# Geolocation Prediction Using Multinomial Regression

Rohit Sharma[1], Paripath Inc

***Abstract*** **–Passive geolocation of communication emitters provides many applications to defense and noncombatant surveillance and security operations. Time Difference of Arrival (TDOA) measurement for stationary emitters may be obtained by sensors mounted on mobile platforms for Wireless Sensor Network (WSN). Complex Ambiguity Function (CAF) of received complex signals can be efficiently calculated to provide required TDOA measurement. TDOA measurements are nonlinear because the emitter uncertainty in the Cartesian domain is non-Gaussian. Moreover, non-line-of-sight (NLOS) signal reception hampers the accurate estimation of geolocation. This coupled with wireless noise (including crosstalk, impulse, and shot noise) makes the deterministic system perform poorly in the field. In this paper, we present a supervised learning methodology that can learn quickly from the simulation data to predict emitter location accurately in the face of changing conditions in the field.**

**Keywords- WSN, CAF, LOS, NLOS.**

## INTRODUCTION

In this paper, we share machine learning methodology for predicting geolocation of a sensor or a group of sensors in wireless sensor networks in the mixed line-of-sight (LOS) and non-line-of-sight (NLOS) environments. With TDOA measurement uncertainty, we face a non-linear estimation problem largely affected by environment and resulting communication loss. Localization is the task of determining the physical coordinates of a sensor node (or a group of sensor nodes) or the spatial relationships among objects. It comprises a set of techniques and mechanisms that allow a sensor to estimate its own location based on information gathered from the sensor's environment. While the Global Positioning System (GPS) is undoubtedly the most well-known location-sensing system, it is not accessible in all environments (e.g., indoors or under dense foliage) and may incur resource costs unacceptable for resource-constrained wireless sensor networks (WSNs). [1] [2]

In this work, we used professional propagation software simulation for extracting geospatial information from the simulation. The propagation software simulation is set up to predict the path loss and path delay between transmitter and receiver including all important parameters of the mobile radio channel. We propose nonlinear polynomial regression as a machine learning method to obtain excellent results on predicting emitter location. Rest of this paper is organized as follows: In the next section we review prior art in the field. Subsequent section defines the problem of geolocation in three-dimensional space with equations. Later sections discuss solutions and results thereof. We start solution section with introducing machine learning to geolocation. Next two sub sections will focus on dataset review and sources of error in measurements. Feature analysis, hypothesis and loss function has been discussed in the next two subsections. Subsequent subsections will discuss data conditioning, training machine learning neurons, generating and saving the model. We close solution section by highlighting model generation and use flow in the software. We present results of our proposed machine learning methodology in the next section. Discussion is closed with a conclusion on our findings and future direction of this work.

### Prior Art

Machine learning is fast becoming ubiquitous in localization research with deep interest from both academia and industry. Several systems are being suggested based on a variety of technologies. A major weakness of many of these systems—such as RFID [5], [6], infrared [3], and ultrasound [4]—is that they require devoted hardware components such as sensors along with considerable infrastructure changes. As a result, these systems require substantial capital investment for deployment in the field.

To avoid significatnt infrastructure changes in wireless network, tremendous amount of research has put into to developing localization systems that require little to no infrastructure change using standards such as Wi-Fi signal strengths [9], [10], [11], [12] and Bluetooth [7], [8] with considerable success. The systems developed using these standards including Wi-Fi signal strengths based on localization is yet to show signs of maturity for being deployed in commercial grade application. Research in this area is mainly focused in two areas: localization using fingerprinting (i.e. by measuring the parameters of the received signal) and localization using deterministic algorithms such as Radio Wave Propagation Model to determine a device's geolocation from the access points in range. Solution to determining an emitter's geolocation has two parts. data collection and data usage. Data collection includes placement of emitters and sensors, choosing parameter measurement type (path delay, frequency, signal strength etc.), whereas data usage includes using adapting measurements

[1] 105 Serra Way, #251 Milpitas CA 95035 USA
Email: rohit@paripath.com, tel: +1-408-372-7405

dataset to predict geolocation of emitter based on. models.

Statistics has been to used find geolocation or objects using signal measurement. First dedicated work was reported in bearings only emitter localization. [13]. In the modern day and age, statistical methods have been combined with linear algebra and computer science to form a new branch called machine learning.

Time Difference of Arrival (TDOA) method has been used as one of the parameters for determining geolocation using fingerprint approach [14], [15]. Using TDOA measurements as one of the parameters is well-matched for predicting the geolocation of high-bandwidth emitters, e.g. radars with tracking algorithm as shown in [16], [17] for monitoring geolocation.

**Problem Statement**

The time difference of arrival method for emitter localization is based on measuring the TDOA for the emitted radar pulses at two sensors, see Figure 1.

$$TDOA = \frac{1}{c}(d_1 - d_2) \qquad \ldots (1)$$

Where, $d_1$ and $d_2$ are distance from emitter E to sensor $S_1$ and sensor $S_2$ respectively and c is speed of light.

$$d_1 = \sqrt[2]{(x_e - x_1)^2 + (y_e - y_1)^2 + (z_e - z_1)^2}$$

$$d_2 = \sqrt[2]{(x_e - x_2)^2 + (y_e - y_2)^2 + (z_e - z_2)^2} \quad \ldots (2)$$

Based on Equation (1) possible emitter positions can be determined when the time difference of arrival is known (measured). Equation (1) does not give a unique position for the emitter since the equation contains three unknowns ($x_e$, $y_e$, $z_e$). In the 3-dimensional case, the possible emitter positions are located on a circular hyperboloid with the sensors at the focal points. In the 2-dimensional case where the sensors and the emitter are in the same plane, the possible emitter positions are located along a hyperbola with the sensors at the focal points.

**Use of Machine Learning**

Machine Learning is an interdisciplinary field of computer science and applied mathematics, which relies on developing a hypothesis of creating the model (as opposed to an algorithm in computer science or methods/formulae in mathematics) and tries to improve it by fitting more data into the model over time. Since we are predicting geolocation of the emitter, regression is a clear choice for prediction. Regression is a class of supervised algorithm that attempts to establish a continuous relationship between a set of dependent variables (geolocation coordinates) and set of other independent variables (signal path delay and signal strength). We develop our hypothesis based on the knowledge of the relationship between dependent and independent variables in TDOA including NLOS identification, as discussed in the last section.

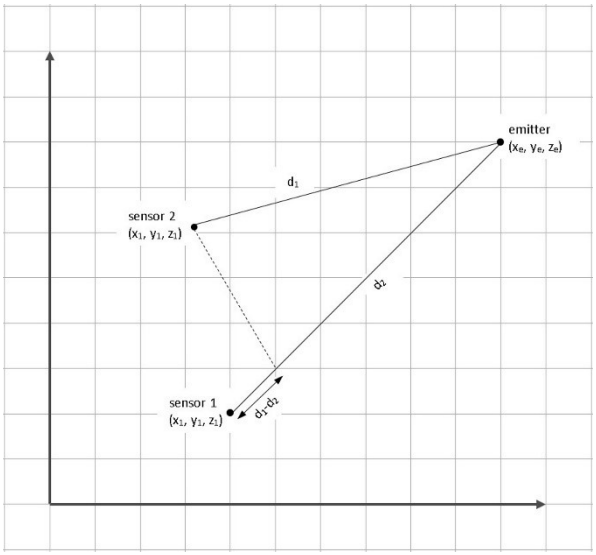


*Figure 1: TDOA Principle*

**Simulation Dataset**

We acquired simulation data from Propagation software package in an ASCII format with point and path statements containing characteristics of the radio channel. The statement point is geolocation coordinate of emitter location. First two fields in the second statement path are a delay of the path (in nanoseconds) and field strength (in dBμV/m). We extracted channel's impulse response from this data to create a new dataset for machine learning framework. The extracted dataset is shown in the table below:

| Delay (ns) | Strength (dBμV/m) | X | Y | Z |
|---|---|---|---|---|
| 2349.565 | 58.71 | 396.0 | 580.0 | 515.46 |
| 926.787 | 74.92 | 736.0 | 980.0 | 515.06 |
| … | … | … | … | … |
| … | … | … | … | … |
| 808.82 | 76.04 | 756.0 | 980.0 | 517.05 |

*Figure 2: Dataset from Channel's Impulse Response*

**Error Sources**

In practice, the estimated distances are not equal to the true distances, because of a number of effects including thermal noise, multipath propagation, interference, and ranging algorithm inaccuracies. Additionally, the direct path between requester and responder may be obstructed, leading to NLOS propagation. In NLOS conditions, the direct signal, either suffers from noticeable signal attenuation due to through material propagation or completely blocked. In case, where signal attenuation is significant source,

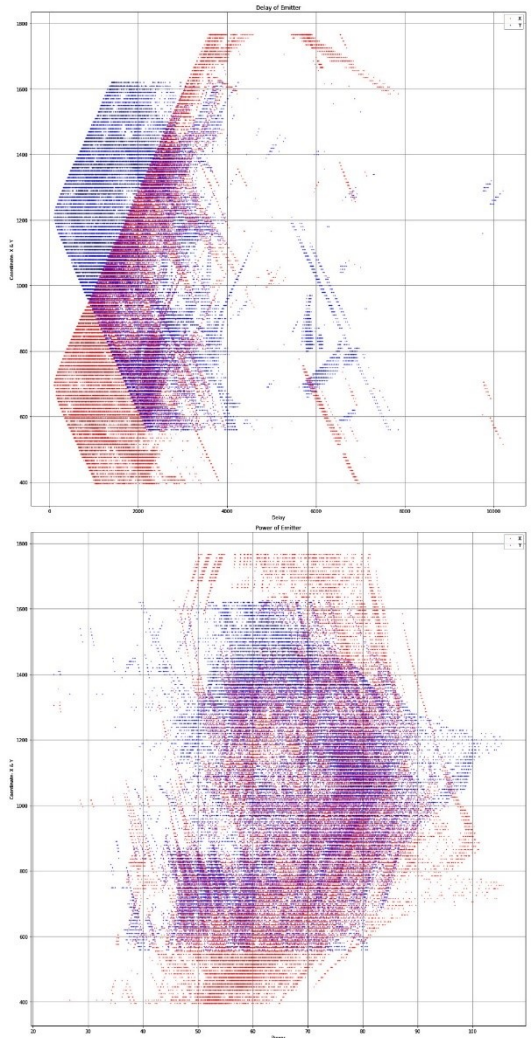


*Figure 3: Plot Visualization for Path Delay and Field Strength*

measurement will falsely reflect large path distance because of reduced propagation speed (i.e., less than the speed of light, c). In cases, where signal is received after multiple reflection, the distance estimate is also more than the actual distance, as it accounts for the distance of a reflected path.

**Feature Analysis**

Once the data was prepared, several data analysis and data visualization techniques were used as a basis for experimentation to come up with a suitable regression model. Parallel Plot is a popular visualization technique used to plot individual data elements across many dimensions. This gives us a sense of the relationship between path delay, field strength and their coordinate. Parallel plots for path-distance vs XY coordinates and field strengths vs XY coordinates are shown in Figure 3. Nature of mixed line-of-sight and the non-line-of-sight environment in the datasets are self-evident since a single and fixed XY co-ordinate shows multiple delay paths captured at the sensor. Minimum distance for a fixed XY pair manifests direct line of sight reception, whereas rest distance more than minimum is attributable to non-line of sight reception.

We used variety of techniques including parallel plots (shown in figure 3), three-dimensional surface plots, Euclidean distance plot, and power-delay profile of the dataset to find meaningful relationships representing nature of NLOS for our hypothesis.

**Hypothesis and Loss Function**

A typical machine regression algorithm has two components: Hypothesis (aka Model) and Loss function (aka cost function). From the simulated dataset, we've three dependent and two independent variables. Based on our knowledge of relationships among dependent variables (path delay and a field strength of signals) and independent variables (geolocation co-ordinates X, Y, Z), we chose the following hypothesis for our machine to train on.

$$X_p = \sqrt[2]{W_x * D^2 + FSW_x(FS^2 - FS)} + k_x$$

$$Y_p = \sqrt[2]{W_y * D^2 + FSW_y(FS^2 - FS)} + k_y$$

$$Z_p = \sqrt[2]{W_z * D^2 + FSW_z(FS^2 - FS)} + k_z$$

Here $D$ is measures path delay and $FS$ is measured field strength of the signal. $W_x$, $W_y$, and $W_z$ are weights of delay parameter in the hypothesis and $FSW_x$, $FSW_y$, and $FSW_z$ are weights of field strength parameter in the direction of X, Y and Z axis respectively. Last three parameters $k_x$, $k_y$, and $k_z$ are intercepts of liner hypothesis in the direction of X, Y, and Z respectively. This gives us a total of nine parameters to tune. In the experimental phase of this work, we investigated 7 different models up to thirty parameters. None of that permutation was any statistically better or worse permitting random noise of order during the training phase.

A loss function or cost function is a function that maps an input data values of one or more variables onto a real number representing loss/cost with associated event. While there are numerous regression loss functions, we focus on the mean squares (MS) function, due to its simplicity and because it makes no assumptions regarding ranging errors.

$$MSE_x = \frac{1}{n}\sum_{i=1}^{n}\left(X_{pred} - X_{sim}\right)^2$$

$$MSE_y = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{pred} - Y_{sim}\right)^2$$

$$MSE_z = \frac{1}{n}\sum_{i=1}^{n}\left(Z_{pred} - Z_{sim}\right)^2$$

$$Loss = \sqrt[2]{MSE_x + MSE_y + MSE_z}$$

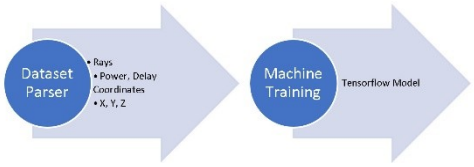


*Figure 4: Model Generation Flow*

Mean square error measures the average of the squares of the errors or deviations—that is, the difference between the predicted geolocation and what is measured during simulation. Next three expressions are mean square error functions of coordinates X, Y, and Z respectively. Fourth expression is the loss function represented as a function of error of first three mean square error functions. Loss function is used during training to minimize overall loss as a mean square error.

**Data Conditioning**

Our simulated dataset consists of over 46,000 samples. We noticed the range of the dataset values was orders of magnitude away. In particular, delay values were under 100 to over 11,000 with a spread of over 1,000. We normalized our dataset using well-known z-scores.

$$z - score = \frac{data - \mu}{\sigma}$$

Here μ and σ are mean and standard deviation of each variable in the dataset. To get ready for the next stage, we randomized simulation dataset to improve the efficiency of our training and to reflect field environment. Split between the training set and test set was chosen at 80% and 20% respectively. A total number of training samples was over 37,000.

**Training and Model Generation**

Compute graph was created with the hypotheses outlined in the previous section. Compute graph consists of with 8 neurons for every coordinate (X, Y, Z) of geolocation with a total of 24 neurons. Once compute graph was ready, we used well-known optimization algorithm gradient descent to optimize loss function described in the previous section. Learning rate is one of the critical hyperparameters to achieve acceptable accuracy. It was fixed at 0.1. Resulting learning network was trained for thousand epochs.

**Using the model**

Once the model was trained, all nine model parameters were saved. These parameters were used in creating a stand-alone model ready for in-house testing and subsequent deployment in the field.

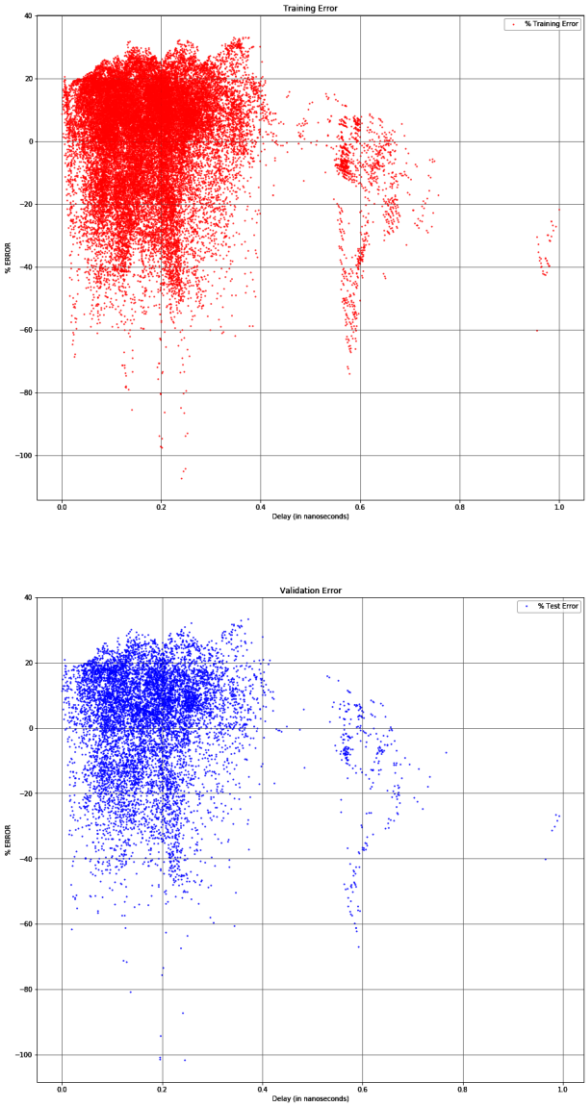


*Figure 5: Percentage Error on Training and Test Dataset*

Testing was performed using this model with the 20% of the dataset set aside during data conditioning phase.

**Results**

As stated earlier, we trained our model for 1000 epochs to optimize our loss function giving us training accuracy of under 2% as shown below:

Training dataset error mean= 1.303 %

The same model was used on the test dataset. And accuracy was found within the same range shown below:

Test dataset error mean= 1.322%

Figure 5 shows charts for the training set and test set separately. The trend in both this chart reflects that data was randomized in a statistically significant manner to show that model is able to generalize well over a large dataset of over 46,000 samples.

## Conclusion

Because the NLOS propagation is ubiquitous in both indoor and outdoor positioning scenarios, a robust algorithm is required to account for the NLOS location estimation error. In this work, we shared a machine learning regression model to minimize NLOS location estimation error for the improvement of location accuracy amid changing field conditions. This model is quick to train and deploy in the field.

## References


[1] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996–2018, 2014.

[2] W. Dargie and C. Poellabauer, Localization. John Wiley & Sons, Ltd, 2010, pp. 249–266

[3] B. Sohn, J. Lee, H. Chae, and W. Yu, "Localization system for a mobile robot using wireless communication with ir landmark," in Proceedings of the 1st International Conference on Robot Communication and Coordination, ser. RoboComm '07. Piscataway, NJ, USA: IEEE Press, 2007, pp. 6:1–6:6. [Online]. Available: http://dl.acm.org/citation.cfm?id=1377868.1377876

[4] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, ser. MobiCom '00. New York, NY, USA: ACM, 2000, pp. 32–43. [Online]. Available: http://doi.acm.org/10.1145/345910.345917

[5] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose, "Mapping and localization with RFID technology," in Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on, vol. 1, April 2004, pp. 1015–1020.

[6] L. Ni, Y. Liu, Y. C. Lau, and A. Patil, "Landmarc: indoor location sensing using active RFID," in Pervasive Computing and Communications, 2003. (PerCom 2003). Proceedings of the First IEEE International Conference on, March 2003, pp. 407–415.

[7] L. Aalto, N. Gothlin, J. Korhonen, and T. Ojala, "Bluetooth and wap ¨ push based location-aware mobile advertising system," in Proceedings of the 2Nd International Conference on Mobile Systems, Applications, and Services, ser. MobiSys '04. New York, NY, USA: ACM, 2004, pp. 49–58. [Online]. Available: http://doi.acm.org/10.1145/990064.990073

[8] R. Bruno and F. Delmastro, "Design and analysis of a bluetooth-based indoor localization system," in Personal Wireless Communications, ser. Lecture Notes in Computer Science, M. Conti, S. Giordano, E. Gregori, and S. Olariu, Eds. Springer Berlin Heidelberg, 2003, vol. 2775, pp. 711–725. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-39867-766

[9] P. Bahl and V. N. Padmanabhan, "Radar: an in-building rfbased user location and tracking system." Institute of Electrical and Electronics Engineers, Inc., March 2000. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=68671

[10] A. Goswami, L. E. Ortiz, and S. R. Das, "Wigem: A learning-based approach for indoor localization," in Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies, ser. CoNEXT '11. New York, NY, USA: ACM, 2011, pp. 3:1–3:12. [Online]. Available: http://doi.acm.org/10.1145/2079296.2079299

[11] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy, "Precise indoor localization using smart phones," in Proceedings of the International Conference on Multimedia, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 787–790. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874078

[12] M. Youssef and A. Agrawala, "The horus wlan location determination system," in Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, ser. MobiSys '05. New York, NY, USA: ACM, 2005, pp. 205–218. [Online]. Available: http://doi.acm.org/10.1145/1067170.1067193

[13] R. Stansfield, "Statistical theory of DF fixing," Journal of IEE, vol. 94, no. 15, pp. 762–770, 1947.

[14] M. Schmidt, "A new approach to geometry of range difference location," IEEE Trans. Aerospace and Electronic Systems, vol. 8, no. 6, pp. 821– 835, November 1972.

[15] K. Ho and Y. Chan, "Solution and performance analysis of geolocation by tdoa," IEEE Trans. Aerospace and Electronic Systems, vol. 29, no. 4, pp. 1311–1322, October 1993.

[16] N. Okello and D. Musicki, "Measurement association and tracking ˇ for emitter localisation using paired UAVs," Journal of Advances in Information Fusion (submitted).

[17] F. Fletcher, B. Ristic, and D. Musicki, "TDOA measurements from two ˇUAVs," in 10th International Conference on Information Fusion, Fusion 2007, Quebec, Canada, July 2007.