# Regression in Wireless Sensor Networks

Muhammad Kashif Ghumman and Tauseef Jamal

imkashifghumman@gmail.com, tauseef.jamal@ulusofona.pt

**Abstract---***In WSN, the main purpose of regression is to locate the nodes by prediction on the basis of readings. This article explains the concept of regression according to WSN perspective and on the basic of these concepts the clustering of nodes through multi-linear regression originates by combing the ideas of locating the nodes through regression and how to utilize nodes parameters in multilinear regression formula.*

## 1. Introduction

A wireless sensor network (WSN) usually contains thousands of devices called sensors capable of computing data, sensing data, transmitting data or communicating with other sensors. These sensor nodes capabilities are increasing day by day. A sensor node consists of four basic components i.e. power unit, sensing unit, a computational unit with some memory and transmitting unit [1]. The basic function of the sensor nodes is to collect signals from their surroundings of any physical phenomenon like temperature, pressure, humidity, noise levels, and movement of physical objects in an environment and sends their observations to their base stations normally sink to convert these observations to valuable information.

In WSNs, where sensor nodes are responsible to react on physical phenomenon happening in the network and send their readings to the sink or to the base station. Several clustering protocols are presented earlier to manage nodes in a group on the basis of their locations in the field and main reason behind is to save energy of nodes from transmitting or receiving redundant data to the sink [2].

## 2. Regression

This section summarize concept of regression related to WSN field in which various protocols and schemes are discussed. In WSN, the main purpose of regression is to locate the nodes by prediction on the basis of readings this section clear the concept of regression according to WSN perspective and on the basic of these concepts the clustering of nodes through multi-linear regression originates by combing the ideas of locating the nodes through regression and how to

utilize nodes parameters in multilinear regression formula is explained.

## 2.1 Basic concept

Regression analysis is a statistical process for estimating the relationships among variables. Regression technique is basically used for predicting data in neural networks and machine learning fields for instance, in deep neural networks regression helps to reduce over fit data by using a technique called dropout [3] in which complex and larger network is mapped into smaller sub-sections exploring its hidden layers to reduce data. But in WSNs, this highly effective technique usage is at a medium level.

Regression is the measure of the relation between the mean value of one variable (e.g., output) and corresponding values of other variables (e.g., time and cost). Regression analysis includes many techniques for modeling and analyzing trend between a dependent variable and independent variable such as Linear Regression, Multi-linear regression, OLS (ordinary least square) Regression, Ridge Regression, LASSO (least absolute shrinkage and selection operator) Regression. In linear regression, we try to build a relationship between two variables using a straight line. We can say that regression analysis is a statistical process for estimating the relationships among variables and return to a former or less developed state of those variables.

## 2.2 Localization of sensor nodes via regression

In order to locate nodes in the field, traditional method involves RSSI (Received Signal Strength Indication) localization but the degree of inaccuracy made it difficult to fit in WSN environment. RSSI localization technique uses radio frequency signals (RF) of the sensor nodes to show their location across the field but exact location cannot be obtained due to the presence of noise which can distort signals. By using regression along with RSSI we can overcome accuracy problem effectively. Nodes whose geographical location are known, we can access the distance of these nodes by RF signal and then calculate location of these nodes by RSSI technique and map them on the field while the nodes whose location are not known or the RF signals are too weak we estimate their unknown nodes location using regression line which is the basic concept of regression. The purpose of efficient Wireless sensor networks is to inference environment accurately by using limited capabilities such as power and bandwidth. In [4] it is explained how to

involve regression in WSN by using nodes readings. In this work, the author uses regression for field estimation instead of early Gaussian field method to divide the space which can be used in various WSN applications. The sensors are deployed on the plane surface and these sensors form a topology based on point to point communication and sense information around its surroundings such as temperature and then stores them at a central point or node they called this node as local kernel point and after all nodes saves their readings at their local kernel points the regression line is applied on the nodes and retrieve that information. This work is not so efficient because they are not mentioning a detailed way to form a topology and communication is only happening between neighbors not the whole group also they use only sensor coordinates information for localization while our proposed technique includes sensors readings along with the sensor coordinates for the formation of groups [5]. Location based consumer applications like Global positioning system in mobiles [8] open a new way to tackle energy efficiency problems in WSN. In this paper, the author introduces a new scheme called "LiReCoFuL" [15] to locate nodes using regression.

Nodes which are present in the network and their locations are already known are divided into three type of categories active nodes which are sensing and processing data, anchor nodes which are not in the sleep state yet and also not doing any work and last target nodes which can be in sleep or awake state but their locations are not yet determined so, the scheme works in a way that sensor nodes are scattered in a house at different locations on the floor and the nodes which are active nodes they are not bothered to take part in locating the target nodes but the anchor nodes are nominated to take part in locating the target nodes. Linear regression function takes target nodes as a dependent variable and anchor nodes as independent variables so with the help of previous data communication history the linear regression function predicts the location of dependent variables on the basis of statics analysis of anchor nodes. The energy level of active nodes didn’t decrease abnormally because they are not using their energy to take part in regression. The main drawback of this work is that if target nodes are predicted on the basis of all nodes except anchor nodes then the precision will be high also, only anchor nodes utilize more energy than other nodes so this methodology is more resource hunger then previous

localizing techniques like Bayesian estimators [7] or maximum likelihood estimators [20] but regression is not involved in them [15]. In a dense wireless sensor network where nodes are close enough to send similar sensed data to their sinks can be highly cost-effective in terms of energy efficiency. So, this highly correlated data can be reduced to save data transmission and increase network lifetime. In this paper, the author exploits two types of correlation spatial and temporal correlation between sensor nodes, spatial correlation focuses on the location of the nodes such as the ones who are close to each other normally sends same type of readings to sink while in temporal correlation we adjust the time frame and notices the readings of the nodes and on the basis of historic data of nodes from time to time we can predict the future readings to come and reduce data transmission [9]. In this paper, a method was proposed to reduce data communication using linear regression as it is considered to be the major reason behind the energy efficiency crises of the WSN. Two modes of nodes were introduced for the sensor nodes, periodically sampling mode and the compressed sampling mode if one node in a group is in periodical sample mode it means all other nodes are in the other mode and can't stores information but the node which is in the first mode starts sensing the information stores its reading for some time and then change its mode and give opportunity to other node in a group so after all nodes stored the sensed information, an algorithm is introduced to predict the future readings of the nodes using regression model and save energy of the nodes by keeping them in sleep state more frequently. This paper detailed the concept of correlation and we will apply spatial correlation in duty-cycling to reduce the number of transmissions [16].

## 3. Analysis of Related Work

### 3.1 Linear Regression

Linear regression uses only one independent variable and only one dependent variable and by using independent variable values in the linear regression formula we can predict the dependent values [14]. This paper presents an algorithm which uses linear regression to divide nodes on the basis of their geographical location iteration-wise and made clusters of that nodes then again did linear regression in the sub-clusters and made even smaller and better clusters until a delta criteria meets. First, the nodes which are present in an area, the location of nodes are pre-determined in the form of their x and

y-coordinates and then put these coordinates' values in the linear regression formula to draw a regression line which crosses between the nodes. This line shows the shortest best possible line which tells the shortest distance across all nodes in the network.

Nodes are being divided in a way that if the node is on the upper side of regression line then it is in upper cluster otherwise in the lower cluster then from both upper and lower group similarity index between the two clusters is measured. In similarity index, the nodes in a specific group measure their distance with its center of that group and calculate its mean similarly the 2nd cluster also measure its mean values with respect to its center after that variance formula is applied in which these mean calculations are added and then divided by the distance between centers of clusters. Similarity index of the clusters shows that in which extend the nodes variate its position with respect to its center if the nodes are sparsely placed then the index value will be higher and vice versa. After that step, these two clusters again undergo in a regression process and are further subdivided into two more sub-clusters on the basis of their location at this stage, there are four sub-clusters the two newly formed sub-clusters again goes under the process of similarity index. Now the similarity index of this iteration and the first iteration is compared if by subtraction the answer didn't show zero or negative then again regression face is repeated in these sub-clusters to further subdivide the nodes and after n iterations the answer comes zero the regression process is halted and groups are made of nodes which are associated with each cluster at that time.

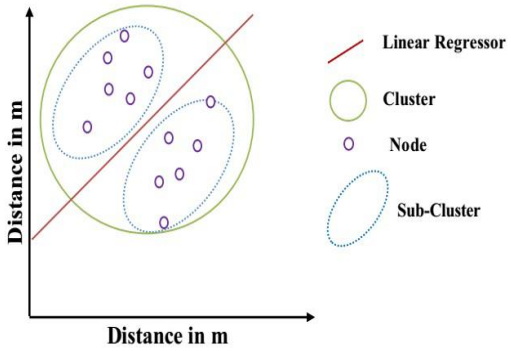


Figure 1: Linear regression showing best possible line on the basis of nodes [12].

This paper proves to be the base groundwork for my work I choose different standards such as type of regression and other criteria but the basic idea of group formulation is similar in a way. After clusters formation, the author compares the clusters made with regression with benchmark methods like k-mean clustering [6] and hierarchical clustering method [13]

and the results were far better than these previous techniques [12].

### 3.2 Multiple Linear Regressions

Multiple linear regressions create the relationship between one dependent variable and two or more independent variables through which dependent variable can be predicted. Simple linear regression can easily be used to reduce data communication overhead by reducing the number to transmissions because we can predict them but too much effort is not applied to check the accuracy of the prediction. This work was different from others because they exploit the multivariate correlation between different readings of the same node on the basis of time difference. Each sensor node can have multiple sensors like light, humidity or temperature values to be sensed by the same node in the network, sensor nodes stores the fixed number of readings of different sensors until they reach a certain threshold, the author appoints each sensor a different variable e.g. light sensor readings of one node will be saved in one variable from time to time similarly temperature and humidity level is also stored because multi-linear regression can take more independent variables we can add number of sensors to the list then the sensor node calculates $\beta$ and $\alpha$ values of each variable in this case 3 variables are considered which means 3 different types of sensors on same node is considered by the author and broadcast it to neighboring nodes, if neighbor nodes have the same $\beta$ and $\alpha$ values then only one node will send to sink because same values means that different nodes measure same measurements and by stop sending repeated $\alpha$ and $\beta$ values to the sink means that we use multivariate correlation to send fewer data and efficiency is increased. This paper makes it possible to learn how we can use node values in multi-linear regression function and integrate into my work and it helps a lot to clear the multi-linear regression concept [11].

### 3.3 LASSO Regression

Least absolute shrinkage and selection operator is the most advanced form of regression in which data prediction is more accurate and it originates from the least square regression instead of choosing all the variables and predict the data, LASSO accepts multiple independent variables for its function and shrunk some of them to zero resulting in less independent variables that are used to predict dependent variables.

In this paper the author first maintains communities of the nodes based on the

similar reading such that the nodes which send similar readings to the sink will form themselves into community structure and then restrict fewer nodes to send data while other remains in the sleep state. Lasso regression is done on the sink when nodes send more than one types of measurements such as temperature, pressure or humidity level to the sink Lasso regression helps to aggregate the data to compress and send to the base station for maximum energy efficiency. Moreover, paper present device management technique in which if one node goes dead before sending its value to sink then data from its neighboring nodes can be predicted by regression [10].

## 4. Conclusions

Regression analysis is a statistical process for estimating the relationships among variables. This paper introduced the concept of regression into WSN. It provided the analysis of related work and how to use node values in multi-linear regression. As a future work we aim to use data aggregation to form the cluster and identify neighbors [17]. We also aim to use the multi-linear regression to measure the Selection Factor (SF) in relay based networks [18] and [19].

## References:


*[1] T. Jamal and SA Butt, Low-Energy Adaptive Clustering Hierarchy (LEACH) Enhancement for Military Security Operations, ISSN 2090-4304 Journal of Basic and Applied Scientific Research, 2017.*

*[2] T. Jamal, SA Butt, "Study of black hole attack in AODV", INTERNATIONAL JOURNAL OF FUTURE GENERATION COMMUNICATION AND NETWORKING, Issue 10, Vol 9, Sep 2017.*

*[3] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.*

*[4] Y. Wang, X. Xu, and X. Tao, "Localization in wireless sensor networks via support vector regression," in Genetic and Evolutionary Computing, 2009.*

*[5] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Regression in sensor networks: Training distributively with alternating projections," in Optics & Photonics 2005.*

*[6] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Systems with Applications, vol. 40, no. 1, pp. 200–210, 2013.*

*[7] G. Chandrasekaran, M. A. Ergin, J. Yang, S. Liu, Y. Chen, M. Gruteser, and R. P. Martin, "Empirical evaluation of the limits on localization using signal strength," in Sensor, Mesh and Ad Hoc Communications and Networks, 2009.*

*[8] T. Jamal and P. Mendes, "Cooperative relaying in user-centric networking under interference conditions," IEEE Communications Magazine, vol. 52, no. 12, pp. 18–24, dec 2014.*

*[9] Paulo Mendes,Waldir Moreira, Tauseef Jamal, Huseyin Haci, Huiling Zhu, "Cooperative Networking In User-Centric Wireless Networks", Springer Lecture Notes in Social Networks, User-Centric Networking: Future Perspectives, ISBN 978-3-319-05217-5, May 2014.*

*[10] T.-Y. Chuang, K.-C. Chen, and H. V. Poor, "Information centric sensor network management via community structure," IEEE Communications Letters, vol. 19, no. 5, pp. 767–770, 2015.*

*[11] C. G. N. De Carvalho, D. G. Gomes, J. N. De Souza, and N. Agoulmine, "Multiple linear regression to improve prediction accuracy in wsn data reduction," in Network Operations and Management Symposium (LANOMS), 2011.*

*[12] N. Hemavathi and S. Sudha, "A novel regression based clustering technique for wireless sensor networks," Wireless Personal Communications, vol. 88, no. 4, pp. 985–1013, 2016.*

*[13] R. Suzuki and H. Shimodaira, "Hierarchical clustering with p-values via multiscale bootstrap resampling," R package, 2013.*

*[14] F. Vanheel, J. Verhaevert, E. Laermans, I. Moerman, and P. Demeester, "Automated linear regression tools improve rssi wsn localization in multipath indoor environment," EURASIP Journal on Wireless Communications and Networking, vol. 2011, no. 1, p. 38, 2011.*

*[15] ——, "A linear regression based cost function for wsn localization," in Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on. IEEE, 2011, pp. 1–5.*

*[16] B. Zhang, Y. Liu, J. He, and Z. Zou, "An energy efficient sampling method through joint linear regression and compressive sensing," in Intelligent Control and Information Processing (ICICIP), 2013.*

*[17] A.Khan, A. Zameer, T. Jamal, and A. Raza, Deep Belief Networks Based Feature Generation and Regression for Predicting Wind Power, 2018, arXiv:1807.11682v1.*

*[18] T. Jamal and P. Mendes. Cooperative Relaying for Dynamci Networks. EU Patent, (EP13182366.8), August 2013.*

*[19] L. Lopes, T. Jamal & P. Mendes, "Towards Implementing Cooperative Relaying" In Technical Report COPE-TR-13-06, CopeLabs University Lusofona Portugal, Jan 2013.*