Signal Coverage Maps Using Measurements and Machine Learning

1 Introduction

New 5G and IoT systems are driving the deployment of dense small cell networks and leading the way to the smart cities of tomorrow. Companies need to use signal coverage maps to optimize and validate their networks but collecting signal measurements to generate those maps is an expensive process. To reduce cost there is a need to apply prediction models to generate coverage maps from limited, sparse measurement data.

In this project, we build a prediction model of signal coverage map based on the data set provided by mySignals[1]. In general, there are two approaches in signal strength prediction: propagation models and data-driven approaches. We design and implement data-driven propagation models in MATLAB using the measurements data. Using the paper[2] as a guide, we implement some models as a baseline using geospatial interpolation techniques, and implement some model using machine learning techniques.

1.1 Baseline: propagation model

We implement the Log Distance Path Loss (LDPL) propagation model, and use it as the baseline.

1.2 Geospatial interpolation

We reproduce the Kriging interpolation method and the de-trending Kriging interpolation method under the guidance of this paper[3].

---

[1] http://www.mysignals.gr/research.php

[2] Alimpertis, Emmanouil , et al. "City-Wide Signal Strength Maps: Prediction with Random Forests." The World Wide Web Conference 2019.

[3] A. Chakraborty, Md S. Rahman, H. Gupta, and S. R Das. [n. d.]. SpecSense: Crowdsensing for Efficient Querying of Spectrum Occupancy. In Proc. of the IEEE INFOCOM '17. 1–9.

1.3 Machine learning

We evaluated a variety of machine learning methods through the regression learning app in MATLAB and found the one with the best effect as the result.

2. Baseline: propagation model

2.1 Determine base station location

In this methods, it is necessary to determine the location of base station. Our dataset does not give specific base station locations. Therefore, the data is needed to determine the possible location of the base station.

Firstly, we make the following assumptions:

1. Each cell has at most one base station;

2. The base station location is the place where the signal strength of the current cell is maximum.

Of course, it may be problematic to directly select the point with the maximum signal intensity of the current cell. After analyzing the data, we found that there was only one data at the point with the maximum signal intensity in some cells, which might not be credible enough. Or the signal strength at that point is constantly floating, up and down. Therefore, we consider setting the evaluation index to determine the location of the base station, determine the score of point I through the formula, and consider that the point with the maximum score is the base station of the current cell. A reasonable base station should have higher signal strength and more data volume support. Therefore, we evaluate the reliability of a point as a base station by the following formula:

$$S(i) = R_{avg}\ (i) + \lambda N(i) + \alpha(i)$$

Where, $S(i)$ represents the fraction, $R_{avg}\ (i)$ represents the average signal intensity measured at the current point, λ is the linear coefficient, which we set as 0.1, $N(i)$ is the data amount of the current point, and $\alpha(i)$ is the fraction complement, determined as follow:

$$\alpha(i) = \begin{cases} 0 & N(i) \geq 3 \\ -20 & N(i) < 3 \end{cases}$$

For example, the data of a potential base station point is: 5 measurements, sigal strength { -53, -56, -54, -55, -57}, then the $S(i)$ is $S = -55 + 0.1 * 5 + 0 = -54.5$.

After calculating the score of each point, select the point with the maximum score as the base station location of the current cell.

2.2 Log Distance Path Loss (LDPL) propagation model

Log Distance Path Loss (LDPL) propagation model is a relatively simple propagation model. As the distance d from the transmitter increases, the lost power decreases logarithmically with the distance, as shown in the formula:

$$P - 10\alpha \log_{10}(d) + N(0, \sigma^2)$$

in which α is the path-loss exponent. we can compute the "perceived" path-loss exponent α_i at observation point si to be:

$$\alpha_i = \frac{P(s_t) - P(s_i)}{10 \log_{10}(d_i)}$$

and $P(s_i)$ are the signal strength at the locations s_i and the assumed transmitter location s_t respectively, and d_i is the distance between s_i from s_t.

Therefore, we have calculation steps of logarithmic loss model:

1. For a specific area (cell), find the corresponding point of base station location

2. Calculate the distance from each point to the base station d_i, and calculate the loss coefficient according to the loss coefficient formula

3. All the loss coefficients are averaged as the current regional loss coefficients

4. The average loss factor is used to calculate the received power at each position after loss

3. Geospatial interpolation

3.1 Ordinary Kriging (OK)

State-of-the-art approaches in data-driven RSS prediction have primarily relied on geospatial interpolation, which however is inherently limited to only spatial features

(lx ,ly). The best representative of this family of predictors is ordinary kriging (OK) and its variants. The method is based on the assumption of homogeneity of geospatial attributes. That is, any point in the space has the same mean and variance, and the measured value of each point can be composed of mean C plus an offset R.

$$E\left[z(x,y)\right] = E[z] = c$$
$$Var\left[z(x,y)\right] = \sigma^2$$

The interpolation formula is shown below, where λ I is the weight coefficient.

$$\hat{z}_0 = \sum_{i=1}^{n} \lambda_i z_i$$

It estimates the value of an unknown point by a weighted sum of data from all known points in space. However, the weight coefficient is not the reciprocal of the distance, but a set of optimal coefficients that can meet the minimum difference between the estimated value $\hat{z}_0$ at the point (x_0, y_0) and the real value z_0.

To calculate this set of optimal coefficients, we need to construct the cost function from the assumptions and constraints, and get the final coefficient calculation formula by Lagrange multiplier method. Where, r_{ij} is a semi-variance function.

$$J + 2\phi(\sum_{i=1}^{n} \lambda_i - 1)$$

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} & 1 \\ r_{21} & r_{22} & \cdots & r_{2n} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \\ -\phi \end{bmatrix} = \begin{bmatrix} r_{1o} \\ r_{2o} \\ \cdots \\ r_{no} \\ 1 \end{bmatrix}$$

The semi-variance function that appears in the coefficient calculation formula is defined as follows. Because half of the variance appears, it is named semi-variance function.

In practice, in order to obtain the expression of the semi-variance function, the spatially similar properties are similar according to the first law of geography. Distance d_{ij} represents spatial similarity, and semi-variance r_{ij} represents attribute similarity.

$$r_{ij} = \sigma^2 - C_{ij} = \frac{1}{2} E\left[(z_i - z_j)^2\right]$$

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$

Kriging interpolation assumes that there is a functional relationship between d_{ij} and r_{ij}, which can be linear, quadratic, exponential or logarithmic.

To confirm this relationship, we calculate the distance and half-variance of any two points on the observational data set, obtain n^2 data pairs of r_{ij} and d_{ij}, plot all d and r as scatter plots, find an optimal fitting curve fitting relation between d and r, and obtain the function relation $r_{ij} = r(d_{ij})$ by reading relevant literature. In this problem we use the exponential function as shown in the figure for semi-variance function fitting.

$$r(d) = C\left(1 - e^{-(h/a)}\right)$$

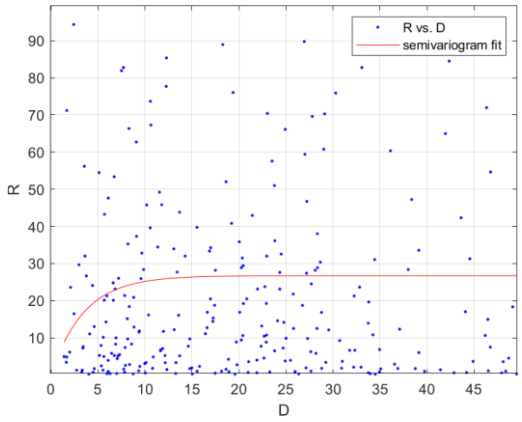

Figure 3.1 Semi-variance function fitting

Thus we get OK interpolation steps:

1. For the observed data, the distance and semi-variance are calculated in pairs;

2. To find the relation between fitting distance and semi-variance of a fitting curve, the corresponding semi-variance can be calculated according to arbitrary distance;

3. Calculate the half-variances between all known points r_{ij};

4. For the unknown point z_0, calculate its half-variance r_{z0} to all known points z_i;

5. The optimal coefficient λ_i is obtained by solving the equations;

6. The weighted sum of the attribute values of the known points is used to obtain the estimated value of the unknown point z_0.

3.2 Ordinary Kriging Detrending (OKD)

In fact, the Kriging interpolation method has the following two shortcomings: in large geographical areas, the assumption of spatial uniformity may not be accurate. Due to different geospatial attributes, each different area may no longer have the same mean and variance.

In the aspect of signal propagation, the mean of signal intensity at each point is not the same. Obviously, the closer the base station is, the greater the signal intensity will be, and the semi-variances between points at the same distance may not be equal.

Therefore, we reproduce an optimized Kriging interpolation method: detrended Kriging interpolation. In this method, the measured value of each point is divided into two parts, one is the signal strength $L(s_i)$ calculated by the Log Distance Path Loss propagation model, and the other is the trendless value $\delta(S_i)$ that satisfies the kriging interpolation conditions:

$$Z(s_i) = L(s_i) + \delta(s_i)$$

Thus, the steps of kriging interpolation method are obtained:

1. Calculate the Log Distance Path Loss model

2. The measured value of each point is subtracted from the Log Distance Path Loss model model to obtain $\delta(s_i)$ satisfying kriging interpolation conditions.

3. Kriging interpolation results are calculated for all $\delta(s_i)$

4. The kriging interpolation results are added to the Log Distance Path Loss model results

3.3 Results and analysis

The results of propagation model method and geospatial interpolation method are shown in the table below:

Table 3.1 Results of Baseline Methods

Method	LDPL	OK	OKD
RMSE	79.68	1.064*10e3	0.532*10e3

It can be seen that the effect of several geospatial interpolation methods is very bad, completely inferior to the results obtained by the propagation model.

We tried to analyze possible causes why interpolation works less well than the simplest logarithmic loss model. This may be due to several reasons:

1. The data set itself is not accurate enough. When we looked at the data set, a lot of it didn't make sense. Even though we removed a lot of anomalous data, there were still multiple widely different measurements at the same coordinates at the same time. And the data set provider also mentioned the low accuracy of coordinate measurements.

2. The hypothesis of Kriging interpolation is not satisfied. The presupposition of Kriging method requires that the measured attributes meet the homogeneity condition. Using this method directly for signal strength actually violates this assumption. Therefore, a kriging interpolation method for trend division is proposed.

3. Geospatial segmentation is not fine enough. During the experiment, we simply segmented the geographical space by the cell to try to satisfy the Kriging interpolation conditions. In the experiment, we found that the semi-variance function fitted by kriging interpolation method in the same cell tended to be equal when the distance was more than 500m, which led to the properties of the semi-variance matrix obtained was not good enough. Therefore, reasonable geographical segmentation is probably better in the diameter of less than 500m or so, and finer segmentation may get better results.

We also thought about possible improvements

The first is more detailed geographic partitioning, as analyzed above. The second is better de-trending. In theory, de-trended Kriging should yield better results, but the experimental results are not as good as the logarithmic loss model. Better results than theoretical propagation models can be obtained if rational geographical partitioning is used and trends are removed well.

4 Machine Learning

4.1 Model

To obtain the signal strength distribution, we use several different machine learning models. These models can be divided into the following categories: tree-based models, linear regression models, ensemble models, and Gaussian process regression (GPR) models.

For the tree-based model, we mainly use fine tree, medium tree, and coarse tree. The main difference is the granularity of the processing. Linear regression models are mainly used: common linear regression model, interactions linear, robust linear, and stepwise linear. The main difference is whether to consider some interaction between data, and the allocation and implementation of steps. But in general, they are all in the category of linear regression. Ensemble models, including boosted trees and bagged trees (mainly random forests), train and predict samples with multiple trees. GPR mainly uses rational quadratic GPR, squared quadratic GPR, Matern 5/2 GPR, and exponential GPR. They use a Gaussian process to perform regression analysis on the data, the difference is the kernel function used.

4.2 Training

We used Regression Learner APP in MATLAB to train the above model. First of all, we compared the performance before and after using Principal Component Analysis (PCA), and found that the data training effect after using PCA was better. Therefore, PCA was first used in subsequent data preprocessing to retain 95% of the effective features and effectively improve the algorithm speed. We look for the best algorithm and parameters with a lower RMSE as the standard.

4.3 Data Selection

We want to look for the selection of data that best characterizes the signals. After reading the description of the data set, we decided to select longitude, latitude, and time as the basic parameters. In addition, it specifically mentions that the cell to which the signal belongs has a great influence on the signal, so the cell ID is also a basic parameter.

Figure 4.1 displays the process of finding the optimal selection of data. In each subfigure, the 13 points on the horizontal axis represent 13 different machine learning methods. Four different color areas represent four different categories. The two

discounts represent train RMSE and test RMSE. We tried longitude, latitude, time, cell ID, LAC (location area code), MNC (mobile network code) and ARFCN (absolute radio-frequency channel number) the selection of these parameters.

As the results show, RMSE is the lowest when the 4 groups of basic parameters are selected. Therefore, selecting these 4 basic parameters as the combination of training data has the best effect.

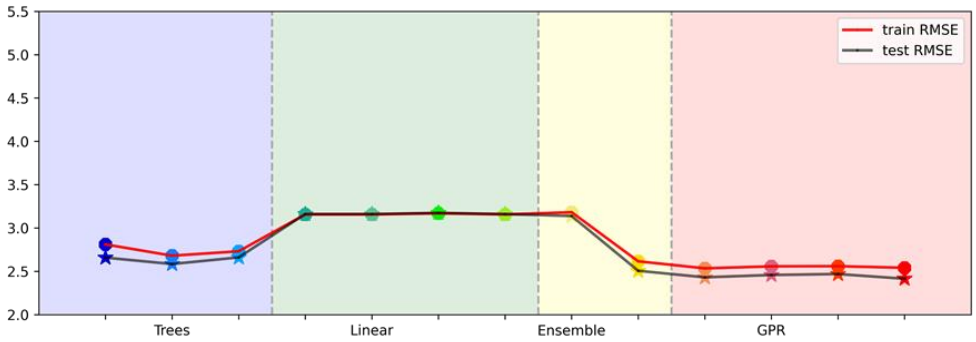

Figure 4.1 (a) Only use longitude, latitude, time, cell ID

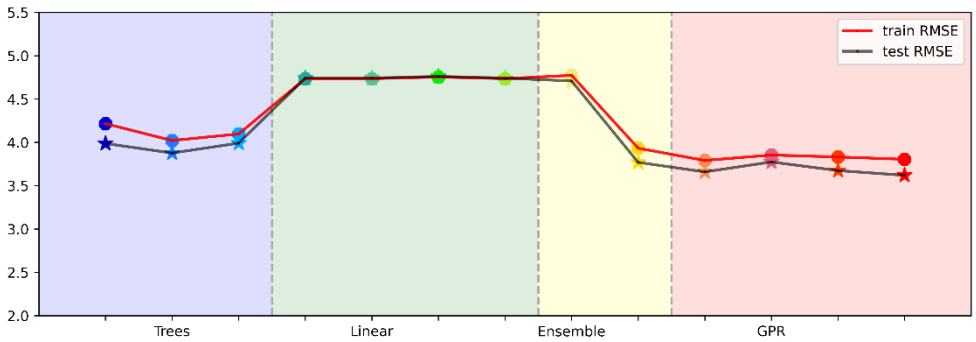

Figure 4.1 (b) Add LAC

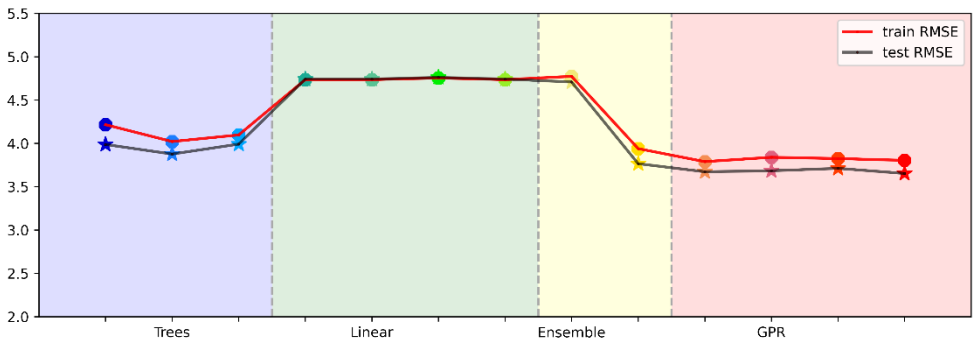

Figure 4.1 (c) Add MNC

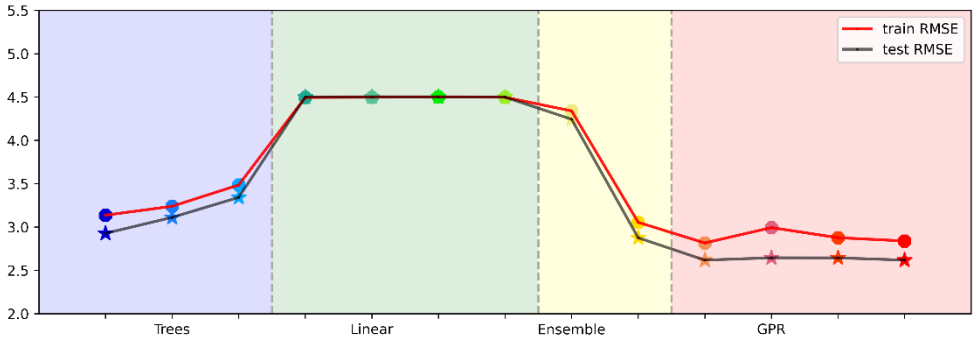

Figure Y (d) Add ARFCN

4.4 Time-based analysis

We want to explore whether signal data correlates with time. We split the time data into months, dates, hours, and minutes for training, and the results are shown in Figure 4.2. We found that RMSE did not decrease, and some even showed high RMSE, indicating that time splitting is not a better representation of the data. Even in the case of time represented as date, two sets of training with the GPR model failed, further demonstrating that time should not be split.

Figure 4.2 (a) Time is expressed in months

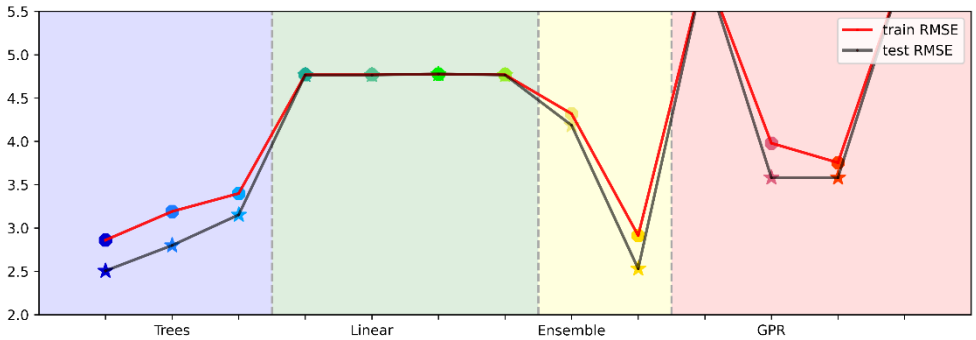

Figure 4.2 (b) Time is expressed in dates

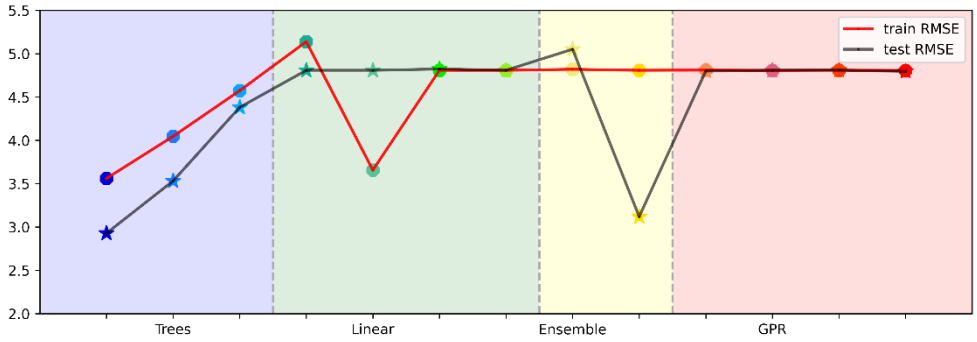

Figure 4.2 (c) Time is expressed in hours

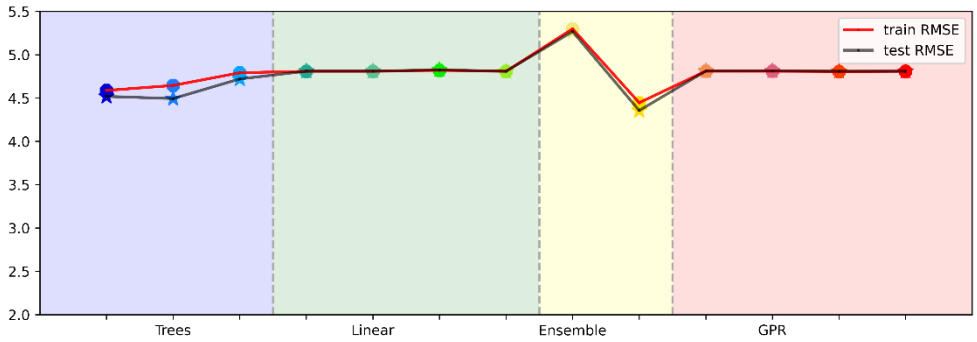

Figure 4.2 (d) Time is expressed in minutes

4.5 Results of Machine Learning

After a series of model selection, parameter adjustment, and training parameter selection optimization, we concluded that: With 1% of the data selected as the training set and the rest as the validation set, and with longitude, latitude, cell ID, and time as the known data, the RMSE obtained using rational quadratic GPR is the minimum 2.431. Figure 4.3 displays the results associated with this training process and the final coverage map.

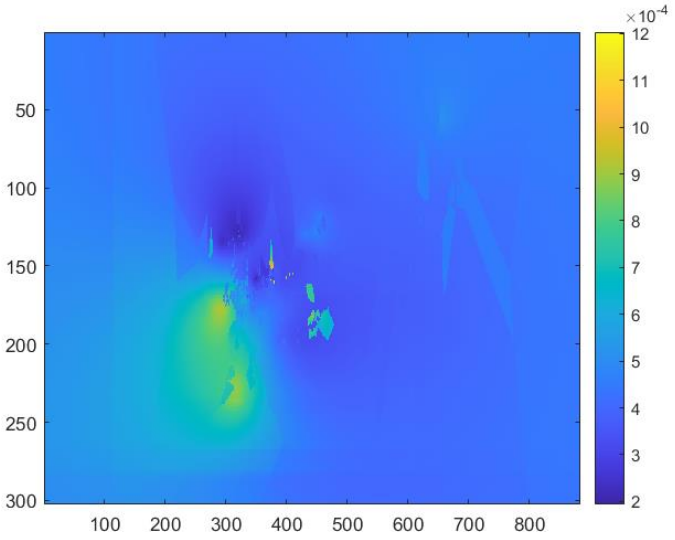

Figure 4.3 (a) Signal strength distribution

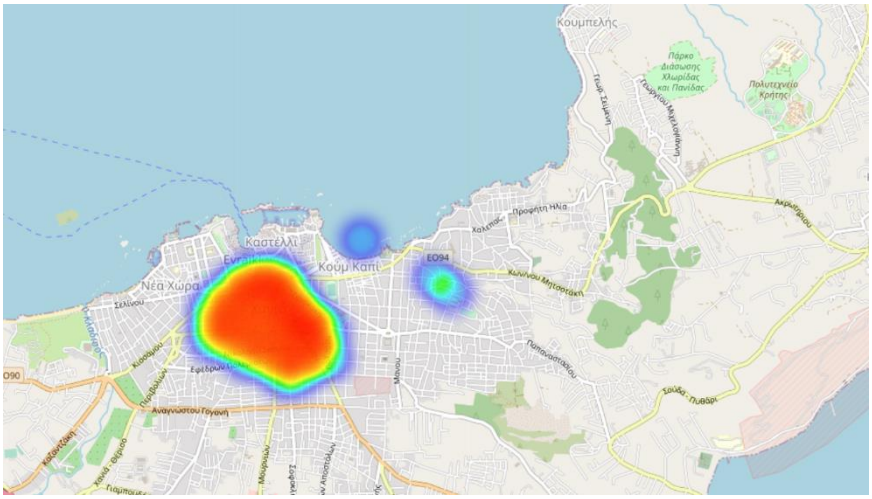

Figure 4.3 (b) Coverage map

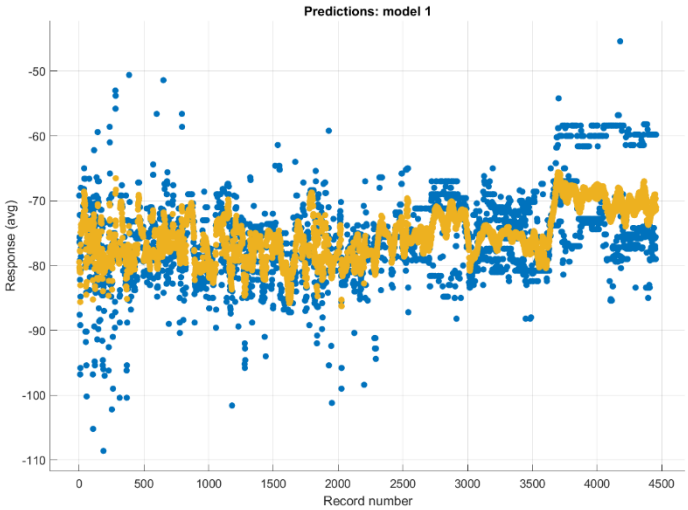

Figure 4.3 (c) Response plot

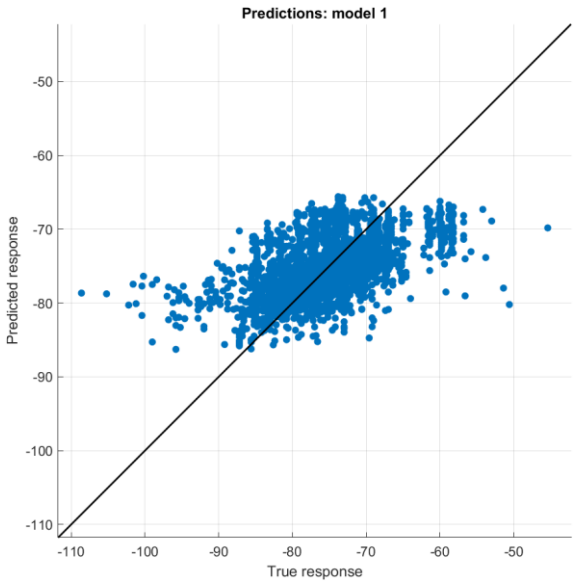

Figure 4.3 (d) Predicted vs. actual plot

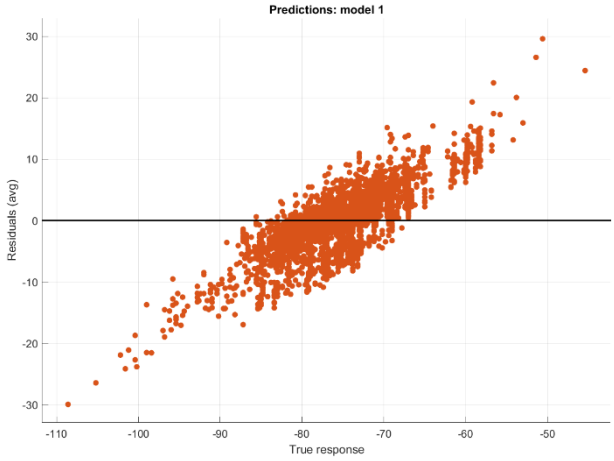

Figure 4.3 (e) Residuals plot

5. Conclusion

Following the references, we reproduce some propagation models and data-driven models for signal strength prediction. The data - driven model includes geospatial interpolation and machine learning. From the results, if the propagation model is taken as the baseline, the geospatial interpolation method gives poor results, while the machine learning method on the other hand gives good results. Due to the lack of ability, we only analyze the possible causes and give the possible improvement methods.