TÉCNICO LISBOA

Estimação de cobertura rádio em sistemas de comunicações móveis ferroviárias

Tiago Ramos de la Cerda Gomes

Dissertação para obtenção do Grau de Mestre em

Engenharia Eletrotécnica e de Computadores

Orientadores: Prof. António José Castelo Branco Rodrigues
Prof. Nuno António Fraga Juliano Cota

Júri

Presidente: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino
Orientador: Prof. António José Castelo Branco Rodrigues
Vogal: Prof. Francisco António Bucho Cercas

setembro 2020

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Agradecimentos

Gostaria de agradecer ao meu co-orientador, Professor Nuno Cota e ao meu orientador, Professor António Rodrigues por toda a orientação e apoio ao longo do desenvolvimento deste trabalho. Toda a orientação, motivação e aconselhamento foram cruciais para superar todas as adversidades e problemas, e concluir esta dissertação totalmente. Além disso, o meu sincero agradecimento à Professora Matilde Pato, incansável no seu apoio, dedicação e disponibilidade na parte final do meu trabalho, salientando que a sua ajuda foi essencial para a concretização deste trabalho. Também gostaria de agradecer à Solvit, em particular, à Rita Beire pela sua ajuda na introdução ao tema desta dissertação, bem como pelos recursos que me disponibilizou para a realização deste trabalho. Gostaria de agradecer ao Instituto de Telecomunicações por me fornecer os meios para a conclusão desta dissertação.

Agradeço à minha família , em especial aos meus pais e à minha irmã, por todas as oportunidades que me têm dado e que sempre me têm apoiado e motivado para fazer melhor ao longo de todo o curso, mesmo quando tudo parecia mais difícil.

À minha namorada pelo incentivo e apoio incondicional que teve ao longo do trabalho.

Por último, mas não menos importante, a todos os meus amigos e colegas que me ajudaram a crescer como pessoa e sempre estiveram presentes para mim durante os bons e maus momentos da minha vida académica. Obrigado.

Resumo

As comunicações móveis ferroviárias apresentam algumas especificidades em termos de planeamento e cobertura rádio que as distinguem das redes de comunicações públicas, sendo a estimação de cobertura rádio uma das principais etapas no planeamento de uma rede rádio de comunicações ferroviárias. Numa altura em que, também estas redes de comunicações se preparam para uma evolução tecnológica, no âmbito do *Future Railway Mobile Communications System* (FRMCS), é fundamental o conhecimento destas especificidades por forma a avaliar a aplicabilidade das diferentes tecnologias às necessidades ferroviárias.

No âmbito de trabalhos anteriores foi demonstrado que a propagação no tipo de ambiente característico da ferrovia introduz alterações significativas em termos de predição de cobertura rádio. Foi igualmente demonstrada a aplicabilidade de modelos convencionais, como o Modelo Okumura-Hata às comunicações ferroviárias, tendo contudo em conta a necessidade de correção e adaptação de alguns parâmetros dos modelos, designadamente os parâmetros que caracterizam o tipo de ambiente.

O objetivo desta dissertação é desenvolver e testar um conjunto de dados, em que para cada ponto efetue a classificação do ambiente, e por conseguinte, selecione qual o modelo adequado à estimação de cobertura rádio, através de uma interface de otimização automática, nomeadamente o *H2O Flow*. Foi esta a metodologia de classificação escolhida tendo sido descrita a sua interface e configuração que maximiza o desempenho do algoritmo, para este tipo de problema.

Palavras Chave

Comunicações rádio em ferrovias; Estimação de sinal; Modelos de Propagação; Okumura-Hata; *H2O Flow*.

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 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 is described its interface and configuration that better maximizes the performance of the algorithm.

Keywords

Railway Communications; Radio signal estimation; Propagation Models; Okumura-Hata; H2O Flow.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

AuC	*Authentication Center*
ANACOM	Autoridade Nacional para as Comunicações
AutoML	*Automatic Machine Learning*
BS	*Base Station*
BSC	*Base Station Controller*
BSIC	*Base Station Identify Code*
BSS	*Base Station Sub-System*
BTS	*Base Transceiver Station*
DM	*Data Mining*
E-GSM	*Extended GSM*
EIR	*Equipment Identity Register*
EIRENE	*European Integrated Railway Enhanced Network*
ESD	*Estimated Standard Deviation*
ETCS	*European Train Control System*
ETSI	*European Telecommunications Standards Institute*
FN	Falsos Negativos
FP	Falsos Positivos
FRMCS	*Future Railway Mobile Communications System*
GBM	*Gradient Boosting Machine*
GLM	*Generalized Linear Model*
GLRM	*Generalized Low Rank Models*
GSM	*Global System for Mobile Communications*

GSM-R *Global System for Mobile Communications - Railway*

ISEL Instituto Superior de Engenharia de Lisboa

HLR *Home Location Register*

ITU-R *International Telecommunication Union - Radiocommunication sector*

LDA *Location Dependent Addressing*

LTE *Long-Term Evolution*

LTE-R *Long-Term Evoluiton - Railways*

ME *Medium Error*

ML *Machine Learning*

MORANE *Mobile Radio for Railways Networks in Europe*

MS *Mobile Station*

MSC *Mobile services Switching Centre*

NSS *Network Sub-System*

OMC *Operation and Maintenance Center*

OSS *Operational Support System*

P-GSM *Primary GSM*

PAMR *Public Access Mobile Radio*

PK Ponto Quilométrico

PMR *Private Mobile Radio*

RE Coeficiente de correlação

REFER Rede Ferroviária Nacional

REFER-Telecom Serviços de Telecomunicações S.A

RMSE *Root Mean Square Error*

SIM *Subscriber Identity Module*

SMOTE *Synthetic Minority Oversampling Technique*

SVM *Support Vector Machine*

UIC *Union Internationale du Chemin-de-Fer*

TP Verdadeiros Positivos

TRX *Transceivers*

VLR *Visitor Location Register*

1

Introdução

Conteúdo

1.1 Enquadramento

O *Global System for Mobile Communications - Railway* (GSM-R) é um sistema de comunicações móveis caracterizado para a rede ferroviária que emergiu da necessidade da criação de um sistema de comunicações digitais sem-fios que cumprisse o objetivo de uniformização tecnológica em toda a rede ferroviária na Europa [4].

Existe uma forte necessidade, nos meios de transporte ferroviários, em assegurar comunicações entre comboios e o GSM-R vem substituir os sistemas antigos de forma a garantir uniformidade entre todas as linhas, sendo este baseado no sistema *Global System for Mobile Communications* (GSM), dada a sua robustez e fiabilidade ao nível de transmissão rádio [5]. Na Europa, o sistema opera na banda dos 876 até 880 MHz (*uplink*) e na banda dos 921 até 925 MHz (*downlink*).

Na instalação de sistemas desta dimensão é necessário que as várias estações base forneçam a maior cobertura possível de forma a minimizar o número de estações base, diminuindo assim o custo de instalação global do sistema e a interferência causada entre as várias estações [1].

A operação do GSM-R para as comunicações ferroviárias ocorre da mesma forma que as redes GSM comuns sendo os comboios, neste caso, os utilizadores em movimento dispondo de antenas no telhado que, por sua vez, comunicam com os mastros das estações base colocados perto da linha ferroviária. Estes são equipados com antenas altamente direcionais por forma a cobrir apenas áreas da faixa ferroviária, reduzindo a interferência de sinal.

Não é fácil o cálculo preciso da atenuação de sinal em cenários com obstáculos, diferentes ambientes, terrenos irregulares entre outros, uma vez que é necessário ter em conta um elevado número de parâmetros. De maneira a resolver este problema utilizam-se modelos de propagação que têm em conta os mecanismos de propagação de sinal em espaço livre com a presença de obstáculos e vários fatores corretivos obtidos através de análises estatísticas em diferentes cenários.

1.1.1 Evolução breve do GSM-R

A entidade *Union Internationale du Chemin-de-Fer* (UIC), em 1992, desenvolveu um projeto europeu denominado por *European Integrated Railway Enhanced Network* (EIRENE) do qual resultaram um conjunto de especificações para a implementação da tecnologia GSM-R, de modo a alcançar o objetivo da uniformização tecnológica em toda a Europa [6].

A equipa responsável por desenvolver o EIRENE, trabalhou em conjunto com o *European Telecommunications Standards Institute* (ETSI) e uma parceria de operadores, fornecedores, fabricantes e vendedores, denominada por *Mobile Radio for Railways Networks in Europe* (MORANE) [4]. Esta parceria trabalhou com o intuito de desenvolver um sistema digital, fiável e flexível, que cumprisse os requisitos impostos pelos intervenientes no setor ferroviário. Contudo, era necessário fazer convergir

num único sistema os diversos esforços de desenvolvimento de sistemas de comunicações sem-fios digitais nos diferentes países [1].

Considerando as particularidades de utilização e tecnológica de cada operador/país, a convergência entre os diferentes sistemas existentes, de geração analógica, torna-se complicada. Sendo assim, tendo em conta a sua robustez e confiabilidade na transmissão de rádio, a escolha dessa tecnologia incidiu sobre a tecnologia GSM, que foi comprovada em centenas de países e inúmeros utilizadores, o que permitiu à tecnologia GSM atingir um enorme sucesso em termos de padrões de comunicação sem-fios [4].

Todavia, tendo em conta as especificidades e os requisitos característicos do uso em caminhos-de-ferro, foi essencial efetuar alterações face à norma. Portanto, com base no padrão GSM *Phase 2+*, foi criado o padrão GSM-R específico para uso nas ferrovias [4]. Os primeiros produtos protótipos desse padrão foram produzidos em 1997. Efetivamente, observa-se que nos últimos anos tem-se averiguado um aumento muito considerável de países com o GSM-R instalado, com ênfase a partir do ano de 2002.

Assim como os subscritores móveis, as operadoras ferroviárias também necessitam de maior capacidade de processamento de dados para conseguir assegurar o bom funcionamento das comunicações móveis ferroviárias. Por conseguinte, embora o GSM-R deva permanecer operacional pelo menos até 2030 [7], está gradualmente a ser substituído por uma versão semelhante das redes *Long-Term Evolution* (LTE) (rede móvel 4G): *Long-Term Evoluiton - Railways* (LTE-R). Esta maior capacidade de processamento face às comunicações móveis ferroviárias gera uma necessidade de alimentar os sistemas de comunicação para novas redes ferroviárias de alta velocidade, bem como oferecer melhores serviços aos passageiros e às operações nas ferrovias, desta forma impulsionando o desenvolvimento da tecnologia. Alguns países já implementaram redes LTE-R pioneiras, como a China e a Coreia [8], mas a arquitetura ainda não foi padronizada pelo ETSI. Grupos da indústria, organizações de normas e operadoras têm feito parceria nos últimos anos para chegar a acordos sobre as especificações do LTE, quando a UIC lançou oficialmente o projeto *Future Railway Mobile Communications System* (FRMCS) em 2013 [9].

1.1.2 O GSM-R em Portugal

A Serviços de Telecomunicações S.A (REFER-Telecom) foi fundada no ano 2000, focando-se atualmente no suporte e operação das comunicações da rede ferroviária nacional e é responsável pelo planeamento, instalação e manutenção do sistema GSM-R em Portugal.

Em 2008, a Rede Ferroviária Nacional (REFER) delegou na REFER-Telecom o seguimento dos estudos, projetos e a obtenção de licenciamento junto da Autoridade Nacional para as Comunicações (ANACOM), tendo em vista a implementação de uma rede de comunicações rádio GSM-R a instalar nas principais linhas da rede ferroviária convencional e também das futuras linhas de alta velocidade.

No decurso da implementação do projeto-piloto de instalação do novo sistema de comunicações ferroviárias GSM-R na Linha de Cascais, a REFER-Telecom, realizou a primeira chamada sobre esse tipo de rede, existindo um forte contributo científico nesta área, nomeadamente nos métodos utilizados para a estimação de cobertura rádio [2].

1.2 Motivação e Objetivos

Hoje em dia presencia-se um grande investimento na modernização das linhas ferroviárias nacionais. Deste modo, é necessário efetuar a estimativa de cobertura de rádio em todas as linhas, envolvendo diversos tipos de ambientes de propagação. Em comunicações móveis ferroviárias essa predição requer maior precisão do que nas redes públicas, dadas as limitações decorrentes dos requisitos de segurança. É, então, necessário a configuração dos modelos de propagação utilizados para os diversos ambientes de propagação e características, existentes nos caminhos-de-ferro. Contudo, este ajuste de parâmetros de um dado modelo é um processo que pode tornar-se difícil, devido ao número de variáveis envolvidas e a dependência entre elas. Torna-se necessário o recurso a técnicas de *Machine Learning* (ML), com recurso à interface *H2O Flow*, que a partir de um conjunto de dados de teste, produz uma solução de parâmetros que minimizam o erro entre a predição e as medidas reais da rede.

Assim, tendo em conta trabalhos desenvolvidos anteriormente nesta área [10] [5], os objetivos desta dissertação serão:

- Estudar os melhores modelos de propagação para estimação e cobertura rádio em comunicações móveis ferroviárias, selecionando-se os modelos que melhor se adaptam aos tipos de ambientes destes sistemas;
- Quantificar o desempenho de cada modelo vs ambiente de propagação, tendo em conta o desvio entre medidas efetuadas e predição de cobertura rádio;
- Aplicar uma metodologia de classificação, com o auxílio da interface *H2O Flow*, permitindo desenvolver um algoritmo em que, para cada ponto, efetua a classificação do ambiente e, consequentemente, a seleção do modelo mais adequado à estimação de cobertura rádio.

1.3 Estrutura

Este relatório divide-se em seis capítulos. No presente capítulo, é fornecido o enquadramento tecnológico deste projeto, assim como o que motivou a sua realização, o seu objetivo e a sua estrutura.

No segundo capítulo são descritos os conceitos fundamentais para a compreensão da área científica considerados para caso de estudo, começando por introduzir todos os elementos que são apontados na propagação em linhas ferroviárias. São apresentadas as características do planeamento do sistema GSM-R, bem como, todo o processo de recolha de medidas nas várias linhas ferroviárias consideradas neste trabalho.

O terceiro capítulo começa por introduzir o modelo de propagação utilizado, explicando os vários parâmetros e fatores corretivos na estimação de cobertura rádio. Além disso, descreve todo o processo de implementação do modelo de propagação ao longo deste trabalho, tendo em conta as medidas rádio recolhidas, os valores estatísticos do modelo e a classificação de cada modelo.

No quarto capítulo é descrita a interface utilizada para a classificação de cada modelo. É feita a descrição do *H2O Flow* e também de todo o processo de dimensionamento necessário para a configuração correta dos algoritmos utilizados.

O quinto capítulo engloba todos os resultados obtidos neste trabalho. Começa por definir o modelo de propagação, com base em alguns conjuntos de testes. São também descritos os testes, realizados em 3 cenários diferentes, que serviram de validação do algoritmo, assim como, um teste final com vista à uniformização do modelo utilizado.

Por fim, no sexto e, último capítulo, são apresentadas as conclusões resultantes do trabalho desenvolvido, bem como, propostas para um trabalho futuro.

2

Conceitos Fundamentais

Conteúdo

Este capítulo fornece uma visão geral dos conceitos fundamentais necessários para entender os sistemas de comunicação ferroviária, ou seja, a arquitetura do sistema, as especificações técnicas das tecnologias GSM-R e os principais parâmetros que afetam o desempenho. Além disso, é retratado como se procede à aquisição dos dados necessários para a resolução do problema, especificamente, a posição geográfica, a configuração da rede e estimação de sinal, e as medidas ao longo da linha ferroviária em foco. No final do capítulo, apresenta-se o estado da arte do trabalho, detalhando as abordagens adotadas e os aspetos inovadores do método utilizado, observando as conclusões retiradas.

2.1 Rede GSM-R

A tecnologia de comunicação ferroviária GSM-R é uma extensão do padrão GSM, atendendo às exigências de comunicação ferroviária, quanto a funcionalidades e robustez. A tecnologia GSM-R foi adotada pela maioria das ferrovias para suportar comunicações operacionais de voz e dados. Dada a maturidade dos atuais sistemas GSM, as metodologias de planeamento e otimização da rede de rádio estão atualmente bem definidas e documentadas e são utilizadas por todos os operadores de redes móveis públicas.

Em termos da rede de rádio as principais divergências, relativamente ao padrão GSM, ocorrem devido ao sistema GSM-R suportar velocidades de até aos 500 km/h, suportando *handovers* e seleção/reseleção de células mais rápidos do que no padrão GSM público. Além disso, ao nível funcional e aplicacional, novas funções foram consideradas de forma a suportar a utilização mais flexível e aplicada às comunicações ferroviárias, tais como o controlo automático dos comboios e as chamadas de emergência, por exemplo [4].

Portanto, existem várias vantagens em utilizar o sistema GSM-R, conforme descrito abaixo [4].

- **Confiabilidade e disponibilidade do GSM-R**
  - Altíssima confiabilidade e disponibilidade de produtos e redes baseadas na experiência de longo prazo provado em sistemas públicos e redes ferroviárias;
  - Redundância superior interna e soluções geográficas redundantes completas;
  - Funcionalidade de cobertura dupla especial, de muito alta confiabilidade para *European Train Control System* (ETCS)[1];
  - Sistema de gestão da rede incluindo falhas em tempo real e gestão de acidentes;
  - Engenharia dedicada e processos de otimização do sistema.

[1] **ETCS**, Sistema Europeu de Controlo de Comboios, é um sistema de controlo da velocidade para substituir os atuais sistemas de controlo de velocidade de cada país.

Efetivamente, o GSM-R é um sistema sem risco operacional, uma vez aprovado e em operação em alta velocidade e em linhas convencionais há mais de 10 anos, assim como sem risco do produto. Garante também a interoperabilidade do sistema, comunicações sem fronteiras do sistema ferroviário, e a eficiência do sistema, uma vez que a comunicação ferroviária é feita a custo reduzido.

2.1.1 Arquitetura de rede

O GSM-R é padronizado pelo ETSI [11] e visa a diminuição da complexidade das respetivas estações base de transmissão. A gestão e manutenção centralizada da rede, assim como, a interligação a outras redes, são conceitos fundamentais desta rede. A arquitetura GSM-R é dividida em três componentes principais, como pode ser visto na Figura 2.1:

- ***Base Station Sub-System* (BSS)**, secção responsável pelo tráfego e sinalização entre terminais móveis e a rede;
- ***Network Sub-System* (NSS)**, componente do sistema GSM-R que manipula o gerenciamento de mobilidade, rastreando a localização móvel e permitindo que os serviços móveis sejam fornecidos aos utilizadores;
- ***Operation and Maintenance Center* (OMC)**, conectado a todos os equipamentos no sistema de comutação e ao *Base Station Controller* (BSC). A implementação do OMC é chamada de *Operational Support System* (OSS), dedicado a fornecedores de serviços de telecomunicações e é utilizado principalmente para suportar processos de rede para manter inventário de rede, configurar componentes de rede, provisionar serviços e gerenciar falhas.

A Figura 2.1 ilustra as principais componentes da arquitetura de uma rede GSM-R, respeitando a norma [12].

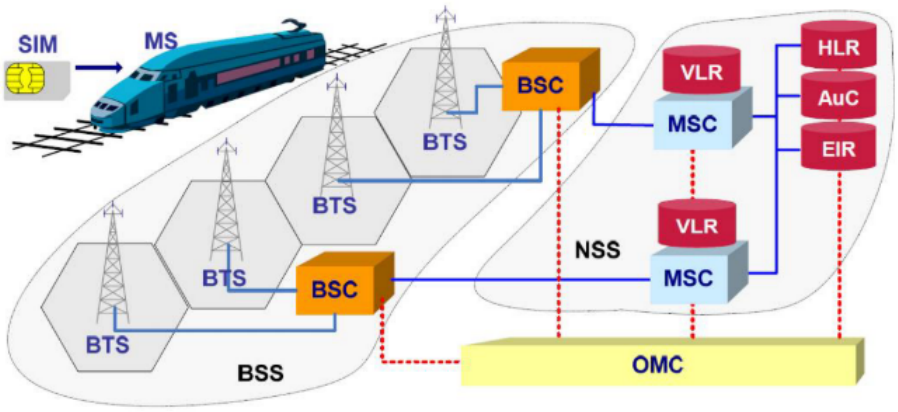


Figura 2.1: Arquitetura de rede GSM-R [1].

Visualizando a figura 2.1, é possível observar que as antenas montadas são as *Base Transceiver Station* (BTS), que contêm o equipamento para transmitir e receber sinais rádio. Para as células setorizadas, a *Base Station* (BS) pode ter vários transceptores, permitindo que ela sirva diferentes frequências/setores. Contudo o GSM-R geralmente tem uma única célula para a área de serviço, portanto a BTS possui vários transceptores trabalhando em frequências diferentes para maximizar a capacidade.

No lado do utilizador do sistema, encontra-se a *Mobile Station* (MS), que são os receptores móveis instalados nos comboios, mas também outros terminais que dependem da rede GSM-R, por exemplo, trabalhadores ferroviários para manutenção ou segurança e serviços de emergência. Este equipamento inclui ainda um cartão inteligente, *Subscriber Identity Module* (SIM), o qual contém informação específica de cada utilizador.

O BSS é constituído por BSCs que controlam as BTSs, contendo cada uma, um número de *Transceivers* (TRX). Por outro lado, o NSS contém os *Mobile services Switching Centre* (MSC), os quais se encontram interligados a um *Visitor Location Register* (VLR). Os VLR são equipamentos que possuem bases de dados com a informação temporária de um determinado utilizador e estão ligados a uma área de serviço assegurada pelo MSC. A gestão dos perfis dos utilizadores ligados à rede é realizada por um conjunto de bases de dados, designadas por *Home Location Register* (HLR), por outro lado, as bases de dados intituladas de *Authentication Center* (AuC) e *Equipment Identity Register* (EIR) são responsáveis pela gestão do mecanismo de segurança e dos equipamentos terminais, respetivamente. Além disso, o NSS liga ao BSS através da interface A do GSM, interface aberta com canais de 64 kbps, uma vez que esta arquitetura é tradicionalmente idêntica à do GSM público [1].

Os utilizadores do terminal móvel podem ser alcançados por meio de numeração funcional e *Location Dependent Addressing* (LDA), sendo este um método de endereçamento que requer informações sobre a localização da estação móvel de modo a atribuir um endereço lógico a uma dada função (controlador) [13].

2.1.2 Espetro da rede GSM-R

A *radio interface* para GSM-R consiste no fluxo de informações que ocorre entre o receptor do comboio (MS) e a estação base (BTS). Em termos de espetro, o ETSI [14] reservou duas bandas de frequência entre 876-880 MHz (ligação ascendente) e 921-925 MHz (ligação descendente), utilizados pelos sistemas EIRENE. Esta banda é denominada de banda GSM-R ou banda UIC. A UIC também definiu a possibilidade de ter uma banda adicional de 200 kHz, banda de guarda, para aumentar a disponibilidade do canal e a banda *Extended GSM* (E-GSM), baseando-se em frequências geralmente reservadas para uso governamental e de defesa. Portanto, uma vez que a banda GSM-R seja insuficiente para as necessidades de comunicação ferroviária numa determinada área, o regulador pode

permitir o uso de frequências extras, desde que disponíveis. O espetro total na banda dos 900 MHz, incluindo a faixa GSM-R pode ser vista na Figura 2.2.

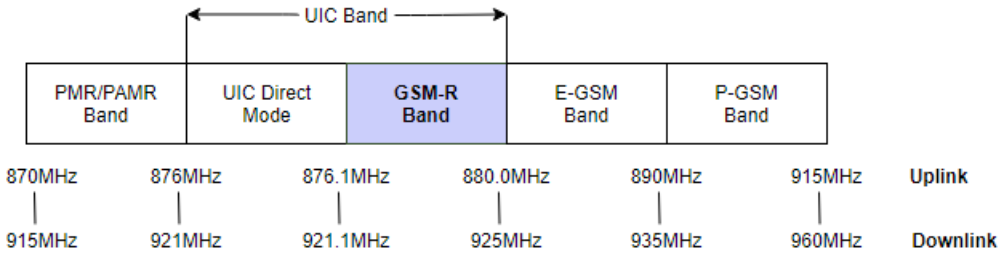


Figura 2.2: Alocação de frequência para GSM-R, na faixa de 900 MHz [1].

Equivalente ao GSM público, cada portadora ocupa um espetro de 200 kHz de banda, o que permite 19 pares de portadoras na banda GSM-R. No extremo inferior da faixa GSM-R foram reservados 100 kHz para o Modo Direto, que permite a comunicação direta entre MS na rede para curtas distâncias [1].

2.1.3 Cobertura GSM-R

Distinguindo a rede GSM da GSM-R, é obrigatório discutir os níveis mínimos de cobertura. No planeamento da rede, o nível de cobertura é definido como a intensidade do campo recebido na antena recetora, localizada no telhado do comboio. A especificação GSM-R define mínimos de cobertura dependendo da velocidade e do tipo de informação recebida. Estes valores estão apresentados na Tabela 2.1 e são definidos apenas para o nível recebido no rádio da locomotiva (*Cab Radio*), considerando uma antena isotrópica a 4 metros de altura [15].

Tabela 2.1: Níveis mínimos de cobertura.

Tipo	Valor Mínimo	Utilização	Velocidade do Comboio
Obrigatório	-98 dBm	Voz e dados de baixa segurança	—
Obrigatório	-95 dBm	ETCS níveis 2/3	$\leq$ 220 km/h
Recomendado	-92 dBm	ETCS níveis 2/3	$\geq$ 280 km/h

Na rede GSM-R, os valores mínimos de cobertura devem obedecer a uma probabilidade de cobertura de pelo menos 95%, a cada 100 m de linha férrea, como pode ser observado na Figura 2.3.

Isto introduz uma diferença significativa em comparação com o sistema GSM, onde o nível de probabilidade de cobertura trata-se da média da cobertura de toda a região. Deste modo, é possível verificar que os requisitos de cobertura para os sistemas GSM-R têm um nível de exigência muito superior ao do GSM.

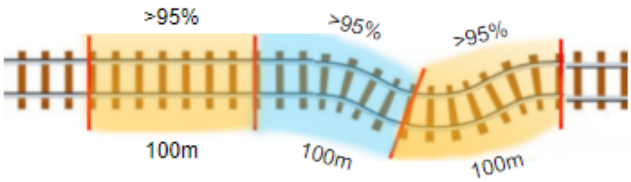


Figura 2.3: Probabilidade de cobertura em GSM-R [2].

É também de salientar que o LDA, já referido anteriormente, é o nome de uma das funcionalidades a ter em atenção no que diz respeito ao dimensionamento de cobertura rádio em GSM-R. Assim sendo, este requer que a rede GSM-R execute uma localização precisa do comboio em relação às fronteiras dos setores de controle.

Este requisito terá implicações no projeto de redes de rádio GSM-R, devido à necessidade de precisão na localização de algumas fronteiras de células, que devem coincidir com as fronteiras do setor de controle. Isto pode ser conseguido através de uma estação base de dois setores no local de comutação desejado, já que esta é a melhor maneira e mais segura de garantir precisão no controle de *handovers*, como se pode observar na Figura 2.4 [2].

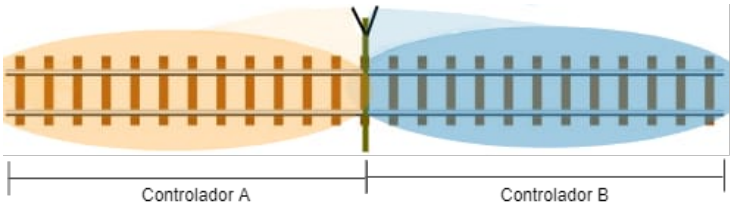


Figura 2.4: Exemplo de aplicação LDA.

A alteração de controlador de forma sucinta, em determinadas regiões da ferrovia, torna-se por isso necessária, de modo a prevenir determinadas limitações no planeamento de cobertura rádio. Este aspeto consiste em forçar o *handover* na área de comutação de endereço, o que implica a colocação de uma estação bi-setorizada na zona de comutação pretendida de forma a assegurar uma maior precisão de *handover* no local. Esta funcionalidade é realizada através de parâmetros específicos do sistema GSM [6].

2.2 Propagação em linhas ferroviárias

A predição de cobertura rádio é um dos passos fundamentais no processo de planeamento da rede rádio GSM-R e tem por base todo o processo de cálculo de ligação, planeamento de cobertura, planeamento de frequências e capacidade, e análise de interferência. Desta forma torna-se muito

importante a correção da predição e a maneira como esta deverá estar adequada ao ambiente em causa. A predição de cobertura rádio tem por base modelos de estimação de sinal, ou modelos de propagação. É com base em modelos de propagação que se efectua a predição de cobertura, apesar de apresentarem resultados que não se exprimem na máxima eficiência da rede projetada [2].

Um dos aspetos fundamentais no dimensionamento de uma rede sem fios centra-se no cálculo da atenuação na propagação de sinal rádio. De modo a diminuir o número de estações base na instalação de sistemas desta dimensão é essencial que as estações base providenciem a maior cobertura possível, fazendo com que se minimize o custo de instalação total do sistema, assim como, a interferência causada entre as várias estações. É improvável determinar, de forma concreta, a atenuação do sinal em cenários reais com obstáculos, diferentes meios, terrenos irregulares, entre outros fatores, devido ao elevado número de parâmetros a considerar. De forma a resolver este problema utilizam-se modelos de propagação que têm em conta os mecanismos de propagação de sinal em espaço livre e na presença de obstáculos, bem como vários fatores corretivos obtidos através de análises estatísticas em diferentes cenários.

2.2.1 Modelos de Propagação

Um modelo de propagação é, então, uma descrição matemática baseado na teoria da propagação eletromagnética e/ou em medidas experimentais que pretende descrever a influência dos fatores que condicionam a propagação nas características do sinal recebido.

Sendo assim, o planeamento das áreas de cobertura das estações base requer a estimação do sinal de modo a conhecer-se o contorno de cada célula. É essencial prever as zonas limites onde o nível de sinal é mínimo e as zonas onde pode haver interferência. A previsão do sinal é feita através de modelos que calculam o valor médio do sinal e a variação em torno desse valor médio.

Os modelos são essencialmente uma mistura de empirismo e aplicação da teoria eletromagnética da propagação.

Por um lado, modelos empíricos são formulados quando se obtêm curvas ou expressões analíticas baseadas em dados medidos, sendo ajustados estatisticamente para os dados recolhidos num dado estudo em causa. Esse tipo de modelo é simples e fácil de usar, alcançando previsões de valores sem recorrer a um grande esforço de processamento computacional, pois não levam em consideração a teoria da propagação de ondas eletromagnéticas. Como esse modelo é baseado em medições, é naturalmente um modelo que apresenta desvios significativos entre a previsão de sinal e as medidas reais que podem afetar negativamente o planeamento da rede de rádio. Além disso, tem em conta todos os fenómenos, tanto conhecidos como os desconhecidos, no entanto deverá ser rigorosamente testado para localizações e frequências diferentes das que serviram para produzir o modelo.

Por outro lado, equações teóricas têm em conta as perdas em espaço livre e as perdas devido ao

plano terra, servem muitas vezes como alicerces a modelos que se recorrem de fatores empíricos para levarem em conta as perdas por difracção, a curvatura da Terra, efeitos atmosféricos e as perdas devido aos edifícios e vegetação.

Posto isto, não existe um modelo de aplicação genérico para todos os tipos de ambientes, frequências e parâmetros. A aplicação de modelos com uma componente empírica requer classificação dos ambientes de propagação. A propagação em zonas com edifícios é fortemente influenciada pelo ambiente envolvente, em especial pelo tamanho e densidade dos edifícios. Como vamos analisar, no próximo capítulo, irá ser utilizado o modelo Okumura-Hata em que os ambientes neste modelos são classificados em três grupos, como se pode observar na Figura 2.5: área urbana, suburbana e rural.

(a) Área urbana **(b)** Área suburbana **(c)** Área rural

Figura 2.5: Ambientes de propagação do modelo Okumura-Hata.

É de salientar que, uma uma classificação imprecisa pode levar a diferentes interpretações por diferentes projetistas, sendo fundamental descrever o ambiente qualitativamente para evitar ambiguidades. O ambiente pode ser observado como composto por um conjunto de diferentes *scattered classes*, independentes umas das outras.

2.3 Predição de Sinal

Uma das principais etapas no planeamento de uma rede rádio é a estimação de cobertura rádio, como já foi mencionado anteriormente. Deste modo, é necessário estimar o nível de sinal de uma rede GSM-R num determinado ponto da linha ferroviária, só sendo possível tendo em conta informações sobre o terreno existente entre a estação base e a estação móvel.

No âmbito da colaboração entre a REFER e o Instituto Superior de Engenharia de Lisboa (ISEL), foi desenvolvida uma aplicação RailWave que tem como propósito a estimação de cobertura rádio em sistemas de comunicações móveis ferroviárias, utilizando o modelo de propagação Okumura-Hata. Esta ferramenta de predição possibilita também, que vários parâmetros especificados por este modelo e os fatores de correção correspondentes, sejam gerados a partir da análise das características do cenário

em estudo. Estas estimativas são cruciais para o planeamento, uma vez que caracterizam um estudo teórico do desempenho do sinal por toda a extensão da linha ferroviária, viabilizando desta forma o planeamento das estações base no decorrer da linha.

Frequentemente, ao tratar-se com ferrovias é vulgar utilizar, para distâncias, a unidade Ponto Quilométrico (PK). Esta não corresponde ao comprimento da linha, e não existe nenhum método numérico para associá-lo à localização geográfica, sendo portanto necessário um ficheiro com essa informação.

Além disso, para determinar a superfície e as características que influenciam a propagação do sinal recorreu-se ao RailWave para extrair a informação geográfica precisa. Para cada BTS, as seguintes informações foram adquiridas para cada ponto da linha ferroviária:

- Distância entre a estação base e o ponto da linha ferroviária;
- Altura efetiva (pela diferença de alturas da base da estação base e do ponto onde se encontra a antena do terminal móvel);
- Distância efetiva atravessada sobre a vegetação;
- Distância percorrida sobre a água (que permite determinar o parâmetro β);
- Altura da ondulação do terreno;
- Altura média da ondulação do terreno;
- Parâmetro adimensional v dos 3 obstáculos principais.

Além destas informações foram também inseridas: a frequência, a altura do móvel e a altura da estação base, de cada estação base. Todas estas informações irão ser analisadas, ao detalhe, no capítulo seguinte.

2.4 Cenários de estudo e medidas rádio

No âmbito deste trabalho, são consideradas três linhas ferroviárias: a linha de Cascais, a linha da Beira Baixa e a linha do Algarve. As três linhas apresentam características diferentes entre si e por isso é esperado que um sinal rádio se comporte de forma diferente em cada uma delas. Isso motivou um estudo de cada uma das linhas e a análise da propagação do sinal em cada um destes cenários. Para o GSM-R, o trabalho nesta área é facilitado, dada a disponibilidade de medidas de cobertura rádio realizadas pela REFER-Telecom nestas linhas, o que permitirá validar os resultados obtidos.

Ao escolher o local (*sites*) para a instalação do equipamento base, considerou-se o local da instalação da REFER, bem como a disponibilidade da localização para esse fim.

2.4.1 Linha de Cascais

A linha de Cascais foi a primeira linha ferroviária com cobertura móvel de uma rede GSM-R. A linha tem uma extensão de 26 km e é caracterizada por um ambiente suburbano e apresenta alguma influência da água, uma vez que a cobertura rádio em determinados troços da linha é realizada por cruzamento com a superfície aquática. Esta proximidade da linha de água tem como impacto uma baixa altitude da linha, não existindo nenhum obstáculo considerável ao longo do traçado da linha.

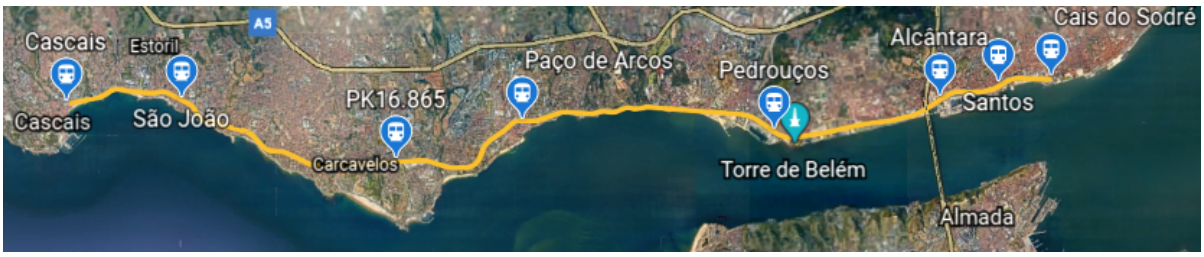


Figura 2.6: Linha ferroviária de Cascais.

Na Figura 2.6, é possível visualizar a linha ferroviária de Cascais na sua totalidade, sendo esta composta por cinco estações base e um repetidor e tem atribuída quatro frequências. Cada estação base inclui duas antenas direcionais orientadas para cada uma das direções da linha.

A tabela 2.2 apresenta as diversas frequências referentes às cinco estações base e ao repetidor da linha de Cascais.

Tabela 2.2: Estações base na linha de Cascais e suas respetivas frequências.

Estação Base	Frequência [MHz]	BSIC
AlcântaraA	921,2	30
AlcântaraB	921,2	30
SantosA (repetidor)	921,2	30
PedrouçosA	922,0	30
PedrouçosB	922,0	30
PaçoArcosA	922,8	30
PaçoArcosB	922,8	30
PK16.865A	923,6	30
PK16.865B	921,2	31
SãoJoãoA	922,0	31
SãoJoãoB	922,0	31

Note-se que cada estação base possui duas antenas que utilizam a mesma frequência, como por exemplo na estação base de Alcântara, ou em frequências diferentes, no caso da estação base de PK16.865. Para identificar uma estação base é muitas vezes utilizado um código usado no GSM,

Base Station Identify Code (BSIC). O código é necessário, pois é possível que as estações base móveis recebam sinal de mais uma estação base na mesma frequência, como acontece com os *sites* de Pedrouços e São João.

2.4.2 Linha da Beira Baixa

A linha da Beira Baixa corresponde à linha que faz a ligação entre a cidade do Entroncamento e o distrito de Guarda, tendo uma extensão total de 270 km. Contudo, considerou-se apenas o troço até Rodão, uma vez que só se efetuou campanha de medidas até este local. O troço tem uma extensão de 63 km, caracterizado por um ambiente rural com terreno montanhoso e com alguns troços da linha caracterizados por um ambiente suburbano. Salienta-se também que o troço encontra-se junto do Rio Tejo, pelo que é necessário ter em conta a influência de água. Na Figura 2.7, é possível visualizar o troço considerado da linha ferroviária da Beira Baixa.

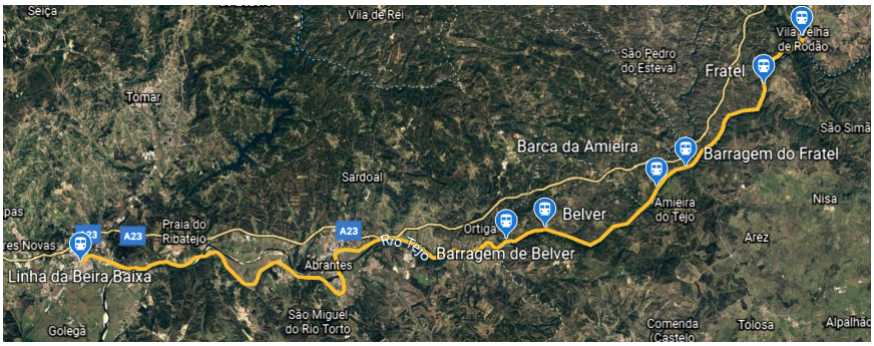


Figura 2.7: Linha ferroviária da Beira Baixa.

Durante a campanha de medidas foram feitas quatro viagens, representadas na Figura 2.8, percorrendo todo o troço da linha. Em cada viagem colocaram-se duas estações base de teste ao longo do troço em frequências diferentes e foi medido o nível de sinal recebido de cada uma delas em cada ponto da linha.

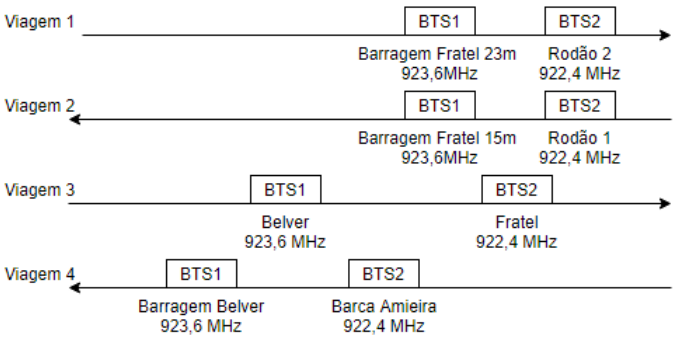


Figura 2.8: Viagens de teste na linha da Beira Baixa.

2.4.3 Linha do Algarve

A linha do Algarve corresponde à linha ferroviária que une a cidade de Lagos a Vila Real de Santo António, tendo uma extensão total de 139,5 km. No entanto, considerou-se apenas o troço representado na Figura 2.9 onde foram retiradas medidas de teste. O troço da linha tem uma extensão de 55 km e é caracterizado por um ambiente suburbano, pela presença de água em alguns pontos da linha e pela existência de muitos obstáculos entre as estações base e os vários pontos da linha.

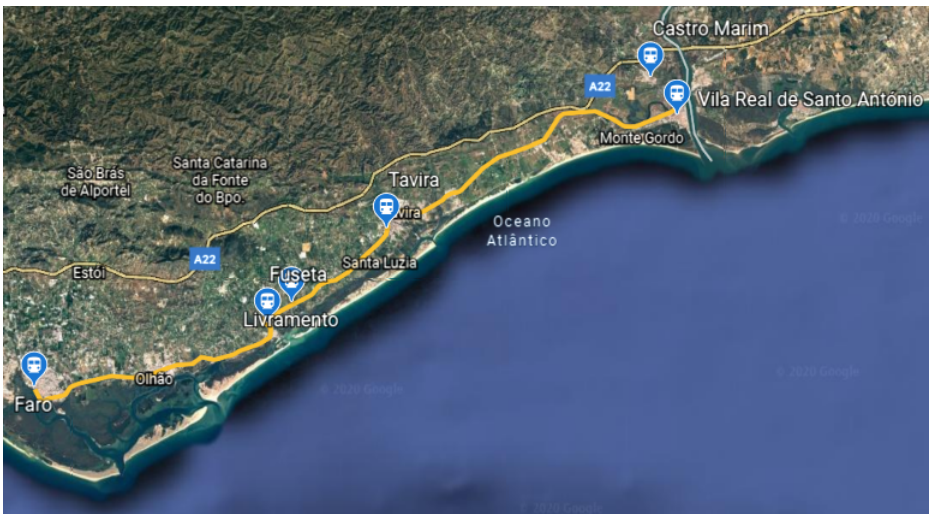


Figura 2.9: Linha ferroviária do Algarve.

Tal como na linha da Beira Baixa, durante a campanha de medidas foram realizadas quatro viagens de teste, colocando estações base em frequências diferentes ao longo da linha e medido o nível de sinal recebido em todos os pontos da linha. Na Figura 2.10 estão ilustradas as quatros viagens realizadas colocando estações base em sete locais distintos ao longo do troço.

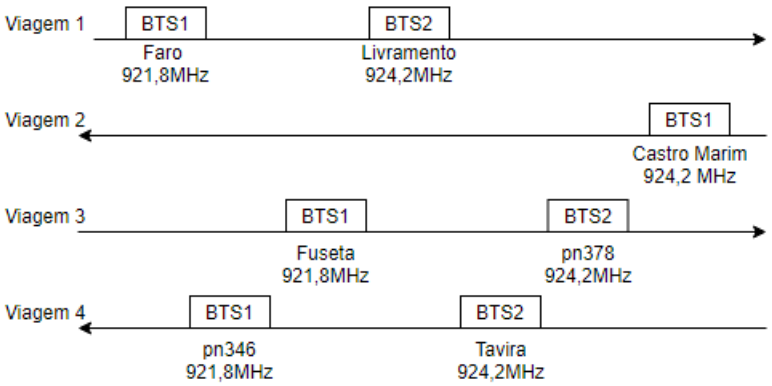


Figura 2.10: Viagens de teste na linha do Algarve.

2.5 Estado de Arte

No trabalho realizado em [5] foram igualmente utilizadas redes neuronais para conseguir obter um modelo que permita uma estimação com maior precisão. Apesar dos bons resultados, verifica-se que neste caso, consequência da própria técnica associada às redes neuronais, o desenvolvimento de um modelo dependerá sempre de medidas previamente realizadas e da significância e validade da amostra de medidas face ao universo de ambientes que existem. Essa dependência permite verificar que a utilização é uma base para o algoritmo de predição tendo em conta um modelo matemático, apesar de determinístico, permite uma melhor acomodação de variações dos tipos de ambientes. Como vantagens, os resultados obtidos concluíram, um reduzido erro de predição para os ambientes e situações cobertas pelas medidas, a validação da utilização das redes neuronais para predição de cobertura rádio e, ao realizar uma classificação prévia do ambiente, uma redução significativa do erro, permitindo realizar uma aprendizagem competitiva. Por outro lado, este modelo indicou ficar impossibilitado com a existência de medidas que representam o universo de aplicabilidade do modelo, bem como com erros elevados em ambientes que não tenham sido cobertos pelas medidas utilizadas no treino da rede.

No trabalho publicado em [10] e no artigo [16] verificou-se que existe a possibilidade de utilização de algoritmos genéticos para otimização automática do modelo Okumura-Hata aos diversos tipos de linhas ferroviárias, tendo em conta os respetivos parâmetros de ambiente. Como vantagens da metodologia proposta salientam-se a redução significativa do erro de predição para o conjunto de ambientes estudados e a validação do modelo utilizado para a estimação da cobertura rádio em ambientes ferroviários. Contudo, ficou claro que esta otimização ao ser efetuada para toda a extensão da linha teria o problema de que, tendo em conta a heterogeneidade de ambiente ao longo de algumas linhas ferroviárias, torna impossível a otimização de um único modelo para todos os diferentes tipos de ambientes.

3

Implementação

Conteúdo

Neste capítulo é apresentado o modelo de propagação proposto e implementado para a estimação de cobertura de rádio na ferrovia, tendo em conta a classificação do ambiente, o qual se divide em três categorias: urbana, suburbana e rural. Esta classificação tem em consideração vários parâmetros, tais como, a ondulação do terreno, a influência da vegetação, a altura e densidade dos edifícios, assim como a densidade de áreas abertas e de água. Além disso, é retratado o método de Deygout, visto que o modelo Okumura-Hata não avalia as perdas por difração oriundo dos obstáculos.

3.1 Modelo de Propagação Okumura-Hata

O modelo de Okumura foi proposto em 1968, baseado em medidas na banda [150,2000] MHz, para distâncias de 1 a 100 km e altura efetiva da antena da BTS de 30 a 1000 m e apresentou os resultados em forma de curvas. Posteriormente, Masaharu Hata em 1980 estabeleceu expressões estabelecidas numa banda mais restrita, que aproximam algumas dessas curvas. Como tal, este é um modelo empírico desenvolvido por Okumura e Hata, sendo baseado em resultados de extensas medições no Japão, fornecendo uma estimativa da perda do sinal em certos ambientes de propagação. Este modelo é útil para se realizar uma estimativa da atenuação de percurso do sinal em diversos ambientes, sendo recomendado para a predição em GSM-R [17]. Além disso, tem conduzido a resultados bastante plausíveis, obtendo erros reduzidos, quando complementado com fatores corretivos próprios ao tipo de ambiente [3], sendo este o modelo mais utilizado na estimação de cobertura rádio durante a fase de planeamento de uma rede móvel.

O valor concluído deste modelo padrão é um ambiente urbano, em terreno plano, sobre o qual são considerados fatores de correção. Os ambientes, neste modelo, são classificados em três grupos [18]:

- **Área urbana:** região de alta densidade de edifícios, sendo estes superiores a 2 pisos;
- **Área suburbana:** alguns obstáculos, com pouca densidade, diante do terminal móvel;
- **Área aberta:** sem obstáculos numa região de 300 a 400 m, diante do terminal móvel.

Para manter as expressões simples, Hata e Okumura assumiram que os emissores estariam localizados em um local elevado. A validade deste modelo encontra-se definida pela Tabela 3.1.

Tabela 3.1: Condições de aplicação do modelo de Okumura-Hata [1].

Parâmetros	Símbolo	Valores Limites
Frequência	f [MHz]	[150,1500]
Distância	d [km]	[1,20]
Altura efetiva da antena de emissão	h_be [m]	[30,200]
Altura da antena móvel	h_m [m]	[1,10]

3.1.1 Altura efetiva

Os resultados alcançados pela aplicação do modelo na predição de cobertura rádio são influenciados pelo método utilizado para a determinação da altura efetiva da antena da estação base. Assim, a altura efetiva da antena da estação base corresponde à diferença das alturas da base da estação base e do ponto onde se encontra a antena do terminal móvel. O cálculo da altura efetiva, baseado no modelo *International Telecommunication Union - Radiocommunication sector* (ITU-R) [19], produz resultados mais aproximados do valor real, visto que este parâmetro tem em conta não só informação sobre a diferença das alturas do emissor e recetor, mas também informação do terreno entre estes dois pontos [3].

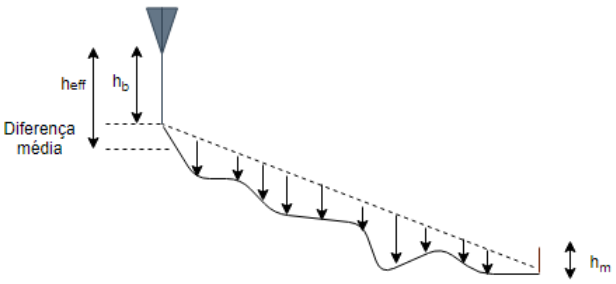


Figura 3.1: Altura efetiva (ITU-R).

Através da Figura 3.1 é possível observar a projeção de uma reta entre a base da BS e o local onde se situaliza a estação móvel para a determinação da altura efetiva, através do modelo ITU-R. São determinadas as diferenças entre a altura do terreno e linha direta que une a base da antena ao local onde se encontra o terminal. A altura efetiva será a soma da altura da antena com a diferença média calculada anteriormente. Foi considerada uma altura mínima de 1 metro, isto é, nos casos em que o resultado obtido é inferior a 1 m, é considerado este valor.

3.1.2 Cálculo da atenuação

O modelo fornece o valor mediano da atenuação de propagação, dependente de frequência, f, distância do móvel à base, d, altura da antena do móvel ao solo, h_m, e altura da estação base, h_{be}. O cálculo da atenuação de sinal é um dos passos fundamentais do planeamento de redes móveis. Assim, este valor é obtido pela seguinte equação:

$$L_{p[dB]} = 69.55 + 26.16log(f_{[MHz]}) - 13.82log(h_{be[m]}) + [44.90 - 6.55log(h_{be[m]})] * log(d_{[km]}) - H_{mu[dB]}(h_m, f) - \sum fatores\ corretivos \tag{3.1}$$

O termo $H_{mu[dB]}$, é um termo de correcção, que depende do ambiente. Assim, na equação 3.2 está representada a correção proposta pelo modelo para um ambiente suburbano básico. Esta correção considera a altura do móvel e a frequência.

$$H_{mu[dB]} = [1.10log(f_{[MHz]}) - 0.70]h_{m[m]} - [1.56log(f_{[MHz]}) - 0.80] \tag{3.2}$$

Tendo em conta este estudo, apenas é considerada a frequência dos 900 MHz do GSM e uma altura do terminal móvel fixa de quatro metros que corresponde à altura da cabine rádio instalada na cabine do maquinista dos comboios. Posto isto, analisando a equação 3.3 temos que o somatório dos fatores corretivos, do modelo Okumura-Hata, é obtido pela influência de vegetação, influência de água, ondulação do terreno, posição na ondulação do terreno, ruas radiais e, por fim, perdas associadas à difração nos obstáculos, através do Modelo de Deygout.

$$\sum fatores\ corretivos = -L_v + K_{mp} + K_{th} + K_{hp} + K_{al} - L_{diff} \tag{3.3}$$

3.1.3 Atenuação pela influência de vegetação

A vegetação pode ser modelada por uma camada dielétrica, com baixas perdas e não muito denso, que se encontra entre o solo e a atmosfera. A atenuação proveniente da vegetação em ambientes urbanos geralmente não é insignificante e, além disso, quando as antenas estão acima do nível das árvores, pode-se usar o modelo para raios refletidos em uma superfície.

Deste modo, esta camada introduz uma atenuação suplementar no sinal, que segundo o modelo de Weissberger [20] pode ser estimada através de:

$$L_{v[dB]} = \begin{cases} 0.063 * f_{[GHz]}^{0.284} * d_{v[m]} & , 0 \leq d_v \leq 14 \\ 0.187 * f_{[GHz]}^{0.284} * d_{v[m]}^{0.588} & , 14 \leq d_v \leq 400 \end{cases} \tag{3.4}$$

em que a variável d_v é a distância efetiva atravessada pela onda dentro da vegetação, expressa em metros. Este modelo de propagação estima a atenuação de sinal rádio devido à presença de uma ou mais árvores numa ligação rádio ponto a ponto.

3.1.4 Atenuação pela influência da água

Devido a razões topográficas e geográficas, muitas linhas ferroviárias são instaladas no litoral, o que implica que, em alguns segmentos de linha, os caminhos de propagação do sinal de rádio incluem

superfícies de água. Deste modo, a previsão de cobertura de rádio em tais ambientes deve considerar este tipo de *clutter*, bem como as implicações no comportamento do sinal devem ser incluídas na propagação. Nestes pontos constata-se uma descida acentuada da atenuação devido à elevada refletividade da água, sendo por isso contabilizada na estimação de cobertura. A partir das medidas efetuadas, e da sua comparação com os resultados da aplicação do modelo anterior, foi proposto em [3] a utilização de uma aproximação, cujos resultados conduziram a melhores desempenhos. Deste modo, a influência da água é contabilizada pelo modelo Okumura-Hata [21] através do cálculo do fator corretivo dos trajetos mistos, K_{mp}, exibido em 3.5.

$$K_{mp}(\beta)_{[dB]} = m_1 * \beta^2 + m_2 * \beta \tag{3.5}$$

em que o parâmetro $\beta = d_s/d$ corresponde à razão entre a distância total percorrida pelo sinal e a distância percorrida na superfície aquática sem que exista o constrangimento da posição da água, como se pode observar na Figura 3.2. Na tabela 3.2 é indicado os parâmetros de calibração, m_k, deste fator corretivo.

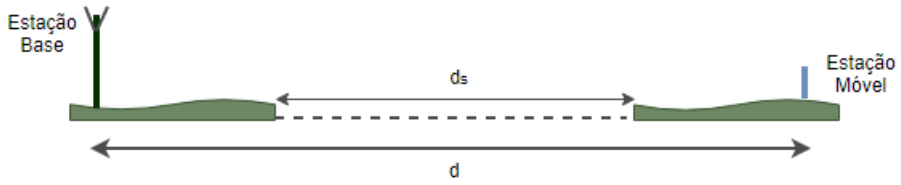


Figura 3.2: Propagação através de trajetos mistos e definição de d_s [3].

Tabela 3.2: Parâmetros de calibração do fator corretivo K_{mp}.

Parâmetro	Valor
$\mathbf{m_1}$	7,9
$\mathbf{m_2}$	12,4

3.1.5 Ondulação do terreno

Em ambientes rurais, muitas vezes é essencial aferir a influência do terreno na estimação de cobertura rádio, uma vez que este dispõe de um relevo acidentado e irregular. Este factor é usado quando, perto do recetor, o terreno apresenta uma ondulação que pode ser enquadrada por 2 valores, tal como se pode observar na Figura 3.3. A atenuação desta ondulação é obtida através da seguinte expressão:

$$K_{th[dB]} = -3log^2(\Delta h) - 0.5log(\Delta h) + 4.5 \tag{3.6}$$

onde Δh representa a altura da ondulação do terreno e é obtida através da diferença entre o percentil 10 e o percentil 90 da respetiva altura do terreno.

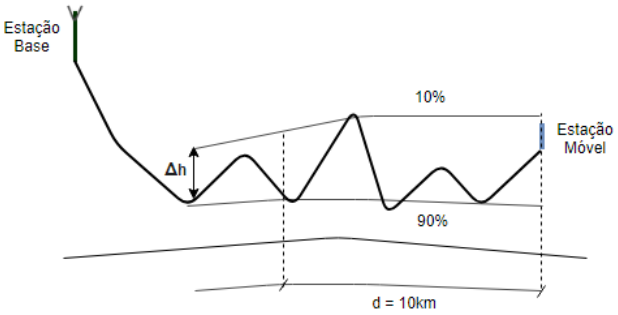


Figura 3.3: Ondulação do terreno, K_{th} [1].

3.1.6 Posição na ondulação do terreno

Uma vez conhecida a posição do terminal móvel na ondulação do terreno, apresentada na Figura 3.4, a atenuação é obtida através da seguinte equação:

$$K_{hp[dB]} = -2log^2(\Delta h_m) + 16log(\Delta h_m) - 12 \tag{3.7}$$

onde Δh_m representa a altura média da ondulação do terreno, cujo valor é obtido através da média entre a diferença entre o percentil 10 e o percentil 90 da altura do terreno.

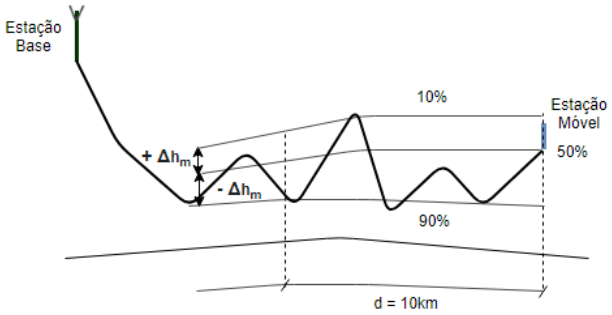


Figura 3.4: Posição na ondulação do terreno, K_{hp} [1].

3.1.7 Ruas radiais

Por fim, este fator corretivo está relacionado com a orientação entre a antena e a linha ferroviária. Quando a orientação da antena é igual à da rua causa uma atenuação baixa que é dada por:

$$K_{al}(d)_{[dB]} = -2.7log(d_{[km]}) + 8.6 \tag{3.8}$$

Conclui-se então que a orientação das antenas em relação ao terminal móvel e a presença de ruas radiais influencia a atenuação do sinal.

3.2 Método de Deygout

Uma vez que o modelo Okumura-Hata não avalia as perdas por difração proveniente dos obstáculos, é considerado o método de Deygout [22]. Deste modo, estabeleceu-se que deveriam ser consideradas as perdas adicionais devido à difração, para efeitos de predição de cobertura rádio em GSM-R, permitindo obter uma maior precisão no cálculo das perdas totais. Este método utiliza a definição do primeiro elipsóide de *Fresnel*, constatando se o elipsóide se encontra impedido e calcula a atenuação respetiva provocada pelo obstáculo. É possível observar a geometria do modelo na Figura 3.5, sendo que este modelo reside numa aproximação admitindo que os obstáculos têm uma geometria em lâmina [23].

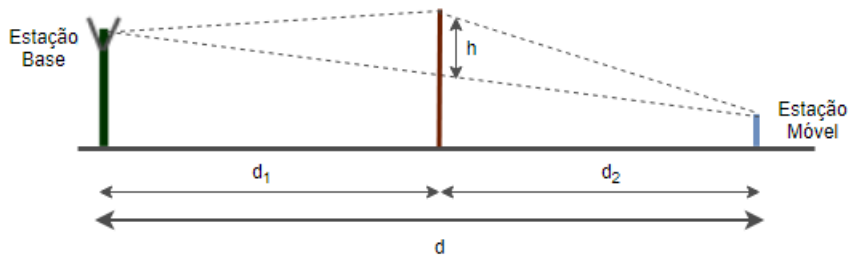


Figura 3.5: Geometria utilizada no cálculo do parâmetro v [1].

Posto isto, é essencial caracterizar o obstáculo por um único parâmetro adimensional v, definido por Fresnel-Kirchhoff, para observar o verdadeiro impacto do obstáculo. Este parâmetro é dado na expressão 3.9.

$$v = h\sqrt{\frac{2d}{\lambda d_1 d_2}} \tag{3.9}$$

em que h corresponde à altura do obstáculo acima (sinal positivo) ou abaixo (sinal negativo) do raio direto entre antenas de emissão e receção, d a distância entre o emissor e o recetor (m), d_1 e d_2 as distâncias do obstáculo a uma e a outra das antenas (m) e λ o comprimento de onda (m).

A atenuação provocada por esse obstáculo é dada por:

$$L_{obs[dB]} = \begin{cases} 0 & ,v \leq -0.7 \\ 6.9 + 20log(\sqrt{(v-0.1)^2)+1} + v - 0.1), & v > -0.7 \end{cases} \tag{3.10}$$

É estimado o obstáculo principal aquele com o maior valor de v e calcula-se a atenuação provocada por este como se este fosse o único obstáculo. Caso exista mais do que um obstáculo em lâmina deverão ser calculados os parâmetros v relativos a cada obstáculo.

No método proposto por Deygout [22] as perdas são obtidas à custa de todos os obstáculos, não apenas o mais importante. Considera-se o obstáculo principal de primeira ordem, o primeiro e último obstáculo de segunda ordem entre as estações base e os vários pontos da linha. No presente trabalho foram considerados apenas os três obstáculos principais, ou seja, os que apresentam maior valor de v, como se pode observar na Figura 3.6.

Além disso, este método divide o percurso em três partes, a primeira parte é a distância até ao primeiro obstáculo principal de segunda ordem, a segunda parte é a distância do primeiro obstáculo principal de segunda ordem até ao obstáculo principal de primeira ordem e a terceira parte é a distância do obstáculo principal de primeira ordem até ao ultimo obstáculo principal de segunda ordem como ilustrado também na Figura 3.6.

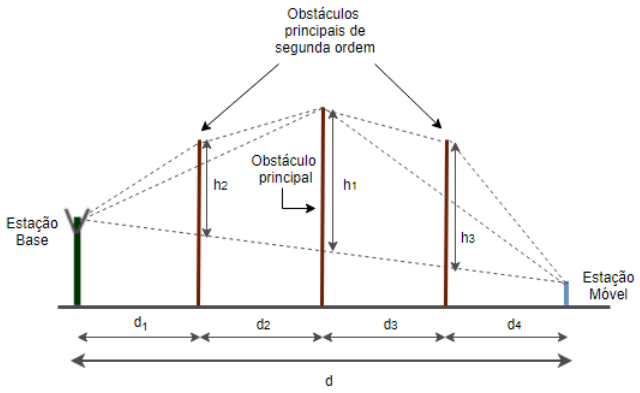


Figura 3.6: Geometria associada ao método de Deygout.

A atenuação total provocada pela difração é obtido através da equação 3.11. Apenas é tida em conta esta atenuação se o valor de v for superior a um determinado limiar pré definido.

$$L_{diff[dB]} = L_{obs_1} + L_{obs_2} + L_{obs_3} \tag{3.11}$$

Deste modo, como é possível observar na equação 3.11 a atenuação por difração no modelo em causa é determinada uma a uma para cada um dos obstáculos, sendo a atenuação total por difração, a soma das três atenuações de cada um deles.

3.3 Algoritmo Desenvolvido

Em [3] demonstrou-se ser válida a utilização do modelo Okumura-Hata para a predição de cobertura rádio em linhas ferroviárias, sendo esta a principal razão da utilização deste modelo. O algoritmo foi desenvolvido em linguagem *Matlab*, uma vez que estamos a tratar com quantidades significativas de dados, derivadas sobretudo da informação geográfica. Irão ser apresentados os parâmetros da linha de Cascais, contudo para as outras duas linhas foi utilizado o mesmo algoritmo.

3.3.1 Modelo de Okumura-Hata

Numa primeira fase, implementou-se o modelo tendo apenas em conta a atenuação calculada pelo modelo Okumura-Hata, sem contabilizar as perdas provocadas pelo somatório dos fatores corretivos. Este cálculo, dependeu da frequência, f, distância do móvel à base, d, altura da estação base, h_{be}, e altura da antena do móvel ao solo, h_m.

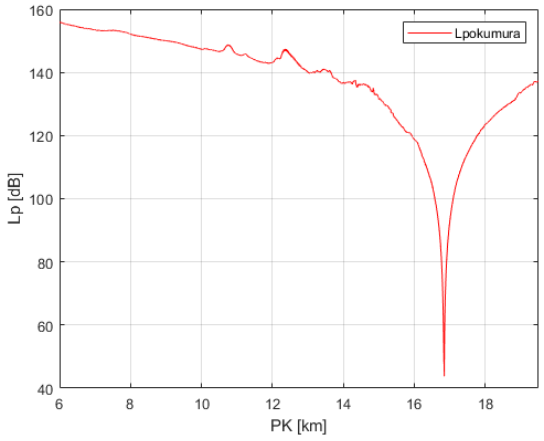


Figura 3.7: Atenuação calculada pelo modelo Okumura-Hata, linha de Cascais.

3.3.2 Fatores corretivos

Posteriormente, implementou-se os diferentes fatores corretivos inerentes ao modelo. Como era de esperar, não existindo nenhum relevo substancial ao longo da linha ferroviária verificou-se pouca influência da vegetação, através do fator corretivo L_v e uma baixa altitude da linha, pelo fator corretivo K_{th}, como consequência da proximidade da linha de água. Por outro lado, uma vez que a cobertura rádio de determinados troços da linha é feita por atravessamento de água, verificou-se uma influência considerável da água, K_{mp}, tendo por isso uma atenuação elevada. O fator corretivo das ruas radiais também apresenta uma atenuação considerável. Na Figura 3.8, referente à linha de Cascais, é possível observar a atenuação provocada por estes fatores corretivos.

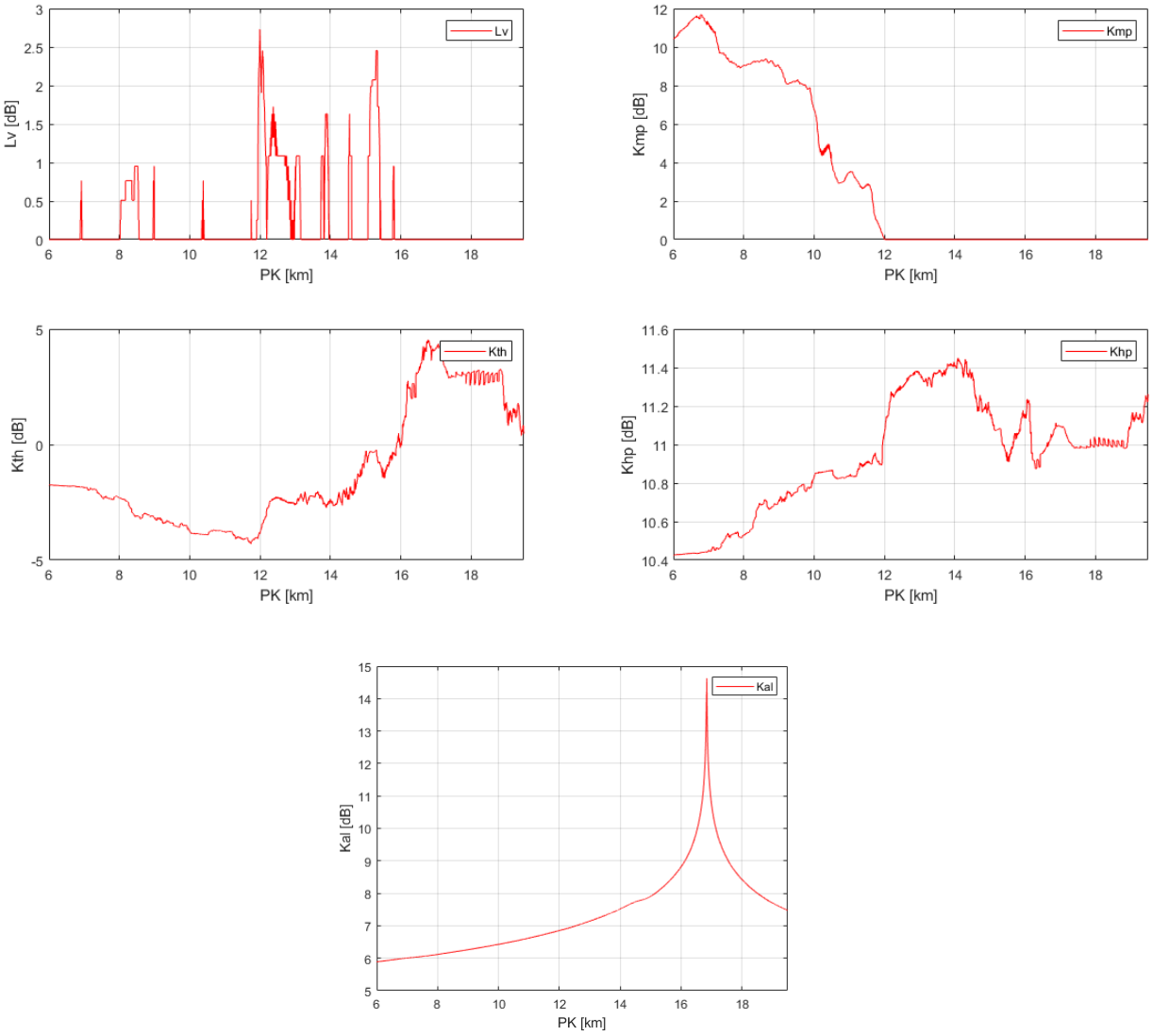


Figura 3.8: Atenuação provocada pelos fatores corretivos, linha de Cascais.

3.3.3 Atenuação provocada pela difração

Como já foi mencionado, o modelo Okumura-Hata não contabiliza as perdas devido à difração proveniente dos obstáculos. Para efeitos da estimação de cobertura rádio em GSM-R considerou-se que deveriam ser contabilizadas estas perdas adicionais, com o intuito de se obter uma maior precisão no cálculo das perdas totais.

Calculou-se, então através do método de Deygout a atenuação provocada pela difração de obstáculos, ilustrada na Figura 3.9.

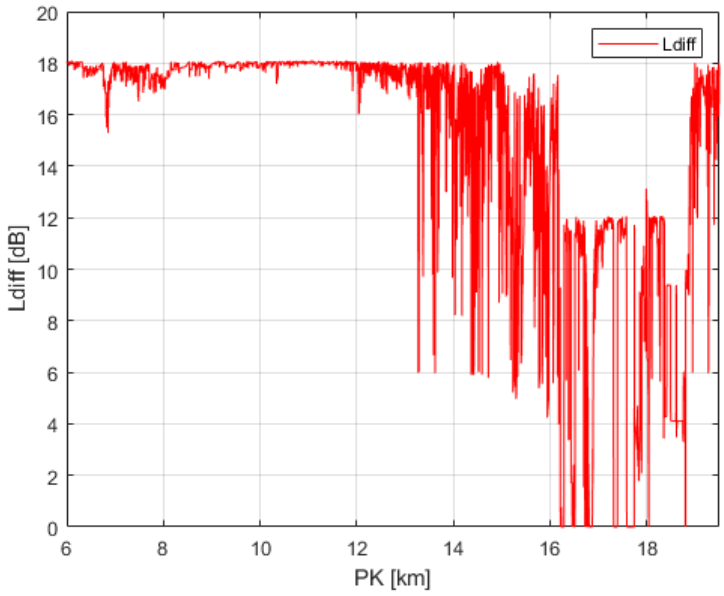
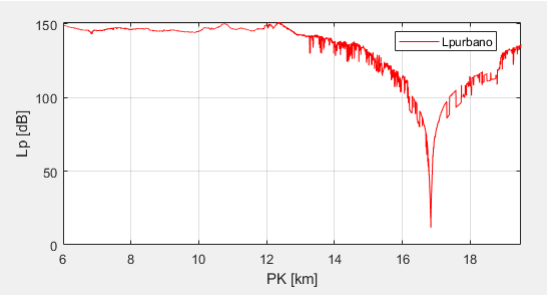


Figura 3.9: Atenuação total provocada pela difração, linha de Cascais.

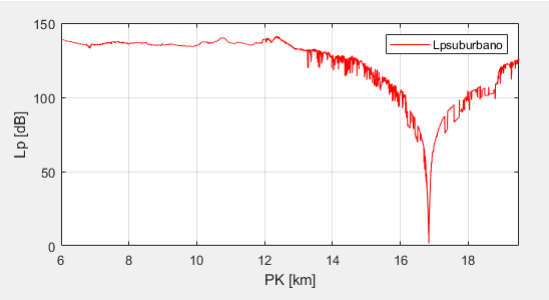
Relembre-se que a atenuação total provocada pela difração é dada pelo somatório das atenuações dos três obstáculos principais.

3.3.4 Atenuação do percurso

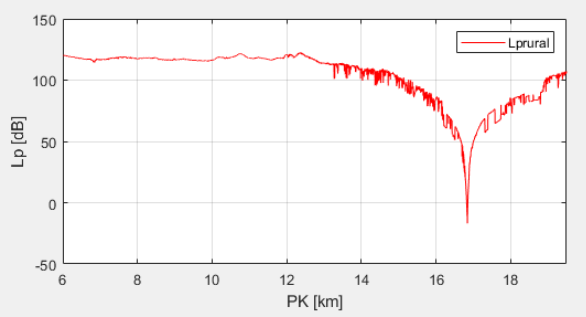
No início deste capítulo, relatou-se que o cálculo da atenuação de sinal é um dos passos fundamentais do planeamento de redes móveis. Por sua vez, procedeu-se à implementação das perdas de propagação (*path loss*) ao longo de PK, tendo em conta os três meios existentes. Chegou-se à conclusão de, como era de esperar, que as perdas associadas ao tipo de ambiente diferem, ilustrada na Figura 3.10. Note-se que a atenuação é maior para o modelo urbano, uma vez que que estamos a ilustrar a linha de Cascais, sendo esta caracterizada por uma urbanização considerável.



(a) *Path loss* do modelo urbano



(b) *Path loss* do modelo suburbano



(c) *Path loss* do modelo rural

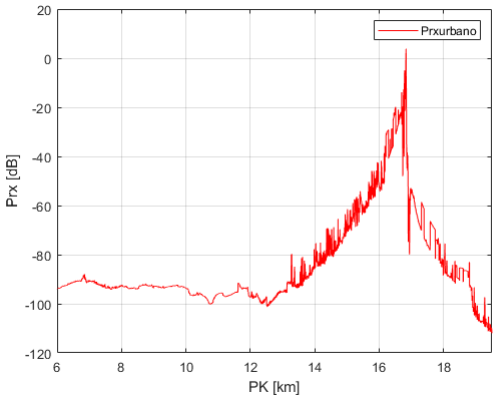
Figura 3.10: Atenuação nos diferentes modelos, linha de Cascais.

3.3.5 Predição de sinal

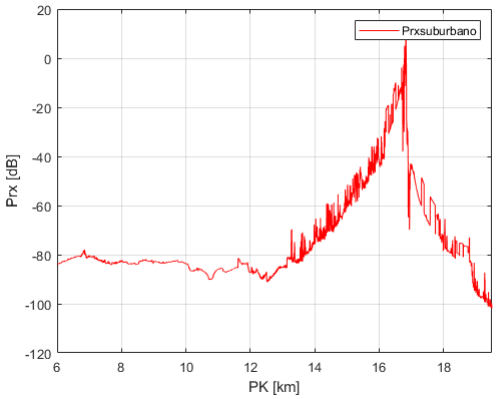
Uma vez implementado todos os parâmetros e atenuações que influenciam a predição de cobertura rádio, procedeu-se à estimação do sinal nos três modelos, através da potência recebida na BS.

Como era de esperar, e de acordo com a Figura 3.10, onde se verifica um decréscimo no valor da atenuação entre o PK13 e o PK17, por sua vez na Figura 3.11 constata-se exatamente o oposto, isto é, um acréscimo no valor da potência recebida do sinal. Ora, isto vai de acordo com o esperado, uma vez que ao existir uma redução no valor da atenuação é de esperar um ganho face ao valor da potência recebida. Por outro lado, a partir do PK17 a atenuação retoma valores mais elevados, confirmando que a partir deste mesmo PK surge uma redução no valor da potência recebida do sinal.

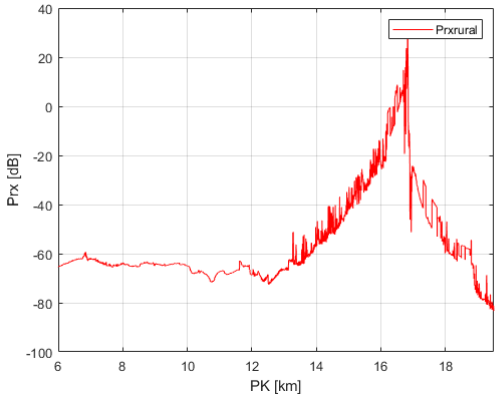
É de salientar também que, no meio rural é onde se verifica melhores picos de potência recebida. Todavia, como vamos analisar a seguir, é o meio onde a predição se desvia mais das medições reais da rede ferroviária.



(a) Modelo urbano



(b) Modelo suburbano



(c) Modelo rural

Figura 3.11: Predição de sinal nos diferentes meios para a linha de Cascais.

3.3.6 Medidas rádio

Para conferir o modelo de predição de cobertura rádio utilizado e corroborar a metodologia implementada na estimação de cobertura rádio ao longo da ferrovia, foram realizadas extensas atividades de medidas rádio pela REFER-Telecom. Efetivamente, a escolha da linha ferroviária para caso de estudo deveu-se à disponibilidade destas medidas de cobertura rádio. Estas foram adquiridas devido à instalação de equipamentos de emissão em diferentes locais por toda a extensão da linha, e equipamento de receção, acrescentado a bordo do comboio, possibilitando a cobertura total da linha.

Na Figura 3.12 é efetuada uma comparação entre a estimação que deriva do modelo implementado e as medidas reais da rede no que diz respeito à linha de Cascais. Para uma melhor análise reduziu-se a extensão da linha ferroviária para os PKs válidos, isto é, os PKs respetivos apenas aos que contêm valores de medidas.

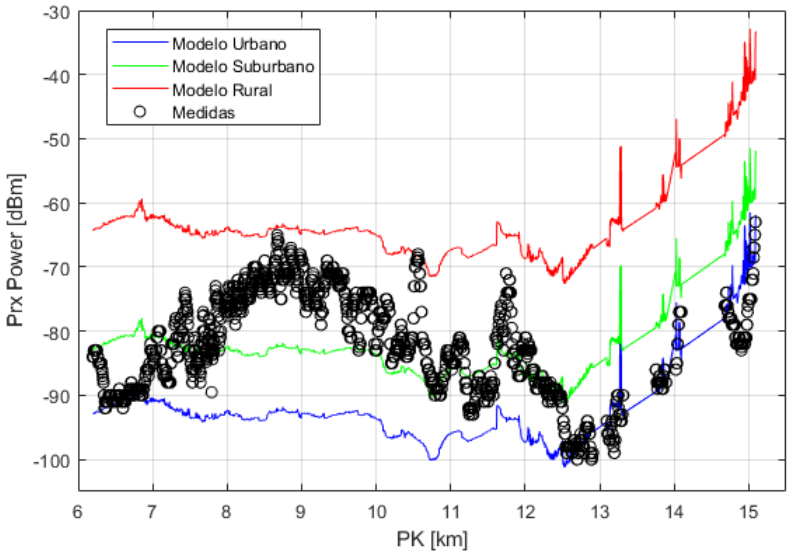


Figura 3.12: Comparação do modelo implementado com as medidas, linha de Cascais.

Note-se, de como foi referido na subsecção anterior, é no modelo rural onde existe a maior discrepância entre a predição de sinal face às medidas reais da rede ferroviária, visto estarmos a analisar a linha de Cascais caracterizada por uma urbanização considerável. Por sua vez, é perfeitamente expectável que sejam as predições dos modelos urbanos e suburbanos as quais se conseguem aproximar mais da curva das medidas.

Salienta-se que entre o PK6 e o PK12 é no modelo suburbano que se consegue uma melhor aproximação entre as duas curvas. A partir do PK12 e até ao fim da linha ferroviária é o modelo urbano onde a predição de sinal é melhor. Além disso, notou-se uma diferença bastante significativa entre a predição e as medidas entre o PK8 e o PK10, que se mantém mesmo após a otimização, como vamos analisar no próximo capítulo, devido à forte presença de água neste troço.

3.3.7 Estatísticas do Erro

De modo a obter uma melhor interpretação e semelhança entre a predição e as medidas calcularam-se estatísticas de primeira ordem e o coeficiente de correlação.

As estatísticas resultantes visam avaliar o erro global da predição do sinal rádio e são traduzidas pelas equações abaixo, nomeadamente o *Medium Error* (ME), o *Root Mean Square Error* (RMSE) e o *Estimated Standard Deviation* (ESD):

$$ME = \frac{1}{n}\sum_{i=1}^{n}|P_{meas_i} - P_{pred_i}| \tag{3.12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|P_{meas_i} - P_{pred_i}|^2} \tag{3.13}$$

$$ESD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|P_{meas_i} - P_{pred_i}| - ME)^2} \tag{3.14}$$

onde P_{meas_i} é o nível de sinal (em dBm) do sinal medido no ponto i, sendo n, o número total de pontos e P_{pred_i}, o valor equivalente da predição.

O cálculo do coeficiente de correlação é dado por:

$$RE = \frac{\sum_{i=1}^{n}(P_{meas_i} - \bar{P}_{meas})(P_{pred_i} - \bar{P}_{pred})}{\sqrt{\sum_{i=1}^{n}(P_{meas_i} - \bar{P}_{meas})^2}\sqrt{\sum_{i=1}^{n}(P_{pred_i} - \bar{P}_{pred})^2}} \tag{3.15}$$

Na Tabela 3.3 podemos comprovar a comparação entre as predições dos três meios e as medidas reais da rede presentes na Figura 3.12.

Tabela 3.3: Estatísticas do modelo implementado face aos três meios.

Modelo	ME [dB]	RMSE [dB]	ESD [dB]	RE
Urbano	11,3896	13,4517	7,1572	0,2513
Suburbano	7,0269	8,5912	4,9429	0,2513
Rural	18,2626	20,1789	8,5828	0,2513

De facto, os valores de ME, RMSE e ESD no modelo rural são muito mais elevados do que nos outros dois modelos. Note-se que é no modelo suburbano onde se encontram os valores mais baixos dos erros, como era de esperar, visto que a linha de Cascais é caracterizada por um ambiente suburbano. Existe uma diferença de 4,3 dB no erro médio, 4,8 dB na raiz do erro quadrático médio e cerca de 2 dB no desvio padrão, face ao ao modelo urbano.

A Figura 3.13 demonstra graficamente as estatísticas da Tabela 3.3, sendo explícito as diferenças entre as estatísticas ME, RMSE e ESD relativamente aos três modelos.

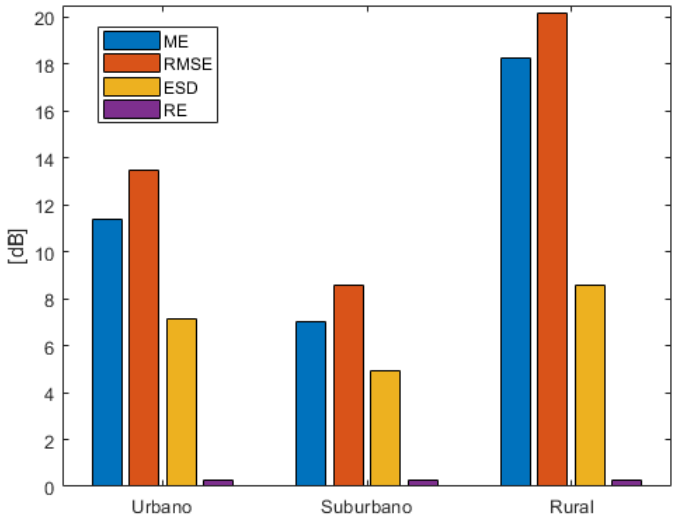


Figura 3.13: Estatísticas do modelo implementado na linha de Cascais.

3.3.8 Classificação

A classificação é o processo de analisar todos os pontos da linha ferroviária e aferir qual dos meios é que apresenta um menor valor de RMSE em cada ponto. Desta forma, é possível atribuir a cada um dos pontos o meio que minimiza o erro. A classificação dos meios é feita segundo o RMSE. A seleção desta medida de erro baseia-se no facto de esta gerar melhor resultados, minimizando as outras estatísticas (ESD e ME) e maximizando o Coeficiente de correlação (RE).

Posteriormente, é feita a avaliação que consiste em avaliar o meio selecionado na classificação e comparar com as medidas reais da rede. Consequentemente, será possível elaborar uma nova predição de sinal que se aproxime da curvatura das medidas ilustrada na Figura 3.12. A avaliação será analisada, com maior detalhe, no capítulo 5 deste trabalho.

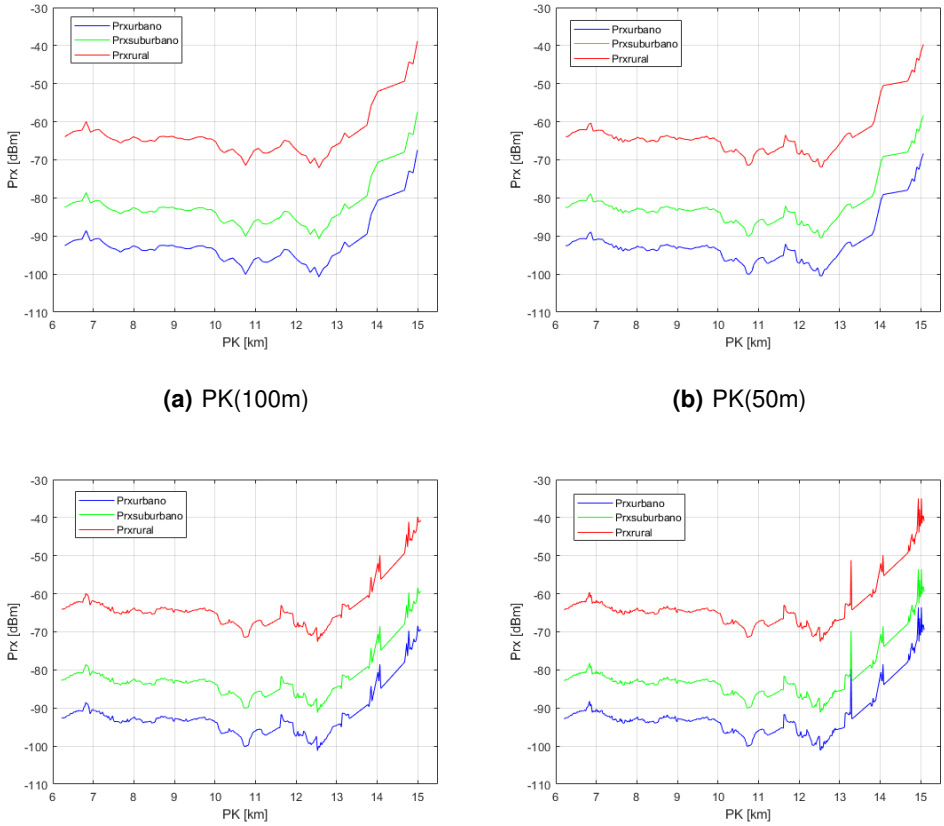
3.3.9 Configuração alternativa de PK

Para minimizar ainda mais os valores de erro agrupou-se os pontos da linha ferroviária em conjuntos de 100 m, 50 m, 20 m e 10 m até se percorrer o comprimento total da linha. Assim, esta configuração de PK teve como objetivo tentar compreender se a quantidade de pontos quilométricos influencia a predição de sinal. Na Tabela 3.4 estão representados o número de segmentos de cada agrupamento que se realizou. Ao agrupar os pontos da linha por diversos segmentos estamos a calcular, consecutivamente, os valores de erro para cada segmento e não tendo em conta os pontos todos da linha.

Tabela 3.4: Número de segmentos dos diversos agrupamentos de pontos.

Percurso	Nº de segmentos
PK(100m)	73
PK(50m)	135
PK(20m)	298
PK(10m)	466

Desta forma, efetuou-se uma nova predição de sinal para cada um dos novos agrupamentos de pontos, retratada na Figura 3.14.



(a) PK(100m)

(b) PK(50m)

(c) PK(20m)

(d) PK(10m)

Figura 3.14: Predição de sinal nos diferentes agrupamentos para a linha de Cascais.

Embora não seja muito percetível, é possível verificar que o número de pontos aumenta de gráfico para gráfico. Uma vez que se está a reduzir o comprimento do agrupamento, é normal que o agrupamento de 10 m (PK10m) seja o que tenha maior número de pontos, em relação aos outros três agrupamentos.

Por enquanto não se pode verificar qual dos agrupamentos apresenta a melhor solução, contudo

espera-se que à partida não seja uma solução fiável, visto que como estamos a repartir a linha em segmentos estamos a reduzir o número de amostras. Isto leva a que o número de pontos contabilizados para o cálculo dos erros seja menor, levando por sua vez a um aumento dos valores estatísticos.

4

H2O Flow

Conteúdo

Procurando uma outra solução de classificação à estimação de cobertura rádio, foi desenvolvido um modelo de *Data Mining* (DM) recorrendo a uma plataforma *open-source*, o *H2O*. O *H2O* é uma interface de aprendizagem automática (ML) e de análise preditiva, em memória, distribuída, rápida e escalável que permite criar modelos de aprendizagem automática em grandes conjuntos de dados e permite também uma fácil produção desses modelos em ambientes empresariais. Tem como missão democratizar a utilização de inteligência artificial.

O código principal do *H2O* é escrito em Java. Os algoritmos são implementados sobre a estrutura distribuída Map/Reduce[1] e utiliza uma *Framework Fork/Join*[2] em Java para *Multi-threading*[3].

Os dados são lidos em paralelo e distribuídos pelo *cluster*, armazenados na memória em formato coluna de forma compactada.

4.1 *H2O Flow*

O *H2O Flow* é uma interface *open-source* do *H2O*, um ambiente interactivo na *web* que permite combinar execução de código, texto, matemática, gráficos num único documento. A interface de utilizador do *H2O Flow* combina perfeitamente a computação de linha de comando com uma interface gráfica moderna. No entanto, em vez de exibir a saída como texto simples, o *Flow* fornece uma interface com o utilizador do género apontar e clicar para cada operação do *H2O*.

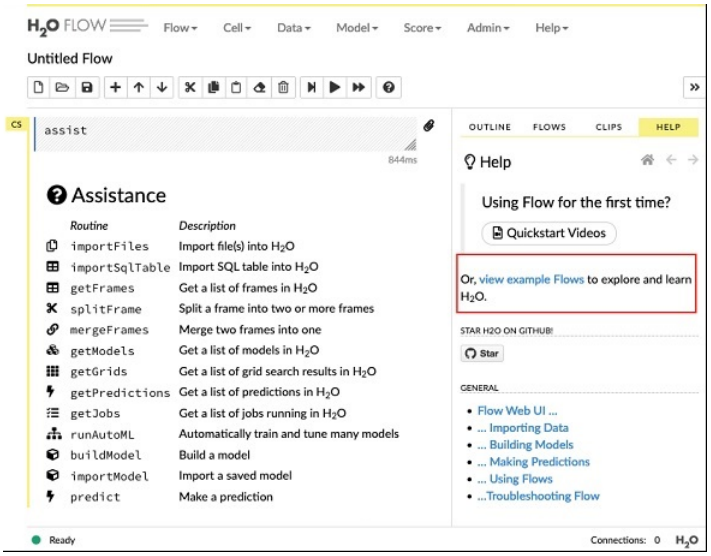


Figura 4.1: Interface de utilizador do *H2O Flow*.

[1] **Map**/**Reduce** é um modelo de programação projetado para processar grandes quantidade de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes.

[2] **Framework Fork**/**Join** fornece ferramentas para ajudar a acelerar o processamento paralelo.

[3] **Multi-threading** é a capacidade que o sistema operacional possui de executar várias *threads* simultaneamente sem que uma interfira na outra.

Os algoritmos disponíveis no *H20 Flow* dividem-se entre aprendizagem supervisionada: problemas de classificação e regressão, e não supervisionada: problemas de *clustering*, deteção de anomalias e outros. Como exemplo, temos *Automatic Machine Learning* (AutoML), *Generalized Linear Model* (GLM), *Gradient Boosting Machine* (GBM), *Support Vector Machine* (SVM) e XGBoost para a aprendizagem supervisionada e, Aggregator, *Generalized Low Rank Models* (GLRM) para não supervisionada.

4.2 *Data Mining*

DM consiste na extração de informação implícita aos dados, previamente desconhecida, e potencialmente útil [24]. A ideia pressuposta ao processo de DM é criar um modelo informático, por exemplo, aplicação, que faculte examinar e extrair padrões dos dados de forma simples, rápida e automática. Em seguida, estes padrões são utilizados para detetar dependências entre os dados, casos específicos, interpretar, compreender, prever ou classificar os novos dados [24].

O processo de DM é antecedido por uma fase designada de pré-processamento, onde é efetuada a limpeza e tratamento dos dados, e procedido por uma etapa de pós-processamento e avaliação dos resultados [24], [25]. Pode-se assim afirmar que, o DM é o conjunto de técnicas e estratégias utilizadas na extração dos padrões dos dados.

4.3 Dimensão dos dados

Um dos problemas questionados em diferentes fases do processo de *Data Mining* é a dimensão dos dados. Esta questão desperta dois pontos essenciais, baixo desempenho dos algoritmos devido à lentidão de execução dos algoritmos e a inexistência de recursos que impossibilitam essa mesma execução. Entende-se como desempenho dos algoritmos, a capacidade destes serem executados e produzirem resultados em tempo considerado adequado. Esse desempenho tende a diminuir proporcionalmente com o tamanho dos dados, uma vez que é necessário mais processamento para produzir um resultado da aplicação de um mesmo algoritmo. De forma a minimizar este problema considerou-se apenas os PK com medições realizadas ao invés de se utilizar todos os pontos de cada linha, para a otimização do algoritmo.

- **Balanceamento dos Dados**

O balanceamento dos dados constitui outro problema, pois afeta a precisão da classificação. Nestes casos, os algoritmos de aprendizagem tendem a especializar-se na classificação da classe maioritária e ignoram os casos da classe minoritária [26]. Uma vez que os dados não se encontram balanceados irão ser utilizadas as seguintes técnicas, todas elas incorporadas no *H2O Flow* que por si efetua automaticamente o balanceamento dos dados:

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

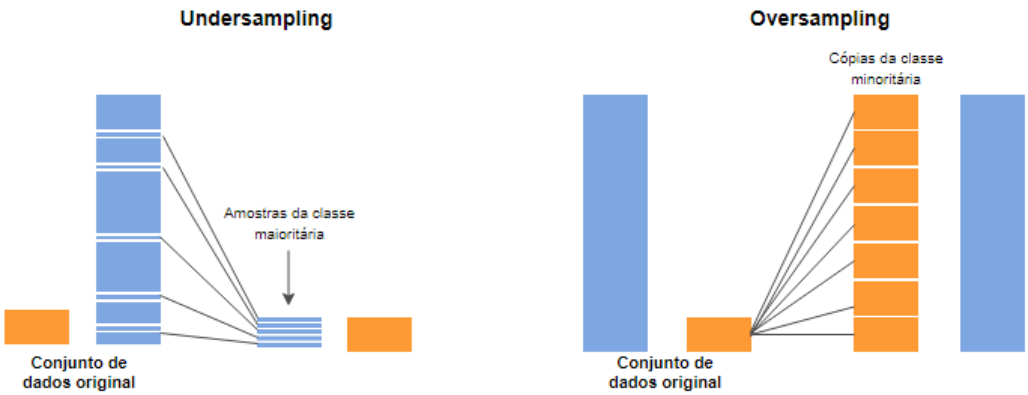


Figura 4.2: Exemplos ilustrativos dos algoritmos *Undersampling* e *Oversampling*.

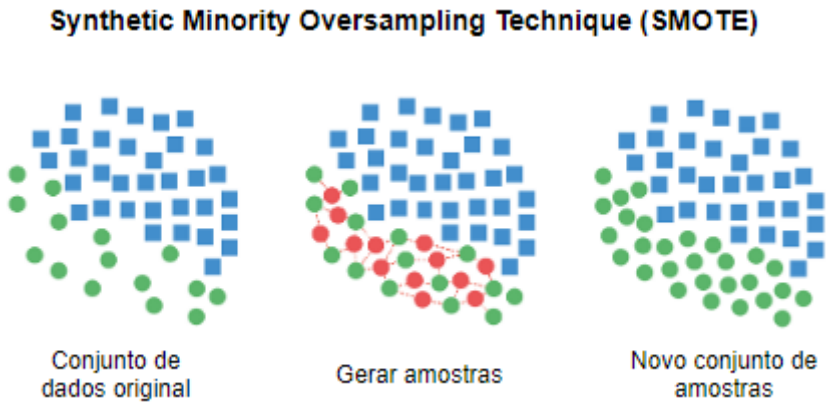


Figura 4.3: Exemplo ilustrativo do algoritmo SMOTE.

A adoção de diferentes técnicas de balanceamento têm impactos diferentes, quando usadas na ordenação dos atributos quanto, à sua importância e na aprendizagem.

O *Undersampling* baseia-se no balanceamento através da redução das amostras da classe maioritária, escolhendo aleatoriamente amostras desta classe até atingir o número de amostras da classe minoritária, ilustrado Figura 4.2. Tem como desvantagem a possibilidade de remover casos da classe maioritária, que podem ser importantes para a classificação desta classe [27].

Por outro lado, o *Oversampling* consiste na duplicação aleatória das amostras da classe minoritária até atingir o número de amostras da classe maioritária, ilustrado Figura 4.2. Tem como principal desvantagem a hipótese de se verificar *overfitting* (sobre-ajuste) do algoritmo. O algoritmo tende a

especializar-se na classificação dos casos duplicados, e diminui a precisão na classificação das amostras que não tem conhecimento prévio [27].

Conforme [28], num conjunto de dados composto principalmente por casos de outra classe, ocorrem regularmente erros na classificação de um caso da classe minoritária. Devido à intenção de classificar estes casos corretamente, os custos associados também são mais elevados. Uma eventual solução que aumente a sensibilidade da classe minoritária é a técnica de *Undersampling*, porém, a solução mais plausível é a junção das técnicas de *Undersampling* da classe maioritária e *Oversampling* da classe minoritária, resultando assim a técnica denominada por SMOTE. Este método permite aumentar o número de amostras no grupo minoritário e gerar amostras sintetizadas por um certo número de amostras adjacentes a cada amostra [27], [28], [29]. As amostras adjacentes consideradas para síntese de novas amostras também englobam o grupo minoritário, ilustrado na Figura 4.3.

4.4 Conjuntos de dados: treino e teste

Uma das etapas do processo do DM centra-se na execução de testes para a validação dos resultados, aplicando algoritmos ao conjunto de dados. Portanto, a validação do modelo centra-se na execução de testes e na comparação dos resultados obtidos com os resultados esperados. Tenciona-se que o modelo seja aplicado a novos casos e que ele seja capaz não apenas focar na classificação dos casos conhecidos, como também classificar novos casos de maneira equivalente ou classificar com melhor desempenho, quando possível. Por outras palavras, considerando que os casos antigos foram utilizados no treino e classificados antecipadamente, procede-se à classificação correta dos novos casos [24].

De forma a otimizar a realização de testes para a validação dos resultados seriam facultados três conjuntos de dados. Um para uso na aprendizagem indicado por conjunto de treino, um para a otimização e verificação dos algoritmos, indicado por conjunto de validação, e outro para testes indicado por conjunto de testes. No entanto, dada a escassez de dados providenciados, é necessário encontrar uma maneira de simular estas condições.

Neste trabalho, foi então utilizado a técnica *hold-out*, visto ser uma boa escolha dada a sua simplicidade, aceitação generalizada e pela sua aplicabilidade ao problema em causa.

- **Hold-out**

O *hold-out* é uma das técnicas mais acessíveis e frequentemente aceite nos procedimentos que envolvem DM. Este reside na divisão do conjunto de dados em dois conjuntos, um para treino e outro para teste. O conjunto de teste será utilizado para verificar os resultados posteriormente. Todavia, os conjuntos de treino e teste devem ser divididos de modo a analisar o seu desempenho em condições

que podem diferir das que se efetua o treino, conseguindo garantir esta divisão selecionando, aleatoriamente, os casos que fazem parte de cada um dos conjuntos. Cada um dos conjuntos tem a sua própria percentagem sendo esta configurável consoante a sua dimensão [30]. Ora, a complexidade desta técnica baseia-se precisamente nisto, saber qual a melhor divisão a efetuar. Do ponto vista da aprendizagem, na maioria dos algoritmos de aprendizagem [24], quanto maior o conjunto de treino, melhor o classificador, até atingir um determinado tamanho e como consequência o classificador piora. Além disso, quanto maior o conjunto de testes, maior a estimativa de erro.

Deste modo, foi utilizada uma distribuição usualmente aceite, de $1/3$ dos dados para conjunto de teste e $2/3$ dos dados para conjunto de treino, uma vez que a classe minoritária tem poucos casos e o conjunto é bastante desequilibrado [24], [31]. Na Figura 4.4 podemos observar uma representação ilustrativa da técnica.

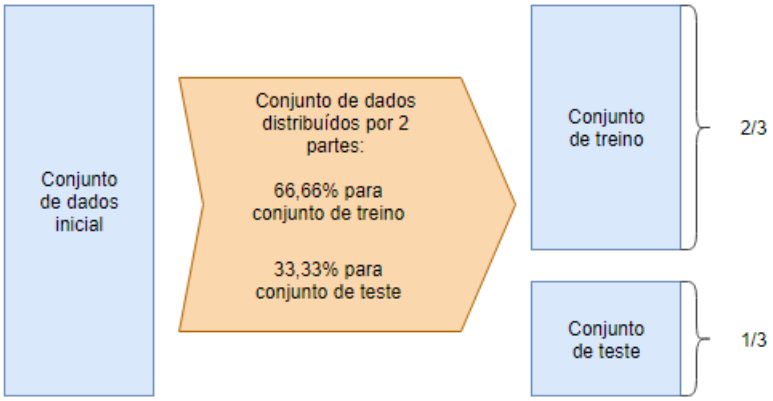


Figura 4.4: Representação da técnica *hold-out*.

Uma vez que a repartição dos casos entre estes dois conjuntos é feita arbitrariamente, é provável que algumas das classes esteja pouca ou nada retratada num dos conjuntos e que a aprendizagem vicie o resultado da classificação. Como maneira de combater este problema, cria-se dois conjuntos que representam todas as classes e conserva-se o relacionamento proporcional entre elas, de modo a obter uma representação maior. No entanto, não há garantia de que esses conjuntos sejam realmente representativos. O dilema é minimizado refazendo o processo de *hold-out* aleatoriamente várias vezes, alcançando vários pares de conjuntos de treino e teste, aos quais são empregues algoritmos de aprendizagem [24]. Contudo, não foi necessário efetuar este processo múltiplas vezes, dado que os resultados obtidos foram satisfatórios e além disso, conseguiu-se analisar o conjunto de dados e perceber o número de pontos de cada classe.

4.5 Algoritmos de Aprendizagem

Por forma a conseguir constatar os padrões que possibilitam minimizar o erro global e ajudar na classificação é necessário aplicar algoritmos ao conjunto de dados. Uma vez adequadamente configurados, os algoritmos têm a capacidade de aprender os padrões dos dados, elaboram uma função aplicada aos novos dados e tentam aproximar-se o mais possível da classificação correta destes. Assim, de acordo com o pretendido, a aprendizagem foi efetuada através do conhecimento prévio da classificação dos casos no conjunto de treino do algoritmo e, posteriormente, esta classificação foi introduzida no processo de aprendizagem. Os algoritmos podem ainda ser repartidos quanto ao tipo de aprendizagem.

Neste trabalho são considerados os tipos de aprendizagem baseados nas árvores de decisão, uma vez que são estes os mais indicados para classes não binárias. Isto, dado que a classificação pretendida é realizada por meio de três classes, ou seja, pretende-de classificar o modelo em três ambientes de propagação diferentes: urbano, suburbano e rural.

Neste processo, entre outras coisas, a capacidade de aprendizagem do algoritmo também deve ser considerada, visto que pode ocorrer os efeitos de *overfitting*[4] e *underfitting*[5]. Portanto, um bom algoritmo é aquele que não se generaliza em excesso.

De forma a compreender qual o melhor algoritmo a utilizar recorreu-se ao algoritmo AutoML, que é um conjunto de métodos de ML e, que de forma automática encontra os melhores algoritmos tendo em conta um conjunto de dados de treino. Assim, após a escolha automática, o AutoML chegou à conclusão que os melhores algoritmos de classificação, para o nosso conjunto de dados de treino em específico, são o GLM e o *XGBoost*.

Por um lado, o GLM estima modelos de regressão (regressão linear e regressão logística para classificação binária) para resultados com base em distribuições exponenciais. Além disso, não pertence à família das árvores [32], o que vai fazer com que não produza resultados fiáveis, como iremos ver adiante.

Por outro lado, o *XGBoost* é um método de *ensemble learning* que oferece uma solução sistemática de modo a construir um modelo de predição combinando a força de vários modelos mais simples, ilustrado na Figura 4.5. Efetivamente, isto torna de facto, o *XGBoost* como uma das melhores soluções para ML [33], visto que mesmo que alguns modelos, isoladamente, obtenham uma predição fraca, quando em conjunto podem formar um modelo com melhor desempenho. O uso do algoritmo mais predominante tem sido com árvores de decisão. No *XGBoost*, as árvores são construídas sequencialmente, de modo que cada árvore seguinte tenha como objetivo reduzir os erros da árvore anterior.

[4] ***overfitting*** ocorre quando o algoritmo adapta-se muito bem ao conjunto de treinos, utilizado na aprendizagem, contudo apresenta fracos resultados nos casos de teste.

[5] ***underfitting*** ocorre quando o algoritmo generaliza muito e assim apresenta uma fraca capacidade para classificar adequadamente os novos casos, tratando-os todos como pertencentes a uma determinada classe.

Cada árvore aprende com seus antecessores e atualiza os erros residuais [33].

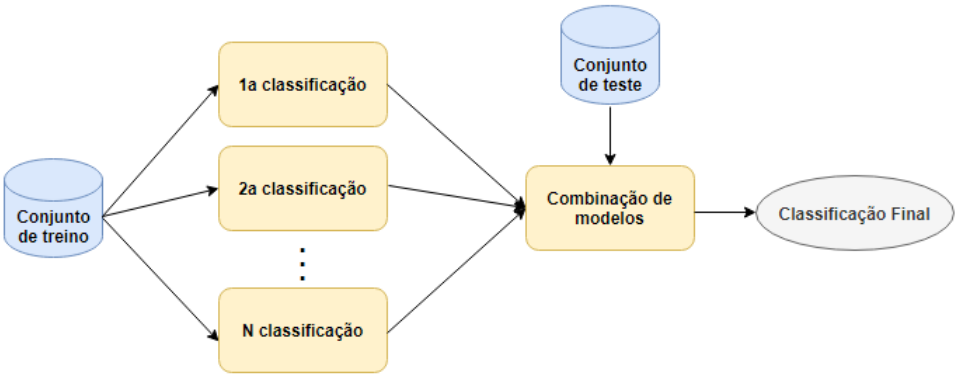


Figura 4.5: Modelo resultante da combinação de vários modelos mais simples.

Como métrica para avaliar e comparar os diversos algoritmos, escolheu-se a matriz confusão. A escolha de um bom método de avaliação é essencial para obter o algoritmo mais adequado e obter o melhor desempenho na classificação do meio.

A matriz confusão, demonstrada na Tabela 4.1, permite averiguar a precisão de um modelo. Com base nesta matriz pode-se ainda verificar a precisão e sensibilidade de um algoritmo. Relembra-se que o objetivo deste trabalho é classificar corretamente todos os casos Verdadeiros Positivos (TP), ou seja, aumentar a sensibilidade e a precisão, mas reduzindo ao máximo classificação incorreta de casos Falsos Negativos (FN) e Falsos Positivos (FP), de modo a conseguir otimizar os valores de precisão e sensibilidade do modelo.

Tabela 4.1: Matriz Confusão para a classe Rural.

		Valor Previsto				
		Rural	**Suburbano**	**Urbano**	**Erro**	**Sensibilidade**
Valor Verdadeiro	**Rural**	TP	FN1	FN2	$\frac{FN1+FN2}{TP+FN1+FN2}$	$\frac{TP}{TP+FN1+FN2}$
	Suburbano	FP1	-	-	-	-
	Urbano	FP2	-	-	-	-
	Precisão	$\frac{TP}{TP+FP1+FP2}$	-	-	-	-

4.6 Implementação no *H2O Flow*

O conjunto de dados é disponibilizado por um ficheiro CSV, que contém informação acerca dos parâmetros que podem influenciar a predição do sinal, assim como, a classificação do meio para cada ponto de PK, ilustrados na Tabela 4.2. A unidade de medida dos parâmetros é o metro.

Tabela 4.2: Sumário dos dados do ficheiro CSV.

Parâmetro	Descrição	Domínio
$\mathbf{d[m]}$	Distância do terminal móvel à estação base.	Numérico
$\mathbf{Água}(\beta)$	Razão entre a distância total percorrida pelo sinal e a distância percorrida na superfície aquática.	Numérico
$\mathbf{h_{be}[m]}$	Altura efetiva dada pelo modelo ITU-R.	Numérico
$\mathbf{\Delta h[m]}$	Altura da ondulação do terreno.	Numérico
$\mathbf{\Delta h_m[m]}$	Altura média da ondulação do terreno.	Numérico
$\mathbf{v_1}$	Parâmetro adimensional referente ao primeiro obstáculo.	Numérico
$\mathbf{v_2}$	Parâmetro adimensional referente ao segundo obstáculo.	Numérico
$\mathbf{v_3}$	Parâmetro adimensional referente ao terceiro obstáculo.	Numérico
$\mathbf{Meio}$	Meio de propagação (urbano, suburbano ou rural).	Enumerado

Na Figura 4.6 está representada a sequência das etapas que serão implementadas pelo *H2O Flow* para treinar o conjunto de dados e para efetuar a predição dos mesmos.

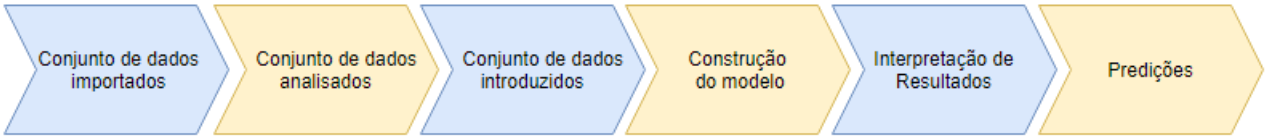


Figura 4.6: Sequência da implementação do *H2O Flow*.

Note-se que, após a introdução dos dados é efetuada a repartição dos conjuntos de treino e teste que servirão para construir o modelo e, posteriormente, na interpretação de resultados é feita a análise e interpretação das classificações obtidas pelos algoritmos GLM e *XGBoost*.

5

Análise de Resultados

Conteúdo

Uma vez estabilizadas as configurações do algoritmo e definido o modelo de propagação, foram realizados alguns testes nas linhas ferroviárias analisadas: Cascais, Beira Baixa e Algarve. Este capítulo viabiliza a configuração final do algoritmo, a análise dos resultados obtidos nas diferentes abordagens seguidas durante este trabalho e também a classificação e a predição obtida através de técnicas de ML com recurso ao *H2O Flow*.

5.1 Configuração do Modelo de Propagação

Como referido anteriormente, o modelo abrange uma série de parâmetros que podem ser calibrados de acordo com o ambiente, dado pela expressão 3.1. O modelo consiste no modelo de Okumura-Hata para os três modelos de propagação com os respetivos fatores corretivos e ainda com perdas adicionais devido à difração, obtido através do método de Deygout.

No começo, foi realizado um conjunto de testes que permitiu definir o modelo de propagação. Foram sugeridas as seguintes alterações ao modelo Okumura-Hata:

- Avaliação e otimização do modelo face aos três meios de propagação;
- Alteração do percurso para diversos agrupamentos;

5.1.1 Avaliação e Otimização do modelo original

A informação geográfica recolhida, juntamente com os parâmetros do modelo, possibilita a composição de uma predição realizada pelo modelo de propagação. Com base no erro entre a predição e as medidas nos diferentes modelos, elaborou-se uma nova predição tendo em conta a otimização dos três modelos face às medidas, em todos os pontos da linha ferroviária.

A avaliação consiste em avaliar o modelo escolhido na classificação, descrita em 3.3.8, e comparar com as medidas reais da rede. Na Tabela 5.1 é possível observar os erros do modelo otimizado em comparação com o modelo anterior, para o primeiro cenário testado: linha de Cascais. Esta linha caracteriza-se pela presença de água ao longo de quase toda a linha e por um ambiente suburbano.

Tabela 5.1: Estatísticas do modelo inicial e do modelo otimizado, linha de Cascais.

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

Sendo a linha de Cascais caracterizada por um ambiente suburbano, é expectável que de entre os três modelos, o suburbano seja aquele que apresenta melhores resultados, comprovado na Tabela 5.1. Note-se ainda, que se conseguiu minimizar os três erros em 3,1 dB, 3,6 dB e 2 dB face ao ME, RMSE e ESD, respetivamente. Por sua vez, maximizou-se o RE em 61%. Na Figura 5.1 ilustra-se a diferença dos erros relativamente ao modelo otimizado e o anterior.

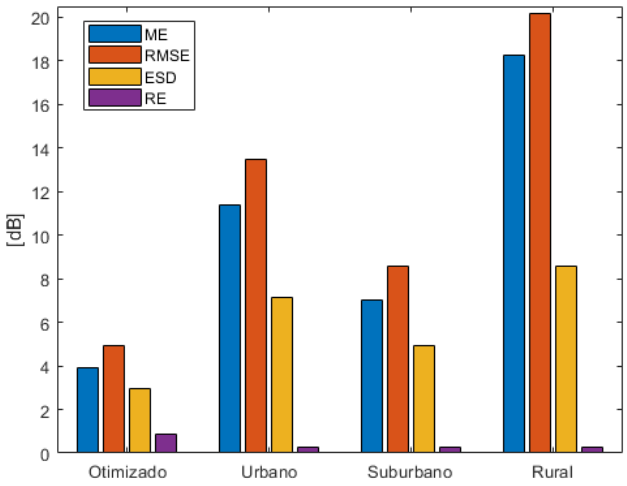


Figura 5.1: Estatísticas da linha de Cascais.

Com a minimização do erro conseguiu-se aproximar as curvaturas entre a predição e as medidas, verificado na Figura 5.2. Obteve-se, assim, uma predição muito mais aproximada face às medidas da linha ferroviária.

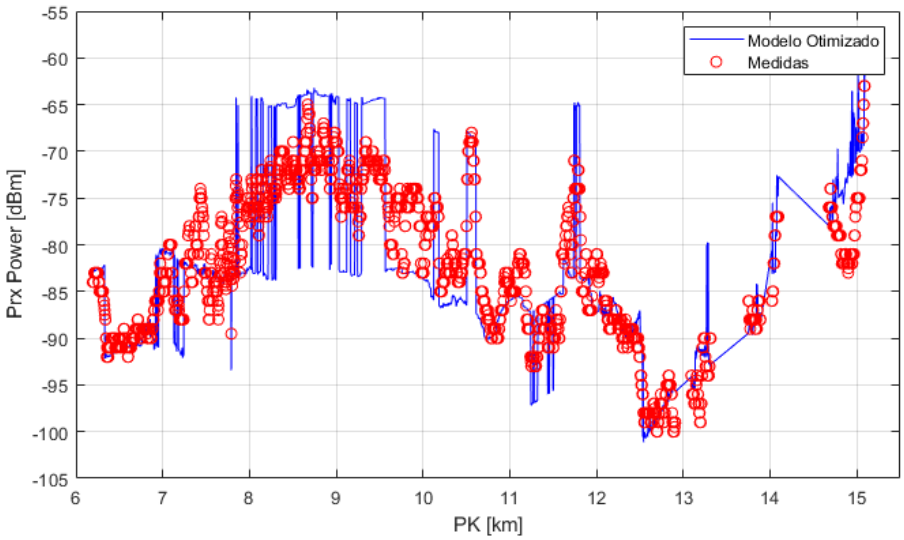


Figura 5.2: Comparação do modelo otimizado com as medidas, linha de Cascais.

O segundo cenário considerado foi a linha da Beira Baixa. A linha caracteriza-se por um ambiente rural com terreno montanhoso e com alguns troços da linha caracterizados com um ambiente suburbano, sendo nestes dois modelos onde se apresentam melhores resultados. Na Tabela 5.2 é possível verificar os erros do modelo otimizado em comparação com o modelo anterior para esta linha.

Tabela 5.2: Estatísticas do modelo inicial e do modelo otimizado, linha da Beira Baixa.

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

Note-se que se conseguiu minimizar os três erros em 6,7 dB, 8,9 dB e 5,9 dB face ao ME, RMSE e ESD, respetivamente. Por sua vez, maximizou-se o RE em cerca de 8%. Na Figura 5.3 ilustra-se a diferença dos erros relativamente ao modelo otimizado e o anterior.

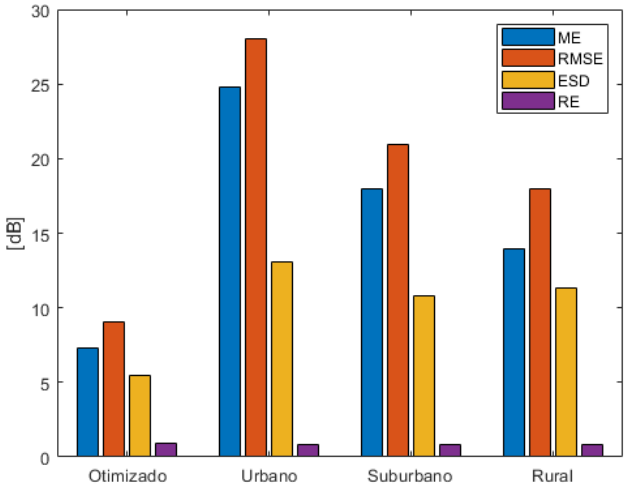


Figura 5.3: Estatísticas da linha da Beira Baixa.

Com a minimização do erro conseguiu-se aproximar as curvaturas entre a predição e as medidas, verificado na Figura 5.4. Obteve-se, assim, uma predição muito mais adjacente face às medidas da linha ferroviária, embora não se tenha conseguido otimizar tanto como na linha de Cascais, visto que a linha da Beira Baixa é caracterizada por um terreno montanhoso e acidentado, tornando mais difícil a estimativa de cobertura rádio em alguns pontos da linha. Na Figura 5.4 está representada a quarta viagem de teste efetuada na linha da Beira Baixa.

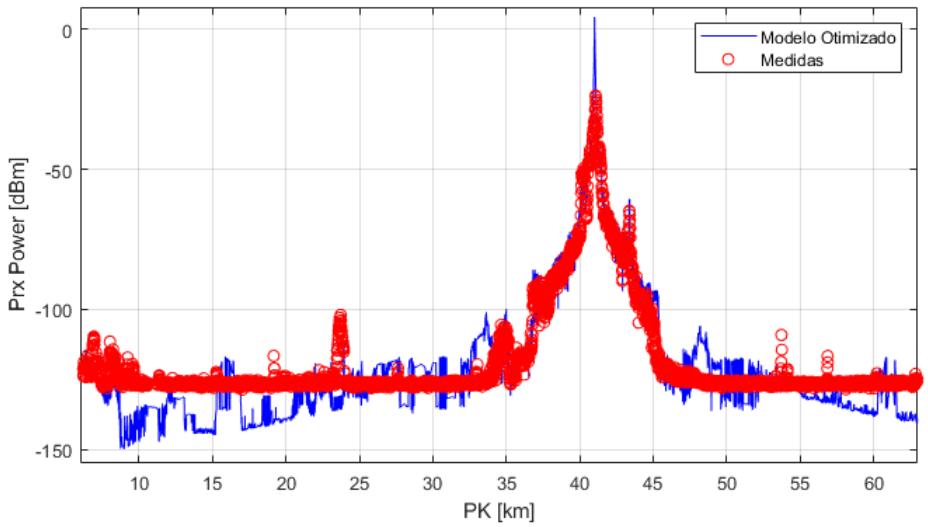


Figura 5.4: Comparação do modelo otimizado com as medidas, linha da Beira Baixa.

O último cenário considerado foi a linha do Algarve. Esta caracteriza-se essencialmente por um ambiente suburbano e pela presença de água em alguns pontos da linha e também pela presença de muitos obstáculos entre as estações base e os diversos pontos da linha. Na Tabela 5.3 é possível verificar os erros do modelo otimizado em comparação com o modelo anterior para esta linha.

Tabela 5.3: Estatísticas do modelo inicial e do modelo otimizado, linha do Algarve.

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

Note-se que os modelos suburbano e rural foram os que apresentaram melhores resultados, dado que a linha do Algarve é caracterizada pela presença de água e obstáculos em diversos pontos da linha. Além disso, conseguiu-se minimizar os três erros em 5,7 dB, 10 dB e 4,7 dB face ao ME, RMSE e ESD, respetivamente. Por sua vez, maximizou-se o RE em cerca de 58%. Na Figura 5.5 ilustra-se a diferença dos erros relativamente ao modelo otimizado e o anterior.

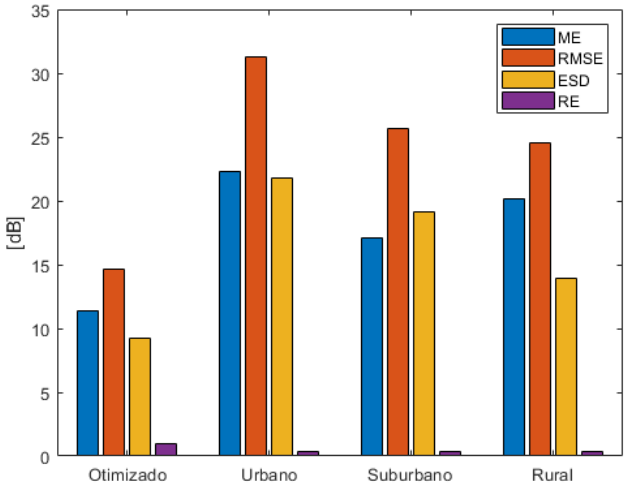


Figura 5.5: Estatísticas da linha do Algarve.

Na Figura 5.6 está representada a quarta viagem de teste efetuada na linha do Algarve. Nota-se nitidamente a diferença do erro obtido, em particular em pontos mais distantes das estações base, contudo, é visível a proximidade das curvaturas face à predição e às medidas ao longo da linha.

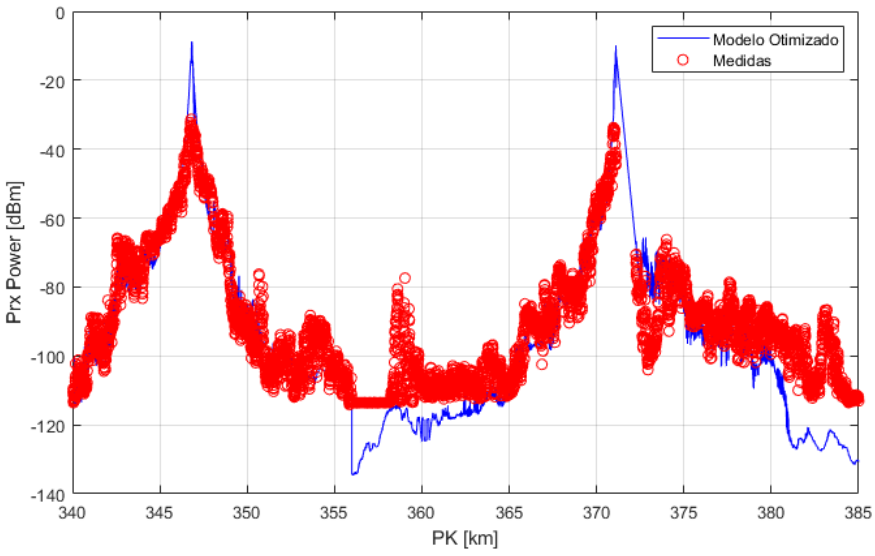


Figura 5.6: Comparação do modelo otimizado com as medidas, linha do Algarve.

5.1.2 Agrupamentos de PK

A alternativa de agrupar os pontos quilométricos em diversos agrupamentos surge na hipótese de se conseguir reduzir o erro RMSE. No entanto, tal não se verificou uma vez que o erro foi menor na utilização de ponto a ponto ao longo da linha e não segmentando a linha por grupos de 100m, 50m, 20m ou 10m, onde é possível confirmar na Tabela 5.4.

Tabela 5.4: Comparação do erro RMSE com o modelo otimizado e os diversos segmentos.

Percurso	RMSE [dB]
Ponto a ponto	4,9169
PK(10m)	5,1737
PK(20m)	5,4024
PK(50m)	5,8729
PK(100m)	6,4223

Note-se que à medida que se aumenta o tamanho do agrupamento o erro RMSE aumenta. A diferença para o PK(10m) não é muito significativa sendo apenas de 0,25 dB, ao contrário para do PK(100m) que já apresenta um erro superior de 1,50 dB face ao modelo otimizado.

5.2 Classificação dos Cenários

Como referido no capítulo 4, recorreu-se a técnicas de ML, com recurso ao *H2O Flow*, de modo a aplicar uma metodologia de classificação, permitindo, para cada ponto, classificar corretamente o ambiente e, consequentemente, selecionar o modelo mais adequado à estimação de cobertura rádio. Uma vez estabelecidos os parâmetros do algoritmo e definido o modelo de propagação, realizaram-se diversos testes em 3 cenários diferentes: a linha de Cascais, a linha da Beira Baixa e a linha do Algarve.

Numa primeira fase pretendeu-se estudar cada cenário individualmente com o objetivo de classificar corretamente os três ambientes em cada uma das linhas ferroviárias. Para cada conjunto de parâmetros é selecionada aleatoriamente pelo *H2O Flow* um conjunto de treino, composto por $2/3$ dos dados, e um conjunto de teste, composto por $1/3$ dos dados. O conjunto de treino é introduzido no AutoML e dá origem à classificação prévia do modelo e que, por sua vez, é aplicada ao conjunto de teste. Para validar o algoritmo é feito a mesma metodologia para as três linhas ferroviárias em questão.

O número de pontos de medida vai variar de linha para linha, uma vez que as linhas têm diferentes comprimentos. Por conseguinte, os conjuntos de treino e de testes apresentam dimensões diferentes e correspondem, respetivamente, a $2/3$ e $1/3$ do número total de pontos de cada cenário, Tabela 5.5.

Tabela 5.5: Número de pontos de cada linha ferroviária.

Cenário	Nº de Pontos	Conjunto de Treino (2/3)	Conjunto de Teste (1/3)
Linha de Cascais	1952	1286	666
Linha da Beira Baixa	4673	3101	1572
Linha do Algarve	6749	4557	2192

5.2.1 Linha de Cascais

O primeiro cenário considerado para treinar é a linha de Cascais. É importante salientar que existe uma grande discrepância no número de pontos relativamente aos três modelos, constatado na Tabela 5.6. Efetivamente, estamos perante um conjunto desequilibrado onde existe o triplo de casos no modelo suburbano em relação ao modelo rural e o dobro em relação ao modelo urbano. Este cenário afeta a credibilidade da classificação obtida e, desta feita o AutoML concluiu que o algoritmo GLM não seria uma boa solução, uma vez que este não conseguiu obter dados suficientes para um conjunto de treino adequado.

Tabela 5.6: Número de pontos para cada modelo, linha de Cascais.

Linha de Cascais	Nº de Pontos
Modelo urbano	534
Modelo suburbano	1078
Modelo rural	340
Total	1952

Na Tabela 5.7 estão representadas as estatísticas do algoritmo GLM, produzidas pelo AutoML, para este cenário. Note-se que o tempo de processamento do AutoML depende do número total de pontos e para este cenário é bastante aceitável. É notório que este algoritmo não foi a melhor solução encontrada pelo AutoML dada a sua posição na seleção do mesmo. O algoritmo atingiu um erro quadrático médio de 0,4228, sendo um valor aceitável, todavia, iremos ver que não é a melhor solução.

Tabela 5.7: Estatísticas do algoritmo GLM para a linha de Cascais.

Parâmetro	Valor
Tempo de processamento	40 min
Posição na seleção do AutoML	129
Distribuição	Multinomial
Erro médio por classe	0,2874
Erro quadrático médio (conjunto de teste)	0,4228

Nas Tabelas 5.8 e 5.9 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente. Nestas matrizes confirma-se a razão pela qual o algoritmo GLM não ser a melhor solução para a classificação correta dos três modelos.

Recorde-se que a sensibilidade é entendida como, de todas as classes verdadeiras, quantas classes é que foram previstas corretamente. Por outro lado, a precisão é dada como, de todas as classes, quantas classes conseguimos prever corretamente.

Tabela 5.8: Matriz confusão para o conjunto de treino para o algoritmo GLM, linha de Cascais.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	125	106	0	0,4589	106/231	**0,54**
	Suburbano	67	604	45	0,1564	112/716	**0,84**
	Urbano	0	72	267	0,2124	72/339	**0,79**
	Total	192	782	312	0,2255	290/1286	-
	Precisão	**0,65**	**0,77**	**0,86**	-	-	-

Tabela 5.9: Matriz confusão para o conjunto de teste para o algoritmo GLM, linha de Cascais.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	53	56	0	0,5138	56/109	**0,49**
	Suburbano	39	303	20	0,1630	59/362	**0,84**
	Urbano	0	32	163	0,1641	32/195	**0,84**
	Total	92	391	183	0,2207	147/666	-
	Precisão	**0,58**	**0,77**	**0,89**	-	-	-

Note-se que é bastante visível a ineficácia deste algoritmo para a classificação do modelo rural, uma vez que o algoritmo obteve uma precisão de 65% e uma sensibilidade de 54% para o conjunto de treino e 58% e 49% para o conjunto de teste, sendo valores muito baixos. Para os outros dois modelos, apesar do algoritmo ter conseguido uma melhor precisão, esta não foi satisfatória. Além disso, apresentou um erro relativamente elevado para a classificação em meio rural, 45,89% e 51,38% para os conjuntos de treino e teste, respetivamente. Ora, isto significa que a classe, neste caso, rural, foi prevista corretamente poucas vezes. Por exemplo, no conjunto de teste, note-se que o algoritmo quando era suposto prever a classe rural, previu corretamente em 53 pontos como classe rural, contudo previu 56 pontos como classe suburbana, originando assim um erro elevado e valores de precisão e sensibilidade baixos para esta classe.

Portanto, o AutoML encontrou o algoritmo *XGBoost* como sendo aquele com melhor desempenho face aos vários erros estatísticos, com ênfase no RMSE. Na Tabela 5.10 estão representadas as es-

tatísticas do algoritmo *XGBoost*. Saliente-se que o AutoML encontrou inúmeros modelos do *XGBoost*, contudo cada um apresenta diferentes parâmetros e valores de entrada configurados automaticamente pelo AutoML, definidos como os hiperparâmetros.

Tabela 5.10: Estatísticas do algoritmo *XGBoost* para a linha de Cascais.

Parâmetro	Valor
Tempo de processamento	40 min
Posição na seleção do AutoML	1
Distribuição	Multinomial
Número de árvores	69
Erro médio por classe	0,0595
Erro quadrático médio (conjunto de teste)	0,2041

Nota-se que para este algoritmo o erro quadrático médio foi menor em relação ao GLM. Nas Tabelas 5.11 e 5.12 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente.

Tabela 5.11: Matriz confusão para o conjunto de treino para o algoritmo *XGBoost*, linha de Cascais.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	231	0	0	0	0/231	**1,0**
	Suburbano	0	715	1	0,0014	1/716	**1,0**
	Urbano	0	0	339	0	0/339	**1,0**
	Total	231	715	340	0,0008	1/1286	-
	Precisão	**1,0**	**1,0**	**1,0**	-	-	-

Tabela 5.12: Matriz confusão para o conjunto de teste para o algoritmo *XGBoost*, linha de Cascais.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	101	8	0	0,0734	8/109	**0,93**
	Suburbano	9	351	2	0,0304	11/362	**0,97**
	Urbano	0	7	188	0,0359	7/195	**0,96**
	Total	110	366	190	0,0390	26/666	-
	Precisão	**0,92**	**0,96**	**0,99**	-	-	-

Como era previsível para o conjunto de treino obteve-se uma precisão de 100%, ocorrendo *overfitting*, visto que o algoritmo conseguiu classificar corretamente todos os pontos presentes no conjunto de treino. Ao invés do GLM, note-se que se conseguiu ótimos valores de precisão para o conjunto de teste, nomeadamente 92% para a classificação em meio rural, 96% para o meio suburbano e 99%

para o meio urbano. Por exemplo, o algoritmo quando era suposto prever a classe urbana, foi capaz de classificar como meio urbano 188 pontos corretamente e apenas 7 pontos incorretamente. De facto, o *XGBoost* atingiu ótimos valores de precisão e sensibilidade, superiores aos do GLM.

5.2.2 Linha da Beira Baixa

O segundo cenário considerado foi a linha da Beira Baixa. A linha caracteriza-se por um ambiente rural com terreno montanhoso. Nesta linha, e como seria de esperar, também existe uma grande discrepância no número de pontos relativamente aos três modelos, constatado na Tabela 5.13. Efetivamente, este cenário afeta a credibilidade da classificação obtida e, por sua vez o AutoML concluiu que o algoritmo GLM não seria uma boa solução.

Tabela 5.13: Número de pontos para cada modelo, linha da Beira Baixa.

Linha da Beira Baixa	Nº de Pontos
Modelo urbano	708
Modelo suburbano	1072
Modelo rural	2893
Total	4673

Na Tabela 5.14 estão representadas as estatísticas do algoritmo GLM, produzidas pelo AutoML, para este cenário. Note-se que o tempo de processamento do AutoML foi mais longo em comparação com a linha de Cascais devido ao maior número total de pontos. É notório que este algoritmo não foi a melhor solução encontrada pelo AutoML dada a sua posição na seleção do mesmo. O algoritmo atingiu um erro quadrático médio de 0,3213, sendo um valor aceitável e melhor do que o da linha de Cascais, todavia, iremos ver que não é a melhor solução.

Tabela 5.14: Estatísticas do algoritmo GLM para a linha da Beira Baixa.

Parâmetro	Valor
Tempo de processamento	47 min
Posição na seleção do AutoML	130
Distribuição	Multinomial
Erro médio por classe	0,2109
Erro quadrático médio (conjunto de teste)	0,3213

Nas Tabelas 5.15 e 5.16 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente. Nestas matrizes confirma-se a razão pela qual o algoritmo GLM não ser a melhor

solução para a classificação correta dos três modelos.

Tabela 5.15: Matriz confusão para o conjunto de treino para o algoritmo GLM, linha da Beira Baixa.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1806	99	4	0,0540	103/1909	**0,95**
	Suburbano	194	464	55	0,3492	249/713	**0,65**
	Urbano	1	109	369	0,2296	110/479	**0,77**
	Total	2001	672	428	0,1490	462/3101	-
	Precisão	**0,90**	**0,69**	**0,86**	-	-	-

Tabela 5.16: Matriz confusão para o conjunto de teste para o algoritmo GLM, linha da Beira Baixa.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	942	42	0	0,0427	42/984	**0,96**
	Suburbano	98	228	33	0,3649	131/359	**0,64**
	Urbano	0	41	188	0,1790	41/229	**0,82**
	Total	1040	311	221	0,1361	214/1572	-
	Precisão	**0,91**	**0,73**	**0,85**	-	-	-

Note-se que é bastante visível, a ineficácia deste algoritmo para a classificação do modelo suburbano, uma vez que o algoritmo obteve uma precisão de 69% e uma sensibilidade de 65% para o conjunto de treino e 73% e 64% para o conjunto de teste, sendo valores muito baixos. Para os outros dois modelos, apesar do algoritmo ter conseguido uma melhor precisão, esta não foi satisfatória.

Portanto, o AutoML encontrou o algoritmo *XGBoost* como sendo aquele com melhor desempenho face aos vários erros estatísticos. Na Tabela 5.17 estão representadas as estatísticas do algoritmo *XGBoost*.

Tabela 5.17: Estatísticas do algoritmo *XGBoost* para a linha da Beira Baixa.

Parâmetro	**Valor**
Tempo de processamento	47 min
Posição na seleção do AutoML	1
Distribuição	Multinomial
Número de árvores	58
Erro médio por classe	0,0603
Erro quadrático médio (conjunto de teste)	0,1865

Nota-se que para este algoritmo o erro quadrático médio foi menor em relação ao GLM. Nas Tabelas 5.18 e 5.19 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente.

Tabela 5.18: Matriz confusão para o conjunto de treino para o algoritmo *XGBoost*, linha da Beira Baixa.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1909	0	0	0	0/1909	**1,0**
	Suburbano	0	713	0	0	0/713	**1,0**
	Urbano	0	0	479	0	0/479	**1,0**
	Total	1909	713	479	0	0/3101	-
	Precisão	**1,0**	**1,0**	**1,0**	-	-	-

Tabela 5.19: Matriz confusão para o conjunto de teste para o algoritmo *XGBoost*, linha da Beira Baixa.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	971	13	0	0,0132	13/984	**0,99**
	Suburbano	21	316	22	0,1198	43/359	**0,88**
	Urbano	0	11	218	0,0480	11/229	**0,95**
	Total	992	340	240	0,0426	67/1572	-
	Precisão	**0,98**	**0,93**	**0,91**	-	-	-

Tal como na linha de Cascais, para o conjunto de treino obteve-se uma precisão de 100%, ocorrendo *overfitting*, visto que o algoritmo conseguiu classificar corretamente todos os pontos presentes no conjunto de treino. Além disso, ao invés do GLM, note-se que se conseguiu ótimos valores de precisão para o conjunto de teste, nomeadamente 98% para a classificação em meio rural, 93% para o meio suburbano e 91% para o meio urbano. Por exemplo, o algoritmo quando era suposto prever a classe urbana, foi capaz de classificar como meio urbano 218 pontos corretamente e apenas 11 pontos incorretamente. De facto, o *XGBoost* atingiu ótimos valores de precisão e sensibilidade, superiores aos do GLM.

5.2.3 Linha do Algarve

O terceiro cenário considerado foi a linha do Algarve. Esta linha caracteriza-se por um ambiente suburbano e pela presença de obstáculos entre as várias estações base e os vários pontos da linha. Nesta linha, e como seria de esperar, também existe uma grande discrepância no número de pontos relativamente aos três modelos, constatado na Tabela 5.20. Tal como sucedido nas outras duas linhas, este cenário afeta a credibilidade da classificação obtida e o AutoML concluiu que o algoritmo GLM não seria uma boa solução.

Tabela 5.20: Número de pontos para cada modelo, linha do Algarve.

Linha do Algarve	Nº de Pontos
Modelo urbano	1611
Modelo suburbano	2456
Modelo rural	2682
Total	6749

Na Tabela 5.21 estão representadas as estatísticas do algoritmo GLM, produzidas pelo AutoML, para esta linha. Note-se que o tempo de processamento do AutoML foi mais longo em comparação com a linha de Cascais e Beira Baixa devido ao maior número total de pontos. É notório que este algoritmo não foi a melhor solução encontrada pelo AutoML dada a sua posição na seleção do mesmo. O algoritmo atingiu um erro quadrático médio de 0,3901, sendo um valor aceitável e melhor até do que o da linha da Beira Baixa, todavia, iremos ver que não é a melhor solução.

Tabela 5.21: Estatísticas do algoritmo GLM para a linha do Algarve.

Parâmetro	Valor
Tempo de processamento	48 min
Posição na seleção do AutoML	176
Distribuição	Multinomial
Erro médio por classe	0,2125
Erro quadrático médio (conjunto de teste)	0,3901

Nas Tabelas 5.22 e 5.23 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente. Nestas matrizes confirma-se a razão pela qual o algoritmo GLM não ser a melhor solução para a classificação correta dos três modelos.

Tabela 5.22: Matriz confusão para o conjunto de treino para o algoritmo GLM, linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1547	246	0	0,1372	246/1793	**0,86**
	Suburbano	233	1164	287	0,3088	520/1684	**0,69**
	Urbano	0	239	841	0,2213	239/1080	**0,78**
	Total	1780	1649	1128	0,2205	1005/4557	-
	Precisão	**0,87**	**0,71**	**0,75**	-	-	-

Tabela 5.23: Matriz confusão para o conjunto de teste para o algoritmo GLM, linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	781	108	0	0,1215	108/889	**0,88**
	Suburbano	96	516	160	0,3316	256/772	**0,67**
	Urbano	0	98	433	0,1846	98/531	**0,82**
	Total	877	722	593	0,2108	462/2192	-
	Precisão	**0,89**	**0,71**	**0,73**	-	-	-

Note-se que é bastante visível, a ineficácia deste algoritmo para a classificação do modelo suburbano, uma vez que o algoritmo obteve uma precisão de 71% e uma sensibilidade de 69% para o conjunto de treino e 71% e 67% para o conjunto de teste, sendo valores muito baixos. Para os outros dois modelos, apesar do algoritmo ter conseguido uma melhor precisão, esta não foi satisfatória.

Tal como nos outros dois cenários, o AutoML encontrou o algoritmo *XGBoost* como sendo aquele com melhor desempenho face aos vários erros estatísticos. Na Tabela 5.24 estão representadas as estatísticas do algoritmo *XGBoost*.

Tabela 5.24: Estatísticas do algoritmo *XGBoost* para a linha do Algarve.

Parâmetro	**Valor**
Tempo de processamento	48 min
Posição na seleção do AutoML	1
Distribuição	Multinomial
Número de árvores	77
Erro médio por classe	0,0876
Erro quadrático médio (conjunto de teste)	0,2549

Repara-se que para este algoritmo o erro quadrático médio foi menor em relação ao GLM. Nas Tabelas 5.25 e 5.26 são apresentadas as matrizes confusão para os conjuntos de treino e teste, respetivamente.

Tabela 5.25: Matriz confusão para o conjunto de treino para o algoritmo *XGBoost*, linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1793	0	0	0	0/1793	**1,0**
	Suburbano	0	1684	0	0	0/1684	**1,0**
	Urbano	0	0	1080	0	0/1080	**1,0**
	Total	1793	1684	1080	0	0/4557	-
	Precisão	**1,0**	**1,0**	**1,0**	-	-	-

Tabela 5.26: Matriz confusão para o conjunto de teste para o algoritmo *XGBoost*, linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	867	22	0	0,0247	22/889	**0,98**
	Suburbano	19	674	79	0,1269	98/772	**0,87**
	Urbano	1	58	472	0,1111	59/531	**0,89**
	Total	887	754	551	0,0817	179/2192	-
	Precisão	**0,98**	**0,89**	**0,86**	-	-	-

Tal como nas outras duas linhas em estudo, para o conjunto de treino obteve-se uma precisão de 100%, ocorrendo *overfitting*, uma vez que o algoritmo conseguiu classificar corretamente todos os pontos presentes no conjunto de treino. Além disso, ao invés do GLM, saliente-se que se conseguiu valores muito satisfatórios de precisão para o conjunto de teste, nomeadamente 98% para a classificação em meio rural, 89% para o meio suburbano e 86% para o meio urbano. Por exemplo, o algoritmo quando era suposto prever a classe rural, foi capaz de classificar como meio rural 867 pontos corretamente e apenas 22 pontos incorretamente. Efetivamente, o *XGBoost* obteve resultados muito satisfatórios com valores de precisão e sensibilidade elevados, superiores aos do GLM.

5.3 Avaliação e otimização através do *H2O Flow*

Com base na classificação obtida pelo *XGBoost*, este elaborou uma nova predição para cada cenário considerado. Posteriormente, foi efetuada a comparação desta nova predição com a predição obtida pelo modelo otimizado de modo a conferir a fiabilidade da utilização de técnicas de *machine learning* para a estimação de cobertura rádio.

5.3.1 Linha de Cascais

O algoritmo *XGBoost* obteve ótimos valores de sensibilidade e precisão em relação à predição obtida através da classificação efetuada. Na Tabela 5.27 está representada a matriz confusão da predição elaborada, para a linha de Cascais. Note-se que se obteve valores de precisão de 98% para os modelos rural e suburbano e 100% para o modelo urbano. Isto significa que o algoritmo *XGBoost* foi capaz de prever grandes parte das classes corretamente, viabilizando desta feita uma boa predição de cobertura rádio.

Tabela 5.27: Matriz confusão da predição da linha de Cascais.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	326	14	0	0,0412	14/340	**0,96**
	Suburbano	8	1068	2	0,0093	10/1078	**0,99**
	Urbano	0	8	526	0,0150	8/534	**0,99**
	Total	334	1090	528	0,0164	32/1952	-
	Precisão	**0,98**	**0,98**	**1,0**	-	-	-

Na Tabela 5.28 é possível observar os erros do modelo otimizado pelo *H2O Flow* em comparação com os três modelos originais e com o modelo otimizado.

Tabela 5.28: Estatísticas da linha de Cascais.

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

Note-se que se conseguiu otimizar o modelo inicial com a predição feita pelo *H2O Flow*. Além disso, esta predição foi muito semelhante à predição obtida pelo modelo otimizado, diferindo 0,03 dB, 0,05 dB e 0,04 dB face ao ME, RMSE e ESD, respetivamente. Na Figura 5.7 é possível verificar que a curva da predição obtida pelo *H2O Flow* sobrepõe-se à do modelo otimizado, o que demonstra a proximidade dos modelos.

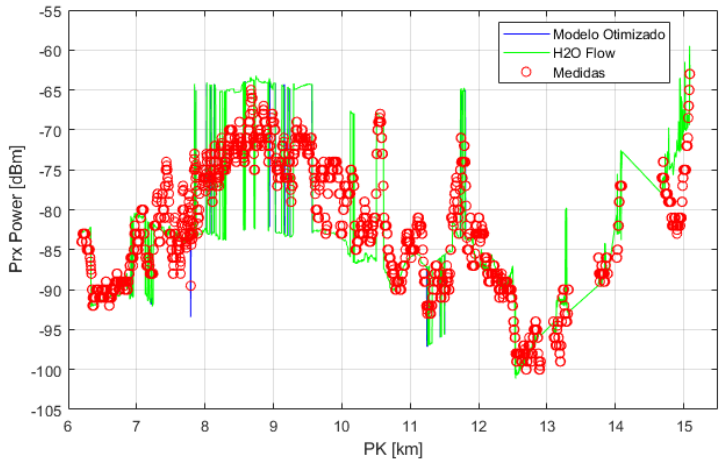


Figura 5.7: Comparação do modelo obtido pelo *H2O Flow* com o modelo otimizado e as medidas, linha de Cascais.

5.3.2 Linha da Beira Baixa

Para a linha da Beira Baixa, o algoritmo *XGBoost* também obteve ótimos valores de sensibilidade e precisão em relação à predição obtida através da classificação efetuada. Na Tabela 5.29 está representada a matriz confusão da predição elaborada, para este cenário.

Tabela 5.29: Matriz confusão da predição da linha da Beira Baixa.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	2879	14	0	0,0048	14/2893	**1,0**
	Suburbano	21	1037	14	0,0326	35/1072	**0,97**
	Urbano	0	19	689	0,0268	19/708	**0,97**
	Total	2900	1070	703	0,0146	68/4673	-
	Precisão	**0,99**	**0,97**	**0,98**	-	-	-

Note-se que se obteve, em relação à predição elaborada pelo algoritmo, valores de precisão e sensibilidade de 99% e 100%, respetivamente, para o modelo rural, dada a maior quantidade de pontos para este modelo. É possível que tenha ocorrido *overfitting* uma vez que este atingiu valores extremamente elevados. Para o modelo suburbano obteve-se valores de precisão e sensibilidade de 97% e, por fim, para o modelo urbano obteve-se valores de precisão e sensibilidade de 98% e 97%, respetivamente. Na Tabela 5.30 é possível observar os erros do modelo otimizado pelo *H2O Flow* em comparação com os três modelos originais e com o modelo otimizado.

Tabela 5.30: Estatísticas da linha da Beira Baixa.

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

Note-se que se conseguiu otimizar o modelo inicial com a predição feita pelo *H2O Flow*. Além disso, esta predição foi muito semelhante à predição obtida pelo modelo otimizado, diferindo 0,07 dB, 0,05 dB e 0,002 dB face ao ME, RMSE e ESD, respetivamente. Na Figura 5.8 é possível verificar que a curva da predição obtida pelo *H2O Flow* sobrepõe-se à do modelo otimizado, o que demonstra a proximidade dos modelos.

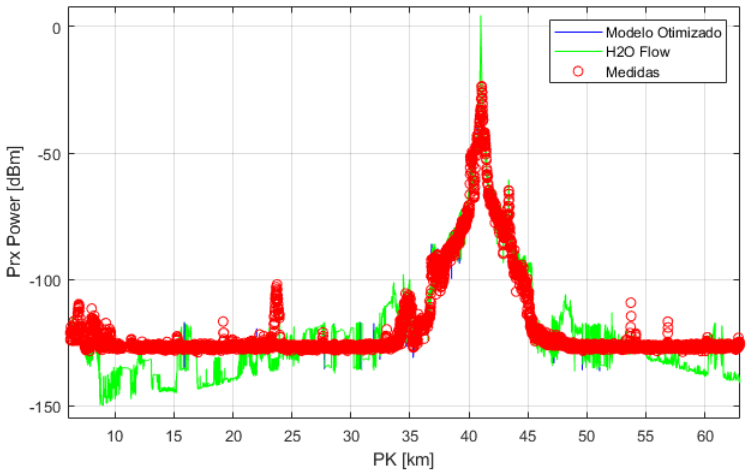


Figura 5.8: Comparação do modelo obtido pelo *H2O Flow* com o modelo otimizado e as medidas, linha da Beira Baixa.

5.3.3 Linha do Algarve

Para a linha do Algarve, o algoritmo *XGBoost* também obteve valores muito satisfatórios de sensibilidade e precisão em relação à predição obtida através da classificação efetuada. Na Tabela 5.31 está representada a matriz confusão da predição elaborada, para este cenário.

Tabela 5.31: Matriz confusão da predição da linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	2659	23	0	0,0086	23/2682	**0,99**
	Suburbano	19	2386	51	0,0285	70/2456	**0,97**
	Urbano	0	67	1544	0,0416	67/1611	**0,96**
	Total	2678	2476	1595	0,0237	160/6749	-
	Precisão	**0,99**	**0,96**	**0,97**	-	-	-

Note-se que se obteve, em relação à predição elaborada pelo algoritmo, valores de precisão e sensibilidade de 99% para o modelo rural. Para o modelo suburbano obteve-se valores de precisão e sensibilidade de 96% e 97%, respetivamente. Por fim, para o modelo urbano obteve-se valores de precisão e sensibilidade de 97% e 96%, respetivamente.

Na Tabela 5.32 é possível observar os erros do modelo otimizado pelo *H2O Flow* em comparação com os três modelos originais e com o modelo otimizado. Note-se que se conseguiu otimizar o modelo inicial com a predição feita pelo *H2O Flow*. Além disso, esta predição foi muito semelhante à predição obtida pelo modelo otimizado, diferindo 1,94 dB, 2,16 dB e 4,68 dB face ao ME, RMSE e ESD, respetivamente. Contudo, foi o cenário que apresentou valores de erro mais elevados em comparação com o

modelo otimizado.

Tabela 5.32: Estatísticas da linha do Algarve.

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

Na Figura 5.9 é possível verificar que a curva da predição obtida pelo *H2O Flow* sobrepõe-se, em grande parte, à do modelo otimizado, o que demonstra a proximidade dos modelos.

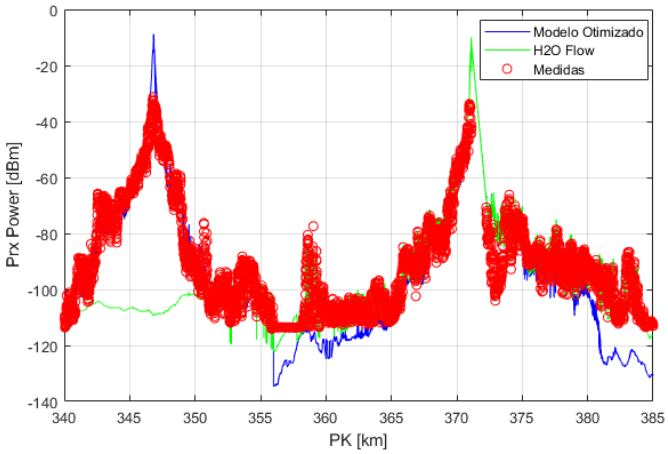


Figura 5.9: Comparação do modelo obtido pelo *H2O Flow* com o modelo otimizado e as medidas, linha do Algarve.

5.4 Classificação e Predição Geral

Depois, de analisadas as linhas separadamente considerou-se criar um novo conjunto de dados que consiste na junção das três linhas, com o objetivo de uniformizar o modelo obtido pelo *XGBoost*, de maneira a que se consiga atingir bons valores de predição para outras linhas ferroviárias. Como referido anteriormente, e se observa na Tabela 5.5, a linha de Cascais tem apenas 1952 pontos e o *ratio* entre a linha da Beira Baixa e esta é de 2,39, por outro lado o *ratio* com a linha do Algarve é de 3,46. De forma a ter uma igual representatividade de todas as linhas, o conjunto original das linhas da Beira Baixa e do Algarve foi aleatoriamente dividido, no *H2O Flow*, em dois conjuntos: um para criar o novo conjunto de treino e um segundo para o conjunto de teste. Então o novo conjunto de treino ficou com a totalidade de

pontos da linha de Cascais (1952), mais, aproximadamente, 45% do conjunto original da linha da Beira baixa (2114) e, aproximadamente, 30% do conjunto original da linha do Algarve (2039). A dimensão do número total de pontos do novo conjunto é apresentado na Tabela 5.33.

Tabela 5.33: Número de pontos para os conjuntos de treino e teste.

Cenário	Nº de Pontos	Conjunto de Treino (2/3)	Conjunto de Teste (1/3)
Modelo Geral	6105	4061	2044
Modelo Urbano	1335	886	449
Modelo Suburbano	2334	1545	789
Modelo Rural	2436	1630	806

O conjunto de dados não está equilibrado, existindo uma discrepância no número de pontos relativamente aos três modelos, constatado na Tabela 5.33, pelo que foi necessário balancear os dados antes da aplicação do algoritmo *XGBoost*. Na Tabela 5.34 estão representadas as estatísticas do *XGBoost*.

Tabela 5.34: Estatísticas do algoritmo *XGBoost*, para o modelo geral.

Parâmetro	Valor
Tempo de processamento	47 min
Posição na seleção do AutoML	1
Distribuição	Multinomial
Número de árvores	74
Erro médio por classe	0,0696
Erro quadrático médio (conjunto de teste)	0,2208

Salienta-se que para o conjunto de teste o *XGBoost* obteve um erro quadrático médio de 0,2208, semelhante aos erros obtidos para cada cenário. Este obteve resultados muito satisfatórios com valores de precisão e sensibilidade muito elevados demonstrados nas Tabelas 5.35 e 5.36.

Tabela 5.35: Matriz confusão para o conjunto de treino para o algoritmo *XGBoost*, para o modelo geral.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1630	0	0	0	0/1630	**1,0**
	Suburbano	0	1545	0	0	0/1545	**1,0**
	Urbano	0	0	886	0	0/886	**1,0**
	Total	1630	1545	886	0	0/4061	-
	Precisão	**1,0**	**1,0**	**1,0**	-	-	-

Tabela 5.36: Matriz confusão para o conjunto de teste para o algoritmo *XGBoost*, para o modelo geral.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	790	33	0	0,0401	33/823	**0,96**
	Suburbano	16	712	39	0,0717	55/767	**0,93**
	Urbano	0	44	410	0,0969	44/454	**0,90**
	Total	806	789	449	0,0646	132/2044	-
	Precisão	**0,98**	**0,90**	**0,91**	-	-	-

Tal como aconteceu em cada cenário independente, para o conjunto de treino também ocorreu *overfitting*, ou seja, obteve-se uma precisão de 100% para os três modelos. Por outro lado, face ao conjunto de teste, o algoritmo obteve valores de precisão e de sensibilidade superiores a 90%, pelo que podemos concluir que o algoritmo conseguiu classificar, de forma geral, corretamente os pontos presentes no conjunto de teste. Efetivamente, o *XGBoost* viabilizou ótimos resultados de classificação dos modelos, independentemente da linha ferroviária em estudo.

Tendo sido definida a classificação para o novo conjunto de dados, efetuou-se a predição, através do algoritmo *XGBoost*, para o conjunto de dados da linha do Algarve que não foram utilizados, servindo desta forma de teste para esta predição. Na Tabela 5.37 está representada a matriz confusão da predição elaborada para este modelo geral aplicada à linha do Algarve.

Tabela 5.37: Matriz confusão da predição do modelo geral, linha do Algarve.

		Valor Previsto					
		Rural	**Suburbano**	**Urbano**	**Erro**	**Razão**	**Sensibilidade**
Valor Verdadeiro	**Rural**	1801	49	11	0,0322	60/1861	**0,97**
	Suburbano	42	1615	149	0,1058	191/1806	**0,89**
	Urbano	30	100	914	0,1245	130/1044	**0,88**
	Total	1872	1764	1074	0,0793	270/4710	-
	Precisão	**0,96**	**0,92**	**0,85**	-	-	-

Note-se que se obteve valores de precisão e sensibilidade acima dos 96% para o modelo rural, uma vez que este apresenta uma maior quantidade de pontos, sendo normal uma melhor predição para este modelo. Em relação aos outros dois modelos, a predição foi satisfatória, com valores de precisão e sensibilidade de 92% e 89%, respetivamente, para o modelo suburbano. Para o modelo urbano, atingiu-se valores de precisão e sensibilidade de 85% e 88%.

5.5 Análise de Parâmetros

Para uma classificação mais precisa é necessário definir quais os parâmetros mais relevantes no cálculo da atenuação do sinal rádio e por sua vez, na estimação de cobertura rádio do sinal. Assim, é importante analisar quais os parâmetros com maior impacto em cada cenário, com vista a perceber que parâmetros podem ter mais influência sobre a classificação de um modelo. Na Tabela 5.38 estão representados as percentagens por parâmetro de cada uma das linhas e do modelo geral.

Tabela 5.38: Valores percentuais para cada parâmetro nos diferentes cenários estudados.

Parâmetro	Cascais	Beira Baixa	Algarve	Modelo Geral
Distância (d)	50,33	34,60	31,19	36,54
Água (β)	10,32	8,40	9,26	12,31
Altura efetiva (h_{be})	10,46	20,94	8,32	10,80
Ondulação (Δ_h)	14,46	9,51	16,49	13,66
Posição na ondulação (Δ_{hm})	7,76	6,41	9,31	7,24
Obstáculo principal (v_1)	2,78	8,25	6,33	12,49
Obstáculo de 2ª ordem (v_2)	1,95	5,69	7,16	3,09
Obstáculo de 3ª ordem (v_3)	1,94	6,20	11,94	3,87

Na Figura 5.10 está representado graficamente a percentagem de cada parâmetro para cada um dos cenários considerados. Através dos resultados ilustrados no gráfico é possível avaliar os parâmetros com maior impacto em cada um dos cenários.

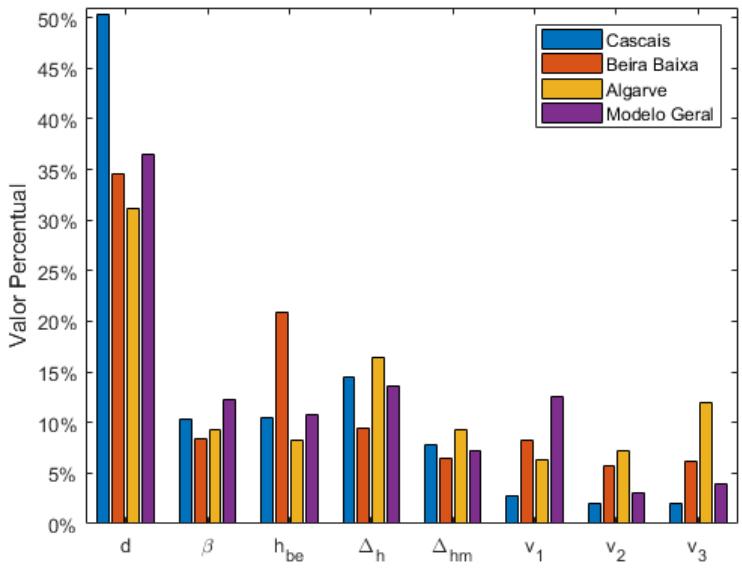


Figura 5.10: Percentagem por parâmetro, para cada cenário considerado.

Através da Figura 5.10 é possível constatar que existem diferenças no que diz respeito à relevância dos vários parâmetros nos quatro cenários considerados. Para o modelo geral, a distância entre a estação base e o terminal móvel é o parâmetro que mais se destaca, tendo maior influência na classificação do modelo. A linha de Cascais, por se tratar de uma zona suburbana com a presença de superfícies aquáticas em quase toda a linha, tem como parâmetros de maior relevância a caracterização da presença de água e dos obstáculos presentes entre a estação base e os vários pontos medidos. A linha da Beira Baixa, por se tratar de uma zona rural, tem como parâmetros de maior relevância altura efetiva e a caracterização da ondulação do terreno. Por fim, a linha do Algarve apresenta um forte impacto da ondulação do terreno sendo considerados mais relevantes os parâmetros da ondulação do terreno e a altura efetiva. Por se tratar de uma zona suburbana com vários obstáculos, é também considerado como parâmetro de maior impacto a caracterização dos obstáculos entre as estações base e os pontos medidos.

5.6 Resumo de Resultados

De modo a conseguir visualizar os resultados obtidos de uma forma simples e rápida, organizou-se na Tabela 5.39 os resultados resumidos relativos à precisão e sensibilidade do algoritmo *XGBoost* para a classificação dos modelos, em todos os cenários considerados.

Tabela 5.39: Resumo de resultados da classificação para cada cenário.

Cenários	**Precisão**			**Sensibilidade**		
	Urbano	Suburbano	Rural	Urbano	Suburbano	Rural
Cascais	99%	96%	92%	96%	97%	93%
Beira Baixa	91%	93%	98%	95%	88%	99%
Algarve	86%	89%	98%	89%	87%	98%
Modelo Geral	93%	90%	97%	90%	93%	96%

Nota-se que o algoritmo *XGBoost* foi uma solução viável para a classificação dos modelos em diversos cenários, uma vez que, de forma geral, obteve-se valores de precisão e sensibilidade superiores a 90%. A linha de Cascais foi o cenário que apresentou melhores resultados, salientando-se ótimos valores de precisão e sensibilidade para os modelos urbano e suburbano, superiores a 95%, uma vez que se trata de uma zona suburbana. A linha da Beira Baixa obteve ótimos valores para o modelo rural, dado que se trata de uma zona rural. Relativamente à linha do Algarve, esta apresentou valores um poucos mais baixos face aos modelos urbano e suburbano, dada a presença de obstáculos entre as estações base e os pontos medidos.

Por fim, considerando o modelo geral, o *XGBoost* obteve valores muito satisfatórios de precisão e sensibilidade, acima dos 90%, o que demonstra a precisão e a viabilidade de técnicas de ML para a classificação dos modelos.

Na Tabela 5.40 estão demonstrados os resultados resumidos relativos à predição elaborada pelos vários modelos, para as três linhas consideradas.

Tabela 5.40: Resumo de resultados da predição para cada cenário.

Cenários	Erros [dB]	Modelo Original			Modelo Otimizado	H2O Flow
		Urbano	Suburbano	Rural		
Cascais	ME	11,3896	7,0269	18.2626	3,9381	3,9681
	RMSE	13,4517	8,5912	20,1789	4,9169	4,9668
	ESD	7,1572	4,9429	8,5828	2,9441	2,9872
Beira Baixa	ME	24,8360	17,9717	13,9847	7,2810	7,3476
	RMSE	28,0728	20,9590	17,9774	9,0895	9,1420
	ESD	13,0855	10,7842	11,2967	5,4411	5,4396
Algarve	ME	22,3371	17,1325	20,2025	11,4022	9,4619
	RMSE	31,2286	25,6574	24,5314	14,6884	16,8459
	ESD	21,8238	19,0993	13,9158	9,2595	13,9376

Pela Tabela 5.40 nota-se que a linha do Algarve apresentou o maior valor de erro, tanto através do *H2O Flow* como através do modelo otimizado do modelo Okumura-Hata. Contudo, obteve-se uma grande otimização por parte dos dois modelos face ao modelo original. Na linha da Beira Baixa e na linha de Cascais, conseguiu-se valores muito próximos através do *H2O Flow* em comparação com o modelo otimizado, produzindo grandes otimizações face ao modelo original.

6

Conclusões

Conteúdo

Este capítulo conclui a presente dissertação, indicando o trabalho desenvolvido, os resultados obtidos e também, aspetos cruciais face ao trabalho a desenvolver no futuro.

6.1 Trabalho desenvolvido

Este trabalho teve como objetivo o desenvolvimento de um algoritmo, com recurso à interface *H2O Flow*, para efetuar a classificação do ambiente e, consequentemente, a seleção do modelo mais adequado à cobertura rádio em ambientes ferroviários. Pretende-se desta forma verificar a aplicabilidade de técnicas de ML, com recurso, ao *H2O Flow* no contexto da estimação rádio em redes GSM-R, testando nas linhas ferroviárias o funcionamento do algoritmo, de modo a perceber qual o modelo de propagação a usar em cada ponto da ferrovia, promovendo a redução do erro global da predição.

Antes de se desenvolver o algoritmo de classificação foi necessário estabelecer qual o modelo de propagação a utilizar. O modelo proposto teve como base o modelo Okumura-Hata, ao qual foram alteradas algumas características, por forma a adaptá-lo à realidade ferroviária, possibilitando aumentar a precisão da predição. De forma a avaliar o modelo de propagação desenvolvido, assim como, verificar a otimização final, foi considerado um conjunto de amostras referentes às medidas rádio recolhidas que retratam as características da propagação ferroviária.

Após ter o modelo de propagação definido, procedeu-se à classificação do tipo de ambiente antes de se proceder à estimação de cobertura. Com base nesta classificação e nos diversos parâmetros que definem o modelo, treinou-se vários conjuntos de amostras, com recurso ao *H2O Flow* com o objetivo de classificar as amostras disponíveis em três cenários distintos: a linha de Cascais, a linha da Beira Baixa e a linha do Algarve. Para estes três cenários foram estimadas diversas métricas para a quantificação do erro de predição.

Por fim, foi elaborado um teste final, em que se criou um novo modelo a partir de um novo conjunto de dados, contendo aproximadamente o mesmo número de pontos das três linhas. Para a previsão utilizou-se os restantes pontos da linha do Algarve, sem se usar os parâmetros utilizados no conjunto de treino, para que se consiga uniformizar o modelo desenvolvido, com vista a que este modelo possa ser utilizado para outras linhas ferroviárias.

6.2 Resultados

A aplicabilidade de técnicas de ML, com recurso ao *H2O Flow*, na classificação dos modelos de propagação, quando aplicados à predição de cobertura rádio em ambientes ferroviários, foi verificado através dos resultados obtidos.

No primeiro teste, foi selecionada, de forma aleatória, para cada cenário em estudo, um conjunto

de dados para treino e um conjunto de dados para teste, compostos pelos vários parâmetros que definem o modelo de propagação. A aplicação do algoritmo *XGBoost* ao conjunto de treino deu origem à classificação prévia para cada ponto da linha ferroviária, que depois foi aplicada ao conjunto de teste. Tendo a classificação definida, obteve-se a predição, através do algoritmo *XGBoost*, em todos os cenários considerados. O algoritmo *XGBoost* permitiu aproximar a predição das medidas, diminuindo as estatísticas ME, RMSE e ESD e aumentando a correlação. Como seria de esperar, a predição efetuada através da classificação dos modelos, com o auxílio do *H2O Flow* não produziu melhores resultados face à predição do modelo otimizado, contudo, atingiu valores de erro muito próximos.

O segundo teste consistiu em padronizar o modelo efetuado para cada linha, ou seja, criou-se um novo conjunto de dados que consistiu na junção das 3 linhas em estudo. Utilizou-se como teste os restantes pontos da linha do Algarve, de modo a atingir bons resultados de classificação e predição para a linha do Algarve através do conjunto de treino das três linhas juntas. Os resultados obtidos neste teste levam a crer que é possível atingir uma boa classificação e predição numa determinada linha ferroviária, utilizando um conjunto de treino de outra linha diferente.

Em conclusão, esta dissertação verificou a aplicabilidade de técnicas de ML e do algoritmo *XGBoost*, através do *H2O Flow*, em diferentes cenários, para a classificação, e consequentemente, a seleção do modelo mais adequado à cobertura rádio em ambiente ferroviários. Esta técnica é, então, uma alternativa viável a considerar em relação ao modelo de propagação Okumura-Hata otimizado, uma vez que ambas as predições foram muito semelhantes.

6.3 Trabalho Futuro

Como trabalho futuro, propõe-se o desenvolvimento de uma classificação através do *H2O Flow* de outro tipo de informação, neste caso, as classes de *clutter*. O objetivo seria criar um conjunto de dados que permitisse classificar corretamente o tipo de ambiente face às diferentes classes de *clutter* existentes, realçando qual ou quais as classes com maior influência na classificação e, por conseguinte, reduzir o erro global na predição de cobertura rádio. Para tal, é necessário estudar e testar diversas técnicas de calibração, analisar os parâmetros de informação geográfica, assim como, a informação de *clutter*, de modo a obter uma classificação mais eficaz.

Bibliografia

[1] REFER Telecom/ISEL, “Metodologia para Planeamento Rádio em GSM-R,” Lisboa, 2009.

[2] 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.

[3] 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.

[4] J. Soure, “Implementação do Sistema GSM-R na Rede Ferroviária Nacional – Projeto-piloto,” ISEC, Outubro 2013.

[5] T. Correia, “Estimação de cobertura rádio em GSM-R através de Redes Neuronais,” Tese de Mestrado, ISEL, Dezembro 2014.

[6] J. Martinho, “Calibração Automática de Modelos de Propagação em Ferrovias,” Tese de Mestrado, IST, Outubro 2016.

[7] Huawei News Room, 2015, “The Future of GSM-R, TETRA with LTE in the Railway Sector,” Acedido em: junho de 2020. [Online]. Available: https://e.huawei.com/uk/news/uk/2015/201512030948.

[8] Project Presentation, Korea Rail Network Authority, “The World’s First LTE-R for 250km/h High-Speed Railway in Republic of Korea,” 2018.

[9] P. Fraga-Lamas, T. Fernández-Caramés, and L. Castedo, “Towards the internet of smart trains: A review on industrial iot-connected railways,” *Sensors*, vol. 17, 06 2017.

[10] A. Beire, “Otimização de Modelo de Propagação utilizando Algoritmos Genéticos: Caso das Comunicações Móveis em Ferrovia,” Tese de Mestrado, ISEL, Dezembro 2013.

[11] UIC ERTMS, “FFFS for Functional Addressing,” 2006.

[12] ETSI, ETS 300 553, “European digital cellular telecommunications system (Phase 2); Layer 1. General requirements,” September 1994.

[13] L. Min and Z. Zhangdui, “Location dependent addressing using GSM-R cellular positioning,” *International Conference on Communication Technology Proceedings, ICCT*, vol. 1, IEEE, 2000.

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

[15] N. Cota, A. Serrador, N. Franco, and J. Neves, “Planeamento Rádio em GSM-R : Metodologia e Caracterização do Sinal,” URSI, Lisboa, 2009.

[16] A. R. Beire, H. Pita, and N. Cota, “Optimizing Propagation Models on Railway Communications Using Genetic Algorithms,” *Procedia Technology*, 2014.

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

[18] L. Correia, “Sistema de Comunicações Móveis - Modelos de Propagação,” Lisboa, Portugal: IST, 2007.

[19] ITU-R Recommendation P.1546, “Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3000 MHz,” October 2001.

[20] M. Weissberger, “An initial critical summary of models predicting the attenuation of radio waves by trees,” pp. 121–123, 1982.

[21] M. Hata, “Empirical Formula for Propagation Loss in Land Mobile Radio Services,” *IEEE Transactions on Vehicular Technology*, vol. 29, no. 3, August 1980.

[22] J. Deygout, “Correction factor for multiple knife-edge diffraction,” *Antennas and Propagation, IEEE Transactions on*, vol. 39, no. 8, August 1991.

[23] Recommendation ITU-R P.526-12, *Propagation by diffraction*, January 2012.

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

[25] V. Santos, “Modelo de data mining para detecção de tumores em exames de rastreio,” Tese de Mestrado, ISEL, Setembro 2013.

[26] 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.

[27] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[28] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal ofArtificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[29] J. Wang, M. Xu, H. Wang, and J. Zhang, “Classification of imbalanced data by using the smote algorithm and locally linear embedding,” in *2006 8th international Conference on Signal Processing*, vol. 3, 2006.

[30] J. Larsen and C. Goutte, “On optimal data split for generalization estimation and model selection,” *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 225–234, 1999.

[31] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *International Joint Conference on Articial Intelligence (IJCAI)*, vol. 14, pp. 1137–1143, 1995.

[32] T. Nykodym, T. Kraljevic, N. Hussami, A. Rao, and A. Wang, “Generalized Linear Modeling with H2O,” *H2O.ai, Inc*, 2016.

[33] “An End-to-End Guide to Understand the Math behind XGBoost,” *Analytics Vidhya, 2018*, Acedido em: julho de 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.