# WIRELESS SENSOR NETWORKS AND FUSION OF CONTEXTUAL INFORMATION FOR WEATHER OUTLIER DETECTION

A. Amidi [a, *], N.A.S. Hamm [a], N. Meratnia [b]

[a] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands - Amidi29183@itc.nl, N.Hamm@utwente.nl
[b] Pervasive System Group, Department of Computer Science (EWI), University of Twente, Enschede, The Netherlands - N.Meratnia@utwente.nl

**KEY WORDS:** Temporal outlier detection, pattern formation, similarity assessment, slopes, contextual information, forecasts, weather.

**ABSTRACT:**

Weather stations are often expensive hence it may be difficult to obtain data with a high spatial coverage. A low cost alternative is wireless sensor network (WSN), which can be deployed as weather stations and address the aforementioned shortcoming. Due to imperfect sensors in WSNs context, provided raw data may be drawn in from of a low quality and reliability level, expectedly that is an emergence of applying outlier detection methods. Outliers may include errors or potentially useful information called events. In this research, forecast values as contextual information are utilized for weather outlier detection. In this paper, outliers are identified by comparing the patterns of WSN and forecasts. With that approach, temporal outliers are detected with respect to slopes of the WSNs and forecasts in the presence of pre-defined tolerance. The experimental results from the real data-set validate the applicability of using contextual information in the context of WSNs for outlier detection in terms of accuracy and energy efficiency.

## 1. INTRDUCTION

### 1.1 Motivation

Recent developments in Earth observation have heightened the need for providing inexpensive data with high spatial and temporal resolution. Due to the high cost of weather stations, wireless sensor network (WSN) can be considered as a reasonable alternative. A WSN can be a weather station network, sending weather information and announcing related events. This system takes the advantage of WSNs which provide local data and can send signals over far distances by using a mesh topology. The system transfers the data while consumes low power. Thus, this system can be installed in locations that are difficult to hardwire or have no access to electricity. Wireless sensor nodes are equipped with sensing, processing, and actuation capabilities (Liu et al., 2003) that are linked by wireless media to transform the required and partially processed data (Díaz-Ramírez et al., 2012).

Quality of data provided by WSNs is critical while, collected raw data may be of low quality because of the WSNs imperfect nature. To assure the quality of data obtained from WSNs, outlier detection refines the measured values. Outliers are those observations that appear to deviate considerably from remainder of that set of data (Hodge et al., 2004). On the other hand, normal data are data that follow the expected patterns. One of the most significant current discussions in WSNs is to extract potentially useful information from captured raw data. Outlier detection techniques are utilized to distinguish normal data from error, event, and malicious attacks on the network (Zhang et al., 2010).

In this study pattern based temporal outlier detection method is proposed. The main idea is to use contextual information for outlier detection. It involves three steps. The first step is pattern formation. The second step is similarity assessment on the basis of slopes. The third step is temporal outlier detection. Experiments are conducted on a real data-set. Set of measurements are utilized to evaluate the applicability of the proposed methodology.

## 2. LITERATURE REVIEW

### 2.1 Comparison of outlier detection methods

In this section outlier detection techniques are classified based on the discipline from which they adopt their ideas. Subsequently, based on the corresponding drawbacks this research is motivated for conducting.

#### 2.1.1 Statistics-based outlier detection techniques

Statistical outlier detection techniques assume a statistical model. The distribution of the sensor data is captured and the instances data are assessed to see how well they fit the model (Zhang, et al., 2010). An instance sensor data is labelled as an outlier since the probability that the data could be generated by the model is very low (Chandola et al., 2007). These techniques assume that majority of sensor data includes normal observations (Zhang, 2010) since WSNs imperfect sensors can

* Corresponding author.

break the assumption especially in the presence of the faulty nodes.

#### 2.1.2 Distance-based outlier detection techniques

Distance-based outlier detection methods utilize fix or flexible thresholds to identify those outliers that are deviated from the majority of the sensor data by using proper techniques for geometric distance measurement (Gogoi et al., 2011). These methods are usually computationally expensive and suffer from the choice of the appropriate threshold.

#### 2.1.3 Clustering-based outlier detection techniques

Clustering-based outlier detection techniques cluster similar sensor data instances into groups with respect to the behavioural similarities. WSNs measurements are labelled as outliers if they do not belong to clusters or if their clusters are significantly smaller than other clusters (Zhang, et al., 2010). These techniques are computationally expensive while computation of pairwise distances should be accomplished for every incoming data (Chandola, et al., 2007). The width of the clusters can either be specified by user or can be derived from the data. As a consequence, assigning the clusters width is challenging since WSNs data are not usually precise and dense and may be scattered spreadly.

#### 2.1.4 Classification-based outlier detection techniques

Classification-based methods identify outliers by using training samples to learn the classification model and assigning WSNs data to outlier or normal classes (Chandola, et al., 2007) in supervised and semi-supervised versions. On the other hand, unsupervised techniques learn the classification model which fits the majority of the data during training (Zhang, et al., 2010). Supervised and semi-supervised techniques assume availability of labelled training samples for model building while they are not necessarily available specifically in real world applications and also new types of normal or outlier data may not be included in the pre-labelled data. The unsupervised methods assume outliers do not belong to frequent patterns which can be broken due to rigid environment, inexpensive sensors, network failure and faulty nodes.

Generally, outlier detection methods in WSNs can be restricted on the basis of prior knowledge based and prior knowledge free methods (Xie et al., 2011). The prior knowledge for outlier detection can be achieved through making assumptions (Palpanas et al., 2003; Subramaniam et al., 2006) and considering experiences (Da Silva et al., 2005; Ioannis et al., 2007). Errors may occur frequently due to imperfect sensors, harsh condition of environment, mechanical errors, and changes in system behaviour, fraudulent behaviour, human error and instrument error (Gogoi, et al., 2011). Thus, utilizing sensor data as a prior knowledge is not always safe and reliable. Furthermore, retrieving prior knowledge from sensor data requires sufficient amount of data (Zhang et al., 2012) which may not be easily available especially in temporary deployments and small networks.

To detect outliers in the absence of sufficient amount of historical and spatial data in the temporary deployments or small networks accurate outliers due to complex weather patterns cannot be identified. Thus, a novel approach was proposed to cover up the abovementioned imperfection. The climatic forecasts are utilized to identify outliers. Experiments from the real data-set are utilized to assess the performance of the proposed methodology.

## 3. STUDY AREA AND DATA SOURCES

The study area is located over the Grand Saint Bernard pass, a harsh mountain environment between Switzerland and Italy since the elevation range in the study site is from 2300 meters to 2500 meters.

Grand-St.-Bernard WSN data-set was observed from 13th September 2007 to 26th October 2007 and recorded at coordinated universal time (UTC) time zone for 45 days. The data-set was provided by 23 sensor nodes. Seventeen of the stations were installed in Switzerland and a few were installed over the Italian border (Figure 1). The proposed methodology is experimented on nodes from the small cluster (25, 28, 29, 31 and 32) and ambient temperature on 2007-09-30. Furthermore, the frequency of the WSN data sampling was two minutes. The accuracy of the temperature sensors' is ±0.3.

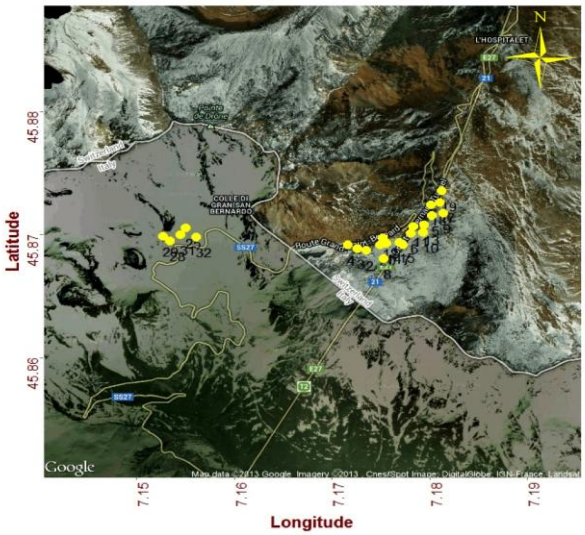


Figure 1. The Grand-St.-Bernard WSN deployment in geographic coordinate system.

To perform temporal outlier detection, climatic products are utilized in this study as contextual information that are provided by the national weather service of Switzerland (MeteoSwiss) that are illustrated by Figure 2. MeteoSwiss forecasts (MFs) are applied for temporal outlier detection. Forecast data are provided by MeteoSwiss that are hourly values 72 hours in advance. Forecasts are instantaneous values but the model time step is 60s only and the real time resolution thus might be about five minutes.

Climatic observations are utilized to evaluate the experimental results which are also provided by MeteoSwiss. Hourly mean data of air temperature two meter above the ground is provided by MeteoSwiss observations (MOs) are obtained from a station located in 'Col du Grand St-Bernard' at an elevation of 2472 meters.

## 4. TEMPORAL OUTLIER DETECTION USING PATTERNS

This study identifies the temporal outliers by participating the contextual information. To identify temporal outliers, MFs and

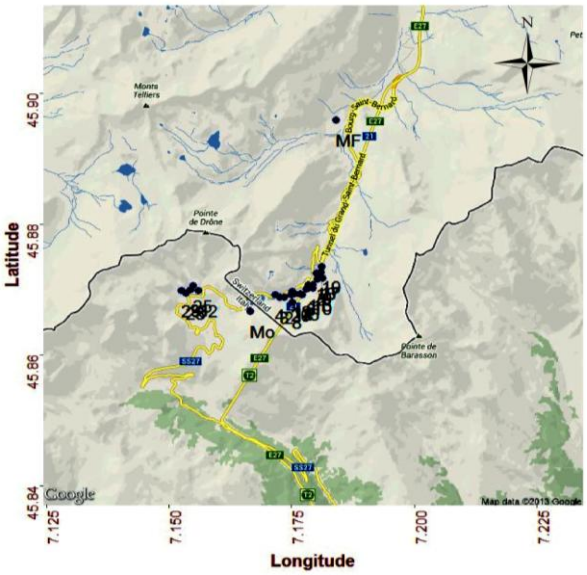


Figure 2. The location of utilized data-sets in this research, where 'MF' presents the MetoSwiss forecast grid point and 'Mo' illustrates the MetoSwiss observation station and two clusters show the WSN small and big clusters in geographic coordinate system.

WSN measurements are compared. The absolute values of WSN and MFs cannot be compared directly due to altitude difference between the climatic and WSNs stations and study area's complex topography. Furthermore, MFs cannot be as local as WSNs measurements for the entire network, which is a key point of WSNs existence. In this research WSN measurements and MFs are compared in terms of temporal patterns. Temporal correlation between successive values of MFs and WSN measurements are utilized to expose temporal patterns in order to identify temporal outliers.

### 4.1 Pattern formation

Behaviour of the temperature on a specific time span at a big picture is called a pattern. Patterns are obtained by querying the values of temperature at consecutive points in time. In order to, form consistent patterns, WSN measurements and MFs temporal resolution should be matched. The mean aggregation function is applied to make the WSN measurements consistent with respect to the MFs. Using the mean aggregation function to provide the similar temporal resolution in WSN yields benefits such as reducing the effects of infrequent tiny deviations and small fluctuations.

### 4.2 Similarity assessment

The main idea of the similarity assessment is to provide a tool to evaluate the produced patterns of WSN measurements and MFs. The similarity assessment is performed by calculating the slopes of WSNs measurements and MFs. Slopes for WSNs measurements are calculated from differences between a current value and previous value within the same time period. On the other hand, slopes for MFs are computed by subtracting two successive hourly forecasts within similar time scale. Using slopes to identify outliers ignores the vertical shifts, global scaling and shrinking (Suntinger et al., 2010) to enhance the effects of complex topography and harsh condition of the study area.

### 4.3 Outlier identification

In WSNs outliers are matter of definitions. In this study, outliers are those observations of WSN on a specific time slice where patterns of WSN and MFs show differences. Patterns are considered similar when the slopes of the MFs are placed within the tolerance of the WSN slope. In other words, if the slope of the MFs at a specific hour lay outside the WSN slope tolerance, a subsequent relevant WSN observation of that hour is labelled as an outlier, otherwise labelled as normal. The tolerance is defined based on the accuracy of the wireless sensor device. Tolerance represents the possible maximum and minimum variation in WSN measurements with respect to the accuracy of the sensors. The possible upper and lower bounds of the WSN measurements are called tolerances.

### 4.4 Correlation analysis

Most data represent some degree of spatial autocorrelation, depending on the scales at which data were observed and then analysed (Fortin, 1999). Additionally, temporal autocorrelation can be expected between successive measurements on the basis of spatial scale. Owing to the low number of wireless nodes (23), the spatial correlation was not explored. Consequently, the temporal correlation effect between contextual information and WSN measurements is investigated.

The Pearson correlation method (Lee Rodgers et al., 1988) is utilized in this research. This measures the strength of the linear relationship between two variables. Among the correlation methods the Pearson is most appropriate for measurements taken from an interval scale (Lund Research Ltd, 2013a) since in this research variables are ambient temperature.

On the other hand, the Spearman rank correlation method (Pirie, 1988) is also applied to limit the assumption on the linearity of the relationship or on data distribution. In Spearman rank order correlation method monotonic relationship between variables is needed (Lund Research Ltd, 2013b).

### 4.5 Accuracy assessment

The detected outliers are assessed via MeteoSwiss observations (MOs) to investigate the suitability of using forecasts. MOs are utilized as a reference data since they do not include the uncertainties of the forecasts. Evaluation is performed based on detection rate (DR) and false positive rate (FPR). DR presents the percentage of correctly detected outliers and FPR presents the percentage of normal data that are incorrectly flagged as outlier.

### 4.6 Evaluation of complexity and energy efficiency

In designing outlier detection methods for WSNs not only the detection accuracy but also the energy consumption issues should be considered. The complexity of applied method for outlier detection is assessed in terms of communication overhead, computation and memory complexity.

## 5. EXPERIMENTAL RESULTS

Temporal correlation between successive values of MFs and WSN measurements are established to identify temporal outliers by means of patterns. Figure 3 reveals the patterns of instantaneous hourly values of WSN measurement against the MFs on 2007-09-30. Wireless nodes show quite different values of the temperature between 12:00 (midnight) to 08:00 and 12:00 (midday) to 16:00. In spite of the considerable differences in absolute values, similar trends were found between the WSN and MFs at 01:00 to 02:00 and 04:00 to 06:00.

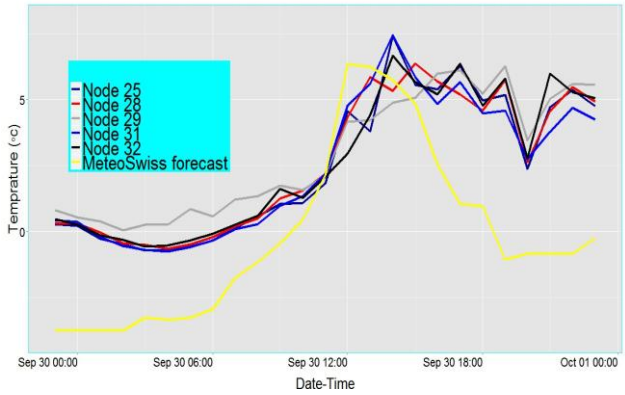


Figure 3. The patterns of WSN ambient temperature measurements and MeteoSwiss forecasts on 2007-09-30.

Temporal outliers are identified by evaluating the similarities of the WSN slopes against the MFs by considering the tolerance. In this study ± 0.6 is selected as tolerance since the accuracy of the ambient temperature sensor was ±0.3. Number of detected outliers in WSN measurements on 2007-09-30 are represented by Table 1.

| ***Node ID*** | **N25** | **N28** | **N29** | **N31** | **N32** |
|---|---|---|---|---|---|
| ***Number of outliers*** | 14 | 12 | 11 | 13 | 14 |

Table 1. Number of outliers detected at different wireless nodes.

The largest number of outliers was detected at nodes 25 and 32 with 14 outliers and smallest number of outliers are identified at node 29 with 11 outliers.

Figure 4 shows the result of accuracy assessment for detected outliers by using pattern approach. The highest detection rate (100%) is at node 29 while the lowest detection rate (70%) is at node 32. The lowest amount of FPR is at node 31 (33%) while the highest rate is at node 25 (52%).

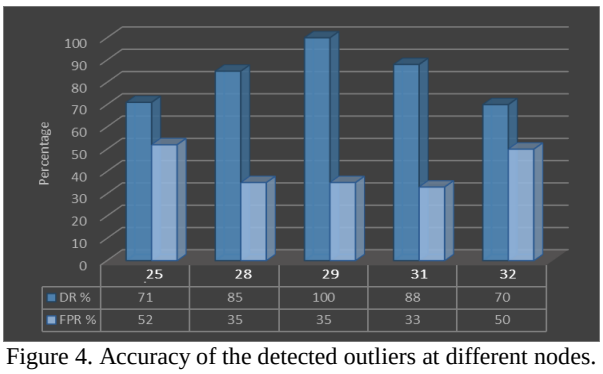


Figure 4. Accuracy of the detected outliers at different nodes.

The Pearson correlation results are presented in Table 2. The highest correlation between nodes and MFs on 2007-09-30 is observed at node 31 and the lowest one is at node 29.

| ***Node ID*** | ***N25*** | ***N28*** | ***N29*** | ***N31*** | ***N32*** |
|---|---|---|---|---|---|
| ***MFs*** | 0.73 | 0.76 | 0.63 | 0.82 | 0.66 |

Table 2. The coefficients of Pearson correlation between wireless nodes and MFs.

The result of Spearman rank correlation method is summarized in Table 3. Similarly, the highest correlation between WSN and MFs is happened at node 31. The lowest correlation is occurred simultaneously by nodes 29 and 32.

| ***Node ID*** | ***N25*** | ***N28*** | ***N29*** | ***N31*** | ***N32*** |
|---|---|---|---|---|---|
| ***MFs*** | 0.77 | 0.77 | 0.72 | 0.84 | 0.72 |

Table 3. The coefficients of Spearman correlation between wireless nodes and MFs.

The temporal outlier detection method is performed at each node and no communication overhead is needed. The computational complexity in temporal outlier detection includes performing the aggregation and similarity assessment in the case of utilizing patterns. For calculating the aggregation the dimension of the WSN vector which includes a number of observations and numbers of variables are important and presented by O(m.d). The computation complexity for MFs can be neglected due this fact no much computation is needed. The memory complexity of temporal outlier detection is mainly about the storing of observations of WSN and MFs and depends on the number of observations and variables O(m.d).

## 6. DISCUSSION AND FUTURE WORK

Due to the fact that MeteoSwiss products were offered at the hourly scale, the detected outliers were represented in hourly resolution as well. Based on the users requirements and interested applications hourly detection and proposed method may be considered (near) real-time.

When a WSN measurement is flagged as an outlier it reveals that the WSN measurement does not follow the expected pattern with respect to the MFs. Consequently, the outlier may occur because of an error in WSN measurements or a weather event which was not forecast. Furthermore, by applying the proposed methodology not only the outliers but also the normal data represent informative information. The occurrence of weather events that are forecasted (normal data) represents useful information. On the other hand, abnormal occurrence (outlier data) may reveal useful information by comparing the forecasted data with observed data. For a specific example, based on forecasts the temperature would drop suddenly in a time span because of the blizzard then two general options can be supposed. Firstly, WSN observation follows the forecast pattern thus the semantic of the WSN observation is known. Secondly, WSN measurement does not represent the similar pattern with respect to the forecast so, regarding to WSN pattern and comparing it with forecast pattern the probable event can be interpreted.

It should draw a careful attention that a high false positive rate may be caused because of the forecasts' age. Forecasts made closer in time should be more accurate. In this study forecasts were provided 36 hours in advance and forecasts for 2007-09-29 are expected to be more accurate than 2007-09-30.

One unanticipated finding was that, the meaningful relation was not observed between detected outliers and coefficient of correlations. It was expected to observe fewer of outliers where the coefficient of the correlation was high. In Pearson method the assumption of linearity is too simple since the slopes may change systematically. The Spearman rank correlation was higher than Pearson correlation that just means the monotonic correlation is larger than the linear correlation. It is suggested that the association of the correlation between WSNs and contextual information is investigated in future studies with respect to the WSNs specifications.

Applying energy efficient methods for outlier and event detection increases the networks lifetime which is a prerequisite for an accurate detection. Lack of energy in nodes can lead to inaccurate observations or even cause incomplete data transmission. The main concern over energy efficiency and complexity are expressed for communication overhead. The communication complexity depends on the transmission rate since the proposed distributed method for temporal outlier detection identifies outliers, locally.

Distinguishing events from errors in outliers is a considerable challenge. Further research should be done to investigate a method for event detection. Besides, further work should also focus on accuracy assessment for distinguishing between errors and events.

## 7. CONCLUSION

In this paper, a novel distributed outlier detection method has been proposed based on using contextual information. The proposed methodology was assessed in terms of accuracy and energy. Experimental results indicated a convincing DR. The design of the methodology enabled each node to identify outliers in an energy friendly manner.

This paper is highlighted the effectiveness and applicability of utilizing contextual information for outlier detection in WSNs. The proposed methodology is appropriate especially in temporary deployment and specific applications where sufficient historical data does not exist. Finally, WSNs potentially are effective tools to substitute costly weather stations.

## REFERENCES

Chandola, V., Banerjee, A., Kumar, V. (2007). Outlier detection: A survey. *To appear, ACM Computing Surveys*.

Da Silva, A. P. R., Martins, M. H., Rocha, B. P., Loureiro, A. A., Ruiz, L. B., Wong, H. C. (2005). *Decentralized* intrusion *detection in wireless sensor networks.* Paper presented at the Proceedings of the 1st ACM international workshop on Quality of service & security in wireless and mobile networks, Montreal, Canada.

Díaz-Ramírez, A., Tafoya, L. A., Atempa, J. A., Mejía-Alvarez, P. (2012). Wireless Sensor Networks and Fusion Information Methods for Forest Fire Detection. *Procedia Technology, 3*, 69-79.

Fortin, M.-J. (1999). Effects of *sampling* unit resolution on the estimation of spatial autocorrelation. *Ecoscience, 6*(4), 636-641.

Gogoi, P., Bhattacharyya, D., Borah, B., Kalita, J. K. (2011). A *survey* of outlier detection methods in network anomaly identification. *The Computer Journal, 54*(4), 570-588.

Hodge, V. J., Austin, J. (2004). A survey of outlier detection *methodologies*. *Artificial Intelligence Review, 22*(2), 85-126.

Ioannis, K., Dimitriou, T., Freiling, F. C. (2007). *Towards intrusion detection in wireless sensor networks.* Paper presented at the Proceedings of the 13th European Wireless Conference, Paris, France.

Lee Rodgers, J., *Nicewander*, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*(1), 59-66.

Liu, J., Chu, M., Reich, J., Zhao, F. (2003). State-centric *programming* for sensor-actuator network systems. *Pervasive Computing, IEEE, 2*(4), 50-62.

Lund Research Ltd. (2013a). Pearson Product-Moment Correlation Retrieved 2013-09-03, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Lund Research Ltd. (2013b). Spearman's Rank-Order *Correlation* Retrieved 2013-09-01, from https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Palpanas, T., Papadopoulos, D., Kalogeraki, V., Gunopulos, D. (2003). *Distributed* deviation detection in sensor networks. *ACM SIGMOD Record, 32*(4), 77-82.

Pirie, W. (1988). *Spearman* rank correlation coefficient. *Encyclopedia of statistical sciences*.

Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., Gunopulos, D. (2006). *Online outlier detection in sensor data using non-parametric models.* Paper presented at the Proceedings of the 32nd international conference on very large data bases, Seoul, Korea.

Suntinger, M., Obweger, H., Schiefer, J., Limbeck, P., Raidl, G. (2010). *Trend-Based Similarity Search in Time-Series Data.* Paper *presented* at the Second International Conference on Advances in Databases Knowledge and Data Applications (DBKDA), France.

Xie, M., Han, S., Tian, B., Parvin, S. (2011). Anomaly detection in *wireless* sensor networks: A survey. *Journal of Network and Computer Applications, 34*(4), 1302-1325.

Zhang, Y. (2010). *Observing the unobservable: distributed online outlier detection in wireless sensor networks.* Ph.D. thesis, University of Twente, Enschede. (10-174)

Zhang, Y., Hamm, N., Meratnia, N., Stein, A., van de Voort, M., *Havinga*, P. (2012). Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science, 26*(8), 1373-1392.

Zhang, Y., Meratnia, N., Havinga, P. (2010). Outlier detection *techniques* for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials, 12*(2), 159-170.