Estimation of Site-specific Radio Propagation Loss with Minimal Information

Team : BoilerSignal
Members : Ahmed P. Mohamed, Manish Kumar Krishne Gowda

1 Member Roles

- **Ahmed P. Mohamed:** PhD Candidate, Purdue University; contributed to feature engineering and training of path loss models.
- **Manish Kumar Krishne Gowda:** PhD student, Purdue University; assisted in extracting regions and preparing the dataset, including computation of geometric, LOS, and Fresnel-based features.

2 Data Selection Method

From the given OBJ file, four distinct spatial regions were extracted to capture the environmental features relevant to signal propagation. Specifically, polygonal building structures were selected within

1. a 50 m spherical region centered around the transmitter (Tx) point
2. a 50 m spherical region centered around the receiver (Rx) point
3. the first Fresnel zone between the Tx and Rx points
4. the third Fresnel zone between the Tx and Rx points.

These regions collectively provide a comprehensive representation of both the immediate surroundings of the transceivers and the key propagation zones influencing the received signal characteristics.

Fresnel zones are widely used in propagation analysis: Reference [2] employs them to evaluate line-of-sight (LOS) versus non-line-of-sight (NLOS) conditions, while Reference [1] leverages these zones to extract additional environmental features for training models to predict path loss (although, they use a LIDAR dataset).Therefore, employing Fresnel zones is well justified in this context. The 50m spheres around the TX and RX points are included to capture the immediate environment, as these regions contain most of the local clutter. A denser or more complex environment increases the number of reflections and scattering, which directly impacts the received signal and consequently the path loss. Including these spheres allows the model to account for local multipath effects more accurately.

To extract the required regions (50m sphere and the Fresnel zones) from a large city mesh .obj file for a given TX−RX pair, we first reads the TX−RX coordinates from the given CSV files. For each pair, we compute the 50m sphere around the rx and tx point, first and third Fresnel zone ellipsoids based on the TX−RX distance and the wavelength. Each triangle in the mesh is sampled, and those intersecting the Fresnel ellipsoids are clipped to generate polygons representing the zone. Duplicate points are removed, and the polygons are ordered to form valid planar shapes. A visualization of the RX zone, TX zone, and the first and third Fresnel zones for a sample TX−RX pair is shown in Figure 1.

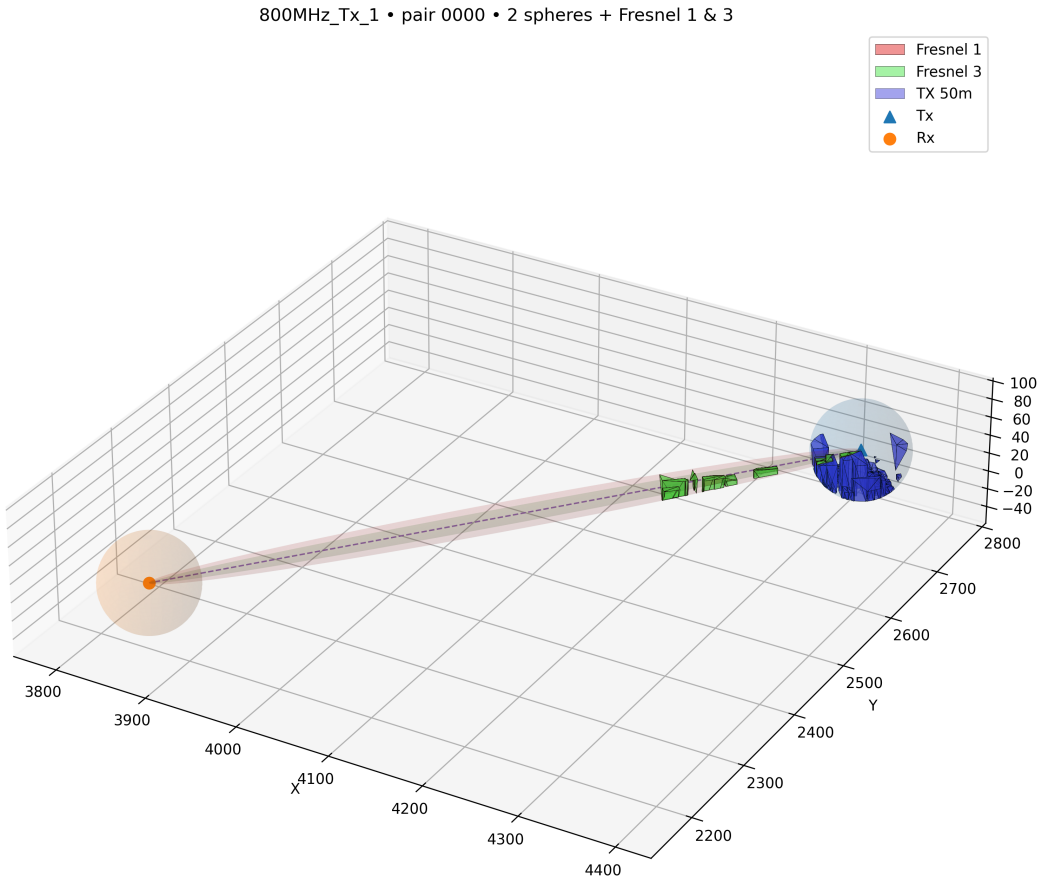

Figure 1: Visualization of the 50m RX zone, 50m TX zone, and the first and third Fresnel zones for a sample TX − RX pair for 800MHz

2.1 Feature Extraction:

2.2 Feature based on coordinate geometry

Multiple features based on coordinate geometry were extracted from the simulated propagation environment to improve the path loss (PL) prediction task. The features were computed based on the 50-meter spheres around the transmitters (Tx) and the corresponding Fresnel regions for each receiver (Rx). These features capture both geometric and environmental aspects of the propagation and are described as follows:

- **3D Euclidean Distance (d_{3D}):** The straight-line distance between the Tx and Rx is calculated using

$$d_{3D} = \sqrt{(x_{Tx} - x_{Rx})^2 + (y_{Tx} - y_{Rx})^2 + (z_{Tx} - z_{Rx})^2}, \tag{1}$$

  where (x_{Tx}, y_{Tx}, z_{Tx}) and (x_{Rx}, y_{Rx}, z_{Rx}) are the 3D coordinates of the transmitter and receiver, respectively.

- **Free-Space Path Loss (FSPL) in dB:** Based on the 3D distance, the free-space path loss is computed as

$$\mathrm{FSPL}_{\mathrm{dB}} = 20\log_{10}(d_{3\mathrm{D}}) + 20\log_{10}(f) - 147.55, \tag{2}$$

  where $d_{3\mathrm{D}}$ is in meters, and f is the carrier frequency in Hz.

- **3GPP UMi Street Canyon Path Loss:** The path loss according to the 3GPP Urban Micro (UMi) Street Canyon model is given by

$$\mathrm{PL}_{\mathrm{UMi}} = 22\log_{10}(d_{3\mathrm{D}}) + 28 + 20\log_{10}(f), \tag{3}$$

  where $d_{3\mathrm{D}}$ is in meters and f is in GHz.

- **Transmitter Height ($h_{\mathbf{Tx}}$):** The absolute height of the transmitter above the ground in meters, which influences the line-of-sight probability and received signal strength.

2.3 Ray-Tracing and Geometric Features for Path Loss Prediction

In addition to distance-based features, several geometric and line-of-sight (LOS) features were extracted to capture the effects of the propagation environment on path loss. These features were computed using ray-tracing on the 3D polygonal representation of the environment, including building and obstruction geometries, for each transmitter-receiver (Tx-Rx) pair. The extracted features are described below:

- **Polygon Count :** The total number of polygons intersected by the Fresnel zone at the third and first Fresnel regions, providing an estimate of environmental complexity along the propagation path.

- **LOS Intersection Count :** The number of polygons intersecting the direct line-of-sight (LOS) ray between Tx and Rx. A higher count indicates a greater likelihood of non-line-of-sight (NLOS) conditions.

- **Distance to First and Last Intersection :** The distance from the transmitter to the nearest and farthest polygon intersections along the LOS ray, providing information about obstruction positions along the path.Equal to -1 when there is LOS.

- **LOS Flag** A binary indicator of whether the Tx-Rx pair has an unobstructed LOS (1) or is blocked (0), determined by the presence of intersecting polygons along the ray.

- **Obstruction Area:** The cumulative surface area of polygons obstructing the Fresnel zones (first and third) and the LOS path, capturing the extent of environmental blockage which correlates with diffraction and shadowing effects.

All features were computed by iterating over polygons only once per Tx-Rx pair, applying the MöllerTrumbore ray-triangle intersection algorithm to detect intersections and calculate distances along the LOS ray. Polygon areas were computed using the standard cross-product method for 3D triangles. In cases of corrupt or missing data files, default values (e.g., 0 polygons, LOS flag 1) were used to maintain dataset consistency.

These geometric and LOS-related features complement the previously described distance- and frequency-based features, providing a rich representation of the propagation environment for machine learning-based path loss prediction.

2.4 Tx and Rx Sphere-Based Environmental Features

To further capture the local propagation environment around the transmitter (Tx) and receiver (Rx), additional features were extracted from the 50-meter spherical regions surrounding each node. These features quantify the density, obstruction, and height of clutter objects within the immediate vicinity of the Tx and Rx. They are described below:

- **Tx Sphere Polygon Count:** The total number of polygons intersecting the 50-meter sphere around the transmitter. This represents the local structural complexity and the number of potential obstacles affecting the transmitted signal.
- **Tx Sphere Obstructing Area :** The cumulative surface area of all polygons within the transmitter sphere, providing a measure of the amount of obstruction and clutter that could cause diffraction, reflection, or shadowing.
- **Average Tx Clutter Height:** The mean of the maximum heights (Z-coordinates) of all polygons within the transmitter sphere, capturing the vertical distribution of obstacles that may affect line-of-sight and propagation.
- **Rx Sphere Polygon Count:** The total number of polygons intersecting the 50-meter sphere around the receiver, indicating the local environmental complexity at the Rx.
- **Rx Sphere Obstructing Area:** The sum of the areas of all polygons within the receiver sphere, quantifying the degree of obstruction near the receiver that may attenuate incoming signals.
- **Average Rx Clutter Height :** The mean of the maximum heights of all polygons within the receiver sphere, describing the vertical clutter distribution around the receiver.

These features were calculated once per Tx-Rx pair using the polygonal 3D environment data, and then broadcasted to all relevant samples in the dataset. They complement the previously described distance-based, LOS, and Fresnel-region features, providing a more comprehensive representation of the surrounding environment for machine learning-based path loss prediction.

3 AI Model and Training Description

As shown in [1] and confirmed through our experiments on this dataset, boosting-based models are particularly effective when the available training data is limited, a requirement of this task. Therefore, we adopt the CatBoost algorithm for the path loss prediction task due to its strong performance with structured tabular data and its robustness to overfitting in low-data regimes.

In this work, the extracted geometric and obstruction-based features are used as input to the CatBoost regression model. The model is trained to predict the measured path loss using the extracted features. CatBoost internally handles feature normalization and missing values, allowing for learning from heterogeneous data sources such as geometric, LOS, and obstruction-based parameters. It builds an ensemble of decision trees sequentially, minimizing prediction error while preventing overfitting through ordered boosting and regularization. It is particularly effective for tabular data such as propagation-loss prediction tasks, as shown in [1].

Initially, we trained a single unified CatBoost model using data aggregated from all frequency bands (800 MHz, 7 GHz, and 28 GHz). However, experimental results revealed that this unified model was suboptimal, as it could not effectively capture the distinct propagation behaviors at

different frequenciesparticularly the variation in free-space loss, diffraction, and blockage characteristics.

To address this, we developed three separate CatBoost models, one per frequency band. Each model was trained exclusively on the corresponding frequency dataset, ensuring frequency-specific featuretarget relationships were learned effectively.

A dispatcher module was implemented to automatically route training and testing samples to the appropriate model based on the frequency label. This modular setup not only improved prediction accuracy but also simplified the model management pipeline during inference. The AI model is shown in Fig. 2

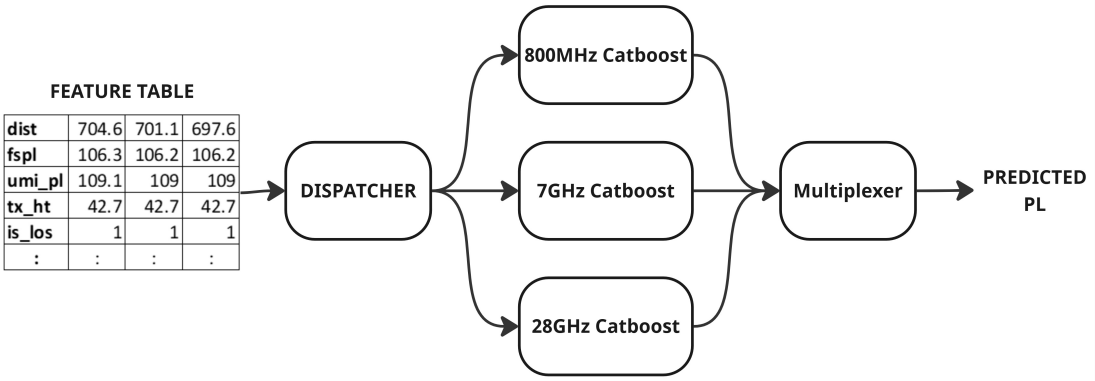

Figure 2: Block diagram of the frequency-specific CatBoost modeling framework. The dispatcher directs samples to the respective model (800 MHz, 7 GHz, or 28 GHz) for training or inference.

The lightweight CatBoost model was trained using frequency-specific datasets as described in Fig. 2. Hyperparameter optimization was performed using the Optuna framework over 20 trials, ensuring efficient exploration of the parameter space while maintaining computational feasibility.

The validation strategy employed an 80/20 trainvalidation split, with early stopping set to 50 rounds to prevent overfitting and reduce unnecessary iterations.

The Optuna tuning process completed in approximately 0.14 minutes (8.4 seconds), while the final model training after parameter selection required only 0.20 seconds, demonstrating the efficiency of the CatBoost implementation for this regression task.

4 Data Subset Selection

The competition provided two related subtasks:

Task 1: Estimation using a small number of receiving points. Transmitting point data may be utilized to enhance accuracy, though fewer is preferable.

Task 2: Estimation using a small number of transmitting points. Receiving point data in the selected transmitting point may be used, though fewer is preferable.

Essentially, both tasks aim to achieve accurate path loss (PL) prediction while minimizing the number of transmitting (Tx) and receiving (Rx) points used for training. To address these requirements, we focused on strategic subset selection for both Tx and Rx points, as described below and submit the same results for both the tasks

4.1 Transmitter Subset Selection

To ensure spatial diversity among transmitters, we employed the Farthest Point Sampling (FPS) algorithm, which iteratively selects Tx points that are maximally distant in 3D space from each other. This approach ensures coverage of the propagation environment with minimal redundancy.

Through experimentation, we observed that reducing the number of Tx points slightly increased the training and testing error, but the degradation was not substantial. Consequently, we found that using as few as one Tx point yielded comparable performance to larger subsets. This indicates that, for the given dataset, the path loss patterns across transmitters were highly correlated, allowing effective learning even with minimal Tx diversity.

4.2 Receiver Subset Selection

For the receiver subset, a random sampling strategy was adopted. The model performance was empirically evaluated across varying subset sizes. It was observed that iteratively reducing the number of Rx points to 33% of previous size (e.g., $900 \rightarrow 300 \rightarrow 100 \rightarrow 33$) retained most of the predictive performance while significantly reducing the dataset size. Consequently, using just 400 Rx points ($\approx$0.02% of the original dataset) was found to achieve an optimal trade-off between accuracy and computational efficiency.

So, for both Task 1 and Task 2, only a single transmitter point (Tx1) from the provided dataset was used for training, while a random subset of 400 receiver points was selected from within the corresponding file.

5 Estimation Accuracy and Generalization performance

We have same submission for both Task 1 and Task 2. The corresponding quantitative outcomes are summarized in Table 1 and Table 2, while the visual comparisons are presented in Figure 3 and Figure 4.

Table 1: Average RMSE per Frequency (Rounded to 3 Decimal Places)

Frequency	Tx 5	Tx 9	Tx 12	Tx 14	Tx 20	RMSE per Freq	Variance per Freq
800MHz	17.830	15.624	18.909	14.770	14.164	16.259	4.190
7GHz	20.259	19.017	20.627	16.562	15.937	18.280	4.589
28GHz	21.406	21.093	23.824	18.889	18.819	20.606	4.419

Table 2: Average RMSE per TxID (Rounded to 3 Decimal Places)

TxID	800MHz	7GHz	28GHz	Average RMSE per TxID	Variance per TxID
Tx 5	17.830	20.259	21.406	19.832	3.335
Tx 9	15.624	19.017	21.093	18.578	7.623
Tx 12	18.909	20.627	23.824	21.121	6.222
Tx 14	14.770	16.562	18.889	16.407	1.441
Tx 20	14.164	15.937	18.819	15.969	8.992

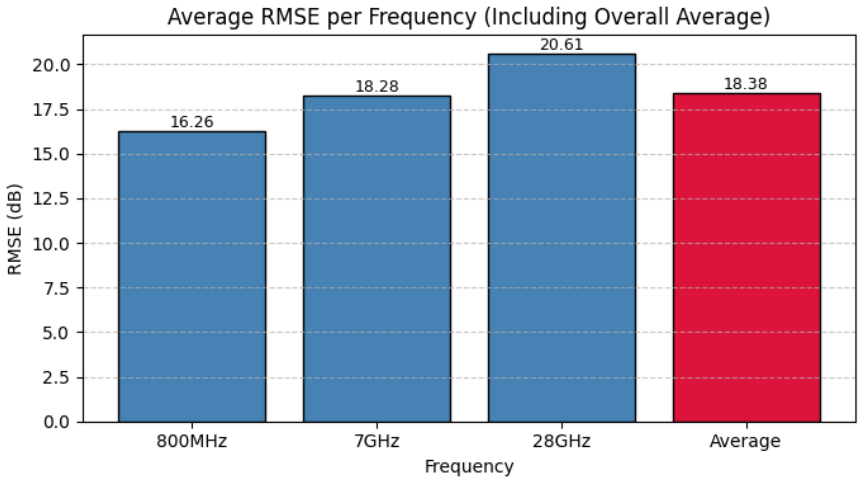

Figure 3: Average RMSE per Frequency (including overall average as a separate bar). Each bar represents the mean RMSE across all transmitters for a given frequency, and the red bar indicates the overall average RMSE across frequencies.

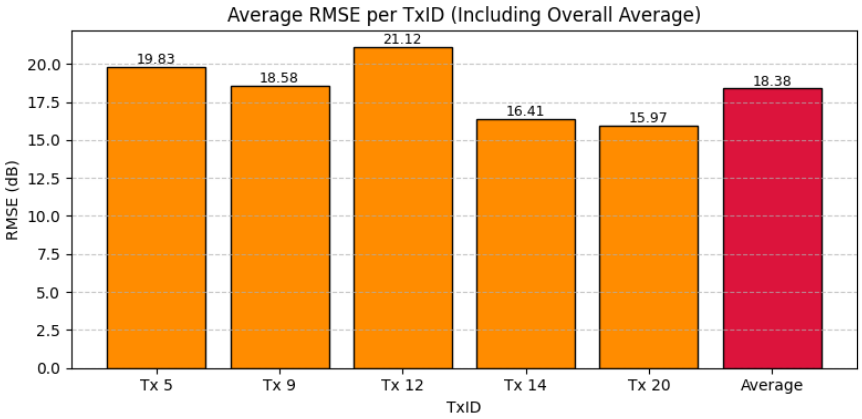

Figure 4: Average RMSE per TxID (including overall average as a separate bar). Each bar represents the mean RMSE across all frequencies for a given transmitter, and the red bar indicates the overall average RMSE across TxIDs.

References

[1] A. P. Mohamed et al., "Simulation-Enhanced Data Augmentation for Machine Learning Pathloss Prediction," ICC 2024 - IEEE International Conference on Communications, Denver, CO, USA, 2024, pp. 4863-4868, doi: 10.1109/ICC51166.2024.10622237.

[2] Y. Zhang, J. V. Krogmeier, C. R. Anderson and D. J. Love, "Large-Scale Cellular Coverage Simulation and Analyses for Follow-Me UAV Data Relay," in IEEE Transactions on Wireless Communications, vol. 23, no. 3, pp. 2396-2412, March 2024, doi: 10.1109/TWC.2023.3298546.