Development of Machine Learning-Based Radio Propagation Models and Benchmarking for Mobile Networks

Seoyoon Chang[1] and Aishwarya Baliga[2]

[1]Seisen International School
[2]Advisor, MOTiV Research co.

ABSTRACT

Path loss prediction plays an important role in assisting network operators to build quality mobile networks by indicating optimal coverage and interference performances. Common empirical models such as the Cost 231-Hata model and Cost 231-Walfisch Ikegami model have been used to predict path loss due to their simplicity and robustness; however, these models do not provide reliable accuracy for small coverage infrastructures, especially in LTE and 5G networks. Therefore, as an alternative method, deterministic models (e.g., ray-tracing models) are widely used for small coverage applications–however, these also require precise 3D digital maps, street structures, and building surface reflection indices, etc. which can be costly and need frequent maintenance. In this paper, an alternative model based on machine learning algorithms is proposed and the algorithms' prediction accuracies are evaluated. The machine learning approach trains models using signal strength data upon which they draw generalizations. The performance of these models is compared to that of the empirical models, and the results reveal that the machine learning models perform significantly better, with Random Forest outperforming other algorithms.

Introduction

Mobile networks are social infrastructures that are indispensable in our daily lives. Whether it be texting or uploading photos onto social media from smartphones, the data is sent through the mobile network system rather than the recipient. The current mobile network has been evolved up to the 5th generation since it has first emerged as an analog communication system in the early 1980s. During the 3rd generation in the early 2000s, analog mobile networks transformed into a digital communication system that allowed for much faster data speed through wider spectrum bandwidths. Additionally, it became possible to engage in a video call whilst talking to a receiver, even though this service was only available in areas with good coverage. During the 4th generation system, which is also referred to as LTE (Long Term Evolution), data speed became further enhanced and became a source wherein numerous multimedia services and platforms were invented and have since prevailed, such as YouTube, Facebook, Instagram, Uber, etc. [1].

Mobile phone users, hereinafter referred to as receiver(s) or UE(s) (User Equipment), can access the internet or make calls when they are within a coverage area provided by a mobile network. Coverage area can be defined as the area where radio signals higher than a certain threshold are transmitted by base stations and can be received without undergoing significant loss. Mobile networks are comprised of multiple base stations that transmit radio signals in order to communicate with UEs. Thus, it can be thought of as the base stations transmitting large radio signals so that the signal strength at any given location is higher than the threshold. However, it is important to note that the coverage area is not only defined by the signal strength but also by the signal-to-noise ratio, which is denoted as SINR [2]. SINR is the ratio of the base station's signal strength that is communicating with the UE's device in relation to the sum of all of the other surrounding bast stations' signal strengths plus the background noise. Therefore, if all of the base

stations transmit radio signals at maximum power, SINR would be degraded rather than increased. The SINR is a significant performance indicator when it comes to measuring the quality of mobile networks as the aim of mobile network operators is to maintain or improve their networks.

The prediction of coverage is a process that calculates the radio signal strength transmitted by each base station. The definition of the radio signal strength differs from mobile network systems, such as RSCP (Reference Signal Code Power) in 3rd generation mobile networks and RSRP (Reference Signal Received Power) in LTE and 5G mobile networks [3]. In this paper, RSRP is used as the signal strength as most of the current mobile network systems operate under the 4th generation.

In order to predict the signal strength, Maxwell's equations need to be considered because the radio signals transmitted via the mobile networks are traveling electromagnetic waves [4]. Despite this, it is virtually impossible to take all the factors affecting the electromagnetic waves in the real world into account. As a result, researchers have investigated methods to simplify signal strength calculation problems by modeling the radio propagation characteristics on earth, from which arose several works such as the Cost 231-Hata model [5] and Walfisch-Ikegami propagation models [6]. These models are well known as representative empirical models that predict signal strength. The advantages of these propagation models include the straightforward prediction of the signal strength; on the other hand, their flaw is that they have less accuracy, especially when it comes to predicting the signal strength in areas closer to base stations, such as within 1 to 2 km in distance. Most of the coverage areas of a single base station in a 4G mobile network are less than this distance, so a new model that can provide better accuracy while maintaining usage simplicity is needed.

The prevailing method used to understand the actual mobile network's coverage is conducted on the streets using signal measurement equipment, such as with RF (Radio Frequency) scanners and special test phones. This activity is called drive tests and is the only way to collect actual network coverage status so far, but not without high costs of resources and time. As a result, many operators have sought for an alternative solution that can replace the drive tests and have arrived at the propagation model. As mentioned earlier, the empirical propagation models do not have comparable accuracy compared to drive tests and deterministic models. For example, ray-tracing models can become very complex and require 3-dimensional building map data. However, since the advent of LTE, coverage prediction methods have undergone a paradigm shift. To minimize usage of time and resources, mobile operators have started to utilize mobile applications as a resource for collecting coverage data. This solution, called crowd source data, has allowed for coverage data collection to become much more convenient and environmentally friendly [7]. Despite the benefits, the key issue remains: how can one convert the actual signal strength data into a propagation model that is simple in terms of use and provides accurate signal strength predictions?

To go about the aforementioned issue, several machine learning-based signal strength prediction models utilizing actual signal strength data have been proposed in this paper. As crowd source data of large areas is usually confidential information belonging to network operators and is also difficult to obtain, drive test data collected in relatively small areas is used instead for training the machine learning models. Since both crowd source data and drive test data are based on similar formats, there is no loss of generality even if the drive test data is used to develop the machine learning-based prediction model.

The rest of this paper is organized as follows: introduction of empirical models, machine learning-based signal strength prediction models, the comparison of all models using actual drive test data, and the evaluation of the models.

Empirical Propagation Models

Computing the path loss and received signal strengths (RSRP) at given locations can have significant implications for network operators; in order to understand the predicted values, though, one must understand the underlying mechanisms of mobile networks. When signals travel from the serving cell (a base station serving the user) to a UE, they are susceptible to interference and barriers that impede propagation. Often regarded as path loss, or attenuation, the

received signal strength at a given location can reduce due to factors such as the aforementioned impediments, as well as operating frequencies, base station and antenna height, and whether the signal travels indoors or outdoors. Moreover, path loss can occur over long distances due to constructive and destructive interference, as well as due to obstacles such as buildings that cause scattering and reflections. Various models have been established to model path loss, with one of the simplest models being free-space path loss. In free-space path loss, signals propagate through free space, traveling in a line-of-space (LOS) path directly from the base station to the receiver [8].

$$PL_{FS} = 20log_{10}(d) + 20log_{10}(f) + 32.44$$

Equation 1 : Free space path loss.

Where,
PL_{FS}: Path loss in free space (dB)
d: Distance between the transmitter and receiver (km)
f: Frequency in MHz.

Yet in reality, the transmission of signals relies on various factors such as the environment, wherein the area can be categorized as urban, suburban or rural, as well as the base station and antenna heights. In order to account for these characteristics, various empirical models have been established to formulate mathematical equations that calculate RSRP at given locations under certain conditions. While there are many traditional empirical models that work for path loss calculations, this paper focuses on two widely-used models used to analyze coverage in LTE networks: the Cost 231-Hata model and the Walfisch-Ikegami model.

The features used for these empirical models have been extracted from multiple datasets, which are initially separated into two datasets: base station data and antenna data. The features selected from the former include the base station's longitude and latitude coordinates, the carrier frequency–the frequency used by the base station's antenna for transmission–and the PCI (Physical Cell Identity)–unique identifiers for the base stations and transmission power [9]. The latter's features are comprised of the antenna's azimuth–the horizontally oriented angle of the antenna equipped at the base station, antenna tilting angle, antenna height and antenna's feeding loss. Using these features, the signal propagation can be calculated by the two empirical models in the following equations.

Empirical Model: Cost 231-Hata Model

$$PL = 46.3 + 33.9log_{10}(f) - 13.82log_{10}(h_b) - ah_r + (44.9 - 6.55log_{10}(h_b)log_{10}d + C$$

Equation 2 [5]: Cost-231 Hata model.

Where,
PL: Path loss (dB)
f: Frequency in MHz
h_b: Base station antenna height above the ground (m)
h_r: Receiver antenna height (m)
d: Distance between the base station and receiver (km)
C: constant (urban are = 3dB)

On the basis of the above information, the Cost 231-Hata model is applicable for systems with operating frequencies ranging from 1500 to 2000 MHz. It is required that the base station height is larger than that of the surrounding

buildings, and the receiver antenna height is assumed as 1.5 meters. The variable d is restricted to distances ranging from 1 to 20 km. Also, the constant C becomes 3 dB as this model is tested for urban areas in Japan. Similarly, the constant a, which taken together with the receiver's antenna height (h_r) represents a correction factor for the mobile antenna height based on the size of the coverage area, is defined in an urban environment as such:

$$a(h_R, f) = 3.2(log_{10}(11.75h_R))^2 - 4.97$$

Equation 3 : Correction factor for the mobile antenna height.

Empirical model: Cost 231-Walfisch-Ikegami Model

Based on models introduced by Walfisch and Bertoni, and Ikegami et al, the Cost 231-Walfisch-Ikegami model takes path attenuation into account, such as rooftop to street diffraction, scatter loss, and multiscreen loss [6]. Based on the case of LOS, the formula for the model can be expressed as Equation 4.

$$PL = 42.6 + 26log_{10}(d) + 20log_{10}(f)$$

Equation 4 : LOS path loss equation

Where,
PL: Path loss (dB)
d: Distance between the transmitter and receiver (km)
f: Frequency in MHz

The distance R is often restricted to values between 0.02 and 5 km.
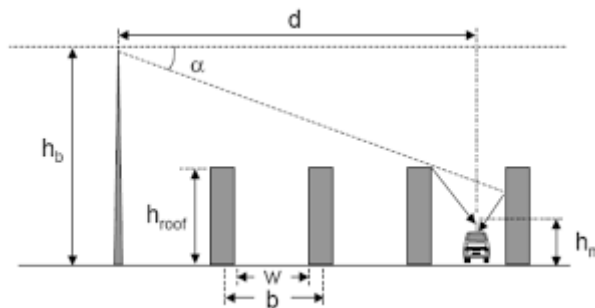
Figure 1. Parameters for Cost 231-Walfisch Ikegami model

In addition to LOS (Line Of Sight), the Walfisch-Ikegami model also takes into account various scattering and diffraction properties due to surrounding rooftops and building.

$$L_{rts} = -16.9 - 10log_{10}(w) + 10log_{10}(f) + 20log_{10}(\Delta h_{mobile}) + L_{ori}$$

Equation 5 : Roof top scattering loss.

Where L_{rts} represents the correction factor for diffraction and scatter from rooftops to the street; w represents the distance between the edges of two adjacent buildings (m); f measures the frequency in MHz. L_{ori} considers ϕ, which is the angle between incidences coming from the base station and the road (measured in degrees).
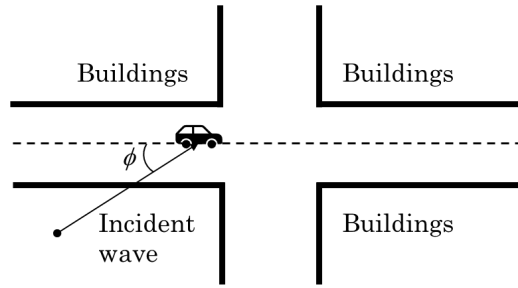
Figure 2. Definition of street orientation angle, ϕ

For $0 \leq \phi < 35$: $L_{ori} = -10 + 0.354 \frac{\phi}{deg}$

For $35 \leq \phi < 55$: $L_{ori} = 2.5 + 0.075(\frac{\phi}{deg} - 35)$

For $55 \leq \phi \leq 90$: $L_{ori} = 4 - 0.114(\frac{\phi}{deg} - 55)$

Furthermore, Δh_{mobile} is equivalent to $h_{roof} - h_{mobile}$, wherein h_{roof} is taken as $h_b \times 0.4$ in this paper and h_m is taken as h_r, or 1.5 meters.

For both the Cost 231-Hata model and Walfisch-Ikegami model, the RSRP can be generalized as Equation 6.

$$RSRP = P_{TX} + G_{AtoUE} - L_{feeding} - PL$$

Equation 6 : Received signal strength equation.

P_{TX} represents the base station's transmission power (dBm), G_{AtoUE} represents the antenna gain toward the UE's location, $L_{feeding}$ is the antenna feeding loss, and PL is the propagation loss which can be calculated from the individual models. The RSRP calculated from both models is compared with machine learning-based models later for benchmarking purpose.

Machine Learning-Based Signal Strength Prediction Models

Understanding network coverage is key for operators so that they can build and maintain high-quality mobile network(s). The conventional way for mobile network operators to comprehend their network coverage is done via drive tests, which have high operational costs and are not eco-friendly. To avoid this, many operators adopt various propagation models, namely empirical models or theoretical models. A recent innovative solution involves collecting coverage data by utilizing mobile applications that are running in the users' mobile phones. More than a sufficient amount of coverage data is generated during communication between the base station and the UE through the mobile applications and are sent to a central data server, which forms a large crowd data source for network coverage. Subsequently, machine learning has garnered attention as it is able to identify patterns in coverage statistics using big data.

In this section, various machine learning regression algorithms used to predict the coverage are explained, including Random Forest, Support Vector Machines (SVM), AdaBoost, and Neural Networks.

Random Forest Model

The Random Forest Regression model consists of a number of decision trees which are trained with the training data subset as shown in the figure below [10].
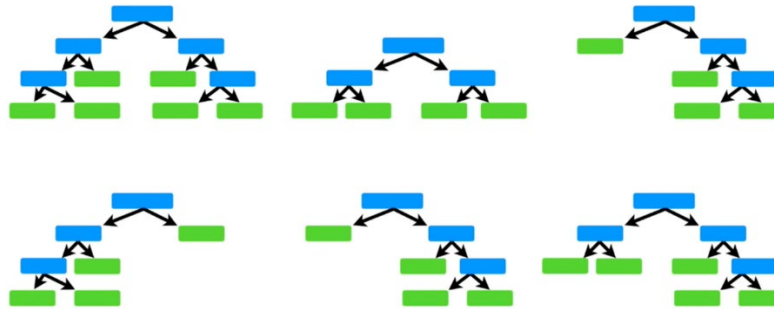
Figure 3. Architecture of Random Forest model comprising of multiple decision trees

Each of the decision trees are trained using different 'bootstrapped' data samples, wherein data is sampled randomly with replacement. The data samples not included in the Bootstrapped dataset can be used to estimate the accuracy of the Random Forest model and is called the 'Out of Bag' dataset. We utilize the Out of Bag dataset to enhance the Random Forest model's accuracy during the training.

The hyper parameters, including the maximum number of estimators and maximum depth, are set to 150 and 30 respectively. Upon training, predictions can be made based on the 7 selected features–frequency, transmission power, base station's antenna height, ground height from sea level to the base station and UE, antenna gain towards the UE, and the distance between the base station and UE.

Support Vector Regression (SVR) Model

Support Vector Machines (SVMs), which are premised on the Vapnik-Chervonenkis theory, are effective computational models in higher dimensions. SVMs are widely used for classification problems in machine learning. The use of SVMs in regression is known as Support Vector Regression (SVR). Contrary to most linear regression models, which aim to minimize the mean squared error, SVR's objective is to find a hyper plane function by which more data samples can be sorted within the decision boundary [11].
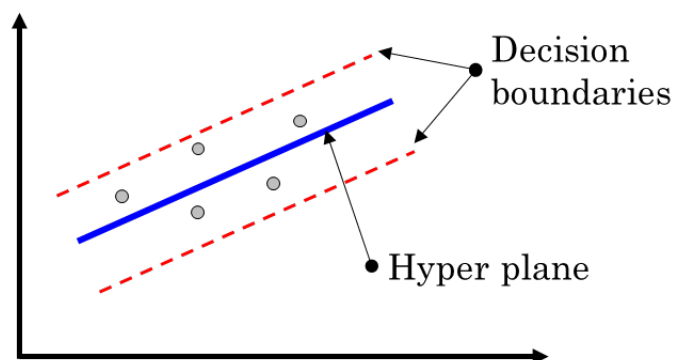

Figure 4. Support Vector Regression

SVR does not depend on the input points' dimensionality [12], which means that even if the dataset for the regression problem has a non-linear shape that cannot be fitted by linearly, SVR can provide a regression function through hyperplanes that are best suited to the dataset. This is possible because of the kernel function, which transforms non-linear datasets to a higher dimensional space so that the non-linear regression function becomes a linear regression problem. In this paper, radial basis function (RBF), is used as the kernel.

AdaBoost Model

Adaptive boosting, or shortly put as AdaBoost, is an ensemble method that utilizes "weak learners", or multiple models, to boost performance in both classification and regression [13]. A weak learner cannot solve non-linear problems but if used collectively, can solve non-linear regression problems. The AdaBoost algorithm involves the use of simple one-level decision trees (a.k.a. decision stump) as weak learners that are added sequentially to the ensemble. In this paper, decision stumps with a single split is also used. Each subsequent model attempts to correct the preceding predictions made by the model in sequence. This is accomplished by increasing the weight of incorrect decisions and decreasing the weight of correct decisions between sequences. An example of AdaBoost-based regression is shown in the figure below. When the AdaBoost was implemented with Python programming language, the number of decision stumps was set to 100.
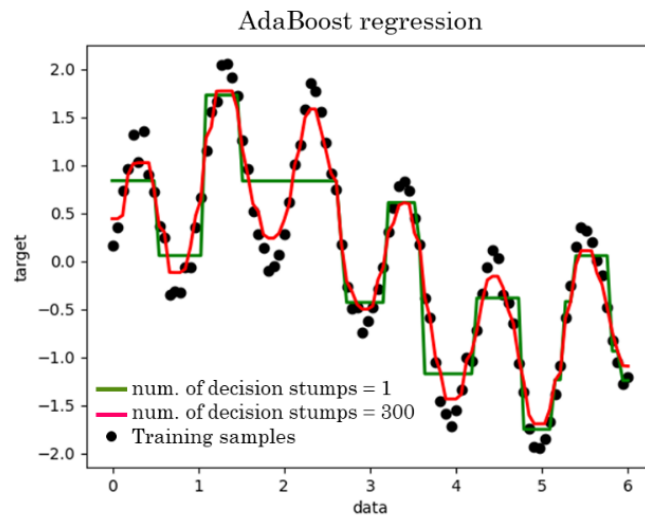
Figure 5. An example of AdaBoost regression

Neural Networks Model

Neural networks, which have been loosely modeled after the human brain, consist of multi-layers of nodes and weights [14]. The multi-layers usually consist of an input layer, one or more hidden layers, and an output layer. The nodes, a.k.a. neurons, are fully connected to those in the next layer with different assigned weights. The number of hidden layers and the number of neurons determine the network size and can impact the model's complexity and accuracy. Features including the number of layers, nodes, and how densely the nodes are interconnected are determined by performing multiple trials with different architecture designs. However, increasing the number of hidden layers will lead to extensive training time and overfitting, especially in the case where the training data set is not large enough.
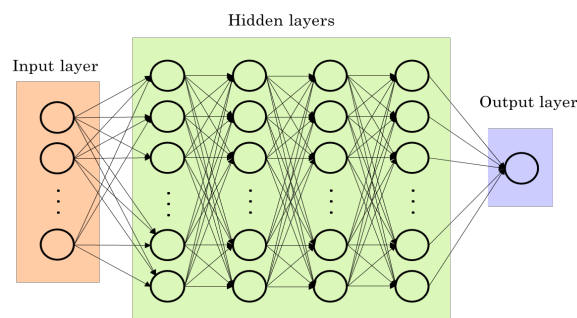
Figure 6. Architecture of the neural networks used to predict the signal strengths

In this paper, the neural network architecture used to predict the signal strength has 4 hidden layers with 10 nodes in each hidden layer. The nodes are fully inter-connected between the layers and there is single output node because the signal strength prediction is a regression problem. The activation function of the output node used is the ReLU function. The loss function and the optimization algorithm applied to minimize the loss function value for training the neural network are MSE (Mean Squared Error) and Adam (Adaptive Moment Estimation). In addition to the aforementioned models, linear regression can also be used to visualize the performance of the predicted RSRP against the actual RSRP by minimizing the residual sum of squares (RSS) [15].

Radio Signal Strength Data Used for Machine Learning Models

While the empirical propagation models are devised to predict the RSRP at the receiver's location using the base station and antenna information, the machine learning-based signal strength prediction models need to first be trained using actual signal strength data before predicting the signal strength. The actual signal strength data used to train the machine learning models in this paper is collected from drive tests conducted in Tokyo, Japan. The equipment utilized to collect the signal strengths of radio waves transmitted by the base stations in mobile networks is called the RF (Radio Frequency) scanner [16]. The RF scanner can measure the RSRPs coming from 6 or more base stations simultaneously; since the RSRP signals from different base stations are distinguishable by PCIs, the RSRPs distributions for each base station can be rearranged. Furthermore, since a GPS receiver can be connected to the RF scanner, the RSRP data can be collected together with GPS information, including longitude and latitude. The GPS information associated with the received RSRP is thus used to calculate the distance between the receiver and the corresponding base station. The drive test data used in this paper is proprietary data of Motiv Research co.


Figure 7. Drive test route where the signal strengths have been collected.

Each RSRP data sample is associated with its corresponding base station and antenna information by matching PCIs and the receiver's location. Since PCI identities range from 0 to 503, it is possible that there are multiple base stations that possess the same PCI. In this case, the base station which is closest to the receiver's location is considered. Once the base station and RSRP data are associated, the antenna gain of the base station facing the receiver's location is calculated. Antennas have horizontal gain patterns and vertical gain patterns, and both patterns must be added to form the overall antenna gain. The antenna gain towards a receiver's location is obtained by calculating the vertical angle, φ, and the horizontal angle, θ. Both angles are determined by the relative altitude difference

between the base station and the receiver as well as the straight distance between them, as illustrated in the figure below.
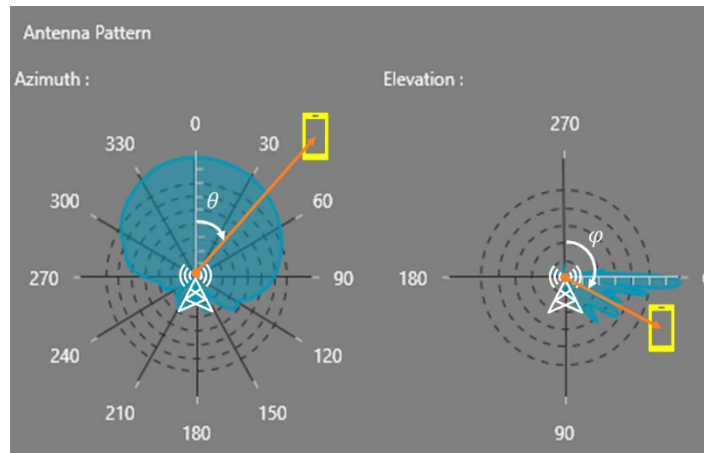

Figure 8. Illustration of the angle orientation towards the receiver in horizontal and vertical patterns.

Lastly, the dataset that will be used for training the machine learning models are formatted below. The training dataset consists of 7 features: frequency, transmission power, base station's antenna height, ground height from sea level to the base station and UE, antenna gain towards the receiver, and distance between the base station and receiver. 70% of the data samples is randomly selected from the training dataset to train the machine learning models and the remaining 30% of is used to test the prediction accuracy of the individual machine learning models. The prediction accuracy of the machine learning models is benchmarked and compared with the two empirical propagation models, the Cost231 Hata model and Walfisch-Ikegami model.

Results: Performances and Benchmarking
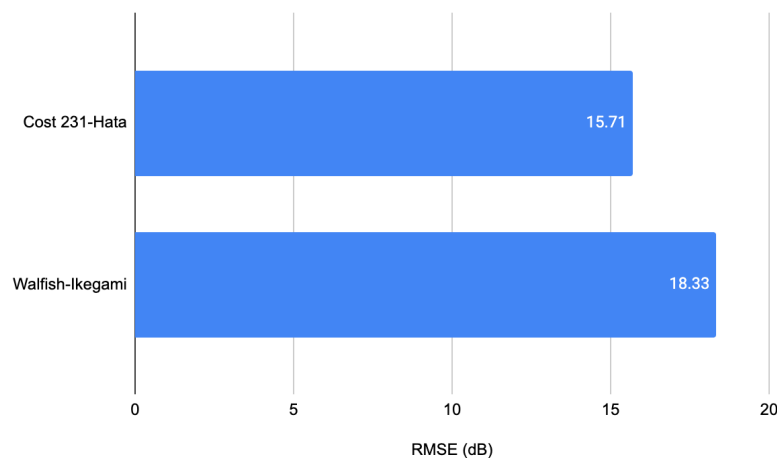
Figure 9. Comparison of the RMSE values for the empirical models (Cost 231-Hata and Walfisch-Ikegami model)
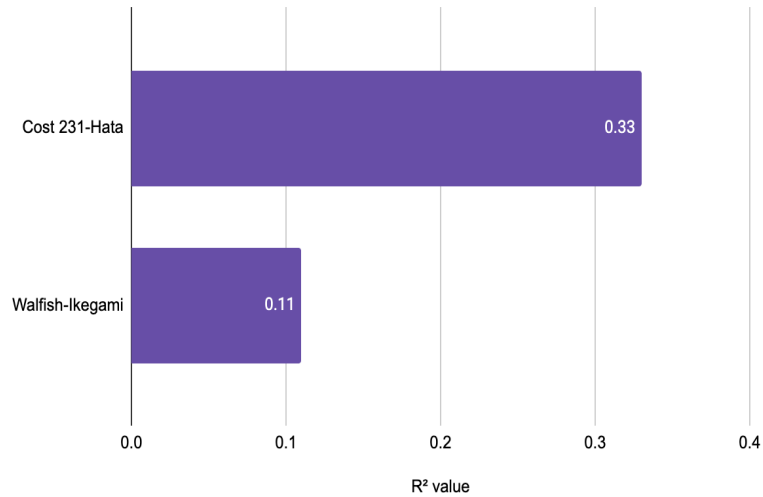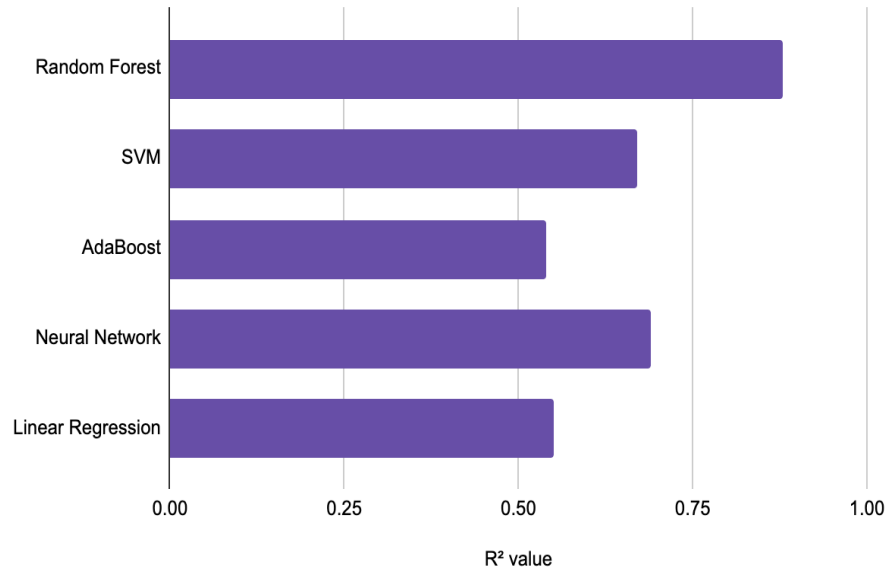
Figure 10. Comparison of the R² values for the empirical models (Cost 231-Hata and Walfisch-Ikegami model)

As demonstrated above, the Cost-231 Hata model outperforms the Walfisch-Ikegami model. Despite the fact that the Walfisch-Ikegami model takes into account more parameters than the Cost 231-Hata model, many of these are default values that were considered due to data on the structure of buildings and roads being unavailable. The Cost 231-Hata model, while still giving a high prediction error (RMSE) of 15.71 and R² value of 0.33, still performs better than its counterpart, which gives an RMSE value of 18.33 and R² value of 0.11. The Coefficient of Determination, also called the R² value, is a metric that examines the proportion of variance in the dependent variable associated with the independent variables [17]. The Root Mean Square Value (RMSE) measures the differences in values between the predicted model and the actual values [18].
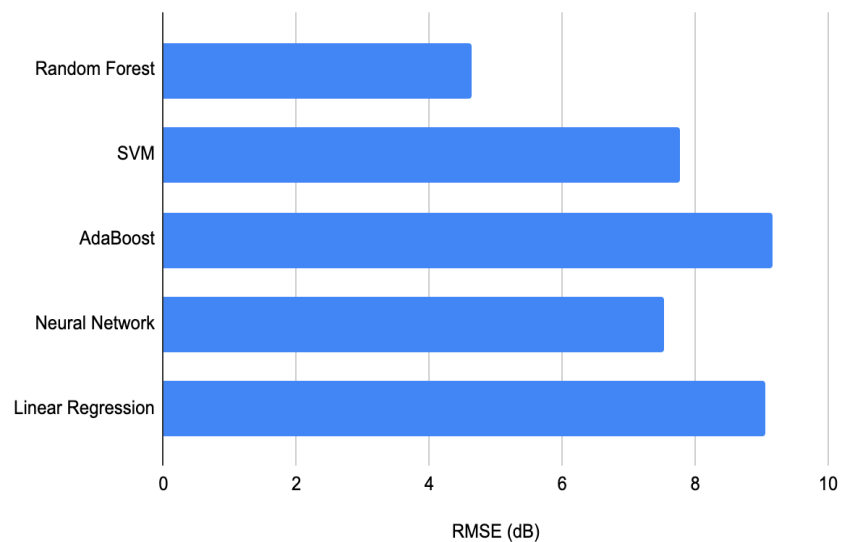
Figure 11. Comparison of the RMSE values for the Machine Learning models

Figure 12. Comparison of the R² values for the Machine Learning models

It is noticeable that the Random Forest machine learning algorithm yields the best results as the path loss prediction accuracy is the highest in comparison to the actual RSRP. The Random Forest model's RMSE value is the lowest being 4.64 dB, and the R² value similarly indicates high correlation, reaching 0.88. The neural network algorithm closely follows the prediction accuracy of the Random Forest model as it is trained with four hidden layers which in turn build complexity for better performance. On the other hand, AdaBoost is the least suitable model as it generates an RMSE value of 9.16 dB and R² value of 0.54, thus being unable to generalize upon the training data for prediction. Overall, the machine learning models show comparatively better performance than both of the empirical models as can be seen by the lower RMSE values. This can likely be attributed to the fact that machine learning models make generalizations based on the given dataset and are suitable for making predictions in specific regions.

Conclusion

In this research paper, base station signal strength prediction was performed with both empirical and machine learning models. The Cost-231 Hata and Walfisch-Ikegami model were used as frameworks for predicting cell coverage under certain constraints and estimated parameters. The performance of these models was compared against the machine learning models trained with a train-test ratio of 70:30; the latter provided better results than both of the empirical models. Of the five algorithmic models used, Random Forest outperformed the rest with Neural Networks closely behind in terms of prediction accuracy. Therefore, using the performance metrics (RMSE and R² value), it can be concluded that the machine learning models perform best compared to empirical propagation models, of which Random Forest provides the highest prediction accuracy. Moreover, it can be concluded that machine learning-based RSRP prediction models are promising candidates for use by network operators as they provide more accurate results than those of the empirical models, especially in smaller and specific areas.

Acknowledgements

References

[1] E. Dahlman, S. Parkvall and J. Skold, "Chapter1. Background of LTE," in *4G LTE/LTE-Advanced for Mobile Broadband*, 2014.

[2] E. Dahlman, S. Parkvall and J. Skold, "Chapter 2. High Data Rates in Mobile Communication," in *4G LTE/LTE-Advanced for Mobile Broadband*, 2014.

[3] A. Elnashar, M. El-saidny and M. Sherif, "Chapter 2.11 LTE Air Interface Procedures," in *Design, Deployment and Performance of 4G LTE Networks*, 2014.

[4] M. Fowler, "Maxwell's Equations and Electromagnetic Waves," University of Virginia, [Online]. Available: http://galileo.phys.virginia.edu/classes/109N/more_stuff/Maxwell_Eq.html.

[5] S. Yuvraj, "Comparison of Okumura, Hata and COST-231 Models on the Basis of Path Loss and Signal Strength," *International Journal of Computer Applications,* vol. 59, no. 11, pp. 37-41, 2012.

[6] M. Piacentini and F. Rinaldi, "Path Loss Prediction in Urban Environment Using Learning Machines and Dimensionality Reduction Techniques," Sapienza Universita Di Roma, Rome, 2009.

[7] M. Knebl, "Three Ways to Improve Your Network Strategy with Crowdsourced Data (White Paper)," SPEEDTEST, 11 March 2020. [Online]. Available: https://www.speedtest.net/insights/blog/improve-network-strategy-crowdsourced-data/.

[8] "Free Space Path Loss: details & calculator," Electronics-notes Incorporating Radio-Electronics.com, [Online]. Available: https://www.electronics-notes.com/articles/antennas-propagation/propagation-overview/free-space-path-loss.php.

[9] A. Elnashar, M. El-saidny and M. Sherif, "Chapter1. LTE Network Architecture and Protocols," in *Design, Deployment and Performance of 4G LTE Networks*, 2014.

[10] J. Starmer, "StatQuest: Random Forests," YouTube, 6 February 2018. [Online]. Available: https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&ab_channel=StatQuestwithJoshStarmer.

[11] J. Hawkins, "Support Vector Regression," [Online]. Available: https://core.ac.uk/download/pdf/81523322.pdf.

[12] J. Starmer, "StatQuest: Support Vector Machines," YouTube, 1 October 2019. [Online]. Available: https://www.youtube.com/watch?v=efR1C6CvhmE&ab_channel=StatQuestwithJoshStarmer.

[13] A. Desarda, "Understanding AdaBoost," towards data science, 17 January 2019. [Online]. Available: https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe.

[14] N. Mandayam, "Radio Propagation Models," Rutgers University, [Online]. Available: http://mobilityfirst.winlab.rutgers.edu/~narayan/Course/Wless/Lecture_3_RadioPropagationModel_Sneha.pdf.

[15] "Scikit learn developers," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[16] "Anritsu: Area Tester," Anritsu, 2021. [Online]. Available: https://www.anritsu.com/ja-jp/test-measurement/products/ml8780a.

[17] "R-squared intuition," Khan Academy, [Online]. Available: https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/a/r-squared-intuition.

[18] J. Moody, "What does RMSE really mean?," 6 September 2019. [Online]. Available: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e.