Smart outlier detection of wireless sensor network

Sahar Kamal[a], Rabie A. Ramadan[b] and Fawzy EL-Refai[c]
[a] *Department of Electronics and Electrical Communications*
Higher Institute of Engineering, [b]*Department of Computer Engineering*
[c] *Department of System and Computer Engineering,*

, [a]*El-Shorouk Academy* ,[b]*Cairo University on leave at Hail University,*
And [c]*El-AZhar University*
[a]Cairo@El-Shorouk City,
[bc]*Cairo@Egypt*

Abstract— **Data sets collected from wireless sensor networks (WSN)are usually considered unreliable and subject to errors due to limited sensor capabilities and hard environmental resulting in a subset of the sensors data called outlier data .This paper proposes a technique to detect outlier data base on spatial-temporal similarity among data collected by geographically distributed sensors . The proposed technique is able to identify an abnormal subset of data collected by sensor node as outlier data. Moreover the proposed technique is able to classify this abnormal observation, an error data set or event affected set. Simulation result shows that high detection rate is achieved compared to conventional outlier detection techniques while preserving low positive false alarm rate.**

Keywords—**wireless sensor network, outlier's detection, fuzzy logic, spatial and temporal similarity.**

I. NTRODUCTION

Wireless sensor network is considered a promising solution for monitoring and measurement of natural physical phenomena such as temperature, humidity, earthquakes, pressure, light ,volt etc. a typical WSN consists of a large number of very small sensors are deployed over a topological area of interest. These sensors are supplied by power resources (batteries, solar cell), measurement unites, processing unites and wirelesses TX/RX unit. Unfortunately, the data collected from sensor nodes are considered inaccurate and may be even unreliable due to measurement errors or superimposed noise on the received data packets in [1]. Duplicated measurement or even missing values are not common in the data set collected by a WSN. A subset of data which appear to be in consistence with the whole data set from which it is collected is called an outlier. Outlier can be defined as in [1] "an outlier is a subset of measurements which appear to be inconsistent with other dataset". On the other hand, outliers as in [2] can be defined as "those subset of observation that is deviated from the normal dataset". Both two definitions can be used as solution to define outlier in data set. Abrupt events such as sudden sensor failure, battery power deployment or even natural arise of physical phenomena are also reasons to which outlier data can be attributed. In order to boost the accuracy and reliability of collected sensor data, an outlier detection Process should be applied and possibly corrected.

There are three sources of outliers due to an environmental changes or error coming from a faulty sensor, it can be defined as (1) errors& noise, (2) events and (3) malicious attacks, the latest one is related to the network security as in [1]. Noise refers to data instance coming from a faulty sensor or noise measurement. Outliers occur frequently are classified as error, while outliers have smaller probability of occurrence are classified as event. Outlier data is normally represented as an arbitrary change and is extremely different from the rest of the data. Noisy data as well as erroneous data should be eliminated or corrected if possible. However, events may arise due to sudden change in the real environment, e.g. air pollution, rainfall, forest fire, chemical spill, etc. Eliminating event outlier from data instance will lead to damage of important hidden information of the data about events as in [3].Outliers that are very close to random errors in terms of size can only be determined through the application of outlier tests. Outlier classification as an event or error is an important matter. Many researches consider outliers and events as similar conditions by considering events as some sort of outliers. Due to the fact that there are existed spatial-temporal similarities between neighboring nodes measurements enable us to classify outlier as event or error. This depends on the fact that error data observations seem to be unrelated while event observations seem to be spatial correlated as in [4].

The main methods to determine outliers can be grouped as, Nearest Neighbor-Based, Statistics-based methods, Cluster-Based and Artificial Intelligence techniques. New approaches are used for outlier detection including Artificial Intelligence techniques such as Fuzzy Logic and Neural Network techniques. Fuzzy logic was suggested by [5] in which it can also be used to geodetic networks for outlier detection. The essential aim of outlier detection algorithms is to detect outliers with high detection rate while keeping false positive rate resource constrains low.

Our work is based on the observation that, in most applications of WSNs, measurements of sensors in the environment tends to be highly correlated for sensors that are geographically close to each other (spatially similarity), and also highly correlated for a period of time (temporally similarity). Using this observation, we take advantage of the spatial and temporal similarity in the sensor data. Frist study, we detect outliers in univariate attribute in WSN. The main contribution of this paper is the uses of Euclidean distance and fuzzy logic to detect outliers in wireless sensor networks .However, spatial and temporal similarity were used, that make it easy to distinguish between error and event. If probability of output of fuzzy logic is exceeding a preselected threshold, the observation is considered as an outlier. The model is tested on real data set from Grand-St-Bernard as in [6] and implemented using MATLAB. This paper achieve high detection rate and keep low false positive alarm rate and computational complexity.

The rest of the paper is organized as follows Section (2), shows a necessary background definition related to outlier detection. Section (3) represents Modeling sensors operation of WSN. The proposed algorithm is presented in section (4) along with the assumptions upon which the proposed technique is built. Section (5) shows experimental results and the performance evaluation of the proposed technique using a realistic data set. Finally, the whole paper is concluded in section (6).

II. RELATED WORK

Recently, there are many researches in outlier detection of WSN to improve reliability and quality of measurement sensor. These researches used different technique to detect outlier such as statistical-based, nearest neighbor-based, clustering-based, classification-based, and spectral decomposition-based approaches. In general, these researches can be those that do not use spatial or temporal correlation data set or those that are based on spatial or temporal correlation only or on both. In 2006, author as in [7], use the spatial correlation exists between neighboring sensor nodes to classify outlying sensors about event boundary. In this model, each node calculates subtract its own measurements and the median from its neighboring measurements. Then outlying node is declared when absolute value of its measurement's deviation degree is greater than a predefine threshold. This technique suffers from low detection rate because it ignores temporal correlation between sensor data reading. As shown by [8], this model used cluster based technique to identify global outlier .Frist, each node clusters the reading and reports cluster summaries then transmit the raw sensor reading to its cluster head. Cluster head collect cluster summaries from all of its nodes before sending them to the sink. An outlier cluster can be declared in the sink if the cluster's average inter-cluster distance is greater than one threshold value of the set of inter-cluster distances. However, these models suffer from the choice cluster width parameter. Additionally, these technique increase computational complexity when compute distance between data instance. In [9] author uses distance similarly to identify global outliers in WSN. Each node use a distance similarly to identify local outliers and then broadcasts abnormal data Instances to all neighboring node for verification .this technique is repeated until all neighboring node agree on the global outliers. This technique increases computational complexity and it isn't adapted to large scale network. In 2007, proposed technique as in [10] uses one class quarter sphere based technique to detect outliers in WSN. This technique takes advantage of temporal correlation to identify local outliers at each node. A measurements sensor that lies outside the quarter sphere is considered as an outlier. Each node transmits only brief information to its parent for global outlier's classification. This technique suffers from low detection rate because it ignored spatial correlation between neighboring nodes. At 2008, author as in [11] use a centered quarter-sphere support vector to detect local outlier in WSN. This technique takes advantage of spatial correlations that exist in sensor data of adjacent nodes to reduce the false alarm rate and to classify outlier as events or errors .but it ignores temporal correlation and increase computational complexity. But in 2009, author as in [12] used outlier detection technique to identify outliers in data set of WSN. This technique takes advantage of spatial temporal correlation exist among sensor data reading. In 2011, author as in [13] proposed outlier detection method in the wireless sensor networks and distinguee between event and error. This technique used to classify the sensor node data as local outlier or cluster outlier or network outlier. This technique consider network outlier or cluster outlier as event and local outlier as error. This algorithm suffers from high computational complexity. In 2012, author as in [14] use advantage of temporal correlation only, spatial correlation only and spatial temporal correlation to detect outlier in WSN by statically model this technique suffers from some computational complexity. This model differs from our model that, our model has the advantage of spatial-temporal similarity combined with fuzzy logic to detect outlier and identify errors and events with high detection rate and relatively low false positive rate than result as in [14]. In 2013, author as in [15] use temporal and spatial properties to identify outliers and distinguish between event and error but with low detection rate and false positive rate than our approach.

III. THE PROPOSED STODM TECHNIQUE

Sensor nodes are assumed to be densely deployed and synchronized in WSN. A subset of sensor is considered as members of the same cluster if they fall within the same radio transmission range of each other. At any time interval Δt, each node reads a data vector s_{ij} where "i" is the time index of the data symbol and "j" is the node spatial ID. The potential of an outlier detection technique is to identify a subset x_i of each sensor set s_i as outliers. A super advantage of a given detection technique is to classify deviation data instance as event or error.

In this section, the proposed approach is introduced in details. Many outlier detection techniques have been developed, however they did not take into account the interesting events. On the other hand, some recently developed researches are focus only in events and did not care about erroneous data. In this paper, a new distance-based approach depends on spatial-temporal similarity combined with fuzzy logic-based approach is proposed to classify outliers, i.e. error data or events. Our methodology consists of the following steps: first step the spatial and temporal similarity is calculated, each one of these is entered as input or (membership function) to fuzzy logic to detect outliers in each node. Second step classify outlier as event or error.

A. *Spatial-Temporal Similarity*

In our proposed algorithm, spatial-temporal similarity is calculated using a two-step process.

First step, the temporal similarity of a given data set of sensor node is calculated on point by point basis and is given by first order difference| s_{i2}-s_{i1}|. The absolute difference is compared to a pre-specified threshold which is calculated according to tolerance of temperature sensor. A data point s_{i2} is considered similar to other points if the absolute first order difference doesn't exceed the threshold. Otherwise dissimilarity is obtained and point of data may be outlier.

Second step, spatial similarity is calculated based on distance between neighboring nodes, We use the Euclidean distance to calculate similarity measure between two point x, y that are in the same transmit ion range and they are in the same close time which is calculated as Eq. (1). Euclidean Distance is a popular choice for univariate and multivariate continuous attributes as [16]. Data instance in point x is considered similar to data point in y if Euclidean distance d(x, y) isn't exceed Preselected threshold. Spatial link is defined as number of spatial similarity to each point with its neighbors as in Eq. (2). Where spatial similarity threshold is calculated by computing mean distance of all data points in the close time.

$$D(x, y) = \sqrt{(x-y)^2} \qquad (1)$$

$$Spatial\ link = \sum_{i=1}^{n} no_{of}\, similarity\ to\ each\ node \qquad (2)$$

Where n is the number of neighboring nodes.

B. *Fuzzy Logic Model*

Recently, many approaches have been tested on decision making theories. Some of the artificial techniques that are used in outlier analysis are Neural Networks, Support Vector Machine and Fuzzy Logic as in [17] .our approach use fuzzy logic as one of artificial techniques to detect outliers in data set of WSN .Fuzzy logic is a logical model providing a general idea about the decision process in the analysis of the data set. The fuzzy logic suggested by [18] is essentially an approach that allows transition values to make a definition between the conventional values such as right/wrong, yes/no, high/low. The main purpose of the method is to bring a certainty to assigning a membership degree to the concepts which are hard to express or have difficult meaning. A fuzzy logic system consists of three main parts, which are fuzzification, rule base and defuzzification. Firstly, fuzzification can be defined as a transfer between a definite system and a fuzzy system and it describes a property of an object in a certain fuzzy set. The objects can belong to 'low, middle, high' property classes with membership functions and each object are assigned to a membership degree between 0 and 1. This technique use temporal and spatial similarity as two inputs or two membership function to fuzzy system. These membership functions are chosen empirically and optimized using a sample input/output data. The most common membership functions include the triangle, trapezoid, Gauss curve and sigmoid. As the membership functions represent the fuzzy set, the selection of their shape and form directly affects the decision process.

Secondly, the rule base combines the membership functions from the fuzzificator with the rule handling data such as 'if, and, although, if not' which is based on the database and stored there. The If-then rules define a connecting antecedent to the consequent (i.e. input to output). These rules are given weights based on their criticality as in [18]. With this approach, measurements can be classified according to their membership degrees by adequate membership e.g.

- If spatial link (low) and temporal similarity (low) then outlier (high)
- If spatial link (low) and temporal similarity (med) then outlier (high)
- If spatial link (high) and temporal similarity l (high) then outlier t (low)
- If spatial link (med) and temporal similarity (med) then outlier (med)

Thirdly, in the defuzzification unit, the rule results that are obtained from the rule handling unit are evaluated in the fuzzificator and turned into definite results as in [18]. Outlier is declared according to the rule results. Fig.1 represents all three stages of fuzzy logic.

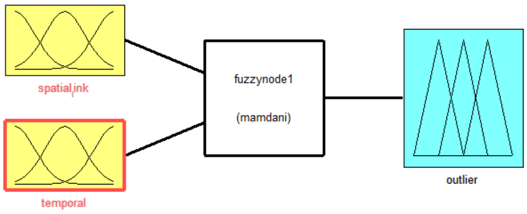

Fig.1. three stage of fuzzy logic

C. *Outlier Classification*

The third step is to classify the degree of outlier value (error or event). In this step, we aim to know the source of the values labeled as outlier. There are two possible options; either this outlier value is due to an error, as a result of low battery or network damage, or due to an event or phenomena in the surrounding environment. Our idea is based on the following observation in the result of this technique *"Error in the sensor* *data are likely to be spatially unrelated while event measurements are probable to be spatially correlated".*

On the other hand, data instance tends to be correlated in both time and space. Hence, we employed this fact by using data from neighboring nodes to assist measuring the spatial similarity, also using time stamps between readings to assist measuring the temporal similarity. In more details, this technique detects the outlier in the previous step and if data instance are declared as outlier produce similar values or values larger than the outlier reading in all nodes, in addition those neighboring nodes readings are within the same time range, and this indicates that it is an interesting event in the physical world. Otherwise, it is likely to be an erroneous data. In our work, we assume that a sensor node (x) is considered to be a neighbor of another node (y) if x is within y's communication range, and vice versa.

IV.EXPERIMENTAL RESULT AND PERFORMANCE EVALUATION

In this section, we investigate the effectiveness of our proposed approach when applied on the real dataset from St.-Bernard wireless sensor network in [6]. We compare the accuracy of our algorithm with another detection method called STGOD method [14], which is based on spatial temporal correlation among neighbor nodes. We evaluate accuracy and the scalability of the proposed method against the STGOD method on real dataset.

A. *Study Area and Data Description*

The proposed outlier detection described in section III is applied to a realistic data set collected from 23 sensor nodes. These nodes are geographically distributed over Switzerland and Italian boarder, representing two clusters. The small cluster, situated in the Italian boarder, contains the five sensor nodes from which data set is obtained. Fig.2 illustrates the geographical distribution of these nodes over the area in which they are deployed. The collected data represent temperature as the attribute of interest. Temperature values are measured over a period 06:00–14:00 on day (30th September 2007) Fig.3 depicts a plot of temperature measurements sensors for all nodes in a small cluster (node25, node28, node29, node31, node32). The measurement tolerance of the deployed sensors is about ±0.3°C.

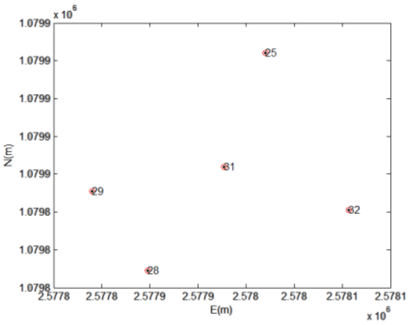

Fig. 2: a small cluster (consist of five nodes) of the Grand St Deployment and Their corresponding metric coordinates (E-N)

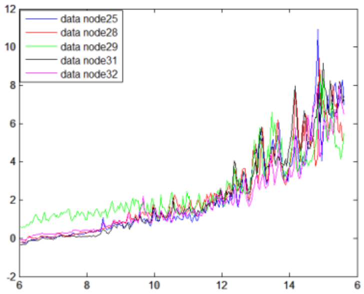

Fig .3: Represent data measurements of each sensor node

B. *Results and Performance Evaluation*

This section is devoted to evaluate the performance of the outlier detection technique proposed in section (III). Two performance metrics are considered. The first is the detection rate (DR) defined as the ratio of the correctly detected outliers to the total number of outliers in a given data set. Another performance metric of interest is the false positive alarm rate (FPR) which is defined as the ratio of normal data points incorrectly classified as outliers to the total number of normal data points. This section shows outliers in each node, detection rate, and false positive rate to each node. To evaluate performance of outlier detection needs a reference dataset. Usually, labeling techniques are utilized to label sensor measurements and classify each data point as normal pattern or anomalous. The choice of the labeling technique powerfully influences the evaluation of outlier detection techniques. There are three labeling techniques are used as in [14], i.e., running average-based, Mahalanonis distance-based, and density-based but our research used first one which fit to the data set as in [14].in this research two software are applied ,statistical model and fuzzy logic Simulink, are implemented by MATLAB. As in Fig.4 and Fig.5 spatial temporal outliers in univariate attribute (temperature) in both node25and node29, which detection rate in node25 is about 92% and FPR is 10.4%, while in node29 detection rate is 93.75% and high false positive rate is 18.33%.

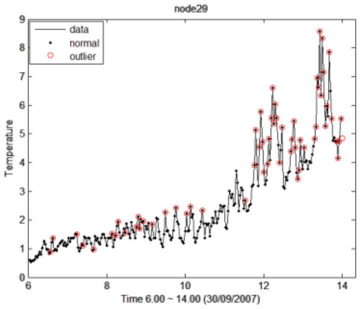

Fig.4.spatial temporal outliers in node25 detected by (STODM)

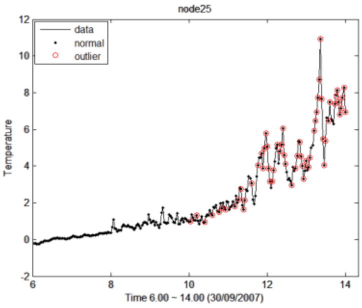

Fig.5.spatial temporal outliers in node25 detected by (STODM)

While in Fig.6, Fig.7and Fig.8 node28, node31 and node32, they have high detection rate 100% and FPR 9.16, 10, 4.5% respectively in each node.

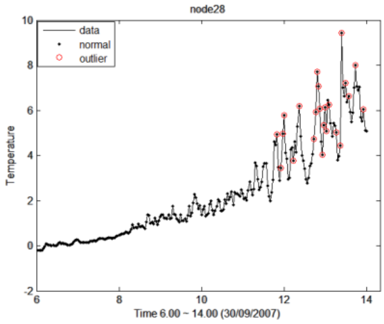

Fig.6. Spatial temporal outliers in node28 detected by (STODM)

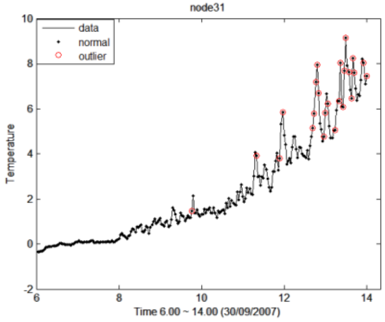

Fig.7. Spatial temporal outliers in node31 detected by (STODM)

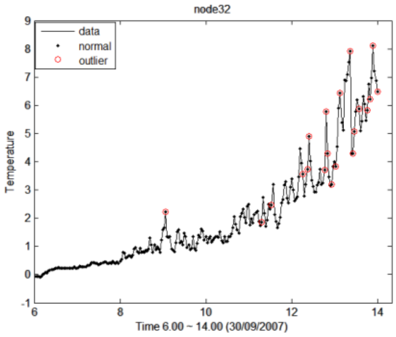

Fig.8. Spatial temporal outliers in node32 detected by (STODM)

Fig.9 shows the result of accuracy assessment for detected outliers by using pattern approach. The highest detection rate (100%) is at node (28, 31, 32) while the lowest detection rate (92%) is at node 25. The lowest amount of FPR is at node 32 (4.5%) while the highest rate is at node 29 (18.33%).

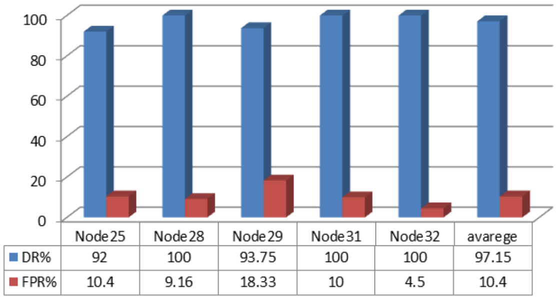

Fig.9. Accuracy of the detected outliers at different nodes.

Extensive ratio on the collected data set shows that both the detection rate and FPR increase when the threshold is decreased. A fixed threshold of on temporal similarly and mean of Euclidean distance of all nodes is computed as threshold of spatial similarity yields an average detection rate of 97.15% and FPR of 10.472%. The relative high FPR is a result of misclassifications of some normal observations while the high detection rate achieved is a result of considering spatial temporal similarly. Table.I show a comparison between the proposed storms with the most frequently used data labeling technique namely the TSOD and the STGOD technique with detection rate and false positive alarm achieved by each algorithm .it can be observed that the proposed algorithm outperforms these techniques in terms of detection rate. Both references models are applied to the same data set as considered in our model. Another advantage of the proposed technique is that it is able to distinguish between errors and events in a given data set obtained from sensor node. Classification of the outlier source is reported in Table.II

TABLE I. SHOWS A COMPARISON BETWEEN OUR APPROACH (STODM) AND STGOD MODEL PROPOSED OF RUNNING AVERAGE IN [14].

Method	DR%	FPR%
STODM	**97.15**	**10.4**
TSOD	**23.4**	**1.7**
STGOD	**72.34**	**10.94**

TABLE II: SHOWS NUMBER OF OUTLIERS AND EVENTS DETECTED AT DIFFERENT NODES USING STODM (OUR MODEL).

Nodes	No of outlier	No of event
Node25	**48**	**5**
Node28	**23**	**4**
Node29	**60**	**5**
Node31	**25**	**5**
Node32	**21**	**4**

(STODM)

V. CONCLUSION AND FUTURE WORK

STODM algorithm proposed in this paper combines the fuzzy logic theory and distance base similarity. Technique to detect outliers and is a new try in area of outlier detection for spatial temporal similarity .The proposed technique is able to identify normal and outlier data .moreover, error and event are also distinguished. High detection rate is achieved compare to conventional techniques while preserving low positive alarm rate and also reduce computational complexity because it uses Euclidian distance to calculate spatial similarity among neighboring nodes.

For future work, we plan to evaluate the algorithm performance on larger dataset. Our aim to build a technique takes multivariate data into account and considers dependencies among the attributes of the sensor data as well as spatial-temporal correlations that exist among the observations of neighboring sensor nodes.

REFERENCES

[1]. Y. Zhang, Nirvana Meratnia, Paul Havinga ,"Outlier Detection Techniques For Wireless Sensor Networks,",A Survey, University of Twente, P.O.Box 217 7500AE, Enschede, The Netherlands, 2010.

[2]. Chandola, V., Banerjee, A. and Kumar, V,"Outlier detection: a survey" ,Technical Report, University of Minnesota , 2007.

[3]. Vipnesh Jha, Om Veer Singh YadavOutlier,"Detection Techniques and Cleaning of Data for Wireless Sensor Networks" ,A Survey, International Journal of Computer Sci ence And Technology.K , 2012.

[4]. Luo X, Dong M, Huang Y,"On distributed fault-tolerant detection in wireless sensor networks", IEEE [10]. Rajasegarar, S., Leckie, C., Palaniswami, M. and Bezdek, J. C,"Quarter sphere based distributed anomaly detection in wireless sensor networks,"Proceedings of IEEE International Conference on Communications, pp. 3864-3869,2007.

[5]. Trans Computer55(1):58–70.R. Nicole, 2006.

[6]. H.Konak, Dilaver A. and Ozturk, E," The effects of observation plan and precision on the duration of outlier detection and fuzzy logic"2005, a real network application, Survey Review, 38, 298, 331-341, 2005.

[7]. Sensor Scope System. http://sensorscope.ep.ch/index.php/Main Page

[8]. S. Subramaniam, Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D,"Online outlier detection in sensor data using nonparametric Models", Seoul, Korea:, VLDB; pp. 187–198M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989, pp. 187–198M, 2006.

[9]. S.Rajasegarar , Leckie C, Palaniswami M, Bezdek JC," Distributed anomaly detection in wireless sensor networks", UK: IEEE, ICCS pp.12–16,2006.

[10]. Branch, J., Szymanski, B., Giannella, C. and Wolf, R ,"In-Network outlier detection in wireless sensor networks", Proceedings of IEEE ICDCS, 2006.

[11]. Y. Zhang, N. Meratnia, and P.J.M. Havinga,"An online outlier detection technique for wireless sensor networks", In Proceedings of the Third IEEE European Conference on Smart Sensing and Context (EuroSSC), pages 25-26,2008.

[12]. Y. Zhang, N. Meratnia, and P.J.M. Havinga,"Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks", In Proceedings of the IEEE 23rd International Conference on Advanced Information Networking and Applications Workshops/Symposia, pages 990-995,2009.

[13]. Mohamed MS, Kavitha T,"Outlier detection using support vector machine in wireless sensor network real time data", 2011,Int J Soft Comput Eng;1(2), 2011.

[14]. Y. Zhang, N.A.S. Hamm, N. Meratnia, A. Stein, M. van de Voort & P.J.M. Havinga," Statistics-based outlier detection for wireless sensor networks", International Journal of Geographical Information Science DOI:10.1080/13658816.2012.654493,2012.

[15]. A. Amidi a, N.A.S. Hamma, N. Meratnia b," wireless sensor networks and fusion of contextual information for weather outlier detection", International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol XL-1/W3, 2013.

[16]. Asmaa Fawzy , Hoda M.O. Mokhtar , Osman Hegazy ,"Outliers detection and classification in wireless sensor networks", Egyptian Informatics Journal 14, 157–164, 2013.

[17]. S. Syed, Cannon M.E,"Fuzzy logic based-map matching algorithm for vehicle navigation system", in urban canyons, ION National Technical Meeting, San Diego, CA, 26-28, 2004.

[18]. Yasemin Sisman1, Aslan Dilaver2 and Sebahattin Bektas1,"Outlier Detection in 3D Coordinate Transformation with Fuzzy Logic",,Acta Montanistica Slovaca Ročník 17, číslo 1, 1-8,2012.