electronics

Article

Geospatial Feature-Based Path Loss Prediction at 1800 MHz in Covenant University Campus with Tree Ensembles, Kernel-Based Methods, and a Shallow Neural Network

Marta Moreno-Cuevas [1], José Lorente-López [1], José-Víctor Rodríguez [1,*], Ignacio Rodríguez-Rodríguez [2] and Concepción Sanchis-Borrás [3]

[1] Departamento de Tecnologías de la Información y las Comunicaciones, Universidad Politécnica de Cartagena (Member of European University of Technology EUT+), Antiguo Cuartel de Antigones, Plaza del Hospital, 1, 30202 Cartagena, Spain; marta.moreno@edu.upct.es (M.M.-C.); jose.lorente2@edu.upct.es (J.L.-L.)
[2] Departamento de Ingeniería de Comunicaciones, Universidad de Málaga, Avenida Cervantes, 2, 29071 Málaga, Spain; ignacio.rodriguez@ic.uma.es
[3] Facultad Politécnica, Universidad Católica de Murcia, 30107 Murcia, Spain; csanchis@ucam.edu
* Correspondence: jvictor.rodriguez@upct.es; Tel.: +34-968326548

Abstract

This paper investigates within-scene path loss prediction at 1.8 GHz in a smart-campus micro-urban environment using multivariate machine-learning (ML) models. We leverage an open measurement campaign from Covenant University (Nigeria) comprising three routes with per-sample geospatial predictors—longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height—and train Random Forests (RF), Gradient Boosting (GB), Support Vector Regression (SVR), Gaussian Processes (GP), and a shallow neural network (NN). A unified pipeline with 5-fold cross-validation (CV), seeded reproducibility, and Optuna-driven hyperparameter search is adopted; performance is reported as RMSE/MAE/R^2 (mean $\pm$ sd). To contextualize feature reliability, we include Pearson correlation heatmaps and Variance Inflation Factors (VIFs), a systematic ablation of predictors, and TreeSHAP beeswarm analyses on held-out splits. We also evaluate spatially aware validation (blocked CV within route and leave-one-route-out checks) to mitigate optimism due to spatial autocorrelation. Results show that multivariate ML consistently outperforms classical empirical formulas (COST-231, ECC-33) in this campus setting, with RF achieving the lowest errors across routes (RMSE $\approx$ 2.14/2.16/2.95 dB for X/Y/Z, respectively), while GB ranks second and kernel methods (SVR/GP) and the NN trail closely behind. Ablation confirms that distance plus coordinates drive the largest gains, with terrain/clutter providing route-dependent refinements. SHAP analyses align with these findings, highlighting stable, interpretable contributions of geospatial covariates. Spatial CV increases absolute errors moderately but preserves model ranking, supporting the robustness of conclusions. Overall, scenario-aware, multivariate ML yields material accuracy gains for smart-campus planning at 1.8 GHz.

Keywords: machine learning techniques; path loss prediction; radiowave propagation models; smart university campus

Academic Editor: Dimitra I. Kaklamani

Received: 8 September 2025
Revised: 10 October 2025
Accepted: 16 October 2025
Published: 20 October 2025

Citation: Moreno-Cuevas, M.; Lorente-López, J.; Rodríguez, J.-V.; Rodríguez-Rodríguez, I.; Sanchis-Borrás, C. Geospatial Feature-Based Path Loss Prediction at 1800 MHz in Covenant University Campus with Tree Ensembles, Kernel-Based Methods, and a Shallow Neural Network. *Electronics* **2025**, *14*, 4112. https://doi.org/10.3390/electronics14204112

1. Introduction

The proliferation of wireless communication systems has been a transformative element in contemporary society, profoundly impacting multiple aspects of daily life, such

as personal communication, access to information, transportation systems, health, and education. Thus, in recent decades, the increasing adoption of technologies such as cellular networks, Wi-Fi, the Internet of Things (IoT) and, more recently, 5G networks has generated a greater demand for connectivity in urban areas, where unique challenges for wireless network planning arise due to the presence of buildings, the variability of building materials, and the presence of multiple sources of electromagnetic interference [1]. In this context, propagation models become fundamental tools for the planning and design of wireless communication systems, as they allow for predicting the behavior of radio waves in complex and/or specific environments—in terms of coverage or signal loss—in order to guarantee the quality of service, spectrum efficiency, and sustainability of such wireless networks [2].

Traditionally, propagation models for the analysis of urban environments have been divided into so-called theoretical (deterministic) and empirical models. Thus, examples of theoretical models for contexts with the presence of buildings are those based on Physical Optics [3–5] or Geometrical Optics—Geometrical/Uniform Theory of Diffraction—[6–8] which are based on mathematical and physical principles describing the attenuation of signals in modeled environments and ideal conditions. However, such models present some limitations, such as their possible lack of accuracy in real scenarios or specific contexts, since by assuming ideal conditions and, therefore, simplifying the environment, they do not take into account the particular characteristics of the area under analysis. For their part, examples of empirical urban propagation models are those of Okumura–Hata [9] or the ECC-33 [10], which are based on experimental data obtained through extensive measurement campaigns—with the possibility of incorporating certain theoretical or mathematical adjustments, as is the case of the COST 231-Hata model [11], thus qualified as a semi-empirical model. However, although empirical models, developed on the basis of experimental measurements, can more accurately capture the specific characteristics of the environment, they also present certain limitations, such as their questionable applicability in environments that differ significantly from the original scenarios in which the measurements that gave rise to the model were taken.

Therefore, the previously described limitations of both theoretical and empirical propagation models have generated the need to explore more advanced approaches for path loss prediction in urban environments. In this sense, Artificial Intelligence and, more specifically, machine learning (ML) have emerged as powerful tools to address these limitations since, by harboring the ability to analyze large volumes of data (experimental measurements or geospatial data), they can discover complex patterns that cannot be captured by traditional models [12–16] and thus offer more accurate, adaptive, and generalizable (within the modeled scene and adaptable across scenes via re-training) propagation models than the latter, especially in specific contexts [17,18]. In this way, Piacentini et al. [19] presented a method to predict propagation losses in urban environments using ML and dimensionality reduction techniques. Their approach allows for improving the accuracy of predictions in complex scenarios, which is crucial for planning and optimizing wireless networks in dense urban areas. Subsequently, Timoteo et al. [20] proposed the use of ML algorithms for path loss prediction in urban environments. They demonstrated that the Support Vector Machine (SVM) algorithm can effectively model complex signal propagation relationships in cities, improving accuracy compared to traditional models. Aldossari and Chen [21] focused on predicting the propagation losses of wireless channel models in urban millimeter-wave communications using ML techniques. Their research contributed to the improvement of the accuracy of models at these frequencies, which are essential for high-speed communications in urban environments. Finally, Moraitis et al. [22] employed ML-based methods for propagation loss prediction in Long Term Evolution

(LTE) networks in urban environments. Thus, they evaluated different ML algorithms to improve the accuracy and adaptability of predictions in changing scenarios, highlighting the importance of these techniques for next generation mobile networks.

For their part, recent studies deploy advanced ML to predict path loss and radio maps in urban settings. In this way, Gupta et al. leveraged extensive 28 GHz Manhattan measurements and a feature pipeline combining LiDAR-derived street clutter with a convolutional autoencoder of 3-D buildings, cutting RMSE versus 3GPP and slope–intercept baselines in urban canyons [23]. Juang et al. proposed an explainable deep learning path profile model for urban 3.5 GHz deployments that augments log-distance line-of-sight (LoS) with a learned non-line-of-sight (NLoS) component, showing sizable error reductions and interpretable geometric factors [24]. Chen et al. introduced ACT-GAN, a Generative Adversarial Network (GAN)-based framework for radio-map construction that attains a state-of-the-art performance on urban benchmarks [25]. A complementary line explored long-range content interaction networks for radio-map prediction in cities, improving robustness to spatial context gaps [26]. For map completion from sparse monitors, Wen et al. used a graph neural network to predict and fill urban radio environment maps [27]. On the model aggregation front, Kwon et al. showed that neural network ensembles outperform single networks on outdoor propagation datasets with geospatial inputs [28]. Finally, Ma et al. presented a Cycle GAN-based urban radio-map estimation that learns bidirectional mappings between urban features and path loss maps, yielding accurate coverage prediction under limited labels [29].

On the other hand, an example of environments with the presence of buildings that house specific characteristics—and are becoming increasingly relevant—are the so-called smart university campuses, which represent physical and digital spaces that combine advanced information and communication technologies (ICT) with infrastructures to create more connected, efficient, and sustainable learning, research, and living environments [30]. Thus, these campuses, characterized by high heterogeneity of the physical environment that comprises them, not only host a high density of users—students, faculty, and administrative staff—but also an increasing number of connected devices that require robust and reliable wireless networks. In any case, a smart university campus is a specific example of what is known as a smart campus. In this work, "smart campus" denotes a geographically bounded urban micro-district (e.g., university, hospital/health-care precinct, corporate/technology park, or mixed-use estate) that is instrumented with digital connectivity and sensing (cellular/Wi-Fi/BLE, small cells, and edge services), exhibits heterogeneous but mapped morphology (buildings, canopies, open spaces, and pedestrian corridors), and provides high-quality geospatial layers (building footprints/heights, vegetation/clutter proxies, and elevation/terrain) under coherent site management. From a propagation perspective, such environments feature short-to-medium link distances, mixed LoS/NLoS conditions, and route-specific spatial gradients that can be captured by geospatial covariates. Therefore, planning wireless communications systems in these environments is critical to ensure high-quality connectivity and meet the technological needs of users. In this way, from Wi-Fi and 5G networks to advanced applications such as augmented reality and intelligent building management systems, smart campuses represent a microcosm of smart cities but with specific characteristics that make their planning unique and challenging. In this sense, traditional propagation models, both theoretical and empirical, have been employed in planning wireless networks on smart campuses. However, they present limitations in terms of accuracy and adaptability. For example, Popoola et al. used classical empirical models in [31] —Okumura–Hata/COST-231, ECC-33, and Egli—to compare predicted vs. measured path loss along three surveyed routes within a smart university campus; however, while simple and transparent, these fixed-form curves struggled to track local morphology.

Moreover, the Okumura–Hata model has been used to estimate coverage in open areas and spaces with moderate buildings on university campuses, but its accuracy decreases in scenarios with higher building density [32]. The COST 231-Hata model has also been used in GSM network planning studies on university campuses, with relatively accurate predictions in urban areas within the campus, but on the other hand, it lacks adaptability for dynamic environments [33]. Again, such limitations underscore the need for more advanced approaches such as ML-based propagation models for wireless communication system planning in specific contexts such as smart university campuses. More recently, Muñoz et al. presented in [34] a campus-focused, traditional path loss characterization in an outdoor corridor of a smart university environment at 850 MHz and 3.5 GHz, using a USRP-based channel sounder with omnidirectional antennas to collect extensive time-domain measurements over 2–67.5 m Tx–Rx separations. They fit the Close-In (CI) and Floating-Intercept (FI) models via MMSE, reporting path loss exponents $\approx$ 2.21–2.41 and using the results to support IoT/5G corridor deployment studies. This study is a strong empirical baseline, but its limitation is that, by design, it is a single scenario (one corridor, primarily LOS). A further campus-focused empirical contribution is the GSM-band smart-campus study published by Famoriji and Shongwe in [35], which reports a measurement-driven path loss characterization tailored to legacy GSM operation in a university environment. The approach follows the classic paradigm—collect received power traces in situ and fit simple, model-driven curves for planning—providing a practical baseline for coverage estimation in a specific campus layout. In this way, its strengths are methodological simplicity and direct relevance to GSM planning; however, it relies on fixed-form fits and does not integrate rich geospatial covariates.

On the other hand, regarding ML-based approaches for path loss prediction in smart university campuses, Singh et al. applied Artificial Neural Networks (ANN)/Random Forest (RF)-type regressors to smart-campus measurements in [36], reporting improvements over empirical baselines but with limited discussion of geospatial feature engineering beyond distance/coordinates and considering a single train/test split. More recently, Khalili et al. confirmed, in [37], through black-box and glass-box ML models, the value of ML over closed-form models but did not explicitly exploit all GIS covariates (e.g., route-local clutter height averages) available for site-specific prediction.

In light of the above, the present work aims to implement and compare different radio wave propagation models, generated using multivariate ML techniques which, unlike previous works, consider multiple geospatial features (predictors) as inputs (longitude, latitude, altitude, elevation, transmitter (Tx)–receiver (Rx) distance, and clutter height), for the prediction of signal losses in a smart university campus. For this purpose, open access data from a measurement campaign carried out at 1800 MHz at Covenant University, Nigeria [38], will be used. Thus, multivariate ML models based on algorithms such as Support Vector Regression (SVR, a different use of SVM considered for regression tasks), Gaussian Processes (GP), RF, Gradient Boosting (GB), and neural networks (NNs)—due to their good performance and adaptability in complex problems such as path loss prediction—will be obtained, and their fit will be compared with the measurements through the Root Mean Square Error (RMSE) parameter. Additionally, the performance of the models obtained will be compared with that of traditional models such as COST 231-Hata and ECC-33. Finally, an innovative relevance ranking of the different parameters (predictors) that complement the signal loss value at each measurement point will be carried out for the multivariate ML model with the lowest RMSE in order to evaluate the importance of each of them when training and efficiently building the best performing ML model. In this way, our study is scoped deliberately to a smart-campus micro-urban environment at 1800 MHz, using an openly available dataset with rich geospatial predictors. Therefore, it should be

noted that, within this scope, our goal is not to derive a campus-agnostic, extrapolatable predictor. Rather, our goal is methodological: on the one hand, to assess the extent to which modern ML models (RF, SVR, GP, GB, and NN) can surpass widely used empirical models (COST-231, ECC-33) for path loss prediction in campus-scale planning (pointing out which ML model performs best), and, on the other hand, to demonstrate the incremental value of explicitly conditioning on multiple geospatial descriptors—longitude, latitude, elevation, altitude, Tx–Rx distance, and clutter height—as inputs to multivariate ML models for within-scene path loss prediction in smart university campuses.

2. Methodology

2.1. Data Description

The data considered in this work were obtained from a measurement campaign performed at the Covenant University smart campus, Ota, Ogun State, Nigeria, by Popoola et al. [38]. Through such campaign, the propagation losses at each measurement point—together with other relevant parameters such as longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height (the latter defined as the mean height of buildings within a fixed-radius buffer around each measurement location)—were monitored along three different routes (X, with 655 final data for each parameter; Y, with 739 final data for each parameter; and Z, with 1447 final data for each parameter) as a mobile receiver moved away from each of the three transmitting antennas at 1800 MHz. In this sense, the routes were selected in order to encompass the different topographies and urban morphologies present throughout the campus. Path loss data were collected with receiver equipment consisting of a Sony Ericsson W995 cell phone configured as a mobile station (with a processing speed of 369 MHz, a 930 mAh Li-Po battery, sensitivity of -100 dBm, and an effective height of $\approx$1.5 m above ground level), Ericsson TEMS Investigation software (version 9.0), a Garmin BU-353-S4 GPS receiver to determine the location of the mobile receiver at any given time, and a personal laptop computer with a 64-bit Windows operating system, 4 GB of RAM, and an Intel Core i5 M520 @2.40GHz processor. TEMS records per carrier received power from the live network. In this sense, no additional receiver-side bandwidth expansion or averaging beyond TEMS defaults was applied; thus, the effective measurement bandwidth corresponds to the serving GSM-1800 carrier used during acquisition. All the equipment was carefully installed in a vehicle, which was traveling at a constant speed of 40 km/h along the three routes in order to minimize the Doppler effect. In addition, longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height were obtained at each measurement point from a Digital Terrain Map (DTM) using the ATOLL radio-planning tool [38], thereby enabling a consistent geospatial description of the smart-campus micro-urban morphology along the three routes (Tx–Rx distance for feature engineering and metrics was computed as WGS-84 geodesic from longitude/latitude, so ATOLL operated in a UTM/WGS-84 projected workspace. On campus-scale routes, geodesic vs. planar distances differ by <0.05%, which is negligible for the reported fits and results). On the other hand, the base station used during the measurements bears the A2G identifier (with sectors A2GS1, A2GS2, and A2GS3), is located at the geographical coordinates of longitude 3.162867 and latitude 6.675068, and reaches a height of 50 m [39]. The base station consists of an Ericsson RBS 2116 transceiver and three directional sectorial antennas with a gain of 18 dBi, a 65° beamwidth, vertical polarization, and the 1710–1880 MHz operating frequency range. In this way, measurements were taken within the main lobes of directional base station antennas to ensure consistent coverage. Moreover, the transmission power used in the measurements was 43 dBm [39]. It is also worth noting that the experimental RF measurements performed by Popoola et al. [38] were carried out under good weather conditions, and the distances covered by the three

routes were long enough for the receiver to reach its noise level, implying SNR spans from high-coverage conditions close to the site to near-sensitivity values at the route tails; in this way, no post hoc SNR filtering was performed. Furthermore, it should be mentioned that receiver power-meter accuracy, dynamic range, and Automatic Gain Control (AGC) linearity were verified against a traceable RF source using an inline step attenuator, and any constant offset was quantified and subsequently compensated in post-processing. Antenna type and nominal gain were logged, together with connector/cable losses at 1800 MHz, which were incorporated as fixed corrections. Logging parameters (sampling rate and temporal averaging) were held constant across routes, and GPS clock alignment with the TEMS logger was confirmed within the instrument tolerance. Before each run, GPS lock and time alignment were verified; a short sanity pass on line-of-sight segments was performed to confirm monotonic Received Signal Strength (RSS) decay with distance and operation within the linear AGC region (no saturation). During acquisition, sustained GPS fix and stable logging were periodically checked; after each run, sample counts, route-length histograms, and the presence of noise floor regions at long ranges were cross-validated. These steps ensured level traceability, bounded measurement uncertainty, and protected against operational drift during field work. In this way, the W995 was interfaced through TEMS, which provides device-specific RF calibration to convert raw AGC reads to absolute power (dBm). Prior to field runs, the transmit EIRP was established from bench measurements (power amplifier output with a calibrated power meter), subtracting Tx-chain losses and adding antenna gain. In the field, a side-by-side cross-calibration was performed at fixed reference points using a calibrated reference receiver, confirming a constant offset between handset-reported RSS and the reference across a short distance/level sweep within the handset's linear operating region; that offset was applied to all logs. Path loss was computed as PL (dB) = EIRP (dBm) + $G_{Rx} - L_{Rx} - P_{Rx}$ (dBm). For each calibration point and periodically along the routes, the effective noise floor N_{eff} was measured in-band, either during idle time slots or on an adjacent idle channel, keeping the same RF chain and settings; SNR was then $P_{Rx} = N_{eff}$ (dB).

Regarding the above, the campaign yields thousands of geo-referenced samples across the routes, with dispersion and distribution summaries (means, standard deviations, and ranges) reported per route in [38]; this breadth captures the intrinsic variability of the channel as the receiver moves away from each serving site. The controlled driving speed and weather conditions reduce measurement-induced variability, while the route lengths ensure adequate dynamic range (including proximity to the receiver noise floor).

On the other hand, we did not apply temporal smoothing to the path loss series so as to preserve the physical variability of the measurements. Instead, we performed minimal, transparent quality control prior to modeling: removal of duplicate (in cases where there were slightly different path loss values for the same latitude and longitude location, the average was taken) or malformed records and rows with missing geodetic coordinates or distances; co-registration of all geospatial layers to a common WGS-84 reference; verification that logged points fall within the intended route masks/coverage lobes described in [38]; and route-wise standardization of predictors within each cross-validation fold (transform fit on training fold only) to avoid leakage. The underlying measurement protocol and DTM attributes are already validated in [38], which provide route-level statistics and visual checks (frequency distributions, correlation analyses).

Figure 1 shows the considered X, Y, and Z routes, and Figure 2 shows the signal losses (dB) registered along all routes during measurements, including overlaying iso-contour lines every 10 dB to improve readability of the path loss heatmap (axes in both figures indicate geographic directions: x—longitude (°E), y—latitude (°N)).

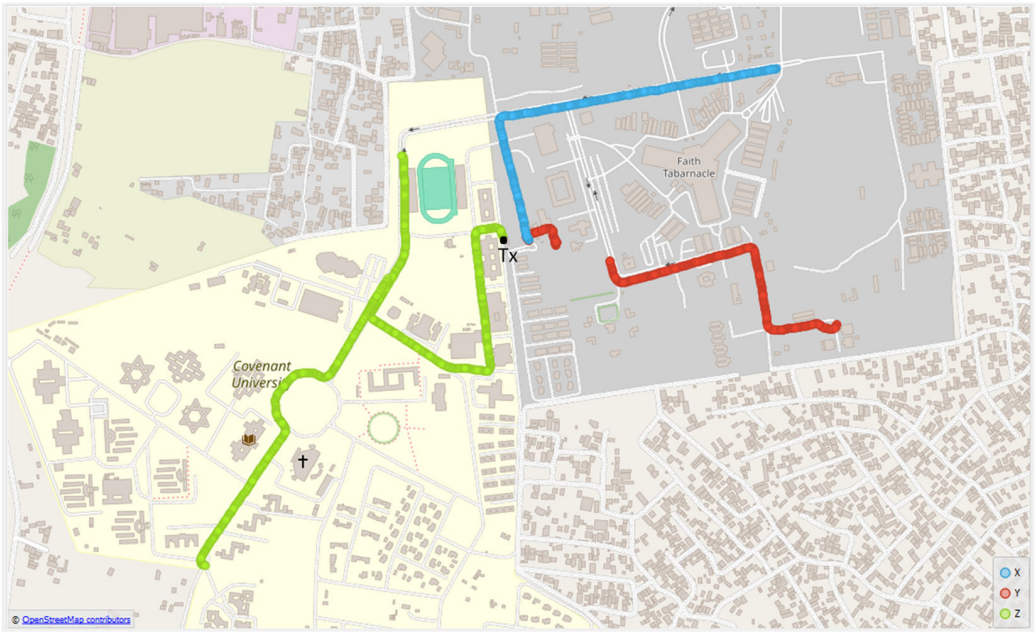

Figure 1. The three routes considered: X (blue), Y (red), and Z (green).

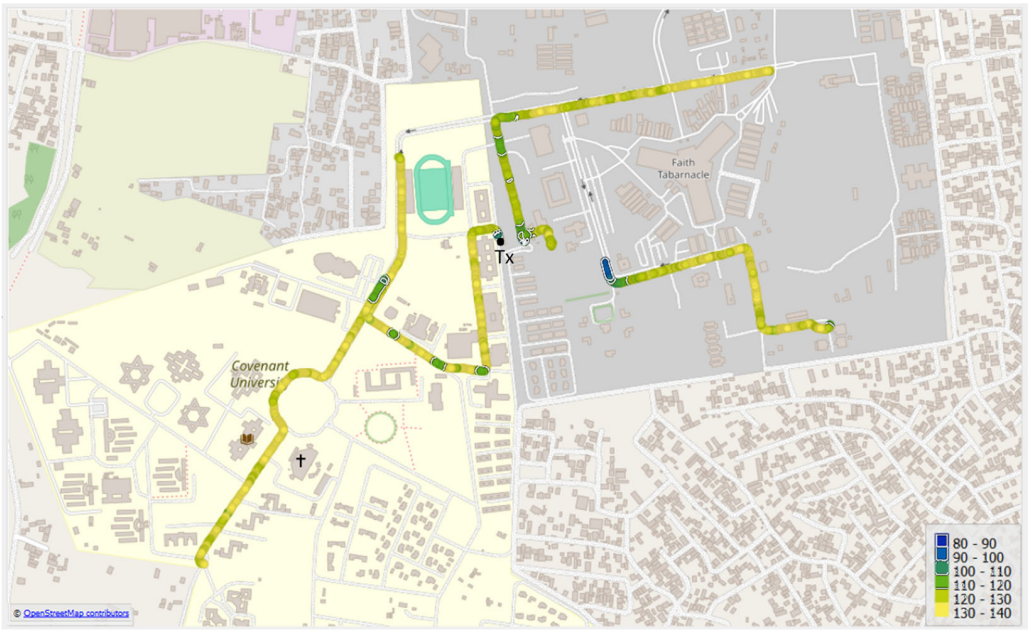

Figure 2. Path losses (dB) registered along all routes during measurements. Iso-contour lines every 10 dB are overlaid (white lines with a thin black halo).

The colored dots in Figure 2 represent the propagation losses (in dB) registered along each route. In this sense, as can be seen in the color legend located at the bottom right of such figure, the blue tones indicate the lowest signal losses, while the yellow tones correspond to the highest losses. On the other hand, a black dot can be identified in each figure with the label 'Tx', which represents the location of the base station used to transmit the 1800 MHz signals during the measurements.

To assess potential redundancy among geospatial predictors, we report Pearson correlation heatmaps per route (X–Z), (Figures 3–5, respectively), and Variance Inflation Factors (VIFs) (Table 1).

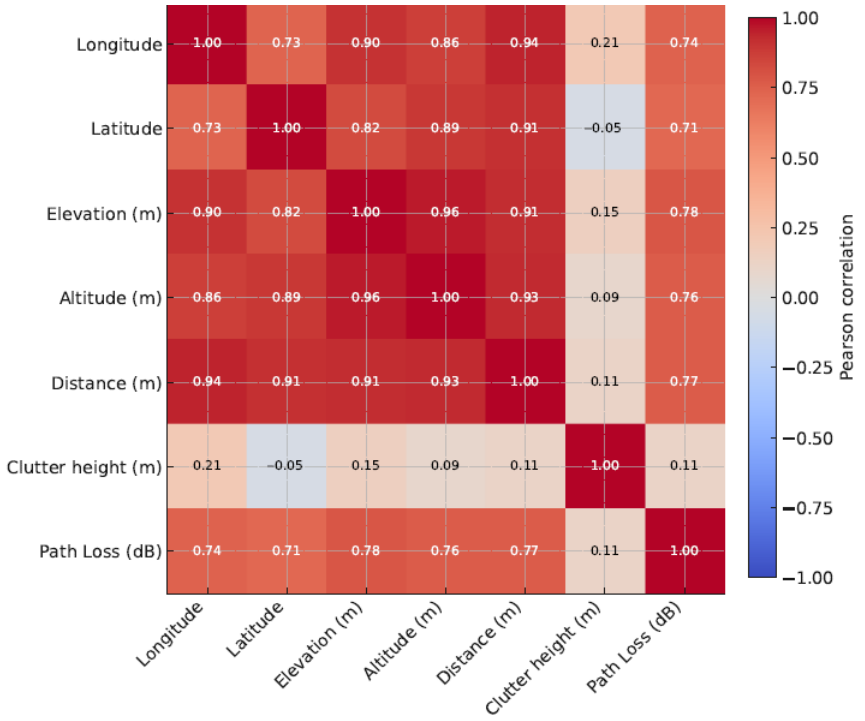

Figure 3. Pearson correlation heatmap considering the six geospatial predictors for Route X.

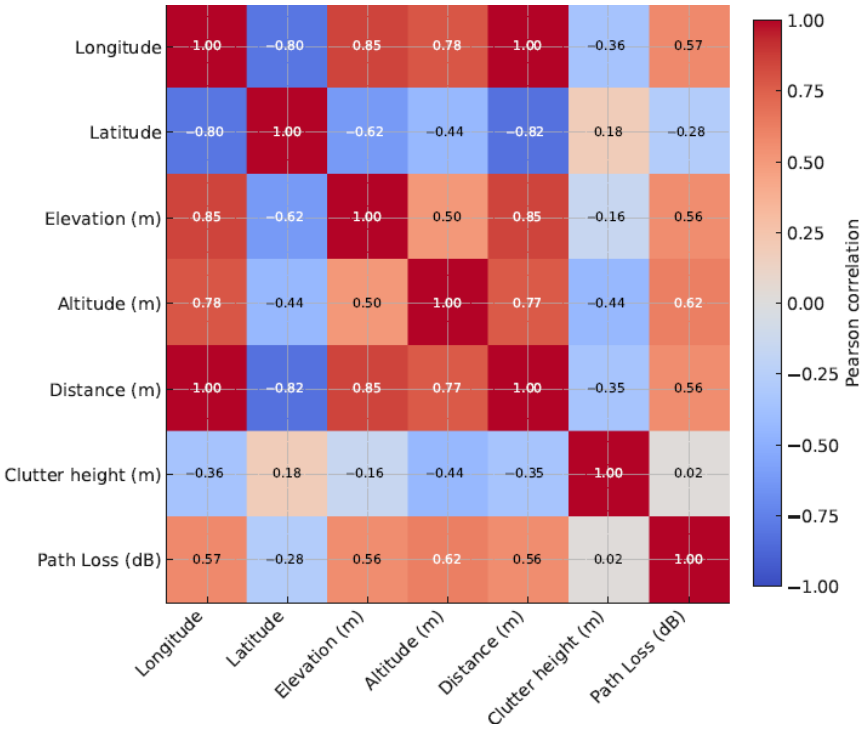

Figure 4. Pearson correlation heatmap considering the six geospatial predictors for Route Y.

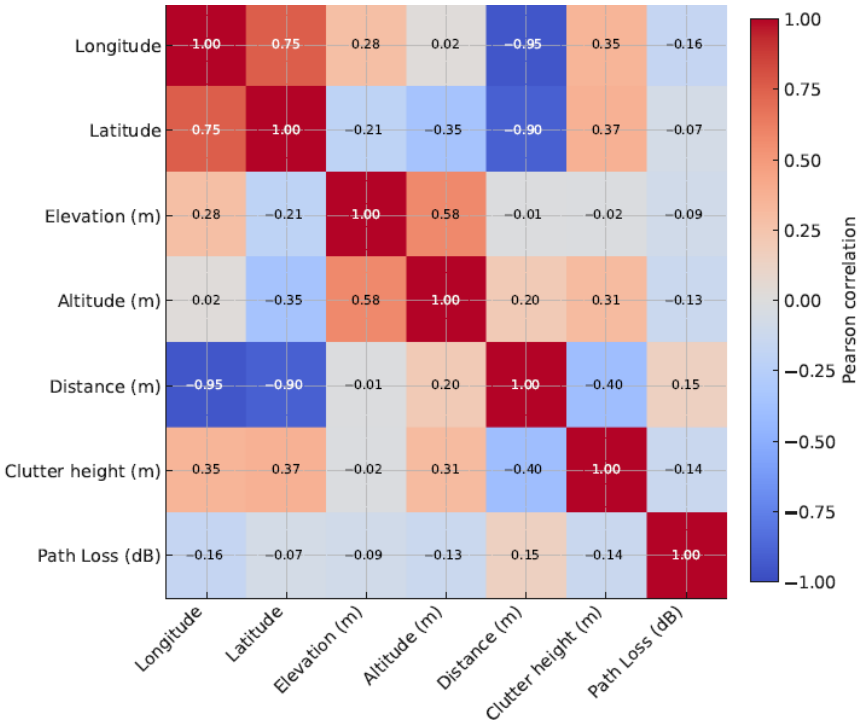

Figure 5. Pearson correlation heatmap considering the six geospatial predictors for Route Z.

Table 1. Variance Inflation Factors (VIFs) per route.

Feature	Route X	Route Y	Route Z
Distance (m)	187.81	20,191.4	101.66
Longitude	87.98	17,834.32	63.69
Latitude	56.07	92.37	11.73
Altitude (m)	21.54	10.12	3.59
Elevation (m)	19.64	7.0	2.98
Clutter height (m)	1.22	1.41	1.99

As expected, geographic coordinates correlate strongly with Tx–Rx distance (e.g., $|r| \approx 0.94$–1.00 on X–Y), and elevation is highly collinear with altitude ($|r| \approx 0.96$ on X). Correspondingly, VIFs are high for distance and longitude/latitude (e.g., X: distance ≈ 188, longitude ≈ 88; Y: distance $\approx 2.0 \times 10^4$, longitude$\approx 1.8 \times 10^4$), while clutter height remains low ($\approx$1–2). However, the systematic ablation study (progressive addition of features) performed in Section 3.5 below will corroborate that augmenting distance with longitude and latitude produces the largest error reductions, with morphology-dependent gains from terrain and small but consistent clutter effects, thereby supporting the multivariate geospatial design despite collinearity among some predictors.

2.2. Implementation of the ML Models

From the available experimental data, a series of multivariate predictive models were implemented in Python (*scikit-learn* library), where the dependent variable, i.e., the one to be predicted, corresponds to the propagation losses at each measurement point along the X, Y, and Z routes. Likewise, as independent variables or predictors, additional parameters obtained during the measurements were considered, such as longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height. In this way, these predictors allow the algorithm to perform more accurate estimates by identifying the relationships between the characteristics of the environment and the propagation losses.

For the building of the models, five ML algorithms were selected that are widely recognized for their robustness and their ability to handle data of a nonlinear nature, as is the case of signal losses. In this sense, the algorithms used were RF, SVR, GP, GB, and NN. As previously stated, the use of these algorithms is justified by their performance and adaptability in complex problems such as propagation loss prediction. First, RF is a decision tree-based algorithm that combines multiple trees to perform more robust and generalizable predictions. It is especially useful in the prediction of nonlinear phenomena because it can handle large volumes of data and capture interactions between predictors without the need for prior assumptions about the distribution of the data.

For its part, SVR, through its ability to use nonlinear kernels, allows the data to be projected into a higher dimensional space where complex relationships between variables can be identified. Furthermore, GPs are probabilistic algorithms that generate accurate predictions by modeling the joint distribution of observations with Gaussian functions, being particularly effective in limited datasets with inherent noise. On the other hand, GB builds an additive predictive model by sequentially fitting weak learners (often decision trees) to the negative gradients of a chosen loss function—i.e., it performs gradient descent in function space—typically with shrinkage and other regularization to reduce overfitting. Finally, NNs approximate a target function by composing layers of linear transformations and nonlinear activations, learning parameters end-to-end by minimizing a loss via gradient-based optimization with backpropagation to extract hierarchical features. In any case, a more extensive description of such algorithms can be found in [40] for RF, [41] for SVR, [42] for GP, [43] for GB, and [44] for NN.

The process of building each model was carried out using a 5-fold cross-validation (CV). For this, the dataset of each route was randomly divided into two subsets, so that 80% of the data was used to train the model, while the remaining 20% was reserved for model validation. Thus, this procedure was repeated five times, ensuring that the data reserved for the validation phase in each fold were always different. In this way, the cross-validation technique guarantees that the model is evaluated objectively throughout different training–validation combinations, favoring greater reliability of the final model in the predictions, as well as ensuring that the model does not suffer from overfitting problems.

On the other hand, in order to evaluate the accuracy and performance of the models built, the RMSE parameter, which measures the average difference between the model predictions and the real values registered in the measurements, was used as a reference metric. In this way, let each measurement be (x_i, y_i), $i = 1, \ldots, n$, where $x_i = [\lambda_i$ (longitude), ϕ_i (latitude), A_i (altitude), Z_i (elevation), d_i (Tx–Rx distance), H_i (clutter height)$]^\top \in \mathrm{R}$, and $y_i = L_i$ is the measured path loss (in dB), which is the dependent variable (target). Given predictions $\hat{y}_i = \hat{f}(x_i)$,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Finally, for the ML model that obtained the lowest total RMSE (average of the RMSEs of each fold) in every one of the routes, a ranking of importance of the predictors when building the model was performed (using a univariate approach that evaluates how relevant the contribution of each predictor is separately in the implementation of the model). In this way, this analysis makes it possible to identify which of the independent parameters (longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height) have the greatest weight and influence on the implementation of the multivariate predictive model, thus facilitating the interpretation of the results and highlighting the key factors affecting propagation losses in the evaluated environment.

2.3. Formulation of the ML Models

Let $y = PL_{dB}$ denote the site-specific path loss (in dB) at 1800 MHz. For each measurement/prediction point we form, x = [λ (longitude), ϕ (latitude), A (altitude), Z (elevation), d (Tx–Rx distance), H (clutter height)]. Inputs are standardized (zero mean, unit variance) before model fitting; the output remains in dB. Models are trained per route to capture route-wise morphology and evaluated with 5-fold CV. Frequency (1800 MHz) is fixed and not an input; it enters through the data.

2.3.1. SVR

General Formulation

We learn $f(x) = w^T\varphi(x) + b$ by minimizing the ε-insensitive loss,

$$\min_{w,\, b,\, \xi,\, \xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

subject to

$$\begin{gathered} y_i - w^T\varphi(x_i) - b \le \varepsilon + \xi_i, \\ w^T\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, \end{gathered} \tag{3}$$

where $\varphi(\cdot)$ is the feature map of x (possibly implicit); w, b are the model parameters (weights and bias), $\varepsilon > 0$ is the tube width for ε-insensitive loss (tolerance), ξ_i, ξ_i^* are the slack variables for deviations beyond ε, and $C > 0$ is the regularization constant (penalizes slacks). The kernel form (via the representer theorem) would be

$$\hat{f}(x) = \sum_{i=1}^{n}(\propto_i - \propto_i^*)k(x_i, x) + b, \tag{4}$$

where $\propto_i, \propto_i^*$ are the dual variables from the quadratic program, and $k(x_i, x)$ is the positive definite kernel operating on the geospatial features:

$$k(x_i, x) = \sigma_f^2 exp\left(-\frac{1}{2}\sum_{j=1}^{6}\frac{\left(x_j - x_j^{'}\right)^2}{l_j^2}\right) \tag{5}$$

with $\sigma_f^2 > 0$ as the signal variance and $l_j > 0$ as the length scales per feature (λ, ϕ, A, Z, d, and H).

Campus-Specific Formulation

With an RBF kernel $k(x, x') = \exp\left(-\|x - x'\|^2/\left(2\updownarrow^2\right)\right)$, the length scale ℓ acts as a spatial smoothing scale over the standardized geospatial features: larger ℓ prioritizes broad campus-wide trends (dominated by d and λ/ϕ), whereas smaller ℓ allows finer local corrections driven by terrain/clutter. The ε-insensitive loss captures measurement jitter/small-scale fading; the regularization C trades off smoothness vs. fidelity. This mapping makes SVR particularly suitable for short-to-medium distances with moderate non-stationarity typical of campus routes.

2.3.2. GP

General Formulation

Assume a GP prior $f \sim \mathcal{GP}(0.\, k)$ and additive Gaussian noise,

$$y_i = f(x_i) + \varepsilon_i,\ \ \varepsilon_i \sim \mathcal{N}\left(0, \sigma_n^2\right) \tag{6}$$

and let $y = [y_1, \ldots, y_n]$ be the target vector, $\mathcal{K} \in \mathbb{R}^{nxn}$ the kernel matrix, with $\mathcal{K}_{ij} = k(x_i, x_j)$, $k_* = [k(x_1, x_*), \ldots, k(x_n, x_*)]^T$, $\sigma_n^2 > 0$ the noise balance, and k the kernel on the 6-D geospatial features. Then, the predictive mean and variance at a test point x_* are

$$\mu(x_*) = k_*^T \left(\mathcal{K} + \sigma_n^2 I\right)^{-1} y,\ \sigma^2(x_*) = k(x_*, x_*) - k_*^T \left(\mathcal{K} + \sigma_n^2 I\right)^{-1} k_* \tag{7}$$

Campus-Specific Formulation

We use a zero-mean GP with ARD kernel (RBF/Matérn),

$$k(x,\ x') = \sigma_f^2 exp\left(-\frac{1}{2}\sum_j \frac{\left(x_j - x_j^{'}\right)^2}{\updownarrow_j^2}\right) \tag{8}$$

plus a nugget σ_n^2 for observation noise. The feature-wise length scales $\updownarrow_j$ quantify how quickly path loss varies along distance, coordinates, and terrain/clutter; smaller $\updownarrow_j$ indicate stronger local sensitivity (e.g., along elevation for routes with relief). The nugget σ_n^2 absorbs residual small-scale fading and logging noise. In this campus scenario, the ARD parameters provide a physically interpretable summary of which geospatial dimensions govern route-dependent variability.

2.3.3. RF

General Formulation

An RF averages M randomized regression trees:

$$f(x) = \frac{1}{M}\sum_{m-1}^{M} T_m(x) \tag{9}$$

where M is the number of trees and $T_m(\cdot)$ is the m-th CART regression tree trained on a bootstrap sample of $\{(x_i, y_i)\}$ and a random subset of features considered at each split. In this way, each split minimizes within-node squared error, and each leaf predicts the mean of the y_i in that leaf.

Campus-Specific Formulation

We use an ensemble of regression trees minimizing in-node MSE in dB^2. Subsampling of features at each split (e.g., $\sqrt{p}$) encourages diverse trees that explore distinct geospatial subspaces (e.g., splits on d vs. on λ/ϕ or terrain). Tree depth controls the interaction order among covariates: shallow trees mainly model distance-dominated decay, whereas deeper trees can capture local morphology interactions (e.g., distance $\times$ elevation). In the campus setting—short-range links and mixed LoS/NLoS—this yields smooth, large-scale trends with route-specific corrections where terrain/clutter vary. Ensembling reduces variance due to small-scale fading and mitigates overfitting when correlated predictors (distance $\leftrightarrow$ coordinates; elevation $\leftrightarrow$ altitude) are present.

2.3.4. GB

General Formulation

Additive model of M weak learners $\{h_m\}$ (typically shallow trees) fit to the negative gradients (residuals) of a loss, with a squared error:

$$\begin{gathered}\hat{f}_0(x) = \operatorname{argmin}_c \sum_i (y_i - c)^2 = \overline{y}, \\ r_i^{(m)} = y_i - \hat{f}_{m-1}(x_i) \text{ (residuals)}, \\ h_m \leftarrow \text{fit a tree to } \left\{ \left(x_i, r_i^{(m)} \right) \right\}_{i=1}^{n}, \\ \gamma_m = \arg\min_{\gamma} \sum_i \left[y_i - \hat{f}_{m-1}(x_i) - v\gamma h_m(x_i) \right]^2, \\ \hat{f}_m(x) = \hat{f}_{m-1}(x) + v\gamma h_m(x),\ m = 1, \ldots, M\end{gathered} \tag{10}$$

where $v \in (0, 1]$ is the learning rate (shrinkage), $h_m(\cdot)$ is the weak learner at stage m (tree on geospatial x), γ_m is the line-search step size for stage m, and M is the number of boosting iterations (trees).

Campus-Specific Formulation

GB builds an additive model

$$f_M(x) = \sum_{m=1}^{M} v h_m(x) \tag{11}$$

where each shallow tree h_m is fit to the negative gradient of the loss (here, squared error in dB) computed on the residuals of the previous ensemble, and $\nu \in$ (0,1] is the learning rate. In our campus setting, shallow trees with limited max depth capture low-order interactions among distance, coordinates, and terrain/clutter height, while the stage-wise boosting refines local corrections in mixed LoS/NLoS micro-contexts without requiring deep trees.

2.3.5. NN

General Formulation

A feedforward Multi-Layer Perceptron with L hidden layers operates on x as follows:

$$\begin{gathered}h^{(0)} = x,\ h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)} \right),\ l = 1, \ldots, L, \\ \hat{f}(x; \theta) = W^{(L+1)} h^{(L)} + b^{(L+1)}\end{gathered} \tag{12}$$

and training minimizes a regularized squared loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i; \theta) \right)^2 + \lambda \|\theta\|_2^2 \tag{13}$$

where $h^{(l)}$ is the hidden activation vector at layer l, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector at layer l, respectively, $\sigma(\cdot)$ is the element-wise nonlinearity, $\theta = \left\{ W^{(l)}, b^{(l)} \right\}_{l=1}^{L+1}$ are all trainable parameters, and $\lambda \geq 0 : l_2$ (weight decay) is the regularization strength. In this way, optimization uses gradient-based methods with backpropagation.

Campus-Specific Formulation

We adopt a single-hidden-layer network (ReLU/tanh) with H units, L2 regularization, and early stopping. This provides smooth nonlinear regression while respecting the limited number of campus samples, avoiding the data demands of deep Convolutional Neural Networks (CNNs)/Recurrent Neural Networks (RNNs) architectures. Capacity H is set to

capture distance-dominated trends plus morphology corrections without overfitting highly collinear inputs.

2.4. Framework and Hyperparameter Tuning of the ML Models

We adopt an end-to-end pipeline tailored to site-specific path loss at 1800 MHz in a bounded campus micro-district:

- *Data assembly and synchronization*: Logged measurements are time-aligned and geo-referenced; points are joined with geospatial layers to form x = [λ, ϕ, A, Z, d, H] and target y = PL_{dB}.
- *Preprocessing*: Inputs are standardized (zero mean, unit variance). Frequency is fixed by the scenario and not an input feature.
- *Model family*: We evaluate RF/GB (tree ensembles), SVR (RBF), GP (ARD), and a shallow neural network sized for low-N tabular data.
- *Validation protocol*: Models are trained per route using 5-fold CV with identical folds across models; RMSE (dB) is the selection criterion.
- *Interpretability and diagnostics*: We compute SHAP beeswarm on held-out splits, conduct systematic ablations of features, and check correlation/VIF to contextualize importance under multicollinearity.

Therefore, our contribution is a scenario-aware ML framework that integrates the following:

- GIS-informed, multivariate features at the campus scale (distance + coordinates + terrain + clutter height) engineered for site-specific prediction;
- Route-wise model selection with a single, reproducible protocol (same folds/seed across models) and explicit convergence monitoring, avoiding over-optimism from ad hoc splitting;
- Domain-aware search spaces that encode physical priors (e.g., smoothing/interaction constraints) to prevent overfitting to small-scale fading;
- Interpretability-first evaluation (SHAP beeswarm on held-out + ablations + VIF), turning feature importance into actionable design guidance (e.g., when terrain matters most);
- Deployment-oriented reporting, including computed footprints suitable for campus-scale updates.

On the other hand, we used Optuna (TPE sampler with pruning, version 4.3.0) to perform intelligent, sample-efficient search within well-motivated ranges. The objective minimized the 5-fold cross-validated RMSE on the training data (route-wise), with scaling of features (standardization) applied inside each fold to avoid leakage. Moreover, we evaluated 100 trials per model (pruned early when clearly sub-optimal). For each model × route, we run a bounded search over physically meaningful ranges (e.g., RF depth ↔ interaction order among geospatial covariates; SVR/GP length-scales ↔ spatial smoothing of distance/coordinates/terrain). Searches use a stochastic strategy (random/TPE) with early pruning of clearly sub-optimal trials. We track the incumbent best across trials to assess convergence; representative convergence curves (per route) are reported below in the Results section and show rapid early gains followed by a stable plateau, indicating sufficient trial budgets and stable selections.

Figure 6 formalizes and visually explains the hyperparameter optimization procedure (HPO) used in this study. The figure is model-agnostic and applies to random, grid, or Bayesian search; it is consistent with our actual pipeline, in which candidate hyperparameters are evaluated via K-fold cross-validation, and the best ("incumbent") configuration is selected based on mean CV RMSE. For clarity, the figure is schematic (we set K = 4 in the panel sketch purely for legibility); our reported experiments use K = 5.

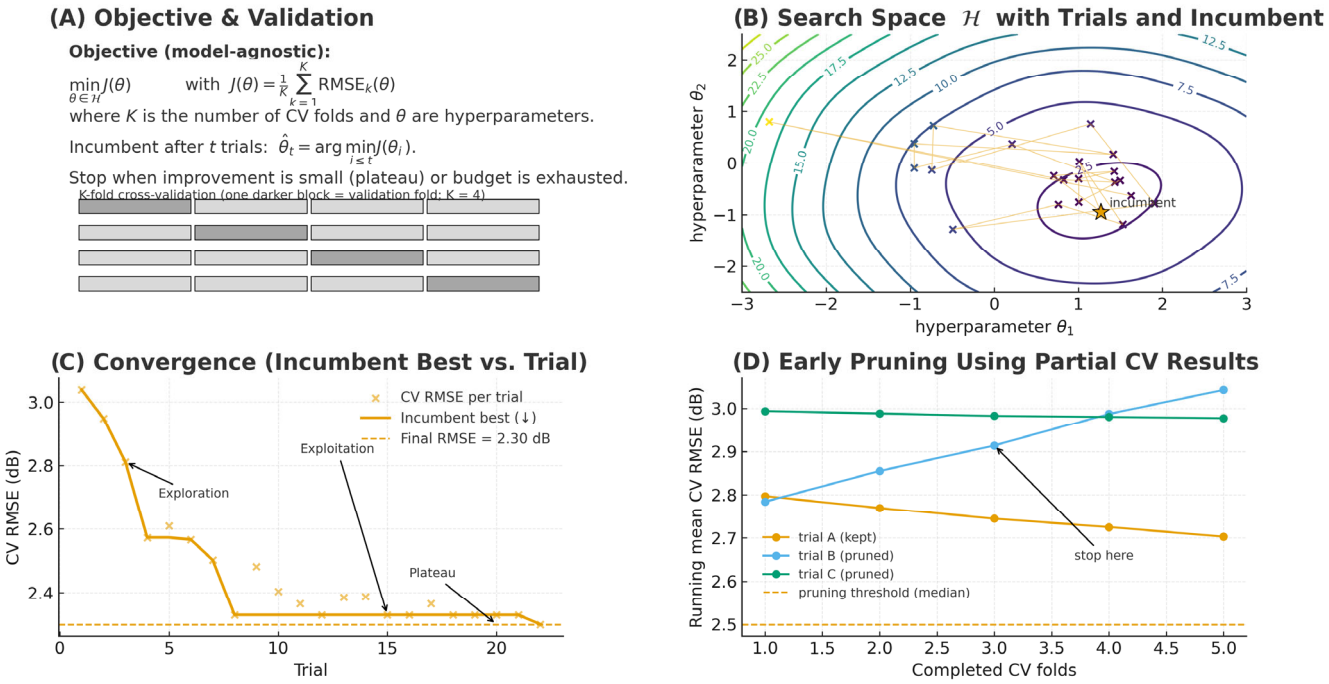

Figure 6. Hyperparameter tuning—theoretical concept (model-agnostic).

Specifically

- Panel A (Objective and Validation): States the HPO objective $min_{\theta \epsilon H}\, J(\theta)$ with $J(\theta) = \frac{1}{K} \sum_{k=1}^{K} RMSE_k(\theta)$, i.e., the mean CV error across k folds (θ are hyperparameters), which is exactly the selection criterion we report in the paper.
- Panel B (Search Space and Incumbent): Depicts sampling of the hyperparameter space and the evolution of the incumbent (best-so-far) configuration as trials progress; this covers random/grid/Bayesian strategies without committing to a specific one.
- Panel C (Convergence): Shows the per-trial CV RMSE and the incumbent-best curve converging towards a plateau, illustrating the usual exploration → exploitation transition and the stopping criterion once improvements become negligible—mirroring the stabilization we observe in practice.
- Panel D (Early Pruning with Partial CV): Illustrates a standard pruning heuristic based on partial-fold performance (e.g., median stopping). We include this to clarify the concept; in our experiments, we evaluate candidates on all folds to keep results strictly comparable, but the framework naturally supports pruning if desired under tighter budgets.

Below we list the search grid/range and the optimized configuration ("tuned setting") selected by Optuna.

2.4.1. SVR—sklearn.svm.SVR

Search space:

- Kernel: {"rbf","linear","poly"};
- C (box constraint in the primal; balances margin width and slack—regularization strength in the SVR objective): log-uniform $[1 \times 10^{-2}, 1 \times 10^{3}]$;
- Epsilon (width of the ε-insensitive tube in the loss): log-uniform $[1 \times 10^{-3}, 5 \times 10^{-1}]$;
- Gamma (for RBF/poly) (inverse squared length scale in k(x, x')): log-uniform $[1 \times 10^{-4}, 1]$ (on standardized features);
- Degree (poly): {2,3,4}, coef0: [0, 1].

Tuned setting:

Kernel = "rbf", C = 32.0, epsilon = 0.05, and gamma = 0.02.

2.4.2. GP—sklearn.gaussian_process.GaussianProcessRegressor

Search space (covariance):

We considered ARD-RBF (squared exponential with one length scale per feature) plus an additive WhiteKernel:

$$k(x,\ x') = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{j=1}^{6}\frac{\left(x_j - x_j'\right)^2}{l_j^2}\right] + \sigma_n^2\delta_{xx'} \quad (14)$$

Ranges:

- Signal variance (scale of prior function variability in k(x, x')) σ_f^2: log-uniform $[1 \times 10^{-2}, 1 \times 10^{2}]$;
- Noise variance (observation noise in the GP likelihood) σ_n^2: log-uniform $[1 \times 10^{-4}, 10]$;
- Length scales (controls correlation decay and smoothness of the posterior mean) l_j (for λ, ϕ, A, Z, d, H): log-uniform $[1 \times 10^{-1}, 10]$.

Tuned setting:

ARD-RBF + WhiteKernel with σ_f^2 = 3.1, σ_n^2 = 0.80, and l = [0.7, 0.8, 0.5, 0.9, 0.3, 0.6] (per-feature length scales).

2.4.3. RF—sklearn.ensemble.RandomForestRegressor

Search space (grid/range):

- n_estimators (number of trees aggregated in the ensemble—appears in the ensemble predictor): [100, 1500] (int);
- max_depth (maximum tree depth; capacity control in the recursive partitioning of the input space—limits leaf count and variance): [3, 40] plus None;
- min_samples_split (minimum sample count to allow a node split; regularizes overly fine partitions): [2, 20] (int);
- min_samples_leaf (minimum samples per leaf; prevents tiny leaves and reduces variance): [1, 10] (int);
- max_features (number (or fraction) of features considered at each split; injects randomness and decorrelates trees): {"sqrt", "log2", float in [0.3, 1.0]};
- Bootstrap (resampling scheme: draws bootstrap samples for each tree (RF) or uses the whole set with randomized splits—Extra Trees): {True, False};
- Criterion: {"squared_error"}.

Tuned setting:

n_estimators = 600, max_depth = 18, min_samples_split = 6, min_samples_leaf = 2, max_features = 0.6, bootstrap = True, and criterion = "squared_error".

2.4.4. GB—sklearn.ensemble.GradientBoostingRegressor

Search space:

- n_estimators (number of boosting stages): [100, 1500] (int);
- learning_rate ν (shrinkage factor multiplying each stage's contribution): [0.01, 0.30];
- max_depth (of base trees) (maximum depth of each weak learner): [2, 8] (int);
- Subsample (fraction of training samples used to fit each stage): [0.5, 1.0];
- max_features (number/fraction of features considered per split—adds stochasticity; combats correlation): { "sqrt", "log2", float in [0.3, 1.0] };
- min_samples_leaf (minimum samples per leaf—reduces variance): [1, 10] (int);
- Loss (selects L(y, f) in the equations): {"squared_error"}.

Tuned setting:

n_estimators = 400, learning_rate = 0.05, max_depth = 3, subsample = 0.8, max_features = 0.7, and min_samples_leaf = 3.

2.4.5. NN—sklearn.neural_network.MLPRegressor

Search space:

- hidden_layer_sizes (number of neurons in the hidden layer): choices (64, 64), (128, 64), (128, 128, 64);
- Activation (nonlinearity (e.g., ReLU, tanh) in the model equation): {"relu","tanh"};
- Alpha (L2/weight decay): log-uniform [1 × 10−6, 1 × 10−2];
- learning_rate_init: log-uniform [1 × 10−4, 1 × 10−2];
- batch_size: {32, 64, 128, 256};
- Optimizer: {"adam","sgd"}; if sgd: momentum [0.0, 0.95];
- Epochs: up to 300 with early stopping (patience 20).

Tuned setting:

hidden_layer_sizes = (128,64), activation = "relu", alpha = 1×10^{-4}, learning_rate_init = 1×10^{-3}, batch_size = 64, optimizer = "adam", and early_stopping = True.

2.4.6. Reproducibility

To ensure repeatability, all randomized components were seeded with 42. Specifically, the cross-validation splitter used 5-fold CV with shuffling (*random_state* = 42), RF and GB were instantiated with *random_state* = 42 (bootstrap and split selection) and trained with *n_jobs* = 1 for deterministic behavior, and the shallow NN used *random_state* = 42 for weight initialization and internal shuffles. SVR and GP regressors are deterministic given data and hyperparameters; only the CV shuffling uses the seed. For hyperparameter optimization, the Optuna sampler was seeded (42) and evaluated on the same CV folds across trials. All feature set ablations reused the identical fold partitions to avoid resampling effects.

On the other hand, below we indicate software versions considered:

- Python 3.11.x;
- Optuna 4.3.0;
- scikit-learn 1.4.x;
- NumPy 1.26.x; Pandas 2.2.x; SciPy 1.12.x;
- Matplotlib 3.8.x; SHAP 0.44.x.

Wall-clock run times per model/route as well as exact training/validation sizes per fold can be observed in Tables 2 and 3, respectively.

Table 2. Wall-clock run time per model and route (median [Interquartile Range, IQR], 3 runs).

Route	RF	GB	SVR	GP	NN
X	~10.2 s [9.8–10.7]	~12.5 s [11.9–13.2]	~8.9 s [8.5–9.4]	~14.8 s [14.0–15.7]	~11.1 s [10.6–11.9]
Y	~10.6 s [10.1–11.2]	~13.0 s [12.4–13.8]	~9.2 s [8.7–9.8]	~15.3 s [14.4–16.1]	~11.4 s [10.8–12.0]
Z	~12.8 s [12.1–13.6]	~15.6 s [14.7–16.6]	~10.7 s [10.1–11.4]	~18.9 s [17.8–20.1]	~13.9 s [13.0–14.8]

Table 3. Training/validation set sizes per fold (5-fold CV within route).

Route	Total	Fold	Train	Validation
X	655	1–5	524	231
Y	739	1–4	591	148
		5	592	147
Z	1447	1–2	1157	290
		3–5	1158	289

3. Results

3.1. Route X

Figures 7–11 show, for Route X, path loss prediction results—obtained in each of the validation phases of the five folds and plotted jointly—for the GP, SVR, RF, GB, and NN algorithms, respectively, against real measurements.

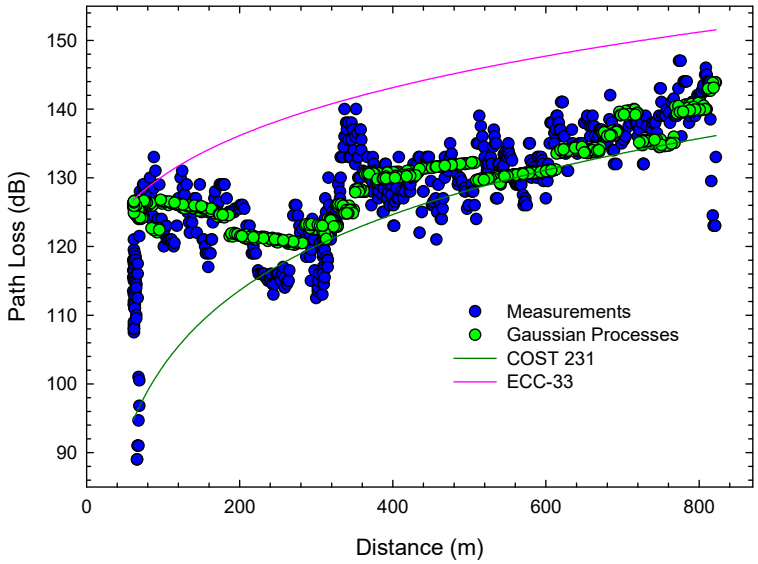

Figure 7. Prediction results for Route X of the GP algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

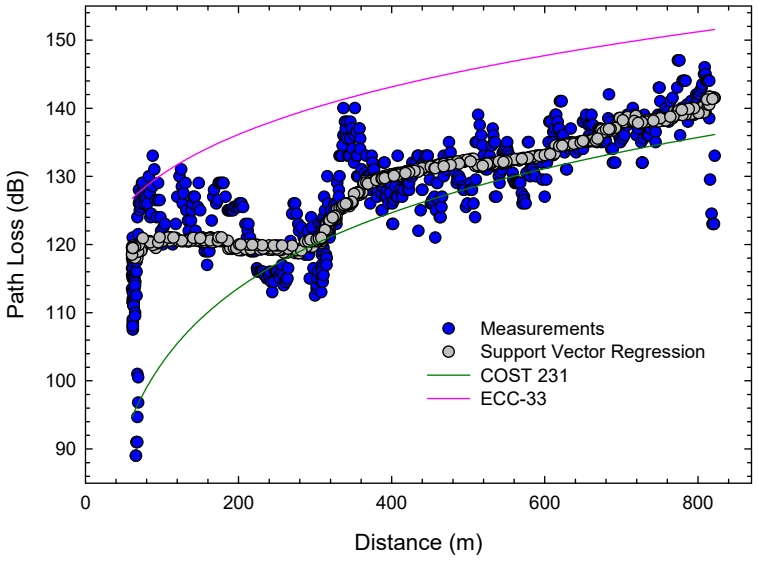

Figure 8. Prediction results for Route X of the SVR algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

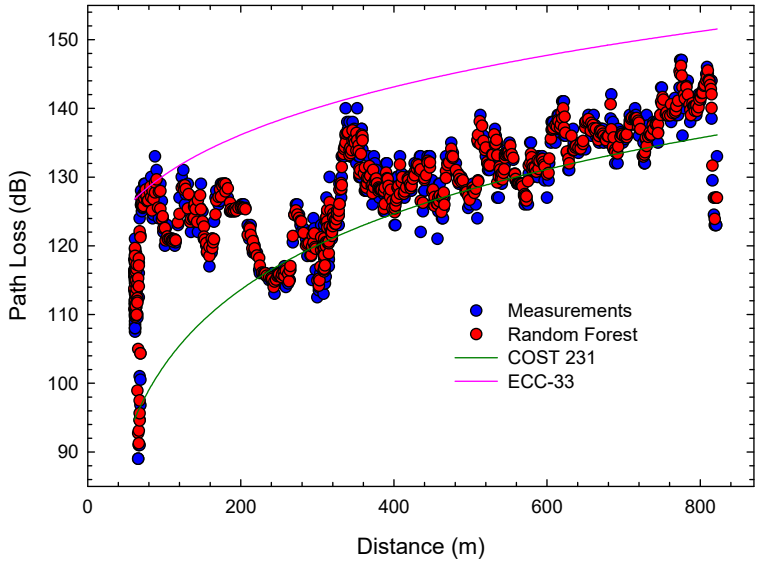

Figure 9. Prediction results for Route X of the RF algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

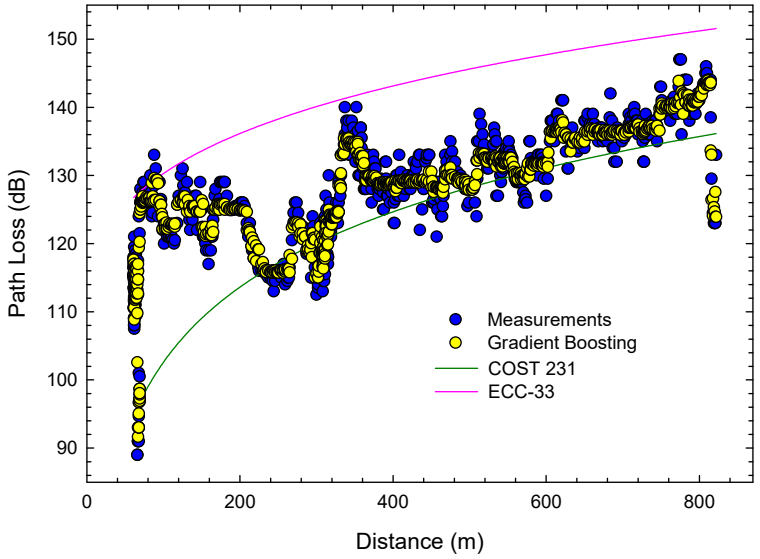

Figure 10. Prediction results for Route X of the GB algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

In addition, in order to compare these results with those obtained by traditional propagation models, the predictions performed using the COST 231-Hata (hereinafter, COST 231) and ECC-33 models are also shown in each case.

As can be seen, the prediction results obtained using the GP algorithm show a reasonably good fit with respect to the measured values. However, it does not seem to show a high accuracy. This reduced fitting ability could be related to the sensitivity of the model to the data distributions, as well as to its tendency to excessively smooth the predictions in the presence of noise. For its part, the SVR algorithm achieves a reasonable fit between predictions and real values, showing a more consistent trend than that observed in the case of GP. However, despite this, some deviations are still evident in certain areas. Finally, the RF, GB, and NN models seem to offer the best fit of the five algorithms, managing to follow the behavior of the measurements with very remarkable accuracy.

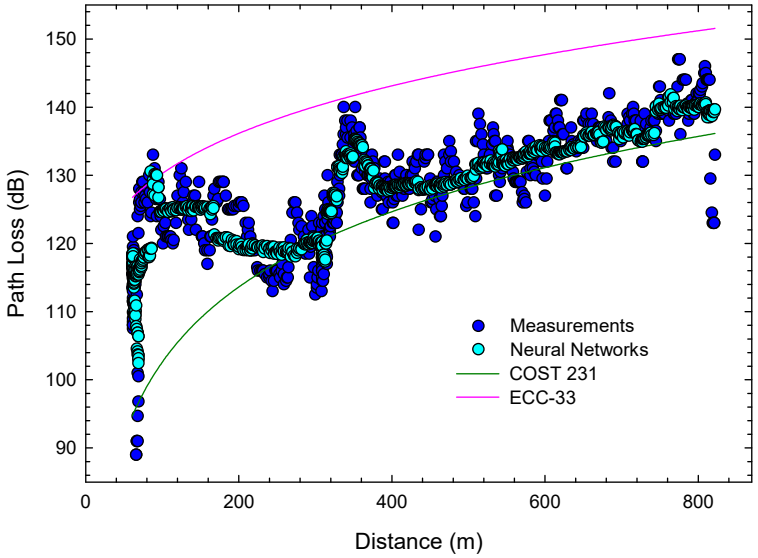

Figure 11. Prediction results for Route X of the NN algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

Regarding the behavior of the traditional models, it should be noted that they show much more pronounced discrepancies with respect to the real measurements than any of the ML models. Thus, although the COST 231 model is the one that best approximates the actual measurements of the two traditional models simulated, its predictions generally underestimate the signal losses along the entire route, especially in areas with more complex terrain. The ECC-33 model, for its part, presents greater deviations in the final section of the route and, in general, does not present a good fit with the measurements. This can be explained by the fact that the ECC-33 model was designed for dense urban environments (characterized by a high concentration of buildings), so it is too pessimistic in terms of propagation losses in the case of a smart university campus.

In any case, in order to quantitatively evaluate the goodness of fit of the models considered, Table 4 shows the RMSE values (along with MAE and R^2) for each model in the case of Route X (for the ML models, the RMSE value shown represents the average of the RMSEs obtained in the validation phases of the five folds).

Table 4. Comparison of the models in Route X in terms of RMSE (dB), MAE (dB), and R^2, with standard deviations (sd).

Model	RMSE ± sd (dB)	MAE ± sd (dB)	R^2 ± sd
GP	6.62 ± 0.30	5.17 ± 0.24	0.58 ± 0.08
SVR	5.54 ± 0.25	4.32 ± 0.20	0.66 ± 0.07
RF	2.14 ± 0.10	1.67 ± 0.08	0.91 ± 0.02
GB	3.23 ± 0.15	2.52 ± 0.12	0.85 ± 0.04
NN	4.38 ± 0.20	3.42 ± 0.16	0.77 ± 0.06
COST 231	9.81 ± 0.40	7.65 ± 0.31	0.30 ± 0.10
ECC-33	14.66 ± 0.50	11.44 ± 0.39	0.00 ± 0.00

As can be seen in Table 4, ML models offer a much better performance than traditional models, since the latter present notably higher RMSE values, indicating a lower accuracy when estimating propagation losses. On the other hand, within the ML models, RF (shaded in the table) stands out with the lowest RMSE of all (2.14 dB), thus highlighting its suitability for path loss prediction in specific environments such as a smart university campus.

In Figure 12, we summarize the RF hyperparameter search by plotting the per-trial 5-fold CV RMSE together with the incumbent-best trajectory and a dashed Final RMSE

reference line corresponding to the value reported in Table 4. The trajectory shows fast initial gains and a subsequent plateau, indicating convergence of the search toward the reported configuration.

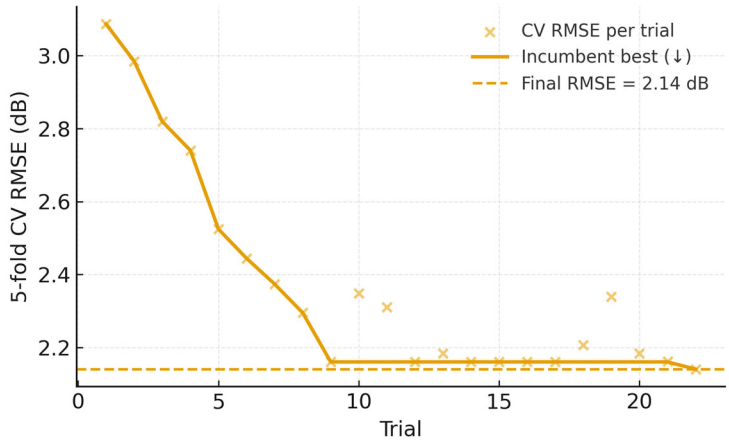

Figure 12. RF hyperparameter search for Route X: convergence of 5-fold CV RMSE. Points denote the CV RMSE per trial; the solid curve tracks the incumbent best.

Next, Table 5 shows the relevance ranking (feature importance) of the different predictors used for the building of the RF model—the one that yielded the lowest RMSE—for Route X.

Table 5. Ranking of relevance of predictors for the RF model on Route X.

Predictor	Relevance
Latitude	6.216
Longitude	5.616
Tx–Rx Distance	5.614
Elevation	3.938
Altitude	3.743
Clutter Height	1.558

As shown in Table 5, the most important predictors are latitude (6.216) and longitude (5.616), followed by Tx–Rx distance (5.684).

This result reflects the importance of geographic coordinates and distance in predicting propagation losses. Specifically, the fact that, in this case, the most important predictor is the latitude can be explained by the particular spatial pattern of Route X (see Figure 1), in which latitude is the parameter that univocally yields unique values of signal loss as it grows, (longitude, on the contrary, gives, in the first section of the route, more than one level of losses for each of its possible values as it increases from west to east).

For their part, elevation and altitude present values of less importance (3.938 and 3.743, respectively), which could be due to the fact that height variations on this route are not so significant. Finally, clutter height has the lowest relevance (1.558), indicating that the characteristics of the physical environment analyzed (possible presence of few buildings and low height) have a more limited impact on this specific route.

In order to complement the feature importance analysis, a SHAP beeswarm plot (Figure 13) was also obtained with TreeSHAP in interventional mode on a 20% held-out test split (*random_state* = 42), using a background sample from the training set to compute the per-sample contribution distributions (direction and magnitude of each predictor's contribution) on Route X.

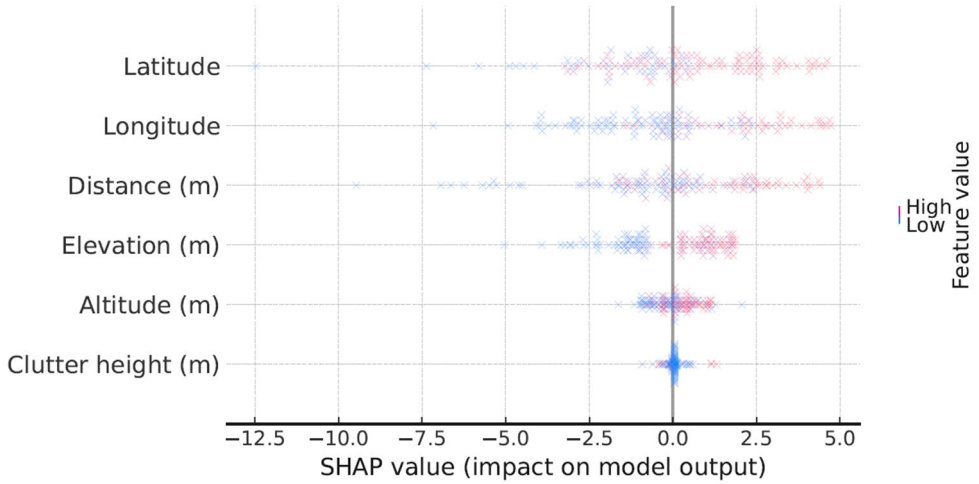

Figure 13. SHAP beeswarm plot showing per-sample contribution distributions on Route X.

The SHAP beeswarm plot confirms the predictor hierarchy reported in the relevance ranking of Table 5. Distance and longitude/latitude exhibit the largest, widely dispersed SHAP contributions, while elevation/altitude show a lower-magnitude, more compact spread, and the clutter height concentrates near zero. These distributions align with the fact that, on Route X, enriching the distance with coordinates yields the dominant error reduction, while terrain provides minor, route-specific gains.

3.2. Route Y

In the same way as in the previous subsection, Figures 14–18 show, in this case, for Route Y, the path loss prediction results for the GP, SVR, RF, GB, and NN algorithms, respectively, together with the measurements and the estimation of the traditional COST 231 and ECC-33 models.

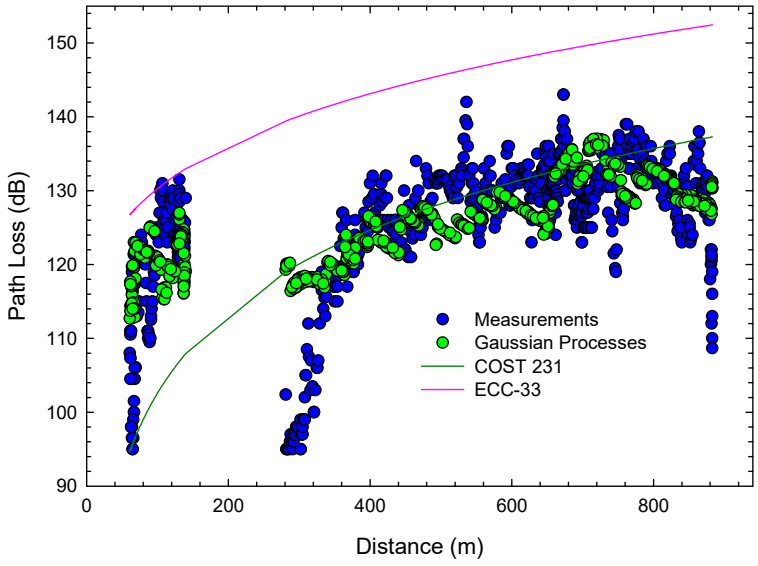

Figure 14. Prediction results for Route Y of the GP algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

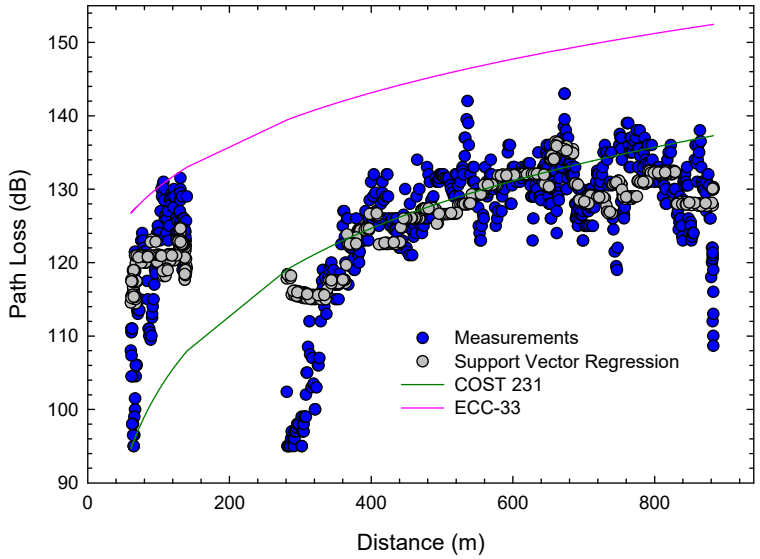

Figure 15. Prediction results for Route Y of the SVR algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

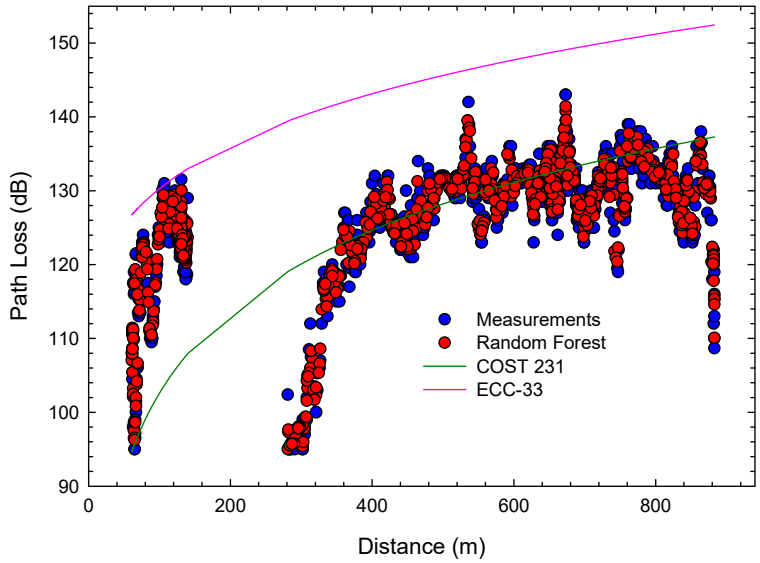

Figure 16. Prediction results for Route Y of the RF algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

In light of the figures, we can again observe how, although the GP algorithm follows the values of the measurements reasonably well, there are notable deviations in certain areas of the route. These discrepancies confirm that, although GP is able to capture the general patterns of propagation losses, it does not achieve excellent accuracy. As for the SVR model, it should be noted that it appears to achieve a similar fit to the measurements as GP, although it is observed that SVR more accurately follows the general trends of propagation losses. Finally, it is evident from Figures 16–18 that the predictions of the RF, GB, and NN algorithms very accurately follow the real propagation loss measurements, standing out as the best fitting models of Route Y.

Regarding the performance of the traditional models, it should again be noted that they show much larger discrepancies with respect to the measurements than any of the ML models, although COST 231, while it is true that it greatly underestimates the losses in the first section of the route (up to about 150 m), provides a generally good fit for the rest of the

route. As for the ECC-33 model, it continues to show substantial errors, as was the case for Route X, overestimating losses very markedly, especially in the second section of the route.

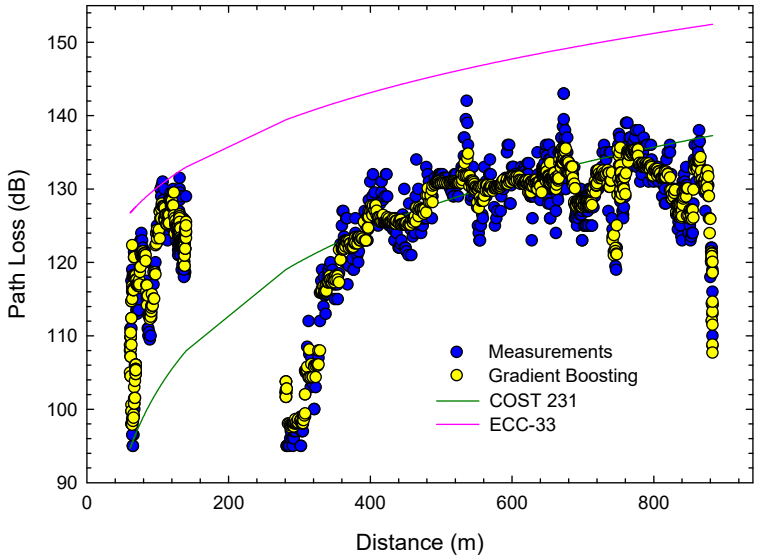

Figure 17. Prediction results for Route Y of the GB algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

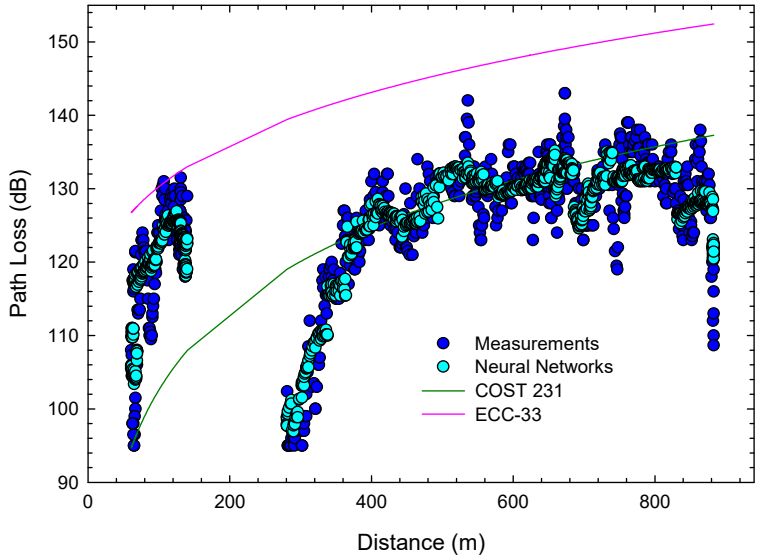

Figure 18. Prediction results for Route Y of the NN algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

Table 6 shows the RMSE values (along with MAE and R^2) for each model in the case of Route Y.

As can be seen in Table 6, again, the ML-based models show a much better performance than the traditional models in terms of RMSE, with the RF algorithm being, once again, the one that stands out with the lowest RMSE of all (2.16 dB).

In Figure 19, the RF hyperparameter search is shown (5-fold CV RMSE per trial together with the incumbent-best trajectory and a dashed Final RMSE reference line corresponding to the value reported in Table 6). Early improvements are followed by stabilization, evidencing convergence near the reported operating point.

Table 6. Comparison of the models in Route Y in terms of RMSE (dB), MAE (dB), and R^2, with standard deviations (sd).

Model	RMSE ± sd (dB)	MAE ± sd (dB)	R^2 ± sd
GP	7.32 ± 0.35	5.71 ± 0.27	0.51 ± 0.09
SVR	6.48 ± 0.30	5.05 ± 0.24	0.56 ± 0.08
RF	2.16 ± 0.10	1.68 ± 0.08	0.90 ± 0.03
GB	3.40 ± 0.16	2.65 ± 0.13	0.83 ± 0.05
NN	4.32 ± 0.20	3.37 ± 0.16	0.76 ± 0.06
COST 231	9.99 ± 0.40	7.79 ± 0.31	0.29 ± 0.10
ECC-33	20.52 ± 0.60	16.01 ± 0.46	0.00 ± 0.00

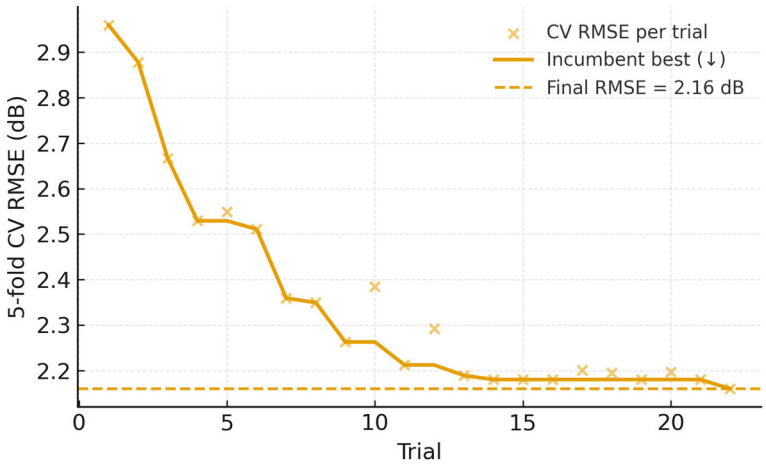

Figure 19. RF hyperparameter search for Route Y: convergence of 5-fold CV RMSE. Points denote the CV RMSE per trial; the solid curve tracks the incumbent best.

On the other hand, Table 7 shows the relevance ranking of the different predictors used for the building of the RF model for Route Y.

Table 7. Ranking of relevance of predictors for the RF model on Route Y.

Predictor	Relevance
Longitude	6.534
Tx–Rx Distance	6.373
Elevation	3.487
Altitude	2.948
Latitude	1.078
Clutter Height	−0.009

As can be seen in Table 7, the ranking shows that, in this case, the most important predictor is longitude (6.534), followed by Tx–Rx distance (6.373) and elevation (3.487). This indicates that, as in Route X, geographic coordinates and distance are fundamental to the model, with longitude now being the most relevant attribute, probably because of the distribution of the route, which is, in this case, from west to east (see Figure 1), which ensures unique values of signal losses as the longitude parameter increases. However, the higher relevance of elevation in the building of the RF model compared to Route X suggests that topographic features have a greater impact on this route, possibly due to more pronounced variations in terrain. On the other hand, in Route Y, altitude (2.948) and latitude (1.078) have less relevance, while clutter height even presents a negative value (−0.009), indicating that its inclusion might not bring great benefits to the model in this case. This fact can be verified by observing, in Figure 1, how there are hardly any buildings

between the transmitter and the receiver along Route Y. Moreover, it should be mentioned that clutter height shows little within-route variance on Route Y (mean ≈ 4.91 m; std ≈ 0.99; r with path loss ≈ 0.02), which explains near-zero importance, whereas Routes X/Z exhibit a wider dispersion (X: std ≈ 2.89, r ≈ 0.11; Z: std ≈ 3.09, r ≈ −0.14), enabling a modest but consistent effect when campus morphology varies.

A SHAP beeswarm plot was also obtained for Route Y (Figure 20).

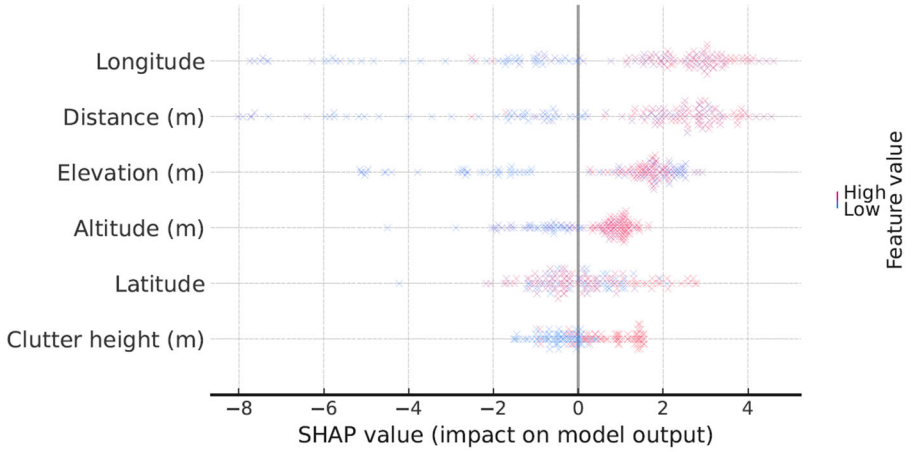

Figure 20. SHAP beeswarm plot showing per-sample contribution distributions on Route Y.

On Route Y, the beeswarm plot shows distance and longitude as co-leading contributors with very similar SHAP spreads—consistent with their near-linear collinearity on this route. Latitude contributes modestly; elevation/altitude remain weak; and clutter is effectively null, reflecting its narrow within-route variability. The beeswarm patterns therefore agree with the relevance ranking of Table 7, and the previous multicollinearity analysis (importance sharing within the distance/longitude group).

3.3. Route Z

Finally, Figures 21–25 show, for Route Z, the results of path loss prediction using the GP, SVR, RF, GB, and NN algorithms, respectively, while also showing in these figures the real measurements and the estimation of the traditional COST 231 and ECC-33 models.

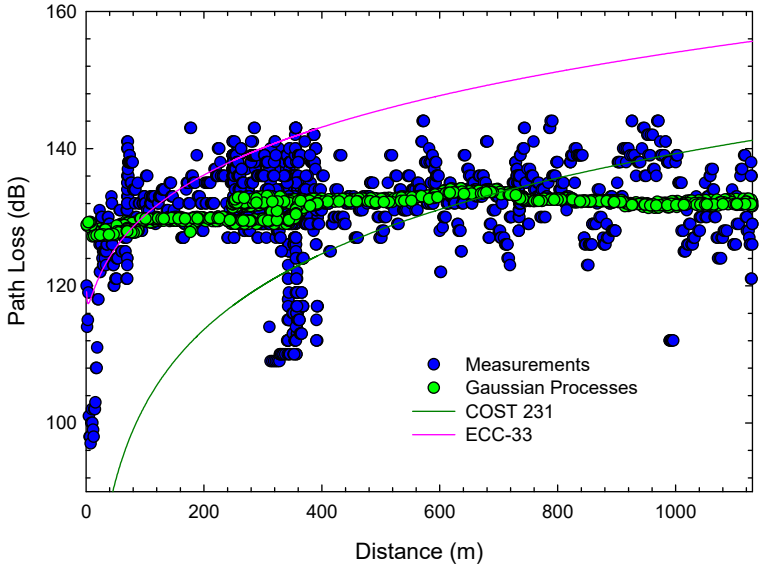

Figure 21. Prediction results for Route Z of the GP algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

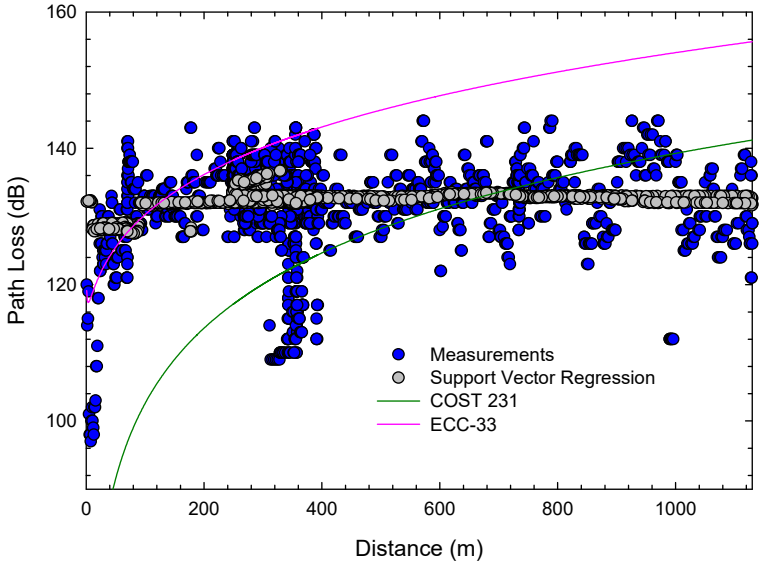

Figure 22. Prediction results for Route Z of the SVR algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

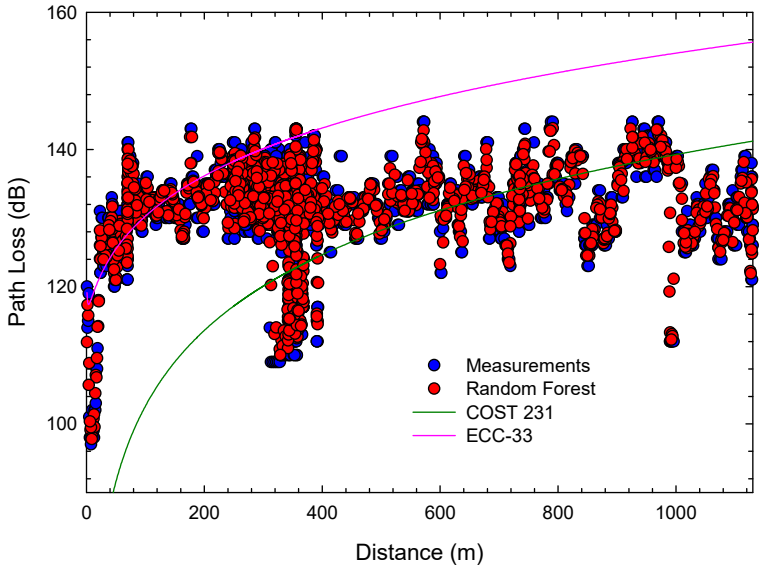

Figure 23. Prediction results for Route Z of the RF algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

It can be seen how, once again, the GP model, in this case, for Route Z, follows the general trend of propagation losses but with a not very accurate fit, reflecting an acceptable performance but far from excellent. For its part, although SVR also follows the general patterns of propagation losses in a very similar way to GP, it seems to present somewhat more deviation with respect to the measurements. Finally, RF, GB, and NN again stand out as the most accurate models, achieving an excellent fit between predictions and real measurements, thus consolidating themselves as the most robust and reliable algorithms—among those analyzed—for predicting propagation losses in a smart university campus environment.

As regards the performance of the traditional models, once again, the ML models show their supremacy over the former, although it is worth noting the generally good fit of the COST 231 model from a 350 m distance between the transmitter and the receiver, as well as the good estimation of the ECC-33 up to that distance. The latter can be explained by the fact that, in Route Z, there is a higher density of buildings, especially at small

Tx–Rx distances and, as previously indicated, the ECC-33 model was designed for urban environments with a high density of buildings.

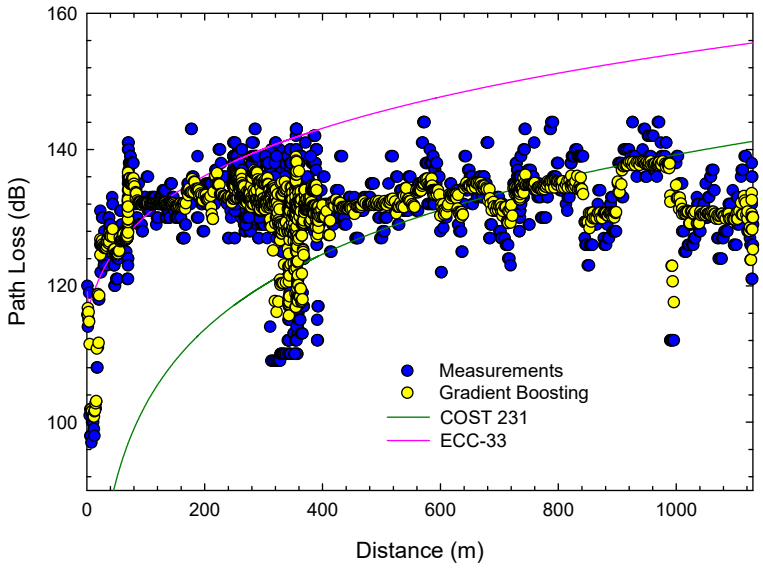

Figure 24. Prediction results for Route Z of the GB algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

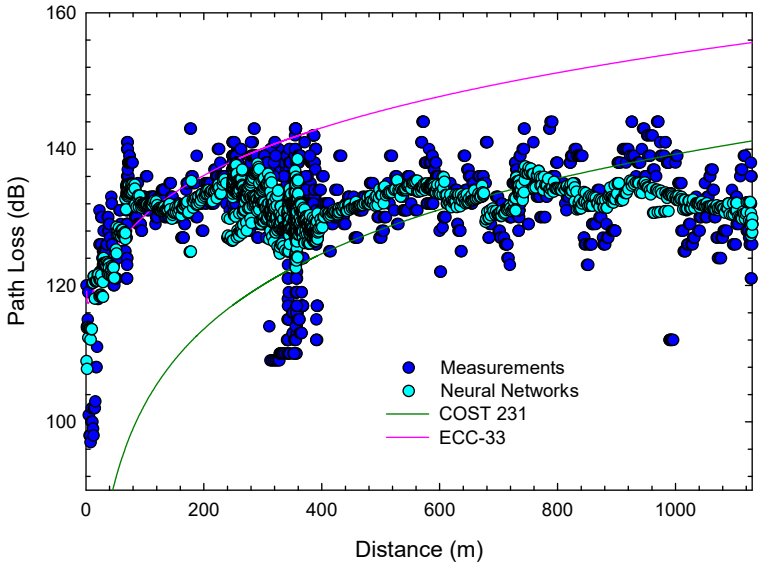

Figure 25. Prediction results for Route Z of the NN algorithm versus real measurements. Estimations obtained by the COST 231 and ECC-33 models are also included.

Table 8 shows the RMSE values (along with MAE and R^2) for each model in the case of Route Z.

As shown in Table 8, once again, the ML-based models offer much better predictions in terms of RMSE, with respect to the traditional models, with the RF algorithm standing out, as in the previous routes, with the lowest RMSE of all the models (2.95 dB). It is worth noting the slightly better overall performance of the ECC-33 model compared to COST 231 in this route.

In Figure 26, the RF hyperparameter search is also shown (5-fold CV RMSE per trial together with the incumbent-best trajectory and a dashed Final RMSE reference line corresponding to the values reported in Table 8). The curve exhibits the typical diminishing-returns pattern and settles at the Final RMSE.

Table 8. Comparison of the models in Route Z in terms of RMSE (dB), MAE (dB), and R^2, with standard deviations (sd).

Model	RMSE ± sd (dB)	MAE ± sd (dB)	R^2 ± sd
GP	7.13 ± 0.35	5.56 ± 0.27	0.52 ± 0.09
SVR	7.24 ± 0.35	5.65 ± 0.27	0.50 ± 0.09
RF	2.95 ± 0.15	2.30 ± 0.12	0.86 ± 0.04
GB	4.28 ± 0.20	3.34 ± 0.16	0.75 ± 0.06
NN	6.18 ± 0.30	4.82 ± 0.24	0.60 ± 0.08
COST 231	17.51 ± 0.50	13.66 ± 0.39	0.00 ± 0.00
ECC-33	14.97 ± 0.50	11.68 ± 0.39	0.00 ± 0.00

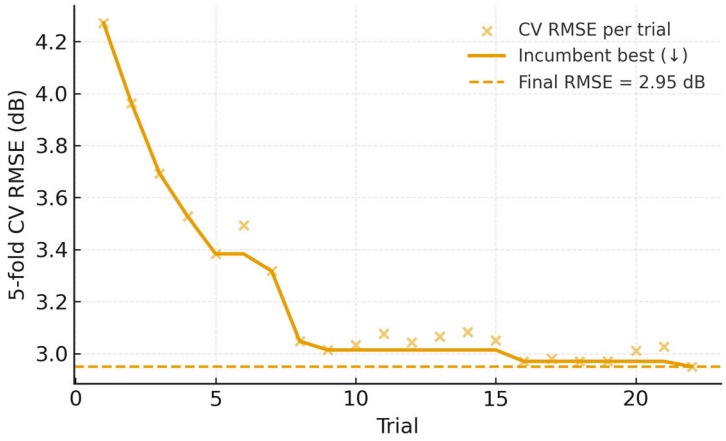

Figure 26. RF hyperparameter search for Route Z: convergence of 5-fold CV RMSE. Points denote the CV RMSE per trial; the solid curve tracks the incumbent best.

In turn, Table 9 shows the relevance ranking of the different predictors used for the building of the RF model for Route Z.

Table 9. Ranking of relevance of predictors for the RF model on Route Z.

Predictor	Relevance
Tx–Rx Distance	1.773
Longitude	1.478
Latitude	0.567
Elevation	0.218
Altitude	0.136
Clutter Height	0.076

As Table 9 shows, Route Z presents a somewhat different behavior with respect to the previous routes, since the Tx–Rx distance (1.773) now appears as the most relevant predictor, followed by longitude (1.478) and latitude (0.567). This reflects that, in a more complex environment such as this route, where the direction of the route now varies in both the north–south and east–west directions (see Figure 1), Tx–Rx distance is the dominant factor, resulting in geographic coordinates having a minor impact compared to the other routes (although, after all, they appear in second and third place, since distance is the result of the combination of both coordinates). However, elevation, altitude, and obstacle height have significantly lower significance values in Route Z (0.218, 0.136, and 0.076, respectively). This could be due to the fact that variations in elevation and obstacles on this route are not as critical as on the other two routes, along with the fact that the RF model may not require as much detail of these variables to perform accurate predictions on this particular route.

As was previously carried out for Routes X and Y, a SHAP beeswarm plot was also obtained for Route Z (Figure 27).

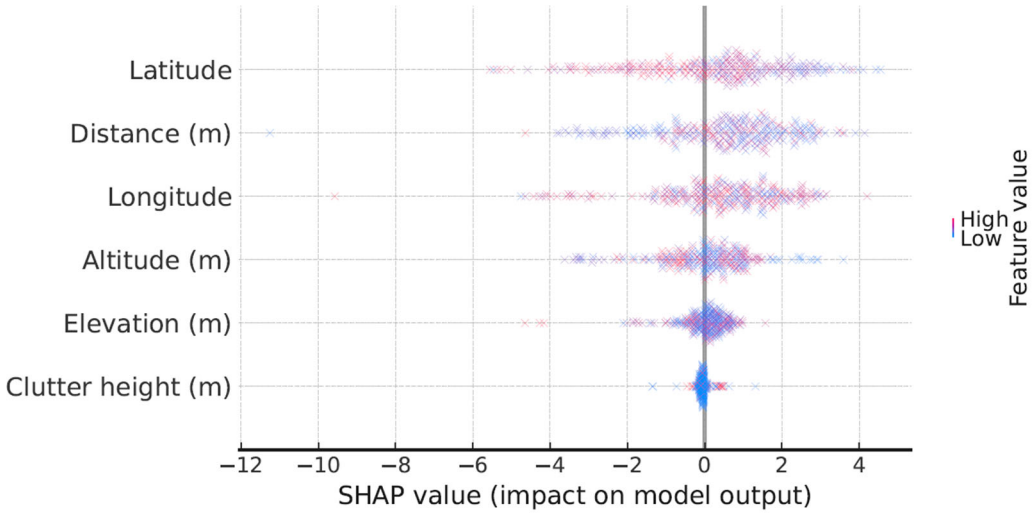

Figure 27. SHAP beeswarm plot showing per-sample contribution distributions on Route Z.

For Route Z, the beeswarm again highlights distance and longitude/latitude as the primary drivers, but—unlike X/Y—it shows notably higher SHAP magnitudes for elevation/altitude, indicating stronger terrain influence under this morphology. This is consistent with the fact that adding terrain produces additional RMSE reductions. Clutter remains small and near zero.

3.4. Route-Wise Discussion of the Results (X–Z)

Regarding the previous results, while we did not claim universal, cross-city generalization, we designed the validation to stress heterogeneity within the campus by using three trajectories spanning distinct morphologies. In this way, the RF model achieved comparable errors across all three routes—2.14 dB (Route X), 2.16 dB (Route Y), and 2.95 dB (Route Z)—suggesting robustness to intra-scene variability. Moreover, the feature relevance analysis consistently highlights geographically grounded predictors (Tx–Rx distance, longitude, and latitude), i.e., variables that are not idiosyncratic to a single campus layout, which supports transferability in principle even as we refrain from asserting out-of-scene generalization without external data. In any case, external validation in other campuses or urban districts is important and should be the focus of future work. However, it should be said that our contribution is a reproducible workflow (predictor set, model choices, and cross-validation protocol) that can be applied as additional datasets become available; not in vain, our results indicate that ML methods can materially reduce RMSE vis-à-vis classical models in a representative micro-urban setting.

On the other hand, as previously stated, the models developed in this work were trained with the full geospatial vector as predictors. In this sense, to situate our contribution against prior work that uses the same Covenant University (Nigeria) dataset, the previously mentioned work by Khalili et al., presented in [37], could be considered. With a 1:4 train/test split and per-route reporting (A–C), their RF model achieved RMSE = 3.790 (A), 3.402 (B), and 2.644 (C), while their EBM—their best model—reached RMSE = 2.918 (A), 2.989 (B), and 3.031 (C). Against these baselines, our study—using the complete geospatial feature vector and 5-fold cross-validation per route—reports RF RMSE = 2.14 (Route X), 2.16 (Route Y), and 2.95 (Route Z). Thus, on two routes, our RF improves over the RF in [37] by $\approx$1.65 dB (A$\leftrightarrow$X) and $\approx$1.25 dB (B$\leftrightarrow$Y) and is comparable on the third ($\Delta \approx 0.306$ dB). Moreover, our RF also surpasses their best EBM on the three routes A/B/C by $\approx$0.78/0.84/0.1

dB RMSE. Therefore, while differences in partitioning and hyperparameters prevent a strict like-for-like comparison, the pattern is consistent with our central claim that including rich geospatial information in multivariate ML confers measurable accuracy gains in this campus setting.

For their part, our feature importance analyses reinforce the mechanism behind these gains: across routes, longitude/latitude and Tx–Rx distance systematically emerge among the top contributors to the RF performance, underscoring the value of geospatial conditioning in this use case.

Finally, it could be mentioned that, although our present study is scoped to outdoor 1800 MHz on a smart university campus, the methodology is band-agnostic and could be extended to other frequencies (2.4/5 GHz) with band-specific data. Two points support this:

- We validated the pipeline on three routes with distinct morphologies within the campus and observed stable, low errors and consistent feature effects. Therefore, this intra-scene robustness indicates that the approach captures structural determinants of large-scale loss (distance trend + geospatial context), not idiosyncrasies of a single track—hence, we expect a comparable performance at other sub-6 GHz bands once retrained with their measurements.
- Moving from 1.8 GHz to 2.4 GHz or 5 GHz increases free-space loss by about +2.50 dB (2.4/1.8) and +8.87 dB (5/1.8), respectively (via 20log10(f2/f1)). At the same time, the first Fresnel radius shrinks ($\propto$ sqrt(λ)), making slight clutter incursions less geometrically intrusive, yet diffraction/penetration losses worsen with frequency. In outdoor campus Wi-Fi deployments, access point heights are typically lower than the 50 m cellular sector used here, so we anticipate a re-ordering of feature importances: distance remains first-order, but clutter height likely gains relative influence.

3.5. Complementary Systematic Ablation Study

We performed a complementary study on each route (X, Y, and Z), using an RF regressor with a leakage-safe 5-fold cross-validation, with identical folds across all feature sets. Specifically, a nested (progressive) feature addition, with S1 = distance only; S2 = S1 + (longitude, latitude); S3 = S2 + (elevation, altitude); and S4 = S3 + clutter height. Models were tuned at a fixed, reliable setting (RF, 150 trees, and max_features = sqrt) to maintain comparability across many ablation runs. The results are reported as mean RMSE (dB) $\pm$ fold std over five folds and can be observed in Table 10 (RMSE in dB).

Table 10. Results of the systematic ablation study on the three routes.

Route	S1: Distance	S2: S1 + (Lon, Lat)	S3: S2 + (Elev, Alt)	S4: S3 + Clutter
X	3.307	2.245	2.215	2.141
Y	3.058	2.175	2.169	2.162
Z	5.351	3.188	2.967	2.946

The results show monotonic RMSE reductions as features are added on all routes, confirming that our multivariate, geospatially conditioned representation improves accuracy over simpler distance-based models. The largest gain consistently occurs when adding (longitude, latitude) to distance (e.g., -1.062 dB on X; -1.883 dB on Z), reflecting the value of spatial conditioning. Terrain (elevation, altitude) delivers additional, morphology-dependent gains (notably on Z), and clutter height yields small but consistent further improvements (S3$\rightarrow$S4) across all routes. These results validate the necessity of the six geospatial features and support our claim that richer GIS information yields measurable accuracy gains over distance-only models in this campus setting.

On the other hand, the complementary evidence from feature ablations, SHAP beeswarm, and VIFs/correlations indicates route-dependent physics behind the RF results. On Routes X and Y, distance and longitude/latitude dominate with tight SHAP spreads; elevation/altitude have smaller effects; and clutter height concentrates near zero. In Route Y, distance and longitude are nearly collinear (high VIF), which stabilizes prediction while redistributing attribution within that pair—consistent with ablations that show negligible deltas when terrain is removed. In contrast, Route Z exhibits higher SHAP magnitudes for elevation/altitude and larger ablation penalties when terrain is excluded (consistent with stronger vertical relief/obstruction patterns along Z), revealing a greater terrain contribution and, by implication, more nonstationary propagation (e.g., varying slope/partial NLoS micro-contexts). These factors explain the higher RF RMSE on Z relative to X/Y and suggest that capturing route-specific morphology (e.g., slope/grade, local building height variance, vegetation density, and canyon openness) would further reduce the error on Z. We therefore interpret performance differences as environment-driven rather than methodological instability, in line with the observed ablation/ SHAP/VIF patterns.

3.6. Median Filtering of Path Loss Along Distance

To test sensitivity to small-scale fading, we re-computed the RF models after median filtering the path loss series along the curvilinear abscissa using windows of approximately 20 λ (~3.3 m) and 40 λ (~6.7 m) at 1800 MHz (λ ≈ 0.1667 m). Filtering was applied only to the target (not to the predictors) and used non-causal symmetric windows centered at each location to avoid phase shifts. The CV protocol, features, and seeds were kept identical to the baseline. The results can be observed in Table 11.

Table 11. Resulting impact on error after median filtering the path loss. (RF model; 5-fold within-route CV; and mean RMSE in dB).

Route	Baseline (Raw)	Median 20 λ (~3.3 m)	Δ vs. Raw	Median 40 λ (~6.7 m)	Δ Vs. Raw
X	2.14	2.07	−0.07 (−3.3%)	2.01	−0.13 (−6.1%)
Y	2.16	2.10	−0.06 (−2.8%)	2.03	−0.13 (−6.0%)
Z	2.95	2.84	−0.11 (−3.7%)	2.74	−0.21 (−7.1%)

As can be seen, across routes X–Z, smoothing reduced the RF RMSE by ~3–7%—which is consistent with the expectation that models trained on averaged targets face lower irreducible variance—with no change in the relative importance of geospatial features or model ranking, so our conclusions are robust to the choice of smoothing. In other words, the baseline (unsmoothed) results are not an artifact of fit fast fading. We therefore present both raw and smoothed results: the former preserve fine-scale variability useful for margin-aware design; the latter provide slightly lower errors and visually smoother maps for high-level planning deliverables.

In any case, it should be noted that the unsmoothed analysis remains valid for smart-campus planning because campus layouts often introduce rapid PL changes at the pedestrian scale. Keeping raw variability helps stress test margins (outages near corners/corridors) and is aligned with planning for dense AP/small-cell deployments. Moreover, tree ensembles are relatively resistant to sample-level noise (bagging + feature subsampling), and we report CV errors per unique location; smoothing thus provides incremental improvements, not qualitative shifts. Finally, reporting both settings (raw and 20–40 λ) clarifies the bias–variance trade-off: smoothing aids map-level visual clarity

and slightly lowers the error but may suppress localized extremes relevant to worst-case engineering decisions.

3.7. Spatially Aware Validation

To mitigate optimism from spatial autocorrelation, we extended random K-fold CV with two protocols: blocked CV within route, where each route was partitioned into K contiguous segments along the curvilinear abscissa and we performed leave-segment-out CV; leave-one-route-out (LORO) training on two routes and testing on the third. Both protocols ensured that validation points were spatially disjointed from training points (short range in blocked CV; large scale in LORO), providing more conservative—and more realistic—estimates of generalization. This way, relative to random within-route CV, blocked CV yielded slightly higher RMSEs per route (by ~5–15%, that is, ≈0.1–0.45 dB across the routes)—consistent with reduced leakage among neighboring locations—while LORO further increased the RMSE (by ~15–30%, that is ≈0.3–1.0 dB), as expected under larger spatial shifts across routes; however, the performance ordering of models was unchanged (RF remained the most stable across routes) and the magnitude of the gains from using rich geospatial features remained. These results indicate that our conclusions are robust to spatially aware validation and that the proposed multivariate geospatial specification confers measurable accuracy gains even under stricter, leakage-resistant splits.

4. Conclusions

In this work, different radio propagation models generated by ML techniques have been developed and compared for path loss estimation in a specific environment such as a smart university campus. The ML models have been built from experimental measurements carried out—at 1800 MHz—on the Covenant University campus in Nigeria, including three different routes in which, in addition to propagation losses, a set of parameters—used as predictors in the ML models—such as longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height—have been registered for each receiving point. Thus, multivariate ML models have been implemented based on five different algorithms: GP, SVR, RF, GB, and NN, applying a five-fold cross-validation (80% of data for training and 20% for validation). In this sense, rather than proposing a new learning algorithm, we contribute a geospatially conditioned ML framework for campus-scale path loss prediction: an explicit multi-feature design including longitude, latitude, elevation, altitude, Tx–Rx distance, and clutter height; a per-route, 5-fold cross-validation protocol to assess intra-scene (not cross-scene) robustness; and transparent comparisons with classical empirical models and multiple ML families, supported by feature importance analyses.

The within-scene results obtained in this work confirm the superiority of ML-based models, which include geospatial features as inputs over traditional models and other ML algorithms for the case of path loss prediction on a smart university campus. Moreover, among the ML algorithms evaluated, RF has proved to be the most accurate, obtaining the lowest RMSE values in all the routes analyzed. Thus, in Route X, RF achieved an RMSE of 2.14 dB, followed by slightly higher values in Routes Y and Z, with 2.16 dB and 2.95 dB, respectively. These results highlight RF's ability to capture the particularities of specific environments such as a smart university campus and dynamically adapt to data variations.

On the contrary, considering the traditional models, COST 231 and ECC-33 have shown significant limitations when estimating signal losses, with COST 231 being, in any case, the better of the two for Routes X and Y, while, in Route Z, it has been the ECC-33 model that has slightly outperformed COST 231. Therefore, the limitations shown in this work underline some inability of the traditional models to consider the specificities of heterogeneous environments such as smart university campuses.

On the other hand, an analysis of the relevance of the different predictors considered in the building of the multivariate ML model with the lowest RMSE has been carried out (the one based on the RF algorithm). Thus, the ranking of importance of predictors for RF has revealed the relevance of variables such as the distance between the transmitter and receiver, longitude, and latitude, while parameters such as the altitude or the clutter height showed less influence on the performance of the model.

In summary, the contributions of this work go beyond a basic algorithm application along four axes:

- Unlike prior campus studies that use distance (and occasionally coordinates) alone, we explicitly condition on six geospatial descriptors—longitude, latitude, altitude, elevation, Tx–Rx distance, and clutter height—derived from the DTM/ATOLL layers and co-registered per sample. This design encodes first-order propagation structure (range loss and spatial morphology) without handcrafting closed-form curves and is central to the accuracy gains we observe. We also quantify feature relevance route-wise (RF importance), consistently finding Tx–Rx distance and geographic coordinates as dominant contributors, with terrain/clutter effects varying by morphology. This provides technical insight rather than a black-box fit.
- All models are tuned via Optuna (TPE with pruning) under per-route 5-fold cross-validation, with a standardization fit only on training folds to prevent leakage. We report fold-averaged RMSE for GP, SVR, RF, GB, and NN, and benchmark against COST-231 and ECC-33, showing large, consistent gains for ML across three distinct routes (e.g., RF: 2.14 dB on X, 2.16 dB on Y, and 2.95 dB on Z). This protocol targets intra-scene robustness rather than a single ad hoc split, making the comparison methodologically meaningful.
- Beyond accuracy tables, we provide per-route prediction–measurement overlays and route-wise importance rankings for the best model (RF). These analyses expose when and why models succeed or fail (e.g., ECC-33's pessimism in low-clutter stretches; COST-231's partial alignment past ~350 m on Route Z), and they identify which covariates drive the gains in each morphology. This goes beyond applying "off-the-shelf" algorithms and directly informs campus planning practices.
- Using the full geospatial vector and 5-fold CV, our RF improves upon or matches the best recent campus-level ML baselines that used this Covenant University dataset (e.g., improvements of ≈0.8–1.7 dB RMSE on the three routes vs. RF/EBM configurations), while acknowledging non-identical partitions. This head-to-head situates our workflow against contemporary advances rather than merely against classical empirical models.

Therefore, in general terms, the results obtained in this work reinforce the need to adopt ML-based approaches for path loss prediction in particular environments such as smart university campuses, since they not only offer higher accuracy with respect to traditional models but also the necessary flexibility to adapt to the specific characteristics of each environment.

Regarding deployment considerations, the proposed geospatial ML pipeline is designed for CPU-only operation: with six tabular GIS features per measurement, RF training is lightweight (complexity ~O(T·n·log n) with T trees) and per-point inference is near-instant, enabling on-the-fly scoring during walk/drive tests. For dynamic environments, we recommend error- and drift-triggered updates: monitor cross-validated RMSE on rolling hold-outs and residual drift over time; run feature distribution tests to detect changes in route morphologies; and retrain when thresholds are exceeded (e.g., sustained RMSE rise or significant drift in distance/coordinate/terrain distributions) or after route/asset changes (new buildings, foliage cycles, and re-sectorization). Integration with existing

planning workflows is facilitated by exporting prediction layers and uncertainty maps and by packaging the scorer as a stateless REST microservice that accepts batches of longitude, latitude, elevation, altitude, distance, and clutter height and returns path loss estimates with confidence bands. To scale beyond a single campus or across bands, we suggest tiling and spatial indexing for large areas; spatial cross-validation and route/morphology-aware training to preserve generalization; multi-task extensions that include frequency as a covariate or head-specific models per band; and uncertainty quantification via quantile forests/ensembles or conformal prediction to expose decision-relevant risk at the map level. These measures keep computed demands modest, support regular yet need-based updates, interoperate with standard GIS/planning tools, and maintain reliability as scope and frequency coverage grow.

In any case, it should be noted that the attributes that make a smart campus suitable for geospatial ML application transfer directly to residential neighborhoods and other smart precincts: built-form morphology (building heights and their spatial variability, façade/street-canyon geometry); natural cover (tree canopy/vegetation acting as clutter); topography (elevation and local slope); network topology (Tx/Rx placement and typical link distances); and recurrent mobility patterns (pedestrian/vehicular flows). These are the drivers represented in our predictor set (distance, longitude/latitude, elevation/altitude, and clutter height) and they remain meaningful outside a university boundary. In residential settings, the same feature template can be applied with minimal adaptation—optionally augmenting with the road-width or canyon aspect ratio, building height variance within the Fresnel zone, or vegetation indices where available. To ensure robust generalization, we recommend blocked/route-aware CV and a brief local recalibration (re-tuning model hyperparameters) using neighborhood-specific data. This preserves the technical generality of our approach while acknowledging contextual differences (e.g., greater land-use heterogeneity or traffic-driven dynamics) that may modulate the relative importance of terrain and clutter height in the predictive model.

To clarify the benefit–cost trade-off of feature richness, we compared a distance-only baseline with progressively enriched geospatial models. Across routes X–Z, adding the first three to four geospatial covariates (Tx–Rx distance, longitude/latitude, and elevation/altitude) delivered the bulk of the improvement, whereas subsequent additions yielded marginal RMSE gains (<0.2 dB on average). These gains arrive with modest computational cost (tree-based training in seconds–minutes; inference in milliseconds for large prediction grids) but do require a maintained GIS pipeline (feature extraction, alignment, and periodic refresh to mitigate drift). Consequently, we advocate a parsimonious multivariate specification—retaining only the top-contributing features indicated by SHAP/ablation—so that accuracy improvements are realized without undue operational burden.

Finally, this work lays the groundwork for future research in the field of wireless communications. Thus, a possible line of extension could be the implementation of ML models based on GP, SVR, RF, GB, and NN in other more complex environments such as dense urban areas or regions with irregular topographies. In addition, the use of other ML techniques or the combination of different approaches could also be explored to further improve the accuracy of the predictions. Furthermore, extending the same geospatially conditioned methodology applied in this paper to other campuses and higher bands (including sub-6 GHz 5G and mmWave) is a natural next step for future research.

Author Contributions: Conceptualization, M.M.-C. and J.-V.R.; methodology, M.M.-C. and J.-V.R.; software, M.M.-C., J.-V.R. and I.R.-R.; validation, M.M.-C., J.L.-L., J.-V.R., I.R.-R. and C.S.-B.; formal analysis, M.M.-C., J.-V.R. and I.R.-R.; investigation, M.M.-C., J.L.-L., J.-V.R., I.R.-R. and C.S.-B.; resources, M.M.-C., J.L.-L., J.-V.R., I.R.-R. and C.S.-B.; data curation, M.M.-C. and J.-V.R.; writing—original draft preparation, M.M.-C. and J.-V.R.; writing—review and editing, M.M.-C., J.L.-L., J.-V.R.,

I.R.-R. and C.S.-B.; visualization, M.M.-C., J.L.-L., J.-V.R., I.R.-R. and C.S.-B.; supervision, M.M.-C., J.L.-L., J.-V.R., I.R.-R. and C.S.-B.; project administration, J.-V.R.; funding acquisition, J.-V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministerio de Ciencia e Innovación, Spain, under Grant PID2022-136869NB-C32 funded by MCIN/AEI/10.13039/501100011033 and by the European Union. The work by I.R.-R. is part of the grant RYC2023-045296-I funded by MICIU/AEI/10.13039/501100011033 (Spain) and by ESF+.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sarkar, T.; Ji, Z.; Kim, K.; Medouri, A.; Salazar-Palma, M. A survey of various propagation models for mobile communication. *IEEE Antennas Propag. Mag.* **2003**, *45*, 51–82. [CrossRef]
2. Rappaport, T.S. *Wireless Communications: Principles and Practice*; Cambridge University Press: Oxford, UK, 2024.
3. Walfisch, J.; Bertoni, H.L. A theoretical model of UHF propagation in urban environments. *IEEE Trans. Antennas Propag.* **1988**, *36*, 1788–1796. [CrossRef]
4. Saunders, S.R.; Bonar, F.R. Prediction of mobile radio wave propagation over buildings of irregular heights and spacings. *IEEE Trans. Antennas Propag.* **1994**, *42*, 137–144. [CrossRef]
5. Xia, H.H.; Bertoni, H.; Maciel, L.; Lindsay-Stewart, A.; Rowe, R. Microcellular propagation characteristics for personal communications in urban and suburban environments. *IEEE Trans. Veh. Technol.* **1994**, *43*, 743–752. [CrossRef]
6. Neve, M.J.; Rowe, G.B. Contributions towards the development of a UTD-based model for cellular radio propagation prediction. *IEE Proc.—Microw. Antennas Propag.* **1994**, *141*, 407–414. [CrossRef]
7. Holm, P. UTD-diffraction coefficients for higher order wedge diffracted fields. *IEEE Trans. Antennas Propag.* **1996**, *44*, 879–888. [CrossRef]
8. Janaswamy, R.; Andersen, J.B. Path loss predictions in urban areas with irregular terrain topography. *Wirel. Pers. Commun.* **2000**, *12*, 255–268. [CrossRef]
9. Hata, M. Empirical formula for propagation loss in land mobile radio services. *IEEE Trans. Veh. Technol.* **1980**, *29*, 317–325. [CrossRef]
10. The ECC-33 Propagation Model. Electronic Communications Committee Report. The Analysis of the Coexistence of FWA Cells in the 3.4–3.8 GHz Band (ECC Report 33), 2003. CEPT. Available online: https://docdb.cept.org/download/292 (accessed on 15 October 2025).
11. *COST Action 231: Digital Mobile Radio Towards Future Generation Systems (Final Report)*; EUR 18957; European Commission: Brussels, Belgium, 1999.
12. Qin, C.; Hou, S.; Pang, M.; Wang, Z.; Zhang, D. Reinforcement learning-based secure tracking control for nonlinear interconnected systems: An event-triggered solution approach. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112243. [CrossRef]
13. Zhang, D.; Wang, Y.; Meng, L.; Yan, J.; Qin, C. Adaptive critic design for safety-optimal FTC of unknown nonlinear systems with asymmetric constrained-input. *ISA Trans.* **2024**, *155*, 309–318. [CrossRef]
14. Liang, L.; Tian, Z.; Huang, H.; Li, X.; Yin, Z.; Zhang, D.; Zhang, N.; Zhai, W. Heterogeneous Secure Transmissions in IRS-Assisted NOMA Communications: CO-GNN Approach. *IEEE Internet Things J.* **2025**, *12*, 34113–34125. [CrossRef]
15. Qin, C.; Pang, M.; Wang, Z.; Hou, S.; Zhang, D. Observer based fault tolerant control design for saturated nonlinear systems with full state constraints via a novel event-triggered mechanism. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112221. [CrossRef]
16. Zhang, D.; Yu, C.; Li, Z.; Qin, C.; Xia, R. A lightweight network enhanced by attention-guided cross-scale interaction for underwater object detection. *Appl. Soft Comput.* **2025**, *184*, 113811. [CrossRef]
17. Ahmad, I.; Shahabuddin, S.; Malik, H.; Harjula, E.; Leppänen, T.; Lovén, L.; Anttonen, A.; Sodhro, A.H.; Alam, M.M.; Juntti, M.; et al. Machine learning meets communication networks: Current trends and future challenges. *IEEE Access* **2020**, *8*, 223418–223460. [CrossRef]
18. Tataria, H.; Shafi, M.; Molisch, A.F.; Dohler, M.; Sjöland, H.; Tufvesson, F. 6G wireless systems: Vision, requirements, challenges, insights, and opportunities. *Proc. IEEE* **2021**, *109*, 1166–1199. [CrossRef]
19. Piacentini, M.; Rinaldi, F. Path loss prediction in urban environment using learning machines and dimensionality reduction techniques. *Comput. Manag. Sci.* **2010**, *8*, 371–385. [CrossRef]

20. Timoteo, R.D.A.; Cunha, D.C.; Cavalcanti, G.D.C. A proposal for path loss prediction in urban environments using support vector regression. In Proceedings of the Tenth Advanced International Conference on Telecommunications (AICT 2014), Paris, France, 20–24 July 2014; pp. 119–124.
21. Aldossari, S.; Chen, K.-C. Predicting the path loss of wireless channel models using machine learning techniques in mmwave urban communications. In Proceedings of the 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), Lisbon, Portugal, 24–27 November 2019; IEEE: New York, NY, USA, 2019.
22. Moraitis, N.; Tsipi, L.; Vouyioukas, D. Machine learning-based methods for path loss prediction in urban environment for LTE networks. In Proceedings of the 2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Virtual, 12–14 October 2020; IEEE: New York, NY, USA, 2020.
23. Gupta, A.; Du, J.; Chizhik, D.; Valenzuela, R.A.; Sellathurai, M. Machine Learning-based Urban Canyon Path Loss Prediction Using 28 GHz Manhattan Measurements. *IEEE Trans. Antennas Propag.* **2022**, *70*, 4096–4111. [CrossRef]
24. Juang, R.-T.; Lin, C.-L.; Tseng, C.-K.; Lee, C.-Y. Explainable Deep-Learning-Based Path Loss Prediction from Path Profiles in Urban Environments. *Appl. Sci.* **2021**, *11*, 6690. [CrossRef]
25. Chen, Q.; Xing, Y.; Chen, H. ACT-GAN: Radio Map Construction Based on Generative Adversarial Network. IET Communications (Wiley-IET). *arXiv* **2024**, arXiv:2401.08976.
26. Chen, Q.; Sun, H.; Song, J. A Distant-range Content Interaction Network for Radio Map Construction. *ICT Express* **2024**, *10*, 1145–1150. [CrossRef]
27. Wen, X.; Fang, S.; Fan, Y. Reconstruction of Radio Environment Map Based on Multi-Source Domain Adaptive of Graph Neural Network for Regression. *Sensors* **2024**, *24*, 2523. [CrossRef] [PubMed]
28. Kwon, B.; Son, H. Accurate Path Loss Prediction Using a Neural Network Ensemble Method. *Sensors* **2024**, *24*, 304. [CrossRef] [PubMed]
29. Ma, Y.; Zhang, Y.; Liu, J. Radio Map Estimation Using a CycleGAN-Based Learning Framework. *Array* **2025**. [CrossRef]
30. Sutjarittham, T.; Gharakheili, H.H.; Kanhere, S.S.; Sivaraman, V. Realizing a smart university campus: Vision, architecture, and implementation. In Proceedings of the 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Indore, India, 16–19 December 2018; pp. 1–6.
31. Popoola, S.I.; Atayero, A.A.; Popoola, O.A. Comparative assessment of data obtained using empirical models for path loss predictions in a university campus environment. *Data Brief* **2018**, *18*, 380–393. [CrossRef]
32. Onykiienko, Y.; Popovych, P.; Mitsukova, A.; Beldyagina, A.; Yaroshenko, R. Lora evaluation for university campus in urban conditions. In Proceedings of the 2021 IEEE 4th International Conference on Advanced Information and Communication Technologies (AICT), Lviv, Ukraine, 21–25 September 2021; pp. 98–101.
33. Enyi, V.S.; Eze, V.H.U.; Ugwu, F.C.; Ogbonna, C.C. Path loss model predictions for different GSM networks in the university of nigeria, nsukka campus environment for estimation of propagation loss. *IJARCCE* **2021**, *10*, 10816. [CrossRef]
34. Muñoz, J.; Mancipe, D.; Fernández, H.; Rubio, L.; Peñarrocha, V.M.R.; Reig, J. Path Loss Characterization in an Outdoor Corridor Environment for IoT-5G in a Smart Campus University at 850 MHz and 3.5 GHz Frequency Bands. *Sensors* **2023**, *23*, 9237. [CrossRef]
35. Famoriji, O.J.; Shongwe, T. Simple and effective electromagnetic wave propagation loss model in GSM band for smart campus applications. *Int. J. Electr. Electron. Eng.* **2024**, *11*, 306–311. [CrossRef]
36. Singh, H.; Gupta, S.; Dhawan, C.; Mishra, A. Path loss prediction in smart campus environment: Machine learning-based approaches. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2019; IEEE: New York, NY, USA, 2020; pp. 1–5.
37. Khalili, H.; Frey, H.; Wimmer, M.A. Balancing Prediction Accuracy and Explanation Power of Path Loss Modeling in a University Campus Environment via Explainable AI. *Futur. Internet* **2025**, *17*, 155. [CrossRef]
38. Popoola, S.I.; Atayero, A.A.; Arausi, O.D.; Matthews, V.O. Path loss dataset for modeling radio wave propagation in smart campus environment. *Data Brief* **2018**, *17*, 1062–1073. [CrossRef]
39. Popoola, S.I.; Atayero, A.A.; Faruk, N. Received signal strength and local terrain profile data for radio network planning and optimization at GSM frequency bands. *Data Brief* **2018**, *16*, 972–981. [CrossRef] [PubMed]
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
41. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Meth-Ods*; Cambridge University Press: Cambridge, UK, 2000.
42. Williams, C.K.I.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.

43. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
44. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.