MARCHE POLYTECHNIC UNIVERSITY

Engineering Faculty

Department of Science, Materials and Environmental Engineering and Urban Planning - SIMAU

Master thesis topic:

"COMPARISON OF DIFFERENT METHODS FOR SENSOR FAULT DIAGNOSIS IN A MUNICIPAL WASTEWATER TREATMENT PLANT".

Author:

Reshad Ahmad Aria Faizy

Supervisor:

Prof. Francesco Fatone

Co-Supervisor:

Pourzangbar Ali

Academic year 2019/2020

Preface

This MSc. thesis is submitted as a requirement for obtaining a Master's Degree in Environmental Engineering at Marche Polytechnic University. It is written based on data analysis obtained from Peschiera Borromeo Wastewater Treatment Plant sensors, the work is conducted by the author, Reshad Ahmad Aria Faizy, under the supervision of Prof. F.Fatone, and co-supervision of Dr. Pourzangbar Ali at the Department of Science, Materials and Environmental Engineering and Urban Planning (SIMAU).

Additionally, would like to express my warm sincere gratitude to Ph.D. Students Marinelli Enrico, and Radini Serena for their kind supports and instructions during the completion of this thesis project.

Reshad Ahmad Aria Faizy

Ancona, Feb 2020

Abstract

Wastewater plants play an important role in removing the contaminants from wastewater and converting it into an effluent that can be reused for various purposes such as irrigation, among others. There are tightening treatment regulations on the quality of effluent regarding the values of the primary variables such as concentrations of ammonia, nitrates and total nitrogen, phosphates and total phosphorus, suspended solids, biochemical and chemical oxygen demand, as well as other process variables like the sludge blanket level. The conventional monitoring of these parameters contains on-line (via the field-installed probes) and off-line (implementing the laboratory experiments) analysis. The real-time monitoring is hard-to-measure, costly, time-consuming, and the field instruments need frequent maintenance that makes the field measurements reliability challenging. In other words, these sensors produce several anomalies during the recording process. Accordingly, to have accurate values of the effluent quality, it is of great importance to detect the outliers among the probe recorded data points. Before final utilization of data derived from the sensors, it is recommended to remove the observations that are not in line with the general data trend.

During the last two decades, different methods for data analyses and outlier detection have been presented by many authors. Among the proposed approaches, artificial neural networks (ANNs), principal component analysis (PCA), fuzzy logic, clustering, fisher discriminant analysis (FDA), independent component analysis, and partial least squares regression (PLS) have attracted much attention. We examined the performance of two different methods including PCA analyzes with T-square (T2) and the Modified Z-score in detecting the outliers and cleaning the data obtained from sensors at the effluent of the Peschiera Borromeo wastewater treatment plant in Milan. The robustness of the proposed methods is evaluated using the benchmark data derived from the laboratory measurements. The MATLAB scripts written based on the T-square and Z-score approaches are implemented to analyze the results and compare the performance of the methods.

The results suggested that both methods showed satisfactory results in the recognition of anomalies. However, the PCA outperforms the modified Z-score on the detection of clustered outliers. The number of outliers detected by the PCA outnumbers those of the moving filter method since the T-squared method is based on the applied radius for the dataset, the effect of the alpha (radius) and the selected interval is highly important. Moving median filter shows acceptable results close to the PCA with less number of outliers detected. The moving median

limit plays a significant role in the number of outliers. Different window sizes and moving median limits are tested to find the best fit where minimum bias is obtained with the maximum number of outliers.

For the future investigation, it is recommended to apply an optimization algorithm to adjust the alpha value in the PCA method that may lead to even better results in the outlier detection.

TABLE OF CONTENTS

Abbreviations

APCA	Adaptive Principle Component Analysis
BOD	Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
CUSUM	Cumulative Sum
DWC	Digital Water City
EDSS	Environmental Decision Support Systems
EPA	Environmental Protection Agency
EWMA	Exponentially Weighted Moving Average
EWPCA	Exponentially Weighted PCA
FDA	Fisher Discriminant Analysis
FDI	Fault Detection and Isolation
GLRT	Generalized Likelihood Ratio Test
MSPC	Multivariate Statistical Process Control
MSPCA	Multi-Scale PCA
OC	Organic Carbon
PCA	Principal Component Analysis
PLS	Partial Least Squares
SBR	Sequencing batch reactor
SOM	Self-organizing maps
SPC	Statistical Process Control
SPE	Sum of Squared Residuals
SVD	Singular Value Decomposition
SVM	Support vector machines
TSS	Total Suspended Solid
UV	Ultraviolet
WWTP	Wastewater Treatment Plant

List of figures

List of tables

Chapter 1 Introduction

1.1 Background

All Wastewater treatment plants are consisting of many complex and complicated processes with a variety of uncertainties and unknowns. Restrictions on water quality requirements became more and more stringent due to the increased awareness about the negative impact of eutrophication on the quality of water bodies (Ansari et al., 2010), therefore more advanced treatment systems are needed to satisfy the (tighter) standards not only for organic carbon (OC) but also for nitrogen (Nitrogen compounds) and phosphorus (nutrient) levels (Vanrolleghem & Lee, 2003, p.). The strict rules and limits on the effluent of WWTPs caused the researchers to consolidate the existing processes and develop new technologies to optimize treatment processes and monitoring of the WWTPs. Among them, sensor monitoring and audit have increasingly been implemented in most plants since these sensors can provide insight into the ongoing bio-processes. (Vanrolleghem & Lee, 2003). Three different kinds of sensors are mainly used in WWTPs, monitoring sensors, automatic control systems, and tools for plant auditing and optimization.

The soft-sensors presented in many case studies are highly effective and inexpensive technologies for extracting the data for WWTP processes and have a better comparison with laboratory data which are routinely acquired in biological wastewater treatment facilities (Haimi et al., 2013). These sensors provide real-time monitoring of the process and the measurement of nutrients concentration. The real-time data of the primary indicators are crucial for the effective utilization of advanced process control and optimization strategies in WWTPs (Haimi et al., 2013).

The proper and precise functionality of the sensors is invaluable for process control and optimization of the plant. Since the sensors are susceptible to faults (close interaction with wastewater) and wrong measures can put the aquatic life at risk, there is a strong interest in monitoring WWTPs in order to early detect and identify any fault or abnormality that might negatively affect the process which is not only important for ensuring plant safety and maintaining effluent qualities but also for the protection of the environment (Aguado & Rosen, 2008; Y. Liu et al., 2016).

According to the (Corominas et al., 2018) literature review on 340 relevant papers, the majority of papers discussed ANN, PCA, and fuzzy logic with an overall of 21%, 13%, and 12% respectively (Figure 1.1). Artificial neural networks (ANN) are mostly recommended in the literature for the prediction of process performance and control. PCA has mainly being used for fault detection and dimensionality reduction and fuzzy logic has been applied for control and prediction purposes (Corominas et al., 2018).

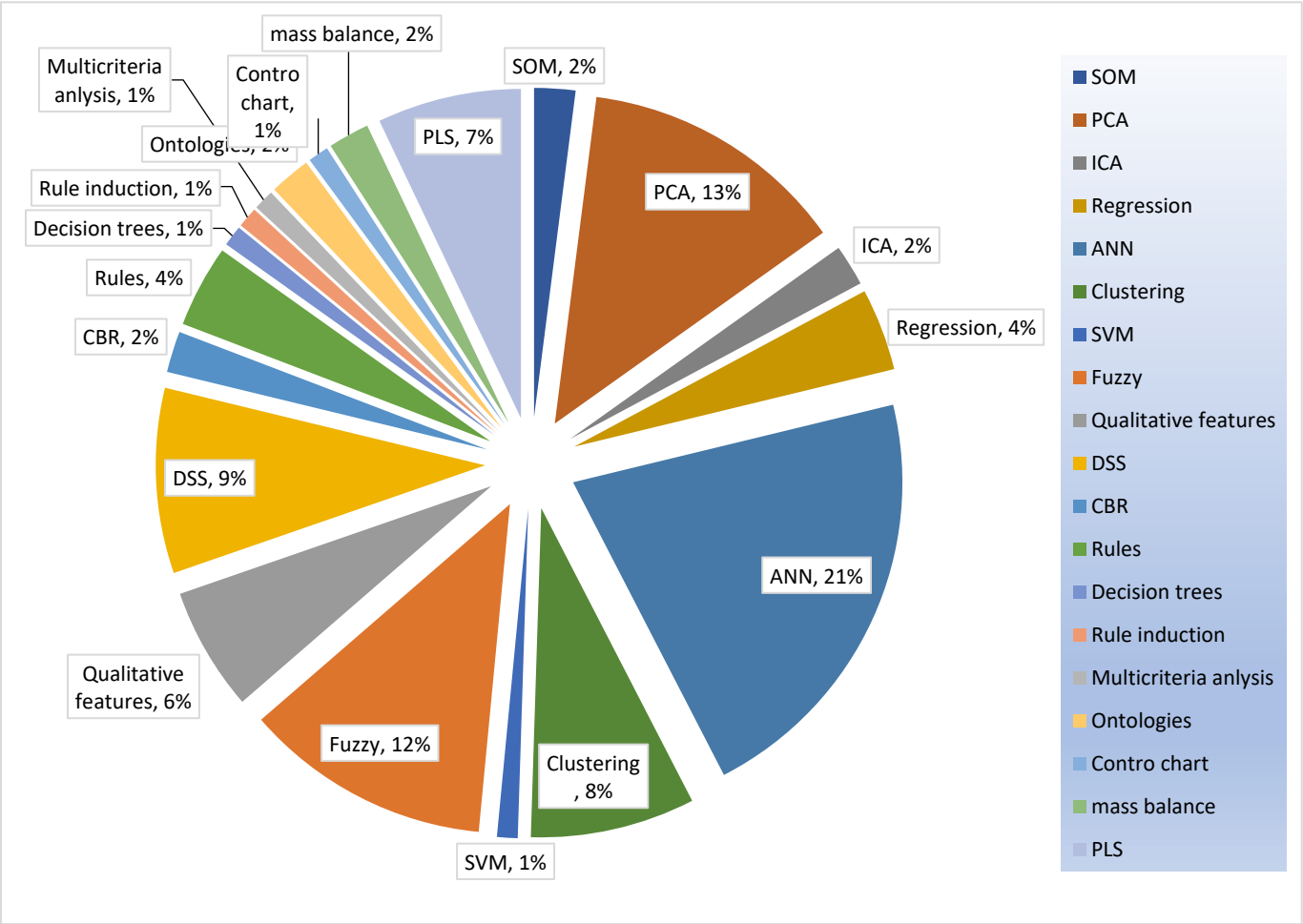

Figure 1.1 Percentage of papers for each of the techniques included in the literature review (Corominas et al., 2018)

Many outlier detections and prognosis methods have been developed throughout history, the timeline for some of the most used technique can be found in figure 1.2.

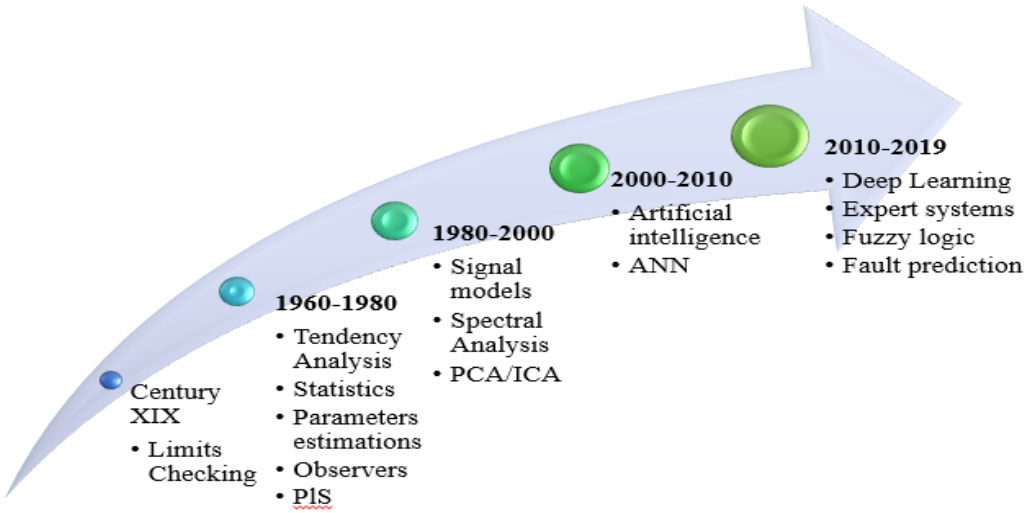

Figure 1.2 Timeline of outlier detection methods

1.2 Objectives and organization of the work

The overall goal of this Master's thesis is to explore the different methods of sensor outlier detection and present deeply two commonly used methods in outlier and noise removal from WWTPs datasets. The data used in this analysis were from Line-2 of the Peschiera Borromeo wastewater treatment plant. The data were constantly recorded over two years and sensors were used at the effluent of the plant to detect and record different compounds such as TSS, NH4, NNOx, and PPO4.

The work done here is part of the project of Digital Water City (DWC) which develops and demonstrates a dozen advanced digital solutions to address current and future water-related challenges. Under this project and to achieve digital solutions the online sensors were installed in Milan Peschiera WWTP to acquire real-time data and they are designed to promote safe water reuse and reduce the risk of microbial contamination of soils and crops during irrigation.

The work for this thesis started with a literature review of research papers on different methods of sensor fault detection and prognosis. In chapter three a short description of the plant and data collected from sensors is presented. In chapter four the results from two outlier detection methods are presented and compared, finally, in the last chapter the results are discussed and some recommendations for further improvement of the models are listed.

Chapter 2 Literature Review

2.1 Literature Review for sensor fault detection in the wastewater treatment plant

2.1.1 Data mining

Data mining is the application of methods to discover structures and patterns in large data sets. From this definition, it will be immediately manifested that the knowledge of data mining has significant overlap with other data analytic disciplines, especially statistics, machine learning, and pattern recognition (Hand & Adams, 2015).

Recently, the term Big Data has come to the superiority in place of data mining that broads the term data mining with a focus on larger and more diverse data sets and sources. Nowadays, Big data is often characterized as data having five "V" s: **V**olume (giant amount of data), **V**elocity (collect, sort, clean, analyze, and interpret data quickly), **V**ariety (different types of data), **V**eracity (correctness of the data), and **V**alue (the cost of data collection and storage relative to the value it produces) (Hand & Adams, (Newhart et al., 2019).

It seems obvious that wastewater treatment plants deal with very large data and optimization of each process is crucial in energy-saving and performance efficiencies. According to the EPA, wastewater treating processes accounts for more than 4% of the United State electricity consumption and by energy optimization of the WWTPs by just 10% could lead to an annual savings of at least $400 million (Asadi et al., 2017). WWTPs are obliged to clean wastewater from different contaminants therefore it uses a huge variety of technologies and methods which demand high energy and to optimize the processes many models were developed by data mining algorithms to have a clear and concise relationship among input and output variables and save some part of the energy used in the cleaning process(Asadi et al., 2017).

2.1.2 Methods for Data cleaning

Nowadays many WWTPs are equipped with digital controls and sensors, in order to have constant updates and on-time information on many processes. The accuracy of sensors is however still highly questionable and fault removal and noise detection of the raw data from WWTP sensors are highly challenging, also considering the strict relationship with risk

management and processes optimization. However, big data analysis in wastewater treatment plants (WWTP) is still widely underutilized (Newhart et al., 2019).

Despite common interest in big data integration at WWTP, most raw data are stored in their original format for potential future performance analyses with very few utilization prospectives. If data received from WWTP operations were analyzed in real-time with data-driven tools, WWTP could effectively detect and respond to process failures, inefficiencies, and abnormalities (Newhart et al., 2019).

Most wastewater treatment plants are operated by fixed upper or lower limits to monitor different processes. The limits are adjusted and based on a WWTP operator's background knowledge of the specific system for the online and offline water quality data (Newhart et al., 2019). Digital sensors are used to monitor water quality in real-time, transmitting a voltage to an electrochemical reaction or physical change inside the sensors as they interact with the environment (constituent concentration, flow rate, etc.). These sensors are calibrated using laboratory analyses which are correlated to voltage or current changes from the sensor. On the other side, solids deposition on sensors, biofilm formation, and precipitates can interfere with the sensor's measurement accuracy and causes some faults and noises (Newhart et al., 2019).

In the last few years, environmental problems have become more challenging in water, air, and heavy metal pollution which are directly related to the increasing number of chemical plants, and among them, wastewater treatment plants account for a high proportion of the pollution (H. Liu et al., 2021). To ensure that wastewater meets the discharge standards, it is required to have precise data about the concentration of some important indicators such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), NH4, P-PO4, N-NO3, and TSS. Due to the above-mentioned reasons, sensors are very susceptible to faults which leads to having inaccurate information about the WWTP discharge, thus to have clean data from sensors many methods have been developed for sensor fault detection and diagnosis which makes it easier to mathematically model to perform process control (H. Liu et al., 2021; Newhart et al., 2019).

Anomaly and fault detection is an important topic in currents researches, and having real-time cleaned data from different sensors is highly recommended to improve the performance of the plant. Many researchers developed different techniques of univariate and multivariate outlier detection. A deep explanation about anomaly detection in multivariate-sensing time-series can be found in Aguado & Rosen (2008) and Ding et al. (2018).

Most of the WWTPs are well equipped with a huge number of developed sensors However, all the processes in WWTPs are seriously interconnected with uncertainties, time-varying, and quite sensitive to disturbances which means that the identification of the location of the faulty sensor is quite difficult (Avella et al., 2011). Many fault detection methods have been developed to overcome the diagnosis difficulties. Multivariate statistical process methods such as principal component analysis (PCA) and partial least squares (PLS) are the two most frequent methods used in the fault detection and diagnosis schemes (Aguado & Rosen, 2008; D. Garcia-Alvarez et al., 2009; Diego Garcia-Alvarez, n.d.; Tao et al., 2013).

PCA techniques are usually followed by Hoteling statistics, T2, and the sum of squared residuals (SPE), or *Q* statistic to detect faults. The *T2* statistic is a measure of the variation in the PCA model (the major variation in the data) and the Q statistic is a measure of the amount of variation (random noise in the data) not captured by the PCA model (Diego Garcia-Alvarez, n.d., 2009).

Many other methods which are recommended by different authors are listed as following additionally according to the (Newhart et al., 2019) classification (Figure 2.1), where the fault detection methods which are implemented in WWTP and those which showed good performance are painted as green :

- Dynamic concurrent kernel partial least squares (H. Liu et al., 2021)
- Online reduced kernel PLS combined with GLRT for fault detection in chemical systems (Fazai et al., 2019)
- Fault detection T2 and SPE charts
- Statistical process control (SPC) charts such as Shewhart, CUSUM, and EWMA (Lee et al., 2004)
- Takagi–Sugeno (T–S) fuzzy logic (Wu & Ho, 2009)
- Neural networks and principal component analysis (NNPCA) (Fuente et al., 2012)
- Artificial Neural Networks (ANNs),
- Object-oriented fuzzy logic fault detection and isolation (FDI) (Genovesi et al., 2000)
- Auto-associative neural networks and ARMA model (Xiao et al., 2017)

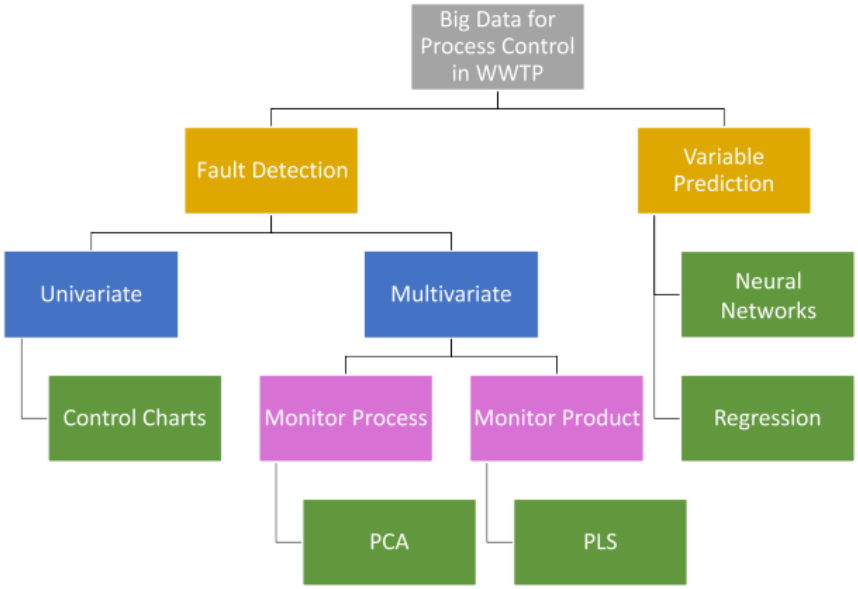

Figure 2.1 Methods mostly used in WWTPs colored as green (Newhart et al., 2019)

2.1.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most popular Multivariate Statistical Process Control (MSPC) monitoring and dimensionality-reduction methods which transforms a large set of variables into a smaller one that still contains most of the information in the large set (Diego Garcia-Alvarez, 2009).

To compute principal components, two methods are popular, Eigen decomposition and singular value decomposition (SVD) which both methods give us the same results. Considering a data matrix of $X \in R^{n \times m}$ containing *n* rows samples of *m* process variables, the first step is to perform standardization which means to make zero mean and unit variance and then the covariance matrix C is going to be calculated (Diego Garcia-Alvarez, 2009):

$$C = \frac{1}{n-1} X^T X \quad (1)$$

By performing SVD decomposition on C matrix:

$$M = U \Sigma V^* \quad (2)$$

Where Σ is a matrix wherein its diagonal are placed the non-negative real eigenvalues of matrix M sorted in decreasing order

$$\Sigma = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & \ldots & \lambda_k \end{bmatrix} (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0) \quad (3)$$

and U matrix contains all eigenvectors of matrix M. By choosing eigenvectors or columns (from left to the right) of U matrix with the highest eigenvalues we can produce the transformation matrix $W \in R^{m \times n}$ which transforms the measured variables into the reduced dimension space (Tao et al., 2013).

$$T = X\,W \quad (4)$$

Columns of matrix W are called loadings and elements of T are called scores. Scores are the values of the original measured variables that have been transformed into the reduced dimension space (Diego Garcia-Alvarez, 2009).

Matrix W contains equally the same number of rows and columns as the input matrix. To decide how many principles components are required to have the representation of almost all the data, simply we need to scree plot the eigenvalues where one plots the eigenvalues in decreasing order and looks for an elbow in the graph. According to figure 2.2a, we can observe a scree (a steep drop in the graph) after the first principle component and then the smooth line which we can only consider the first feature which contains maximum variance.

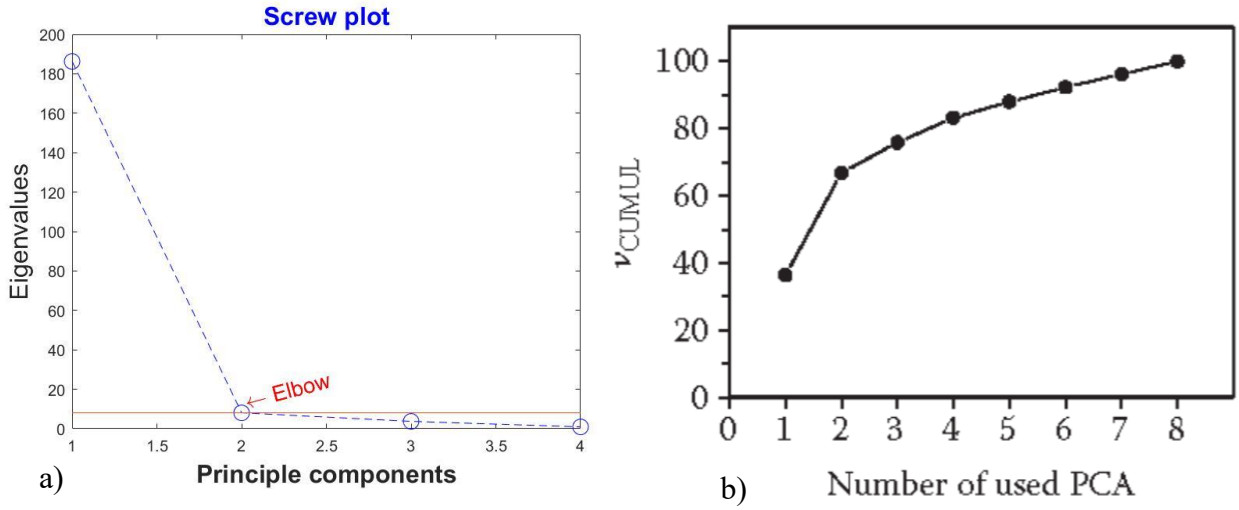

Figure 2.2 a) screw plot, b) *cumulative variance vs the number of components (Häggblom, n.d.)*

A simple method, illustrated by the **Error! Reference source not found.** is to plot the c umulative variance of scores (the variance for each new PC added to the variance of previous PCs) against the PC number, cumulative percent variance (CPV) principal is calculated as

following where k is the number of the principal components of PCA model and λ_j is the *jth* eigenvalue of covariance matrix:

$$CPV = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \times 100\ \% \tag{5}$$

To have the maximum of the variance in data, it is suggested that the PCs should explain at least 80 % to 90%, of the total variance (Häggblom, n.d.).

The result from PCA can be interpreted by Hotelling's T^2 and square prediction error (SPE) or Q method. T^2 represents the major variation in the data and Q represents the random noise in the data (Diego Garcia-Alvarez, 2009). T^2 is calculated by the following equation:

$$T^2 = x^T \, \mathrm{P} \, \Lambda_a^{-1} \, P^{-1} \mathrm{x} \tag{6}$$

where Λ_a is a diagonal matrix constructed by the first Eigen-values of Σ, which can preserve the maximum variance of Σ (Xiao et al., 2017).

The data which is processed considered normal if matches this criterion:

$$T^2 \leq T_a^2 = \frac{(n^2-1)a}{n(n-a)} F_a(a, n-a) \tag{7}$$

$F_a(a, n-a)$ is the Fisher-Snedecor distribution, $n-a$ degrees of freedom and a the level of significance varies between 90% and 95% (Diego Garcia-Alvarez, 2009).

Principal Component Analysis Summary:

1. Organize the data set as an n x m matrix, where m is the number of measured variables and n is the number of trials.
2. Normalize the data to have zero mean and unity variance.
3. Calculate the eigenvectors–eigenvalues of the covariance matrix.
4. Select the first eigenvectors as the principal components.
5. Transform the original data using the principal components (projection).
6. Use T- square method for outlier detection

Many statistical methods assume that data are normally distributed which is symmetric, unimodal, and bell-shaped, and only consider the mean and variance of the data. In the multivariate case additionally, covariance is considered. When the data are normally distributed, exact inferences can be made about the mean, variance, and covariance but when

the data does not meet the normality assumption (the data usually not distributed normally in WWTP) it is more difficult to identify the distribution of the statistic (Newhart et al., 2019).

2.1.4 Research Papers summary

This subsection represents the most relevant research papers which are reviewed about the data analyzes and sensor fault detection techniques. Recently due to the importance of fault detection and data cleaning of WWTPs before data analyses and utilization for optimization and control, many kinds of research have been done and the most recommended techniques are PCA, ANN, and PLS. Meanwhile, according to (Corominas et al., 2018) the majority of the 340 reviewed papers discuss ANN (21%) and PCA (13%) techniques for prediction of the process and fault detection respectively.

Multilayered perceptron artificial neural networks (MLP-ANNs) are the simplest and have many positive aspects like the high potential for process simulation, accuracy in prediction, and efficient process control of wastewater treatment plants therefore, are the most commonly used techniques in WWTPs network (Güçlü & Dursun, 2010).

Moreover, the ANN can predict plant performance with a correlation coefficient (R) between the observed and predicted output variables. Additionally, ANN provides an effective analyzing and diagnosing tool to understand and simulate the non-linear behavior of the plant, and can be used as a valuable performance assessment tool for plant operators and decision-makers (Nasr et al., 2012).

Despite ANN which is used for prediction purposes, Principal Components Analysis (PCA) is mostly used for detection tasks (Diego Garcia-Alvarez, 2009), different types of fault and noise can be detected with PCA such as Toxicity shock, inhibition, Bulking, etc. (Fuente et al., 2012).

In the following table, some of the outlier detection and prognosis methods are shortly listed and a brief introduction about the procedure applied on each method is presented.

Table 2-1 Summary of the reviewed papers with a short description of the procedure.

Author	The topic of the paper	Sensor fault detection method		Data used	Procedure	Important Notes
(Garcia-Alvarez et al. 2009)	**Fault Detection And Diagnosis Using Multivariate Statistical Techniques In A Wastewater Treatment Plant**	Multivariate Statistical Techniques	Principal Components Analysis (PCA) (detection tasks)	By following the historical databases to analyze the statistical model (PCA model)	● Accounting a data matrix $X \in R^{n \times m}$ with n samples ● Then determining the PCA to form the covariance matrix R ● To detect the faults, the FDA considers by accounting the collected data during various faulty conditions and apply a discriminant function that takes into consideration the similarity between the actual data and the data concerning each class	● Different types of fault and noise can be detected are such as Toxicity shock, inhibition, Bulking. ● Q statistic detects toxicity shock fault better than the T2 (Hotelling statistics) statistic
			Fisher discriminant analysis (FDA) (diagnosis tasks)	Using collected data from the operational plant		FDA provides an optimum lower-dimensional representation on basis of variance
(Fuente et al. 2012)	**Fault Detection In A Wastewater Treatment Plant Based On Neural Networks And PCA**	Neural network PCA (integrates neural networks (NN) and principal component analysis (PCA))	Neural networks are used to model the nonlinear dynamic system in nominal operating conditions. the residuals are evaluated by the PCA technique, using the classical charts, T2 and Q, to detect the faults,	The influent is used with a range of 15000−35000m3/d from dry and rainy weather events data files for biological wastewater treatment process with two refluxes: external, internal	The real characteristic of the plant is compared with the output neural network with a similar operational condition by using the residuals. Then, these residuals are calculated by the PCA technique, using the classical charts, T2 and Q for fault detection and the contribution evaluation to perform the fault isolation. The neural networks are employed to eliminate the non-linearities and the dynamic characteristics of the process	● The plant model was simulated for 28 days and 100 days respectively in closed-loop and open-loop configuration to determine the steady-state. ●The model is developed by using MATLAB and SIMULINK. ● The application of PCA from residual data and two control charts are built to determine the actual behavior ● If there is any fault, the contribution charts are used for root cause analysis for any particular variable. ● The results show that the NNPCA method is more reliable than classical PCA
(Garcia-Alvarez 2014)	**Fault Detection Using Principal Component Analysis (PCA) In A Wastewater Treatment Plant (WWTP)**	(PCA)	Classical Statistical Process Control (SPC) uses typical control charts, such as Shewhart charts, cumulative sum (CUSUM) charts, an exponentially weighted moving average (EWMA) chartsmonitoring a single variable.	Simulation of wastewater treatment plant based on previously recorded database	● In open-loop condition, the plant model has to be analyzed for 100 − 150 days ● In close-loop, the plant is simulated for 14 days, and faults are caused on the 7th day. The	● Faults are basically the disturbances in the plant which can be determined by several available methods to deal with such open issue like multi-scale PCA (MSPCA), adaptive PCA (APCA), recursive PCA, exponentially weighted PCA (EWPCA), dynamic PCA and Nonlinear PCA using auto-associative neural networks

			Multivariate statistical process control (MSPC)		samples for monitoring plant are taken 100 times per day. ● For principal calculation, the CPV approach is used with a 95% maximum variance level. ● For apprehending a more variable process, the most effective option will be seven principal components for a given data.	● The performance for detecting the toxicity shock fault is always better in Q statistic than T2 statistic
(Genovesi & Steyer 2000)	**Integrated Fault Detection And Isolation: Application To A Winery's Wastewater Treatment Plant**	integrated object-oriented fuzzy logic fault detection isolation (FDI)	Residual Generation Methods: 1-Sensor Fault (SF) 2- Sub Process Fault (SPF) 3-Process Fault (PF).	Data obtained with important information like symptoms from the industrial wine production plant	● From this plant, the biological wastewater process is considered for fault detection and isolation (FDI) strategy. ● For pilot plant operation of this plant 1 m^3 anaerobic digestion is performed. ● The control strategy is split into three main levels: sensor level, subprocess level (i.e., the local loops), and process level. ● The object-oriented framework is implemented for detection and isolation algorithms	● The approach with only one FDI method is not effective for ideal industrial application. ● The fuzzy logic based on the FDI approach is more accurate than a threshold comparison algorithm.
(Wu & W. C. Ho, 2009)	**Fuzzy Filter Design For Itˆo Stochastic Systems WithApplication To Sensor Fault Detection**	Takagi–Sugeno (T–S) fuzzy system (a weighted sum of some simple linear stochastic subsystems)	fuzzy-rule-dependent fault detection filters fuzzy-rule-independent fault detection filters	Data of Stochastic Systems	Formation of a dynamical system linked a residual generator. Then, a weighing function matrix is integrated with the fault for enhancing the fault detection performance	● The fuzzy model portrays the local linear input/output relations of the system. the fuzzy-rule-dependent filter is less conventional than the fuzzy-rule-independent filter, particularly for disturbance attenuation performance level. ● The fundamental idea of fault detection is to form a residual evaluation function by following residual signal for comparison with a predefined threshold.

(Güçlü & Dursun 2010)	**Artificial Neural Network Modeling Of A Large-Scale Wastewater Treatment Plant Operation**	Artificial Neural Networks (ANNs),	Back-propagation with momentum (BPM) learning algorithm Multilayered perceptron artificial neural networks (MLP-ANNs) are the simplest and therefore most commonly used neural network architectures	The central wastewater treatment plant in Ankara is designed for 3,900,000 populations equivalent to the average dry and storm weather flow rate is 765,000 and 1,530,000 m3/day respectively. The influent and effluent parameters are chosen from common parameters of aerobic wastewater treatment plants through a literature review.	The ANN application has three modeling steps: training, validation, and testing. These steps are adopted mainly for efficient connection within network blocks to eliminate any possible error and to obtain better performance output.	● The biggest advantage of using a neural network model is that it has a distinct ability to learn non-linear functions without knowing the prior information from important variables. ● The ANN modeling has many positive aspects like the high potential for process simulation, accuracy in prediction, and efficient process control of wastewater treatment plants. ● The output results can be compared using errors based on the difference between the original observed values and predicted values.
(Xiao et al. 2017)	**Fault Diagnosis And Prognosis Of Wastewater Processes With Incomplete Data By The Auto-Associative Neural Networks And ARMA Model**	auto-associative neural networks shallow and deep structure	Kernel Density Estimation (KDE) to alleviate the Gaussian assumption	The simulated model is performed into two platforms MATLAB/SIMULINK from process parameters collected from two wastewater treatment plants. The model consists of five compartment biological tanks (5999 m3) and a secondary settler (6000 m3).	Firstly, a shallow and a deep ANN are executed to model SPE statistics to recognize the missing values for reconstruction efficiently. Then, the obtained SPE can be integrated into the ARMA model for multi-step-ahead prediction. Thirdly, KDE can be used to reconfigure the control limit of fault diagnosis effectively.	● Hallow and deep AANNs achieved better performance relatively than the KPCA-based model ● the non-parametric technique to estimate probability density functions. ● Integrated framework with ANN and an ARMA for faultdiagnosis is used. ● An ANN helps to measure the missing data from another sensor. ● Multivariate optimization like the quasi-Newton method can be implemented to determine more than one missing value. ● The proposed method is good for simulation studies for receiving more effective results and can be applied for existing and new wastewater treatment plant along with BSM1.

			ARMA model (Auto-Regressive and Moving Average Model)			
(Liu et al., 2020)	**Monitoring Of Wastewater Treatment Processes Using Dynamic Concurrent Kernel Partial Least Squares**	dynamic concurrent kernel partial least squares (DCKPLS)		This study is based on nutrient removal for biological wastewater treatment plant with four biological processes two clarifiers, one sludge thickening tank, and a dewatering system. For controlling the monitoring system, six input variables and three output variables were considered.	utilizing augmented matrices	● Multivariate statistical process monitoring (MSPM) approaches are adopted for detecting the faults ● By applying the MSPM, the model can be more stable and efficient in faults detection and diagnostician. ● This approach is very important and helpful for quick fault detection techniques in the wastewater treatment plant.
		CKPLS			1. Apply KPLS on scaled X and Y to get the score matrix T, loading matrix P, and Q 2. Determine the predictable output then calculate singular value decomposition 3. Evaluate PCA with Ay principal components on unpredictable output results 4. Calculate the quality irrelevant variations and run PCA with Ax principal components to receive input principal score matrix Tx and process residual	

Chapter 3 Material and methods

3.1 Peschiera Borromeo WWTP

3.1.1 General description

Peschiera Borromeo WWTP is located in the peri-urban area of Milan, in Via Roma - Cascina Brusada. It has a treatment capacity of about 566000 PE and treats daily an average flow rate of 216000 m3/d. The plant has two separate treatment trains (i.e., Line 1 and Line 2), which treat the wastewater coming from the two sewer network sectors of the peri-urban region of Milan as described in section 2.1. Sewage coming from the area managed by Metropolitana Milanese are treated in Line 1, whereas the wastewater coming from the sewer sector controlled by CAP is treated in Line 2. Line 1 includes coarse screening, pumping station, fine screening, grit and oil removal, primary sedimentation, biological treatment for organic carbon removal, tertiary filtration combined with nutrient removal in BIOFOR reactor, and chemical disinfection with peracetic acid. Line 2 includes coarse screening, pumping station, fine screening, a compact SEDIPAC unit for grit and oil removal coupled with primary sedimentation, a BIOFOR unit for organic and nutrient loads removal combined with tertiary filtration, and a final disinfection treatment with UV.

The digital management of the plant is performed by remote control and a SCADA system for the continuous acquisition of online data measured at the WWTP. Laboratory analyses are performed periodically for influent and effluent characterization, as well as to control specific processes. Data from laboratory analyses will be uploaded and managed by specific software (e.g., WaterLims).

The dataset of process parameters is managed in the SCADA system by remote control. Equipment status and related alarms on electro-mechanical units are continuously monitored. It allows rapid intervention in case of anomaly detection.

Offline data about cumulative energy consumptions, chemical supply, sludge, and waste production and disposal are stored in internal management systems. Maintenance operations, internal reports, and emergency procedures follow specific and documented protocols.

Peschiera Borromeo WWTP is already equipped with a monitoring network of conventional sensors to verify effluent quality, in terms of NH4, N-NO3, PO4, and TSS. Other probes for the measurement of REDOX, Dissolved Oxygen, Temperature, and N-NO3, are installed in the biological unit.

In this thesis we are focused on probes data from wastewater treatment Line 2, detailed below:

Screening and pumping units are equipped with level radars, and alarms in case of malfunction of the electromechanical equipment. Energy meters are installed to measure dynamically the real energy consumption.

On the compact SEDIPAC unit, flow meters are installed to measure the effluent flow rates, as well as the sludge extraction. All the equipment for sludge extraction, oil removal system, and sludge conveyor are equipped with alarms. Energy meters measure electricity consumption. On the internal back-flush, that is sent back to the SEDIPAC, chemicals are dosed for phosphorus precipitation, and the related electromechanical equipment is provided with alarms.

BIOFOR reactor for biologic and nutrient removal combined with filtration is divided into 10 modules, 5 dedicated to pre-denitrification and 5 voted to organic removal and nitrification. In the aerobic compartments REDOX, Temperature, and Dissolved Oxygen are measured online with sensors, while the anoxic zones are provided with REDOX probes. In the internal recycle an N-NO3 analyzer is installed and the recycle flow rate is also measured. Backwashing is monitored with a flow meter, the flux is activated alternatively by temporization or by pressure signals from sensors installed on the surface of the filter. Energy meters are installed to monitor electricity demand.

In the UV disinfection unit, sensors are installed to monitor the UV light intensity. Maintenance operations are supported by a counter system with a threshold of a maximum of 10000 working hours for each lamp. Specific energy meters are installed to monitor the UV unit.

In the final effluent, a set of probes are installed to monitor in real-time several parameters. Online measures are available for NH4, PO4, N-NO3, and TSS measurements. A flow meter is installed to control the amount of treated water discharged.

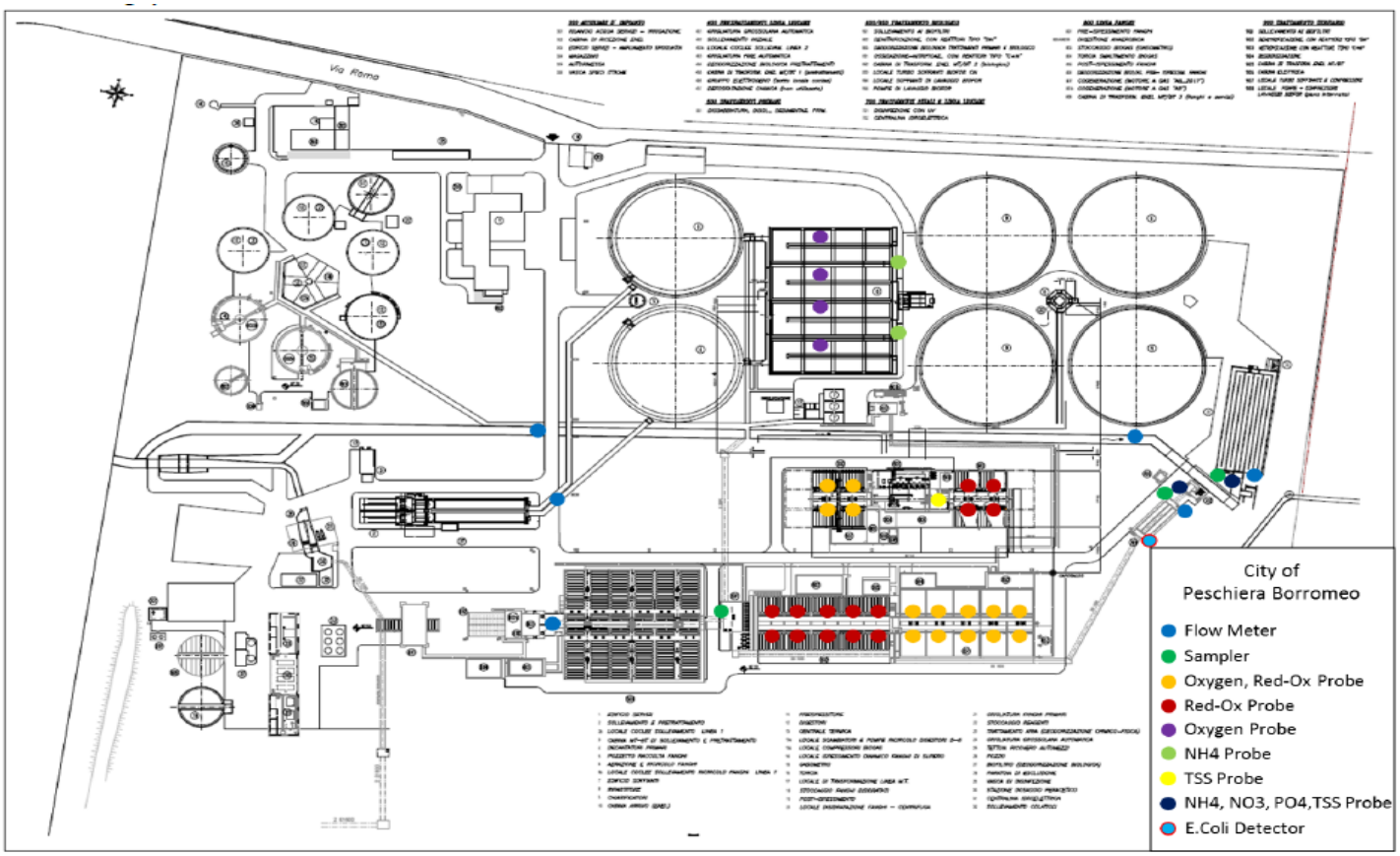

Figure 3.1 Online monitoring system and remote control

Moreover, the plant is equipped by many facilities to reduce odorous effluents and two CHP unit is used for heat and electricity production.

The water treatment lines are equipped with 47 instruments to monitor the performance of the plant, divided into Flow Meters, Probes, and Samplers.

Table 3-1 Instruments to monitor the performance of the plant

N°	Type	Description
6	**Flow Meter**	In Continuous monitoring the flow of water inlet/Outlet/by-passed
3	**Sampler**	Programmable automatic sampler
14	**Oxygen, Red-Ox**	In Continuous Monitoring Aerobic condition of Nitrification Reactors
14	**Red-Ox**	In Continuous Monitoring Anaerobic condition of De-Nitrification Reactors
4	**Oxygen**	In Continuous Monitoring Aerobic condition of activated sludge reactors
2	**NH4**	In Continuous monitoring, NH4 outlet of the first stage activated sludge reactors (colorimetric analyzer)
1	**TSS**	In Continuous monitoring TSS inlet of tertiary Biofor treatment (optical sensor)
2	**TSS, NH4, NO3, PO4**	In Continuous monitoring the Outlet of the two treatment lines to verify the respect of the limit and the overall performance TSS (optical sensor); NH4 and PO4 (colorimetric analyzers); NO3 (optical sensor)
1	**Bacteria detector**	Bacteria/non-bacterial detector (BACMON by Grundfos) to adjust in continuous light bulb intensity of UV final disinfection
-	**E. Coli analyzer**	Fluid ion will provide FLUID ALERT system within DWC project activities development

3.1.2 Sampling and periodical lab monitoring

The following samples are taken on a weekly basis, as 24h average:

- Inlet wastewater (both lines): COD (chemical oxygen demand), BOD5 (Biological oxygen demand), Ammonia (N basis and NH4 basis), total nitrogen, nitrate (NO3 basis), total phosphorous, TSS (total suspended solid), metals (Al; As; Cd; Cr; Mn; Ni; Pb; Cu; Zn; Fe), pH, conductivity, chlorides, phosphate, sulphate;
- Outlet treated water (both lines): COD (chemical oxygen demand), BOD5 (Biological oxygen demand), Ammonia (N basis and NH4 basis), total nitrogen, nitrate (NO3 basis), total phosphorous, TSS (total suspended solid), metals (Al; As; Cd; Cr; Mn; Ni; Pb; Cu; Zn; Fe), pH, conductivity, chlorides, phosphate, sulphate, E. Coli;
- At the inlet of the biological treatment (Biofor line 2): pH, conductivity; TSS (total suspended solid), COD (chemical oxygen demand), total nitrogen, total phosphorous.

The lab is inside the WWTP, certificated UNI CEI EN ISO/IEC 17025:2005, is available also for extra analyses (process optimization). The results are available within 24h (5 days for BOD5) and registered in software for data management (Water LIMS). The same system provides to send email and alert for any parameters that exceed the legislation limit.

3.2 Methods for probe data analysis

Data from any sensor may include observations that don't appear to belong with the rest of the dataset. If we understand in detail the mechanism that produces each observation, we might well be able to define most of these anomalous results. For example, an instrument or sensor may have been calibrated improperly, something may have been stacked on the surface of the sensor, an uncontrollable event may have affected the result, a recording error may have occurred or measurement may have been read incorrectly. In fact, we would act on such information by classifying the anomalous observations as outliers and setting them aside by removing them. Then our analysis would focus on the process that we intended to study (Iglewicz & Hoaglin, 1993).

The probe data obtained from Peschiera Borromeo WWTP were analyzed for outliers' removal, with two different outlier detection methods, The Moving median filter based on Modified Z-score and T-square method.

3.2.1 Modified Z-score

Standard scores or Z-scores are usually applied to screen data for outliers. The Z-scores are based on the well-known property of the normal distribution that if X is distributed as $N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ is distributed as N (0, 1). Therefore we are tempted to use the Z-scores of the observations $x_1, x_2, \ldots, x_n$ as a method for labeling outliers (Ivanushkin et al., 2019):

$$z_i = (x_i - \bar{x})/s, \text{ where } s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \quad (1)$$

Where $\bar{x}$ and s indicate the mean and standard deviation of the sample, respectively.

One popular rule labels Z-scores that exceed 3 in absolute value ($|z_i| > 3$) as outliers. This approach has been used in most statistical software packages and even simple calculators. But surprisingly this method fails in sample size data as small as n = 10, therefore it is not appropriate to think of Z_i as being approximately normally distributed. Moreover, the disadvantage of this method is that the outliers influence the mean values and the standard deviation as the z-score method relies on these values for measuring the central tendency and large outliers can lead the results to be inadequate. So it is concluded that Z-scores are not satisfactory for outlier labeling, especially in small data sets (Iglewicz & Hoaglin, 1993; Ivanushkin et al., 2019).

The selection of an inappropriate distributional model may produce outliers. A common mistake among the researchers is to assume that the data are normally distributed when, in fact, they come to a large extent from different distribution such as WWTPs. Such data may usually contain observations that seem to be outliers due to the wrong choice of distribution. By applying a proper statistical model, usually it can be realized that these observations are not truly outliers, but they are only unlikely observations from the normal distribution. In such situations, the result for outlier detection can lead to a more suitable statistical model (Iglewicz & Hoaglin, 1993).

As Z-score is affected by sample size, therefore, the alternative to Z-scores is resistant estimators. To be successful, the estimators should not be unduly affected by changes in a fair proportion of the sample. Such estimators are said to have a high breakdown bound or point (Iglewicz & Hoaglin, 1993). Iglewicz & Hoaglin (1993) in a very thorough and quite readable book on outliers' detection, stated that "the breakdown point of an estimator is defined as the largest proportion of the data that can be replaced by arbitrary values without causing the

estimated value to become infinite. Thus, the sample means, standard deviation, and range have breakdown points of zero, as one observation moved to infinity would make these estimators infinite".

The Modified z-score uses median and estimator MAD (the median of the absolute deviation about median) instead of the mean and standard deviation of the sample data therefore it reduces the influence of the outliers on the score (Iglewicz & Hoaglin, 1993; Ivanushkin et al., 2019). Median Absolute deviation (MAD) from the median was rediscovered and popularized by Hampel (1974).

$$MAD = \text{median}_i\{|x_i - \tilde{x}|\} \tag{2}$$

This estimator also has an approximately 50% breakdown point and is slightly easier to compute. The median and MAD now leads to a modified Z-scores, defined as follows:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \tag{3}$$

where 0,6745 – is the 0,75th quartile of standard normal distribution (Ivanushkin et al., 2019), this means that for normally distributed data, one-half of the data is within 2/3 of a standard deviation unit of the mean (Important Z-Scores, n.d.). the constant is needed because E(MAD) = 0.6745σ for large n (Iglewicz & Hoaglin, 1993; Ivanushkin et al., 2019).

The M score or median (M) is, like the mean, a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers. (Leys et al., 2013).

Observations will be considered as outliers when $|M_i| > D$. And the suggested value by the author is D = 3.5 (Iglewicz & Hoaglin, 1993).

For a better understanding of the value 0,6745, we will briefly check the relationship between MAD and standard deviation which has been explicitly described by (Hoaglin et al., 1983) and presented in Wikipedia. In order to use the MAD as a consistent estimator for the estimation of the standard deviation σ, one takes ("Median Absolute Deviation," 2020):

$$\hat{\sigma} = k \cdot \text{MAD},$$

where k is a constant scale factor, which depends on the distribution.

For normally distributed data k is taken to be:

$$k = 1/(\Phi^{-1}(3/4)) \approx 1.4826,$$

i.e., the reciprocal of the quantile function Φ^{-1} (also known as the inverse of the cumulative distribution function) for the standard normal distribution $Z = (X - \mu)/\sigma$. The argument 3/4 is such that $\pm$MAD covers 50% (between 1/4 and 3/4) of the standard normal cumulative distribution function ("Median Absolute Deviation," 2020), i.e.

$$\frac{1}{2} = P(|X - \mu| \leq \text{MAD}) = P\left(\left|\frac{X - \mu}{\sigma}\right| \leq \frac{\text{MAD}}{\sigma}\right) = P\left(|Z| \leq \frac{\text{MAD}}{\sigma}\right).$$

Therefore, we must have that:

$$\Phi\left(\frac{\text{MAD}}{\sigma}\right) - \Phi\left(-\frac{\text{MAD}}{\sigma}\right) = \frac{1}{2}.$$

Noticing that:

$$\Phi(-\text{MAD}/\sigma) = 1 - \Phi(\text{MAD}/\sigma),$$

we have that $\frac{\text{MAD}}{\sigma} = \Phi^{-1}\left(\frac{3}{4}\right) = 0.67449$ from which we obtain the scale factor

$$k = 1/\Phi^{-1}(3/4) = 1.4826.$$

Another way of establishing the relationship is noting that MAD equals the half-normal distribution median:

$$\text{MAD} = \sigma\sqrt{2}\text{erf}^{-1}(1/2) \approx 0.67449\sigma.$$

This form indicates the portable error in probable error ("Median Absolute Deviation," 2020).

In this thesis project, in order to have a better result due to the high amount of outliers which not only affect the median but also MAD, the moving median and moving MAD have been applied and the different appropriate window sizes for each compound are considered.

The Moving median which is based on a modified Z-score operates over a moving window of values. For each window, it was selected the middle value (in terms of rank) found in the window. If the window value is an even number, the average value of the two middle-ranking values will be used.

The moving median of a data set can be calculated as follow:

$$z(n) = \text{median}([\ldots x(n-2), x(n-1), x(n), x(n+1), x(n+2) \ldots]) \quad (4)$$

Those ns that do not exist are not considered in the median calculation. For example, if we assume the window size as 5:

$$z(1) = \text{median}(x(1-2), x(1-1), x(1), x(1+1), x(1+2)])$$

$$z(1) = \text{median}(x(-1), x(0), x(1), x(2), x(3)])$$

As there is no minus index data in our dataset, therefore the moving median for the first index data will be:

$$z(1) = \text{median}([\, x(1), x(2), x(3)])$$

$$z(2) = \text{median}([\, x(1), x(2), x(3), x(4)]) \text{ and so on.}$$

As we are calculating the moving median for each data, therefore we will obtain the number of moving median the same as our dataset.

3.2.2 T- Squared (T^2) method

Outliers are the set of objects that are far similar to the remainder of the data which highly affect the accuracy and applicability of the data. Outlier detection is extremely important aside from having more accurate data but can be used for direct detection of fraud, criminal activities in e-commerce, and detecting suspicious activities in businesses (Bolton & Hand, 2002).

According to (Zhang & Wang, 2006), there is no single universally applicable or generic outlier detection approach. Therefore, many approaches have been proposed to detect outliers, these approaches can be classified into four major categories, distribution-based which described in (Hawkins, 1980), distance-based discussed in (Knorr & Ng, 1998), density-based explained in (Breunig et al., 2000) and clustering-based approaches (Zhang & Wang, 2006). Among them, the clustering technique is more popular (Jain & Dubes, 1988).

Clustering-based approaches consider clusters of small sizes as clustered outliers (i.e., clusters containing significantly fewer points than other clusters) (Loureiro et al., 2004). The advantage of the clustering-based approaches is that they do not have to be supervised (Jayakumar & Thomas, 2013).

In this master thesis, the clustering outlier detection based on the square of Mahalanobis distance has been used. In statistics, Mahalanobis distance is a measure introduced by P. C. Mahalanobis (1936), which is based on the idea of measuring how many standard deviations away a point is from the mean (Jayakumar & Thomas, 2013). This distance grows as P moves away from the mean along each principal component axis and takes into account correlations of the data set. Generally, this distance is calculated as follow:

$$\textbf{(Mahalanobis distance)}_{\,i} = \sqrt{(X_i - \bar{X})^T S^{-1} (X_i - \bar{X})} \tag{5}$$

Where $\bar{X}$ is the data mean matrix and S is the data covariance matrix.

In test statistics, the squared Mahalanobis distance was proposed by Harold Hotelling which is known as T-square distribution and it is given as follow (Jayakumar & Thomas, 2013):

$$T_i^2 = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X}) \tag{6}$$

In this method, the data is first mapped from the X region to the Y region. In the X region, the coordinate axes are correlated and dependent but wherein Y space, there is no dependence of the coordinate axes. For this purpose and to map the data into a new dimensional space the Principle Component Analysis (PCA) technique is used.

Then, the data is transferred from Y dimensional space to Z dimensional space, which contains a spherical space with a radius equal to one (figure 3.2), assuming that the data are normally distributed, otherwise we consider the radius of the sphere to be alpha (α).

$$\|Z\| = X . S^{-1} . X^T \leq \alpha \tag{7}$$

Where $X = \begin{Bmatrix} X(1,1) & X(1,2) & \cdots & X(1,P) \\ X(2,1) & \dots & \cdots & \dots \\ \dots & \dots & \cdots & \dots \\ \dots & \dots & \cdots & \dots \\ X(n,1) & \dots & \cdots & X(n,P) \end{Bmatrix}_{n \times P}$, covariance matrix

$S = \begin{Bmatrix} S(1,1) & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & S(P,P) \end{Bmatrix}_{P \times P}$, $X^T = \begin{Bmatrix} X(1,1) & X(2,1) & \cdots & X(n,1) \\ X(1,2) & \dots & \cdots & \dots \\ \dots & \dots & \cdots & \dots \\ \dots & \dots & \cdots & \dots \\ X(1,P) & X(2,P) & \cdots & X(n,P) \end{Bmatrix}_{P \times n}$

And $S(1,1) = \frac{1}{n-1} [X(1,1) \times X(1,1) + X(2,1) \times X(2,1) + \dots + X(n,1) \times X(n,1)]$

The data that is inside this sphere (i.e. smaller than α) is listed as normal data and the data located outside the sphere are considered as outliers.

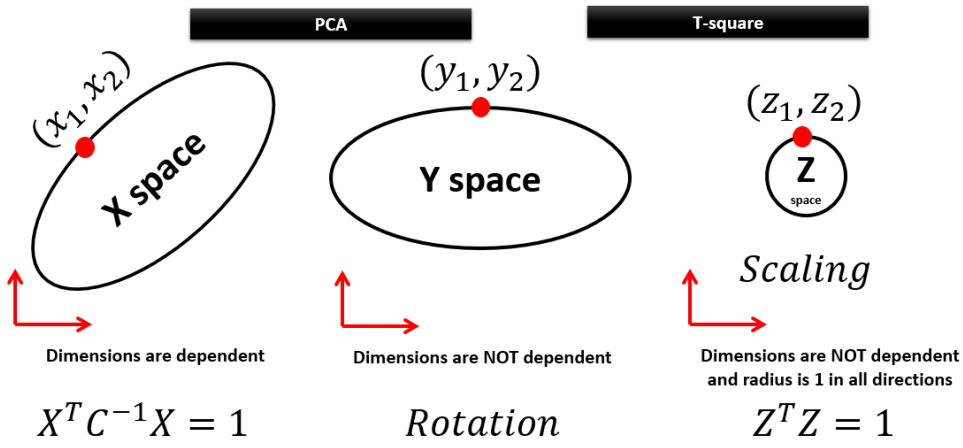

Figure 3.2 PCA with T-squared

Determining the range of alpha depends on the distribution of the data and the expert on the recorded data. For normally distributed data, if the alpha is equal to 1, it means that 68.27% of the data are normal and the rest of the data can be discarded and could be outliers (Figure 3.3). If the alpha is 2, so we consider 95.45% of the data to be normal and the rest as the outlier.

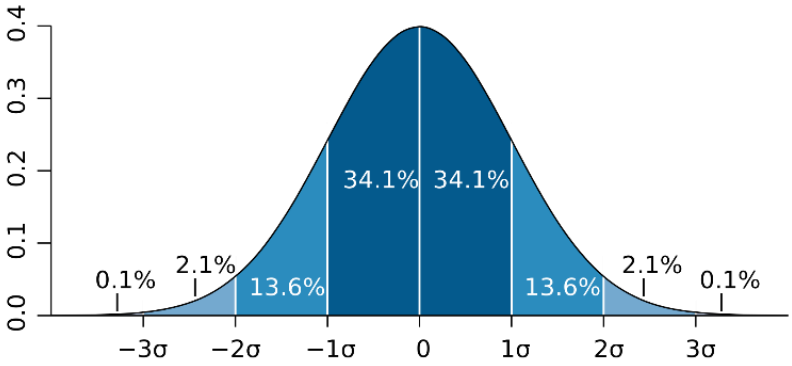

Figure 3.3 Normal distribution diagram ("Normal Distribution," 2021)

If the data have a normal distribution and are statistically normalized (the mean is zero and the standard deviation is one) it means by subtracting the mean of the data and dividing it by standard deviation, statistically normalized data is obtained. Therefore, with Alpha equal to one, two, and three we obtained that 68.27%, 95,45 %, and 99.73% of the data is normal respectively. However, for not normally distributed data (in our case), it is better to calculate the average of normal data and compare it with the number, for example, 0.95. If the average normal data is less than 0.95, therefore recommended increasing the alpha.

3.3 The procedure applied to Sensor data

Peschiera Borromeo WWTP is located in the peri-urban area of Milan, It has a treatment capacity of about 566000 PE, and treats daily an average flow rate of 216000 m^3/d. The plant has two separate treatment trains (i.e., Line 1 and Line 2), which treat the wastewater coming from the two sewer network sectors of the peri-urban region of Milan. Since line two was selected for experimental activities a set of probes are installed to monitor in real-time several parameters. Online measures are available for NH4 (from AMTAX sc sensor), PO4 (Phosphax sc sensor), N-NO3 (Nitratax sc sensor), and TSS (Solitax ts-line sc sensor) measurements.

To check the reliability of the sensor data, a preliminary modification and clean-up of the data for both method (moving median filter based on modified Z-score and T-squared) were performed as follow:

- All the blank dates were eliminated
- Missing values removed;
- The cells that contained zero values were eliminated;
- When duplicates values in the same date occurred, the second value was eliminated;
- To compare sensor data with lab data the daily average on sensor data is employed and the values lower than the lab threshold set to the minimum threshold.
- Lab data that does not have recorded sensor data were eliminated

Data analysis was performed using the programming language MATLAB R2020a. The preliminary cleanup was very effective in the elaboration of procedures, used in the phase of "data processing", to detect anomalies and clean the raw dataset and avoid errors from the programing language MATLAB.

Several MATLAB built-in functions were tried to clean the data such as rmoutliers, filloutliers, isoutlier, and smoothdata (some of the results can be found in appendix 1). Since these functions are designed in such a way that time-series data follows a particular seasonal trend and the data which stand alone and far from the recognized cyclic pathway can be identified and removed. Whereas the dataset acquired from the probes was not characterized by a definite trend. Therefore, two other methods were applied to detect the outliers.

Raw data from sensors were collected at Peschiera-Borromeo WWTP from 2018 till 2020 (table 3-3), simultaneously laboratory data were provided with an irregular sampling

period (table 3-4). Figure 3.4 – Figure 3.7 are presented the raw data collected by sensors with corresponding laboratory measurements. By visual observation, it can be seen the presence of unreliable measurements in the figures.

Table 3-2 Lab data properties

The properties of measured data by the laboratory					
Parameter	**Data points**	**First measurement (month/day/year)**	**Last measurement (month/day/year)**	**Min of data**	**Max of data**
NH4	129	01/08/2018 January	07/30/2020 July	0.5	13.5
NNO3	102	01/08/2018 January	07/23/2020 July	2.1	13.0
PPO4	102	01/08/2018 January	07/23/2020 July	0.2	1.3
TSS	129	01/08/2018 January	07/30/2020 July	5.0	38.0

All the sensor data provided are from the 4th of January 2018, except TSS (12th of November 2018). The sensor data contains zero recorded values that have been removed before preliminary clean-up. Lab data which are after the last date of the sensor measurement (14th of July 2020) are not considered in calculations

Table 3-3 Raw sensor data properties before preliminary clean-up

The properties of measured data by Sensors					
Parameter	**Data points**	**First measurement (month/day/year)**	**Last measurement (month/day/year)**	**Min of data**	**Max of data**
NH4	61117	01/04/2018 January	07/14/2020 July	0.0	52.042
NNO3	73708	01/04/2018 January	07/14/2020 July	0.0	46.192
PPO4	55014	01/04/2018 January	07/14/2020 July	0.0	11.54
TSS	65376	11/12/2018 November	07/14/2020 July	0.67	37805.38

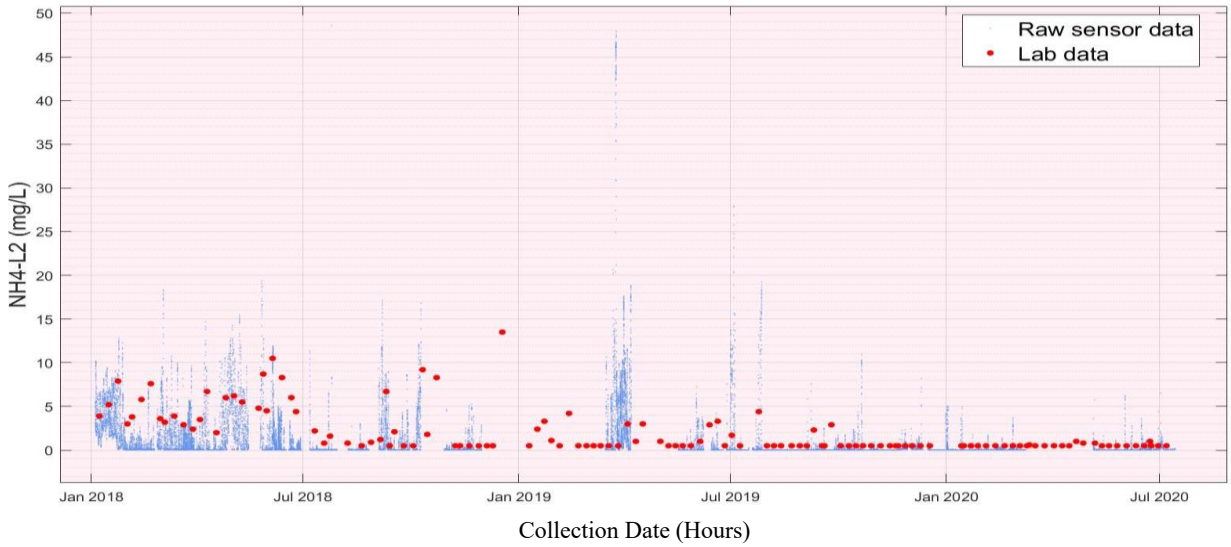

Figure 3.4 Raw sensor data (blue) and lab data (red) for NH4 line 2

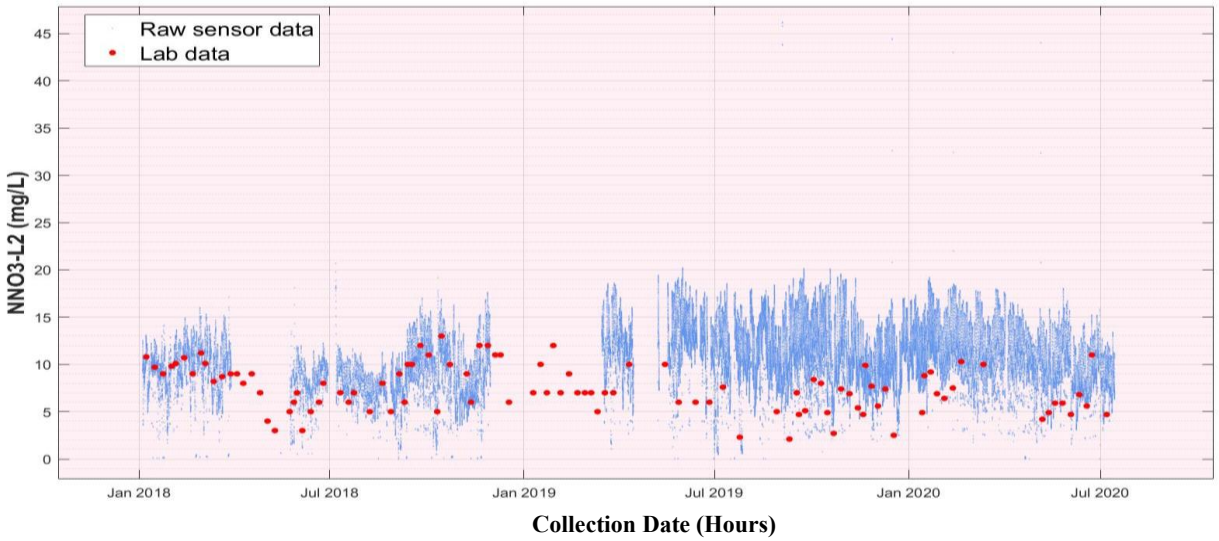

Figure 3.5 Raw sensor data (blue) and lab data (red) for NNO3 line 2

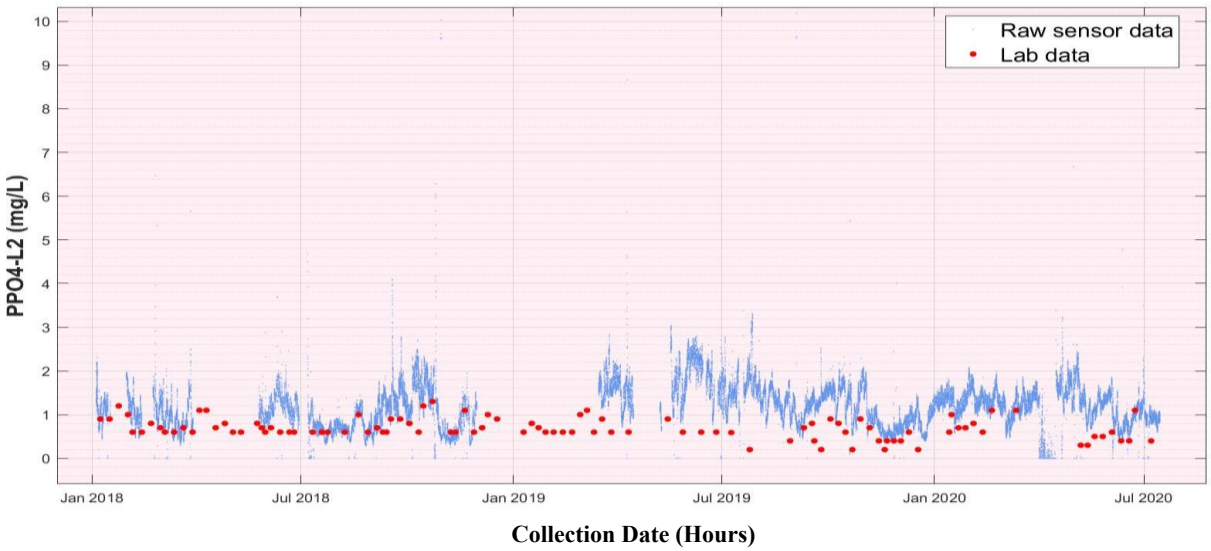

Figure 3.6 Raw sensor data (blue) and lab data (red) for PPO4 line 2

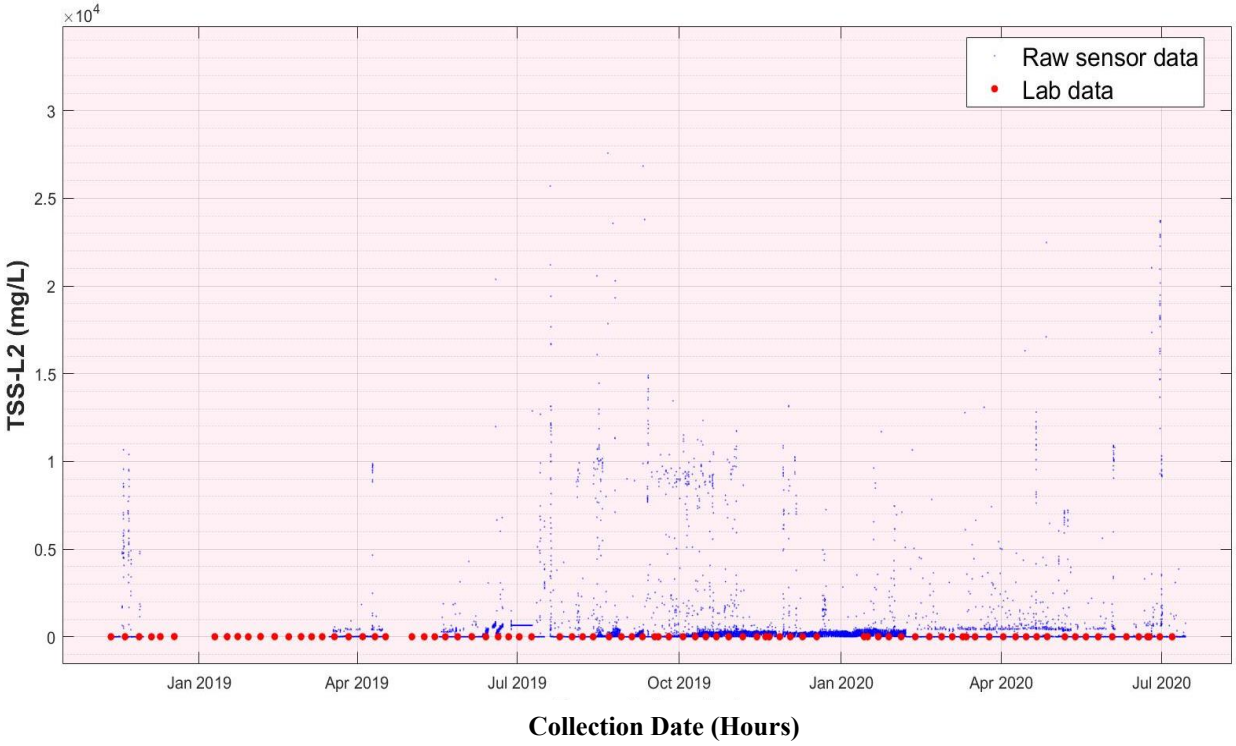

Figure 3.7 Raw sensor data (blue) and lab data (red) for TSS line 2

According to (Samuelsson, 2017), different statistics parameters can be considered to evaluate the sensor performance such as accuracy, precision, bias, trueness, repeatability, long-term stability, reproducibility, response time, calibration uncertainty, non-linearity, measurement noise, coefficient of variation, and limit of detection and quantification. The most relevant standard guideline for online sensors at water resource recovery facilities (WRRFs), is ISO 15839:2003 ("Water quality- On-line sensors equipment for water – Specifications and performance tests") (Samuelsson, 2017).

To compare and evaluate the accuracy of the final and selected methods (Pourzangbar et al., 2017) recommends calculating the following parameters, Correlation Coefficient (CC), the Root Mean Square Error (RMSE), the Scatter Index (SI), and the BIAS as given by equations (8-11), where O_i and P_i denote the observed and predicted outputs, respectively. N is the number of observed data, P_m and O_m are the corresponding mean values of the predicted and observed outputs.

$$CC = \frac{\sum_{i=1}^{N}\left(O_i - \bar{O_m}\right)\left(P_i - \bar{P_m}\right)}{\sqrt{\sum_{i=1}^{N}\left(O_i - \bar{O_m}\right)^2 \times \sum_{i=1}^{N}\left(P_i - \bar{P_m}\right)^2}} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}} \quad (9)$$

$$SI = \frac{\sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}}}{\bar{O_m}} \times 100\% \quad (10)$$

$$\text{BIAS} = \frac{\sum_{i=1}^{N}(P_i - O_i)}{N} \quad (11)$$

The CC shows how the predicted and measured data points are close to the best fit line, the Root Mean Square Error (MSE) is the standard deviation of the residuals (prediction errors). RMSE is a measure of how these residuals are spread out, or how concentrated the data is around the best fit line. The SI gives the percentage of expected error for the parameter. BIAS shows the predictions are underestimated (when BIAS ≤ 0) or overestimated (when BIAS ≥ 0) (Teshnehdel et al., 2020).

In this thesis project, raw sensor data were compared with laboratory measurements to evaluate the bias produced by on-line sensors and the mean absolute relative bias of cleaned (after applying the outlier detection methods) and raw data is compared. Sensor data were pre-processed before the bias calculation (preliminary clean-up).

Samples for laboratory analyses were based on a composite 24-hour collection and therefore the hourly lab result is not available. Thus, sensors' data measurements were daily averaged to be comparable with laboratory measurements, and on those days that lab data is not available the sensor daily average is removed accordingly. Then, the mean absolute relative bias of the sensor measurement was calculated by the difference between the daily average of sensor data and the lab measure divided by the lab measure:

$$\boldsymbol{Relative\ Bias} = \frac{\sum_{i=1}^{N}\left|\frac{(S_i - L_i)}{L_i}\right|}{N} \ggg \frac{\sum_{i=1}^{N}\left|\frac{Sensor\ data_{daily\ average} - Lab\ measure}{Lab\ measure}\right|}{N} \quad (12)$$

In this formula, if the bias approaches zero, it means that the sensor detected the same value as the lab data, which shows the highest accuracy.

The following formula is used to check the improvement of the bias:

$$Bias\ accuracy\ improvement = \frac{Bias_{Raw\ data} - Bias_{cleaned\ data}}{Bias_{Raw\ Data}}\ \% \qquad (13)$$

If the relative bias accuracy improvement approaches 100%, it means that the cleaned bias is close to zero which shows the highest cleaning rate of the data which sensor data is similar to the lab measurement.

Chapter 4 Analyses and Results

4.1 NH4 sensor data cleaning

The data acquired for NH4-N and NH4 is from AMTAX-sc online analyzer which is used in the water industry. The measure is performed through a gas selective electrode (GSE) that uses liquid to gas- phase conversion. The analyzer is equipped with an autonomous system for automatic self-calibration and cleaning. Instrumentation includes a humidity sensor to detect leakage and automatically initiate a safe shutdown. The accuracy of the sensor is 3 % + 0.05 mg/L (using standard solutions).

The NH_4^+ sensor is characterized by several flat periods (value ~0) lasting from a few hours to several days; the majority of these flat periods correspond to days with laboratory NH_4^+ concentrations lower than the detection limit therefore all zero measurements were deleted before applying data cleaning method.

After preliminary clean up (which is explained in chapter 3) of the data in Microsoft Excel, the moving median filter and T-squared method is applied.

Several window size ranges and moving median limits were attempted to find the best interval to reduce the bias, although there were no correlations between constant window size and rolling moving median limit, therefore random iteration is performed to find the best fit (Table A).

The initial bias of the raw sensor data was 0.3127 compared to the laboratory data and after cleaning of the data with moving median it reduced to 0.3052, which shows 2.42% ($\frac{0.3127-0.3052}{0.3127}$%) improvement in accuracy of cleaned data compared to the raw data. Based on the equation of (12) in chapter three, as the bias tends to zero the quality of the sensor data increases.

Figure 4.1 illustrates the moving median filter applied on NH4 raw sensor data and the points which are located far away from the median are considered as outliers. Figure 4.2 shows a visual image of the raw data with rolling median and cleaned data with cleaned rolling median, the daily average of both cleaned and raw data vs lab data is presented in Figure 4.3 and cleaned sensor data with lab data can be seen in Figure 4.4.

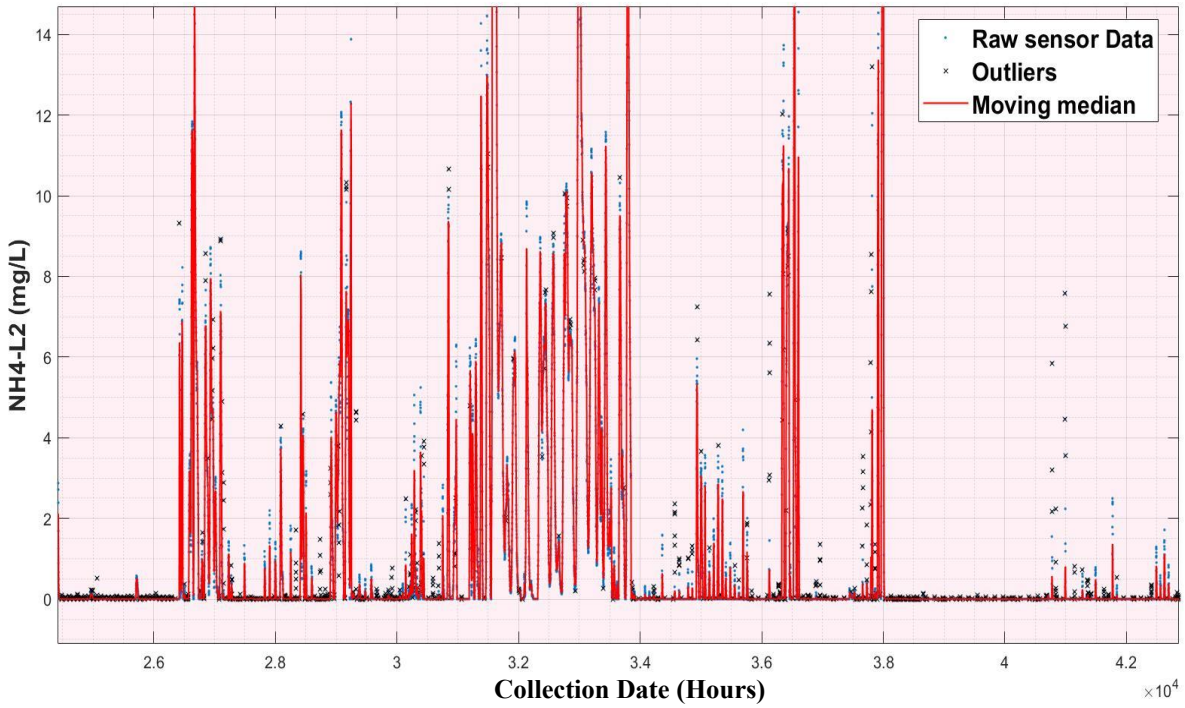

Figure 4.1 Moving median filter outlier detection on NH4 sensor raw data (window size 135, moving median limit 2.5)

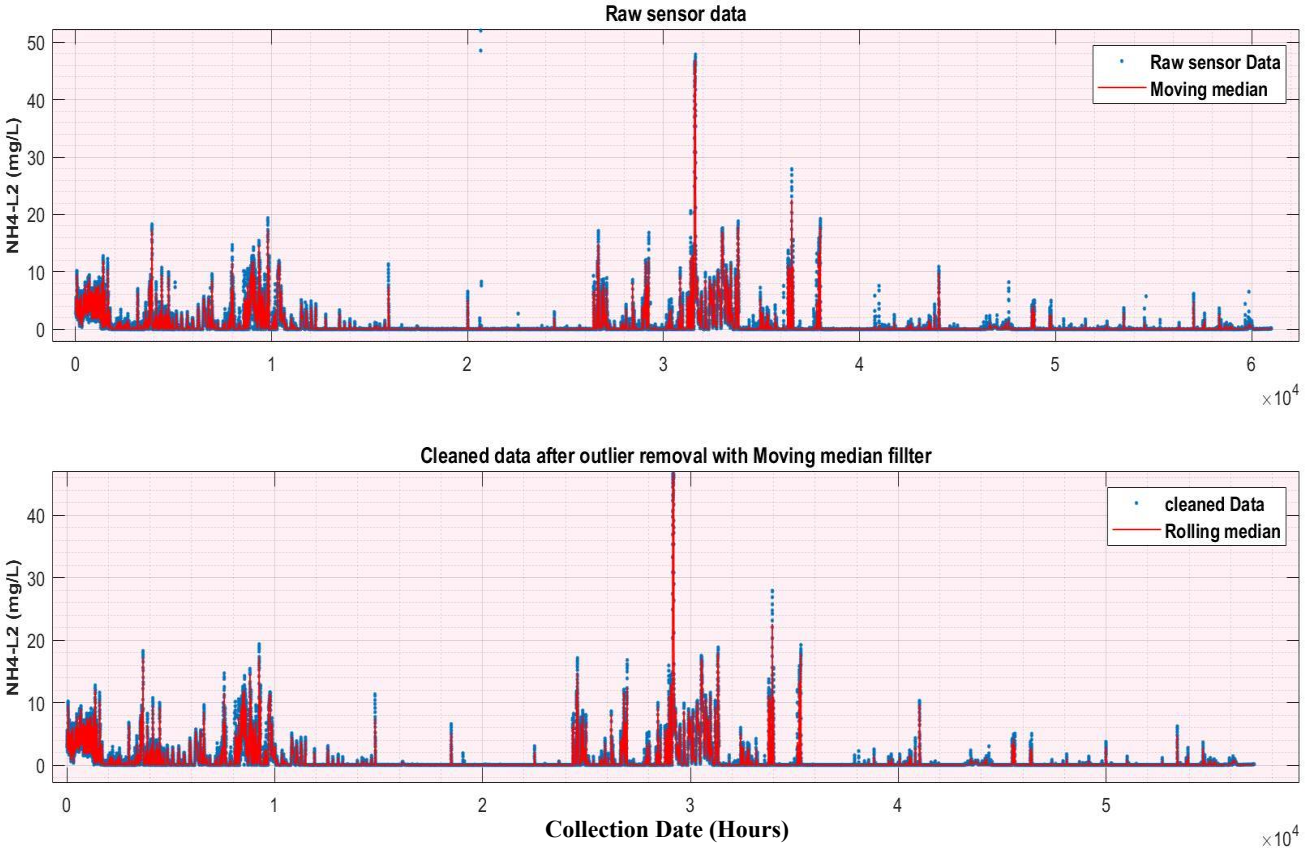

Figure 4.2 Moving median vs raw and cleaned data

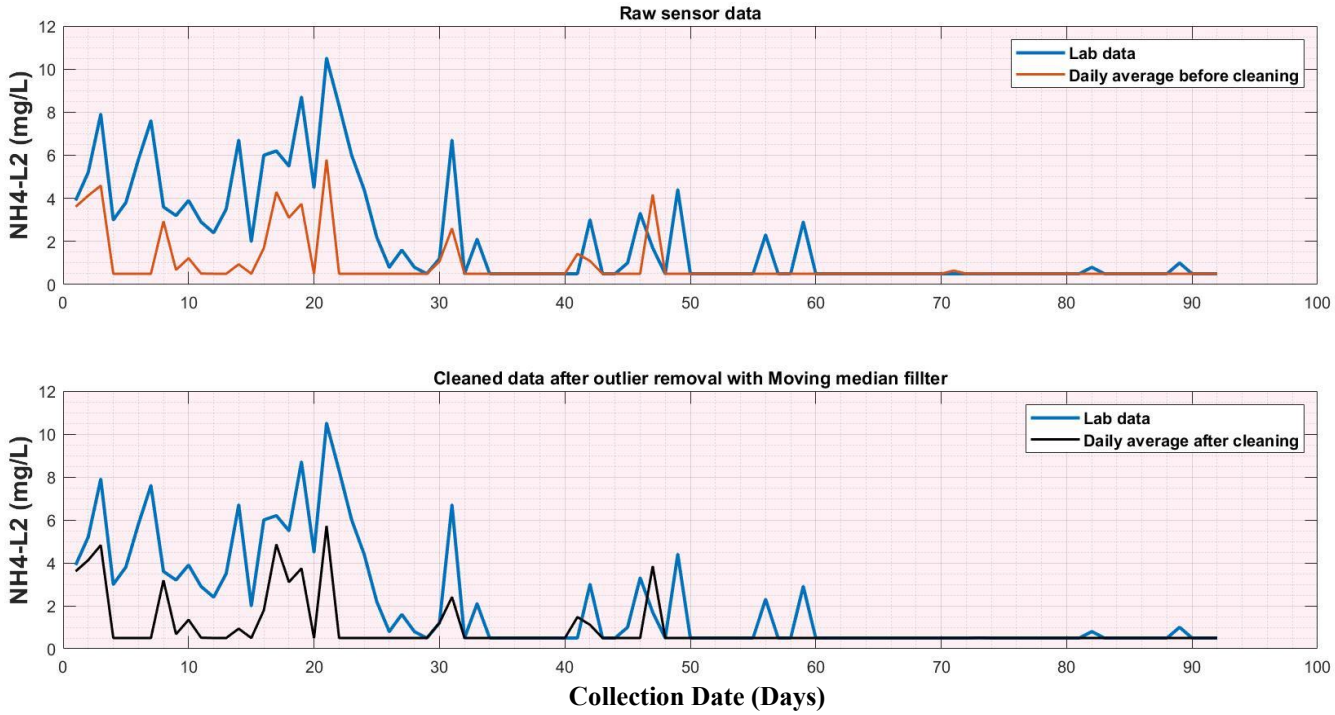

Figure 4.3 Daily average of raw and cleaned data vs lab data

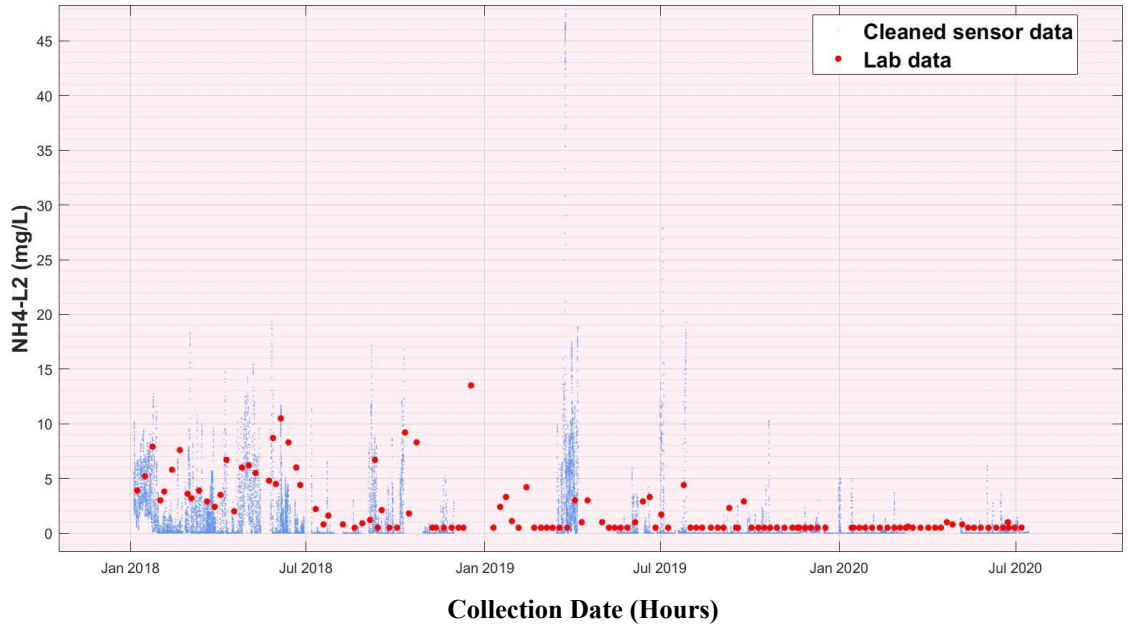

Figure 4.4 Cleaned sensor data by moving median filter vs lab data

Meanwhile, the T-squared with 4000 data interval shows a slightly better result than the rolling median filter with 3.05% BIAS improvement but it shows a daily average of sensor data slightly different from the rolling median filter which is away from the lab data (Figure 4.5).

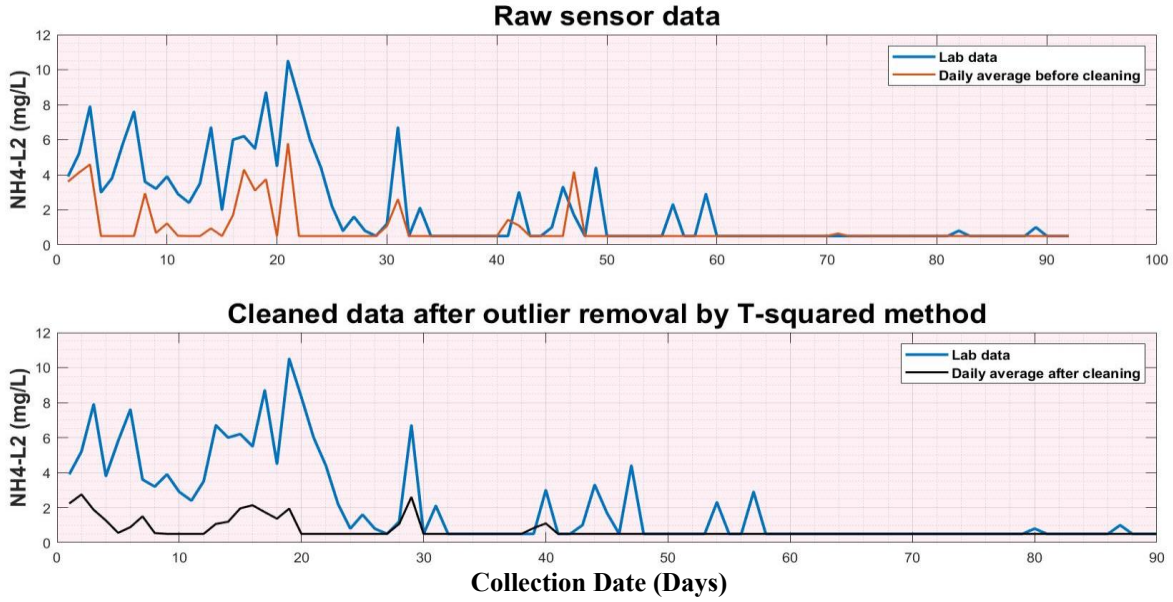

Figure 4.5 Daily average of raw and cleaned data vs lab data after applying of T-squared method

The T-Squared method does not follow the fluctuation of the graph but as it is explained in chapter 3, it has a specific radius of alpha, and any data out of the circle is considered as an outlier (Figure 4.6).

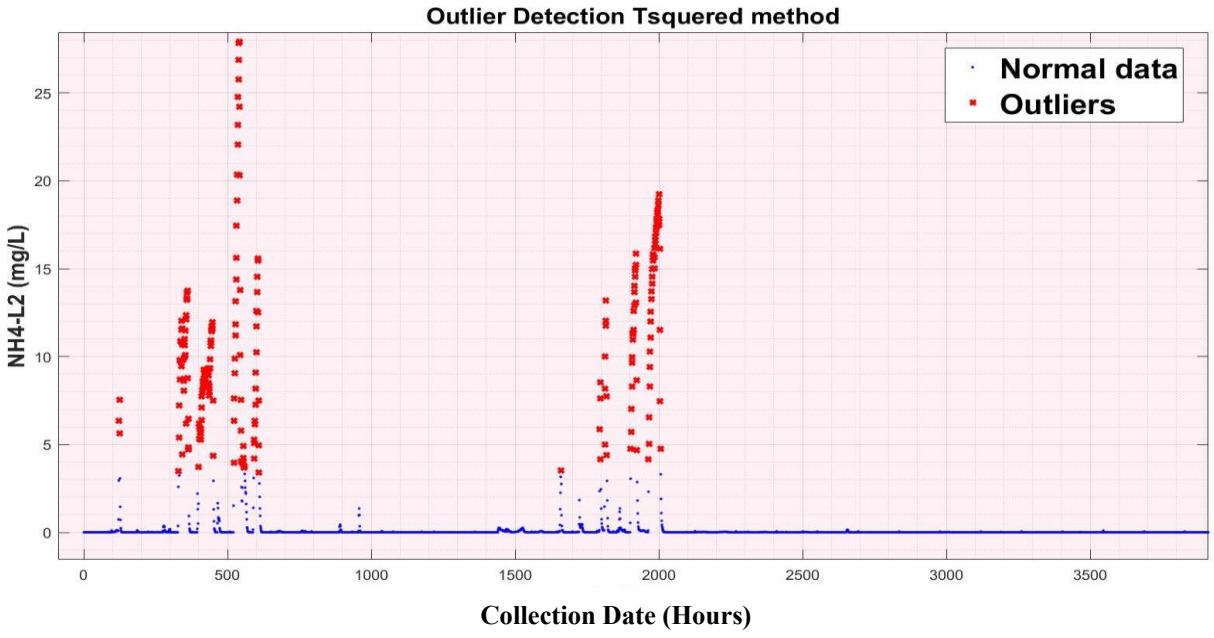

Figure 4.6 Outliers detected by T-squared method

The PCA (T-squared) method mostly removed the highest values in the dataset which resulted in the reduction of daily average and a better bias result than moving median filter.

4.2 N-NOx sensor data cleaning

The data acquired for NNOx is from Nitratax sc online analyzer. Nitratax sc is a process sensor for continuous measurement of nitrate in water. The measurement method uses ultraviolet (UV) light absorption below 250 nm. The probe has a two-beam absorption photometer that compensates interferences by turbidity and organic matter. Detection limit and accuracy for standard solutions are 0.1 mg/L NO2+3-N and ± 3 % of measured value +0.5 mg/L respectively.

NNOx probe sensor recorded data slightly better than NH4 with few peaks and no extremely high or low values as shown in figure 3.5. The actual relative bias for this probe after preliminary clean-up is 0.61. Moving median displayed a very satisfactory result (Figure 4.7) in outlier detection although it only improved the bias accuracy by 1.752% (Cleaned data bias=0.59).

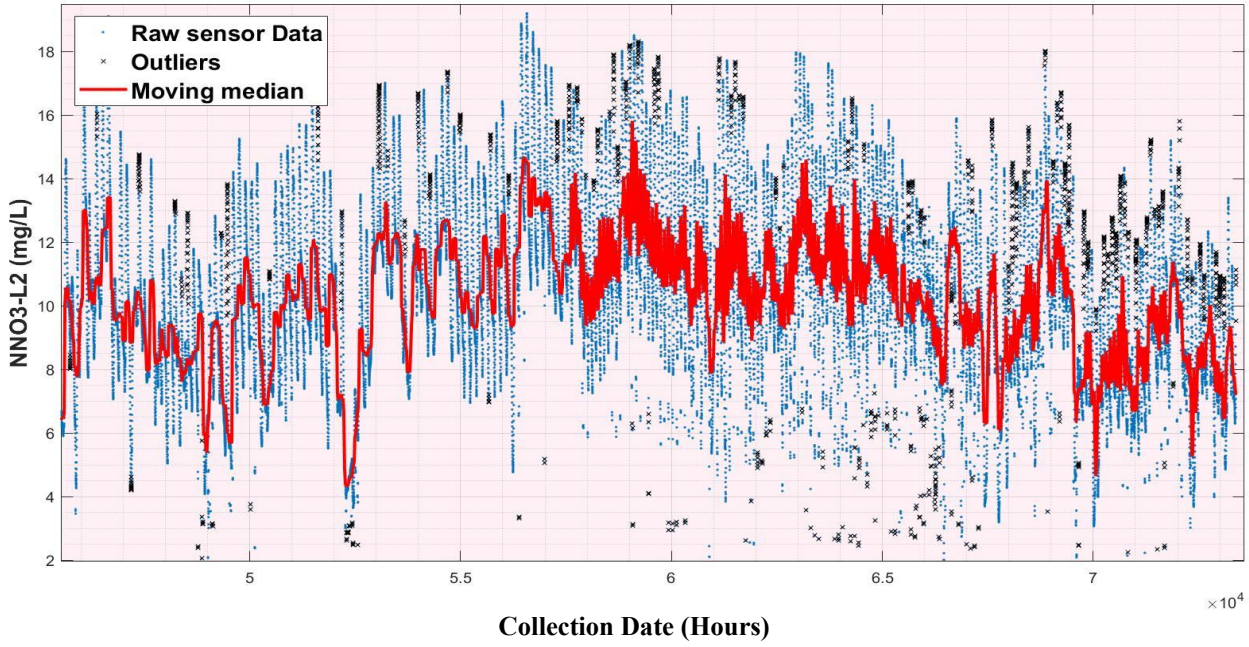

Figure 4.7 Moving median outlier detection on NNOx sensor raw data (window size 135, moving median limit 2.5)

The effect of moving median in the removal of outliers can be seen in Figure 4.8 that the NNO3 concentration range reduced from 50 mg/L to 25 mg/L which all data away from the median is removed. The daily average of the raw and cleaned sensor data is compared with lab data in Figure 4.9, which in some points it showed a closer trend to lab measurement. Although outlier removing techniques works well in outlier detection, after the daily average of the sensor

data, the effect of outlier removal is very small, only in cases where a very high value is existing in recorded data.

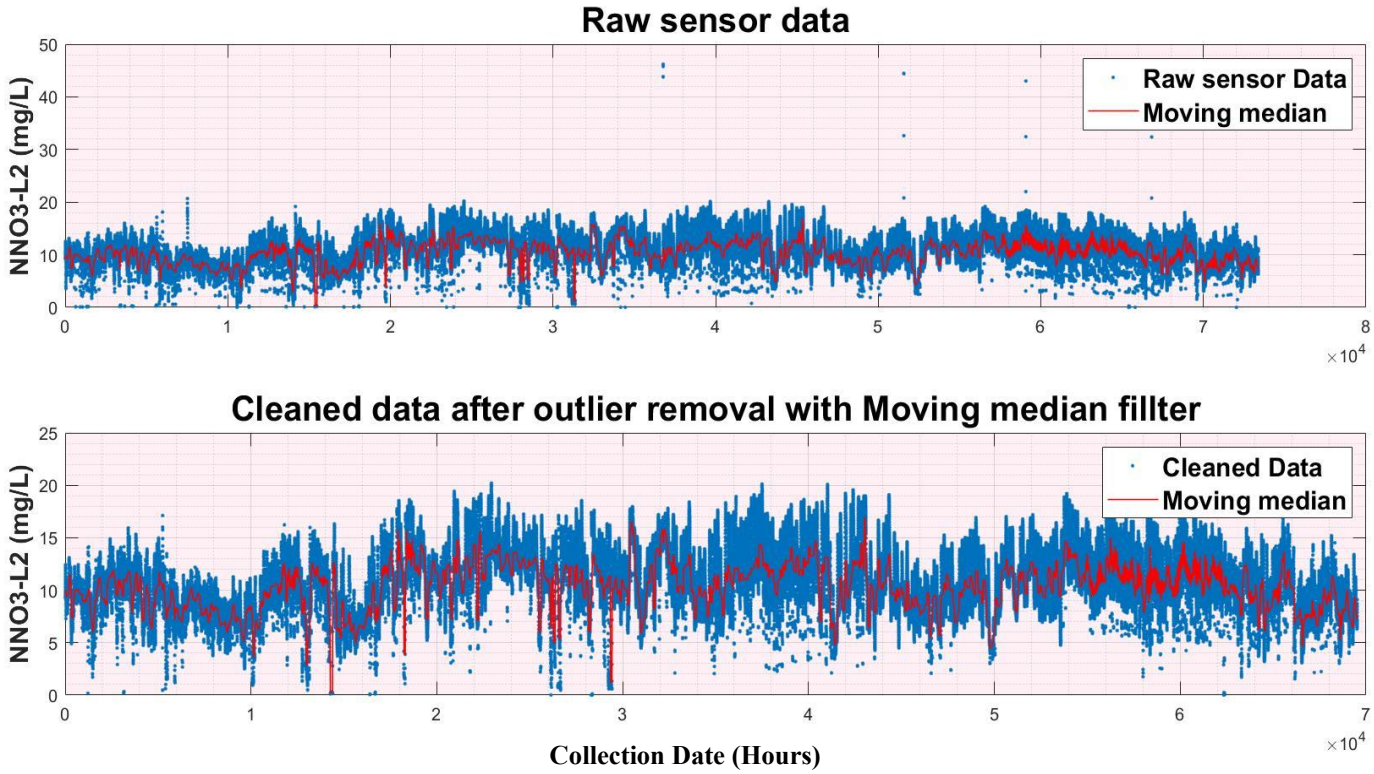

Figure 4.8 Moving median with raw and cleaned data NNOx

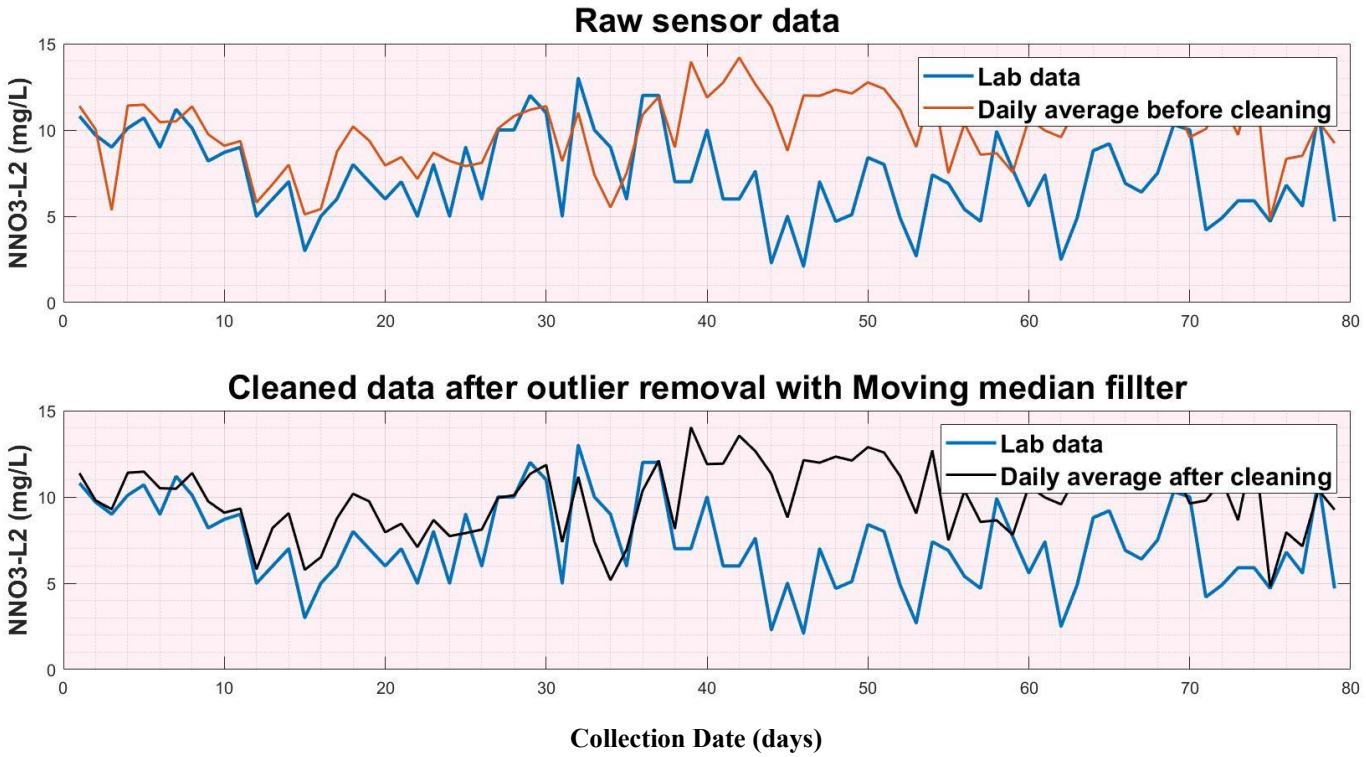

Figure 4.9 Daily average of raw and cleaned data vs lab data

Although the NNO3 recorded data has an acceptable daily average compared to the lab data, the moving median filter can reduce bias more and more if applying more trials in finding a better fit than the current window size and moving median limit. In both cleaned and uncleaned sensor data, the second half of the graph shows that recorded data mostly located above the laboratory measurements (Figure 4.10). The daily average obtained after PCA is quite similar to Figure 4.9.

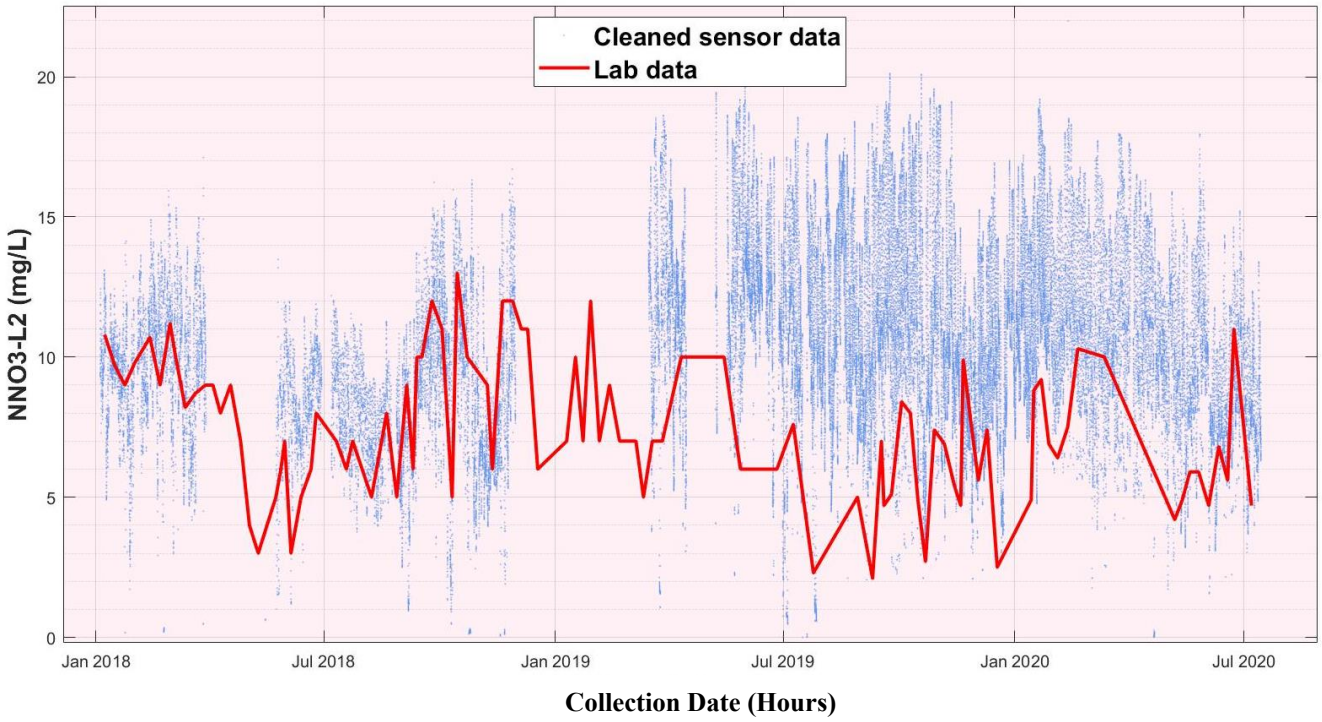

Figure 4.10 Cleaned sensor data by moving median filter vs lab data

In the second part of outlier detection, the PCA application showed a slightly better result in data cleaning of the NNOx probe with 2.164% bias accuracy improvement. The relative bias decrease from 0.61 to 0.5968. the outliers were detected in both very high and very low recorded values (Figure 4.11).

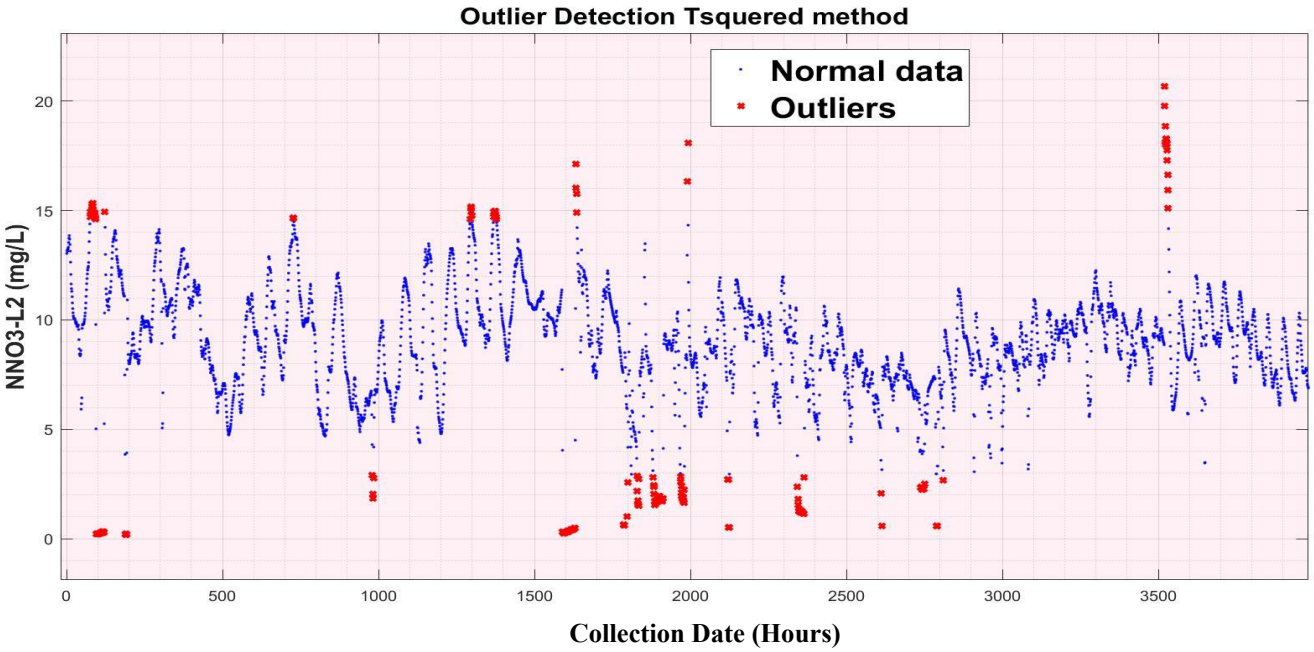

Figure 4.11 Outliers detected by T-squared method

4.3 PPO4 sensor data cleaning

The data acquired for PPO4 is from the Phosphax sc sensor which is an analyzer for the online measurement of ortho-phosphate in water. The measurement principle is based on Molybdovanadate yellow colorimetric method. The analyzer performs a zero-point calibration automatically without the use of a standard solution. Accuracy is maintained by compensating for the background color of the sample at the beginning of every measurement cycle. The detection limit is 0.05 mg/L PO4-P and the accuracy is around ± 2 % ± 0.05 mg/L, with standard solutions.

The raw data obtained from the sensor is quite close to the lab data with few inaccuracies. The relative bias from raw data is 0.996 which shows a quite low performance of the sensor than NH4 and NNOx. After the preliminary cleaning of the data and applying the moving filter the relative bias is reduced to 0.987 with a 0.8% improvement in bias accuracy.

The data set contains very low records (0,004 more than a thousand values) which highly affect the daily average. Moreover, on those dates that all the recorded data are above or below the lab data, it is reasonable that the daily average of the raw data is not getting closer to the lab data and the effect of outliers is negligible as it is shown in Figure 4.14. for example, all recorded sensor data from July 2019 till January 2020, are above the lab data although many outliers were removed (Figure 4.15). The moving median showed a low influence on the data

therefore after the removal of outliers the trend of the data in both cleaned and raw data looks similar (Figure 4.13) and there is no difference in daily average compared to the lab data (Figure 4.14).

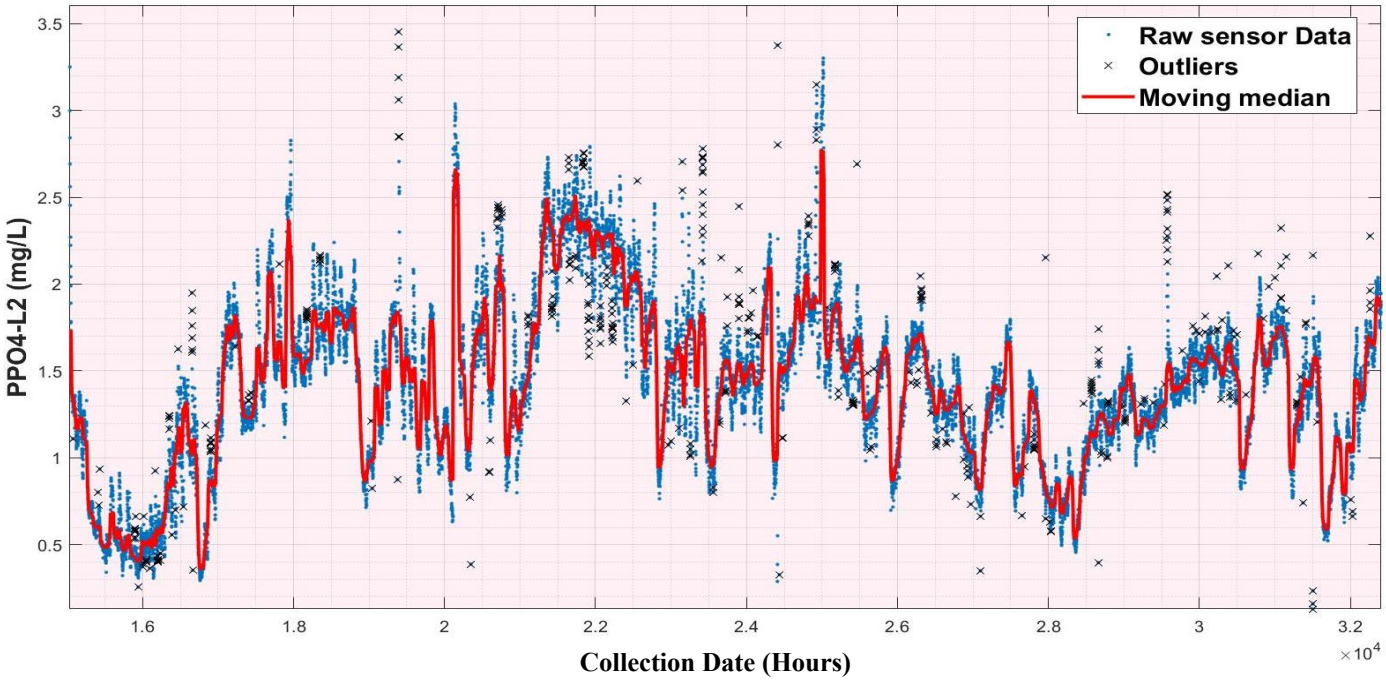

Figure 4.12 moving median filter (red line)on PPO4 raw data with. Outliers (Black crosses), normal data (blue dots). (window size 135, moving median limit 2.5)

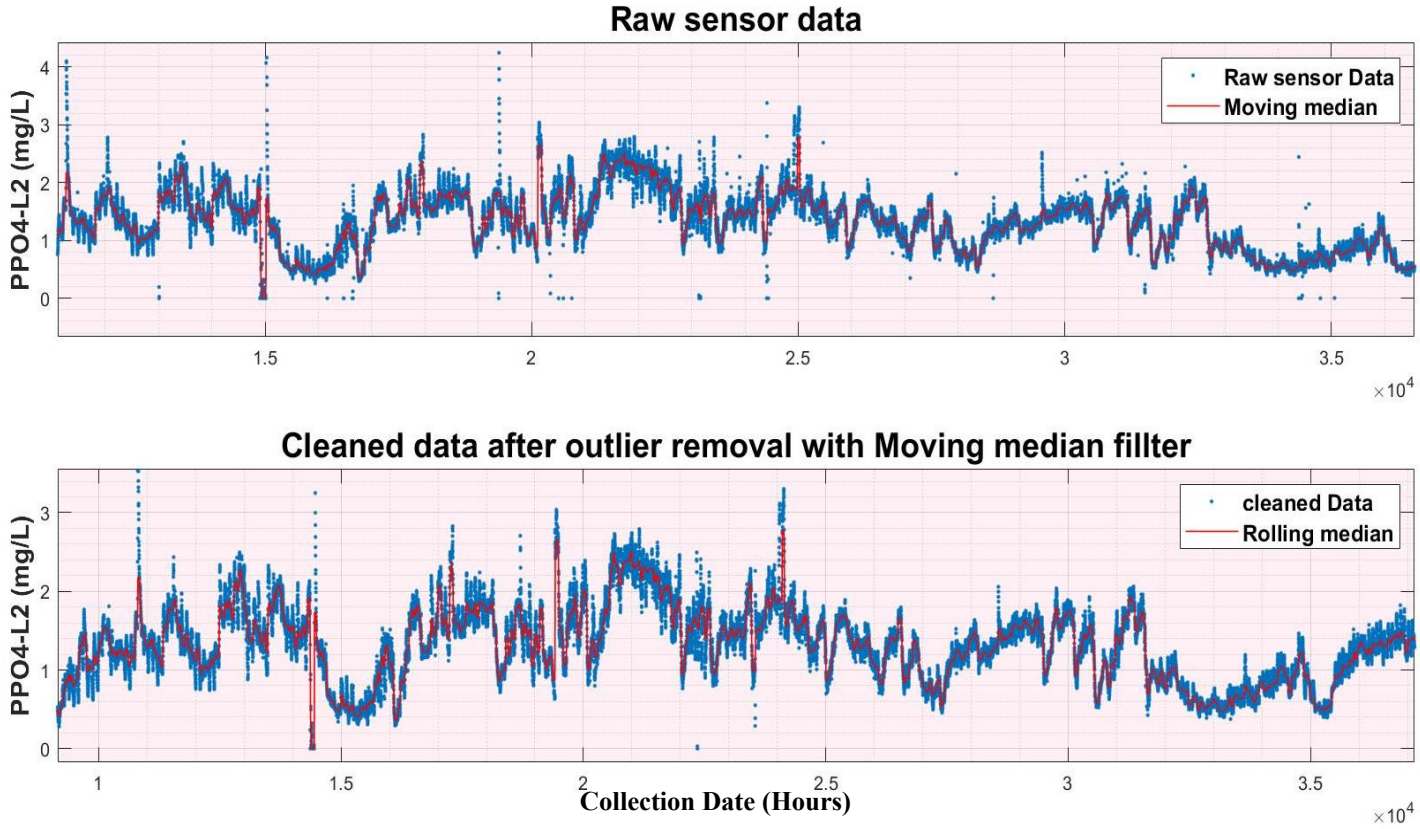

Figure 4.13 Moving median with raw and cleaned data PPO4

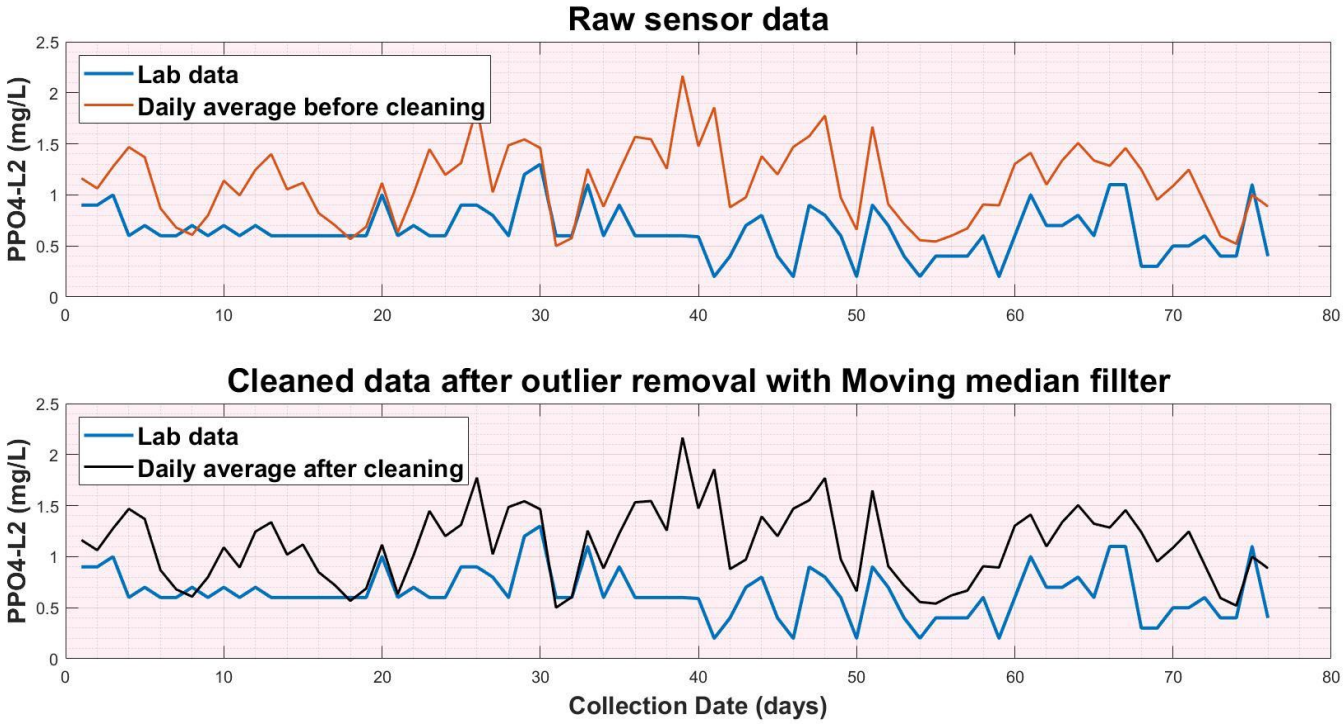

Figure 4.14 Daily average of raw and cleaned data vs lab data

In the following Figure 4.15) there are unrecorded data in April and December of 2018 and January and February of 2019 which may influenced the upcoming recorded data and are mostly recorded above the lab measurements (may be calibration error).

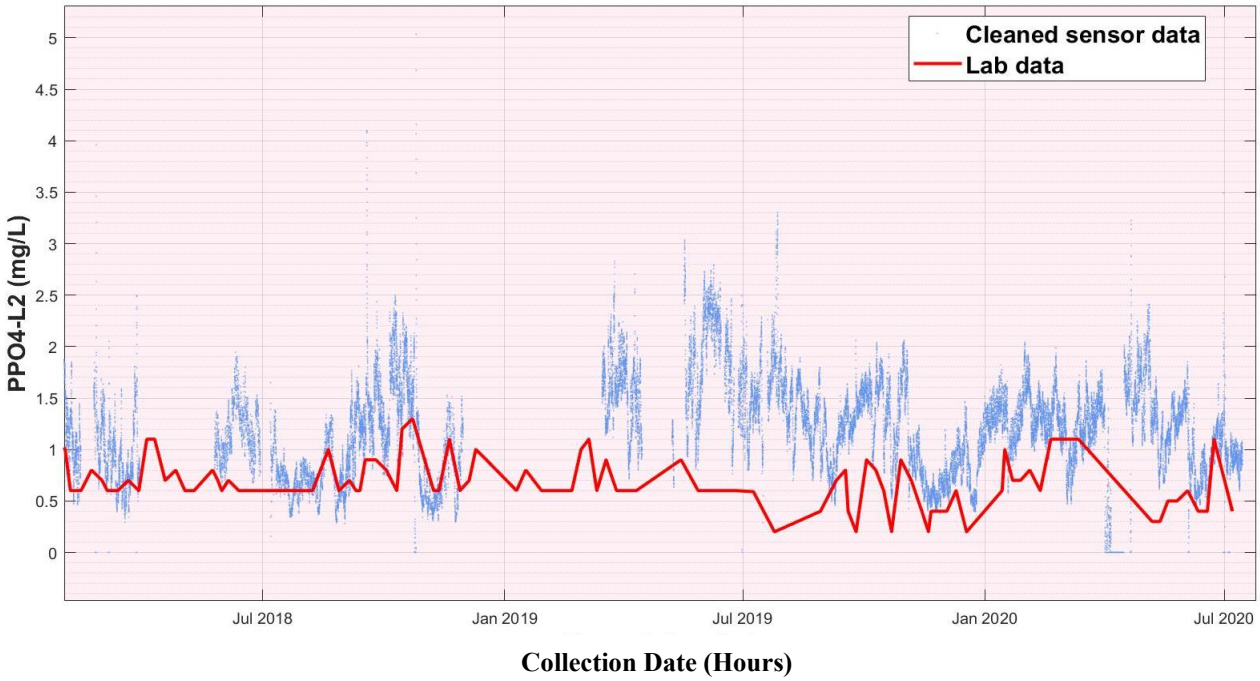

Figure 4.15 Cleaned sensor data by moving median filter vs lab data

The **T-squared method** displayed a slightly better result than the moving median filter due to the removal of a large number of upper points (Figure 4.16). The outliers which are detect caused a better relative bias of 0.9855 compared to the moving median and 1.01% bias improvement. The daily average obtained from cleaned data based on PCA is quite similar to the applied moving median filter. The PCA applied on PPO4 with the interval of 4000 data with a random value of alpha for each interval. The result can be improved by more trials and error method on the selection of alpha to find the best fit alpha for each interval. Meanwhile, the size of the intervals plays a key role, therefore lower interval size leads to a more accurate result.

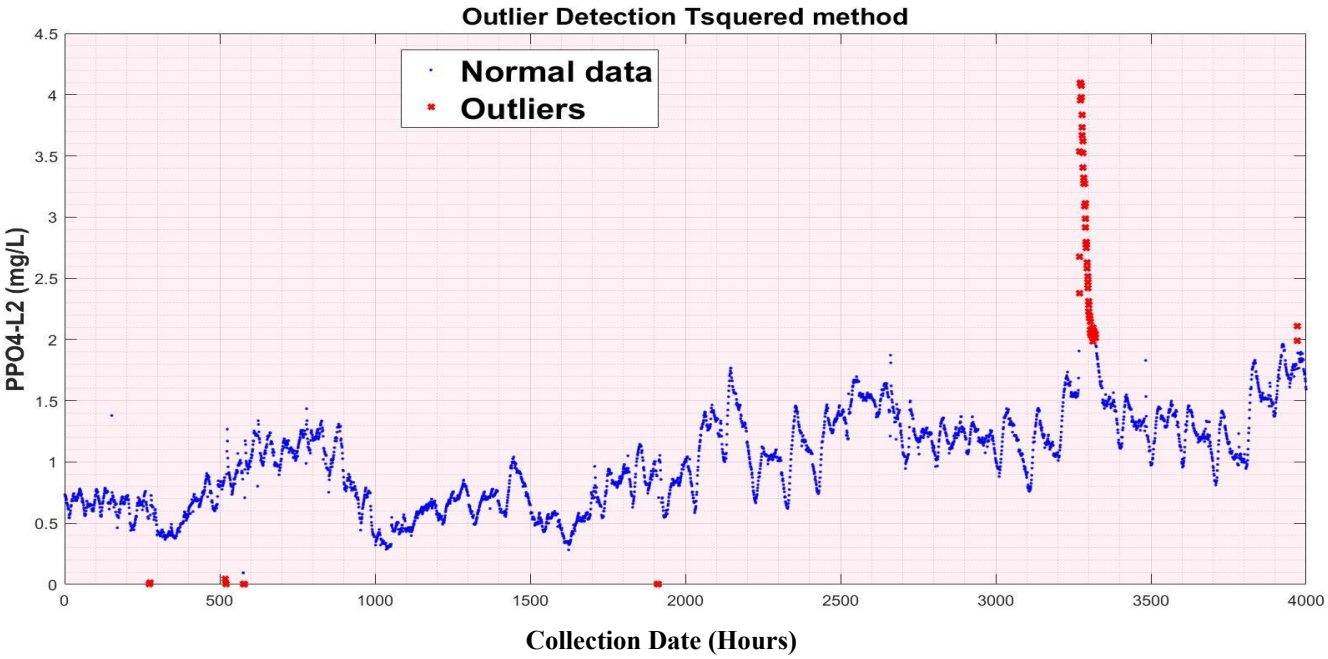

Figure 4.16 Outliers detected by T-squared method (4000 intervals)

4.4 TSS sensor data cleaning

The TSS dataset is obtained from the Solitax ts-line sc probe which is installed in WWTP for continuous monitoring and measurement of turbidity and suspended solids. The measuring principle consists of a dual-beam infrared/scattered light photometer and the measurement is in accordance with DIN 38414). The measuring range for TSS is 0.001 - 50 g/L (i.e. 0.001 - 50,000 mg/L).

TSS probe data contains extremely high values which affect the daily mean therefore The raw data relative bias is a very big relative bias of 50.2, which means completely inappropriate data. The moving median filter displayed acceptable result in outlier detection and in Figure 4.17 the redline (i.e. moving median filter line) follow a very smooth trend and

mostly all points above the line are detected as outliers. After applying the moving median filter the relative bias reduced by 71.37 % to 14.4, although many of the normal data are above the laboratory maximum measurement that is why the relative bias is still far away from zero.

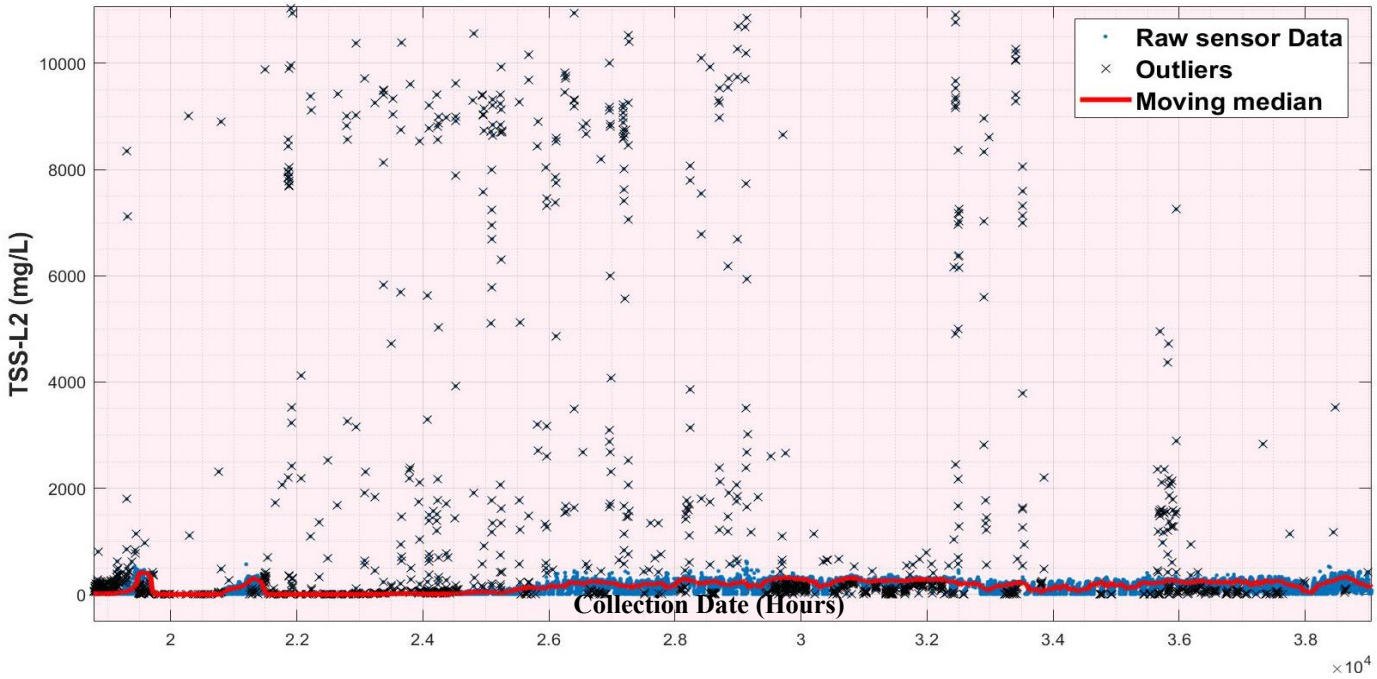

Figure 4.17 moving median filter (red line)on TSS raw data with. Outliers (Black crosses), normal data (blue dots). (window size 245, moving median limit 2.5)

Figure 4.18 gives a better visualization of the effect of moving median on the data, the y axis (TSS concentration) dropped dramatically and in cleaned data, the data fluctuates between the range of 0 to 700. The daily average of the raw data contains high peaks and fluctuations and in the cleaned data it can be seen that the last two peaks are removed (Figure 4.19).

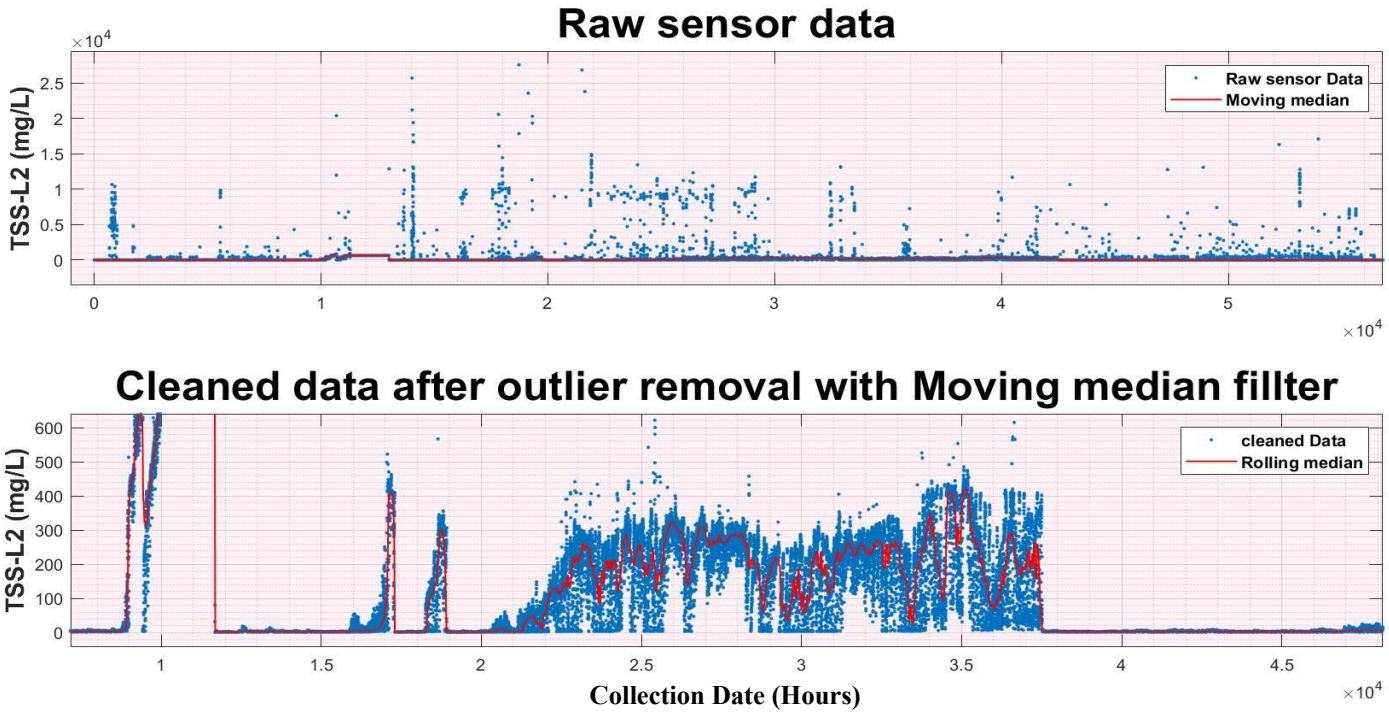

Figure 4.18 Moving median with raw and cleaned data TSS

Figure 4.19 Daily average of raw and cleaned data vs lab data (TSS)

In Figure 4.20 can be seen that still a huge amount of the data are located above the 100 mg/L, which can be removed by more trial and error method on moving median filter especial by reducing the moving median limit and obtain better relative bias.

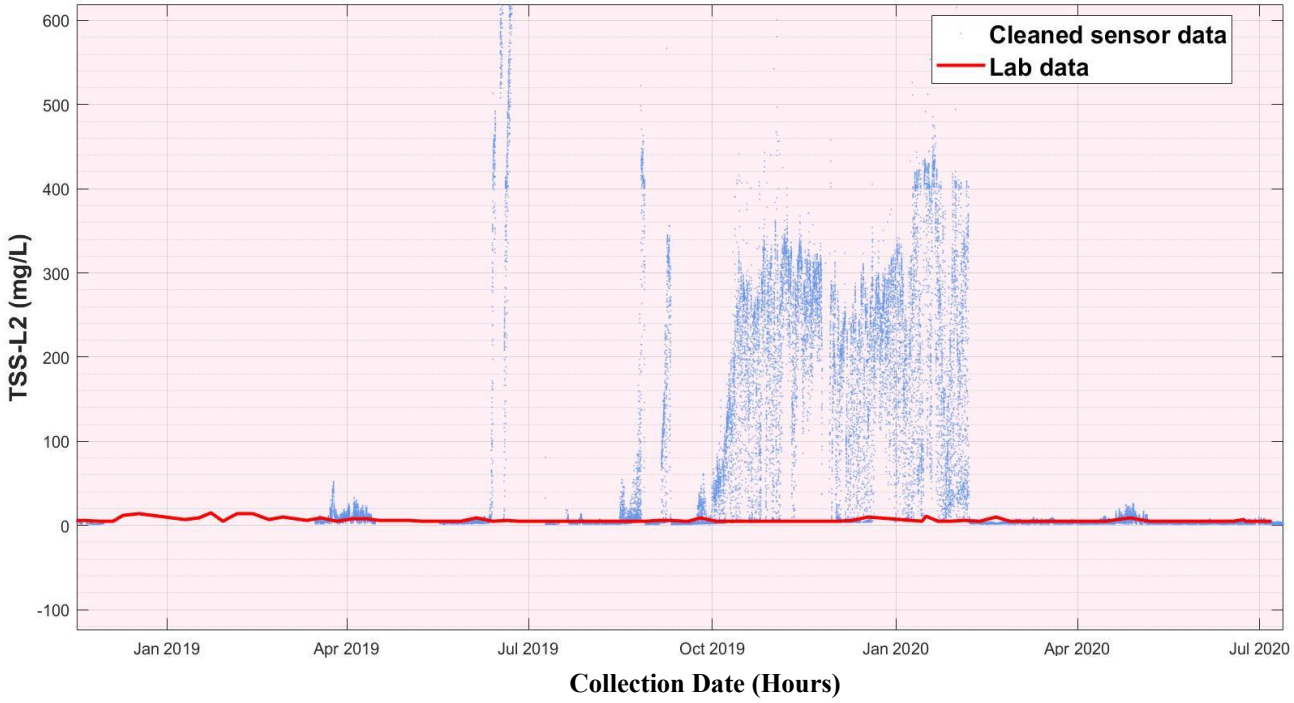

Figure 4.20 Cleaned sensor data by moving median filter and lab data (blue dots)

Although the moving median showed a good result on the TSS probe for outlier detection and the sample median (not rolling median) and MAD (median absolute deviation) is tried to check the result. Despite other compounds (showed an increase in relative bias), the TSS probe displayed twice better results than the moving median filter with the relative bias of 6.03 which is an 87.976% improvement in the relative bias of raw sensor data. The daily average range reduced from 6000 to 200 with few fluctuations in the graph (Figure 4.21). based on this filter we can reach up to 98.5% relative bias improvement although around 30 % of the data will be considered as outliers (with a median limit of 0.1).

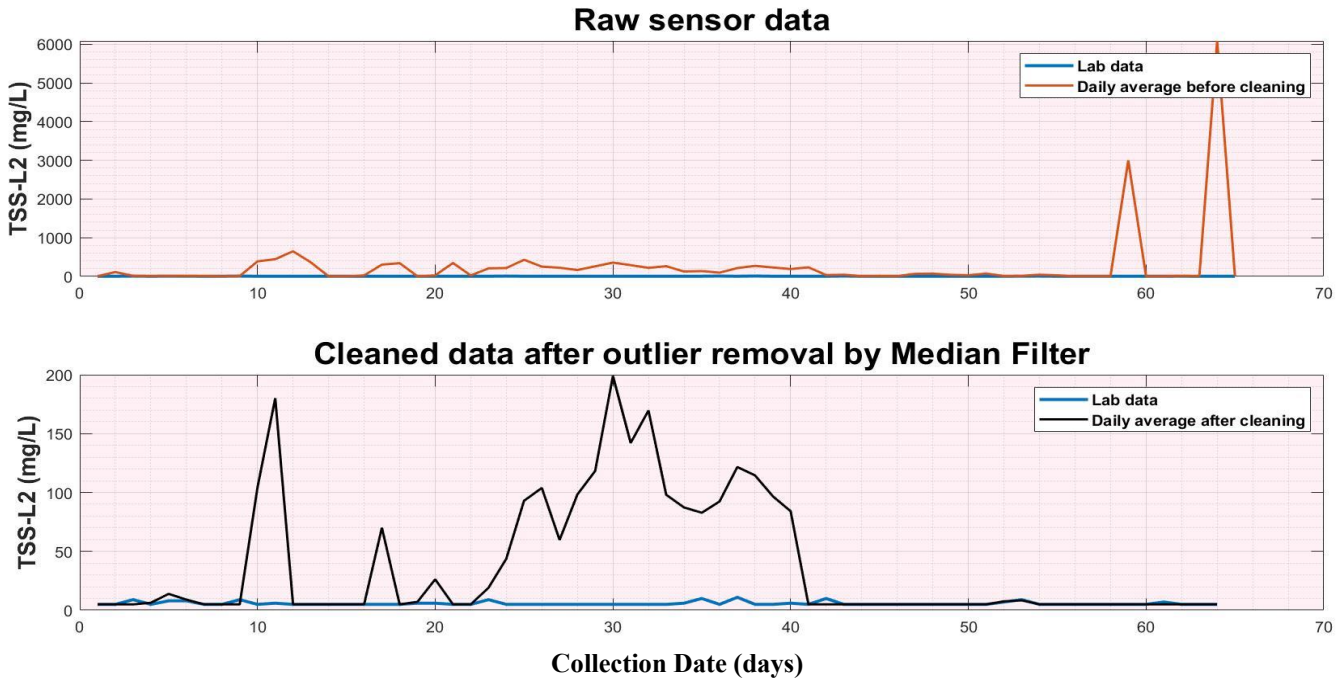

Figure 4.21 Raw and cleaned data daily average by the median filter (median limit 0.5)

The T-squared outlier detection method showed better results for the TSS probe the same as for the other compounds. After application of the T-squared method on the TSS probe the relative bias reduced to 14.174 which is a 71.766% bias improvement. For obtaining such a result very low alpha values were selected, although it showed a better result than moving median filter, it captured some low values that may not be real outliers (Figure 4.22). the daily average obtained with this method is quite similar to the moving median filter.

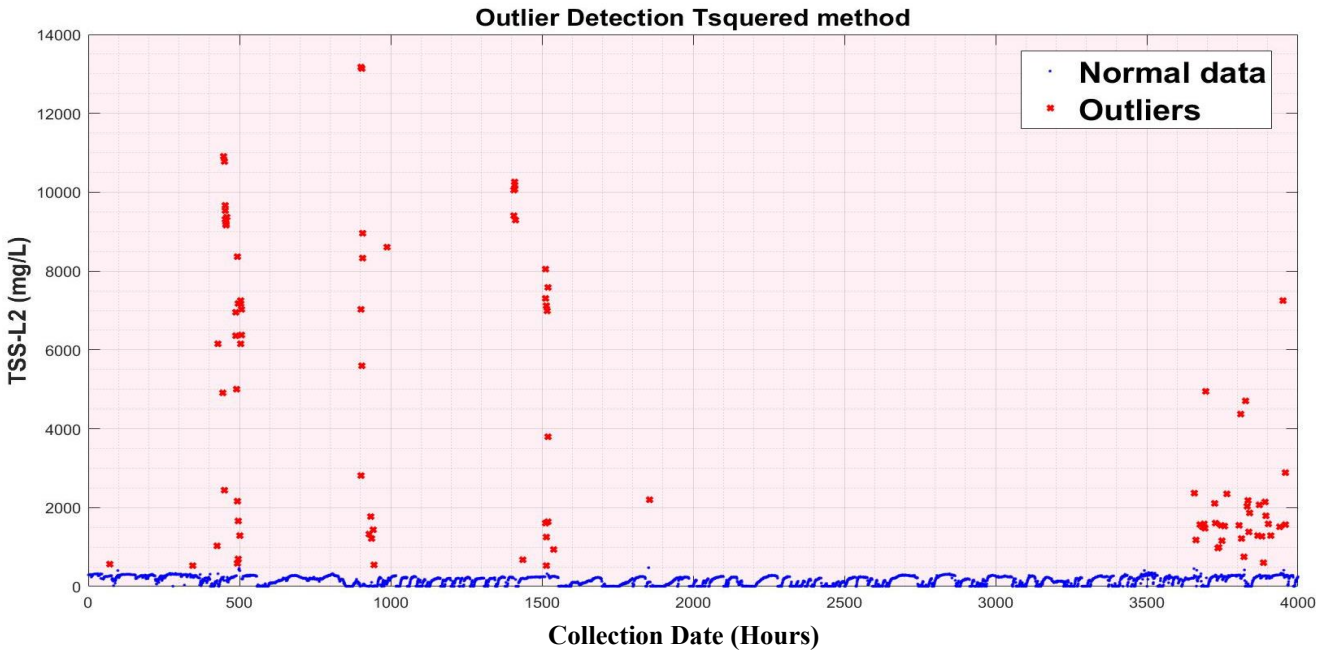

Figure 4.22 Outliers detected by T-squared method (4000 intervals)

4.5 Summary of the results

The application of the Moving median filter and PCA on probe data of the Peschiera Borromeo WWTP for outlier detection showed satisfactory results. The data obtained from probes required a preliminary clean up (discussed in chapter 3) to have an accurate result and avoid errors from the programming language MATLAB.

The number of outliers detected by PCA and moving median filter in order to achieve satisfactory result was quite different for each parameter. in the following table the amount of outliers detected by each method is presented:

Table 4-1 Amount of outliers detected by each method

Description \ Compounds		NH4_L2	NNOx_L2	PPO4_L2	TSS_L2
Total sensors data after preliminary clean-up		61018	73382	54991	65036
Total data detected as Outliers	**T-square Method**	7415	2108	3934	4322
	Modified z-score method	3894	3854	1535	7658
Percentage of Outliers detected	**T-square Method**	12%	3%	7%	7%
	Modified z-score method	6%	5%	3%	12%

After preliminary clean-up, the moving median filter was applied on each probe separately, in order to find the best fit and obtained better results different window size and moving median limit is tested. Meanwhile, PCA based on the T-squared method was tested on the data and different alpha values were selected for each interval of data to maximize the outlier detection process. For comparison of the methods, the mean relative bias based on laboratory measurements was calculated (equation 12). In order to check the performance of each method, the relative bias improvement was calculated based on equation (13).

It can be observed that both methods reach paragonable results for all the probes analyzed, with the T-squared method fitting better for all sensors.

Best performances, in terms of increased bias accuracy, were obtained from TSS sensor cleaning, since they were the ones most affected by elevated outliers.

The overall result for each method is presented in the following table:

Table 4-2: Mean relative bias of raw data (with outliers), data cleaned by the moving median filter, and data cleaned by PCA method.

	RAW DATA	MOVING MEDIAN FILTER		T-SQUARED METHOD	
	Mean Bias	Mean Bias	Increased Bias accuracy	Mean Bias	Increased Bias accuracy
NH_4	0.3127	0.3052	2.419%	0.3032	3.0432%
NOx	0.6100	0.5994	1.752%	0.5968	2.1636%
PO_4	0.9956	0.9877	0.80%	0.9855	1.0102%
TSS	50.1994	14.368	71.378%	14.1735	71.7656%

Chapter 5 Conclusions and discussion

In this thesis, real-time data from different sensors were analyzed in order to evaluate their applicability in real-time control systems for health risk minimization. Particularly, in this study, probe data from sensors were checked for outliers to provide real-time data to Early Warning Systems (DWC project), which will be managed to assure health-risk control in Peschiera Borromeo WWTP, the monitoring networks aim to reduce the health risks related to microbiological contamination in bathing water sites.

The water management sector usually relies on treatment processes and removal efficiencies, using a huge number of sensors and meters, alarms, and automatic control tools. In recent years, technological progress allowed the digitalization of the water sector providing new sensors, always more precise and reliable, and tools for decision support. For effective utilization of sensor data in WWTP operation, procedures and standard protocol need to be developed for the acquisition of reliable data.

In this work, real-time measurements of TSS, NH_4, PO_4, and NO_3 concentrations were processed mathematically to detect outliers in the measurements, and long periods in which the sensor measurements are outside the expected operating range of the system. Two cleaning methods, moving median filter based on modified z- score and PCA, were applied and deeply discussed for outlier detection and removal. The cleaning procedure based on PCA (T-squared) showed better results in the identification of outliers than the Moving median filter. The T-squared method was based on the value of alpha while moving median considered two parameters moving median limit and window size.

To compare the accuracy of the two methods, a relative bias was calculated considering laboratory data measurements, which were considered as true values. It was noted that both the applied cleaning procedure slightly reduced the relative bias of data compared to the raw (uncleaned) data

Both methods showed satisfactory results on data which had sudden peaks and high recorded values, PCA showed significantly better result on detection of continuously recorded data (clustered outliers) while moving median filter could detect both high and low trends. The amount of outliers detected by PCA is higher than the moving filter method. Since the T-squared method is based on the applied radius for the dataset, the effect of the alpha (radius) and the selected interval is highly important and it is possible to obtain even better results if the optimization algorithms will be used for future data mining on this project.

Moving median filter showed acceptable results close to the PCA with the fewer number of outliers detected. The moving median limit plays a significant role in the number of the outliers, Different window sizes and moving median limits were checked to find the best fit where minimum bias is obtained with the maximum number of outliers.

To improve the accuracy and result of the models some recommendations are as follow:

Further recommendations:

- In order to further improve the accuracy of the cleaned data compared to lab data (i.e. reduce the Bias), it is recommended to substitute all the measured probe data lower than the lab detection limit with the value of the lab threshold.
- to the lab threshold.
- Remove any data which is extremely higher than maximum lab measurement.
- Remove all equal recorded values fixed for one or more than one day (constant recorded values for long period).
- Use of optimization algorithm for moving median filter to find the best median limit and window size with maximum real outliers.
- Use of different data intervals for the T-squared method (interval of 4000 data is used in the project) and apply optimization algorithm to find the best alpha for each interval to reduce the bias.

References

Aguado, D., & Rosen, C. (2008). Multivariate statistical monitoring of continuous wastewater treatment plants. *Engineering Applications of Artificial Intelligence*, 12.

Ansari, A. A., Singh, G. S., Lanza, G. R., & Rast, W. (2010). *Eutrophication: Causes, consequences and control* (Vol. 1). Springer.

Asadi, A., Verma, A., Yang, K., & Mejabi, B. (2017). Wastewater treatment aeration process optimization: A data mining approach. *Journal of Environmental Management*, *203*, 630–639. https://doi.org/10.1016/j.jenvman.2016.07.047

Avella, A. C., Görner, T., Yvon, J., Chappe, P., Guinot-Thomas, P., & de Donato, Ph. (2011). A combined approach for a better understanding of wastewater treatment plants operation: Statistical analysis of monitoring database and sludge physico-chemical characterization. *Water Research*, *45*(3), 981–992. https://doi.org/10.1016/j.watres.2010.09.028

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, *17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.

Corominas, Ll., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, *106*, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023

Ding, N., Gao, H., Bu, H., Ma, H., & Si, H. (2018). Multivariate-Time-Series-Driven Real-time Anomaly Detection Based on Bayesian Network. *Sensors*, *18*(10), 3367. https://doi.org/10.3390/s18103367

Fazai, R., Mansouri, M., Abodayeh, K., Nounou, H., & Nounou, M. (2019). Online reduced kernel PLS combined with GLRT for fault detection in chemical systems. *Process Safety and Environmental Protection*, *128*, 228–243. https://doi.org/10.1016/j.psep.2019.05.018

Fuente, M. J., Garcia-Alvarez, D., Sainz-Palmero, G. I., & Vega, P. (2012). Fault detection in a wastewater treatment plant based on neural networks and PCA. *2012 20th Mediterranean Conference on Control & Automation (MED)*, 758–763. https://doi.org/10.1109/MED.2012.6265729

Garcia-Alvarez, D., Fuente, M. J., Vega, P., & Sainz, G. (2009). Fault Detection and Diagnosis using Multivariate Statistical Techniques in a Wastewater Treatment Plant.* *This work was supported in part by the national research agency of Spain (CICYT) through the project DPI2006-15716-C02-02 and the regional government of Castilla y Leon through the project VA052A07. *IFAC Proceedings Volumes*, *42*(11), 952–957. https://doi.org/10.3182/20090712-4-TR-2008.00156

Garcia-Alvarez, Diego. (n.d.). *FAULT DETECTION USING PRINCIPAL COMPONENT ANALYSIS (PCA) IN A WASTEWATER TREATMENT PLANT (WWTP)*. 7.

Garcia-Alvarez, Diego. (2009). *FAULT DETECTION USING PRINCIPAL COMPONENT ANALYSIS (PCA) IN A WASTEWATER TREATMENT PLANT (WWTP)*.

Genovesi, A., Harmand, J., & Steyer, J.-P. (2000). Integrated Fault Detection and Isolation: Application to a Winery's Wastewater Treatment Plant. *Appl. Intell.*, *13*, 59–76. https://doi.org/10.1023/A:1008379329794

Güçlü, D., & Dursun, Ş. (2010). Artificial neural network modelling of a large-scale wastewater treatment plant operation. *Bioprocess and Biosystems Engineering*, *33*(9), 1051–1058. https://doi.org/10.1007/s00449-010-0430-x

Häggblom, K.-E. (n.d.). *Basics of Multivariate Modelling and Data Analysis*. Abo Akademi University. Retrieved January 17, 2021, from https://www.users.abo.fi/khaggblo/MMDA/MMDA6.pdf

Haimi, H., Mulas, M., Corona, F., & Vahala, R. (2013). Data-derived soft-sensors for biological wastewater treatment plants: An overview. *Environmental Modelling & Software*, *47*, 88–107. https://doi.org/10.1016/j.envsoft.2013.05.009

Hand, D. J., & Adams, N. M. (2015). Data Mining. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (pp. 1–7). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06466.pub2

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). Springer.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis* (Vol. 3). Wiley New York.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. ASQC Quality Press.

Important z-scores. (n.d.). Retrieved January 19, 2021, from http://www.math.uni.edu/~campbell/stat/normfact.html

Ivanushkin, M. A., Volgin, S. S., Kaurov, I. V., & Tkachenko, I. S. (2019). Analysis of statistical methods for outlier detection in telemetry data arrays, obtained from "AIST" small satellites. *Journal of Physics: Conference Series*, *1326*, 012029. https://doi.org/10.1088/1742-6596/1326/1/012029

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jayakumar, G. D. S., & Thomas, B. J. (2013). A new procedure of clustering based on multivariate outlier detection. *Journal of Data Science*, *11*(1), 69–84.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. *VLDB*, *98*, 392–403.

Lee, J.-M., Yoo, C., Choi, S. W., Vanrolleghem, P. A., & Lee, I.-B. (2004). Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, *59*(1), 223–234. https://doi.org/10.1016/j.ces.2003.09.012

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). *Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median*. 3.

Liu, H., Yang, J., Zhang, Y., & Yang, C. (2021). Monitoring of wastewater treatment processes using dynamic concurrent kernel partial least squares. *Process Safety and Environmental Protection*, *147*, 274–282. https://doi.org/10.1016/j.psep.2020.09.034

Liu, Y., Xiao, H., Pan, Y., Huang, D., & Wang, Q. (2016). Development of multiple-step soft-sensors using a Gaussian process model with application for fault prognosis. *Chemometrics and Intelligent Laboratory Systems*, *157*, 85–95. https://doi.org/10.1016/j.chemolab.2016.07.002

Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: A data cleaning application. *Proceedings of KDNet Symposium on Knowledge-Based Systems for the Public Sector*.

Median absolute deviation. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Median_absolute_deviation&oldid=993308047

Nasr, M. S., Moustafa, M. A. E., Seif, H. A. E., & El Kobrosy, G. (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. *Alexandria Engineering Journal*, *51*(1), 37–43. https://doi.org/10.1016/j.aej.2012.07.005

Newhart, K. B., Holloway, R. W., Hering, A. S., & Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water Research*, *157*, 498–513. https://doi.org/10.1016/j.watres.2019.03.030

Normal distribution. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Normal_distribution&oldid=1002951236

Pourzangbar, A., Losada, M. A., Saber, A., Ahari, L. R., Larroudé, P., Vaezi, M., & Brocchini, M. (2017). Prediction of non-breaking wave induced scour depth at the trunk section of breakwaters using Genetic Programming and Artificial Neural Networks. *Coastal Engineering*, *121*, 107–118. https://doi.org/10.1016/j.coastaleng.2016.12.008

Samuelsson, O. (2017). *Fault detection in water resource recovery facilities*. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-329777

Tao, E. P., Shen, W. H., Liu, T. L., & Chen, X. Q. (2013). Fault diagnosis based on PCA for sensors of laboratorial wastewater treatment process. *Chemometrics and Intelligent Laboratory Systems*, *128*, 49–55. https://doi.org/10.1016/j.chemolab.2013.07.012

Teshnehdel, S., Mirnezami, S., Saber, A., Pourzangbar, A., & Olabi, A. G. (2020). Data-driven and numerical approaches to predict thermal comfort in traditional courtyards. *Sustainable Energy Technologies and Assessments*, *37*, 100569. https://doi.org/10.1016/j.seta.2019.100569

Vanrolleghem, P. A., & Lee, D. S. (2003). On-line monitoring equipment for wastewater treatment processes: State of the art. *Water Science and Technology*, *47*(2), 1–34. https://doi.org/10.2166/wst.2003.0074

Wu, L., & Ho, D. W. C. (2009). Fuzzy Filter Design for Itoˆ Stochastic Systems With Application to Sensor Fault Detection. *IEEE TRANSACTIONS ON FUZZY SYSTEMS*, *17*(1), 10.

Xiao, H., Huang, D., Pan, Y., Liu, Y., & Song, K. (2017). Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. *Chemometrics and Intelligent Laboratory Systems*, *161*, 96–107. https://doi.org/10.1016/j.chemolab.2016.12.009

Zhang, J., & Wang, H. (2006). Detecting outlying subspaces for high-dimensional data: The new task, algorithms, and performance. *Knowledge and Information Systems*, *10*(3), 333–355. https://doi.org/10.1007/s10115-006-0020-z

Appendix 1

Window	Median limit	BIAS	Improved Bias	Window	Median limit	BIAS	Improved Bias
14.0	1.5	0.305	2.419%	19.0	3.0	0.312	0.35%
13.0	1.7	0.305	2.40%	12.0	2.1	0.312	0.28%
13.0	2.0	0.305	2.40%	20.0	3.0	0.312	0.22%
13.0	1.5	0.306	2.29%	25.0	3.0	0.312	0.19%
15.0	1.6	0.306	2.22%	10.0	3.7	0.312	0.18%
15.0	1.7	0.306	2.21%	25.0	3.7	0.312	0.15%
14.0	1.7	0.306	2.20%	25.0	3.5	0.312	0.14%
15.0	1.8	0.306	2.11%	25.0	3.6	0.312	0.14%
15.0	1.5	0.306	2.09%	20.0	2.4	0.312	0.12%
15.0	2.2	0.308	1.62%	25.0	3.4	0.312	0.11%
40.0	2.2	0.308	1.62%	11.0	2.2	0.312	0.09%
16.0	3.0	0.308	1.39%	25.0	1.9	0.313	-0.08%
17.0	3.0	0.309	1.30%	25.0	2.4	0.313	-0.08%
13.0	2.1	0.309	1.28%	28.0	3.0	0.314	-0.47%
15.0	3.0	0.309	1.10%	28.0	2.3	0.314	-0.51%

Table A. window size and median limit attempt on NH4 raw data to find the best fit

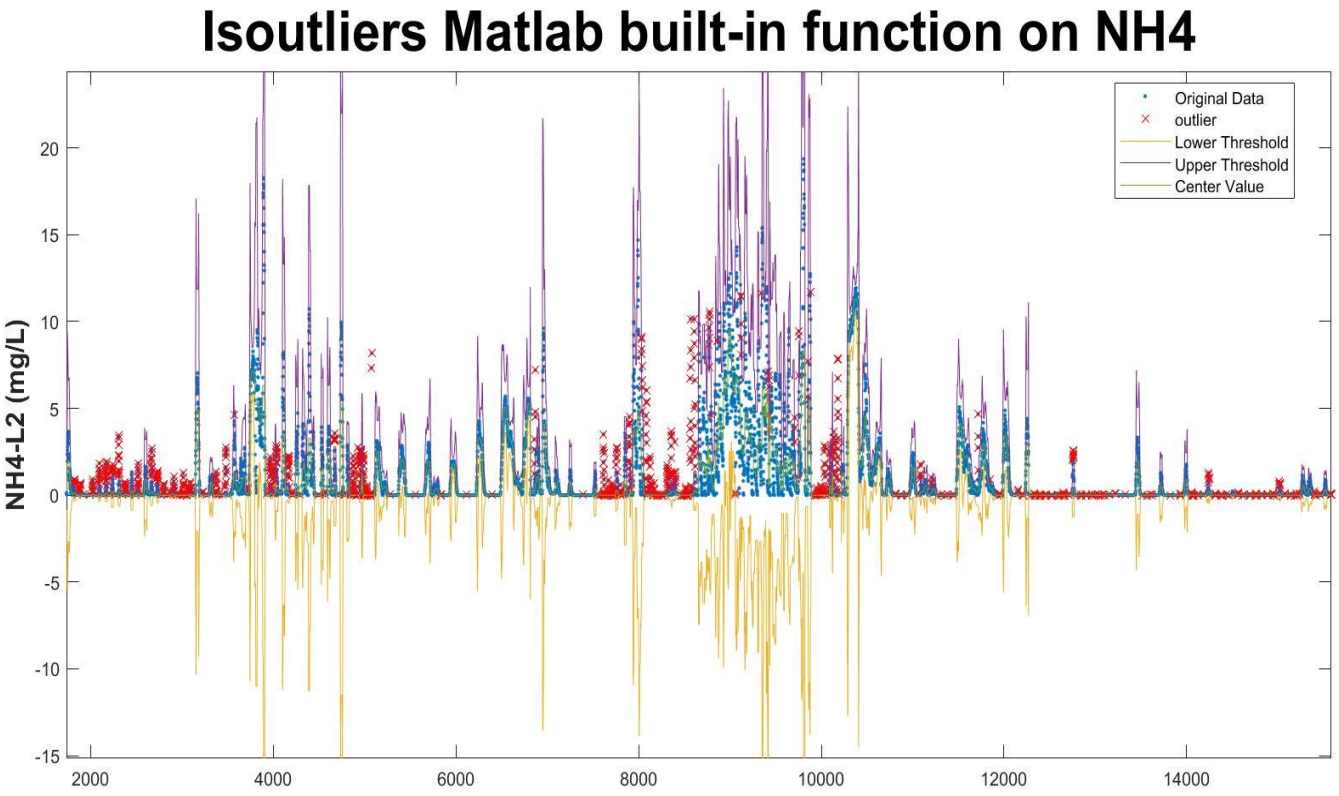

Figure A. isoutliers function to detect outliers

MATLAB CODES:

1. Moving median filter based on modified z- score for NH4 sensor data

```
clc;
clear;
close all;
```

Load data

```
[num,txt,raw] = xlsread('NH4_L2.xlsx');
data = num;

data = data(1:end, 1);
D = 1.5
Window_size = 14

MAD = movmad(data,Window_size);
Median = movmedian(data,Window_size);
M = zeros();
outlier = zeros();

for i=1:length(data)
    M(i) = (0.6745*(data(i)-Median(i)))/MAD(i);
    M = M';
    outlier(i) = abs(M(i))>D;
end
```

```
ind = find(outlier);  % all outlier index
ind1 = find(~outlier);  %all not outlier index
dates_outliers = txt(ind'); %picks the date of outliers
outliers = [ind', data(ind)];
all_normal_data = [txt(ind1'),num2cell(data(ind1))];
```

Daily average of the raw data

```
TT = readtimetable('NH4_L2.xlsx');
DailyAvg_actualdata = retime(TT,'daily','mean');
[numb,text,both] = xlsread('labNH4.xlsx');
lb = string(text);
lt = datetime(lb,'InputFormat','MM/dd/yyyy');
pickdata = [numb(:,1),DailyAvg_actualdata.Var1(DailyAvg_actualdata.Time(lt))];
pickdata(pickdata <0.5) = 0.5;
rows = any(isnan(pickdata),2);
pickdata(rows,:) = [];
```

daily average of the cleaned data

```
b = string(txt(ind1'));
t = datetime(b,'InputFormat','MM/dd/yyyy h:mm:ss a');
```

```
T = timetable(t, data(ind1));
DailyAvg = retime(T,'daily','mean');
pickcleaned_data = [numb(:,1),DailyAvg.Var1(DailyAvg.t(lt))];
pickcleaned_data (pickcleaned_data  <0.5) = 0.5;
c_rows = any(isnan(pickcleaned_data ),2);
pickcleaned_data (c_rows,:) = [];
```

relative BIAS calculation

```
Accuracy = zeros();
for i = 1:length(pickcleaned_data)
    Accuracy (i) = abs(((pickcleaned_data(i,2)-
pickcleaned_data(i,1))/(pickcleaned_data(i,1))));

end
BIAS_cleaned = mean(Accuracy);

Accuracyuncleaned = zeros();
for i = 1:length(pickdata)
    Accuracyuncleaned (i) = abs(((pickdata(i,2)-pickdata(i,1))/(pickdata(i,1))));

end
BIAS_real = mean(Accuracyuncleaned);

% compare = BIAS_real > BIAS_cleaned
% if the compare value increases means better result
compare = strcat(num2str(((BIAS_real-BIAS_cleaned)/(BIAS_real)*100)), '%');
```

plot of the outliers

```
A = (1:length(data))';
plot(A,data,'.',outliers(:,1),outliers(:,2),'xk','MarkerSize',4)
hold on
plot(Median,'-r','LineWidth',1)
legend('Raw sensor Data', 'Outliers','Moving
median','location','best','fontweight','bold','FontSize',15)
ylabel('NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')

% %save the cleaned data
filename = 'cleaned_data.xlsx';   %save the edited data in this file
writecell(all_normal_data,filename,'Sheet',1)
writematrix(outliers,filename,'Sheet',2)
```

plot cleaned data with raw data

```
z = string(all_normal_data);
y = double(z(:,2));
Cmedian = movmedian(y,Window_size);

tiledlayout(2,1)
ax1 = nexttile;
plot(ax1,data,'.')
```

```
hold on
plot(ax1, Median,'-r','LineWidth',0.8)
title(ax1,'Raw sensor data','fontweight','bold','FontSize',15)
ylabel(ax1,'NH4-L2 (mg/L)','fontweight','bold','FontSize',10)
legend('Raw sensor Data','Moving median','location','best','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')

ax2 = nexttile;
plot(ax2, y,'.')
hold on
plot(ax2, Cmedian,'-r','LineWidth',0.8)
title(ax2,'Cleaned data after outlier removal with Moving median
fillter','fontweight','bold','FontSize',10)
ylabel(ax2,'NH4-L2 (mg/L)','fontweight','bold','FontSize',10)
legend('cleaned Data','Rolling
median','location','best','fontweight','bold','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```

plot daily average of both cleaned and uncleaned data

```
tiledlayout(2,1)
ax3 = nexttile;
plot(ax3,pickdata(:,1),'-','LineWidth',2)
hold on
plot(ax3, pickdata(:,2),'LineWidth',1.5)
title(ax3,'Raw sensor data','fontweight','bold','FontSize',30)
ylabel(ax3,'NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
legend('Lab data','Daily average before cleaning','location' ,
'best','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')

ax4 = nexttile;
plot(ax4, pickcleaned_data(:,1),'-','LineWidth',2)
hold on
plot(ax4, pickcleaned_data(:,2),'-k','LineWidth',1.5)
title(ax4,'Cleaned data after outlier removal with Moving median
fillter','fontweight','bold','FontSize',30)
ylabel(ax4,'NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
legend('Lab data','Daily average after cleaning','location' ,
'best','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```

plot of uncleaned sensor with lab

```
figure
plot(TT.Time,TT.Var1,'.', 'color','#6495ED', 'MarkerSize',.5)
hold on
plot(lt, numb,'.r', 'MarkerSize',15)
grid on
ylabel('NH4-L2 (mg/L)','fontweight','bold','FontSize',15);
xlabel('Recorded period','fontweight','bold','FontSize',15);
grid minor
```

```
legend('Raw sensor data', 'Lab data','location','best','fontweight','bold','FontSize',15)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```

plot of cleaned sensor with lab

```
figure
plot(T.t,T.Var1,'.', 'color','#6495ED', 'MarkerSize',.5)
hold on
plot(lt, numb,'-r', 'LineWidth',2)
grid on
ylabel('NH4-L2 (mg/L)','fontweight','bold','FontSize',15);
xlabel('Recorded period','fontweight','bold','FontSize',15);
grid minor
legend('Cleaned sensor data', 'Lab data','location','best','fontweight','bold','FontSize',15)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```

2. T-squared method for NH4 sensor data

```
clc;
clear;
close all;
```

load the data

```
[num,txt,raw] = xlsread('NH4_L2.xlsx');
data = num;
% for i=1:size(data,1)
%     if (isnan(data(i)))
%         data(i)=0;
%     end
% end
% deleterow = false(size(data, 1), 1);
% for n = 1:size(data, 1)
%   if data(n, 1) == 0
%       deleterow(n) = true;
%    end
% end
% data(deleterow, :) = [];
% data;

date{1}=txt(1:4000,1);     % I import only first 4000 date and time
date{2}=txt(4001:8000,1);
date{3}=txt(8001:12000,1);
date{4}=txt(12001:16000,1);
date{5}=txt(16001:20000,1);

date{6}=txt(20001:24000,1);
date{7}=txt(24001:28000,1);
date{8}=txt(28001:32000,1);
date{9}=txt(32001:36000,1);
```

```
date{10}=txt(36001:40000,1);

date{11}=txt(40001:44000,1);
date{12}=txt(44001:48000,1);
date{13}=txt(48001:52000,1);
date{14}=txt(52001:56000,1);
date{15}=txt(56001:61018,1);
date = date';

x{1}=data(1:4000,1);  % I import only first 4000 data points
x{2}=data(4001:8000,1);
x{3}=data(8001:12000,1);
x{4}=data(12001:16000,1);
x{5}=data(16001:20000,1);

x{6}=data(20001:24000,1);
x{7}=data(24001:28000,1);
x{8}=data(28001:32000,1);
x{9}=data(32001:36000,1);
x{10}=data(36001:40000,1);

x{11}=data(40001:44000,1);
x{12}=data(44001:48000,1);
x{13}=data(48001:52000,1);
x{14}=data(52001:56000,1);
x{15}=data(56001:61018,1);

x=x';
n=zeros(size(x));
m=cell(size(x));
S=cell(size(x));
z=cell(size(x));
is_normal=cell(size(x));
outlier_indices=cell(size(x));
NormalData_indices=cell(size(x));
Outliers=cell(size(x));
Result=cell(size(x));
alpha=cell(size(x));
alpha{1}=0.4;   % this parameter is important since our dataset is not normal. If the dataset
is normal, it is 1. However, here we must set it as a value that leads to
mean(is_normal{1})=0.95
alpha{2}=0.4;
alpha{3}=0.4;
alpha{4}=0.10;
alpha{5}=0.1;

alpha{6}=15;
alpha{7}=15;
alpha{8}=50;
alpha{9}=1;
alpha{10}=1;

alpha{11}=10;
alpha{12}=40;
alpha{13}=10;
alpha{14}=10;
alpha{15}=10;

for i=1:numel(x)
```

```
    n(i)=size(x{i},1);
    m{i}= mean(x{i});
    S{i}=0;
    for j=1:n(i)
        S{i}=S{i}+(x{i}(j,:)-m{i})'*(x{i}(j,:)-m{i});
    end
    S{i}=S{i}/(n(i)-1);

    for j=1:n(i)
        z{i}(j,:)=(x{i}(j,:)-m{i})*inv(S{i})*(x{i}(j,:)-m{i})';
    end

    is_normal{i}=(z{i}<=alpha{i});
end

for j=1:numel(x)
    outlier_indices{j}    = find(is_normal{j}==0);
    NormalData_indices{j} = find(is_normal{j}==1);
    Outliers{j}           = x{j}(is_normal{j}==0);
    Result{j} = [outlier_indices{j}, Outliers{j}];
    Normal_data{j} = [x{j}(is_normal{j}==1)];
    Normal_dates{j} = [date{j}(is_normal{j}==1)];
end
Normal_data = Normal_data';
Normal_dates = Normal_dates';
```

plot outliers

```
for j=1:numel(x)
    dim1=1;
    figure;
    plot(NormalData_indices{j},x{j}(is_normal{j},dim1),'.b','MarkerSize',5);
    hold on;
    plot(outlier_indices{j},x{j}(~is_normal{j},dim1),'xr','LineWidth',2,'MarkerSize',5);
    hold on;
    xlabel('Indices','fontweight','bold','FontSize',15)
    ylabel('NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
    legend({'Normal data','Outliers'},'location','best','fontweight','bold','FontSize',20)
    title('Outlier Detection Tsquered method','FontSize', 16);
    set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
    set(gca,'xminorgrid','on','yminorgrid','on')
end

    k=0;
for i=1:size(Result,1)
    j = size (Result{i,1},1);
    all_outliers (k+1:j+k,:) = Result{i,1};
    k = size(all_outliers,1);
end

    k=0;
for i=1:size(Normal_data,1)
    j = size (Normal_data{i,1},1);
    all_normaldata (k+1:j+k,:) = Normal_data{i,1};
    k = size(all_normaldata,1);
end

    k=0;
```

```
for i=1:size(Normal_dates,1)
    j = size (Normal_dates{i,1},1);
    all_normaldates (k+1:j+k,:) = Normal_dates{i,1};
    k = size(all_normaldates,1);
end
```

save cleaned data

```
filename = 'cleaneddata.xlsx';   %save the edited data in this file
writematrix(all_normaldata,filename,'Sheet',1)
writecell(all_normaldates,filename,'Sheet',2)
```

Daily average of the raw data

```
TT = readtimetable('NH4_L2.xlsx');
DailyAvg_actualdata = retime(TT,'daily','mean');
[numb,text,both] = xlsread('labNH4.xlsx');
lb = string(text);
lt = datetime(lb,'InputFormat','MM/dd/yyyy');
pickdata = [numb(:,1),DailyAvg_actualdata.Var1(DailyAvg_actualdata.Time(lt))];
pickdata(pickdata <0.5) = 0.5;
rows = any(isnan(pickdata),2);
pickdata(rows,:) = [];

% dialy average of the cleaned data
b = string(all_normaldates);
t = datetime(b,'InputFormat','MM/dd/yyyy h:mm:ss a');
T = timetable(t, all_normaldata);
DailyAvg = retime(T,'daily','mean');
pickcleaned_data = [numb(:,1),DailyAvg.all_normaldata(DailyAvg.t(lt))];
pickcleaned_data (pickcleaned_data  <0.5) = 0.5;
c_rows = any(isnan(pickcleaned_data ),2);
pickcleaned_data (c_rows,:) = [];
```

relative BIAS calculation for cleaned data

```
Accuracy = zeros();
for i = 1:length(pickcleaned_data)
    Accuracy (i) = abs(((pickcleaned_data(i,2)-
pickcleaned_data(i,1))/(pickcleaned_data(i,1))));

end
BIAS_cleaned = mean(Accuracy);

% relative BIAS calculation for uncleaned data
Accuracyuncleaned = zeros();
for i = 1:length(pickdata)
    Accuracyuncleaned (i) = abs(((pickdata(i,2)-pickdata(i,1))/(pickdata(i,1))));

end
BIAS_real = mean(Accuracyuncleaned);

% compare = BIAS_real > BIAS_cleaned
% if the compare value increases it means better result
```

```
compare = strcat(num2str(((BIAS_real-BIAS_cleaned)/(BIAS_real)*100)), '%');
```

plot daily average

```
figure
tiledlayout(2,1)
ax3 = nexttile;
plot(ax3,pickdata(:,1),'-','LineWidth',2)
hold on
plot(ax3, pickdata(:,2),'LineWidth',1.5)
title(ax3,'Raw sensor data','fontweight','bold','FontSize',20)
ylabel(ax3,'NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
legend('Lab data','Daily average before cleaning','location' ,
'best','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')

ax4 = nexttile;
plot(ax4, pickcleaned_data(:,1),'-','LineWidth',2)
hold on
plot(ax4, pickcleaned_data(:,2),'-k','LineWidth',1.5)
title(ax4,'Cleaned data after outlier removal by T-squared
method','fontweight','bold','FontSize',20)
ylabel(ax4,'NH4-L2 (mg/L)','fontweight','bold','FontSize',15)
legend('Lab data','Daily average after cleaning','location' ,
'best','fontweight','bold','FontSize',10)
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```

plot of uncleaned sensor with lab

```
figure
plot(TT.Time,TT.Var1,'.b', 'MarkerSize',.5)
hold on
plot(lt, numb,'.y', 'MarkerSize',15)
grid on
ylabel('NH4-L2 (mg/L)');
xlabel('Recorded period');
grid minor
legend('Raw sensor data', 'Lab data','location','best')
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
% plot of cleaned sensor with lab
figure
plot(T.t,T.all_normaldata,'.k', 'MarkerSize',.5)
hold on
plot(lt, numb,'.y', 'MarkerSize',15)
grid on
ylabel('NH4-L2 (mg/L)');
xlabel('Recorded period');
grid minor
legend('Cleaned sensor data', 'Lab data','location','best')
set(gca,'color','#FFFOF5','YGrid','on','XGrid','on')
set(gca,'xminorgrid','on','yminorgrid','on')
```