Estimation of radio coverage in railway mobile communications systems

Tiago Ramos de la Cerda Gomes*, António Rodrigues* and Nuno Cota†,
*Instituto Superior Técnico
†Instituto Superior de Engenharia de Lisboa

***Abstract*—Railway communications features some specifications in terms of radio coverage and planning a radio mobile communication network that distinguish them from public communications networks, since estimation of radio coverage is one of the main stages in the planning of a radio railway communications. Nowadays these communications networks are also preparing for technological developments under the Future Railway Mobile Communications System (FRMCS), thus the knowledge of these specific features is essential in order to evaluate the applicability of the different technologies to the railway's needs. In previous works, it was shown that the propagation in different type of characteristic environment of the railway introduces significant changes in terms of prediction of radio coverage. The applicability of conventional models such as the Okumura-Hata Model to railway communications has also been demonstrated, but with several issues regarding the need to correct and adapt some parameters of the models, namely those that characterize the type of environment. The purpose of this dissertation is to develop and test a set of data in order to select the most suitable model for the estimation of radio coverage through an automatic optimization interface, namely H2O Flow. This is the classification methodology chosen in which I describe its interface and configuration that better maximizes the performance of the algorithm.**

***Index Terms*—Railway Communications, Radio signal estimation, Propagation Models, Okumura-Hata, H2O Flow**

I. Introduction

Global System for Mobile Communications - Railway (GSM-R) is a featured mobile communications system for the rail network, that has emerged from the need to create a digital communications system to accomplish the goal of technological standardisation across the rail network in Europe [1]. There is a strong need, in the means of rail transport, to ensure communications between trains. The GSM-R replaces the old systems to ensure uniformity between all lines, which is based on the Global System for Mobile Communications (GSM) system, given its robustness and reliability at the radio transmission level [2].

It's not easy to accurately calculate signal attenuation in scenarios that contain obstacles, different environments, irregular terrain, among others, since it's necessary to take into account a high number of parameters. In order to solve this problem, are used propagation models that take into account the signal propagation mechanisms in free space with the presence of obstacles and various corrective factors, which are obtained through statistical analysis in different scenarios. Thus, it is necessary to estimate radio coverage on all lines, involving various types of propagation environments. The configuration of the propagation models used for the different propagation environments and characteristics existing on the railways is then imperial. However, this adjustment of parameters of a given model is a process that can become difficult, due to the number of variables involved and the dependence between them. It becomes necessary to appeal to Machine Learning (ML) techniques, using the H2O Flow interface, which, based on a set of test data, produces a solution of parameters that minimize the error between the prediction and the actual measurements of the network.

The objective of this dissertation is to study the best propagation models for estimation and radio coverage in railway mobile communications, selecting the models that best adapt to the types of environments in these systems. In addition, it's to apply a classification methodology, with the aid of the H2O Flow interface, allowing the development of an algorithm in which, for each point, it performs the classification of the environment and, consequently, the selection of the most appropriate model for the estimation radio coverage.

This work is organized as follows: Section II describes the planning of the GSM-R system and present us with all the elements that belong in the propagation of railway lines. In Section III the propagation model used and its implementation throughout this work. In Section IV the interface used to classify each model, H2O Flow and in Section V large-scale propagation results are presented and discussed. Finally, conclusions are drawn in Section VI.

II. Railway Communications

A. GSM-R Basic Concepts

The railway communication technology GSM-R is an extension of the standard GSM, respecting the railway communication requirements, in terms of functionality and robustness. GSM-R technology has been adopted by most railways to ensure operational voice and data communications. Given the maturity of GSM current systems, the planning and optimization methodologies of the radio network are currently well defined, documented and are used by all public mobile network operators.

In terms of the radio network, the main divergences regarding the GSM standard are due to the GSM-R reaches speeds of up to $500km/h$, it also supports handovers and a faster cell selection/re-selection than in the public GSM standard. In addition, at the functional and application level, new functions were considered in order to support a more flexible use of

railway communications, such as automatic train control and emergency calls, for example [1].

B. GSM-R Network Architecture

GSM-R is standardized by European Telecommunications Standards Institute (ETSI) [3] and aims to reduce the complexity of the respective transmission base stations. Centralized network management and maintenance, as well as interconnection to other networks, are fundamental concepts of this network. The GSM-R architecture is divided into three main components, as can be seen in Figure 1:

- **Base Station Sub-System (BSS)**, section responsible for traffic and signaling between mobile terminals and the network;
- **Network Sub-System (NSS)**, component of the GSM-R system that handles mobility management, tracking mobile location and allowing mobile services to be provided to users;
- **Operation and Maintenance Center (OMC)**, connected to all equipment in the switching system and to Base Station Controller (BSC). The implementation of OMC is called Operational Support System (OSS), dedicated to telecommunications service providers and is used primarily to support network processes to maintain network inventory, configure network components, provision services and manage failures.

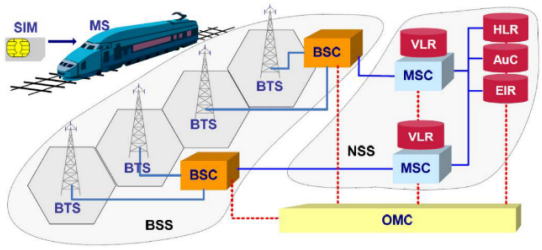

Fig. 1: GSM-R network architecture [4].

On the user side of the system, there is a Mobile Station (MS), which are several mobile receivers installed on trains, but also other terminals that depend on the GSM-R network, for example, maintenance or security railway workers and emergency services. This equipment also includes a smart card, Subscriber Identity Module (SIM), which contains information specific to each user.

C. Spectrum of the GSM-R network

The radio interface for GSM-R consists of the information flow that occurs between the train receiver (MS) and the base station (Base Transceiver Station (BTS)). In terms of spectrum, ETSI [5] has reserved two frequency bands between 876-880 MHz (uplink) and 921-925 MHz (downlink), used by European Integrated Railway Enhanced Network (EIRENE) systems. This band is called the GSM-R band or the Union Internationale du Chemin-de-Fer (UIC) band. The UIC already has defined the possibility of having an additional 200 kHz band, guard band, to increase channel availability and also the Extended GSM (E-GSM) band, based on frequencies generally reserved for government and of defense. Therefore, once the GSM-R band is insufficient for the railway communication needs in a given area, the regulator can allow the use of extra frequencies, as long as they are available. The total spectrum in the 900 MHz band, including the GSM-R band can be seen in Figure 2.

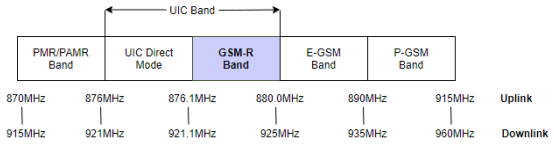

Fig. 2: Frequency allocation for GSM-R, in the 900 MHz band [4].

D. GSM-R coverage

It's mandatory to discuss the minimum levels of coverage when comparing the GSM and GSM-R. In network planning, the level of coverage is defined as the intensity of the field received at the receiving antenna, located on the roof of the train. The GSM-R norm defines coverage minimums depending on the speed and type of information received. These values are presented in the Table I and are defined only for the level received on the locomotive radio (Cab Radio), considering an isotropic antenna 4 meters high [6].

TABLE I: Minimum coverage levels.

Type	Minimum Value	Usage	Train Speed
Mandatory	$-98dBm$	Voice and non-safety critical data	—
Mandatory	$-95dBm$	ETCS levels 2/3	$\leq 220km/h$
Recommended	$-92dBm$	ETCS levels 2/3	$\geq 280km/h$

In the GSM-R network, the minimum coverage values must obey a coverage probability of at least 95%, for every $100m$ of railroad tracks, as can be seen in Figure 3. This introduces a significant difference compared to the GSM system, where the level of coverage probability is the average coverage for the entire region. In this way, it's possible to verify that the coverage requirements for GSM-R systems have a much higher level of demand than that of GSM.

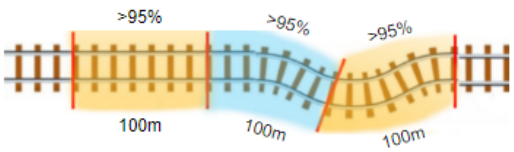

Fig. 3: Coverage probability on GSM-R [7].

E. Propagation on railway lines

Prediction of radio coverage is one of the fundamental steps in the planning process of the radio network GSM-R and

is based on the whole process of link calculation, coverage planning, frequency and capacity planning, and interference analysis. In this way, it becomes very important to correct the prediction and the way it should be adapted to the environment in question. The radio coverage prediction is based on signal estimation models, or propagation models. It's based on propagation models that the coverage prediction is made, despite presenting results that don't manifest in the maximum efficiency of the projected network [7].

One of the fundamental aspects when projecting a wireless network focuses on calculating the attenuation in the propagation of radio signals. In order to reduce the number of base stations in the installation of systems of this size, it's essential that the base stations provide the greatest possible coverage, making it possible to minimize the total cost of installation, as well as the interference caused between the various stations. It's unlikely to determine, in a concrete way, the attenuation of the signal in real life scenarios with obstacles, different means, uneven terrain, among other factors, due to the large number of parameters to be considered. In order to solve this problem, propagation models are used to take into account the signal propagation mechanisms in free space and in the presence of obstacles, as well as various corrective factors obtained through statistical analysis in different scenarios.

III. Okumura-Hata Model

This model is useful to estimate the signal path attenuation in different environments, being recommended for prediction in GSM-R [8]. In addition, it has led to very plausible results, obtaining reduced errors, when complemented with correction factors specific to the type of environment [9], being the most used model in estimating radio coverage during the planning phase of a mobile network.

A. Path loss estimation

The model provides the median value of the propagation attenuation, depending on frequency, f, distance from mobile terminal to base station, d, height from mobile terminal to the ground, h_m, and base station height, h_{be}. The calculation of the signal attenuation is one of the fundamental steps in planning mobile networks. Thus, this value is obtained by the following equation:

$$\begin{aligned} L_{p[dB]} = 69.55 + 26.16log(f_{[MHz]}) - 13.82log(h_{be[m]}) + \\ [44.90 - 6.55log(h_{be[m]})] * log(d_{[km]}) - \\ H_{mu[dB]}(h_m, f) - \sum correction\ factors \end{aligned} \tag{1}$$

The term $H_{mu[dB]}$, is a correction term, which depends on the environment. Thus, the equation 2 represents a correction proposed by the model for a basic suburban environment. This correction considers the height of the mobile terminal and the frequency.

$$\begin{aligned} H_{mu[dB]} = [1.10log(f_{[MHz]}) - 0.70]h_{m[m]} - \\ [1.56log(f_{[MHz]}) - 0.80] \end{aligned} \tag{2}$$

Thus, by analyzing the equation 3 we have the sum of the correction factors, of the Okumura-Hata model, which is obtained by the influence of vegetation, water influence, terrain undulation, along path correction and, finally, losses associated with diffraction in the obstacles, through the Deygout Model.

$$\begin{aligned} \sum correction\ factors = -L_v + K_{mp} + K_{th} \\ + K_{hp} + K_{al} - L_{diff} \end{aligned} \tag{3}$$

B. Measurements

In order to check the radio coverage prediction model used and corroborate the methodology implemented in the estimation of radio coverage along the railroad, extensive radio measurement activities were carried out by REFER-Telecom. In fact, the choice of the railway line for the case study was due to the availability of these radio coverage measures. These measurements were acquired due to the installation of emission equipment in different locations along the entire length of the line, and reception equipment, added on board the train, allowing full coverage of the line.

C. Error Statistics

In order to obtain a better interpretation and similarities between the prediction and the measures, first order statistics and the correlation coefficient were calculated.

The resulting statistics aim to assess the global error of radio signal prediction and are translated by the equations below, namely Medium Error (ME), Root Mean Square Error (RMSE) and Estimated Standard Deviation (ESD):

$$ME = \frac{1}{n}\sum_{i=1}^{n}|P_{meas_i} - P_{pred_i}| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|P_{meas_i} - P_{pred_i}|^2} \tag{5}$$

$$ESD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|P_{meas_i} - P_{pred_i}| - ME)^2} \tag{6}$$

where P_{meas_i} is the signal level (in dBm) of the signal measured at i, n being the total number of points and P_{pred_i}, the equivalent value of the prediction.

The correlation coefficient provides a measure of the degree of linear relationship between measured and predicted variables and is calculated as,

$$RE = \frac{\sum_{i=1}^{n}(P_{meas_i} - \bar{P}_{meas})(P_{pred_i} - \bar{P}_{pred})}{\sqrt{\sum_{i=1}^{n}(P_{meas_i} - \bar{P}_{meas})^2}\sqrt{\sum_{i=1}^{n}(P_{pred_i} - \bar{P}_{pred})^2}} \tag{7}$$

D. Classification

Classification is the process of analyzing all points on the railway line and determining which means has the lowest value of RMSE at each point. In this way, it is possible to assign to each of the points the means that minimizes the error. The classification of the media is done according to RMSE. The selection of this error is based on the fact that it generates better results, minimizing the other statistics (ESD and ME) and maximizing Coeficiente de correlação (RE).

Subsequently, the assessment is made, which resides in evaluating the model selected in the classification and compares it with the actual measurements of the network. Consequently, it will be possible to elaborate a new signal prediction that approximates the curvature of the measures. The evaluation will be analyzed in more detail in chapter 5 of this work.

IV. H2O Flow

Looking for another classification solution to the radio coverage estimation, a Data Mining (DM) model was developed using a open-source platform, H2O. H2O is a machine learning (ML) and predictive, in-memory, distributed, fast and scalable interface that allows you to create machine learning models on large data sets and also allows easy production of these models in business environments. Its mission is to democratize the use of artificial intelligence.

A. Data Mining

DM is the extraction of implicit information from the data, previously unknown, and potentially useful [10]. The assumed idea of the DM process is to create a computer model, for example, an application, which makes it possible to examine and extract patterns from the data in a simple, fast and automatic way. These patterns are then used to detect dependencies between data, specific cases, interpret, understand, predict or classify the new data [10].

B. Data Dimension

One of the problems questioned at different stages of the Data Mining process is the size of the data. This issue raises two essential points, low performance of the algorithms due to the slow execution of them and the lack of resources that makes the same execution impossible. In order to minimize this problem, only Ponto Quilométrico (PK) with measurements was considered instead of using all points of each line, for the optimization of the algorithm.

1) Data Set Balancing: Balancing the data is another problem, as it affects the precision of the classification. In these cases, the learning algorithms tend to specialize in classifying the majority class and ignore the cases of the minority [11]. Since the data is not balanced, the following techniques will be used, all of which are incorporated in H2O Flow which automatically balances the data:

- Undersampling;
- Oversampling;
- Synthetic Minority Oversampling Technique (SMOTE).

C. Datasets: train and test

One of the steps in the DM process focuses on running tests to validate the results, applying algorithms to the data set. Therefore, the validation of the model focuses on the execution of tests and on the comparison of the results obtained with the expected results. It is intended that the model is applied to new cases and that it's able to not only focus on the classification of known cases, but also to classify new cases in an equivalent way or to classify with better performance, when possible.

In this work, the hold-out technique was then used, as it is a good choice, given its simplicity, widespread acceptance and its applicability to the problem in question. Thus, a usually accepted distribution of $1/3$ of data for the test set and $2/3$ of the data for the training set was used, since the minority class has few cases and the set is quite unbalanced [10]. In Figure 4 we can see an illustrative representation of the technique.

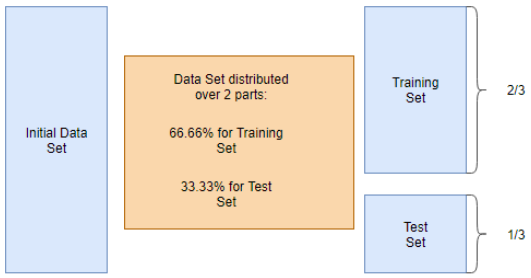

Fig. 4: Representation of the hold-out technique.

D. Machine Learning Algorithms

In this work, the learning types are considered based on decision trees, since these are the most suitable for non-binary classes. So, given that the intended classification is carried out by three classes, that is, it's intended to classify the model in three different propagation environments: urban, suburban and rural.

In order to understand which is the best algorithm to use, the Automatic Machine Learning (AutoML) algorithm was used, which is a set of ML methods and automatically finds the best algorithms taking into account a set of data training. Thus, after the automatic choice, AutoML came to the conclusion that the best classification algorithms, for our specific training data set, are Generalized Linear Model (GLM) and XGBoost.

As a form to evaluate and compare the different algorithms, the confusion matrix was chosen. The choice of a good evaluation method is essential to obtain the most appropriate algorithm and obtain the best performance in the classification of the model.

E. H2O Flow implementation

The data set is made available by a CSV file, which contains information about the parameters that can influence the signal prediction, as well as the classification of the medium for each point of PK. The unit of measurement of the parameters is the meter.

Figure 5 shows the sequence of steps that will be implemented by H2O Flow to train the data set and to make its prediction.

Fig. 5: H2O Flow implementation sequence.

After entering the data, the training and test sets will be distributed, which will serve to build the model and, later on, in the interpretation of results is made an analysis and interpretation of the classifications obtained by the algorithms GLM and XGBoost.

V. Results

Once the algorithm settings were stabilized and the propagation model was defined, some tests were carried out on the railway lines analyzed: Cascais, Beira Baixa and Algarve. This chapter enables the final configuration of the algorithm, the analysis of the results obtained in the different approaches followed during this work and also the classification and prediction obtained through ML techniques using H2O Flow.

A. Propagation Model Configuration

As previously mentioned, the model covers a series of parameters that can be calibrated according to the environment, given by the expression 1. The model consists of the Okumura-Hata model for the three propagation models with the respective correction factors and with additional losses due to diffraction, obtained through the Deygout method.

1) Evaluation and optimization of the original model: The geographic information collected, together with the model parameters, allows the composition of a prediction made by the propagation model. Based on the error between the prediction and the measurements in the different models, a new prediction was made taking into account the optimization of the three models in relation to the measurements, at all points of the railway line.

The evaluation consists in evaluating the model chosen in the classification, described in III-D, and comparing it with the real measurements of the network. In Table II it's possible to observe the errors of the optimized model compared to the previous model, for the first tested scenario: Cascais line. This line is characterized by the presence of water along almost the entire legth of the line and a suburban environment.

TABLE II: Statistics of the initial model and the optimized model, Cascais line.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	11,3896	13,4517	7,1572	0,2513
Suburban	7,0269	8,5912	4,9429	0,2513
Rural	18,2626	20,1789	8,5828	0,2513
Optimized	3,9381	4,9169	2,9441	0,8609

As the Cascais line is characterized by a suburban environment, it's expected that among the three models, the suburban will be the one with the best results, as shown in Table II. It should also be noted that it was possible to minimize the three errors by 3.1 dB, 3.6 dB and 2 dB regarding ME, RMSE and ESD, respectively. In turn, RE was maximized by 61%.

With the minimization of the error, it was possible to approximate the curvatures between the prediction and the measures.This can be verified in Figure 6. A much more adjacent prediction was thus obtained compared to the measurements of the railway line.

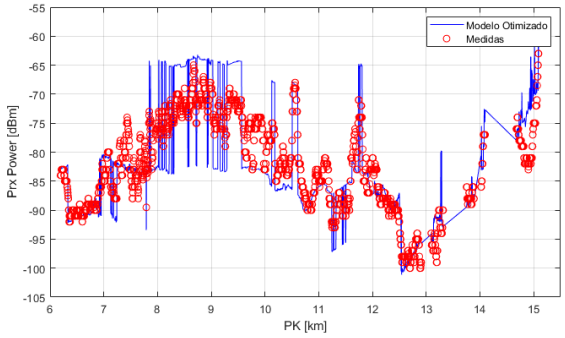

Fig. 6: Comparison of the optimized model with the measurements, Cascais line.

The second scenario considered was the Beira Baixa line. The line is characterized by a rural environment with mountainous terrain and with some sections of the line characterized by a suburban environment, with these two models showing the best results. In Table III it is possible to check the errors of the optimized model compared to the previous model for this line.

TABLE III: Statistics of the initial model and the optimized model, Beira Baixa line.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	24,8360	28,0728	13,0855	0,8415
Suburban	17,9717	20,9590	10,7842	0,8415
Rural	13,9847	17,9774	11,2967	0,8415
Optimized	7,2810	9,0895	5,4411	0,9209

Note that it was possible to minimize the three errors by 6.7 dB, 8.9 dB and 5.9 dB regarding ME, RMSE and ESD, respectively. In turn, RE was maximized by around 8%.

With the minimization of the error, it was possible to approximate the curvatures between the prediction and the measures, verified in Figure 7. Thus, a much more adjacent prediction was obtained compared to the measurements of the railway line, although it was not possible to optimize as much as in the Cascais line, since the Beira Baixa line is characterized by a mountainous and rugged terrain, making it more difficult to estimate the radio coverage in some points of the line. Figure 7 shows the fourth test trip on the Beira Baixa line.

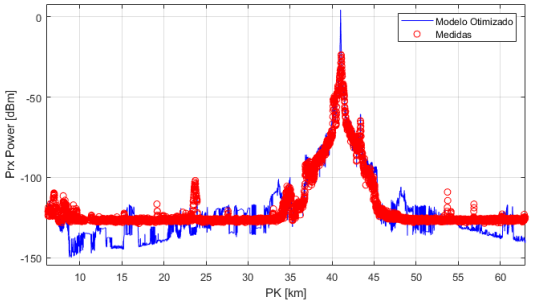

Fig. 7: Comparison of the optimized model with the measurements, Beira Baixa line.

The last scenario considered was the Algarve line. This is essentially characterized by a suburban environment and the presence of water in some points of the line and also by the presence of many obstacles between the base stations and the different points of the line. In Table IV it's possible to check the errors of the optimized model compared to the previous model for this line.

TABLE IV: Statistics of the initial model and the optimized model, Algarve line.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	22,3371	31,2286	21,8238	0,3454
Suburban	17,1325	25,6574	19,0993	0,3454
Rural	20,2025	24,5314	13,9158	0,3454
Optimized	11,4022	14,6884	9,2595	0,9231

It should be noted that the suburban and rural models showed the best results, given that the Algarve line is characterized by the presence of water and obstacles at different points on the line. In addition, it was possible to minimize the three errors by 5.7 dB, 10 dB and 4.7 dB regarding ME, RMSE and ESD, respectively. In turn, RE was maximized by around 58%.

Figure 8 shows the fourth test trip on the Algarve line. The difference in the error obtained can be clearly seen, particularly at points more distant from the base stations, however, the proximity of the curvatures to the prediction and measurements along the line is visible.

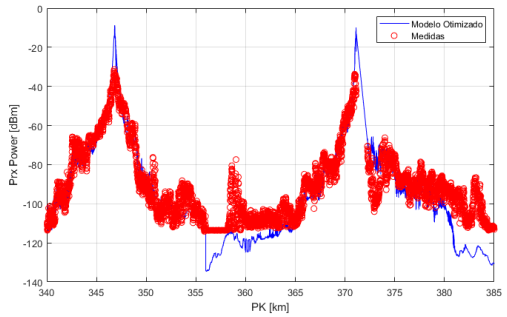

Fig. 8: Comparison of the optimized model with the measurements, Algarve line.

B. Scenarios Classification

Once the parameters of the algorithm were established and the propagation model was defined, several tests were carried out in 3 different scenarios: the Cascais line, the Beira Baixa line and the Algarve line.

In the first phase, it was intended to study each scenario individually with the purpose to correctly classify the three environments on each of the railway lines. For each set of parameters, a training set, consisting of $2/3$ of the data, and a test set, consisting of $1/3$ of the data, are selected randomly by H2O Flow. The training set is introduced in AutoML and gives rise to the previous classification of the model, which in turn is applied to the test set. To validate the algorithm, the same methodology is used for the three railway lines in question.

The number of measurement points will vary from line to line, as the lines have different lengths. Therefore, the training and test sets have different dimensions and correspond, respectively, to $2/3$ and $1/3$ of the total number of points in each scenario, Table V.

TABLE V: Number of points on each railway line.

Scenario	Number of points	Training set (2/3)	Test set (1/3)
Cascais's line	1952	1286	666
Beira Baixa's line	4673	3101	1572
Algarve's line	6749	4557	2192

1) Cascais Line: The first scenario considered for training is the Cascais line. It's important to note that there is a large discrepancy in the number of points in relation to the three models. Effectively, we are facing an unbalanced set where the cases triple in the suburban model compared to the rural model and twice as much in relation to the urban model. This scenario affects the credibility of the classification obtained and because of this AutoML concluded that the GLM algorithm would not be a good solution, since it was unable to obtain enough data for an adequate training set.

Therefore, AutoML found the XGBoost algorithm to be the one with the best performance in view of the various statistical errors, with an emphasis on RMSE. Tables VI and VII show the confusion matrices for the training and test sets, respectively.

TABLE VI: Confusion matrix for the training set for the XGBoost algorithm, Cascais line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	231	0	0	0	0/231	**1,0**
	Suburban	0	715	1	0,0014	1/716	**1,0**
	Urban	0	0	339	0	0/339	**1,0**
	Total	231	715	340	0,0008	1/1286	-
	Precision	**1,0**	**1,0**	**1,0**	-	-	-

As was expected for the training set, a precision of 100% was obtained, occurring overfitting, since the algorithm was able to correctly classify all the points present in the training set. It should be noted that excellent precision values were obtained for the test set, namely 92% for the classification in rural areas, 96% for the suburban environment and 99% for

TABLE VII: Confusion matrix for the test set for the XGBoost algorithm, Cascais line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	101	8	0	0,0734	8/109	**0,93**
	Suburban	9	351	2	0,0304	11/362	**0,97**
	Urban	0	7	188	0,0359	7/195	**0,96**
	Total	110	366	190	0,0390	26/666	-
	Precision	**0,92**	**0,96**	**0,99**	-	-	-

the environment urban. For example, when the algorithm was supposed to predict the urban class, it was able to classify 188 points correctly as urban and only 7 points incorrectly.

2) Beira Baixa Line: The second scenario considered was the Beira Baixa line. The line stands out for a rural environment with mountainous terrain. In this line, and as expected, there is also a large discrepancy in the number of points in relation to the three models. Effectively, this scenario affects the credibility of the classification obtained and, in turn, AutoML concluded that the GLM algorithm would not be a good solution.

Therefore, AutoML found the XGBoost algorithm to be the one with the best performance in view of the various statistical errors. Tables VIII and IX show the confusion matrices for the training and test sets, respectively.

TABLE VIII: Confusion matrix for the training set for the XGBoost algorithm, Beira Baixa line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	1909	0	0	0	0/1909	**1,0**
	Suburban	0	713	0	0	0/713	**1,0**
	Urban	0	0	479	0	0/479	**1,0**
	Total	1909	713	479	0	0/3101	-
	Precision	**1,0**	**1,0**	**1,0**	-	-	-

TABLE IX: Confusion matrix for the test set for the XGBoost algorithm, Beira Baixa line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	971	13	0	0,0132	13/984	**0,99**
	Suburban	21	316	22	0,1198	43/359	**0,88**
	Urban	0	11	218	0,0480	11/229	**0,95**
	Total	992	340	240	0,0426	67/1572	-
	Precision	**0,98**	**0,93**	**0,91**	-	-	-

Just like on the Cascais line, for the training set a precision of 100% was obtained, occurring overfitting, since the algorithm was able to correctly classify all the points present in the training set. In addition, it should be noted that excellent precision values were obtained for the test set, namely 98% for classification in rural areas, 93% for suburban areas and 91% for the urban environment. For example, when the algorithm was supposed to predict the urban class, it was able to classify 218 points correctly as urban and only 11 points incorrectly.

3) Algarve Line: The third scenario considered was the Algarve line. This line is characterized by a suburban environment and by the presence of obstacles between the various base stations and the various points on the line. In this line, and as expected, there is also a large discrepancy in the number of points in relation to the three models. As in the other two lines, this scenario affects the credibility of the classification obtained and AutoML concluded that the GLM algorithm would not be a good solution.

As in the other two scenarios, AutoML found the XGBoost algorithm to be the one with the best performance in view of the various statistical errors. Tables X and XI show the confusion matrices for the training and test sets, respectively.

TABLE X: Confusion matrix for the training set for the XGBoost algorithm, Algarve line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	1793	0	0	0	0/1793	**1,0**
	Suburban	0	1684	0	0	0/1684	**1,0**
	Urban	0	0	1080	0	0/1080	**1,0**
	Total	1793	1684	1080	0	0/4557	-
	Precision	**1,0**	**1,0**	**1,0**	-	-	-

TABLE XI: Confusion matrix for the test set for the XGBoost algorithm, Algarve line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	867	22	0	0,0247	22/889	**0,98**
	Suburban	19	674	79	0,1269	98/772	**0,87**
	Urban	1	58	472	0,1111	59/531	**0,89**
	Total	887	754	551	0,0817	179/2192	-
	Precision	**0,98**	**0,89**	**0,86**	-	-	-

Such as in the other two lines under study, for the training set a precision of 100% was obtained, occurring overfitting, since the algorithm was able to correctly classify all the points present in the training set. Furthermore, it should be noted that very satisfactory precision values were obtained for the test set, namely 98% for classification in rural areas, 89 % for suburban areas and 86% for urban areas. For example, when the algorithm was supposed to predict the rural class, it was able to classify 867 points as rural environment correctly and only 22 points incorrectly.

C. Evaluation and optimization through H2O Flow

Based on the classification obtained by XGBoost, it's elaborated a new prediction for each scenario considered. Subsequently, this new prediction was compared with the prediction obtained by the optimized model in order to check the reliability of the use of machine learning techniques to estimate radio coverage.

1) Cascais Line: The XGBoost algorithm obtained excellent values of recall and precision in relation to the prediction obtained through the classification. Table XII shows the confusion matrix of the prediction, for the Cascais line. Note that the precision values of 98% were obtained for the rural and suburban models and 100% for the urban model. This means that the XGBoost algorithm was able to correctly predict large parts of the classes, thus enabling a good prediction of radio coverage.

In Table XIII it is possible to observe the errors of the model optimized by H2O Flow in comparison with the three original models and with the optimized model.

Note that it was possible to optimize the initial model with the prediction made by H2O Flow. In addition, this

TABLE XII: Confusion matrix of the Cascais line prediction.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	326	14	0	0,0412	14/340	**0,96**
	Suburban	8	1068	2	0,0093	10/1078	**0,99**
	Urban	0	8	526	0,0150	8/534	**0,99**
	Total	334	1090	528	0,0164	32/1952	-
	Precision	**0,98**	**0,98**	**1,0**	-	-	-

TABLE XIII: Cascais line statistics.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	11,3896	13,4517	7,1572	0,2513
Suburban	7,0269	8,5912	4,9429	0,2513
Rural	18,2626	20,1789	8,5828	0,2513
Optimized	3,9381	4,9169	2,9441	0,8609
H2O Flow	3,9681	4,9668	2,9872	0,8546

prediction was very similar to the prediction obtained by the optimized model, differing 0.03 dB, 0.05 dB and 0.04 dB from ME, RMSE and ESD, respectively . In Figure 9 it's possible to verify that the prediction curve obtained by H2O Flow overlaps with the optimized model, which demonstrates the proximity of the models.

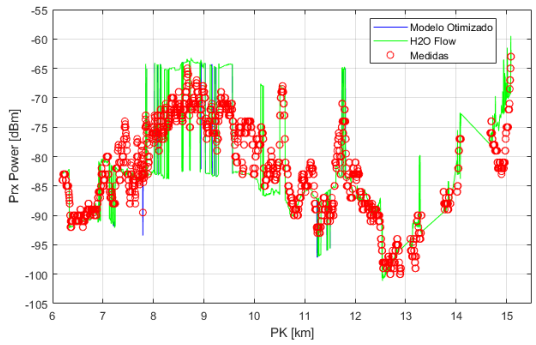

Fig. 9: Comparison of the model obtained by H2O Flow with the optimized model and the measurements, Cascais line.

2) Beira Baixa Line: For the Beira Baixa line, the XGBoost algorithm also obtained excellent values of sensitivity and precision in relation to the prediction obtained through the classification performed. Table XIV shows the confusion matrix of the prediction made for this scenario.

TABLE XIV: Confusion matrix of the Beira Baixa line prediction.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	2879	14	0	0,0048	14/2893	**1,0**
	Suburban	21	1037	14	0,0326	35/1072	**0,97**
	Urban	0	19	689	0,0268	19/708	**0,97**
	Total	2900	1070	703	0,0146	68/4673	-
	Precision	**0,99**	**0,97**	**0,98**	-	-	-

Notice that, in relation to the prediction made by the algorithm, precision and recall values of 99% and 100%, respectively, were obtained for the rural model, given the greater number of points for this model. It's possible that overfitting occurred as it reached extremely high values. For the suburban model, precision and recall values of 97% were obtained, and finally, for the urban model, precision and recall values of 98% and 97%, respectively. In Table XV it's possible to observe the errors of the model optimized by H2O Flow in comparison with the three original models and with the optimized model.

TABLE XV: Estatísticas da linha da Beira Baixa.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	24,8360	28,0728	13,0855	0,8415
Suburban	17,9717	20,9590	10,7842	0,8415
Rural	13,9847	17,9774	11,2967	0,8415
Optimized	7,2810	9,0895	5,4411	0,9209
H2O Flow	7,3476	9,1420	5,4396	0,9194

Notice that it was possible to optimize the initial model with the prediction made by H2O Flow. In addition, this prediction was very similar to the prediction obtained by the optimized model, differing 0.07 dB, 0.05 dB and 0.002 dB compared to ME, RMSE and ESD, respectively. In Figure 10 it's possible to verify that the prediction curve obtained by H2O Flow overlaps with the optimized model, which demonstrates the proximity of the models.

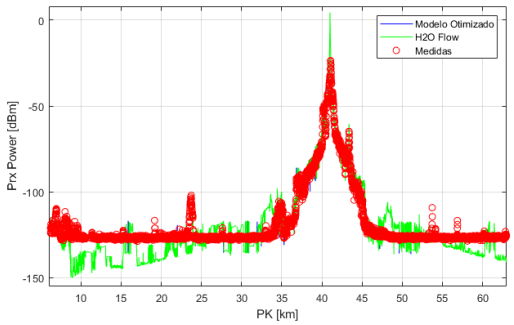

Fig. 10: Comparison of the model obtained by H2O Flow with the optimized model and the measurements, Beira Baixa line.

3) Algarve Line: For the Algarve line, the XGBoost algorithm also obtained very satisfactory values of recall and precision in relation to the prediction obtained through the classification performed. Table XVI shows the confusion matrix of the prediction made for this scenario.

TABLE XVI: Confusion matrix of the Algarve line prediction.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	2659	23	0	0,0086	23/2682	**0,99**
	Suburban	19	2386	51	0,0285	70/2456	**0,97**
	Urban	0	67	1544	0,0416	67/1611	**0,96**
	Total	2678	2476	1595	0,0237	160/6749	-
	Precision	**0,99**	**0,96**	**0,97**	-	-	-

Note that, in relation to the prediction made by the algorithm, precision and recall values of 99% for the rural model were obtained. For the suburban model, precision and recall values of 96% and 97%, respectively, were obtained. Finally, for the urban model, precision and recall values of 97% and 96%, respectively, were obtained.

In the Table XVII it's possible to observe the errors of the model optimized by H2O Flow in comparison with the three

original models and with the optimized model. Notice that it was possible to optimize the initial model with the prediction made by H2O Flow. Furthermore, this prediction was very similar to the prediction obtained by the optimized model, differing 1.94 dB, 2.16 dB and 4.68 dB compared to ME, RMSE and ESD, respectively . However, it was the scenario that showed higher error values compared to the optimized model.

TABLE XVII: Algarve line statistics.

Model	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urban	22,3371	31,2286	21,8238	0,3454
Suburban	17,1325	25,6574	19,0993	0,3454
Rural	20,2025	24,5314	13,9158	0,3454
Optimized	11,4022	14,6884	9,2595	0,9231
H2O Flow	9,4619	16,8459	13,9376	0,5675

In Figure 11 it's possible to verify that the prediction curve obtained by H2O Flow largely overlaps with the optimized model, which demonstrates the proximity of the models.

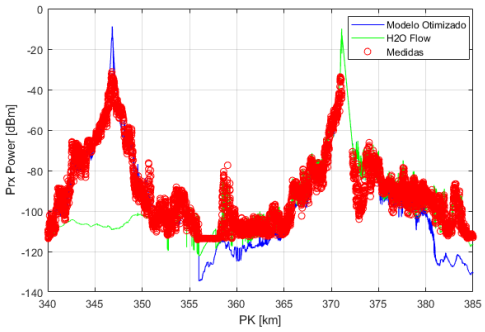

Fig. 11: Comparison of the model obtained by H2O Flow with the optimized model and the measurements, Algarve line.

4) General Classification and Prediction: After analyzing the lines separately, a new set of data that consists of the joining of the three lines was considered, with the objective of standardizing the model obtained by XGBoost, in order to achieve good prediction values for other railway lines. As previously mentioned, and as seen in the Table V, the Cascais line has only 1952 points and the ratio between the Beira Baixa line and this previously mentioned is 2.39, on the other hand the ratio with the Algarve line is 3.46. In order to have an equal representation of all the lines, the original set of the Beira Baixa and Algarve lines was randomly divided, in the H2O Flow, into two sets: one to create the new training set and a second for the test set. So the new training set took all the points of the Cascais line (1952), plus, approximately 45% of the original set of the Beira Baixa line (2114) and, approximately 30% of the original set of the line of the Algarve (2039). The dimension of the total number of points in the new set is shown in Table XVIII.

Due to the unbalanced data set, there is a discrepancy in the number of points in relation to the three models, found in Table XVIII. It was necessary to balance the data before applying the XGBoost algorithm.

TABLE XVIII: Number of points for training and test sets.

Scenario	Number of points	Training Set (2/3)	Test Set (1/3)
General Model	6105	4061	2044
Urban Model	1335	886	449
Suburban Model	2334	1545	789
Rural Model	2436	1630	806

It should be noted that for the test set, XGBoost obtained an average squared error similar to the errors obtained for each scenario. This obtained very satisfactory results with very high precision and sensitivity values demonstrated in Tables XIX and XX.

TABLE XIX: Confusion matrix for the training set for the XGBoost algorithm, for the general model.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	1630	0	0	0	0/1630	**1,0**
	Suburban	0	1545	0	0	0/1545	**1,0**
	Urban	0	0	886	0	0/886	**1,0**
	Total	1630	1545	886	0	0/4061	-
	Precision	**1,0**	**1,0**	**1,0**	-	-	-

TABLE XX: Confusion matrix for the test set for the XGBoost algorithm, for the general model.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	790	33	0	0,0401	33/823	**0,96**
	Suburban	16	712	39	0,0717	55/767	**0,93**
	Urban	0	44	410	0,0969	44/454	**0,90**
	Total	806	789	449	0,0646	132/2044	-
	Precision	**0,98**	**0,90**	**0,91**	-	-	-

As in each independent scenario, overfitting also occurred for the training set, that is, a precision of 100% was obtained for the three models. On the other hand, compared to the test set, the algorithm obtained precision and recall values above 90%, so we can conclude that the algorithm was able to classify, in general, correctly the points present in the test set. Effectively, XGBoost made possible excellent classification results for the models, regardless of the railway line.

Having defined the classification for the new data set, the prediction was made, using the XGBoost algorithm, for the data set of the Algarve line that were not used previously, serving as a test for this prediction. Table XXI shows the confusion matrix of the prediction elaborated for this general model applied to the Algarve line.

TABLE XXI: Confusion matrix of general model prediction, Algarve line.

		Predicted Values					
		Rural	**Suburban**	**Urban**	**Error**	**Rate**	**Recall**
Actual Values	**Rural**	1801	49	11	0,0322	60/1861	**0,97**
	Suburban	42	1615	149	0,1058	191/1806	**0,89**
	Urban	30	100	914	0,1245	130/1044	**0,88**
	Total	1872	1764	1074	0,0793	270/4710	-
	Precision	**0,96**	**0,92**	**0,85**	-	-	-

Note that precision and recall values above 96% were obtained for the rural model, since it has a higher number of points, and a better prediction for this model is normal to expect. In relation to the other two models, the prediction was satisfactory, with precision and recall values of 92% and 89%,

respectively, for the suburban model. For the urban model, precision and recall values of 85% and 88% were achieved.

D. Parameter Analysis

For a more accurate classification, it's necessary to define which parameters are most relevant in calculating the radio signal attenuation and, in turn, in estimating radio signal coverage. Table XXII shows the percentages by parameter of each lines and the general model.

TABLE XXII: Percentage values for each parameter in the different scenarios studied.

Parameter	Cascais	Beira Baixa	Algarve	General Model
Distance (d)	50,33	34,60	31,19	36,54
Mixed Paths, (β)	10,32	8,40	9,26	12,31
Effective height, (h_{be})	10,46	20,94	8,32	10,80
Terrain undulation, (Δ_h)	14,46	9,51	16,49	13,66
Position in terrain undulation, (Δ_{hm})	7,76	6,41	9,31	7,24
Main obstacle (v_1)	2,78	8,25	6,33	12,49
2nd order obstacle(v_2)	1,95	5,69	7,16	3,09
3rd order obstacle (v_3)	1,94	6,20	11,94	3,87

Through Table XXII it's possible to note that there are differences in regard to the relevance of the various parameters in the four scenarios considered. For the general model, the distance between the base station and the mobile terminal is the parameter that stands out the most, having the greatest influence on the classification of the model. The Cascais line, as it is a suburban area with the presence of water surfaces in almost the entire line, has parameters of greater relevance such as the characterization of the presence of water and the obstacles present between the base station and the various measured points. The Beira Baixa line, as it is a rural area, parameters of greater relevance such as effective height and the characterization of the terrain undulation. Finally, the Algarve line has a strong impact on the terrain undulation and the effective height. As it is a suburban area with several obstacles, the characterization of obstacles between base stations and measured points is also considered as the parameter of greatest impact.

VI. Conclusion

The applicability of ML techniques, using H2O Flow, in the classification of propagation models, when applied to the prediction of radio coverage in railway environments, was verified through the results obtained.

In the first test, a set of data for training and a set of data for testing, composed of the various parameters that define the propagation model, were selected for each scenario under study. The application of the XGBoost algorithm to the training set originated the previous classification for each point on the railway line, which was then applied to the test set. Having the classification defined, the prediction was obtained, through the XGBoost algorithm, in all considered scenarios. The XGBoost algorithm made it possible to approximate the measurement prediction, decreasing the statistics ME, RMSE and ESD and increasing the correlation. As expected, the prediction made through the classification of the models, with the help of H2O Flow did not produce better results compared to the prediction of the optimized model, however, it reached very close error values.

The second test consisted in creating a standardize model for each line, this means that a new set of data was created with the intend of joining the 3 lines under study. The remaining points of the Algarve line were used as a test, in order to achieve good classification and prediction results for the Algarve line through the training set of the three lines together. The results obtained in this test lead us to believe that it's possible to achieve a good classification and prediction on a given railway line, using a training set from another different line.

In conclusion, this dissertation verified the applicability of ML techniques and the XGBoost algorithm, through H2O Flow, in different scenarios, for the classification, and consequently, the selection of the most appropriate model suitable for radio coverage in railway environments. This technique is, therefore, a viable alternative to consider in relation to the optimized Okumura-Hata propagation model, since both predictions were very similar.

References

[1] J. Soure, "Implementation of the GSM-R System in the National Railway Network - Pilot Project," ISEC, October 2013.

[2] T. Correia, "Estimation of radio coverage in GSM-R through neural networks," ISEL, December 2014.

[3] UIC ERTMS, "FFFS for Functional Addressing," 2006.

[4] REFER Telecom/ISEL, "Methodology for Radio Planning in GSM-R," Lisbon, 2009.

[5] ETSI, ETSI EN 301 515 v2.3.0, "Global System for Mobile Communication (GSM); Requirements for GSM operation on railways."

[6] N. Cota, A. Serrador, N. Franco, and J. Neves, "Radio Planning in GSM-R: Methodology and Characterization of the Signal," URSI, Lisbon, 2009.

[7] N. Cota, A. Serrador, P. Vieira, J. Neves, and A. Rodrigues, "An Enhanced Radio Network Planning Methodology for GSM-R Communications," in Conftele 2013 - 9th edition of the Conference on Telecommunications, Castelo Branco, Portugal, 2013.

[8] *UIC, GSM-R Procurement Guide Version 5.0*, February 2007.

[9] N. Cota, A. Serrador, P. Vieira, A. R. Beire, and A. Rodrigues, "On the Use of Okumura–Hata Propagation Model on Railway Communications," *Wireless Personal Communications Symposium (WPMC2013)*, Atlantic City, New Jersey, USA, 2013.

[10] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.