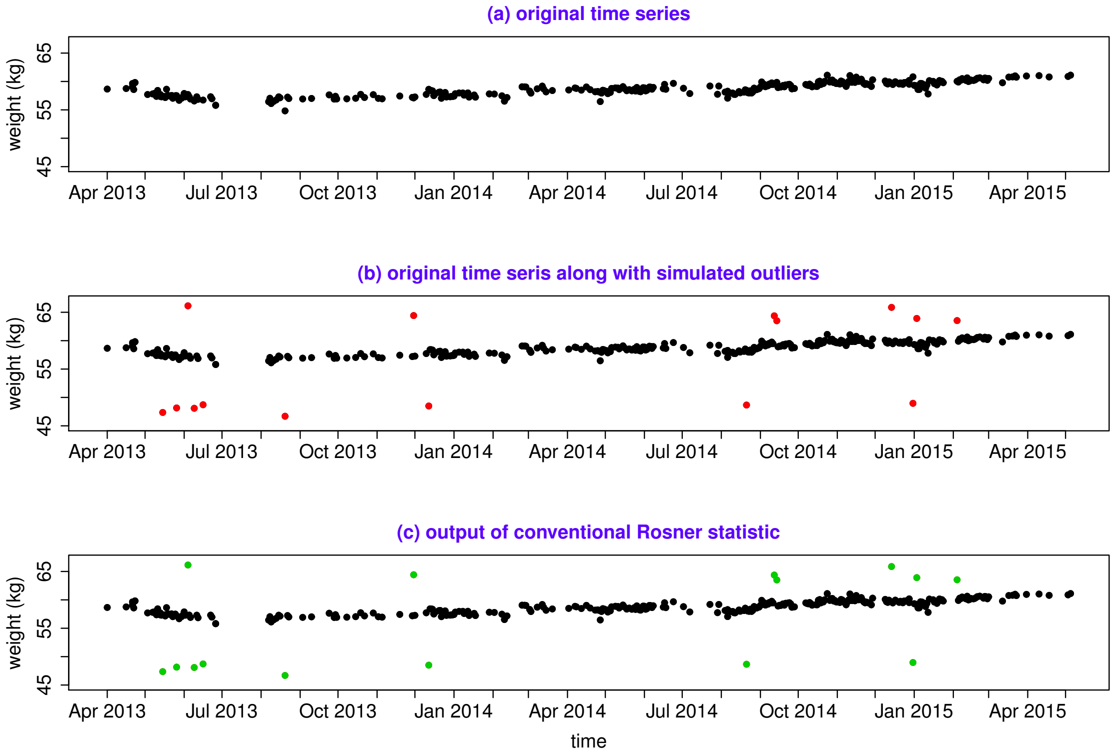


Figure 6.6 *(a) original time series before adding simulated outliers, (b) original time series along with simulated outliers, represented by red dots, and (c) output of conventional Rosner statistic* $(\alpha = 0.99)$ *where green dots are the detected outliers.*

curve shown in Figure 6.1(b), the best statistical performance of the algorithm was obtained by $\alpha = 0.99$. Based on Table 6.2, the average specificity of the conventional Rosner statistic in the best case was equal to 0.99 that is almost ideal. In Figure 6.6 the performance of the algorithm can be inspected where all of the outliers were picked out correctly without inaccurate detection of even any single point.

The computation time of the conventional Rosner statistic was another strong point of this algorithm. As shown in Table 6.3, taking only 0.003 seconds on average for each weight time series announces an extremely fast algorithm.

Weaknesses

The most important downside of the Rosner statistic is its low power in detection of outliers in case the dynamics of the weight time series increases. Having a time series that follows an upward trend, containing two long gaps of measurements like what has been shown in Figure 6.7, has led to significantly low true positive

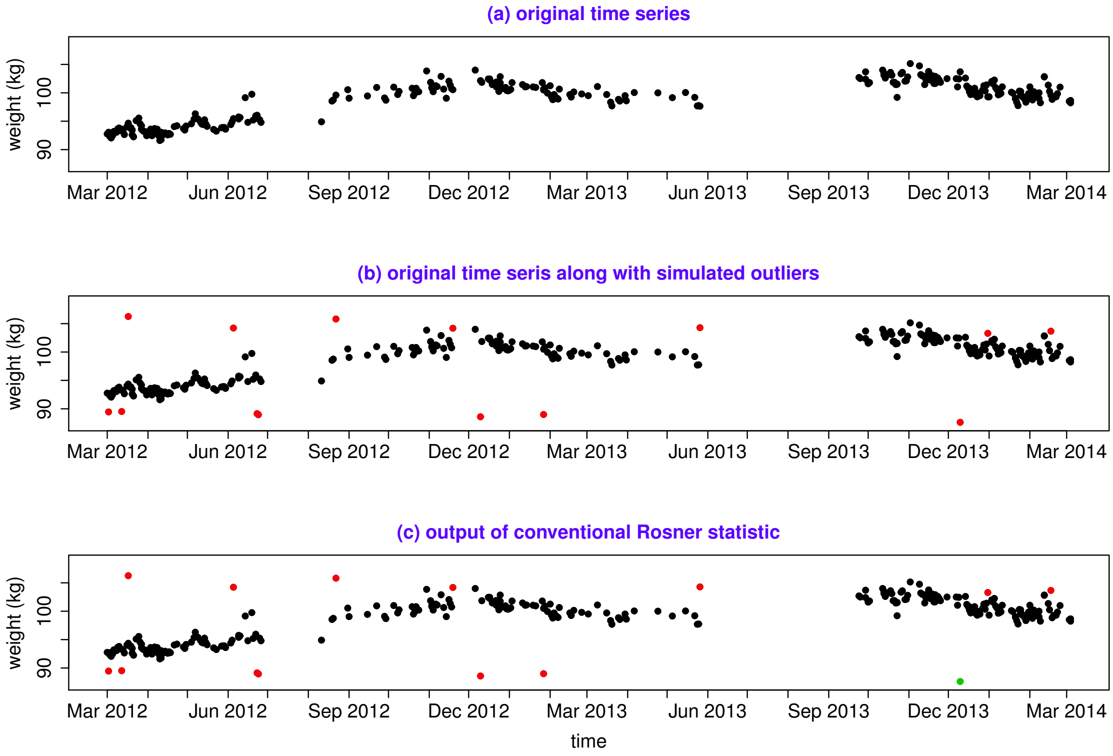


***Figure* 6.7** *(a) original time series before adding simulated outliers, (b) original time series along with simulated outliers, represented by red dots, and (c) output of conventional Rosner statistic (*$\alpha = 0.99$*) where green dots are the detected outliers.*

rate. In the mentioned figure only one of the simulated outliers was recognized correctly. This suggest that conventional Rosner statistic would not be the best choice for dynamically variable weight time series such as the case shown in Figure 6.7. The reason for such a degraded performance may originate in the violation of the Rosner statistic assumption implying on normality of the time series distribution. To investigate the issue in more detail, histograms of the clean weight time series used in Figure 6.6(a) and Figure 6.7(a) were sketched in Figure 6.8(a) and Figure 6.8(b), respectively. Although none of the histograms show normal distribution, there is a big difference between these two histograms. That is, the above histogram contains only one part despite the bottom one that comprises two separate parts. That might be the justification of Rosner statistic inability in finding outlying values for the case depicted in Figure 6.7 compared with the one in Figure 6.6.

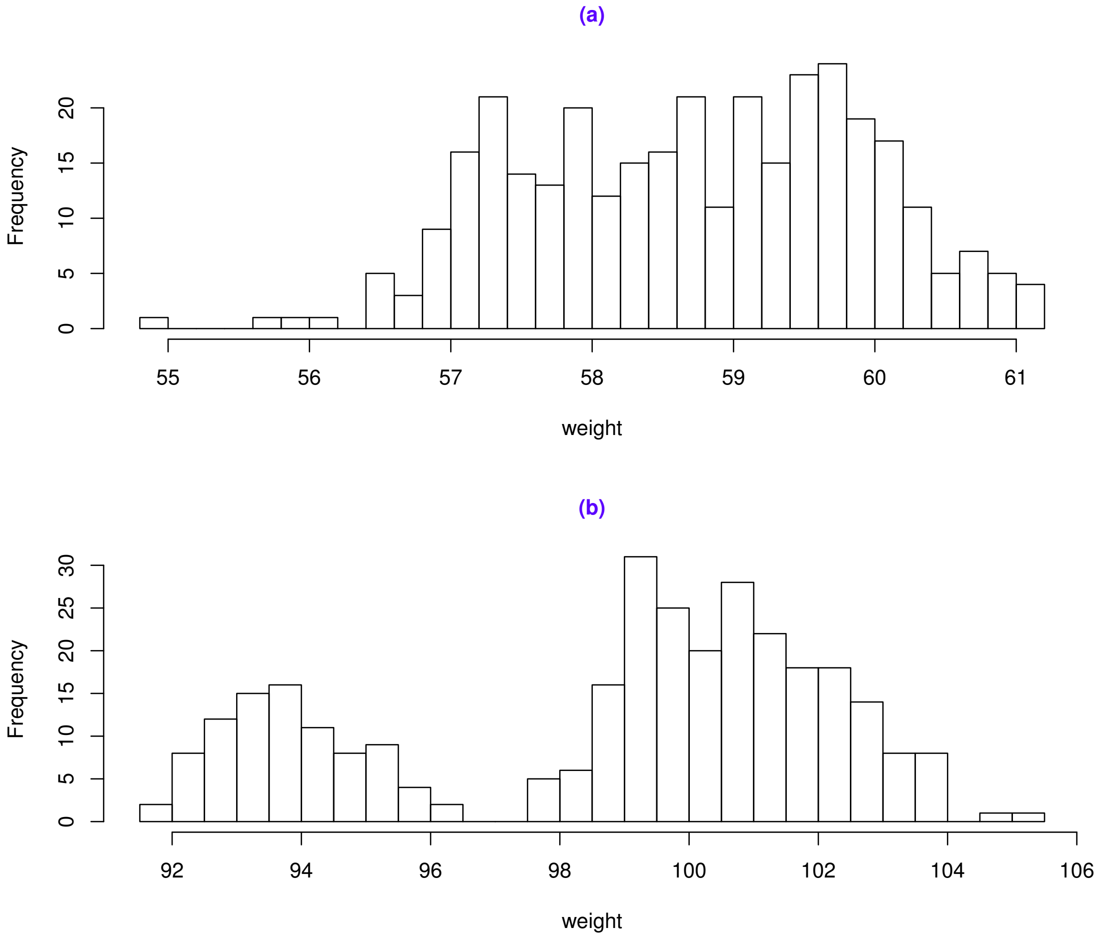


Figure 6.8 *(a) histogram of the time series shown in Figure 6.6(a), and (b) histogram of the time series shown in Figure 6.7(a).*

6.1.4 Results of Simulated Test Set: Windowed Rosner Statistic

The idea of implementing windowed Rosner statistic came into the consideration after observing the weak points of the conventional Rosner statistic. Based on the ROC curve depicted in Figure 6.1(e), the Type-I error rate (α) that gives the highest statistical performance was equal to 0.99. Besides, the algorithm performance is also affected by another controlling variable named window length (L). By keeping $\alpha =$ 0.99 and testing different window length values, the best statistical performance was obtained by $L = 60$. The average sensitivity and specificity values obtained by the aforementioned controlling parameters were equal to 0.932 and 0.987, respectively.

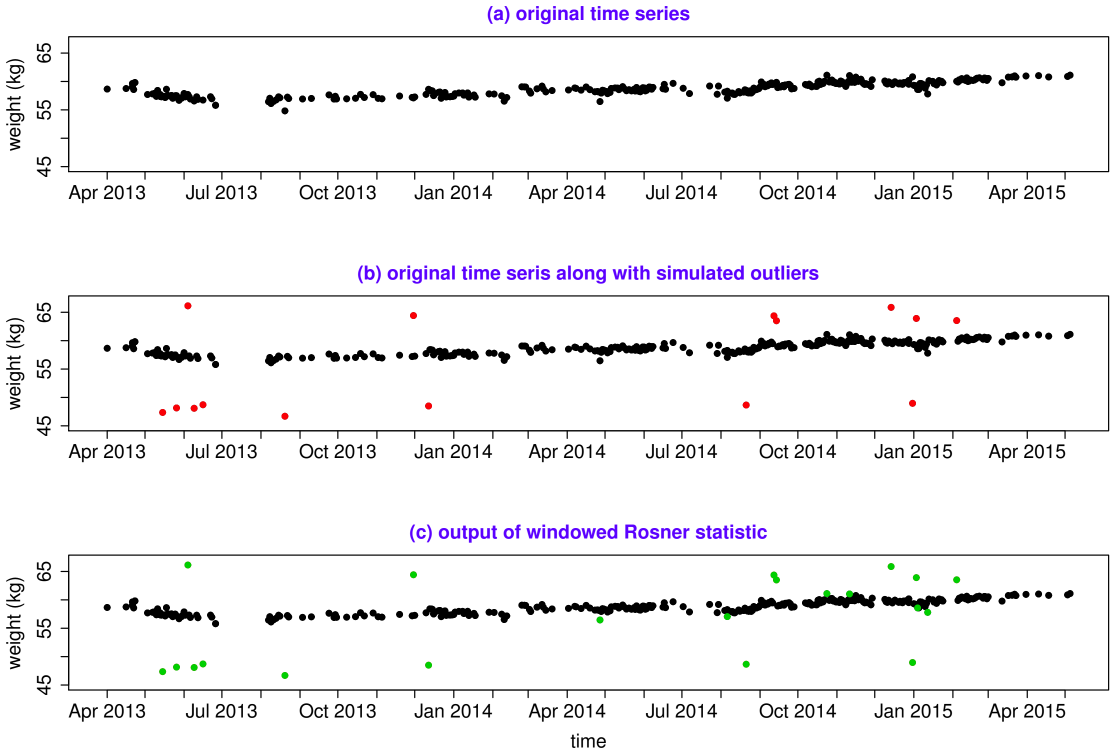


Figure 6.9 *(a) original time series before inserting simulated outliers, (b) original time series after inserting simulated outliers, represented by red dots, and (c) output of windowed Rosner statistic ($\alpha = 0.99$ and $L = 60$) where green dots represent the points that the algorithm identified as outliers.*

The ROC curve in Figure 6.1(f) shows how different values of L affected the statistical performance of windowed Rosner statistic. In this figure it is quite clear that lengthening the window has led to performance improvement until reaching the peak performance. It should be mentioned that increasing the window length to values beyond $L = 60$ has resulted in performance deterioration although they are not included in Figure 6.1(f) due to visual complexity. In following paragraphs the strength and weaknesses of the windowed Rosner statistic using $\alpha = 0.99$ and $L = 60$ are discussed.

Strengths

Considering the output of windowed Rosner statistic, in Figure 6.9(c) the true positive rate of the algorithm seems quite acceptable since there is no outlying point that remained undetected. Based on the Table 6.2 the average sensitivity of the algorithm equals 93 percent. Moreover, the average specificity of 99 percent suggests

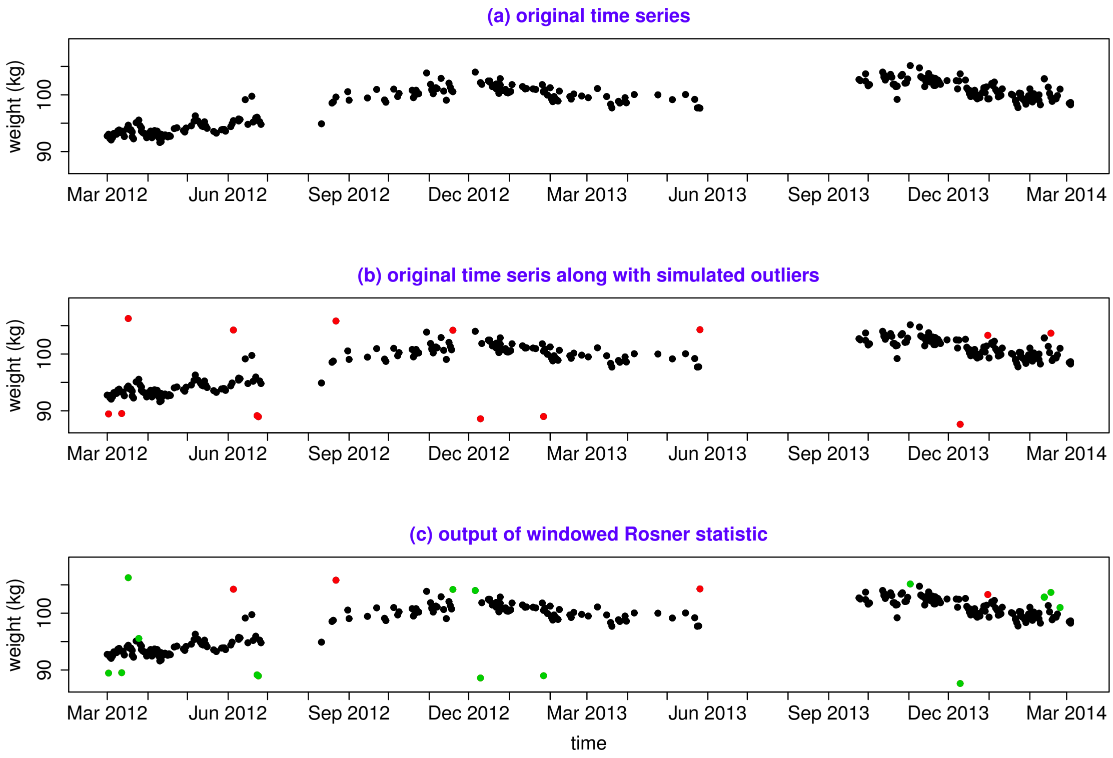


Figure 6.10 *(a) original time series before inserting simulated outliers, (b) original time series after inserting simulated outliers, represented by red dots, and (c) output of windowed Rosner statistic ($\alpha = 0.99$ and $L = 60$) where green dots represent the points that the algorithm identified as outliers.*

a moderately low rate of false positive detections although it can still be improved.

The low processing time of windowed Rosner statistic is on the other hand another strong point of this algorithm. According to Table 6.3 the average duration of processing for each time series in the simulated test set was about 0.019 seconds. Therefore, windowed Rosner statistic can also be counted as a fast outlier detection algorithm.

Weaknesses

By looking into Figure 6.10, it is comprehensible that the majority of outliers were identified by the algorithm except the points lying on the edges of measurement gaps. Presence of "level shifts" is the reason for inability of windowed Rosner statistic in detection of the remained outliers. That is, when there is an outlier right before

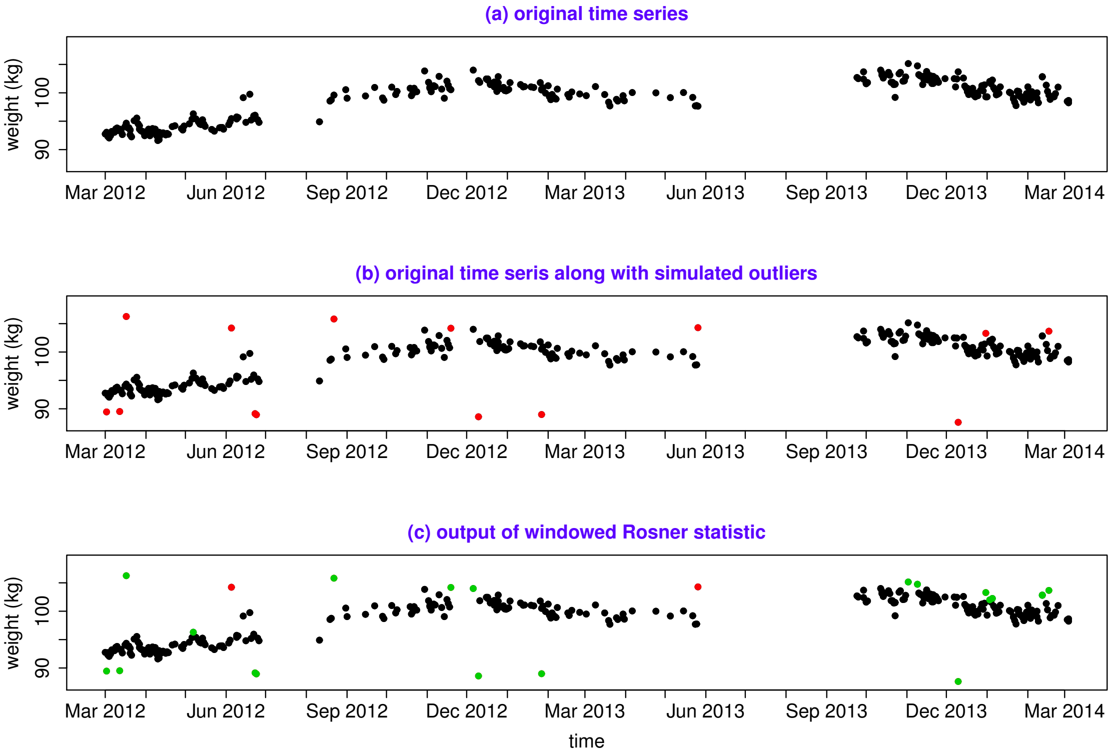


Figure 6.11 *(a) original time series before adding simulated outliers, (b) original time series along with simulated outliers, represented by red dots, and (c) output of windowed Rosner statistic ($\alpha = 0.99$ and $L = 30$) where green dots represent the points that the algorithm identified as outliers.*

the measurement gap (started on June 2013) followed by a level shift, the detection power of windowed Rosner statistic deteriorates.

One solution for the above-mentioned challenge can be shortening the window length L in case of encountering gaps of measurements followed by level shifts. In other words, specifically for the time series similar to the one depicted in Figure 6.10, shortening the window length would decrease the dominance of level-shifted points in each window. Having fewer level-shifted points in each window can lead to better performance at the end. This was proved in Figure 6.11(c) where the time series was tested using $L = 30$. It is quite clear that shorter window length showed more power in detection of outliers on the edges of measurement gaps comparing to the output shown in Figure 6.10(c) where the window length (L) is 60.

6.2 Results of Real Test Set

In this section the results of the outlier detection methods related to the real test set are depicted and discussed. The real test set included the weight time series that originally contain outliers. The suspected outlying values in each time series of this test set were annotated by visual inspection. The criteria for labeling a point as an outlier was the amount of weight difference in comparison with adjacent weight measurements considering their time difference. At the end, the rate of true and false detections were calculated to show the goodness of classification.

ROC curves of each of the examined techniques corresponding to real test set can be observed in Figure 6.12. For ARIMA technique and conventional Rosner statistic, critical value (C) and Type-I error rate (α) were the examined controlling variables as shown in Figure 6.12(a) and Figure 6.12(b), respectively. For each of the two techniques, moving MAD and windowed Rosner statistic, there are two ROC curves because of having two controlling variables. For moving MAD, threshold value (θ_m) and window length (k) and for windowed Rosner statistic, Type-I error rate (α) and window length (L) are the controlling parameters, as illustrated in Figure 6.12(c) to Figure 6.12(f). The best parameters showing the highest average statistical sensitivity and specificity of each algorithm were then identified based on ROC curves. By definition the best controlling parameters are the ones that fall closer to the point (FPR=0,TPR=1) on the upper-left corner of the ROC curves. These parameters are depicted in Table 6.4. Accordingly, Table 6.5 shows the average statistical performance of the implemented techniques in terms of sensitivity and specificity related to the chosen controlling parameters.

Average processing time of the studied techniques for each time series in real test set can be reviewed in Table 6.6. The efficiency of each technique in terms of computation time related to real weight time series can be discussed accordingly.

Two sample weight time series of real test set were intentionally selected for further investigation and explanation of the strengths and weaknesses of the implemented methods. The first sample weight time series depicts a case where a subject seemed to weigh an additional object of about 7 to 8 kilograms in a few of the time instants (Figure 6.13(a)). By careful tracking of the time intervals between the measurements it was found that those measurements cannot be the true weight values of the user. Instead, they might be representing the weight of the subject plus a back back or a pet. Indeed, these values were annotated as being suspected outliers and can

Table 6.4 *The best controlling parameters of the implemented algorithms regarding real test set.*

Variable	Value
ARIMA critical value (C)	2.5
moving MAD threshold value (θ_m)	4
moving MAD window length (k)	20
Rosner Type-I error rate (α)	0.99
windowed Rosner Type-I error rate (α)	0.1
windowed Rosner window length (L)	60

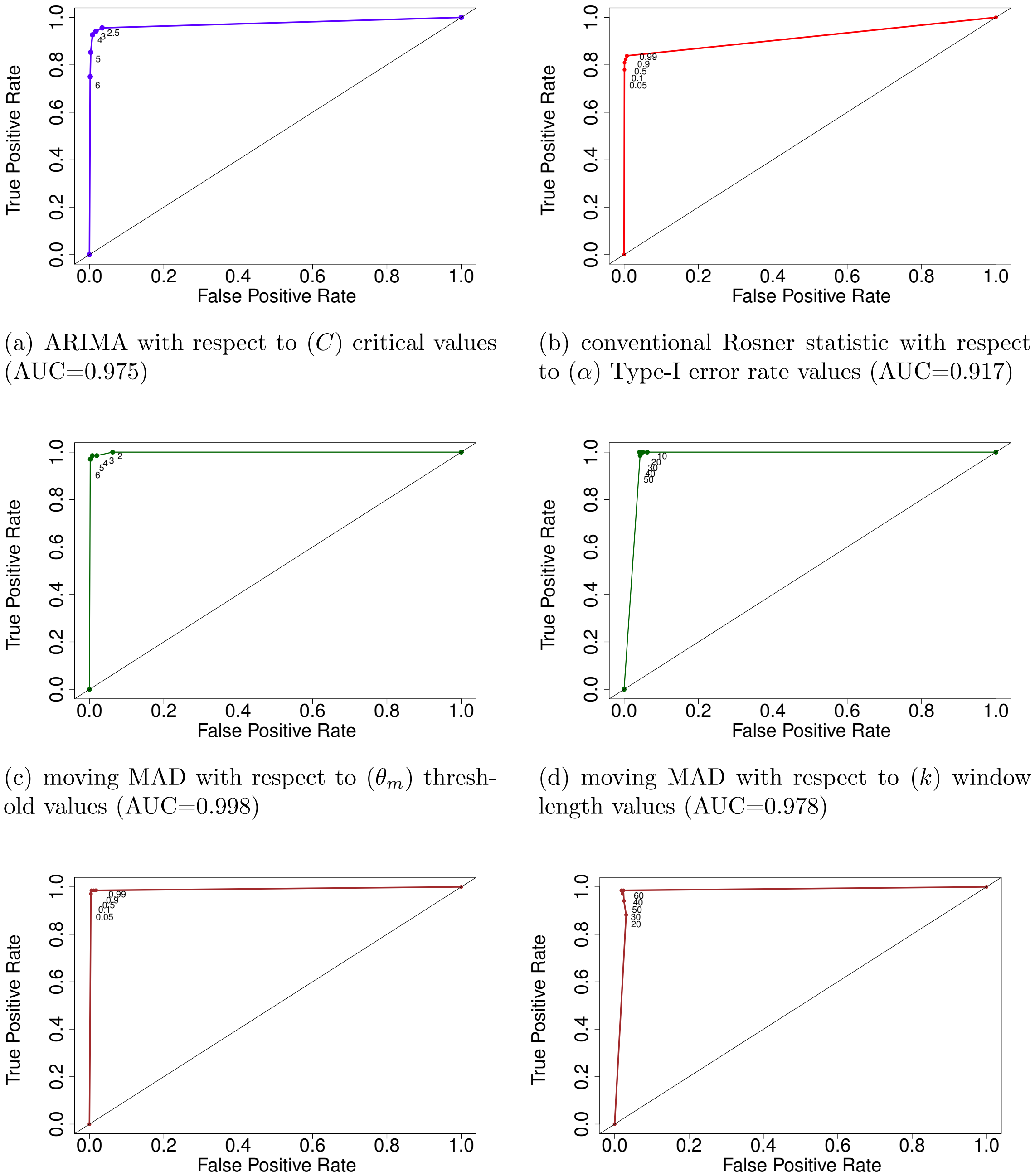
Table 6.5 *The best statistical performance of implemented outlier detection techniques in real test set.*

Methods	Average Sensitivity	Average Specificity
ARIMA	0.956	0.966
moving MAD	1.000	0.994
conventional Rosner statistic	0.838	0.992
windowed Rosner statistic	0.985	0.995

Table 6.6 *Average processing time of implemented outlier detection techniques for each time series in real test set.*

Methods	time (s)
ARIMA	$5,714.670$
moving MAD	0.029
conventional Rosner statistic	0.004
windowed Rosner statistic	0.038

be spotted by red dots in Figure 6.13(b). The second selected sample weight time series visualizes a case where there were a few weight measurements with more than 20 kilograms difference compared to the rest of measurements. Those values are counted as obvious outliers that may introduce conditions in which someone other than the main user had used the weight scale (Figure 6.14(a)). Those outliers were annotated in Figure 6.14(b) with red dots.



(a) ARIMA with respect to (C) critical values (AUC=0.975)

(b) conventional Rosner statistic with respect to (α) Type-I error rate values (AUC=0.917)

(c) moving MAD with respect to (θ_m) threshold values (AUC=0.998)

(d) moving MAD with respect to (k) window length values (AUC=0.978)

(e) windowed Rosner statistic with respect to (α) Type-I error rate values (AUC=0.991)

(f) windowed Rosner statistic with respect to (L) window length values (AUC=0.976)

Figure 6.12 *ROC curves of the real test set corresponding to (a) ARIMA approach with respect to critical value (C), (b) conventional Rosner statistic with respect to Type-I error rate (α), (c) moving MAD with respect to threshold value (θ_m), (d) moving MAD with respect to window length (k), (e) windowed Rosner statistic with respect to Type-I error rate (α), and (f) windowed Rosner statistic with respect to window length (L). For ARIMA, five critical values $C = \{6, 5, 4, 3, 2.5\}$ were tested. For moving MAD five threshold values $\theta_m = \{6, 5, 4, 3, 2\}$ as well as five window length $k = \{10, 20, 30, 40, 50\}$ were examined. conventional Rosner statistic was assessed by five Type-I error rate values $\alpha = \{0.05, 0.1, 0.5, 0.9, 0.99\}$. Windowed Rosner statistic was also tested by five Type-I error rate values $\alpha = \{0.05, 0.1, 0.5, 0.9, 0.99\}$ as well as five window length $L = \{20, 30, 40, 50, 60\}$.*

6.2.1 Results of Real Test Set: ARIMA Technique

The best classification performance of ARIMA technique was observed by setting the critical value (C) equal to 2.5 (Fig 6.12). The average values of sensitivity and specificity corresponding to the chosen critical value were 0.956 and 0.966, respectively. The following paragraphs explains the assessment of its performance over real test set.

Strengths

The cases shown in Figure 6.13 and Figure 6.14, reveal the power of ARIMA algorithm in identification of annotated outliers. Almost all of the red dots were replaced by green dots which means the true positive rate of the algorithm was quite acceptable. It should be noted that the controlling parameters chosen for ARIMA approach made it extremely sensitive in this section. The power of ARIMA algorithm stems in its ability in considering sequential aspects of the time series shown in Figure 6.13(a) that allowed this algorithm to properly find all the outlying values.

Weaknesses

One of the notable drawbacks of highly sensitive ARIMA approach is its elevated false positive rate, as depicted in Figure 6.13(c). In real test set, ARIMA technique wrongly marked 487 points as outliers among total number of 14112 points. In other words, 3.45 percent of the non-outlying values were misclassified. The rate of false positives can be reduced by lowering the sensitivity of the algorithm but that leads to reduction of true positive rate.

Another weak point of the ARIMA approach is its low power in detecting the outliers at the very beginning of the time series. This happened in three different real cases where the outlying value remained undetected. One of the mentioned cases can be observed in Figure 6.14(c) where the first point of the time series was annotated as outlier although in the output that point was not detected by the algorithm.

Computation time of the ARIMA approach imposes another limitation for large scale usage of this technique. According to Table 6.6, the amount of time spent for cleaning the time series in real test set was on average $5,714.670$ seconds. The

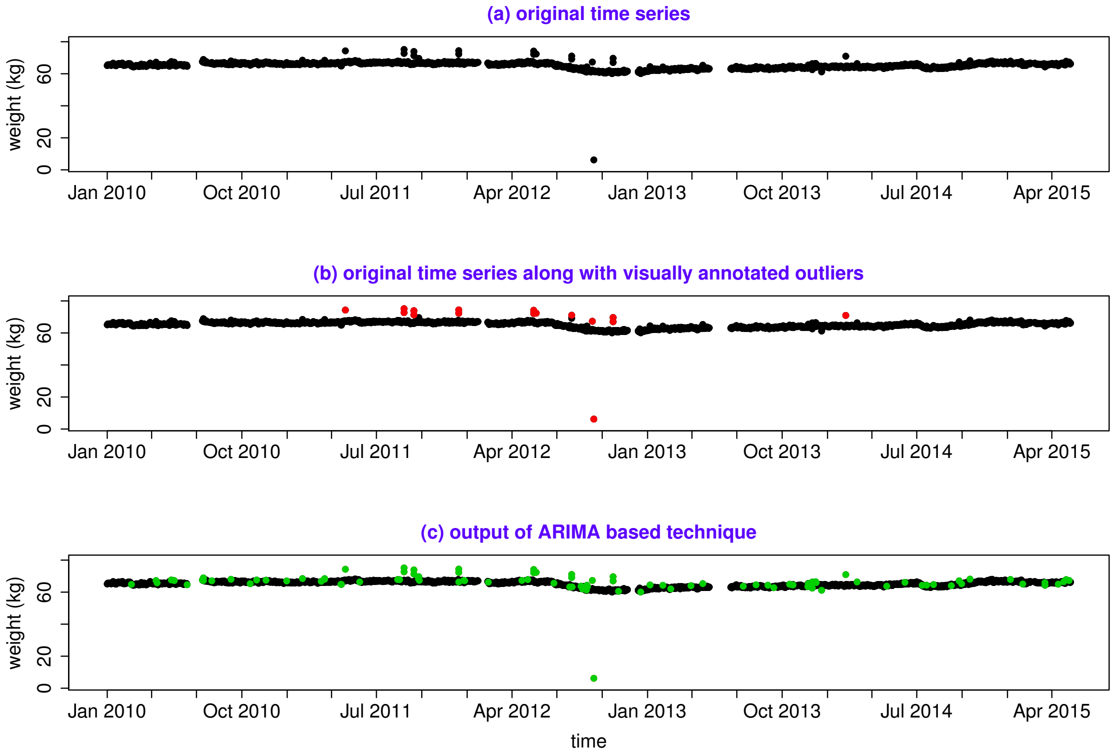


Figure 6.13 *(a) original time series, (b) original time series after visually annotating the suspected outliers, showed by red dots, and (c) output of ARIMA approach ($C = 2.5$) where green dots represent the points that the algorithm identified as outliers.*

extremely low speed of ARIMA algorithm would not allow us to apply it for bigger sets of weight time series data.

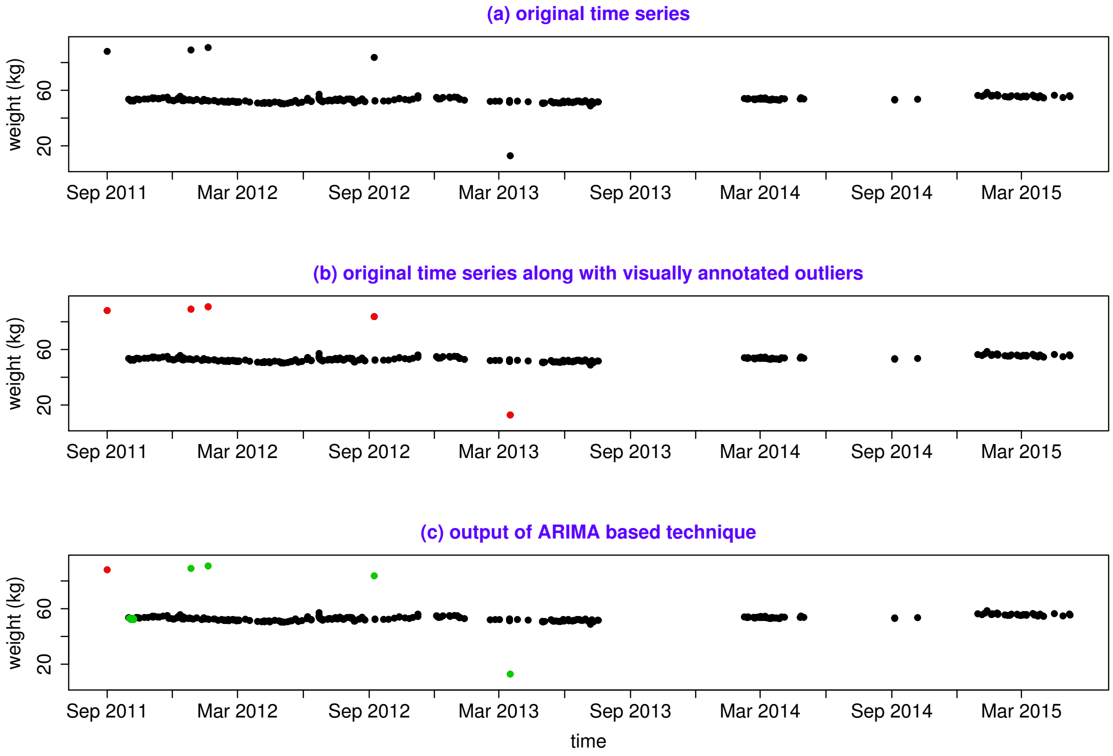


Figure 6.14 *(a) original time series, (b) original time series after visually annotating the suspected outliers, represented by red dots, and (c) output of ARIMA approach ($C = 2.5$) where green dots represent the points that the algorithm identified as outliers.*

6.2.2 Results of Real Test Set: Moving MAD

Based on Figure 6.12(c) and Figure 6.12(d), and the ROC analysis described formerly, the best threshold value (θ_m) and the best window length (k) were equal to 4 and 20. The average sensitivity and specificity values obtained by the chosen controlling variables were equivalently 1.000 and 0.994. The next paragraphs clarify the advantages and disadvantages of the moving MAD.

Strengths

Figure 6.15 and Figure 6.16 depict two cases where the annotated outliers were all detected correctly. Sufficiently sensitive controlling variables helped detecting all of the annotated outliers. In addition, Figure 6.16 clearly shows how all of the outlying values were spotted no matter if there were at the beginning or ending side of the time series. As can be seen in Table 6.5 the true positive rate and the true negative rate of the algorithm are 1 and 0.99, respectively.

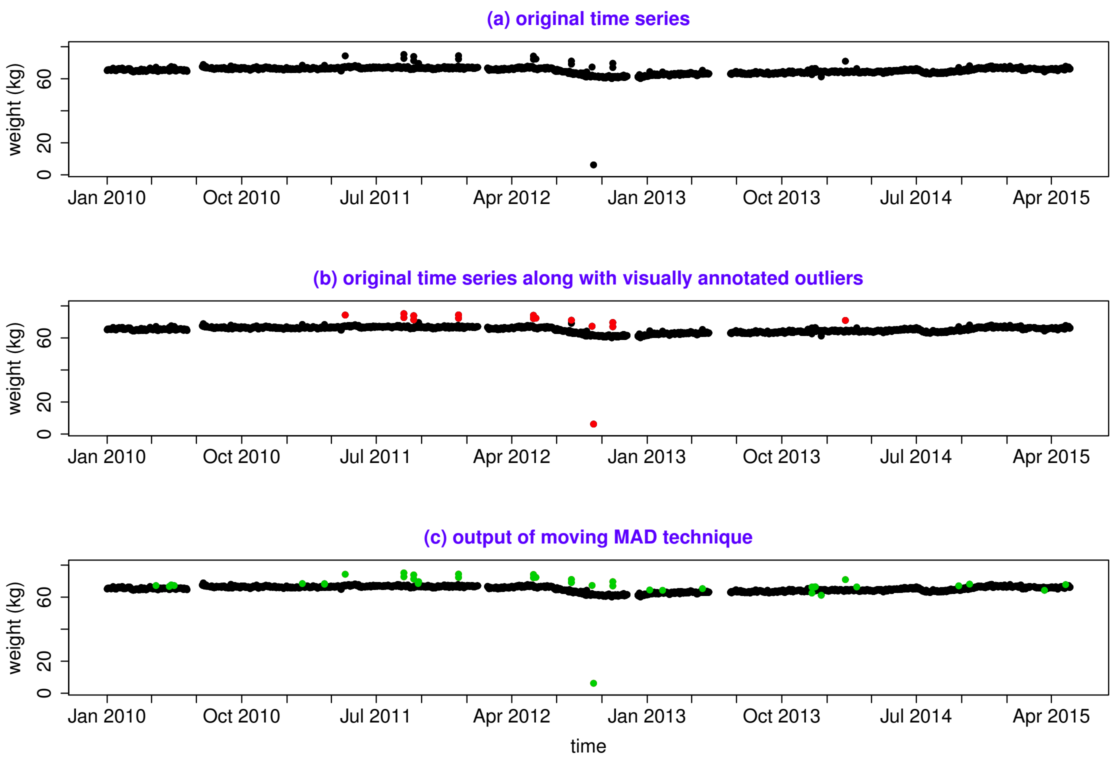


Figure 6.15 *(a) original time series, (b) original time series after visually marking the suspected outliers, showed by red dots, and (c) output of moving MAD ($\theta_m = 4$ and $k = 20$) approach where green dots are the detected outliers.*

The computation time for each of the time series in real test set was on average 0.029 seconds depicted in Table 6.6. Therefore, moving MAD can be considered as a significantly fast outlier detection algorithm. The reason for such a short computation time is the very few number of mathematical operations included in the algorithm.

Weaknesses

According to Figure 6.15(c), the false positive rate of the algorithm needs to be improved since there are a number of wrongly detected points. These wrongly detected points are within the normal variations of body mass; however, because of setting highly sensitive controlling parameters, the true negative rate of the results dropped down. This can be solve by reduction of sensitivity of the algorithm although that may cause neglecting some of the outliers.

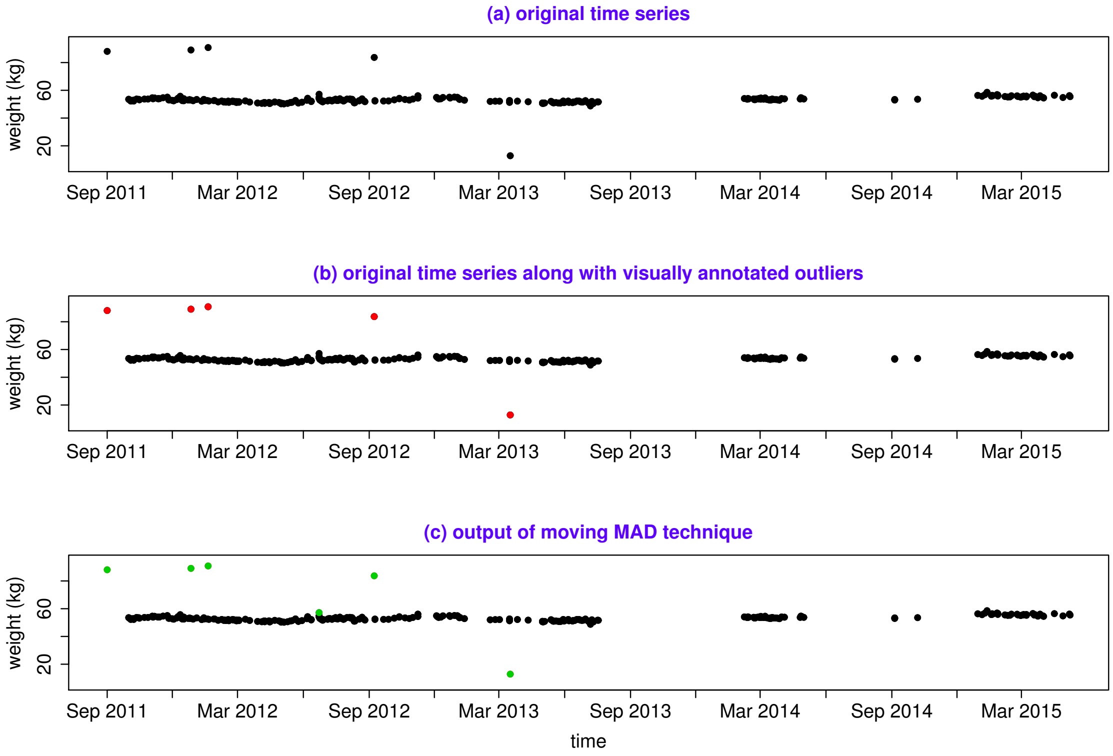


Figure 6.16 *(a) original time series, (b) original time series after visually marking the suspected outliers, showed by red dots, and (c) output of moving MAD ($\theta_m = 4$ and $k = 20$) approach where green dots are the detected outliers.*

The key solution in improving the specificity of moving MAD might be providing the capability of involving sequential aspects of the time series into outlier detection procedure. This cannot be done though, unless by adding heavy mathematical operations that may lead to computational complexity.

6.2.3 Results of Real Test Set: Conventional Rosner Statistic

Considering the ROC curve shown in Figure 6.12(b), the best Type-I error rate value (α) was equal to 0.99 that subsequently resulted in 0.838 and 0.992 as the average sensitivity and specificity values. Choosing such a high Type-I error rate value means having more freedom in selecting points as outliers with less cautiousness compared to lower Type-I error rate values. The benefits and drawbacks of the algorithm are discussed next.

Strengths

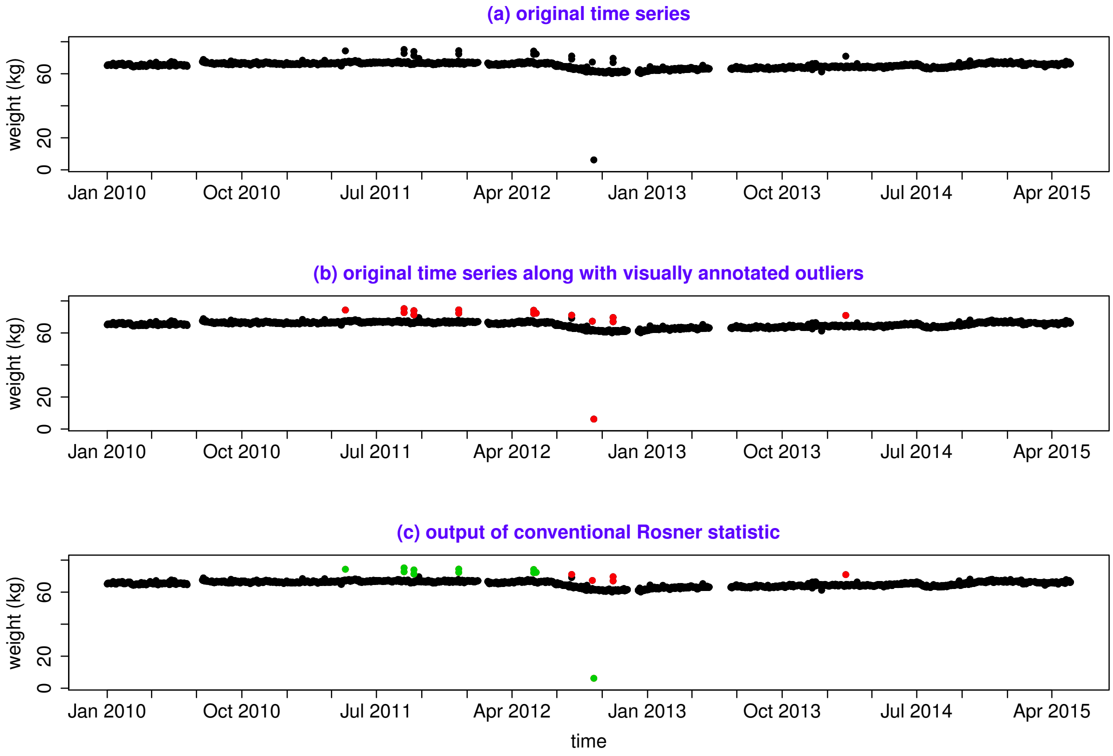


Figure 6.17 *(a) original time series, (b) original time series after annotating the suspected outliers, represented by red dots, and (c) output of conventional Rosner statistic ($\alpha = 0.99$) where green dots represent the points that the algorithm identified as outliers.*

The outputs of conventional Rosner statistic shown in Figure 6.17(c) and Figure 6.18(c) suggest considerably low rate of false positives. In other words, conventional Rosner statistic performed reasonably robust in terms of specificity because of having very few wrong detections of non-outlying values.

Another advantage of this technique would be its markedly low computation time. In more detail, for each of the weight time series in real test set, 0.004 seconds on average elapsed to get the output (Table 6.6). This notably small amount of computation time provides the conventional Rosner statistic great applicability in case of dealing with large amount of data.

Weaknesses

The lack of high enough sensitivity in detection of the whole set of outliers is the main deficiency of conventional Rosner statistic. By taking Figure 6.17(c) into consideration, it is quite obvious that a few of the outliers which are represented by

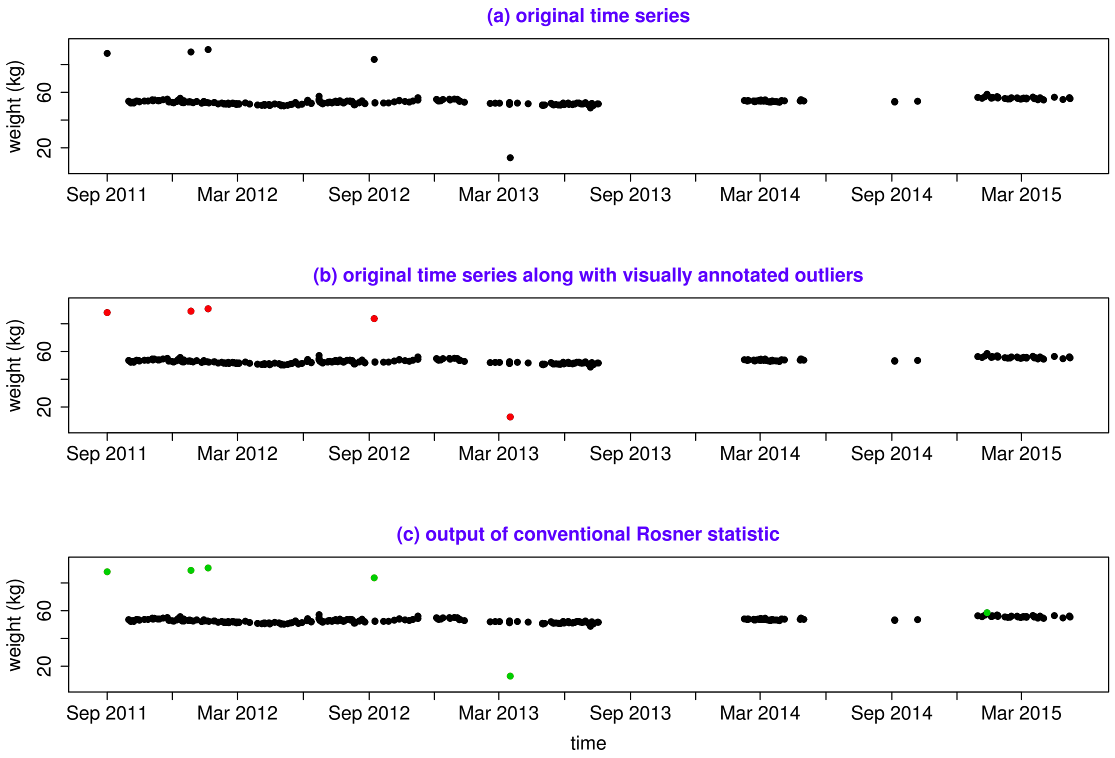


Figure 6.18 *(a) original time series, (b) original time series after annotating the suspected outliers, represented by red dots, and (c) output of conventional Rosner statistic ($\alpha = 0.99$) where green dots represent the points that the algorithm identified as outliers.*

red dots are undetected. The reason for that limitation may stem in the occurrence of a slight level shift after April 2012. The whole level of the time series as well as the outliers there slightly fell by a few kilograms. This resulted in a condition in which the outliers after that level shift lie in an almost similar level as the non-outlying values before the level shift. The non-outlying values masked the outliers after the mentioned level shift therefore they were not regarded as outliers.

Another main weakness of this algorithm originates in the principle assumptions of Rosner statistic. In detail, the maximum number of suspected outliers has to be defined for the algorithm in advance. That means, if the user underestimates the number of outliers, then some of the outliers will be obviously neglected. The other limitation of Rosner statistic arises from restriction of the number of suspected outliers. The assumptions is number of outliers cannot be more than 10 percent of the time series length. Hence, in case of facing with highly corrupted data, conventional Rosner statistic may not operate very well.

6.2.4 Results of Real Test Set: Windowed Rosner Statistic

The windowed Rosner statistic was implemented for the sake of compensating the weak points of conventional Rosner statistic. Based on the ROC analysis, the highest average sensitivity and specificity were reached using 0.1 and 60 for the two controlling parameters, Type-I error rate (α) and window length (L). Correspondingly, the peak average sensitivity and specificity values were equal to 0.985 and 0.995, respectively. The two cases in Figure 6.19 and Figure 6.20 demonstrate the algorithm performance in detection of annotated outliers. The strong and weak points of the windowed Rosner statistic are presented in the succeeding paragraphs.

Strengths

As can be seen in Figure 6.19 and Figure 6.20, all of the suspected outliers were detected correctly. Evidently, windowed Rosner statistic is powerful enough to detect the points lying either before or after the minute level shift which occurred after April 2012 in Figure 6.19(c). Moreover, considering the case in Figure 6.20, windowed Rosner statistic idealistically identified all of the outlying values. The specificity of the algorithm is also exemplary since there is only one false detection (Figure 6.20(c)).

Another advantage of windowed Rosner statistic is its low computation time. For each of the time series in real test set 0.038 seconds was spent on average to get the outputs (Table 6.6). Windowed Rosner statistic can also be classified as computationally fast outlier detection technique.

Weaknesses

The number of wrong detections or the false positive rate of the windowed Rosner statistic should be improved for the cases where level shifts, missing data, and fluctuating weight dynamics exist. For instance, the FPR rate of algorithm regarding the weight time series depicted in Figure 6.20(c) clearly suffered from presence of level shift that occurred near April 2012. As can be seen there are quite a number of normal data points which were wrongly detected as outliers. Such a number of wrong detections may impose limitations in applicability of this algorithm for outlier detection in time series of weight measurements.

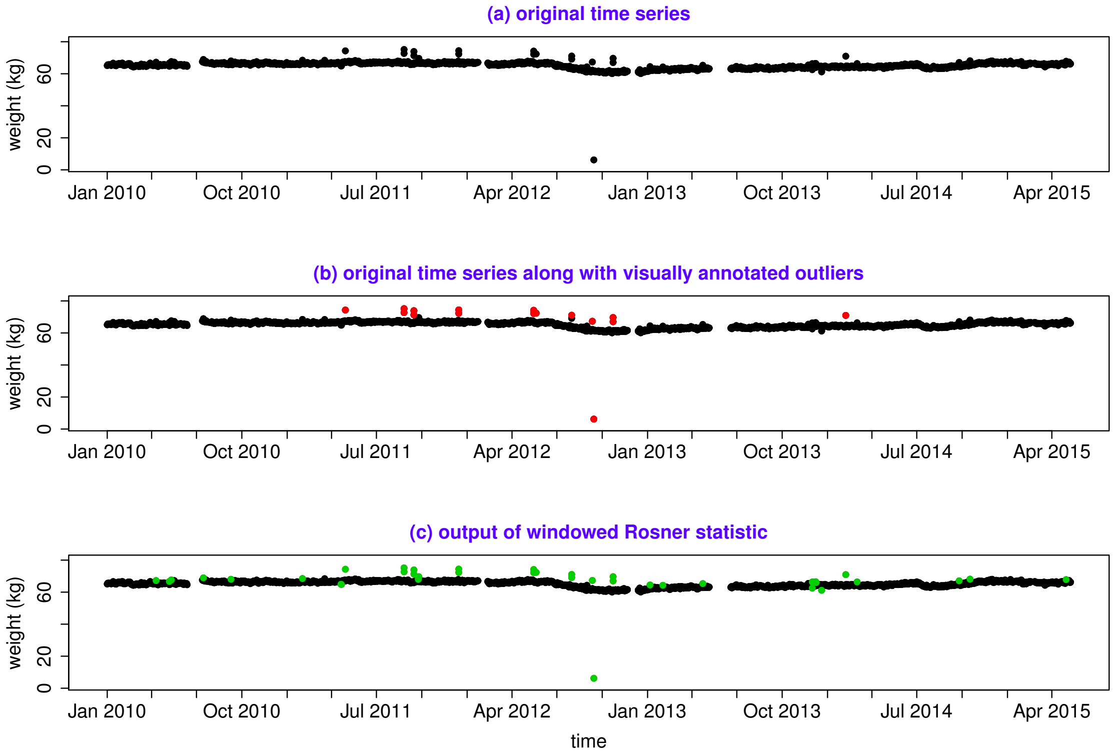


Figure 6.19 *(a) original time series, (b) original time series after marking the suspected outliers, showed by red dots, and (c) output of windowed Rosner statistic ($\alpha = 0.1$ and $L = 60$) where green dots are the detected outliers.*

A potential trick for reducing the number of false detections can be taking the advantage of the initial guess about the number of suspected outliers in each window. In other words, since the algorithm is repeating the conventional Rosner statistic inside each sliding window, by adaptively reducing the number of suspected outliers in each window (as explained in 4.3) the number of false detections can be restricted. For example, considering the weight time series shown in Figure 6.19(b), there is not any suspected outlier during the year 2010 although the algorithm made a few false detections. If the number of suspected outliers in the windows sliding within the year 2010 can be set to either zero or near zero, the algorithm makes less false detections and accordingly the FPR rate is controlled in the corresponding sliding windows.

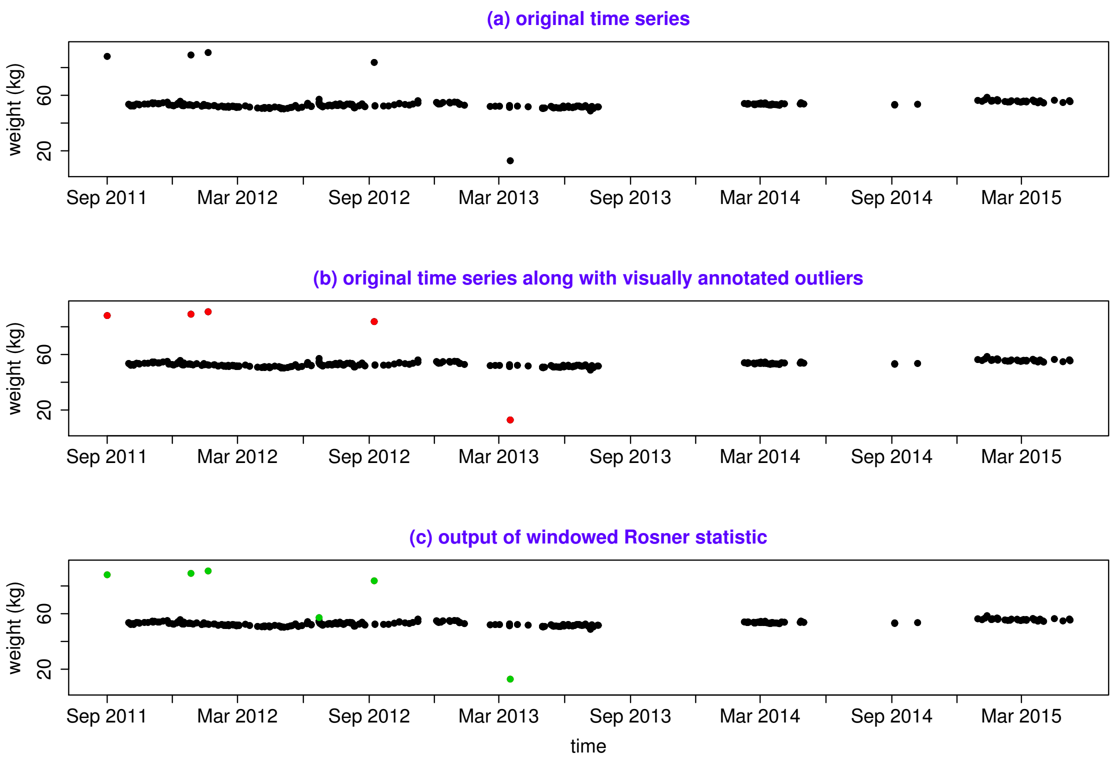


Figure 6.20 *(a) original time series, (b) original time series after marking the suspected outliers, showed by red dots, and (c) output of windowed Rosner statistic ($\alpha = 0.1$ and $L = 60$) where green dots are the detected outliers.*

6.3 Performance Comparison Summary

In this section a comparison of the studied techniques is presented considering all the discussed strengths and weaknesses found in simulated and real test sets.

As far as sensitivity of the outlier detection algorithms are concerned, moving MAD and windowed Rosner statistic showed quite high performances over real test set. The ARIMA technique performed slightly less powerful because of being unable to identify the outlying points at the beginning of a few real time series. In the simulated test set again the most sensitive method was moving MAD. The second best method was ARIMA technique, while windowed Rosner statistic was the third. Conventional Rosner statistic could not be considered as a highly sensitive algorithm in comparison with the rest of the studied techniques. The violation of primary assumptions of conventional Rosner statistic significantly deteriorated its performance.

Regarding the specificity, i.e. the ability of making as few wrong detections as possible, the three algorithms, moving MAD, conventional Rosner statistic, and windowed Rosner statistic showed significantly strong performances over the real test set. However, the ARIMA technique performed scarcely less well. Regarding the simulated test set, conventional and windowed Rosner statistic algorithms operated the best. Afterward, ARIMA approach and moving MAD were placed.

In the case of outliers lying on the edges of measurement gaps, ARIMA algorithm was the only method able to detect every single outlier. This suggests the superiority of ARIMA technique in comparison with the rest of studied methods.

Implementation of windowed Rosner statistic truly enhanced the weaknesses of conventional Rosner statistic. Improvement in sensitivity without losing much of specificity was optimally achieved.

From computational complexity viewpoint, the ARIMA approach ranked as the heaviest algorithm among the studied techniques. In contrast, the rest of the techniques performed extremely fast with average processing times less than 1 second for each time series in both simulated and real test sets.

None of the tested algorithms showed robustness to parameter selection as there were quite varying performance results corresponding to different controlling parameters. Although it was not considered in this thesis, the best way of evaluating selected controlling parameters is done via cross validation. In other words, after finding the best set of controlling parameters via ROC analysis, there should have been a testing phase in which the algorithms were tested over a set of unseen data.

7. CONCLUSIONS AND FUTURE WORK

This thesis was based on a data set that includes weight time series of 10,000 randomly selected anonymous weight scale users from all over the world. The necessity of self-weighing, vraiability of body mass, and properties of weight time series were explained in Chapter 2. A thorough explanation of the outlier detection in univariate time series was presented in Chapter 3. Subsequently, four point-wise outlier detection techniques namely ARIMA technique, moving MAD, conventional Rosner statistic, and windowed Rosner statistics were described in Chapter 4. Furthermore, a few of the behavioral patterns of the users in terms of self-weighing were investigated and reported in Chapter 5.

Regarding the behavioral patterns of the studied self-weighers, after exploring the measurement times during the day, it was comprehended that subjects tended to monitor their weight mostly in the morning between 5 to 10 AM. Interestingly, the most repeated time interval of self-weighing was between 12 to 36 hours that means a large number of consecutive recordings had a time difference between 12 to 36 hours. Moreover, the studied population tended to weigh less during the weekend days compared to the weekdays. The same phenomenon happened for monthly self-weighing frequency in November and December every year. In other words, people measured their weight less during the two mentioned months compared to the rest of the months in each year.

The thorough evaluation of the outlier detection techniques unveiled that in general moving MAD operated better comparing to other studied methods. The highest average sensitivity along with the second highest average specificity in real test set reveal the power of moving MAD in handling the outlier detection. Reaching such a high diagnostic performance probably originates in the robustness of outlier detection utilizing median deviations as compared to arithmetic mean deviations. Windowed Rosner statistic performed slightly less powerful than moving MAD, and ARIMA method slightly less powerful than windowed Rosner statistic. The conventional Rosner statistic did not seem to be contextually appropriate for outlier detection in weight time series data.

Identification of the outlier detection ground truth can be one of the biggest challenges in biomedical time series analysis, particularly the time series recorded by self-monitoring devices. The lack of reliable ground truth and the potential misclassification of the non-outlying values can sometimes become costly. For instance, in case of treating heart failure patients, removing even a single true weight measurement that indicates the hypervolemia (or body fluid overload) would lead to dangerous outcomes. Therefore, specific enough algorithms are necessary in clinical use.

In addition, defining a ground truth based on visual inspection may sometimes lead to faulty decisions. One solution for this issue can be recognition and utilization of highly accurate statistical models that represents the normal range of weight variation as a function of time using large-scale data sets of weight time series. The data set used in this study can be an appropriate choice for extraction of such models. In future the applicability of these data-driven models will be further investigated.

Outlier detection methods assessed in this thesis were not working on a real-time basis. In other words, weight measurements were gathered first before being fed to the outlier detection algorithms. Outlier detection can also be done on a real-time basis where every weight measurement that is recorded by digital weight scale is immediately classified and labeled on the fly (either as an outlier or a normal data point). This case has not been considered in this thesis and is postponed to future studies.

To sum up, the main objective of this thesis was addressing the problem of outlier detection in time series of weight measurements. As an overall conclusion, based to the acquired results, none of the presented methods was able to ideally solve the outlier detection problem. That is mainly because of lacking appropriate physiologically representative models of normal and abnormal weight variations. Hence, more realistic results can be obtained by incorporating models of temporal dynamics of weight variation into the outlier detection process.

As an overall self evaluation of the thesis work, the primary objective of the thesis was ultimately met, i.e. comparison of univariate statistical based outlier detection techniques in time series of weight measurements. The methodology can be improved though, specifically the final validation of optimized parameters. In other words, the parameter optimization should typically be done over training data, while the performance evaluation of the chosen parameters over test data. Lack of this approach in the final performance evaluation might have slightly biased the final conclusions. The greatest challenge of this thesis work was addressing the outlier detection in time series of weight measurements for the first time in history of biomedical time series analysis. At last, the main strengths of this thesis was comparison of studied techniques using both simulated outliers and real outliers. Evaluating simulated outliers helped understanding the weak and strong points of each method in a thorougher manner.

REFERENCES

[1] W. H. Organization, "Data and statistics," 2016.

[2] J. Levi, L. M. Segal, J. Rayburn, and A. Martin, *State of Obesity: Better Policies for a Healthier America: 2015*. Trust for America's Health, 2015.

[3] L. E. Burke, J. Wang, and M. A. Sevick, "Self-monitoring in weight loss: a systematic review of the literature," *Journal of the American Dietetic Association*, vol. 111, no. 1, pp. 92–102, 2011.

[4] 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.

[5] R. R. Wing, D. F. Tate, A. A. Gorin, H. A. Raynor, and J. L. Fava, "A self-regulation program for maintenance of weight loss," *New England Journal of Medicine*, vol. 355, no. 15, pp. 1563–1571, 2006.

[6] C. R. Pacanowski, F. Bertz, and D. A. Levitsky, "Daily self-weighing to control body weight in adults," *SAGE Open*, vol. 4, no. 4, p. 2158244014556992, 2014.

[7] M. L. Butryn, S. Phelan, J. O. Hill, and R. R. Wing, "Consistent self-monitoring of weight: a key component of successful weight loss maintenance," *Obesity*, vol. 15, no. 12, pp. 3091–3096, 2007.

[8] A.-L. Orsama, E. Mattila, M. Ermes, M. Van Gils, B. Wansink, and I. Korhonen, "Weight rhythms: weight increases during weekends and decreases during weekdays," *Obesity facts*, vol. 7, no. 1, pp. 36–47, 2014.

[9] S. B. Racette, E. P. Weiss, K. B. Schechtman, K. Steger-May, D. T. Villareal, K. A. Obert, and J. O. Holloszy, "Influence of weekend lifestyle patterns on body weight," *Obesity*, vol. 16, no. 8, pp. 1826–1830, 2008.

[10] M. Tuomisto, T. Terho, I. Korhonen, R. Lappalainen, T. Tuomisto, P. Laippala, and V. Turjanmaa, "Diurnal and weekly rhythms of health-related variables in home recordings for two months," *Physiology & behavior*, vol. 87, no. 4, pp. 650–658, 2006.

[11] W. A. Van Staveren, P. Deurenberg, J. Burema, L. C. De Groot, and J. Hautvast, "Seasonal variation in food intake, pattern of physical activity and change in body weight in a group of young adult dutch women consuming self-selected diets.," *International journal of obesity*, vol. 10, no. 2, pp. 133–145, 1985.

[12] Y. Ma, B. C. Olendzki, W. Li, A. R. Hafner, D. Chiriboga, J. R. Hebert, M. Campbell, M. Sarnie, and I. S. Ockene, "Seasonal variation in food intake, physical activity, and body weight in a predominantly overweight population," *European journal of clinical nutrition*, vol. 60, no. 4, pp. 519–528, 2006.

[13] M. Kobayashi and M. Kobayashi, "The relationship between obesity and seasonal variation in body weight among elementary school children in tokyo," *Economics & Human Biology*, vol. 4, no. 2, pp. 253–261, 2006.

[14] S. Mehrang, E. Helander, M. Pavel, A. Chieh, and I. Korhonen, "Outlier detection in weight time series of connected scales," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 1489–1496, IEEE, 2015.

[15] 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), 2015 37th Annual International Conference of the IEEE*, pp. 1616–1620, IEEE, 2015.

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

[17] J. C. Seidell, T. Visscher, and R. T. Hoogeveen, "Overweight and obesity in the mortality rate data: current evidence and research issues.," *Medicine and science in sports and exercise*, vol. 31, no. 11 Suppl, pp. S597–601, 1999.

[18] 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.

[19] S. A. Schroeder, "We can do better improving the health of the american people," *New England Journal of Medicine*, vol. 357, no. 12, pp. 1221–1228, 2007.

[20] 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.

[21] R. C. Baker and D. S. Kirschenbaum, "Self-monitoring may be necessary for successful weight control," *Behavior Therapy*, vol. 24, no. 3, pp. 377–394, 1993.

[22] C. S. Carver and M. F. Scheier, *Attention and self-regulation: A control-theory approach to human behavior*. Springer Science & Business Media, 2012.

[23] M. T. McGuire, R. R. Wing, M. L. Klem, and J. O. Hillf, "Behavioral strategies of individuals who have maintained long-term weight losses," *Obesity research*, vol. 7, no. 4, pp. 334–341, 1999.

[24] M. T. McGuire, R. R. Wing, M. L. Klem, W. Lang, and J. O. Hill, "What predicts weight regain in a group of successful weight losers?," *Journal of consulting and clinical psychology*, vol. 67, no. 2, p. 177, 1999.

[25] D. Levitsky, J. Garay, M. Nausbaum, L. Neighbors, and D. Dellavalle, "Monitoring weight daily blocks the freshman weight gain: a model for combating the epidemic of obesity," *International journal of obesity*, vol. 30, no. 6, pp. 1003–1010, 2006.

[26] J. J. VanWormer, J. A. Linde, L. J. Harnack, S. D. Stovitz, and R. W. Jeffery, "Self-weighing frequency is associated with weight gain prevention over 2 years among working adults," *International journal of behavioral medicine*, vol. 19, no. 3, pp. 351–358, 2012.

[27] Z. Cooper and C. G. Fairburn, "A new cognitive behavioural approach to the treatment of obesity," *Behaviour research and therapy*, vol. 39, no. 5, pp. 499–511, 2001.

[28] Z. Cooper, C. G. Fairburn, T. Wadden, A. Stunkard, *et al.*, "Cognitive-behavioral treatment of obesity.," *Handbook of obesity treatment*, pp. 465–479, 2002.

[29] N. M. Albert, "Fluid management strategies in heart failure," *Critical care nurse*, vol. 32, no. 2, pp. 20–32, 2012.

[30] H. Kataoka, "A new monitoring method for the estimation of body fluid status by digital weight scale incorporating bioelectrical impedance analyzer in definite

heart failure patients," *Journal of cardiac failure*, vol. 15, no. 5, pp. 410–418, 2009.

[31] N. M. Metheny, *Fluid and electrolyte balance.* Jones & Bartlett Publishers, 2011.

[32] N. J. Rehrer, "Fluid and electrolyte balance in ultra-endurance sport," *Sports Medicine*, vol. 31, no. 10, pp. 701–715, 2001.

[33] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice.* OTexts, 2014.

[34] J. W. Kemnitz, J. R. Gibber, K. A. Lindsay, and S. G. Eisele, "Effects of ovarian hormones on eating behaviors, body weight, and glucoregulation in rhesus monkeys," *Hormones and Behavior*, vol. 23, no. 2, pp. 235–250, 1989.

[35] D. R. Shahar, P. Froom, G. Harari, N. Yerushalmi, F. Lubin, and E. Kristal-Boneh, "Changes in dietary intake account for seasonal changes in cardiovascular disease risk factors," *European journal of clinical nutrition*, vol. 53, no. 5, pp. 395–400, 1999.

[36] C. E. Matthews, P. S. Freedson, J. R. Hebert, E. J. Stanek, P. A. Merriam, M. C. Rosal, C. B. Ebbeling, and I. S. Ockene, "Seasonal variation in household, occupational, and leisure time physical activity: longitudinal analyses from the seasonal variation of blood cholesterol study," *American journal of epidemiology*, vol. 153, no. 2, pp. 172–183, 2001.

[37] D. J. Casa, L. E. Armstrong, S. K. Hillman, S. J. Montain, R. V. Reiff, B. S. Rich, W. O. Roberts, and J. A. Stone, "National athletic trainers' association position statement: fluid replacement for athletes," *Journal of athletic training*, vol. 35, no. 2, p. 212, 2000.

[38] B. M. Popkin, K. E. D'Anci, and I. H. Rosenberg, "Water, hydration, and health," *Nutrition reviews*, vol. 68, no. 8, pp. 439–458, 2010.

[39] H. S. Teng, K. Chen, and S. C. Lu, "Adaptive real-time anomaly detection using inductively generated sequential patterns," in *Research in Security and Privacy, 1990. Proceedings., 1990 IEEE Computer Society Symposium on*, pp. 278–284, IEEE, 1990.

[40] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[41] D. Cheboli, *Anomaly detection of time series.* PhD thesis, University of Minnesota, 2010.

[42] P.-N. Tan, M. Steinbach, V. Kumar, *et al.*, *Introduction to data mining*, vol. 1. Pearson Addison Wesley Boston, 2006.

[43] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification. 2nd,” *Edition. New York*, 2001.

[44] A. K. Jain and R. C. Dubes, *Algorithms for clustering data.* Prentice-Hall, Inc., 1988.

[45] S. Basu, M. Bilenko, and R. J. Mooney, “A probabilistic framework for semi-supervised clustering,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59–68, ACM, 2004.

[46] V. Barnett, “The ordering of multivariate data,” *Journal of the Royal Statistical Society. Series A (General)*, pp. 318–355, 1976.

[47] V. Bamnett and T. Lewis, “Outliers in statistical data,” 1994.

[48] R. J. Beckman and R. D. Cook, “Outlier..........s,” *Technometrics*, vol. 25, no. 2, pp. 119–149, 1983.

[49] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

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

[51] H. E. Solberg and A. Lahti, “Detection of outliers in reference distributions: performance of hornâs algorithm,” *Clinical chemistry*, vol. 51, no. 12, pp. 2326–2332, 2005.

[52] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications.* Springer Science & Business Media, 2013.

[53] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, “Towards parameter-free data mining,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 206–215, ACM, 2004.

[54] A. Arning, R. Agrawal, and P. Raghavan, "A linear method for deviation detection in large databases.," in *KDD*, pp. 164–169, 1996.

[55] F. J. Anscombe, "Rejection of outliers," *Technometrics*, vol. 2, no. 2, pp. 123–146, 1960.

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

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

[58] 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.

[59] J. López-de Lacalle, "tsoutliers r package for detection of outliers in time series,"

[60] L. Davies and U. Gather, "The identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782–792, 1993.

[61] R. K. Pearson, "Exploring process data," *Journal of Process Control*, vol. 11, no. 2, pp. 179–194, 2001.

[62] R. K. Pearson, "Outliers in process modeling and identification," *Control Systems Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 55–63, 2002.

[63] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*, pp. 131–146, Springer, 2005.

[64] E. Acuna and C. Rodriguez, "A meta analysis study of outlier detection methods in classification," *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, 2004.

[65] C. P. Quesenberry and H. A. David, "Some tests for outliers," *Biometrika*, vol. 48, no. 3/4, pp. 379–390, 1961.

[66] Withings, "Withings services terms and conditions," 2015.

[67] J. L. de Lacalle, *tsoutliers: Detection of Outliers in Time Series*, 2015. R package version 0.6.

[68] R. J. Hyndman, *forecast: Forecasting functions for time series and linear models*, 2016. R package version 7.1.

[69] R. J. Hyndman and Y. Khandakar, “Automatic time series forecasting: the forecast package for R,” *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008.

[70] S. P. Millard, *EnvStats: An R Package for Environmental Statistics*. New York: Springer, 2013.

[71] C. E. Metz, “Basic principles of roc analysis,” in *Seminars in nuclear medicine*, vol. 8, pp. 283–298, Elsevier, 1978.

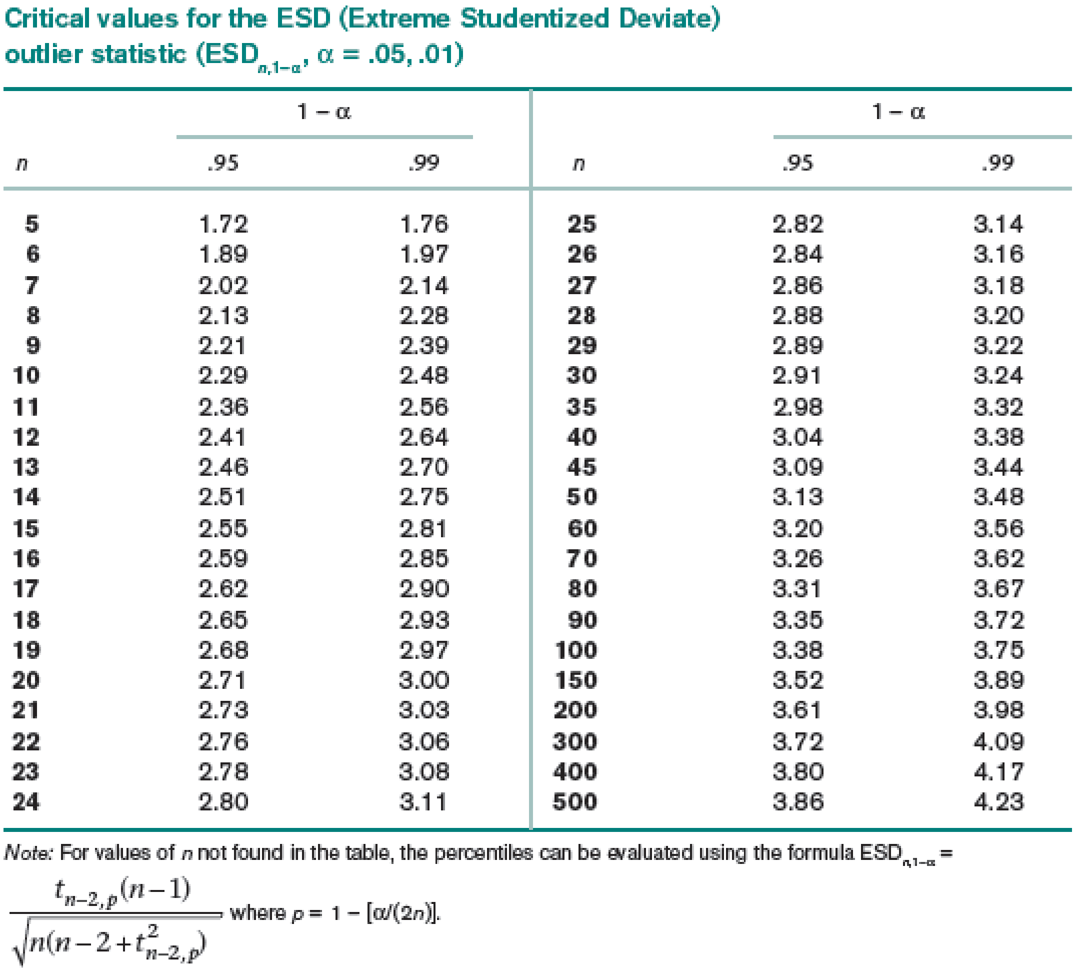
[72] M. H. Zweig and G. Campbell, “Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine.,” *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[73] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[74] P. F. Griner, R. J. Mayewski, A. I. Mushlin, and P. Greenland, “Selection and interpretation of diagnostic tests and procedures. principles and applications.,” *Annals of internal medicine*, vol. 94, no. 4 Pt 2, p. 557, 1981.

APPENDIX A

The critical values of Extreme Studentized Deviate (ESD statistics) based on an approximation provided by [65, 50].



Critical values for the ESD (Extreme Studentized Deviate) outlier statistic (ESD$_{n,1-\alpha}$, α = .05, .01)

	$1-\alpha$			$1-\alpha$	
n	*.95*	*.99*	*n*	*.95*	*.99*
5	1.72	1.76	**25**	2.82	3.14
6	1.89	1.97	**26**	2.84	3.16
7	2.02	2.14	**27**	2.86	3.18
8	2.13	2.28	**28**	2.88	3.20
9	2.21	2.39	**29**	2.89	3.22
10	2.29	2.48	**30**	2.91	3.24
11	2.36	2.56	**35**	2.98	3.32
12	2.41	2.64	**40**	3.04	3.38
13	2.46	2.70	**45**	3.09	3.44
14	2.51	2.75	**50**	3.13	3.48
15	2.55	2.81	**60**	3.20	3.56
16	2.59	2.85	**70**	3.26	3.62
17	2.62	2.90	**80**	3.31	3.67
18	2.65	2.93	**90**	3.35	3.72
19	2.68	2.97	**100**	3.38	3.75
20	2.71	3.00	**150**	3.52	3.89
21	2.73	3.03	**200**	3.61	3.98
22	2.76	3.06	**300**	3.72	4.09
23	2.78	3.08	**400**	3.80	4.17
24	2.80	3.11	**500**	3.86	4.23

Note: For values of *n* not found in the table, the percentiles can be evaluated using the formula ESD$_{n,1-\alpha}$ =

$$\frac{t_{n-2,p}(n-1)}{\sqrt{n(n-2+t^2_{n-2,p})}}$$ where $p = 1 - [\alpha/(2n)]$.

Figure A.1 *Critical values of extreme studentized deviate (ESD statistics).*