TÉCNICO LISBOA

Calibração Automática de Modelos de Propagação em Ferrovias

João Pedro Rebelo Martinho

Dissertação de Mestrado em Engenharia Eletrotécnica e de Computadores

Orientadores: Prof. António José Castelo Branco Rodrigues

Prof. Nuno Cota

Prof. Hélder Pita

Júri

Presidente: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Orientador: Prof. António José Castelo Branco Rodrigues

Vogal: Prof. António João Nuno Serrador

Novembro 2016

Agradecimentos

Ao longo da realização desta dissertação de mestrado contei com inúmeros apoios e incentivos de professores, familiares e amigos, aos quais não posso deixar de agradecer.

Ao Prof. António Rodrigues, pela sua orientação, apoio e motivação.

Ao Prof. Nuno Cota, pelas suas ideias e críticas construtivas, bem como pelos recursos que me disponibilizou para a realização deste trabalho.

Ao Prof. Hélder Pita, pelo saber que me transmitiu, pela disponibilidade e pela colaboração no elucidar de questões inerentes à elaboração deste trabalho.

À Rita Beire, pela sua ajuda na introdução ao tema desta dissertação, bem como pela sua disponibilidade na discussão de soluções.

A todos os meus amigos.

Em especial ao Gonçalo Alho, pelo apoio constante durante a fase de elaboração deste trabalho.

Ao Gonçalo Beirão, pela disciplina que me transmitiu.

À Carina, pelo amor e compreensão.

E, por último, aos elementos da minha família, por toda a força que me transmitiram, amor e apoio incondicional, e por continuarem a ser o modelo de coragem, no qual eu me inspiro. A eles, dedico este trabalho.

Resumo

A previsão de cobertura do sinal de rádio é uma etapa fundamental no planeamento de uma rede rádio de comunicações móveis. Em ambiente ferroviário, essa previsão exige uma precisão e rigor superiores comparativamente aos das redes públicas, dadas as limitações decorrentes dos requisitos de segurança. Torna-se, portanto, essencial a calibração dos modelos de propagação utilizados para os diferentes tipos de ambientes e características, presentes na ferrovia. O processo de ajuste dos parâmetros de um dado modelo, implica o recurso a técnicas de otimização automática, que a partir de amostras de teste, produzem soluções de parâmetros que minimizam o erro existente no ajuste das curvas.

A utilização de algoritmos genéticos demonstrou-se ser válida, na otimização de parâmetros de calibração de modelos de propagação, quando aplicados à predição de cobertura rádio em ferrovias. No entanto, foi destacada a dificuldade em obter uma otimização global, em termos de modulação do comportamento do sinal, para os diferentes ambientes, bem como a não utilização da informação de clutter.

O desenvolvimento de um algoritmo de clustering, capaz de agrupar um conjunto de medidas caracterizantes de um dado cenário ferroviário, em subconjuntos que partilhem semelhanças morfológicas, possibilita a otimização dos parâmetros de configuração, para cada grupo obtido.

Através da calibração de modelos de propagação, para os diferentes tipos de ambientes e características, combinando técnicas de Data Mining, como algoritmos genéticos e de clustering, produzem-se soluções de parâmetros que minimizam, em 10%, o desvio padrão do erro de predição, comparativamente aos valores obtidos através de uma otimização global.

Palavras-chave: *Clustering*; *Clutter*; Algoritmos Genéticos; modelos de propagação; comunicações rádio em ferrovias.

Abstract

The radio signal coverage prediction is one of the key steps in planning a radio mobile communication network. In rail environment, this estimate requires precision and higher accuracy compared to public networks, given the constraints arising from safety requirements. It is therefore essential to calibrate the propagation models used for different kinds of environments and characteristics of the railroad. The process of adjusting the parameters of a given model requires the use of automatic optimization techniques, which from test samples, produce parameters solutions that minimize the error in the setting of the curves.

The use of genetic algorithms has shown to be valid on the optimization of calibration parameters in propagation models, when applied to radio coverage prediction in railways. However, it highlighted the difficulty in obtaining an overall optimization in terms of signal modulation behavior for different types of environments as well as the non-utilization of clutter information.

The development of a clustering algorithm, able to group a set of measures of a given rail scenario, into subsets that, in the context of spreading radio, share geographic similarities / morphological characteristics, enables the optimization of configuration parameters, for each of the obtained groups.

The calibration of propagation models for the different types of environments and characteristics, by combining techniques of Data Mining, such as Genetic Algorithms and Clustering, produce solutions of parameters that decreases the standard deviation of radio prediction error by 10%, compared to values obtained through a global optimization.

Keywords: Clustering; Clutter; Genetic Algorithms; calibration of propagation models; railway communications.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AuC	*Authentication Center*
ADN	Ácido Desoxirribonucleico
AG	Algoritmos Genéticos
ANACOM	Autoridade Nacional para as Comunicações
AP	Algoritmo Proposto
BSC	*Base Station Controllers*
BSS	*Base Station Sub-System*
BTS	*Base Transceiver Stations*
EIR	*Equipment Identity Register*
EIRENE	*European Integrated Railway Radio Enhanced Network*
EM	*Expectation Maximization*
ESD	*Error Standard Deviation*
ETL	*Export Transform and Load*
ETSI	*European Telecommunications Standard Institute*
GSM	*Global System for Mobile Communications*
GSM-R	*Global System for Mobile Communications Railways*
HLR	*Home Location Register*
KMP	*K-Means* Personalizado
LDA	*Location Dependent Addressing*
MND	*Mutual Neigbor Distance*
MORAINE	*MObile radio for RAilway Networks in Europe*
ME	*Mean Error*
MS	*Mobile Station*
MSC	*Mobile services Switching Centre*
NSS	*Network Sub-System*

OMC	*Operation and Maintenance Center*
QoS	*Quality of Service*
RE	Coeficiente de correlação
RMSE	*Root Mean Square Error*
SIM	*Subscriber Identity Module*
SOM	*Self Organizing Maps*
TRX	*Transceivers*
UIC	*Union Internationale du Chemin-de-Fer*
VLR	*Visitor Location Register*

Lista de Equações

Capítulo 1

Introdução

O presente capítulo fornece uma visão global da dissertação, onde é abordado o enquadramento deste projeto, a sua motivação e objetivos, bem como a sua estrutura.

1.1 Enquadramento

A livre circulação transfronteiriça em caminhos-de-ferro confrontou-se com grandes problemas ao longo dos tempos, os quais se deviam ao facto de a ferrovia utilizar predominantemente sistemas proprietários, fechados e não interoperáveis. Surgiu, assim, a 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, o que determinou a conceção de um sistema de comunicações móveis especifico para a rede ferroviária.

Em 1992 [1], a entidade UIC *(Union Internationale du Chemin-de-Fer*), iniciou o desenvolvimento de um projeto europeu denominado EIRENE (*European Integrated Railway Radio Enhanced Network*) [2], que originou um conjunto de especificações para a implementação da tecnologia GSM-R (*Global System for Mobile Communications – Railway*), de modo a responder ao objetivo de uniformização tecnológica [3].

Estas especificações, tendo por base a norma GSM, foram validadas pelo MORANE (*MObile radio for RAilway Networks in Europe*), e aumentaram os requisitos em termos de qualidade de serviço das redes rádio. A escolha do GSM como tecnologia base utilizada para o desenvolvimento da nova geração de sistemas de rádio comunicações, deveu-se à sua grande robustez e fiabilidade ao nível da transmissão rádio, acrescidas de motivos técnicos e económicos. Existem, no entanto, diferenças entre os sistemas GSM e GSM-R, no que respeita à componente rádio. As principais relacionam-se com o facto de o sistema GSM-R permitir velocidades até $500\,km/h$, suportando *handovers* e seleção de células mais rápida do que na norma original. Por outro lado, foram adicionadas novas funcionalidades ao sistema de modo a permitir uma utilização mais flexível e aplicada às comunicações ferroviárias. A rede GSM-R diferencia-se de outras redes móveis pelo facto de ser uma rede fechada, operando em frequências exclusivas, com equipamentos terminais próprios e funcionalidades específicas para a exploração ferroviária.

Relativamente ao espectro de frequências, em 1995 o ETSI (*European Telecommunications Standard Institute*) [4] reservou duas faixas de frequência entre $876-880\,MHz$ (*uplink*) e $921-925\,MHz$ (*downlink*) para a utilização pelos sistemas EIRENE.

Em Portugal, desde o ano 2000, a entidade responsável pela gestão das comunicações ferroviárias é a REFER Telecom, ficando esta com a responsabilidade de aplicar o sistema europeu à ferrovia portuguesa [5].

Em Fevereiro de 2008, a REFER delegou na REFER Telecom a prossecução dos estudos, projetos e a obtenção de licenciamento junto da ANACOM (Autoridade Nacional para as Comunicações) [6], tendo em vista a implementação de uma rede de comunicações rádio GSM-R, a instalar nas principais linhas da rede ferroviária convencional, e também nas futuras linhas de alta velocidade [5].

Em 2012 [7], a REFER Telecom, respeitando as normas de interoperabilidade ferroviária da Comunidade Europeia, efetuou a primeira chamada sobre a rede GSM-R num troço da linha ferroviária de Cascais.

1.2 Motivação e Objetivos

A previsão de cobertura do sinal de rádio é uma das principais etapas no planeamento de uma rede rádio de comunicações móveis. Quando se tratam de comunicações móveis em ferrovias, essa estimativa exige uma precisão e rigor superiores comparativamente à das redes públicas, dadas as limitações decorrentes dos requisitos de segurança. Torna-se, por esta razão, essencial a calibração dos modelos de propagação utilizados para os diferentes tipos de ambientes e características, presentes na ferrovia. No entanto, este ajuste de parâmetros de um dado modelo, tendo como base métodos iterativos lineares tradicionais, é um processo que pode tornar-se muito complexo, dado o número de variáveis envolvidas e a dependência entre elas. Este processo implica a necessidade do recurso a técnicas de otimização automática (AG – Algoritmos Genéticos), que a partir de amostras de teste, produzem soluções de parâmetros que minimizam o erro existente no ajuste das curvas.

Tendo por base as metodologias propostas nos trabalhos [8] e [9], bem como os respetivos resultados obtidos, os quais se apresentam descritos no Capítulo 2.6 – Estado da Arte, é proposto associar as vantagens da utilização de modelos de propagação com base na predição de cobertura rádio, a técnicas de agrupamento (*clustering*) que permitam obter previamente, uma classificação dos tipos de ambiente, de forma a reduzir o erro global na predição. Para tal, é necessário estudar e testar diversas técnicas de classificação, analisar os parâmetros de caracterização da localização geográfica, bem como a informação de *clutter*, de modo a obter uma classificação mais eficaz e determinar o número e as características finais dos tipos de ambientes.

Para os diversos tipos de ambientes / classes, utilizam-se os modelos de estimação mais apropriados, incluindo o modelo de Okumura-Hata [10] e [11], o qual demonstrou bons resultados na predição de cobertura rádio em ferrovias [12], e utilizando a informação de *clutter* para melhorar a precisão dos modelos.

1.3 Estrutura

Este relatório, realizado no âmbito da disciplina de Dissertação de Mestrado em Engenharia Eletrotécnica e de Computadores, é composto por 5 capítulos.

No presente capítulo, é fornecido o enquadramento tecnológico deste projeto, bem como o que motivou a sua realização, o seu objetivo e a sua estrutura.

O Capítulo 2 apresenta os fundamentos teóricos da área científica do projeto, nos quais se encontra uma descrição global do sistema GSM-R, as características e requisitos da propagação em ferrovias, a informação de *clutter*, o princípio de funcionamento dos algoritmos genéticos, os conceitos fundamentais de *clustering* e a explicação relativa às técnicas de *clustering* escolhidas. É também

apresentada uma revisão do Estado da Arte.

O Capítulo 3 descreve a metodologia utilizada para a realização do AP (Algoritmo Proposto), isto é, a implementação de uma calibração automática de modelos de propagação, para os diferentes tipos de ambientes e características, utilizando algoritmos genéticos e de *clustering*, respetivamente.

O Capítulo 4 fornece a configuração final do algoritmo desenvolvido, bem como a análise dos resultados obtidos.

O Capítulo 5 conclui a presente dissertação, fornecendo aspetos fundamentais relativos ao trabalho a desenvolver no futuro.

.

Capítulo 2

Fundamentos Teóricos

Este capítulo aborda os conceitos teóricos das tecnologias envolvidas no presente trabalho, nomeadamente o sistema GSM-R, os modelos de propagação e respetivos requisitos, o *clutter*, os AG e o *clustering.* O capítulo é finalizado com uma revisão do Estado da Arte.

2.1 GSM-R

2.1.1 Introdução

Os métodos de cobertura e otimização do sistema GSM, cuja maturidade de documentação é inquestionável, são utilizados por todos os operadores de redes públicas de comunicações móveis. Os objetivos e considerações, estabelecidos por esta metodologia, encontram-se afastados da realidade nas comunicações em caminhos-de-ferro, quer em termos de QoS (*Quality of Service*), quer em termos de arquitetura, cobertura rádio, etc. Conclui-se, por esta razão, que os métodos acordados no planeamento e cobertura de rede rádio GSM não são favoráveis na comunicação celular em ferrovias.

Posteriormente às recomendações das respetivas instituições de regulamentação, referidas anteriormente, foi aplicado um investimento significativo na migração da rede analógica de comunicações móveis dos operadores, para GSM-R. Esta norma, tendo como base uma tecnologia robusta, segura e de acesso rápido, satisfaz as necessidades especiais dos operadores de infraestruturas ferroviárias, em termos de comunicações profissionais de voz e dados.

2.1.2 Arquitetura

Tal como a totalidade das arquiteturas projetadas em comunicações móveis, a do sistema GSM-R é planeada visando a diminuição da complexidade das respetivas estações base de transmissão, prevenindo o pior caso, sendo este, por exemplo, a possibilidade de criação ou sectorização de células, cujo preço é pouco acessível. A gestão e manutenção centralizada, bem como a interligação a outras redes, são características capitais desta rede. A Figura 1 ilustra os principais constituintes da arquitetura de uma rede GSM-R, respeitando a norma [13].

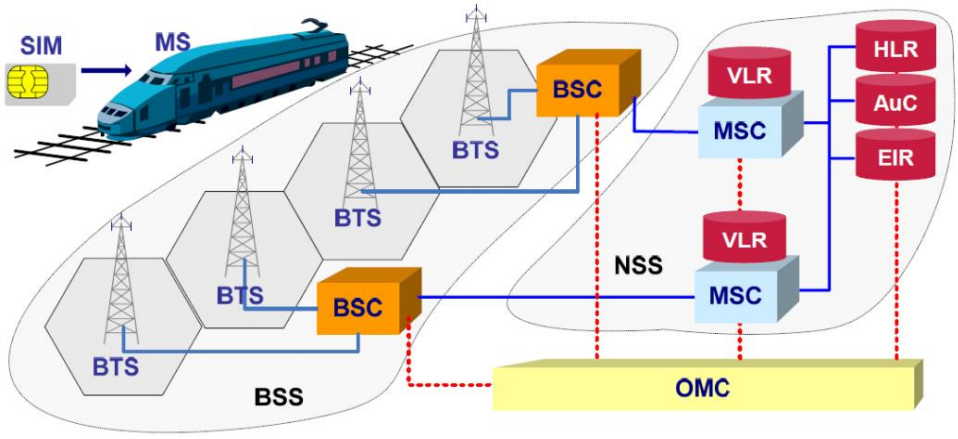

Figura 1 – Arquitetura de uma rede GSM-R [14].

Na extremidade esquerda da Figura 1 encontra-se o equipamento terminal, designado por MS (*Mobile Station*), cuja função é a ligação à rede de acesso rádio através da interface rádio. Este elemento inclui um cartão inteligente, SIM (*Subscriber Identity Module*), o qual contém informação específica de um dado assinante.

A rede do operador é repartida em dois subsistemas independentes e de funcionamento distinto, o BSS (*Base Station Sub-System*) e o NSS (*Network Sub-System*). O NSS, responsável pelo controlo de chamadas, é constituído por MSCs (*Mobile services Switching Centre*), os quais se encontram interligados a um VLR (*Visitor Location Register*). Os VLR são equipamentos que possuem bases de dados com a informação temporária de um determinado assinante e estão ligados a uma área de serviço abrangida pelos MSCs. A gestão dos perfis dos assinantes ligados à rede é realizada por um conjunto de bases de dados, designadas por HLR (*Home Location Register*), por outro lado, as bases de dados intituladas de AuC (*Authentication Center*) e EIR (*Equipment Identity Register*) são responsáveis pela gestão do mecanismo de segurança e dos equipamentos terminais, respetivamente.

O BSS, responsável por todas as funcionalidades referentes à transmissão, é constituído por BSCs (*Base Station Controllers*), os quais têm a função de controlar desde uma a mais BTSs (*Base Transceiver Stations*), que, por sua vez, são compostas por um dado número de TRXs (*Transceivers*).

Sendo esta arquitetura tradicionalmente idêntica à do GSM público, estes dois subsistemas estão interligados através da interface A de GSM, constituída por canais de $64kbps$.

O OMC (*Operation and Maintenance Center*) realiza a monitorização da totalidade da rede, abrangendo a sua configuração, monitorização de desempenho, gestão de assinantes, etc [14].

2.1.3 Cobertura

A definição dos níveis mínimos de cobertura é uma das distinções mais significativas, das especificações da rede GSM para a GSM-R. Em GSM-R, os níveis mínimos de cobertura são dependentes da velocidade e do tipo de informação transmitida. Os valores apresentados na Tabela 1 são definidos considerando a situação de rádio de cabine, com uma antena, considerada isotrópica, instalada a $4m$ de altura [15].

TIPO	VALOR MÍNIMO	UTILIZAÇÃO	VELOCIDADE
OBRIGATÓRIO	$-98\ dBm$	Voz e dados de baixa segurança	---
OBRIGATÓRIO	$-95dBm$	ETCS níveis 2/3	$\leq\ 220km/h$
RECOMENDADO	$-92dBm$	ETCS níveis 2/3	$\geq\ 280km/h$

Tabela 1 – Níveis mínimos de cobertura dependendo da velocidade e do tipo de informação transmitida.

Nos sistemas GSM, a probabilidade de cobertura trata-se da média da cobertura de toda a região. Por outro lado, em GSM-R, tal como se pode verificar, através da Figura 2, os valores mínimos de cobertura devem respeitar uma probabilidade de cobertura superior a 95%, a cada $100m$ de segmento de ferrovia. Voltando-se assim, a verificar um nível de exigência muito superior, relativamente aos requisitos de cobertura para sistemas GSM [15].

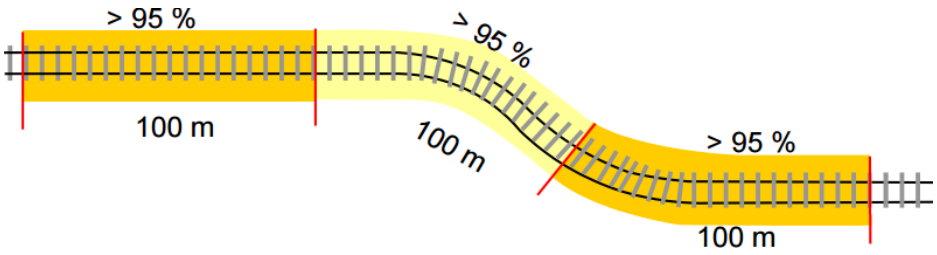

Figura 2 – Probabilidade de cobertura por 100m de linha férrea.

LDA (*Location Dependent Addressing*) é o nome de uma das funcionalidades mais importantes no dimensionamento de cobertura rádio em GSM-R. O seu propósito, dependendo da localização do utilizador, é o de atribuir um endereço lógico a uma dada função (controlador). Assim sendo, a estrutura de comando de circulação da ferrovia poderá impor que, a nível celular, os limites das células sejam coerentes com a estrutura de identificadores definidos (Figura 3) [15].

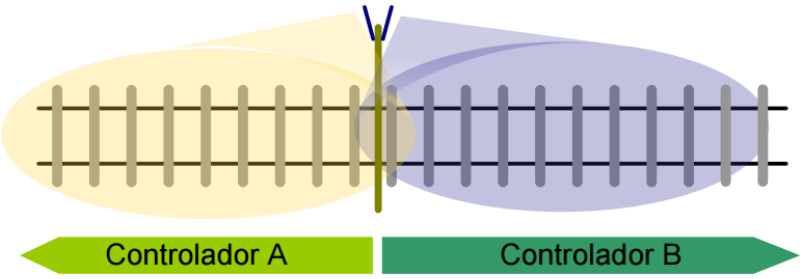

Figura 3 – Atribuição de um endereço lógico a um dado controlador.

A alteração de controlador de forma concisa, em determinadas zonas da linha férrea, torna-se, por isso, necessária. Como tal, o *handover* deve ser forçado na área de comutação de endereço, o que implica a colocação de uma estação bi-setorizada na respetiva zona de comutação, assegurando, assim, uma maior precisão de *handover* no local. Esta funcionalidade é realizada através de parâmetros específicos do sistema GSM.

2.2 Propagação em Ferrovias

2.2.1 Introdução

O presente subcapítulo é, na ausência de outra referência, baseado em [16].

O cálculo da atenuação na propagação de sinal rádio é um dos passos fundamentais no projeto de qualquer rede sem fios. 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.

A estimação de atenuação do sinal pode ser feita através de uma abordagem determinística, ou seja, utilizar a modelação matemática dos mecanismos de propagação considerados para determinar o comportamento do sinal, ou através de uma abordagem estatística. Devido ao elevado número de parâmetros a considerar, é impossível calcular ao certo a atenuação do sinal em cenários reais com obstáculos, diferentes ambientes, terrenos irregulares, entre outros fatores.

Para 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.

A maioria dos modelos fornecem a mediana ou os valores médios do sinal, por isso, é necessário conhecer as estatísticas do sinal, de modo a determinar a sua variação. O problema da estimativa do sinal não pode ser exclusivamente abordado de uma maneira determinística. Uma estimativa correta do sinal, e o desenvolvimento de modelos relativos à mesma, implica o conhecimento sobre todos os fatores que influenciam a propagação numa dada comunicação móvel.

Os modelos podem ser divididos em duas categorias: empírica e teórica. Os modelos empíricos são baseados em medidas, visando alcançar as melhores equações de ajuste. Estes têm como vantagem, a contabilização de todos os fatores que influenciam a propagação, no entanto, necessitam de validação em ambientes diferentes dos que foram utilizados para estabelecer o respetivo modelo.

Os modelos teóricos são uma aproximação à realidade, não tendo em conta todos os fatores e permitindo uma alteração fácil dos respetivos parâmetros. Estes demonstram uma elevada dependência da resolução do banco de dados geográfico.

Atualmente os modelos contemplam ambas as abordagens. A utilização de um dado modelo requer uma classificação prévia do ambiente, o qual se divide em três categorias: rural, suburbana e urbana. Esta classificação tem em consideração vários parâmetros, tais como, a ondulação do terreno, a densidade da vegetação, a altura e densidade dos edifícios, bem como a densidade de áreas abertas e de água.

2.2.2 Modelo Okumura-Hata

Entre 1962 e 1965, na cidade de Tóquio, foram realizados dois grandes testes, com várias estações emissoras, transmitindo em várias bandas, numa grande variedade de ambientes de propagação, com o objetivo de explorar as maiores influências na propagação das ondas. Em 1968, Okumura propôs um modelo empírico baseado em medidas na banda de $150 - 2000\text{MHz}$ e apresentou o respetivo resultado em forma de curvas. Masaharu Hata, em 1980, publicou equações estabelecidas numa banda mais restrita, que aproximam algumas dessas curvas.

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:

- Área aberta: quando não existem obstáculos numa região de 300 a $400m$, diante do terminal móvel;
- Área suburbana: quando existem alguns obstáculos, com pouca densidade, na região próxima do terminal móvel;
- Área urbana: quando se trata de uma região de alta densidade de construção, com edifícios constituídos por mais que 2 andares.

Apesar de o modelo original ser válido para os intervalos que se apresentam na coluna esquerda da Tabela 2, posteriormente à formulação de Hata, este ficou mais restrito (coluna à direita).

$\boldsymbol{f}\ [\boldsymbol{MHz}]$	$[\mathbf{150, 2000}]$	$[\mathbf{150, 1500}]$
$\boldsymbol{d}\ [\boldsymbol{km}]$	$[\mathbf{1, 100}]$	$[\mathbf{1, 20}]$
$\boldsymbol{h\_be}\ [\boldsymbol{m}]$	$[\mathbf{30, 1000}]$	$[\mathbf{30, 200}]$
$\boldsymbol{h\_m}\ [\boldsymbol{m}]$	$[\mathbf{1, 10}]$	$[\mathbf{1, 10}]$

Tabela 2 – Intervalos para os quais o modelo Okumura-Hata é válido.

2.2.2.1 Altura Efetiva

A Figura 4 ilustra como é que a altura efetiva da antena da estação base, h_{be}, é determinada, onde h_{bs}, é a altura da antena da estação base, em relação ao solo, h_b, a altura da antena da referida estação base e h_{ga} é a altura do terreno relativamente ao solo.

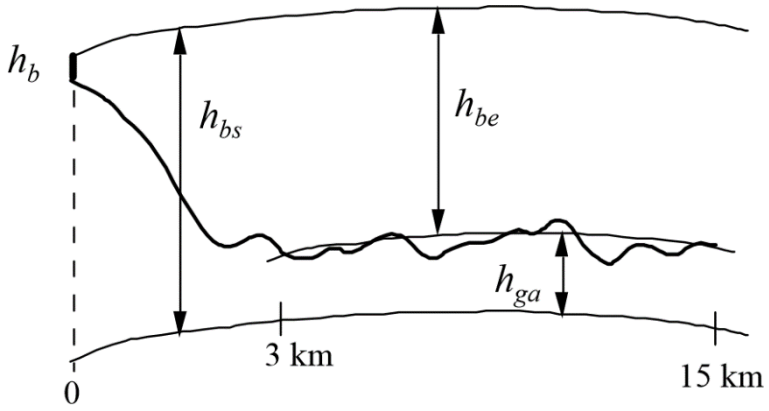

Figura 4 – Altura efetiva da antena da estação base.

2.2.2.2 Atenuação

O modelo fornece o valor mediano da atenuação, o qual é influenciado por parâmetros como a frequência, f, a distância do terminal móvel à estação base, d, e a altura da antena do terminal móvel, h_m. O valor mediano da atenuação dado pela seguinte equação:

$$L_{p[dB]} = 69.55 + 26.16\ log\left(f_{[MHz]}\right) - 13.82\ log\left(h_{be[m]}\right) + \left[44.90 - 6.55\ log\left(h_{be[m]}\right)\right] log\left(d_{[km]}\right) - H_{mu[dB]}(h_m, f) - \sum factores\ correctivos \quad (1)$$

onde, para um ambiente suburbano básico:

$$H_{mu[dB]} = \left[1.10\ log\left(f_{[MHz]}\right) - 0.70\right]h_{m[m]} - \left[1.56\ log\left(f_{[MHz]}\right) - 0.80\right]. \quad (2)$$

2.2.2.1 Fatores Corretivos

Ruas radiais

Este fator corretivo é considerado tendo em conta a orientação entre a antena e a linha da ferrovia, sendo que quando esta é igual em ambos os elementos, o valor da sua atenuação é dado por:

$$K_{ac}(\theta)_{[dB]} = 2.1\ log\left(d_{[km]}\right) - 6.3. \quad (3)$$

Ondulação do terreno

A altura da ondulação do terreno, Δh_b, tal como se pode observar na Figura 5, é obtida através da diferença entre o percentil 10 e o percentil 90 da respetiva altura do terreno.

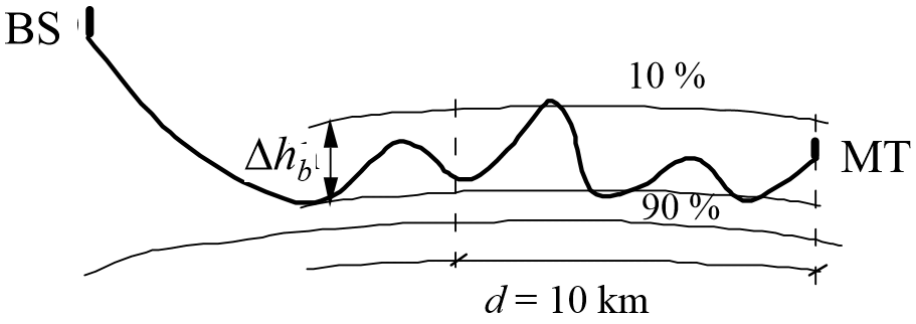

Figura 5 – Ondulação do terreno.

A atenuação desta ondulação é dada por:

$$K_{th}(\Delta h_b)_{[dB]} = -3\ log^2(\Delta h_{b[m]}) - 0.5\ log(\Delta h_{b[m]}) + 4.5 \quad (4)$$

No entanto, quando se tem conhecimento da localização do terminal móvel, na referida ondulação do terreno, esta atenuação é obtida através da seguinte equação:

$$K_{hp}(\Delta h_{b[m]})_{[dB]} = -2\ log^2(\Delta h_{b[m]}) + 16\ log(\Delta h_{b[m]}) - 12 \quad (5)$$

onde Δh_b é, neste caso, 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.

Declive médio do terreno

A Figura 6 apresenta a identificação do ângulo que representa o declive médio do terreno.

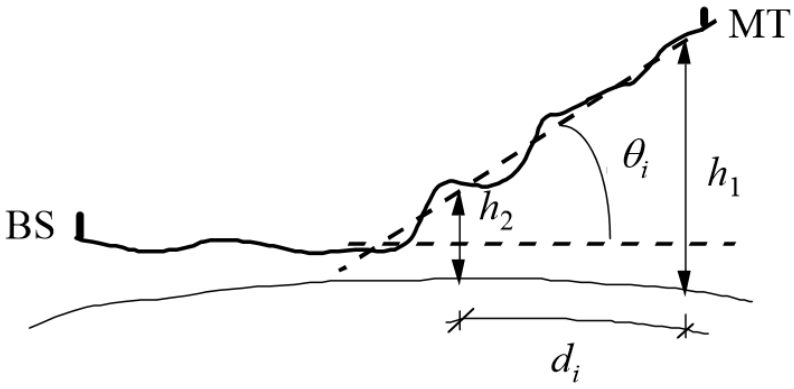

Figura 6 – Declive médio do terreno.

O fator corretivo que tem em conta esta característica do terreno é dado por:

$$K_{sp}(\theta)_{[dB]} = \begin{cases} -0.0025\ \theta^2_{[mrad]} + 0.204\ \theta_{[mrad]},\ (d < 10km) \\ -0.648\ \theta^{1.09}_{[mrad]},\ (d < 30km) \\ -0.0012\ \theta^2_{[mrad]} + 0.840\ \theta_{[mrad]},\ (d < 60km) \end{cases} \quad (6)$$

Trajetos mistos

O parâmetro $\beta = \frac{d_s}{d}$ descreve a relação entre a distância do percurso onde existe água, d_s e a distância total do percurso, entre a estação base e o terminal móvel. A Figura 7 apresenta os dois cenários possíveis, em termos de trajetos mistos.

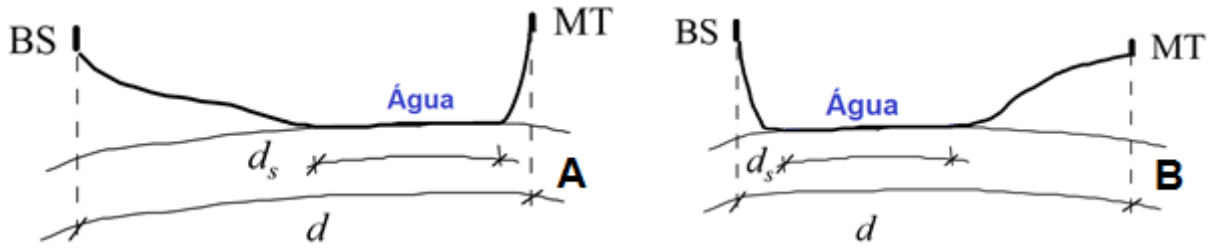

Figura 7 – Trajetos mistos.

O fator corretivo que suporta este tipo de percurso é dado por:

$$K_{mp}(\beta)_{[dB]} = \begin{cases} \begin{cases} -11.9\beta^2 + 4.7\beta\ , & d > 60km \\ -7.8\beta^2 + 5.6\beta\ , & d < 30km \end{cases} & A \\ \begin{cases} -12.4\beta^2 + 27.2\beta\ , & d > 60km \\ -8.0\beta^2 + 19.0\beta\ , & d < 30km \end{cases} & B \end{cases} \quad (7)$$

Onde A (cenário ilustrado à esquerda, na Figura 7) considera a situação em que a localização da água se encontra longe da estação base, relativamente ao terminal móvel, sendo B a situação inversa.

2.2.3 Modelo Deygout

Visto que o modelo do Okumura-Hata não contabiliza as perdas por difração, devido aos obstáculos, para efeitos da predição de cobertura rádio em GSM-R, estas perdas adicionais são determinadas através da utilização de um modelo que consiste numa aproximação, admitindo que os obstáculos têm uma geometria em lâmina, conforme o modelo considerado na recomendação P.526 [17].

O modelo de Deygout [18] deve ser usado quando as dimensões do obstáculo são muito superiores ao comprimento de onda, sendo a sua geometria a apresentada na figura seguinte.

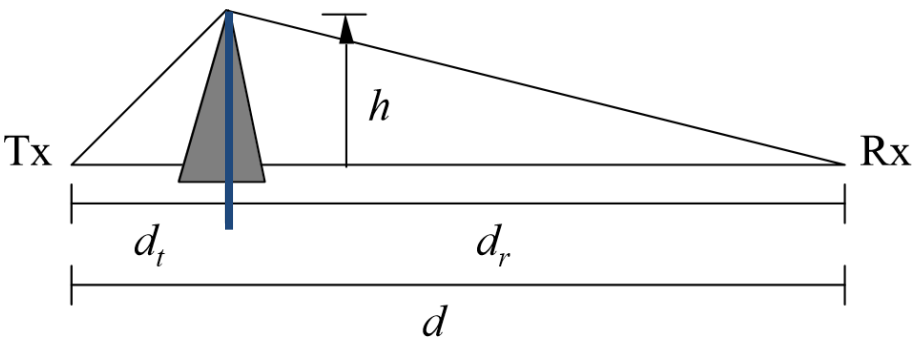

Figura 8 – Geometria do método Deygout.

A atenuação é dada por: $L_{ke[dB]} = 6.4 + 20\log\left(v + \sqrt{v^2 + 1}\right), v > -0.7$ (8)

sendo $v = h\sqrt{\frac{2d}{\lambda d_t d_r}}$ (9)

é o parâmetro definido por Fresnel-Kirchhoff, onde h é a altura do obstáculo, quer seja esta acima (sinal positivo) ou abaixo (sinal negativo) do raio direto entre as antenas de transmissão e receção, d é a distância total da ligação, d_t é a distância entre a estação base e o obstáculo, d_r é a distância entre o obstáculo e o terminal móvel e λ é o comprimento de onda.

2.3 Informação de *Clutter*

2.3.1 Introdução

O presente subcapítulo é, na ausência de outra referência, baseado em [19], [20] e [21].

Quando ondas de rádio se propagam no vácuo, isto é, em contexto de propagação em espaço livre, os únicos fatores a considerar são a frequência e a distância. Para todos os outros casos é necessário considerar o ambiente em que a onda de rádio se propaga, seja este caracterizado por gases, chuva, neve, areia, qualquer tipo de edifícios, vegetação, colinas, corpos de água, etc.

Os dados meteorológicos e geoclimáticos, juntamente com as características morfológicas do terreno, superfície e base de dados *land use* (base de dados de *clutter*) são recursos que podem ser usados por modelos de propagação, para melhorar a eficácia da previsão de perdas entre as antenas de transmissão e receção. Cada entidade física que um sinal de rádio encontra, depois de sair da antena de transmissão, afeta a força e a direção do sinal. As entidades físicas que afetam o sinal podem ser agrupadas em quatro categorias:

1. A atmosfera (ou outros meios gasosos) promove a refração e a dispersão das ondas de rádio; a refração provoca uma mudança na direção da onda de rádio, enquanto a dispersão geralmente enfraquece a onda.

2. As características do terreno (colinas e montanhas) bloqueiam as ondas de rádio, obrigando-as a dispersar sobre o topo ou em torno dos lados, enfraquecendo, assim, o sinal. As ondas de rádio também sofrem reflecção e dispersão aquando em contacto com a superfície do terreno.

3. Tal como o terreno, estruturas como edifícios, casas, torres, etc., bloqueiam as ondas de rádio. As ondas são refletidas e dispersas em torno das estruturas dos edifícios.

4. Folhas e ramos de árvores, assim como outros tipos de vegetação, também enfraquecem as

ondas de rádio, dispersando-as, provocando um efeito semelhante, causado por edifícios.

2.3.2 Classes de *Clutter*

Clutter refere-se a uma classificação das características superficiais que influenciam a propagação de ondas rádio. O *clutter* é geralmente produzido a partir de imagens de satélite multiespectrais onde classes distintas de características superficiais podem ser delineadas através de homogeneidade espectral, entre outras características. Para certas classes como água, florestas e terras de agricultura, torna-se necessário o emprego de técnicas de classificação supervisionada, sendo este um processo iterativo. Os resultados são verificados e reverificados de modo a obter uma classificação de elevada precisão. A maioria das características do ambiente construído, porém, são classificadas manualmente, utilizando o método de foto-interpretação.

A Figura 9 ilustra um cenário representado por uma comunicação entre uma BTS e dois móveis. Ambos os móveis estão situados à mesma distância da BTS, no entanto, um encontra-se atrás de um edificio e outro, numa área aberta perto de uma lagoa.

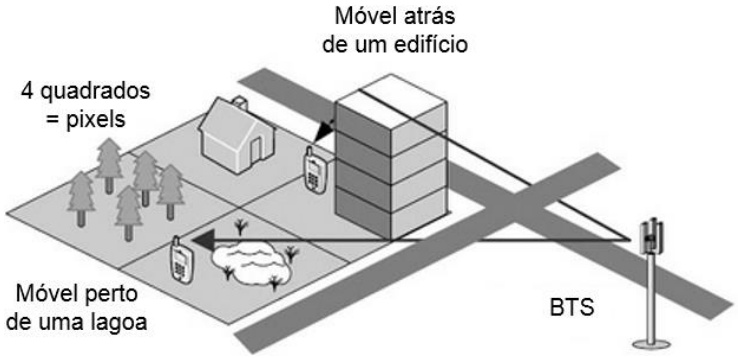

Figura 9 – Comunicação entre uma BTS e dois móveis, afectada pelo *clutter* local.

Nesta situação, os sinais recebidos pelos móveis são distintos, devido ao facto de serem afectados pelo *clutter* local. De modo a contabilizar o descrito no parágrafo anterior e as características do presente cenário, utiliza-se a informação de *clutter*, na qual cada pixel (quadrado) está associado a um código que define as características desse mesmo quadrado.

A Tabela 3 apresenta a descrição das classes de *clutter*, acompanhadas pelo respectivo código, utilizadas na implementação da estratégia proposta.

1	*Sea*	Áreas de águas costeiras, incluindo oceanos, baías e estuários.
2	*Inland water*	Áreas de água aberta permanente; corpos de água naturais e feitos pelo Homem, que podem ser estáticos ou fluidos (rios, barragens, reservatórios e lagos).

3	*Wetland*	Áreas de terra aberta ou de vegetação, periodicamente inundada ou coberta com água estagnada superficial.
4	*Barren*	Áreas que não contêm vegetação.
5	*Grass/Agriculture*	Campos agrícolas tipicamente caracterizados pela sua forma geométrica e usados para a produção de culturas anuais (pomares, vinhas, pastagens para o gado, colheitas de feno etc); terra cultivada, lotes não desenvolvidos, parques e campos de golfe.
6	*Rangeland*	Vegetação pouco densa e dispersa e áreas de relva mista.
7	*Woodland*	Sem continuidade e de densidade mista (30% - 60% de cobertura formada por copas de árvores); com árvores geralmente menores do que 5 metros.
8	*Forest*	Cobertura contínua de plantação e/ou espécies de árvores nativas com uma altura média superior a 5 metros.
9	*Village*	Pequenas áreas construídas dentro de ambiente rural, que incluem tanto classes suburbanas como urbanas.
10	*Suburban*	Áreas residenciais, principalmente compostas por casas de um andar, com uma cobertura das árvores média (<30%). As superfícies impermeáveis são responsáveis por 20% a 49% da cobertura total.
11	*Dense Suburban*	Áreas de densas estruturas residenciais misturadas com zonas comerciais e com estruturas residênciais de 2 a 4 andares. Cobertura de árvores (<5%). Superfícies impermeáveis são responsáveis por 50% a 79% da cobertura total.
12	*Urban*	Áreas extremamente desenvolvidas consistindo num misto de estruturas comerciais e residências multi familiares (apartamentos, etc). A cota de superfície impermeável vai desde 80% a 90% da cobertura total. A altura média dos edifícios é inferior a 40 metros.
13	*Dense Urban*	Áreas dentro do perímetro urbano, densamente povoadas com características geralmente indistintas entre si; alturas de edifícios com menos de 40 metros.
14	*Core Urban*	Áreas dentro do perímetro urbano, densamente povoadas com características geralmente indistintas entre si; edifícios com altura média de 40 metros.
15	*Building Blocks*	Grupos de edifícios estreitos em geral que podem ser paralelos e separados por um espaço aberto; (principalmente apartamentos ou blocos de escritórios).
16	*Industrial*	Áreas industriais/comerciais/institucionais incluíndo edifícios de

		grandes áreas, com altura geralmente abaixo dos 20 metros e separados por ruas mais largas do que 20 metros.
17	*Airport*	Pistas de aterragem e superfícies lisas.
18	*Open In Urban*	Áreas com pouca ou nenhuma vegetação (áreas pavimentadas) dentro do espaço urbano, incluindo corredores de transporte.
19	*Unclassified*	Áreas sem classificação.

Tabela 3 – Classes de *clutter*.

2.4 Algoritmos Genéticos

2.4.1 Introdução

AG são algoritmos de pesquisa heurística adaptativa, tendo como base conceitos e ideias evolutivas provenientes da seleção natural e genética [22]. Como tal, estes representam uma exploração inteligente de uma pesquisa aleatória usada para resolver problemas de otimização, direcionando a pesquisa para a região de melhor desempenho dentro do respetivo espaço de pesquisa. As técnicas básicas de AG são projetadas de modo a simular processos em sistemas naturais, necessários para a evolução, nomeadamente os princípios da "sobrevivência do mais apto", estabelecidos na teoria da evolução descrita por Charles Darwin. Uma vez que na natureza, a concorrência entre indivíduos por recursos escassos, resulta em indivíduos dotados e aptos a dominar sobre os mais fracos.

Os AG simulam a sobrevivência do mais forte entre indivíduos, através de gerações consecutivas, para resolver um problema. Cada geração é constituída por uma população de cadeias de caracteres, as quais são análogas a um dado cromossoma, existente num ADN (Ácido Desoxirribonucleico), constituído por uma dada codificação (genótipo). Cada indivíduo, submetido a um processo de evolução, representa um ponto num espaço de pesquisa e, também, uma possível solução para o problema (fenótipo).

2.4.2 Princípio de Funcionamento

O presente subcapítulo é, na ausência de outra referência, baseado em [8], [22] e [23].

Um AG trata-se de um algoritmo probabilístico, o qual mantém uma população $P(t) = \{x_1^t, \dots, x_n^t\}$ para a iteracção t. A Figura 10 apresenta o pseudo-código de um AG.

Início

$t \leftarrow 0$

Inicializar aleatoriamente $P(t)$

Avaliar $P(t)$

Enquanto (**não** condição-terminar)

Fazer

$t \leftarrow t + 1$

$P(t) \leftarrow$ Seleção $P(t-1)$

Alterar $P(t)$

Avaliar $P(t)$

Fim

Figura 10 – Pseudo-código de um AG.

Cada elemento de P representa uma possível solução do problema, sendo cada indivíduo, avaliado segundo um dado critério. Em seguida, gera-se uma nova população a partir desta, sendo substituído, pelos seus descendentes, um subconjunto de indivíduos selecionados, possuindo os mais dotados uma maior probabilidade de serem incluídos nesta seleção. Estes descendentes são obtidos pela aplicação de operadores genéticos. Existem operadores genéticos unários (mutação) e de ordem superior (cruzamento), sendo que os primeiros originam novos indivíduos pela simples modificação de um indivíduo, enquanto os últimos geram novos indivíduos pela combinação de partes de vários indivíduos da população atual. Por fim, tal como já foi referido no parágrafo anterior, após várias gerações, o algoritmo converge para um ponto ótimo.

A população inicial para o AG pode ser gerada através de diversos processos, sendo a criação de cada cromossoma da população, com valores aleatórios, o método mais utilizado. No entanto, qualquer informação prévia relativamente à solução final desejada deve ser utilizada na criação da população inicial.

A implementação de um AG deve ser caracterizada por uma representação genética para as soluções de um dado problema, por um processo de criação da população inicial, por uma dada função de classificação, simuladora de um determinado ambiente com o objetivo de avaliar indivíduos em meios de "*fitness*". Esta implementação deve ainda ser caracterizada por operadores genéticos, de modo a alterar a composição dos descendentes da população e também, por valores destinados a vários parâmetros que um AG utiliza (tamanho da população, probabilidade da utilização de operadores genéticos, etc).

Codificação de cromossomas

Tal como já foi referido anteriormente, cada indivíduo de uma dada população, apresenta uma potencial solução para o problema, sendo esta representada por um cromossoma próprio. O AG clássico, proposto por John Holland, codifica os cromossomas através de combinações binárias, respeitando o teorema fundamental de AG. Este teorema afirma, que para esquemas (modelo que identifica um subconjunto de *strings* com semelhanças em certas posições de uma *string*) curtos, de ordem baixa e de aptidão acima da média, aumentam exponencialmente nas próximas gerações.

Apesar de Holland defender que este tipo de implementação (codificação binária) alcança bons desempenhos, maximizando o paralelismo implícito inerente ao AG, em várias aplicações práticas, este tipo de codificação pode atingir desempenhos não satisfatórios. Esta conclusão é defendida por Michalewicz, ao afirmar, que quando aplicada a problemas numéricos de elevadas dimensões, que requeiram uma solução de alta precisão, uma representação binária pode resultar num desempenho insatisfatório. Como tal, este sugere a utilização de valores *float* em casos semelhantes ao descrito anteriormente. Pode, por isso, concluir-se que a escolha da codificação a aplicar é essencial aquando da utilização de AG num determinado projeto.

A estrutura de um dado cromossoma deve ser simples e apresentar todas as soluções no interior de um espaço de pesquisa, sendo devido a este facto que os *arrays* são tipicamente utilizados como estruturas de dados.

No problema a desenvolver, o respetivo *array* é preenchido com os desvios relativamente aos parâmetros originais do modelo de Okumura-Hata. Através da utilização de desvios, em vez dos valores reais dos parâmetros, torna-se possível controlar a gama de valores que os mesmos podem tomar, de modo a não permitir uma distorção significativa do modelo de propagação.

A dimensão do indivíduo é diretamente proporcional ao número de parâmetros do respetivo modelo.

Operadores genéticos

Os operadores genéticos são responsáveis por transformar a população, através de sucessivas gerações, procurando aumentar as capacidades de adaptação dos indivíduos, mantendo as características que foram adquiridas pelas gerações anteriores.

Seleção

A seleção tem o objetivo de realçar as melhores soluções (indivíduos) numa dada população, as quais são copiadas para a próxima geração. Esta escolha é feita de modo a que os indivíduos mais adaptados ao meio ambiente tenham uma probabilidade maior de se reproduzirem. Os métodos de seleção mais comuns são os seguintes:

- **Proporcional à Aptidão do Indivíduo**: este método de seleção gera uma probabilidade de sobrevivência, cujo valor é diretamente proporcional à quantidade de *fitness*, através da qual é definido se o indivíduo é melhor ou pior, comparativamente com o resto da aptidão dos

indivíduos. Assim, a possibilidade de sobrevivência torna-se proporcional à aptidão do indivíduo, sendo esta probabilidade dada por:

$$\Pr[x] = \frac{f(x)}{\sum_{y \in P} f(y)}\,, \tag{10}$$

onde x corresponde ao indivíduo, P refere-se à população, y representa outro indivíduo da mesma população e $f(x)$ trata-se da função de *fitness.*

- **Classificação**: neste método, a seleção é realizada tendo como base a classificação de cada indivíduo, dentro de uma dada população, relativamente à sua aptidão. Através desta abordagem, os indivíduos com maiores capacidades de adaptação, são impedidos de dominar prematuramente relativamente ao resto dos indivíduos, aumentado, assim, a diversidade da população.
- **Roleta**: neste tipo de seleção, cada indivíduo da população ocupa uma porção da roleta, proporcional ao seu valor de aptidão, tal como ilustra a Figura 11. Deste modo, os indivíduos com maior capacidade de adaptação possuem uma porção maior da respetiva roleta, tendo, assim, uma probabilidade maior de serem escolhidos, quer para passar à geração seguinte, quer para gerar descendentes, comparativamente aos indivíduos menos aptos.

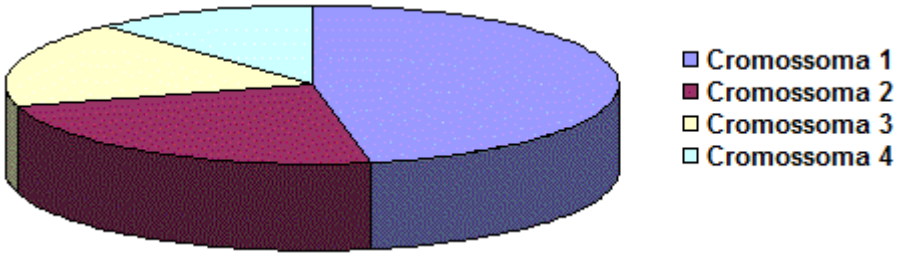

Figura 11 – Método da Roleta.

O número de vezes que a roleta é manipulada corresponde ao número total de indivíduos de uma dada população. A porção da roleta ocupada por cada indivíduo, dependendo da sua capacidade de adaptação, é dada por:

$$P_j = \frac{Adaptação_j}{\sum_{i=1}^{N} Adaptação_j} \times 100\%. \tag{11}$$

Através deste método, os indivíduos mais aptos são realçados, em detrimento dos indivíduos menos capazes.

Cruzamento

O método de cruzamento baseia-se no fenómeno de recombinação entre dois cromossomas diferentes, sendo este inspirado no conceito de reprodução sexuada. Neste método, dois indivíduos são escolhidos a partir de um conjunto de indivíduos de aptidão elevada, com o objetivo de produzir descendentes, através da troca de segmentos do seu respetivo código. Assim, é simulado o efeito de reprodução artificial de "descendência", cujo resultado é proveniente da recombinação de segmentos de código

dos progenitores.

Um dos métodos mais utilizados para implementar este tipo de seleção é o cruzamento num único ponto. Neste método de cruzamento, é definido um ponto de permuta num local específico ou aleatório, nos cromossomas dos dois indivíduos progenitores (Px e Py da Figura 12) e, em seguida, um dos indivíduos contribui com todo o seu código, localizado antes desse ponto, e o outro contribui com toda a sua informação, a partir da localização do seu ponto, produzindo assim um novo descendente.

Px: 101011 | 1010 $\quad\quad$ $D1$: 101011 | 1110

Py: 010100 | 1110 $\quad\quad$ $D2$: 010100 | 1010

Figura 12 – Cruzamento num único ponto.

Outro método vulgarmente utilizado é o cruzamento em dois pontos, sendo este semelhante ao método descrito anteriormente, mas em vez de ter apenas um ponto de cruzamento, este utiliza dois. A Figura 13 ilustra o resultado da produção de dois descendentes, utilizando este tipo de cruzamento.

Px: 101 | 011 | 1010 $\quad\quad$ $D1$: 101 | 100 | 1010

Py: 010 | 100 |1110 $\quad\quad$ $D2$: 010| 011 | 1110

Figura 13 – Cruzamento em dois pontos.

Os métodos de cruzamento referidos anteriormente podem ser utilizados aquando da utilização de representação real (valores em *float*), no entanto, existem operadores genéticos específicos para esse tipo de codificação [24].

Mutação

O operador de mutação tem o objetivo de modificar, aleatoriamente, um ou mais genes de um dado cromossoma, sem comprometer os progressos já realizados pela pesquisa sucessiva do AG. A probabilidade de mutação de um dado gene é definida como taxa de mutação, cuja probabilidade de ocorrência é exígua.

Utilizando codificação binária, este operador realiza a mutação de um ou mais genes, escolhidos aleatoriamente, invertendo os seus respetivos valores situados no interior de um dado cromossoma.

Em representação real, a mutação pode ser concebida utilizando diferentes métodos, tais como a mutação uniforme, a mutação de *Gaussian*, entre outros. Na mutação uniforme, o método seleciona aleatoriamente um gene dentro de um cromossoma e modifica-o, substituindo-o por um novo valor, escolhido aleatoriamente, o qual deve estar dentro do intervalo de valores aceites pelo respetivo gene. No método de mutação de *Gaussian*, todos os genes presentes no interior de um cromossoma, são alterados através da adição de ruído, o qual segue uma distribuição de *Gaussian*, sendo esta descrita por: $x' = x + N(0, \sigma)$ (12),

onde $N\ (0, \sigma)$ corresponde a um *array* constituído por variáveis aleatórias gaussianas, com média igual a 0 e variância σ.

2.5 Clustering

2.5.1 Introdução

Na ausencia de outra referência, o presente capítulo é baseado maioritariamente em [25], [26] e [27].

Clustering é uma das tarefas mais úteis no processo de exploração de dados (*Data Mining*), visando a descoberta de grupos e de distribuições e padrões interessantes, em dados subjacentes. O processo de *Clustering* encontra-se enraizado a muitas áreas, incluindo *Data Mining*, estatísticas, biologia e *Machine Learning.* O processo de agrupamento de um conjunto de objetos em grupos (*clusters*) de objetos semelhantes é designado por *Clustering.* Um *cluster* é uma coleção de elementos de dados que apresentam semelhanças entre elementos do mesmo grupo e disparidades entre elementos de outros *clusters*. Esta técnica visa a obtenção de grupos homogéneos e os mais separados possíveis e tem sido amplamente utilizada em diversas aplicações, tais como na realização de estudos de mercado, reconhecimento de padrões, análise de dados e processamento de imagem.

Por exemplo, considere-se uma base de dados constituída por registos de itens comprados por um dado conjunto de clientes. Um procedimento de *clustering* pode agrupar o conjunto dos clientes de modo a que, os clientes com padrões de compra semelhantes, pertençam ao mesmo *cluster*. Sendo o principal objetivo, revelar a organização dos padrões em grupos que possibilitem descobrir semelhanças e diferenças, bem como extrair conclusões úteis, relativas aos respetivos clientes.

No processo de organização por *clusters*, não existem classes predefinidas, nem exemplos que revelem o tipo de relações desejável, entre os elementos de dados. Sendo, por isso, referido como um processo não supervisionado. Por outro lado, a classificação trata-se do processo de atribuição de um elemento de dados desconhecido, a um grupo específico, de um dado conjunto de grupos predefinidos.

2.5.2 Fases de um Processo de *Clustering*

As etapas fundamentais de qualquer processo de *clustering*, apresentam-se ilustradas na Figura 14.

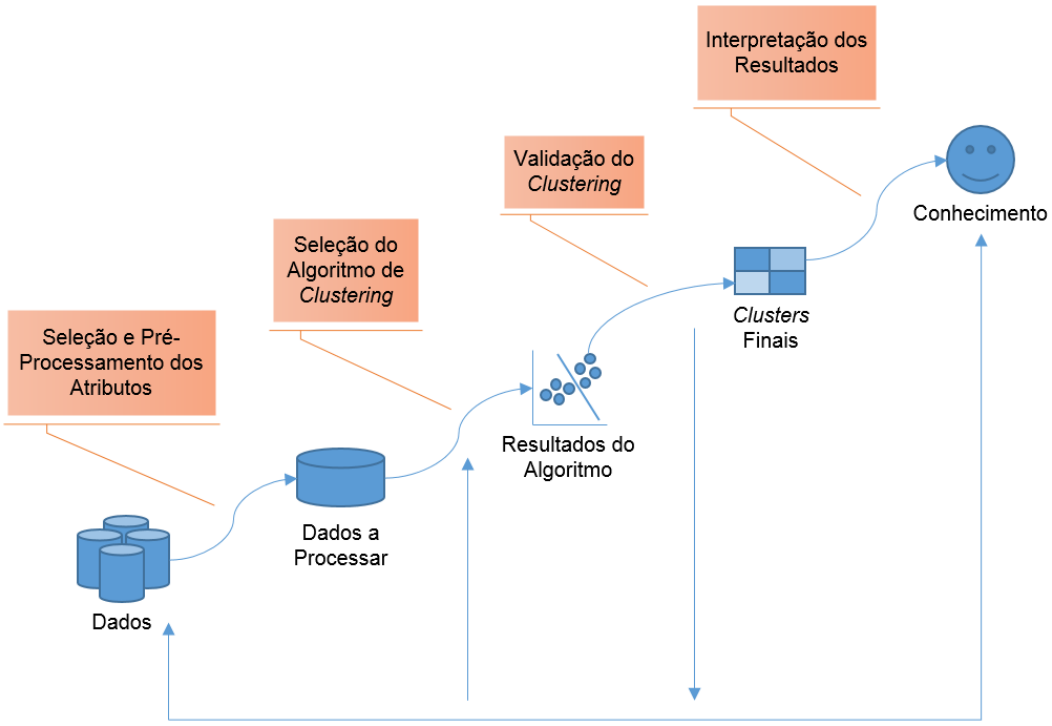

Figura 14 – Fases de um processo de *clustering*.

Um processo de *clustering* aplicado a um conjunto de dados, pode resultar em partições diferentes, dependendo do critério específico, usado para o agrupamento. Como tal, existe a necessidade de pré-processamento, antes de ser aplicada uma dada técnica de *clustering* a um conjunto de dados.

A fase de seleção e pré-processamento dos atributos, dos elementos de dados, tem o objetivo de selecionar corretamente as características, sobre as quais o *clustering* deve ser executado, de modo a codificar o máximo de informação possível, relativamente ao objetivo final, e a melhorar a qualidade do agrupamento. A normalização trata-se de uma técnica de pré-processamento de dados, utilizada para redimensionar um dado conjunto de atributos, num intervalo de valores específico. Uma normalização prévia dos dados, é bastante vantajosa e, particularmente necessária, para métricas de distâncias sensíveis a variações de amplitude, ou de escala (exemplo – distância euclidiana) dos atributos dos respectivos elementos de dados. A normalização anula a possibilidade de, elementos constituídos por atributos de valores elevados, superarem elementos caracterizados por atributos de valores menores, através da uniformização das amplitudes e/ou das escalas, desses mesmos atributos. As técnicas de pré-processamento de dados, tais como, *Z-Score, Min-Max e Decimal Scaling* [28], são aplicadas a dados brutos, tornando-os limpos, livres de ruído e consistentes, através de transformações lineares, que melhorando a precisão dos algoritmos de *clustering,* possibilitam a construção de *clusters* de boa qualidade.

O passo referente à escolha de um algoritmo de *clustering*, centra-se na definição de uma medida de proximidade e de um critério de agrupamento. Esta definição caracteriza um algoritmo de *clustering*. A medida proximidade quantifica o quão "semelhantes" dois elementos de dados são. O critério de *clustering* tanto pode ser expresso por meio de uma função de custo ou algum outro tipo de regras.

A exactidão dos resultados do algoritmo de *clustering* é verificada utilizando técnicas apropriadas (índices de validação). Visto que os algoritmos de *clustering* definem conjuntos que não são previamente conhecidos, independentemente dos métodos de *clustering*, a partição final dos dados requer algum tipo de avaliação.

A interpretação dos resultados, tipicamente, integra os resultados do *clustering* com outras evidências experimentais, com o objectivo de as analisar e retirar informação útil.

2.5.3 Categorias de Algoritmos de *Clustering*

Uma multitude de métodos de *clustering* são propostos na literatura. Os algoritmos de *clustering* são classificados de acordo com:

- O tipo de dados de entrada para o algoritmo;
- O critério de *clustering* que define a similaridade entre os pontos de dados;
- A teoria e os conceitos fundamentais em que se baseiam as técnicas de análise de *clustering*.

Assim, de acordo com o método adoptado para a definição dos *clusters*, os algoritmos são classificados nos seguintes tipos:

- *Clustering* de partição: decompõem um conjunto de dados, num conjunto de *clusters* disjuntos, determinando um número inteiro de partições que otimizam um dado critério de convergência.
- *Clustering* hierárquico: foca-se, sucessivamente, em fundir *clusters* menores para criar *clusters* maiores, ou em dividir *clusters* maiores. O resultado do algoritmo é uma "árvore de *clusters*", isto é, um dendrograma, que indica como os *clusters* estão relacionados. Ao "cortar" o dendograma num dado nível, obtém-se o *clustering* dos elementos de dados, em grupos separados.
- *Clustering* baseado em densidade: agrupa objetos vizinhos, de um conjunto de dados, em *clusters*, com base em condições de densidade.
- *Clustering* baseado em grelha: para a obtenção de dados espaciais. Este quantifica o espaço num número finito de células e, em seguida, realiza todas as operações no espaço quantificado.

Para cada uma das categorias acima, há uma vasta riqueza de subtipos e diferentes algoritmos para contruir os *clusters*. Em termos gerais, os algoritmos de *clustering* são baseados em critérios que avaliam a qualidade de uma determinada partição. Mais especificamente, estes assumem como critério alguns parâmetros (tais como o número de *clusters*, a densidade de *clusters*, etc) e definem a melhor estratégia de agrupamento, de um conjunto de dados, segundo os respectivos parâmetros.

2.5.4 Algoritmos de *Clustering*

K-Means

Um dos algoritmos iterativos mais comuns é o algoritmo *K-Means*, amplamente utilizado pela sua simplicidade de implementação e velocidade de convergência. A complexidade temporal deste algoritmo depende do número de objetos e *clusters* a criar, sendo $O(t \times k \times n)$, onde t corresponde ao número de iterações, k, ao número de *clusters* e n, ao número de objectos.

O algoritmo é aplicado a um conjunto de elementos de dados, n, definindo previamente um valor K, relativo ao número *clusters* a construir. Depois, são produzidos K *centroids*, com o objectivo de minimizar a função objectiva, que é a distância média de cada elemento, ao *centroid* mais próximo. Uma implementação típica do algoritmo começa com uma seleção aleatória dos K *centroids*, atribuindo de forma iterativa, cada elemento de dados, ao *centroid* mais próximo, atualizando, simultâneamente, as novas posições dos *centroids* até que a convergência seja alcançada.

Uma desvantagem de *K-Means*, trata-se da possibilidade de atingir um mínimo local da função objectiva, em detrimento do mínimo global desejado, o que significa que a convergência é alcançada, mas a solução não é a ideal. No entanto, esta limitação é superada, executando o algoritmo múltiplas vezes, com diferentes *centroids*, selecionando a partição com o menor erro de *clustering*.

Em Anexos – Anexo A, encontra-se uma explicação do algoritmo *K-Means*, baseada num exemplo numérico.

Fuzzy C-Means

No algoritmo *K-Means*, cada elemento pode ser classificado num único *cluster* (*clustering* exclusivo), e os *centroids* são atualizados com base nos elementos classificados. O algoritmo *Fuzzy C-Means* considera que todos os elementos têm um determinado grau de pertença para cada *cluster*, e os respectivos *centroids* são calculados com base nesses graus.

Enquanto no algoritmo *K-Means*, um *centroid* (centro de um dado *cluster*) é calculado através da média dos elementos presentes nesse *cluster*, o *Fuzzy C-Means* determina o centro, através de uma média ponderada de todos os elementos, utilizando, como pesos, as probabilidades de pertença de cada elemento. Os elementos caracterizados por um valor elevado, relativo à probabilidade de pertencerem uma dada classe, possuem pesos maiores, os quais traduzem uma influência maior sobre o *centroid*.

O processo de atribuição de elementos aos *centroids* é semelhante ao algoritmo *K-Means*. A atualização de *centroids* é repetida até que a convergência seja alcançada.

Hierárquico

Um algoritmo de *clustering* hierárquico cria uma árvore hierárquica de semelhanças entre os elementos (dendrograma). O seu principio de funcionamento é baseado em *clustering* de aglomeração, sendo o algoritmo inicializado, através da atribuição de cada elemento ao *cluster* específico. As distâncias entre agrupamentos são definidas, utilizando uma métrica de distância (por exemplo, a euclidiana) ou de semelhança (por exemplo, a correlação). Em seguida, o algoritmo funde os dois *clusters* mais próximos e actualiza a totalidade das distâncias, ao *cluster* recém-formado, através de um método de ligação. Este passo é repetido até que haja apenas um *cluster* que contenha todos os elementos.

Este processo define uma sequência de partições aninhadas, na qual cada uma contém uma partição com menos um *cluster*, comparativamente à partição anterior. Para obter uma partição constituída por K agrupamentos, o processo deve ser finalizado em $K - 1$.

Expectation Maximization

O algoritmo de *clustering Expectation Maximization*, estima as densidades de probabilidade das classes, utilizando o algoritmo de *Expectation Maximization (*EM*)*. O resultado é um conjunto estimado de K distribuições multivariadas, sendo cada uma defininda por um *cluster*. Cada elemento de dados é atribuido ao *cluster* com a máxima probabilidade condicional.

Diferentes considerações sobre o modelo correspondem a diferentes restrições sobre as matrizes de covariância de cada distribuição. Quanto menos rígidas forem as restrições, mais flexível é o modelo, no entanto, são necessárias mais amostras para a obtenção de boas estimativas dos parâmetros adicionais.

Self Organazing Maps

Através da aplicação de *Self Organazing Maps* (SOM) ao conjunto de dados, os *clusters* podem ser definidos por pontos sobre uma "grelha ajustada" aos dados. Usualmente, o algoritmo utiliza uma grelha bidimensional num espaço dimensional mais elevado, no entanto, para o *clustering* é típico utilizar-se uma grelha unidimensional.

O agrupamento utilizando SOM é bastante útil na visualização dos dados, devido à representação espacial da grelha, facilitada pela sua baixa dimensionalidade, revelando informações úteis acerca dos dados [29].

2.5.5 Técnicas de Validação de *Clustering*

As técnicas de validação dos resultados de *clustering* visam responder a questões como: "quantos grupos existem no conjunto de dados?", "a configuração de *clustering* resultante é a mais adequada para o conjuntos de dados?", "existirá uma partição melhor para o conjunto de dados?".

Um dos desafios mais importantes na análise de *clusters* é a avaliação dos resultados do *clustering*, de modo a encontrar o esquema / configuração que melhor se adapta aos dados subjacentes.

O objetivo dos métodos de *clustering* foca-se em descobrir grupos significativos, presentes num conjunto de dados. A determinação do número ideal de *clusters*, no qual se encaixa um conjunto de dados, é um dos problemas de *clustering* mais desafiantes.

A visualização do conjunto de dados é uma verificação fundamental dos resultados de *clustering*. No entanto, para grandes conjuntos de dados multidimensionais (por exemplo: mais de três dimensões) a interpretação dos resultados torna-se visualmente impossível. Incentivando a utilização de um índice de validação de *clustering*.

O procedimento de avaliar os resultados de um algoritmo de *clustering* é conhecido sob o termo de validação do *clustering*. Em termos gerais, existem três abordagens que possibilitam investigar a validação dos resultados de *clustering*:

- A primeira abordagem, designada de validação externa, compara a partição gerada pelo algoritmo de *clustering* com uma estrutura previamente especificada, sendo esta imposta ao conjunto de dados, de modo a reflectir a referida estrutura, na organização dos *clusters* do respetivo conjunto de dados.
- A segunda é baseada no cálculo de propriedades dos *clusters* resultantes, tais como a compacidade e a separação. Esta abordagem é designada de validação interna porque não necessita de informações adicionais sobre os dados.
- A terceira é baseada em comparações de partições geradas pelo mesmo algoritmo de *clustering*, assumindo diferentes parâmetros ou subconjuntos de dados. Esta é designada de validação relativa e também não requer informações adicionais.

As duas primeiras abordagens são baseadas em testes estatísticos e a sua principal desvantagem é o seu alto custo computacional. Por outro lado, a terceira abordagem visa encontrar o melhor esquema de organização por *clusters*, através do qual, um algoritmo de *clustering* pode ser definido, em determinadas condições e parâmetros.

2.5.6 Cenário em Alta Dimensão

A análise de *clusters* em cenários de alta dimensão torna-se bastante desafiante, devido à grande variação no comportamento dos atributos dos elementos de dados, sobre as diferentes localizações

dos dados. Com o aumento da dimensão, as distâncias vão perdendo a sua eficácia, bem como a sua significância estatística, em virtude de atributos irrelevantes. O princípio centra-se no facto de, os atributos caracterizados por frações exíguas, permanecerem relevantes com o aumento da dimensão dos dados, proporcionando a perda de definição das distâncias, bem como o aumento do efeito de concentração, devido ao comportamento dos atributos irrelevantes. Os efeitos de concentração referem-se à situação, em que uma quantidade elevada de atributos ruidosos ou não correlacionados, provoca um cenário em que todas as distâncias entre pontos, se tornam semelhantes [30].

Em algoritmos de *clustering* baseados em distância, o ruído e o efeito de concentração são problemáticos de duas maneiras:

1. Um aumento do ruído causado por atributos irrelevantes, pode causar erros na distância de representação e, consequentemente, promover uma representação errada das distâncias entre objectos.

2. O efeito de concentração, incentivado por dimensões irrelevantes conduzem a uma redução da significância estatística dos resultados provenientes de algoritmos baseados em distâncias.

Uma das premissas para abordar estes problemas, passa por controlar o tamanho da dimensão dos dados, seleccionando os atributos, considerados como os mais influentes, bem como por aplicar uma função de proximidade, que ofereça melhor contraste de dados, no cálculo da distância entre pontos.

2.6 Estado da Arte

No trabalho desenvolvido em [8] foi estudada a aplicabilidade dos algoritmos genéticos (AG) à otimização multivariável do modelo de propagação em ferrovia. A metodologia utilizada na modelação do problema e na implementação do algoritmo conduziu a resultados muito acima dos obtidos pelos métodos normais, permitindo obter uma otimização do conjunto de parâmetros ajustáveis do modelo de propagação Okumura-Hata para um conjunto de medidas rádio obtidas previamente. A metodologia proposta obteve como principais vantagens, a redução significativa do erro de predição para o conjunto de ambientes estudados, a conservação das características básicas do modelo, bem como do seu significado teórico, através de uma modelação metodológica dos parâmetros de calibração e, ainda, a validação do modelo utilizado para a estimação da cobertura rádio em ambientes ferroviários. Como desvantagens da metodologia proposta destacam-se a dificuldade na otimização global do modelo, em termos de modulação do comportamento do sinal, para os diferentes tipos de ambientes e a não utilização da informação de *clutter* de forma generalizada.

A metodologia proposta em [9] teve por base a utilização de redes neuronais para realizar a predição de cobertura radio, não utilizando um modelo de estimação de cobertura radio. Os resultados obtidos

concluíram como vantagens, um erro de predição diminuto para os ambientes e situações cobertas pelas medidas, a validação da utilização das redes neuronais para predição de cobertura rádio e, ao efetuar uma classificação prévia do ambiente, uma redução significativa do erro, permitindo realizar uma aprendizagem competitiva. Como desvantagens, esta técnica revelou ficar comprometida com a existência de medidas que representem 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.

Capítulo 3

A Associação de *Clustering* a Otimização

Este capítulo apresenta a implementação de uma calibração automática de modelos de propagação, para os diferentes tipos de ambientes e características, utilizando algoritmos genéticos e de *clustering*, respetivamente.

3.1 Introdução

Em [8] demonstrou-se ser válida a utilização de AG na otimização de parâmetros de calibração de modelos de propagação, quando aplicados à predição de cobertura rádio em caminhos-de-ferro. Na realização desta metodologia, foi destacada a dificuldade em obter uma otimização global, em termos de modulação do comportamento do sinal, para os diferentes tipos de ambientes. Acrescentou-se, também, a desvantagem da não utilização da informação de *clutter* de forma generalizada. Como tal, foi proposto o desenvolvimento de um algoritmo, capaz de agrupar um conjunto de medidas obtidas em ambiente ferroviário, constituído pela respetiva informação de *clutter*, em subconjuntos que partilhem semelhanças geográficas/morfológicas. Possibilitando, assim, a aplicação dos AG, a cada um dos grupos obtidos. O diagrama de blocos, ilustrado pela Figura 15, é inspirado numa macro visão dos procedimentos sequencialmente efetuados, pelo algoritmo desenvolvido.

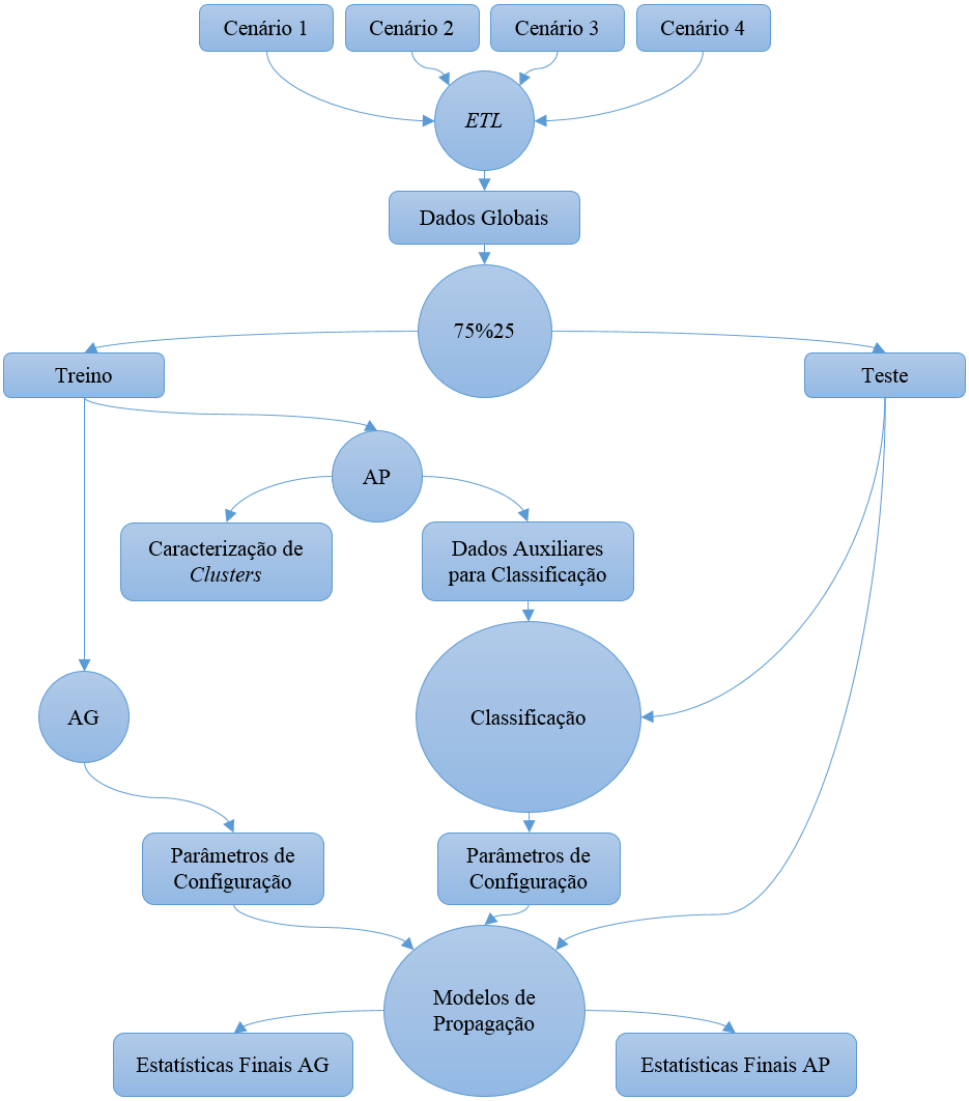

Figura 15 – Diagrama de blocos da implementação do algoritmo desenvolvido.

A informação recolhida em [8], possibilitou a criação dos cenários apresentados na figura acima. Esta informação é submetida a um processo de ETL (*Export*, *Transform and Load*), isto é, os dados são exportados, transformados e, posteriormente, armazenados num ficheiro contentor de dados *Matlab*. Após este processo, é realizada uma divisão aleatória para as fases de treino (75%) e de teste (25%), cuja implementação tem o objetivo de validar o algoritmo desenvolvido. Na fase de treino, é selecionada a informação a ser consumida pelos AG (Algoritmos Genéticos de [8]), sendo a informação resultante, um conjunto de parâmetros de configuração, otimizados para o caso global, a ser utilizado pelo respetivo modelo de propagação, na fase de teste. Por outro lado, é, também, selecionada a informação a ser consumida pelo AP (Algoritmo Proposto), sendo a informação resultante, a caracterização dos *clusters*, construídos através do algoritmo de *clustering K-Means* Personalizado (KMP). De AP resulta, ainda, um conjunto de dados auxiliares do processo de classificação. Após a classificação dos elementos de teste, é retornado um conjunto de parâmetros de configuração, previamente otimizados para um dado *cluster*, a ser utilizado pelo modelo de propagação, na fase de teste. Depois, a informação geográfica dos dados de teste, juntamente com os parâmetros de configuração, de ambos os algoritmos (AG e AP), são introduzidos no modelo de propagação e é elaborada uma predição final. De modo a comparar a predição, proveniente de ambos os algoritmos, com as medidas previamente realizadas, utilizam-se estatísticas de primeira ordem (o erro médio absoluto, ME, a raiz do erro quadrático médio, $RMSE$ e o desvio padrão do erro, ESD), bem como o coeficiente de correlação (RE). No fim, são comparadas as estatísticas finais de AG, com as de AP.

O presente capítulo tem o objetivo de descrever o procedimento das etapas inerentes à implementação da estratégia proposta, quer relativamente ao *clustering*, quer à combinação dos AG, com o algoritmo desenvolvido. O capítulo é finalizado com a apresentação do método de interpretação dos resultados de *clustering*, sendo este baseado na informação de *clutter* por *cluster* e na comparação das estatísticas finais, provenientes de cada um dos caminhos ilustrados pela Figura 15.

O algoritmo foi desenvolvido em *Matlab*, devido às elevadas dimensões dos dados utilizados, quer para a otimização, quer para o *clustering*.

3.2 Informação Geográfica e ETL dos Elementos de Dados

Tal como já foi referido anteriormente, a predição de cobertura do sinal rádio é uma etapa indispensável no planeamento de uma rede rádio. Em [8] foram obtidas estimativas, referentes a um estudo teórico do comportamento do sinal, ao longo dos caminhos-de-ferro, permitindo assim, o planeamento das BTS ao longo da linha. Em ambiente ferroviário, a métrica utilizada, quer para distâncias, quer para referenciar uma dada ocorrência ou instalação, é designada por PK (Ponto Quilométrico). Não existindo nenhum método numérico de associar um PK a um ponto geográfico, torna-se necessária a utilização de um ficheiro com essa informação.

As características geográficas/morfológicas dos cenários estudados (Algarve, Cascais, Sintra e Vendas Novas), recolhidas por BTS, intrínsecas a cada ponto da linha férrea, são as seguintes:

- Distância entre a BTS e o ponto da linha;
- Altura efetiva da antena da BTS;
- Parâmetros relativos aos 3 obstáculos principais;
- Distância percorrida sobre vegetação;
- Distância percorrida sobre água;
- Altura da ondulação do terreno;
- Altura média da ondulação do terreno.

A esta informação foi adicionada a altura da antena do móvel, a frequência e a informação de *clutter*, a qual pode ser visualizada, quer ao longo da linha, quer espacialmente. As Figuras 16, 17, 18 e 19 ilustram estes dois tipos de visualização, relativos aos quatro cenários estudados.

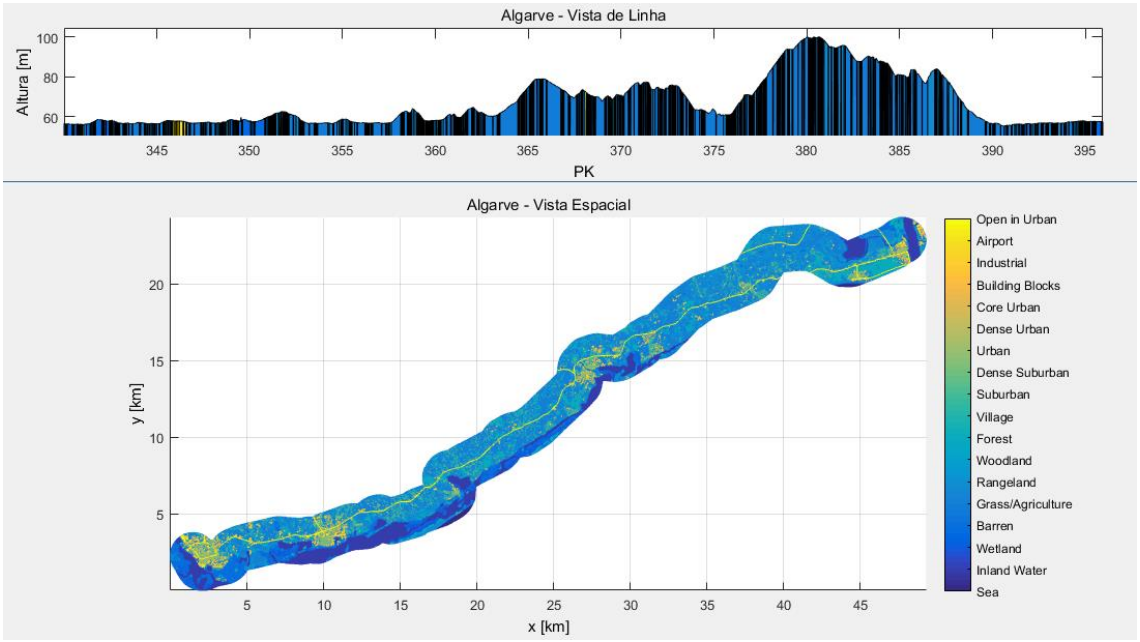

Figura 16 – Informação de *clutter* da linha do Algarve.

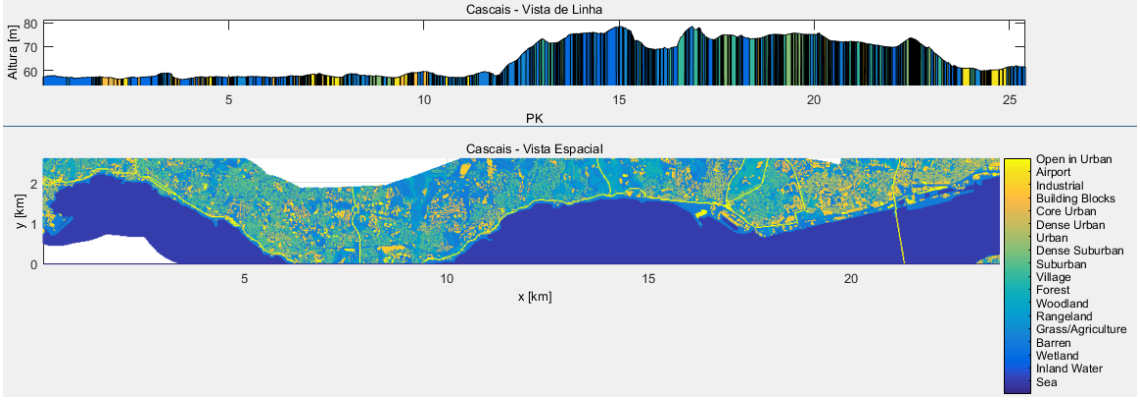

Figura 17 – Informação de *clutter* da linha de Cascais.

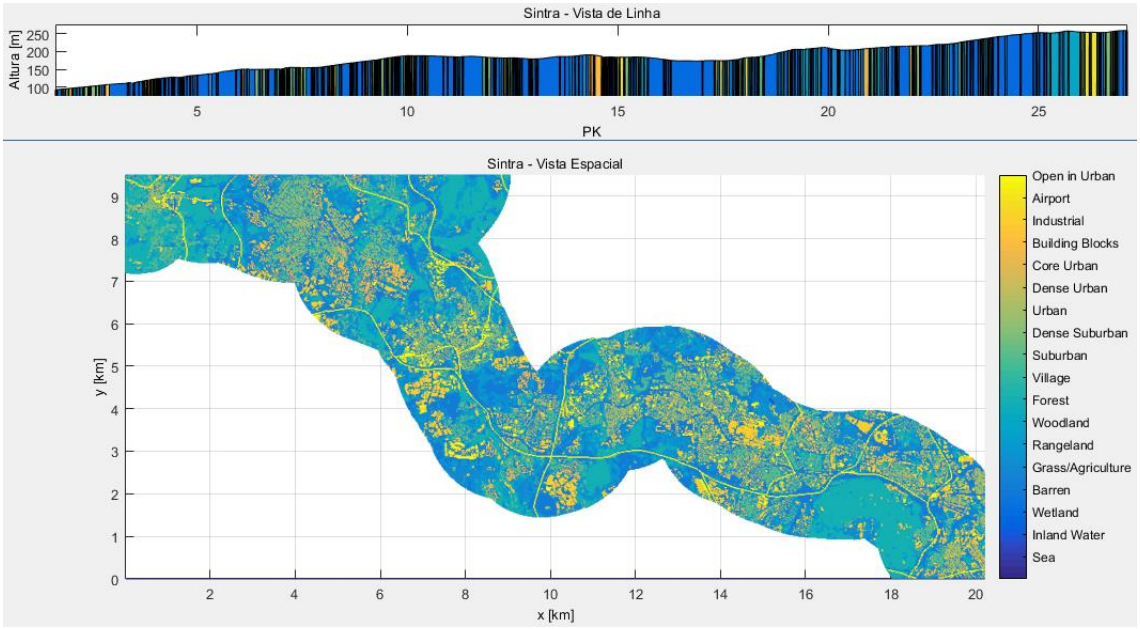

Figura 18 – Informação de *clutter* da linha de Sintra.

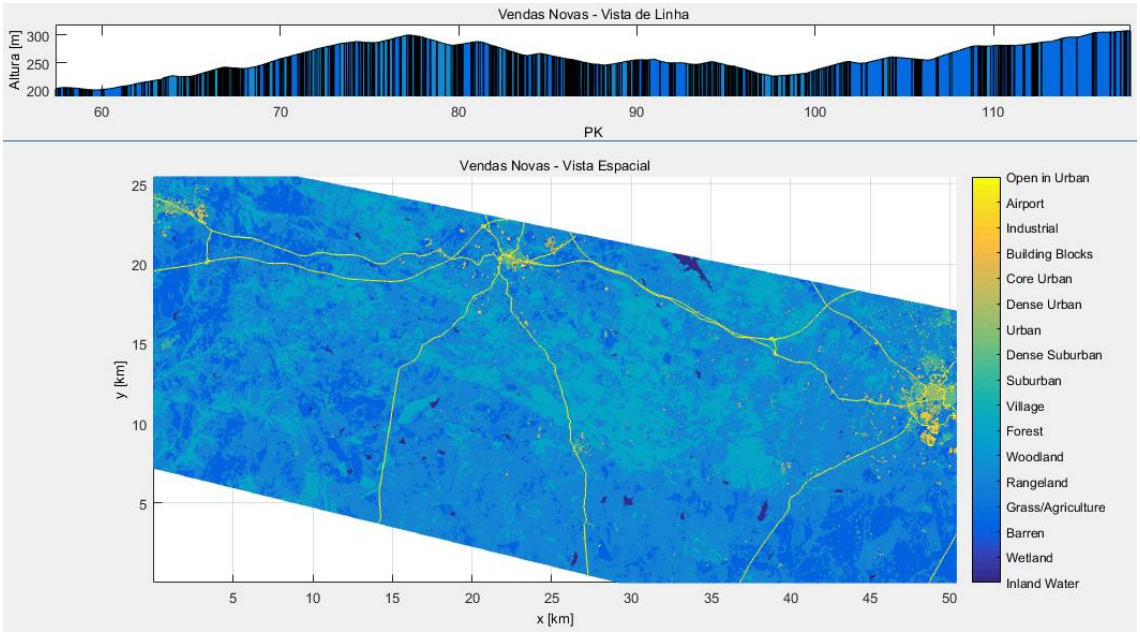

Figura 19 – Informação de *clutter* da linha de Vendas Novas.

Tal como já foi referido anteriormente, a informação de *clutter* utilizada, é constituída por 19 classes, as quais se apresentam descriminadas, em cada uma das figuras, através de uma barra de cores.

A informação recolhida, descrita até então, é posteriormente sujeita a um processo de ETL [31]. Tal como o nome sugere, este processo tem o objetivo de extrair, transformar e armazenar dados, provenientes de uma fonte externa, para um certo ficheiro contentor de dados. Neste caso, a informação é extraída dos respetivos ficheiros contentores, sendo convertida para um formato matricial,

de modo a poder ser transformada. Esta transformação é realizada de modo a que, cada elemento de dados, seja representado por uma linha e por um determinado número de colunas, equivalente ao número de atributos correspondentes. A estrutura resultante, referente ao conjunto de dados, representa-se por uma matriz de n-por-p (n elementos por p atributos). Estes atributos correspondem às características da informação anteriormente descrita (distancia, altura das antenas, classes de *clutter*, etc). Após esta fase, os dados são armazenados num ficheiro *Matlab* (Dados Globais), de modo a facilitar a seleção dos respetivos elementos de dados, quer para o AG, quer para o AP.

3.3 Processo de Aprendizagem

A realização de uma divisão aleatória dos dados, armazenados no respetivo ficheiro *Matlab*, tem o objetivo de validar o algoritmo desenvolvido. Para a fase de treino é aleatoriamente amostrada 75% da informação, sendo os restantes 25%, direcionados para a fase de teste.

A fase de treino apresenta-se ilustrada pela Figura 20, construída através de um recorte parcial do diagrama de blocos global.

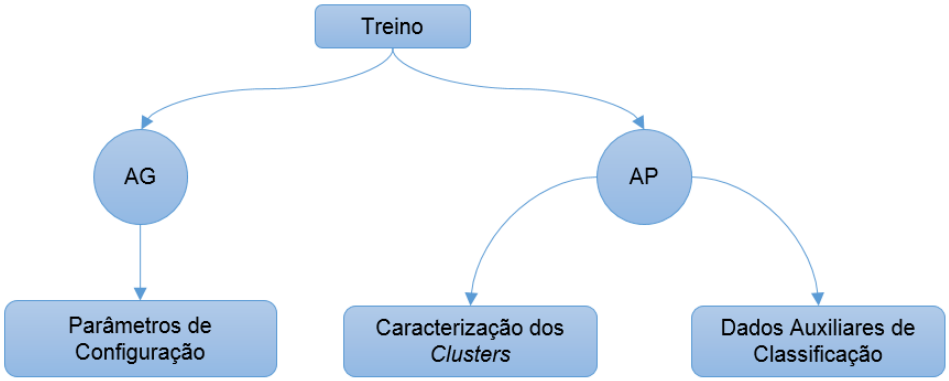

Figura 20 – Diagrama de blocos da fase de treino.

Durante o processo de aprendizagem, é selecionada a informação a ser consumida pelo AP, sendo a informação resultante, os dados a classificar, os quais são constituídos por *K* conjuntos de parâmetros de configuração do respetivo modelo e pela localização dos *K centroids*. Deste processo, também resulta a caracterização de cada *cluster* construído através do algoritmo de *clustering* desenvolvido (KMP).

Por outro lado, é selecionada a informação a ser consumida pelo AG, a qual corresponde aos elementos de dados, de treino, caracterizados pela totalidade dos atributos referidos no subcapítulo 3.2 – Informação Geográfica e ETL dos Elementos de Dados, à exceção da informação de *clutter*. E, tal como já foi referido anteriormente, a informação resultante deste processo é um conjunto de parâmetros

de configuração, otimizados para o caso global. No presente subcapítulo são relatadas as diferentes estratégias aplicadas, até alcançar a mais vantajosa para o respetivo objetivo, isto é, é relatado o caminho percorrido até à solução final, de cada uma das etapas do algoritmo desenvolvido.

A Figura 21 ilustra um "*zoom in"* do processo AP, sendo o objetivo deste subcapítulo, explicar o seu princípio de funcionamento, bem como as etapas inerentes à execução do mesmo.

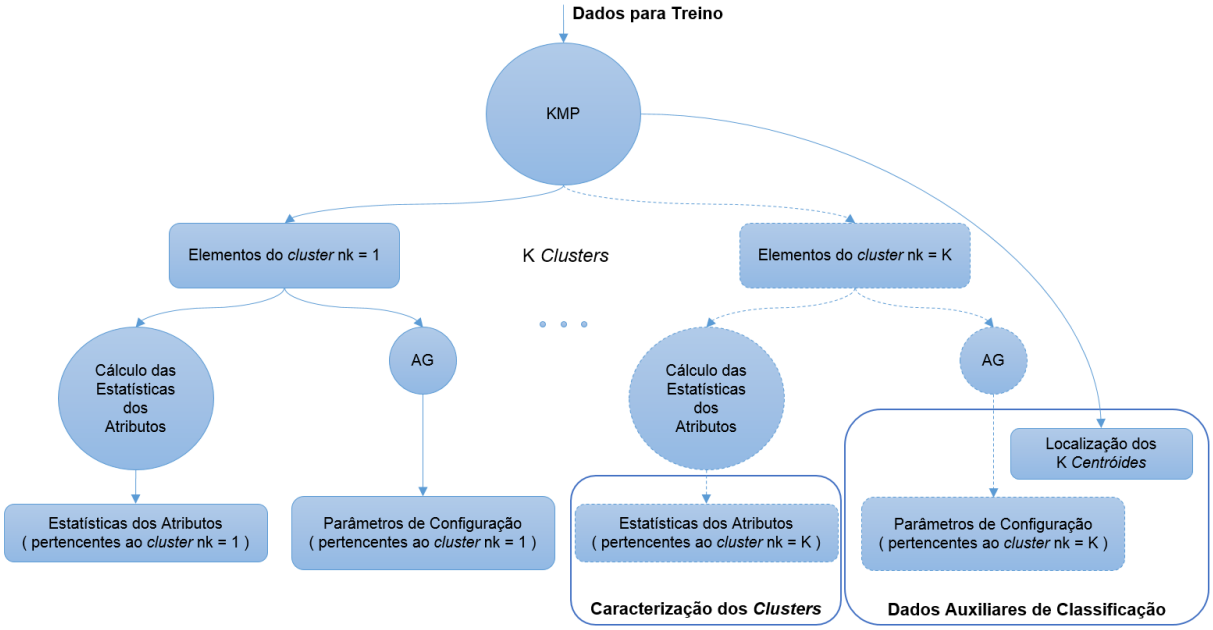

Figura 21 – *Zoom in* do processo AP.

3.3.1 *K-Means* Personalizado

Os elementos de dados, descritos no subcapítulo anterior, são armazenados num ficheiro *Matlab*, constituído pelos dados globais, através da qual é realizada a seleção e pré-processamento dos mesmos, para o consumo do algoritmo de *clustering* desenvolvido (KMP). A seleção dos elementos de dados, caracterizados por um certo número de atributos, visa recolher as particularidades mais influentes de cada ponto, para o agrupamento.

Para a implementação deste algoritmo é utilizada a função *kmeans,* do *Matlab,* devido à sua velocidade de convergência e simplicidade. Os argumentos de entrada tratam-se de uma matriz X e de um valor inteiro positivo K. As linhas de X correspondem a observações / pontos e as colunas, a variáveis / atributos; enquanto K, corresponde ao número de *clusters* a produzir. A função retorna um vetor constituído pelos índices dos *clusters* correspondentes a cada elemento de dados [32].

O nome do presente subcapítulo é inspirado no facto de terem sido adicionadas ao algoritmo original *K-Means*, as estratégias de seleção e de pré-processamento do conjunto dos elementos de dados, bem como o método de validação do *clustering*, de modo a estimar o valor de *K*. Tornando-se, assim, um algoritmo *K-Means* personalizado, ajustado às características dos dados utilizados, com o objetivo

de obter o melhor agrupamento possível.

Seleção e pré-processamento de dados

Ao longo da realização deste capítulo, são implementadas várias estratégias de agrupamento, nomeadamente na seleção dos atributos dos elementos de dados a serem agrupados. A aplicação de pré-processamento aos atributos sobre os quais o algoritmo de *clustering* é executado, tem o objetivo de melhorar a qualidade do agrupamento. Como tal, é realizada a normalização dos atributos dos elementos de dados. *Min-Max* é o método que apresenta melhores resultados, comparativamente com as três técnicas de normalização, referidas anteriormente, para o tipo de dados utilizado, e, como tal, é o escolhido para o pré-processamento de dados. A normalização *Min-Max* é o processo de transformar atributos de elementos de dados, em valores compreendidos entre 0,0 e 1,0. Sendo o menor (*Min*) valor, definido como 0.0 e o maior (*Max*) como 1.0. Proporcionando, assim, uma maneira fácil de comparar valores, medidos através de diferentes escalas ou através de unidades de medida diferentes. A normalização de um valor, de um dado atributo, é traduzida pela seguinte equação:

$$MinMax(X_{ij}) = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \quad (13)$$

A utilização de normalização tem o objetivo de uniformizar os atributos dos elementos de dados, atribuindo-lhes pesos iguais, de modo a que elementos causadores de ruído, possam ser anulados, aumentando, assim, a validade dos dados e, consequentemente, a precisão do resultado de *clustering*.

A implementação de várias estratégias de agrupamento tem o objetivo de alcançar uma seleção ótima, que promova o melhor resultado estatístico final, isto é, a maior diminuição do erro na estimativa do valor da atenuação, comparativamente aos valores obtidos em [8].

O primeiro conjunto de dados a ser selecionado é caracterizado pelo conjunto de atributos que provou em [8], ser o mais influente na predição do sinal rádio. Este conjunto é constituído pela distância (d), pela altura efectiva da antena da estação base, obtida segundo ITUR (hbe) e pelo parâmetro do obstáculo principal, calculado através do método de Deygout ($v1$).

Os testes realizados consideram todas as combinações possíveis entre os três atributos:

1. d;
2. $v1$;
3. hbe;
4. d e $v1$;
5. d e hbe;
6. $v1$ e hbe;
7. $d, v1$ e hbe.

A construção do gráfico representado pela Figura 22 tem o objetivo de facilitar a visualização dos resultados, onde são apresentadas as estatísticas provenientes de AP, correspondentes a este

conjunto de dados (T1).

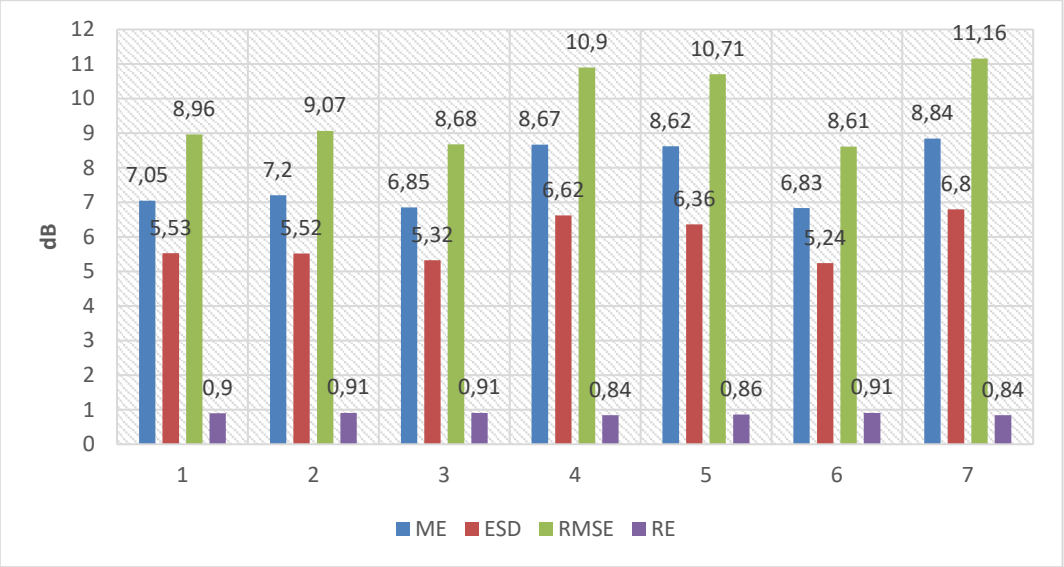

Figura 22 – Estatísticas de T1.

Através da análise do gráfico acima, conclui-se que o melhor resultado, para este conjunto de dados, consiste em selecionar os atributos $v1$ e hbe. A Figura 23 apresenta a comparação entre as estatísticas resultantes ($ME, ESD, RMSE$ e RE), correspondentes a esta selecção, provenientes de AP, e as estatísticas finais, provenientes de AG.

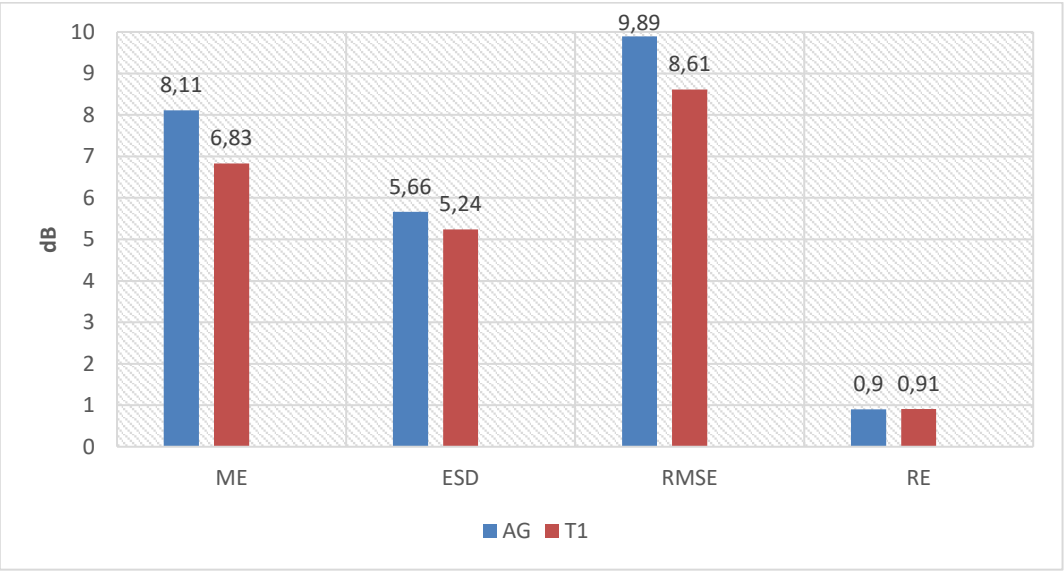

Figura 23 – Comparação entre as estatísticas AG e as melhores estatísticas de T1.

Esta estratégia comprova que, apesar da não utilização da informação de *clutter*, através da aplicação de um algoritmo de *clustering*, de modo a agrupar o conjunto de dados utilizado, tendo como base os parâmetros $v1$ e hbe, em subconjuntos que partilhem semelhanças geográficas/morfológicas, e da

aplicação dos algoritmos genéticos, a cada um dos *clusters* obtidos, produzem-se soluções de parâmetros que minimizam o erro da estimativa do valor da atenuação, comparativamente aos valores obtidos, utilizando o algoritmo desenvolvido em [8]. Adiciona-se, também, o facto de que, à exceção das combinações 4, 5 e 7, os resultados das restantes combinações, apresentam melhorias, relativamente aos resultados, utilizando apenas os AG.

O segundo conjunto de atributos a ser selecionado e, depois, submetido a um processo de *clustering*, trata-se da informação de *clutter*, sendo esta, tal como já foi referido anteriormente, constituída por 19 classes de *clutter*.

A motivação para a utilização da informação de *clutter* como atributos do agrupamento, deve-se ao facto de esta possuir a classificação das características do terreno, dos cenários em estudo. A Figura 24 apresenta a comparação entre as estatísticas resultantes, correspondentes a este teste (T2), provenientes de AP, e as estatísticas finais, provenientes de AG.

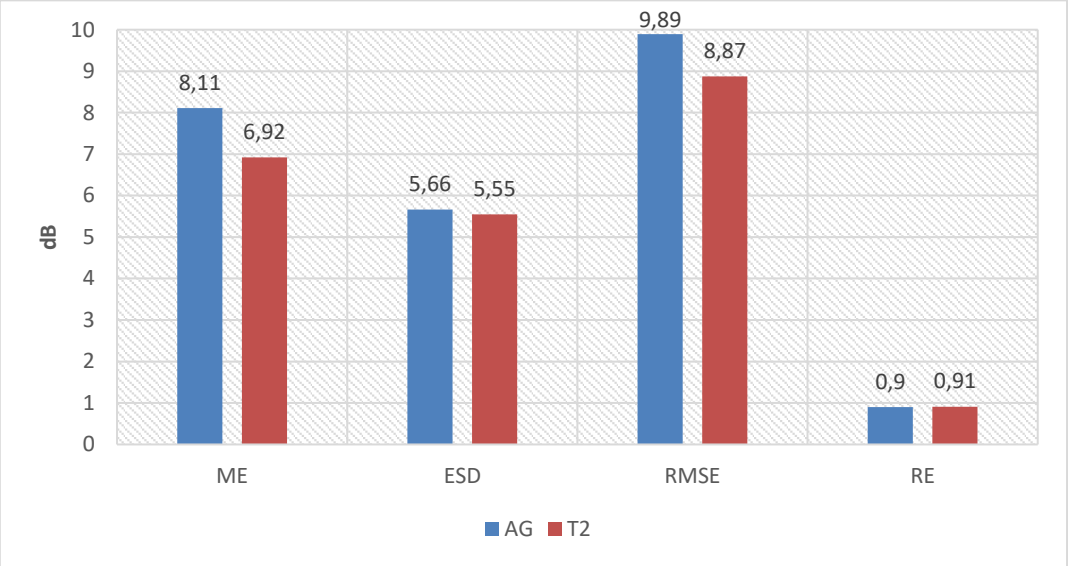

Figura 24 – Comparação entre as estatísticas AG e as estatísticas de T2.

Analisando o gráfico acima, observa-se uma ligeira subida no valor das estatísticas ME, ESD e RMSE, comparativamente com os valores obtidos em T1. No entanto, e destacando a possível existência de erros cartográficos, na construção da respetiva base de dados de *clutter*, mantém-se uma melhoria de resultados, comparativamente aos obtidos através do treino com AG, comprovando que a utilização da informação de *clutter* conduz a uma melhoria da precisão do modelo, diminuindo o erro global na predição do sinal rádio.

Os atributos utilizados no primeiro conjunto de dados, em conjunto com a informação apresentada no segundo, formam o terceiro conjunto de dados a ser selecionado e, posteriormente, agrupado. Os testes realizados consideram todas as combinações possíveis, admitindo a informação de *clutter* como base:

1. 19 Classes de *clutter* e d;
2. 19 Classes de *clutter* e $v1$;
3. 19 Classes de *clutter* e hbe;
4. 19 Classes de *clutter*, d e $v1$;
5. 19 Classes de *clutter*, d e hbe;
6. 19 Classes de *clutter*, $v1$ e hbe;
7. 19 Classes de *clutter*, $d, v1$ e hbe.

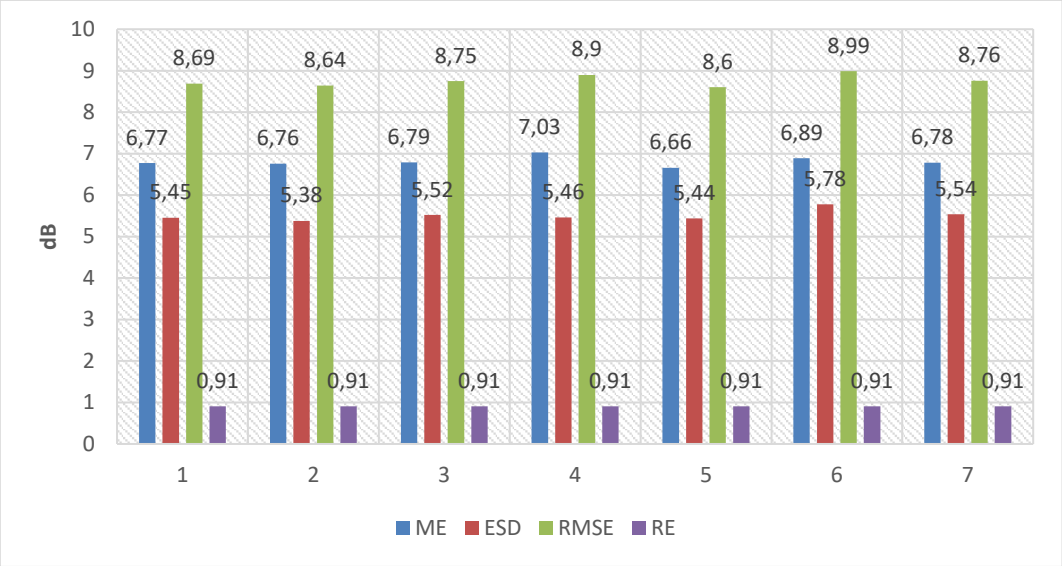

Figura 25 – Estatísticas de T3.

Analisando o gráfico acima, conclui-se que todas as combinações, através da utilização da informação de *clutter* como atributo, apresentam melhores resultados, comparativamente com os obtidos, utilizando apenas os AG.

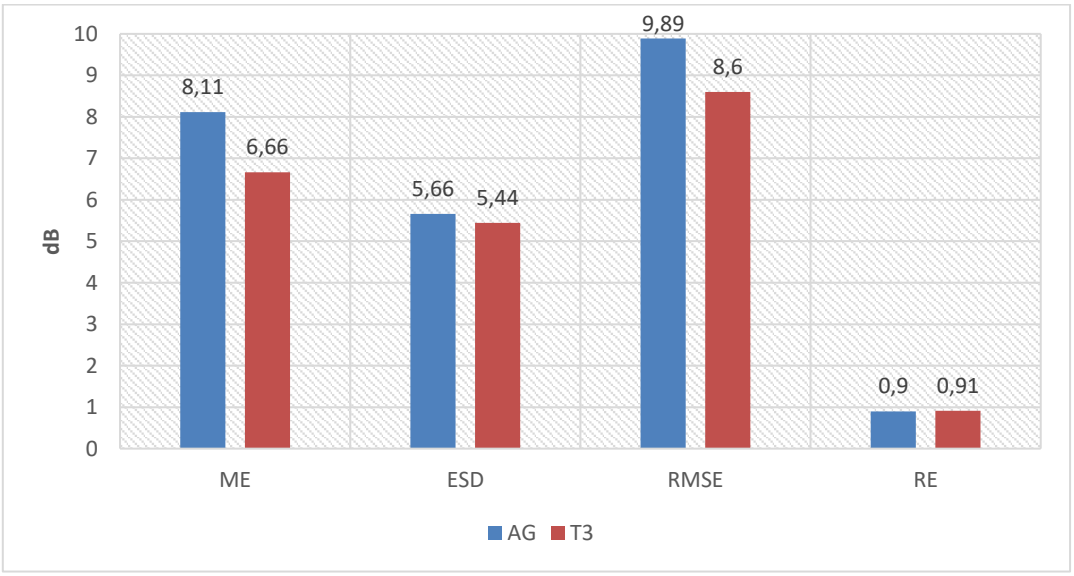

Figura 26 – Comparação entre as estatísticas AG e as melhores estatísticas de T3.

Após observar as estatísticas obtidas utilizando a informação de *clutter*, em conjunto com os atributos relativos ao modelo, decidiu-se realizar uma limpeza dos dados referentes à informação de *clutter*, isto é, eliminar as influências abaixo de um determinado *threshold*, nas imagens (pixels) constantes, em cada ponto, de cada percurso de propagação (ferrovia). Esta limpeza, em conjunto com uma redução das classes de *clutter*, visa evitar os possíveis efeitos de concentração, causados por atributos irrelevantes, de modo a melhorar os resultados obtidos até então. Os testes realizados, relativos à exclusão baseada num dado limite, incluíram as percentagens de 10%, 15% e 20%. A definição de um *threshold* de 15%, é a que apresenta melhores resultados de *clustering* e, como tal, é a escolhida para a limpeza das classes de *clutter*.

As estratégias de redução da dimensão dos dados, aos quais é aplicado o algoritmo de *clustering* desenvolvido, são baseadas nas definições das classes de *clutter* utilizadas. Esta redução consiste na criação de novos atributos, construídos através da junção de classes de *clutter* que partilhem semelhanças, isto é, que apresentem características, que em contexto de propagação de sinal rádio, possam ser consideradas semelhantes.

A Tabela 4 apresenta as classes finais de *clutter*, obtidas após o processo de redimensionamento de dados.

1	*Water*	1	*Sea*
		2	*Inland water*
		3	*Wetland*
2	*Vegetation*	7	*Woodland*
		8	*Forest*
		10	*Suburban*
3	*Urban*	9	*Village*
		11	*Dense Suburban*
		12	*Urban*
		13	*Dense Urban*
		14	*Core Urban*
		15	*Building Blocks*
		16	*Industrial*

4	*Open*	4	*Barren*
		5	*Grass/Agriculture*
		6	*Rangeland*
		17	*Airport*
		18	*Open In Urban*

Tabela 4 – Classes finais de *clutter*.

O critério de decisão utilizado para a junção das classes 1, 2 e 3, responsável pela criação do atributo "*Water*", consiste no facto destas classes serem caracterizadas por áreas cobertas de água. O atributo "Vegetation", formado através da interligação das classes 7, 8 e 10, é caracterizado por áreas cobertas de vegetação e/ou espécies de árvores. O atributo "*Urban*" surgiu da união de todas as classes compostas por edifícios e características urbanas, bem como, da consideração de que edificios com uma altura variável de 20 a 40 metros, provocam o mesmo efeito na comunicação rádio, do ponto de vista da antena móvel, que tem 4 metros de altura. A decisão da criação do atributo "*Open*", constituído pelas classes 4, 5, 6, 17 e 18, assenta no facto de estas possuírem como característica comum, áreas de terreno aberto, sem obstáculos e com pouca, a nenhuma, vegetação. A classe de *clutter* 19, designada "*Unclassified*", foi eliminada devido ao facto de introduzir irrelevância, em termos de características, no conjunto global de dados, sendo, maioritariamente, representada por uma influência igual a 0.

As 4 classes de *clutter* formam o quarto conjunto de atributos dos dados a agrupar. A Figura 27 apresenta a comparação entre as estatísticas resultantes desta seleção (T4), provenientes de AP, e as estatísticas finais, provenientes de AG.

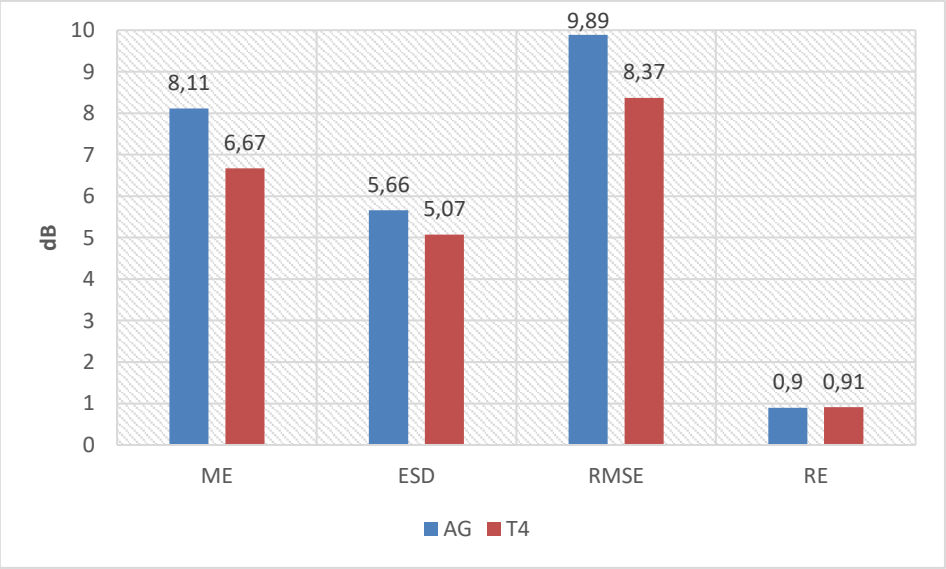

Figura 27 – Comparação entre as estatísticas AG e as estatísticas de T4.

Os resultados estatísticos de T4, ilustrados no gráfico acima, refletem uma diminuição do erro global na predição do sinal rádio, comparativamente com os resultados obtidos em T2, utilizando as 19 classes de *clutter*. Validando, assim, a limpeza efetuada, bem como a redução da dimensão dos dados, através da junção das classes de *clutter* consideradas semelhantes, em contexto de propagação rádio.

O primeiro conjunto de dados, caracterizado pelos atributos distância, d, altura efectiva da antena da estação base, hbe, e parâmetro do obstáculo principal, $v1$, em conjunto com as 4 classes de *clutter*, resultantes do processo de redução dos dados, formam o último conjunto de dados a ser selecionado e, posteriormente, agrupado. Admitindo a informação de *clutter* como base, os testes realizados consideram todas as combinações possíveis. A análise destes resultados é apresentada no Capítulo 4 – Resultados, no qual se encontra a configuração final do algoritmo de *clustering* desenvolvido.

Validação do *clustering*

A validação do *clustering* traduz o quão bem o algoritmo de agrupamento descobriu os *clusters* do conjunto de dados [26].

A técnica de validação dos resultados do *clustering* visa, através de múltiplas execuções do KMP, aplicando um certo número de configurações de agrupamento, descobrir a configuração que produz resultados com melhor qualidade, isto é, a que constrói um agrupamento caracterizado por grupos o mais compactos e separados possiveis.

O método de validação utilizado, tem o objectivo de estimar o melhor valor de *K*, isto é, a quantidade de *clusters* que o KMP deve construir, de modo a obter *clusters* capazes de discriminar, com qualidade, o conjunto de elementos de dados, X, a agrupar. No entanto, com o objectivo de, primeiro, visualizar a distribuição dos pontos de dados, de modo a verificar se existem *clusters* naturais intrínsecos ao conjunto de elementos de dados, apresenta-se a Figura 28.

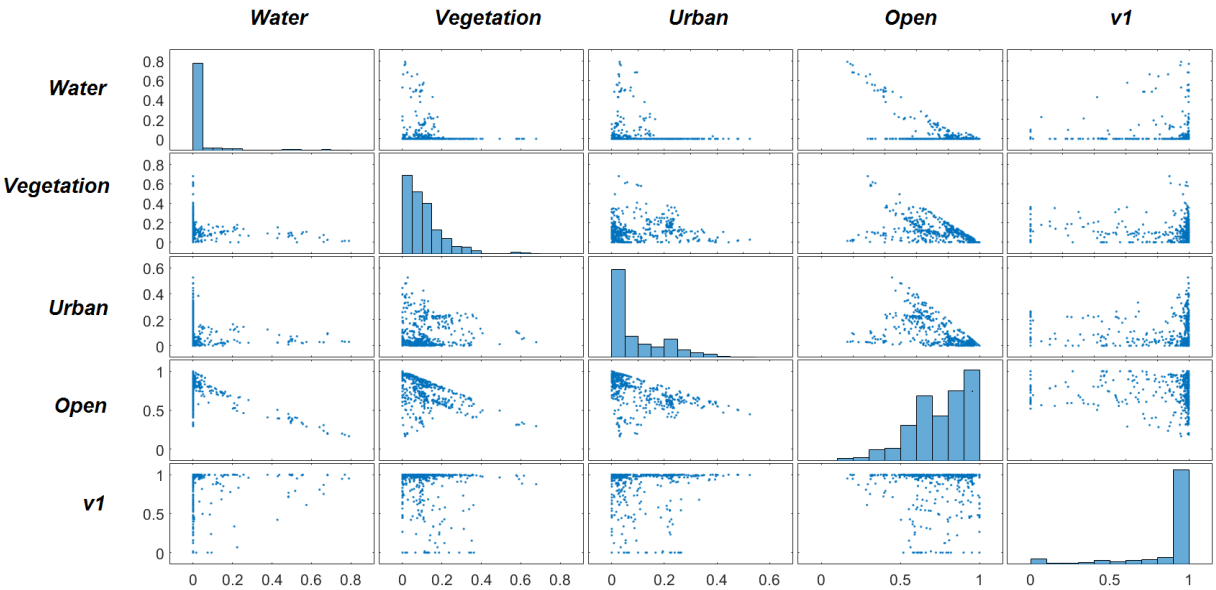

Figura 28 – Distribuição dos elementos do conjunto de dados utilizado para *clustering*.

A matriz de figuras apresentada acima, ilustra, por cada linha i, uma coluna j, correspondente à comparação entre a distribuição dos elementos de dados do atributo i de X, com a do atributo j de X. Por exemplo, na segunda coluna da linha "*Water*" da matriz de figuras, apresenta-se a comparação da distribuição dos elementos de dados do atributo "*Water*", com a distribuição dos elementos de dados do atributo "*Vegetation*". Ao longo da diagonal da matriz de figuras, apresentam-se os histogramas relativos à distribuição dos elementos de dados, de cada atributo de X.

A Figura 28 enfatiza a dificuldade na pesquisa, visual, de formações de *clusters* naturais, em conjuntos de dados caracterizados por dimensões elevadas. Após ter-se provado ser bastante dificil, a visualizaçao de *clusters* naturais no conjunto dos elementos de dados definidos para o processo de *clustering*, aposta-se num índice de validação, relativo, de *clustering*.

O índice de validação escolhido trata-se da percentagem de variância explicada, a qual avalia a relação entre a variância entre *clusters* e a variância dentro dos *clusters*. Quanto menor for o valor deste índice, maior é a dispersão no interior de um *cluster*. Quanto maior for o valor da variância explicada, menor é a dispersão no interior de um *cluster* (mais compactos são).

A premissa da abordagem implementada consiste em escolher o melhor esquema de *clustering*, a partir de um conjunto de regimes definidos, caracterizados por diferentes valores de *K*. O objectivo é obter o valor que melhor se ajuste aos dados definidos. O procedimento de identificação do melhor esquema de *clustering*, baseado na percentagem de variância explicada, é constituido pelas seguites etapas:

- Executar o KMP para um intervalo de valores de *K*, definido entre um valor mínimo e um máximo ($nk = \{1,2,3,4,5,6\}$).
- Para cada um dos valores de nk, executar o KMP r vezes, guardando o melhor valor, correspondente à percentagem de variância explicada, obtido por cada valor de *K* (nk).

Visualmente, este resultado pode ser obtido traçando um gráfico da percentagem de variância explicada, em função de *K*. O valor de *K*, para o qual ocorra uma mudança significativa do valor do índice, tipicamente, apresenta-se ilustrado por um "cotovelo", sendo esta a localizaçao correspondente ao número de *clusters* subjacentes ao conjunto de dados. Caso não seja possivel identificar visualmente, uma mudança significativa no valor do indice de validação, é escolhido o numero de *clusters* que corresponder a um resultado da percentagem da variância explicada, superior a 90%.

Inicialização dos K *centroids*

Após determinado o valor de K, são iniciadas as etapas inerentes à execução do algoritmo de *clustering* KMP, começando pela inicialização dos K *centroids*. A escolha dos *centroids* iniciais é um passo fundamental do algoritmo base de *K-Means*. Quando os *centroids* são escolhidos aleatoriamente, diferentes execuções do algoritmo de *clustering*, produzem diferentes resultados relativamente à soma

do erro quadrático. Acrescentando o facto de, os *clusters* resultantes serem, tipicamente, pobres, quer em termos de coesão, quer em termos de extracção de informação [33].

Arthur e Vassilvitskii propõem em [34] uma etapa de inicialização ponderada, através do desenvolvimento do algoritmo de inicialização *K-Means*++, cujas etapas, inerentes à sua execução, são apresentadas em seguida, sendo $D(x)$, a menor distância a partir de um elemento de dados, até ao *centroid* mais próximo, previamente definido.

1. Selecionar o primeiro *centroid*, $C1$, aleatoriamente de X.
2. Selecionar um novo *centroid* Ci, considerando $x \in X$ com probabilidade $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$
3. Repetir o passo 2, até obter K *centroids*.

De acordo com Arthur e Vassilvitskii, *K-Means* ++ melhora o tempo de execução do algoritmo de Lloyd (*K-Means*), bem como a qualidade da solução final. Arthur e Vassilvitskii demonstraram, usando um estudo de simulação para diversas orientações de *cluster*, que o *K-Means*++ alcança uma convergência mais rápida, obtendo *clusters* mais compactos, comparativamente ao algoritmo de Lloyd. Como tal, é o método de inicialização utilizado, na execução do KMP.

Atribuição de elementos aos *centroids*

Nesta etapa é realizada o agrupamento, propriamente dito, sendo esta executada até à convergência do algoritmo de *clustering*, isto é, até à ausência de novas atribuições de elementos aos *centroids*.

A atribuição dos elementos aos *centroids* mais proximos é realizada através de uma função de medida de proximidade, de modo a quantificar a noção de "mais próximo", para os respectivos elementos do conjunto dados utilizado. O objetivo do *clustering* é, geralmente, expresso por uma função objectiva, dependente das proximidades entre os pontos e os *centroids* [35].

Tal como já foi referido anteriormente, em cenários de altas dimensões, o rácio entre o ponto mais próximo e o mais distante, aproxima-se de 1, isto é, os pontos tornam-se uniformemente afastados uns dos outros. Em [36] é fornecida demonstração teórica e, também, experimental, relativamente à análise da dependência da norma L_m, do valor de m. É demonstrado que os contrastes relativos, das distâncias a um ponto de consulta, dependem fortemente da métrica L_m utilizada. Assim, para um conjunto de dados caracterizado por uma dimensão, d, elevada ($d \geq 3$), torna-se vantajosa a utilização de valores exíguos, de m. O que significa que, a métrica L_1 (distância de *Manhattan*), em aplicações de alta dimensão, oferece maior contraste de dados, comparativamente a L_2 (distância euclidiana). A Figura 29 ilustra a diferença entre as duas métricas referidas.

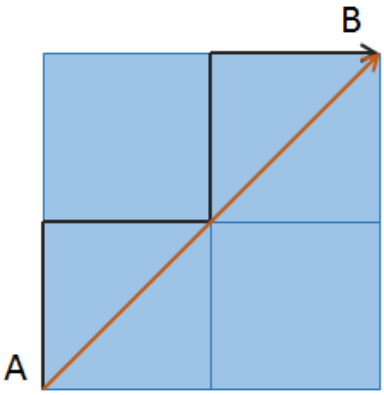

Figura 29 – Distância euclidiana vs distância de *Manhattan.*

Sendo o caminho a laranja, referente à distância euclidiana entre o ponto A e o ponto B. Por outro lado, o caminho a preto corresponde à métrica de *Manhattan*, muitas vezes designada por "*city block distance*".

Informação resultante do *clustering*

Os indices dos *clusters* correspondentes a cada elemento de dados, bem como a localização dos *K centroids*, trata-se da informação proveniente do algoritmo de *clustering* KMP. Os *clusters* resultantes são caracterizados pelos atributos dos elementos pertencentes a esses *clusters*. Aos quais são aplicados cálculos estatisticos com o objectivo de avaliar a variação de cada atributo, em cada *cluster*.

3.3.2 Otimização

A informação geográfica recolhida, juntamente com os parâmetros do modelo, possibilita a elaboração de uma predição, realizada pelo modelo de propagação, descrito no subcapítulo 2.2 – Propagação em Ferrovias. Com base no erro entre a predição e as medidas, o algoritmo de otimização atribui novos parâmetros ao modelo. Os novos parâmetros geram uma nova predição, que é novamente avaliada pelo algoritmo. Este processo repete-se até que seja atingida uma condição de paragem, quer por ter sido atingido um determinado valor de erro, quer por ter sido alcançado o número máximo de iterações.

O descrito no parágrafo acima, trata-se do processo de otimização desenvolvido em [8], sendo este aplicado globalmente, à totalidade dos dados de treino, bem como, parcialmente, por *cluster*.

A informação resultante da otimização global trata-se de um conjunto de parâmetros de configuração do modelo, otimizados para a totalidade dos elementos de dados.

A otimização por *cluster* é obtida, tirando partido dos índices dos *clusters* correspondentes a cada elemento, provenientes do KMP, de modo a filtrar os atributos necessários, correspondentes à informação contida em cada *cluster*, para a realização do processo de otimização. A informação resultante da otimização por *cluster* trata-se de um conjunto de parâmetros de configuração do modelo,

otimizados para os elementos presentes nesse mesmo *cluster*.

No decorrer do processo de aprendizagem, introduziu-se o efeito de réplicas nos AG, isto é, são executados múltiplas vezes, com o objetivo de encontrar o conjunto de parâmetros de configuração do modelo, correspondente ao melhor *fitness*, ou seja, o mais otimizados possíveis, tanto em termos globais, como em termos parciais (por *cluster*). O mesmo conceito é aplicado ao KMP, visando diminuir erros de agrupamento, de modo a obter a solução com os *clusters* mais compactos possíveis.

3.4 Processo de Teste

Para a fase de teste (Figura 30), são direcionados os restantes 25%, da informação aleatoriamente amostrada. Após a classificação dos elementos de teste, nos respetivos *clusters*, a informação geográfica previamente recolhida, juntamente com os parâmetros de configuração, provenientes de ambos os algoritmos, são introduzidos no modelo de propagação e é elaborada uma predição final. No fim, é realizada uma comparação da predição, proveniente de ambos os algoritmos, com as medidas previamente realizadas, utilizando estatísticas de primeira ordem e o coeficiente de correlação.

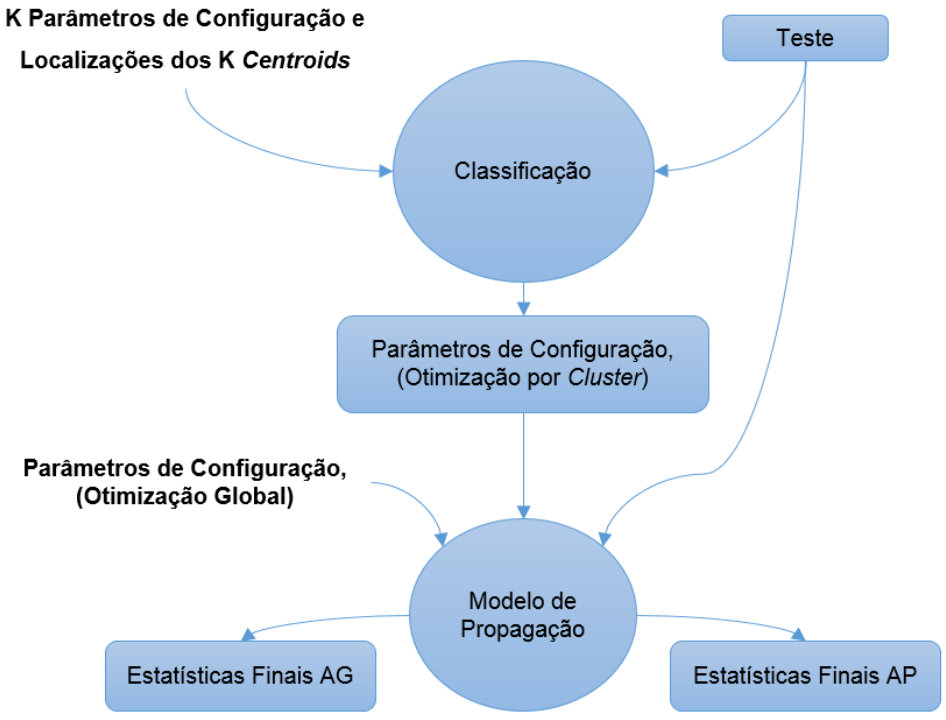

Figura 30 – Diagrama de blocos da fase de teste.

3.4.1 Classificação

A classificação [37] é o processo de encontrar um modelo que descreva e distinga um elemento de dados, com o objectivo de usar esse mesmo modelo para predizer a categoria de elementos, cuja descrição / etiqueta é desconhecida. O modelo derivado é baseado na análise de um conjunto de dados de treino.

Tal como foi referido anteriormente, os *clusters* resultantes, após a aplicação do *clustering* ao conjunto de dados, são caracterizados pelos atributos dos elementos pertencentes a esses *clusters*. Em seguida, um elemento desconhecido pode ser classificado num *cluster* específico, com base na semelhança entre os seus atributos e os dos já definidos *clusters*.

O cálculo da distância entre um elemento de teste e as *K* localizações dos *centroids* é realizado, utilizando a mesma métrica que foi usada no decorrer do processo de agrupamento. A menor, das *K* distâncias resultantes, revela o grupo a que o elemento de teste é pertencente. Sendo o conjunto de parâmetros de configuração, correspondentes ao *cluster* resultante da classificação, aplicado ao respetivo elemento de teste.

3.4.2 Modelo de Propagação

O modelo do Okumura-Hata não contabiliza as perdas devido à difração resultante dos obstáculos, portanto, para tais efeitos, considerou-se um modelo que os contabiliza. O modelo utilizado, para o cálculo da predição de cobertura rádio em GSM-R, é composto pelo modelo de Okumura-Hata com os respetivos fatores corretivos, pelo método de Deygout, de modo a contabilizar as perdas adicionais devido à difração, permitindo obter uma maior precisão no cálculo das perdas totais e, ainda, pelo método baseado na recomendação ITU-R P.1546, a qual demonstrou ser a mais benéfica [8] para a determinação da altura efetiva da antena da estação base.

Relativamente às estatísticas resultantes, considera-se um conjunto de parâmetros de configuração do modelo, mais otimizado, quanto menor for o desvio resultante da predição, calculada através desses parâmetros, relativamente às medidas reais. Para uma melhor comparação entre a predição e as medidas calculam-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 seguintes equações:

$$\text{Erro médio absoluto} = ME = \frac{1}{n}\sum_{i=1}^{n}\left|P_{measi} - P_{predi}\right| \quad (14)$$

$$\text{Raiz do erro quadrático médio} = RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|P_{measi} - P_{predi}\right|^2} \quad (15)$$

$$\text{Desvio padrão do erro} = ESD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|P_{measi} - P_{predi}\right| - ME\right)^2} \quad (16)$$

onde P_{measi} é o nível de sinal (em dBm) do sinal medido no ponto i , sendo n, o número total de pontos e P_{predi}, o valor correspondente da predição. O cálculo do coeficiente de correlação é dado por:

$$RE = \frac{\sum_{i=1}^{n}(P_{measi} - \bar{P}_{meas})(P_{predi} - \bar{P}_{pred})}{\sqrt{\sum_{i=1}^{n}(P_{measi} - \bar{P}_{meas})^2}\sqrt{\sum_{i=1}^{n}(P_{predi} - \bar{P}_{pred})^2}} \quad (17)$$

Métodos de interpretação dos resultados

Os *clusters* resultantes, após aplicado o *clustering* a um conjunto de dados, são caracterizados pelos atributos dos elementos pertencentes a esses *clusters*. Possibilitando a classificação de um elemento desconhecido num *cluster* específico, com base na semelhança entre os seus atributos e os dos já definidos *clusters*. Assim, torna-se possível a extracção de conhecimento útil, relativamente aos dados iniciais. De modo a avaliar a variação de cada atributo, presente em cada *cluster*, são aplicados cálculos estatisticos, nomeadamente a média e o desvio padrão.

Para um conjunto de dados, a média é a soma das observações divididas pelo número de observações. Esta identifica a localização central dos dados. O desvio padrão mede a difusão, isto é, a variação do conjunto de dados, bem como a relação da média com o resto dos dados. Se os elementos, de um dado atributo, estiverem situados perto da média, indicando uma presença uniforme de uma dada característica, o desvio padrão será pequeno. Por outro lado, se muitos elementos de dados, de um atributo, estiverem longe da média, o desvio padrão será grande, traduzindo uma difusão elevada, relativamente à presença de uma dada característica. Em casos extremos, se todos os elementos, de um atributo, apresentarem valores iguais, o desvio padrão será zero [38].

.

Capítulo 4

Resultados

Este capítulo fornece a configuração final do algoritmo desenvolvido, bem como a análise dos resultados obtidos.

4.1 Configuração Final

Os atributos utilizados no primeiro conjunto de dados (distância, d, altura efectiva da antena da estação base, hbe, e parâmetro do obstáculo principal, $v1$), em conjunto com as 4 classes de *clutter*, resultantes do processo de redução dos dados, formam o conjunto de dados, final, a ser selecionado e, posteriormente, agrupado. Os testes realizados consideram todas as combinações possíveis, admitindo a informação de *clutter* como base:

1. 4 Classes de *clutter* e d;
2. 4 Classes de *clutter* e $v1$;
3. 4 Classes de *clutter* e hbe;
4. 4 Classes de *clutter*, d e $v1$;
5. 4 Classes de *clutter*, d e hbe;
6. 4 Classes de *clutter*, $v1$ e hbe;
7. 4 Classes de *clutter*, $d, v1$ e hbe.

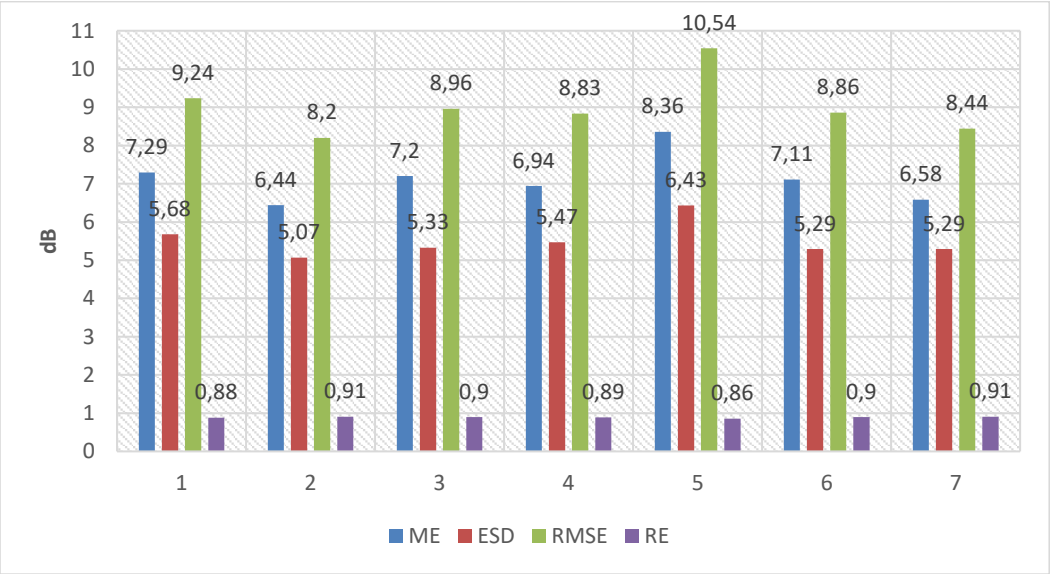

Figura 31 – Estatísticas de T5.

Através da análise do gráfico acima, conclui-se que é obtido o melhor resultado, através da seleção das 4 classes de *clutter*, juntamente com o parâmetro $v1$. A Figura 32 apresenta a comparação entre as melhores estatísticas resultantes ($ME, ESD, RMSE$ e RE), correspondentes a esta seleção, provenientes de AP, e as estatísticas finais, provenientes de AG.

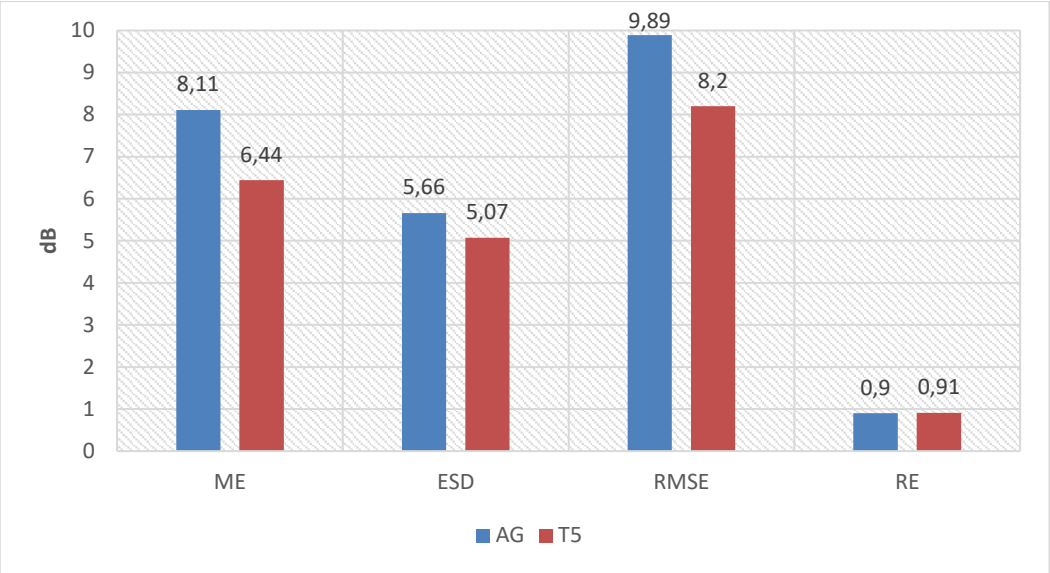

Figura 32 – Comparação entre as estatísticas AG e as melhores estatísticas de T5.

A combinação das 4 classes de *clutter*, com o parâmetro $v1$, promove o melhor resultado estatístico final, isto é, a maior diminuição do erro na estimativa do valor da atenuação, comparativamente aos valores obtidos em [8].

Para a estimação do valor de *K* que melhor se ajuste aos dados definidos, utiliza-se o índice de validação, descrito no Capítulo 3 – A Associação de *Clustering* a Otimização, sendo este baseado na percentagem de variância explicada. Executando o KMP, múltiplas vezes, para o intervalo de valores de *K*, previamente definido, e guardando o melhor valor correspondente à variância explicada, obtido por cada valor de *K* (nk), traçou-se o gráfico representado pela Figura 33, da variância explicada, em função de *K*.

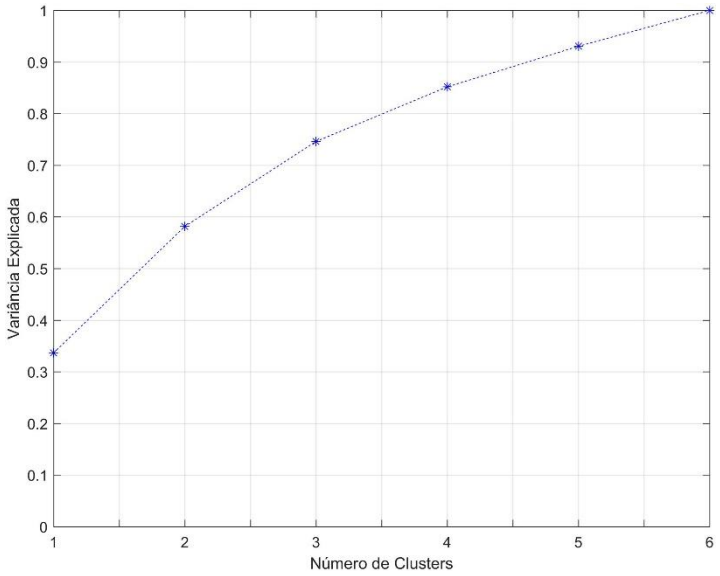

Figura 33 – Estimação do valor de *K*.

Analisando o gráfico acima, verifica-se que a partir de $K = 5$, apesar de, visualmente, não ser facilmente percetível, o valor da variância explicada, sofre uma mudança significativa, e o facto de corresponder a um resultado superior a 90%, conclui-se ser esta, a localização correspondente ao número de *clusters* subjacentes ao conjunto de dados. Para o conjunto de dados a agrupar, constituído pelas 4 classes de *clutter* e pelo parâmetro relativo ao obstáculo principal $v1$, o valor de *K* resultante, do processo de validação, é igual a 5.

4.2 Análise dos Resultados

Após terem sido descobertos os argumentos ideais, representativos da configuração final do algoritmo de *clustering* KMP, é realizada a respetiva seleção e pré-processamento dos elementos de dados, a serem agrupados em 5 *clusters*. Inicializando os 5 *centroids* através do método de Arthur e Vassilvitskii e realizando a atribuição dos elementos aos *centroids* mais próximos, com base na distância de *Manhattan*, são construídos 5 *clusters*.

4.2.1 Análise de *Clusters*

Os *clusters* resultantes, tal como já foi referido anteriormente, são caracterizados pelos atributos dos elementos pertencentes a esses *clusters*. Assim, torna-se possível a extração de conhecimento útil, relativamente aos dados iniciais. De modo a avaliar a presença de cada atributo, em termos de variação de dados, são aplicados cálculos estatísticos, nomeadamente a média e o desvio padrão, aos valores iniciais (reais) dos atributos presentes em cada *cluster*. A média fornece a localização central de um conjunto de dados. O desvio padrão descreve a dispersão dos dados, bem como a sua distribuição em torno da média. Um desvio padrão de valor exíguo, indica que os dados se encontram agrupados em torno da média. Um valor maior, revela a existência de dispersão de dados [39].

As figuras seguintes ilustram a caracterização dos atributos dos elementos de dados, presentes em cada *cluster*, sendo as 4 classes de *clutter*, os primeiros atributos a serem analisados.

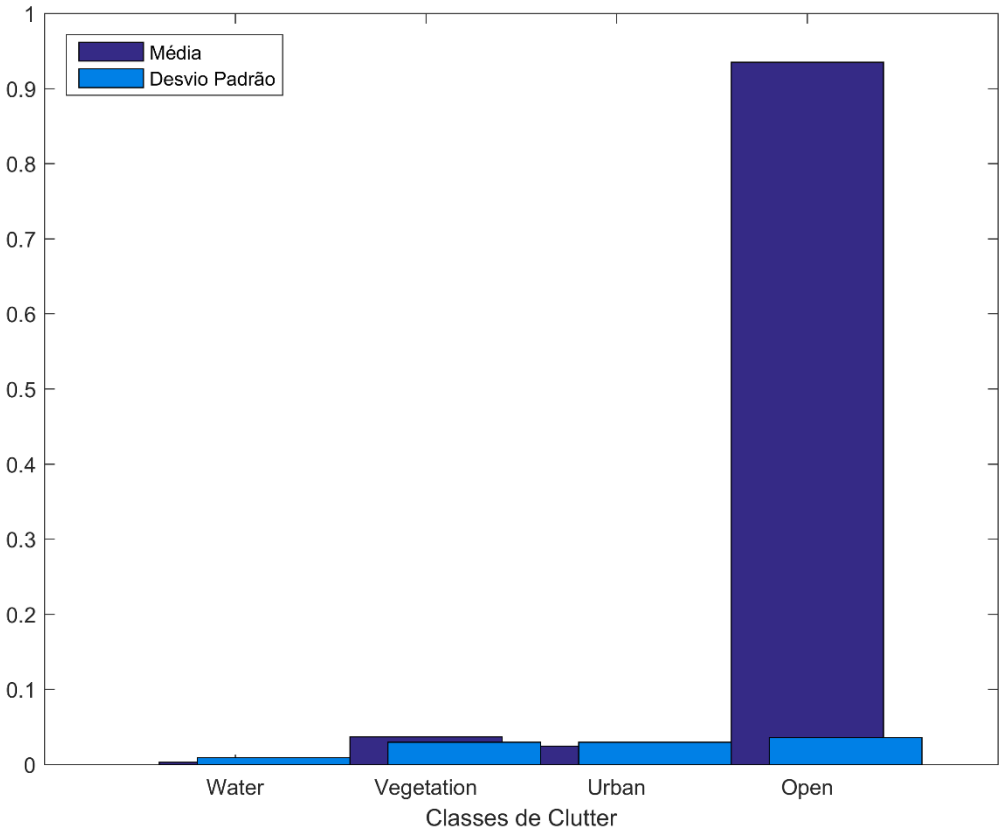

Figura 34 – Atributos dos elementos de dados presentes no *cluster* 1.

Avaliando os valores resultantes, dos cálculos das estatísticas aplicados a cada atributo, bem como a relação existente entre ambos (média e desvio padrão), conclui-se que o primeiro *cluster* é caracterizado predominantemente pela presença de áreas de terreno aberto, apresentando em média, uma influência de, aproximadamente, 90%. Esta, representada pelo atributo "*Open*", traduz-se numa presença uniforme dos dados, tendo em conta o valor diminuto do respetivo desvio padrão.

Os atributos representados por desvios padrões superiores aos valores das respetivas médias ("*Water*" e "*Urban*"), revelam uma presença bastante dispersa e tendo em conta o valor médio de ambos, são considerados nulos. A presença de vegetação é, também, considerada nula.

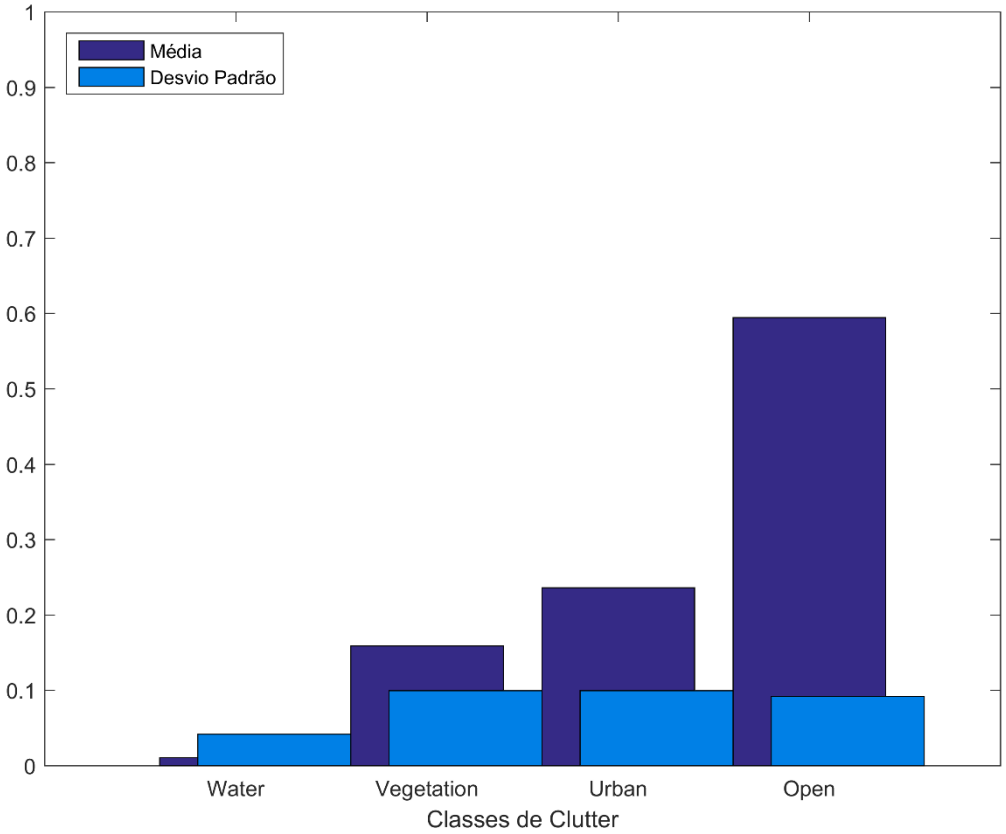

Figura 35 – Atributos dos elementos de dados presentes no *cluster* 2.

No segundo *cluster*, apesar de, em média, a influência da vegetação ser superior a 15%, visto que este atributo possui um desvio padrão elevado, relativamente ao seu valor médio, conclui-se que o atributo "*Vegetation*" apresenta uma presença dispersa de áreas cobertas por plantação e / ou copas de árvores. Analisando a relação desvio padrão / média, do atributo "*Urban*", conclui-se que a presença de áreas com características urbanas é ligeiramente uniforme, possuindo uma influência superior a 20%. Este *cluster* é caracterizado predominantemente pela presença de áreas de terreno aberto, centralizados num valor médio de 60%, e pela ausência de áreas cobertas de água.

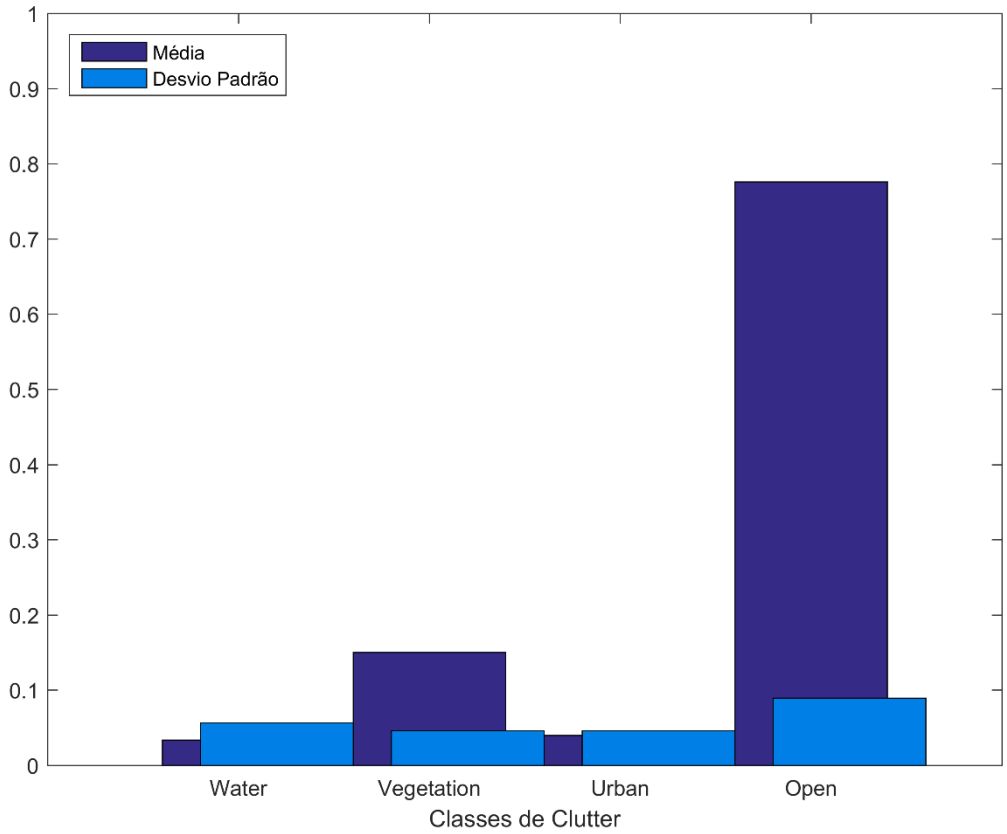

Figura 36 – Atributos dos elementos de dados presentes no *cluster* 3.

O terceiro *cluster* é caracterizado maioritariamente por áreas de terreno aberto, apresentando em média, uma influência superior a 75%. Esta, representada pelo atributo "*Open*", traduz-se numa presença uniforme dos dados, tendo em conta o valor mínimo do respetivo desvio padrão. Os atributos "*Water*" e "*Urban*", representados por desvios padrões superiores aos valores das respetivas médias, são considerados nulos. O atributo "Vegetation" possui um desvio padrão pequeno, relativamente ao seu valor médio, traduzindo uma influência de aproximadamente 15%, de áreas cobertas por plantações e/ou árvores.

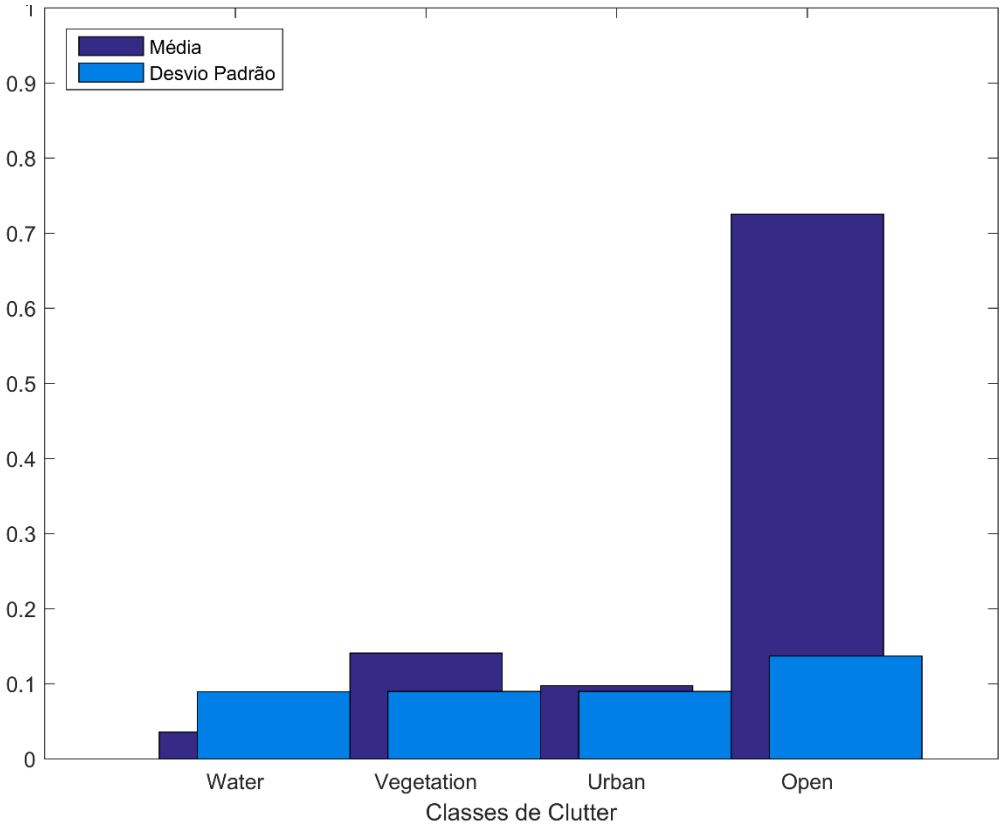

Figura 37 – Atributos dos elementos de dados presentes no *cluster* 4.

No quarto *cluster*, a influência da vegetação apresenta um valor médio superior a 15%, no entanto, visto que este atributo possui um desvio padrão elevado, relativamente ao seu valor médio, conclui-se que "*Vegetation*" apresenta uma presença dispersa de áreas cobertas de vegetação. O desvio padrão referente à influência do atributo "*Urban*" apresenta-se perto do seu valor médio, indicando também, uma elevada dispersão de dados, revelando uma presença bastante dispersa de áreas cobertas por edifícios ou de características urbanas. Este *cluster* é caracterizado maioritariamente por áreas de terreno aberto, apresentando em média, uma influência de, aproximadamente, 70%. Esta, representada pelo atributo "*Open*", traduz-se numa presença uniforme dos dados, tendo em conta o valor diminuto do respetivo desvio padrão

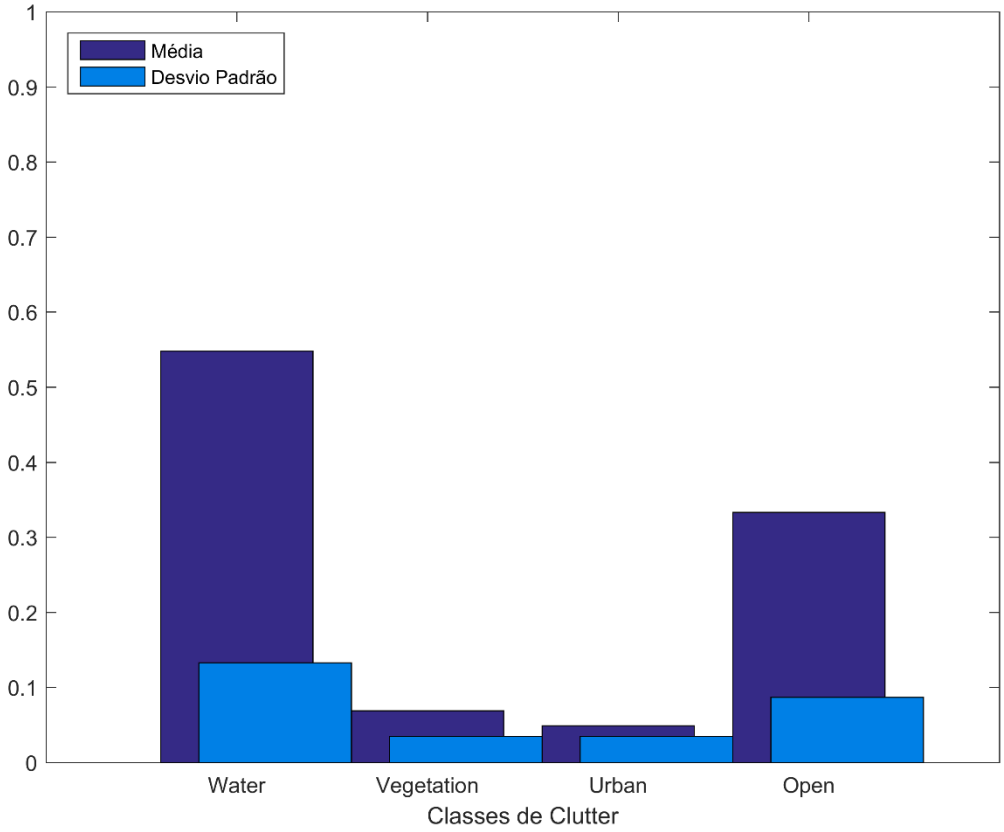

Figura 38 – Atributos dos elementos de dados presentes no *cluster* 5.

O quinto *cluster* é caracterizado maioritariamente por áreas cobertas de água, apresentando em média, uma influência superior a 50%. Esta, representada pelo atributo "*Water*", traduz-se numa presença uniforme de áreas inundadas, tendo em conta o valor exíguo do respetivo desvio padrão. A presença de áreas caracterizadas por terrenos abertos apresenta, em média, uma influência superior a 30%. Esta, representada pelo atributo "*Open*", traduz-se numa presença uniforme dos dados, tendo em conta o valor diminuto do respetivo desvio padrão.

O desvio padrão referente à influência do atributo "*Vegetation*" apresenta-se perto do seu valor médio, indicando uma elevada dispersão, relativamente à presença de áreas de plantação e / ou árvores. O atributo "Urban", segundo o mesmo critério de avaliação, apresenta, também, uma elevada dispersão de dados, revelando uma presença bastante dispersa de áreas dentro do perímetro urbano.

No conjunto total dos elementos de dados, os valores que o parâmetro representativo da presença de obstáculos, $v1$, assume, variam de -1.8 a 0 e, por isso, a análise da sua variação, por *cluster*, é realizada singularmente, numa escala apropriada, diferente da apresentada anteriormente na análise dos atributos referentes às classes de *clutter*.

Para uma melhor análise, em termos de visualização do atributo, realiza-se uma categorização do

parâmetro v1, dividindo-o em 4 categorias:

- *Ingored Obstacles:* valores para os quais os obstáculos são considerados desprezáveis;
- *Low Obstacles:* valores para os quais os obstáculos são considerados pequenos;
- *Medium Obstacles:* valores para os quais os obstáculos são considerados médios;
- *High Obstacles:* valores para os quais os obstáculos são considerados elevados.

Esta divisão é baseada num estudo prévio, o qual consistiu na realização de um histograma (Figura 39), com o objetivo de visualizar a distribuição dos pontos de $v1$, ao longo do intervalo de valores assumidos por este parâmetro.

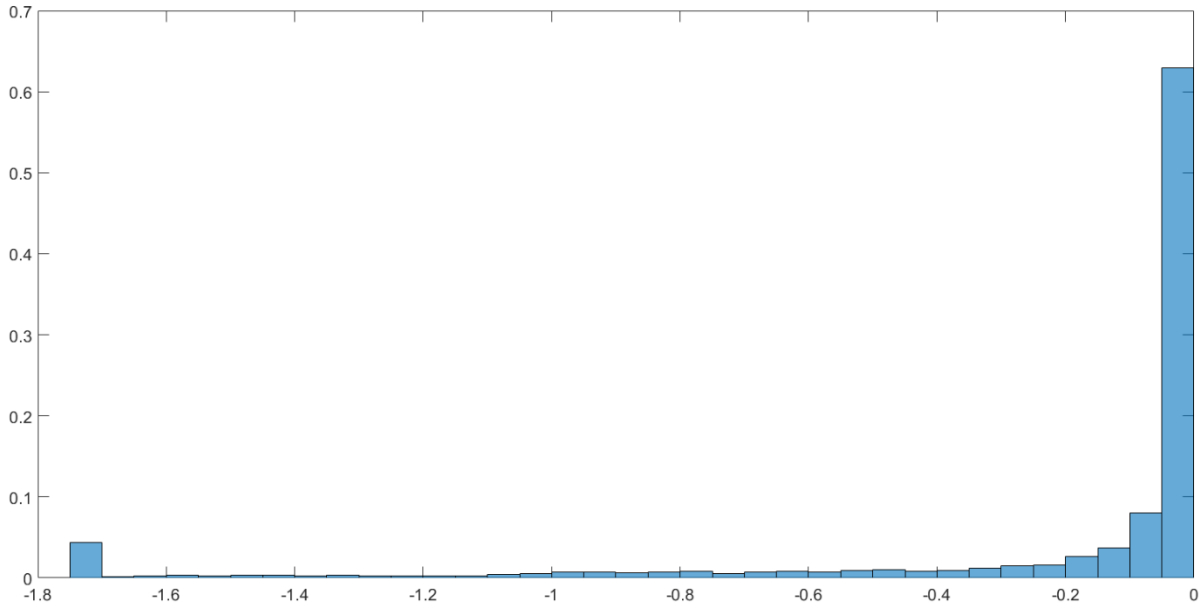

Figura 39 – Histograma de $v1$.

A altura de cada barra é o número relativo de observações, isto é, a relação: $número\ de\ observações\ no\ intervalo\ da\ barra\ /\ número\ total\ de\ observações$. Para efeitos de estimação do valor da atenuação e tendo em conta a distribuição ilustrada acima, consideraram-se os seguintes intervalos, em cada categoria:

- *Ingored Obstacles:* $-1.8 < v1 < -0.7$
- *Low Obstacles:* $-0.7 < v1 < -0.2$
- *Medium Obstacles:* $-0.2 < v1 < -0.05$
- *High Obstacles:* $-0.05 < v1 < 0$

As figuras seguintes ilustram a caracterização dos *clusters*, relativamente à presença de obstáculos, através da análise dos cálculos estatísticos (média e desvio padrão) de cada categoria, do parâmetro $v1$.

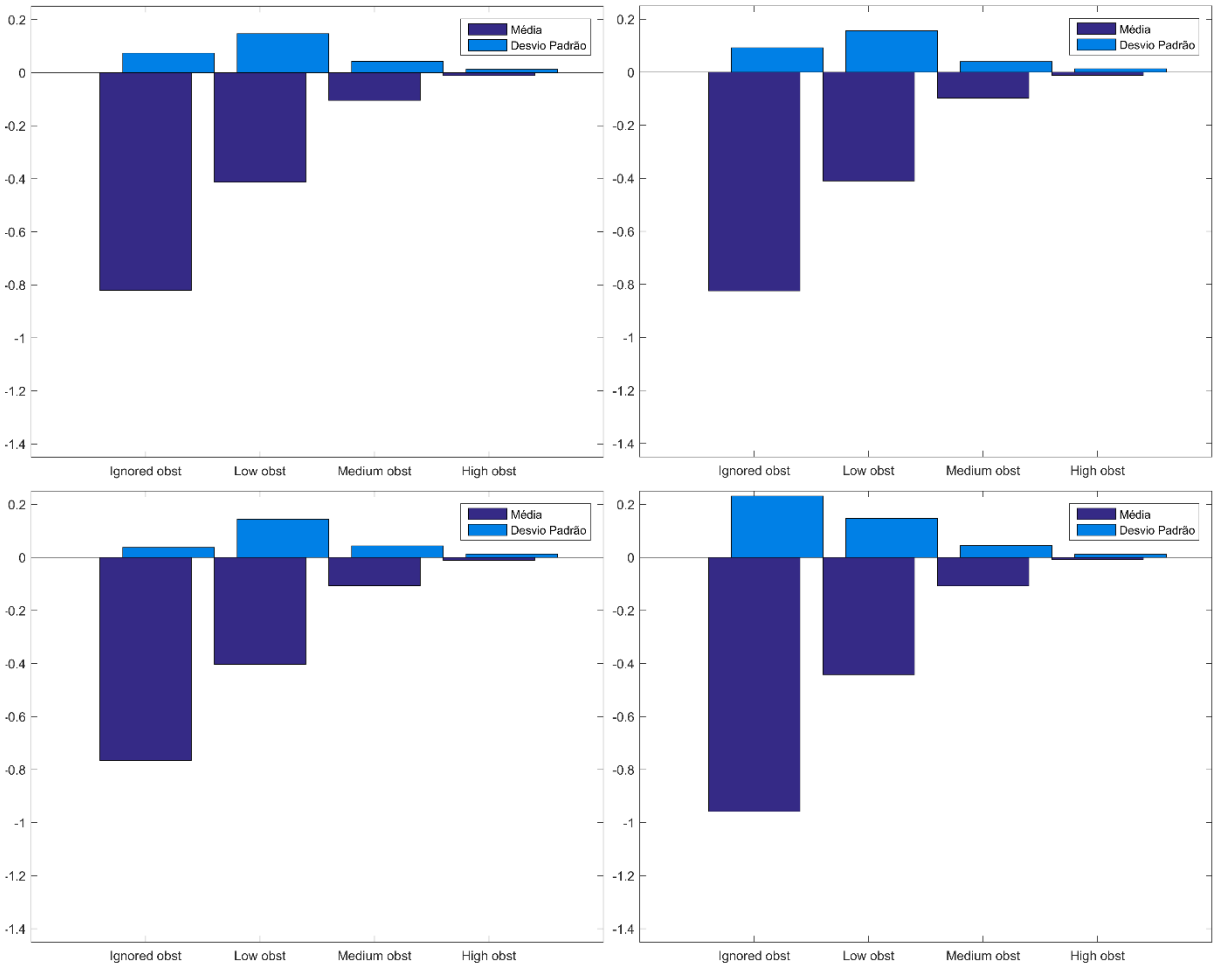

Figura 40 – Análise da presença de obstáculos nos *clusters* 1, 2, 3 e 5.

Analisando a figura acima, conclui-se que a presença das 4 categorias de obstáculos, previamente definidas, representadas pelo valor médio e pelo desvio padrão de cada uma, mantém-se constante nos *clusters* que se apresentam ilustrados. O *cluster* 4, ilustrado pela Figura 41, é o único, dos 5 *clusters* construídos, que apresenta uma categoria predominante ("*Ignored obst*"), sendo esta descrita pela indicação de que não existem obstáculos. A categoria definida apresenta-se caracterizada por um valor médio de $-1{,}4$ e por um desvio padrão, diminuto relativamente à média, de 0.2, o que indica uma presença uniforme dos elementos de dados em torno de $v1 = -1.4$. Este valor revela que a totalidade do lóbulo de transmissão do sinal rádio se encontra desobstruído.

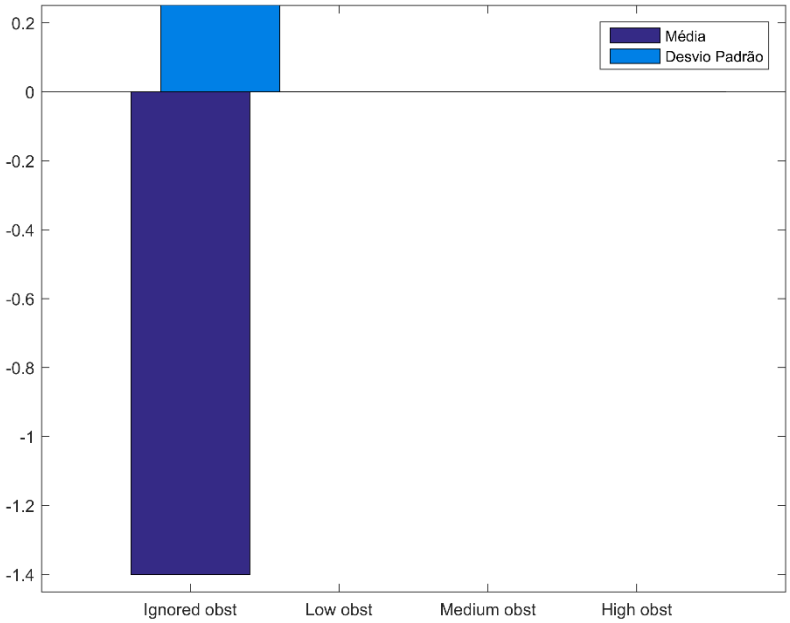

Figura 41 – Análise da presença de obstáculos no *cluster* 4.

Os cálculos estatísticos, relativos à caracterização da presença de obstáculos, através do parâmetro v1, não revelaram ser muito discriminativos na extração de informação útil, em termos de análise dos valores reais, utilizados para o processo de *clustering*. No entanto, a inclusão de v1 na seleção ótima de atributos do conjunto de elementos de dados, presente na configuração final do algoritmo de *clustering*, revelou ser vantajosa, em termos de melhoria de resultados através da associação de KMP a AG.

Através da comparação das medidas efetuadas, com a predição do sinal rádio obtida através dos AG e com a obtida através do AP, torna-se possível a visualização da melhoria do ajuste das curvas. A Figura 42, referente a um percurso de ferrovia de Cascais, trata-se de um exemplo ilustrativo da comparação entre os pontos relativos às medidas efetuadas (a vermelho) e as curvas correspondentes à predição do sinal rádio, quer utilizando os AG (a azul), quer utilizando a associação proposta de KMP a AG (a magenta).

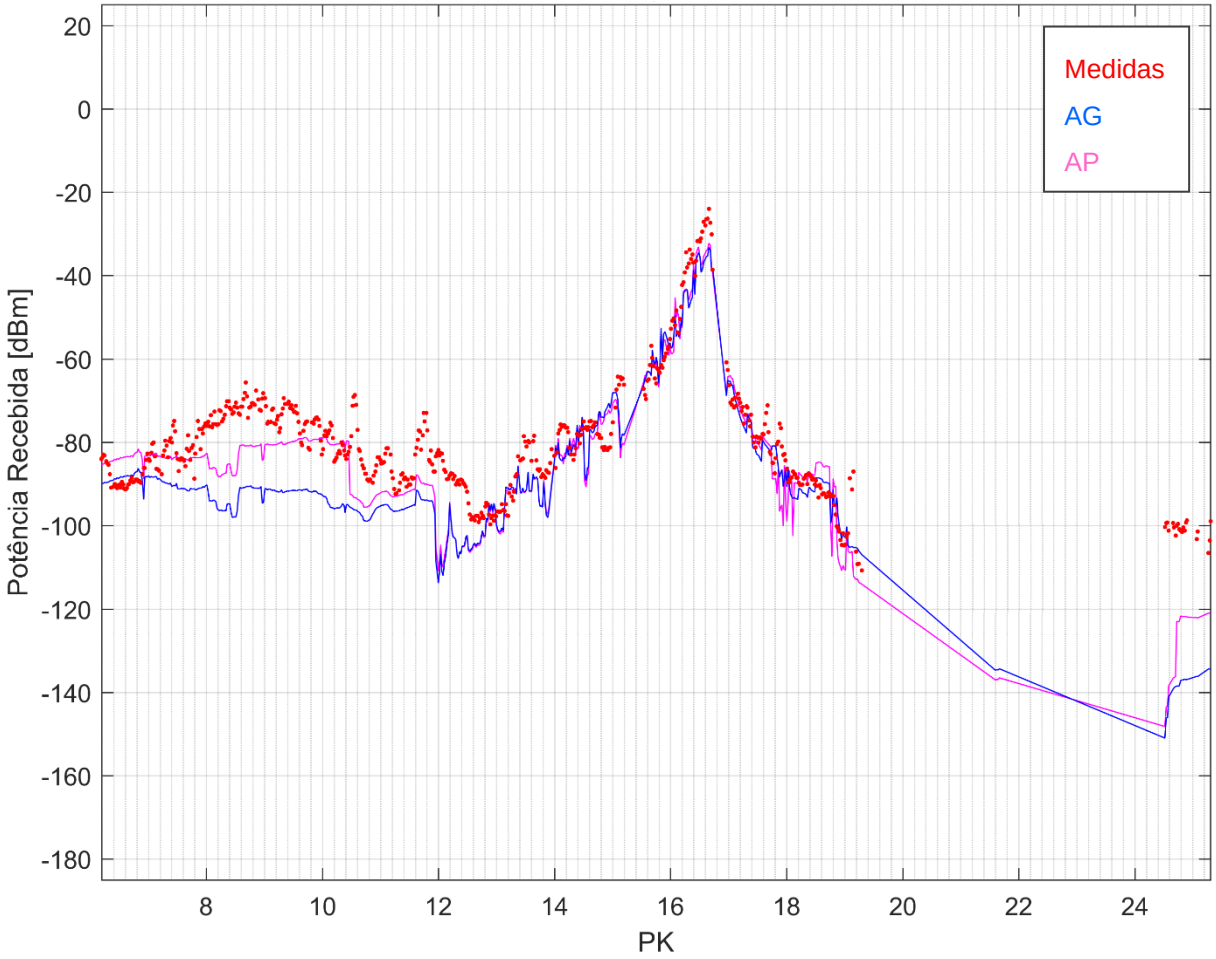

Figura 42 – Comparação entre os pontos das medidas e as curvas de predição.

Observando a figura acima, verifica-se uma melhoria (do PK 8 ao PK 12), em termos de acompanhamento do sinal referente às medidas, do ajuste de curvas resultante do algoritmo realizado, relativamente ao ajuste resultante, utilizando os AG.

.

Capítulo 5

Conclusões

Este capítulo conclui a presente dissertação, fornecendo, também, aspetos fundamentais relativos ao trabalho a desenvolver no futuro.

5.1 Algoritmo Desenvolvido

5.1.1 Resultados

O objetivo desta dissertação foca-se em associar as vantagens da utilização de modelos de propagação com base na predição de cobertura rádio, a *clustering*, de modo a obter uma classificação prévia dos tipos de ambiente, promovendo a redução do erro global na predição.

A combinação das 4 classes de *clutter*, com o parâmetro $v1$, promove o melhor resultado estatístico final, isto é, a maior diminuição do erro na estimativa do valor da atenuação, comparativamente aos valores obtidos em [8]. Sendo a configuração de *clustering* que melhor se ajusta ao conjunto de dados utilizado, definida pelos seguintes pontos:

- Mapeamento das 19 classes de *clutter*, num número mínimo de classes (4), filtrando influências inferiores a 15%;
- Normalização do atributo, referente ao parâmetro $v1$, utilizando o método *Min-Max*;
- Seleção dos atributos dos elementos do conjunto de dados:
    - 4 classes de *clutter (Water, Vegetation, Urban e Open) e p*arâmetro $v1$.
- Determinação do valor de *K*, utilizando a variância explicada como método de validação do *clustering*;
- Inicialização dos *K centroids* através do algoritmo *K-Means++*;
- Realização do agrupamento utilizando como função de proximidade, a métrica de *Manhattan*.
- Múltiplas execuções do algoritmo KMP;

A aplicação desta configuração de *clustering*, com o objetivo de agrupar o conjunto dos elementos de dados, em subconjuntos que partilhem semelhanças geográficas/morfológicas, associada à aplicação dos AG, de modo a otimizar o conjunto de parâmetros de configuração do modelo, para os elementos presentes em cada um dos grupos obtidos, produz soluções de parâmetros que minimizam o erro da estimativa do valor da atenuação, comparativamente aos valores obtidos, utilizando o algoritmo desenvolvido em [8].

Acrescentando o facto de, através desta associação, ter sido alcançado um desvio padrão do erro de predição do sinal rádio de, aproximadamente, $5{,}1\ dB$. A redução desta estatística revela a possibilidade de uma redução no número de estações base, no planeamento da rede e, por consequência, uma redução nos custos de implementação.

5.1.2 Limitações

A força do *K-Means* reside na sua simplicidade e elegância, no entanto, uma das grandes limitações

deste algoritmo de *clustering*, trata-se do facto de este assumir que os *clusters* a construir são de natureza esférica e de tamanhos semelhantes. A distância do *centroid* de um *cluster* até ao seu elemento mais distante, é análoga ao raio do *cluster*, sendo o processo iterativo, de encontrar elementos de dados mais próximos ao centro do *cluster*, semelhante ao estreitamento do seu raio, visando a obtenção de *clusters* compactos, como esferas [40]. Esta característica torna-se uma desvantagem quando, num conjunto de dados, o formato de um *cluster* for, por exemplo, uma elipse.

Este problema é demonstrado pela Figura 43, na qual se apresentam dois *clusters* bem separados, ilustrados por marcadores de duas formas diferentes (triângulos e quadrados).

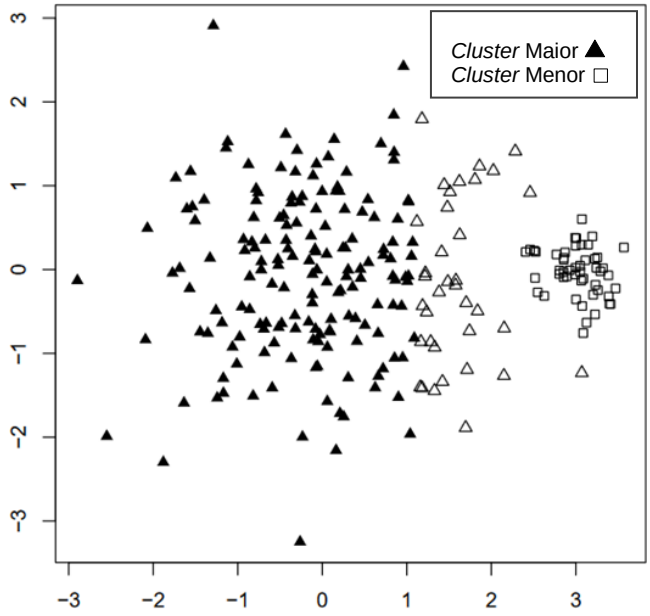

Figura 43 – Ilustração de uma grande desvantagem de *K-Means*.

Quando o algoritmo de *K-Means* é aplicado ao conjunto de dados caracterizados pelos *clusters* presentes na Figura 43, um vasto número de elementos, caracterizados por verdadeiras associações ao *cluster* maior, é classificado, erradamente, no *cluster* de tamanho menor [35].

5.2 Trabalho Futuro

De modo a responder às limitações do algoritmo *K-Means*, propõem-se, como trabalho futuro, a utilização de SOM (Self-Organizing Maps), os quais possuem uma tendência exígua de atingir um mínimo local da função objetiva, em detrimento do mínimo global, comparativamente com *K-Means* [41].

SOM trata-se de uma técnica não supervisionada de visualização de dados, que pode ser usada para visualizar conjuntos de elementos de dados caracterizados por elevadas dimensões, em

representações de dimensões inferiores (tipicamente bidimensionais). Uma das principais vantagens do SOM, em termos de visualização, assenta no facto de este mapeamento preservar as relações topológicas, intrínsecas aos dados originais.

O método de visualização do SOM é usualmente baseado em *heatmaps* (“mapas de calor"). Um *heatmap* ilustra a distribuição de uma variável ao longo do SOM, isto é:

- Se se imaginar um SOM, como uma sala lotada de pessoas;
- Se se estiver a observar essas pessoas, de cima (num género de miradouro) para baixo;
- E se cada pessoa possuir um cartão colorido, representativo da sua idade;

O resultado seria um heatmap do SOM. As pessoas de idades semelhantes, idealmente, apresentar-se-iam agregadas na mesma área. O mesmo pode ser repetido para a idade, peso, etc.

A visualização de diferentes *heatmaps* permite explorar a relação entre as variáveis de entrada [42].

Referências

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

[2] International Union Of Railways. [online]. http://www.uic.org/, acedido em: Setembro de 2015.

[3] GSMR - Info. [online]. http://www.gsmr-info.com/, acedido em: Setembro de 2015.

[4] http://www.etsi.org/images/files/ETSITechnologyLeaflets/GSMfor%20railways(GSM-R).pdf, acedido em: Setembro de 2015.

[5] REFER Telecom, “NetRail”, Vol. 3, Junho 2011.

[6] ANACOM. ANACOM - Autoridade Nacional de Comunicações. [online]. www.anacom.pt.

[7] http://www.iptelecom.pt/centro-de-imprensa/primeira-chamada-GSM-R-na-rede-ferroviaria-nacional, acedido em: Setembro de 2015.

[8] Beire, Ana; “Otimização de modelos de propagação utilizando Algoritmos Genéticos: Caso das Comunicações Móveis em Ferrovia”, ISEL, Dezembro, 2013.

[9] Correia, Tiago; “Estimação de cobertura rádio em GSM-R através de Redes Neuronais”, ISEL, Dezembro, 2014.

[10] Okumura, Y.; Ohmori, E.; Kawano, T.; Fukuda, *K*. “Field Strength and its Variability in VHF and UHF Land-Mobile Radio Service”. Review of the Electrical Communication Laboratory, Vol. 16, Nº 9-10, Outubro 1968, 16, pp. 825-73.

[11] Hata, Masaharu. “Empirical Formula for Propagation Loss in Land Mobile Radio Services”. IEEE Transactions on Vehicular Technology, Vol. VT-29, Nº 3, Agosto 1980, 29, pp. 317-25.

[12] Cota, Nuno; Serrador, António; Vieira, Pedro; Beire, Ana; Rodrigues, António; "On the Use of Okumura-Hata Propagation Model on Railway Communications," in Wireless Personal Multimedia Communications Symposium (WPMC2013), Atlantic City, New Jersey, USA, 2013.

[13] ETSI, ETS 300 553 Digital cellular telecommunications system (Phase 2); layer 1. General requirements.

[14] Metodologia para planeamento de rádio em GSM-R, ISEL. Departamento de Engenharia Eletrónica e Telecomunicações e de Computadores do Instituto Superior de Engenharia de Lisboa; Refer Telecom; Lisboa, 2009.

[15] Cota, Nuno; Serrador, António; Franco, Nuno e Neves, José, “Planeamento Rádio em GSM-R: Metodologia e Caracterização do Sinal”, URSI, Lisboa, 2009.

[16] Correia, Luís; “Sistemas de Comunicações Móveis – Modelos de Propagação”. Lisboa, Portugal: IST, 2007.

[17] Recommendation ITU-R P.526-12, "Propagation by diffraction," Janeiro 2012.

[18] J. Deygout, "Correction factor for multiple knife-edge diffraction," Antennas and Propagation, IEEE Transactions on, vol. 39, no. 8, pp. 1256-1258, Agosto 1991.

[19] http://www.teleres.com.au/Terrain, acedido a Agosto de 2016.

[20] Pahl, John; “Interference Analysis: Modelling Radio Systems for Spectrum Management”; pp. 100-156, Abril 2016.

[21] Anderson, Harry; Hicks Ted; Kirtner, Jody; “The Application of Land Use / Land Cover (*Clutter*) Data to Wireless Communication System Design”; EDX Wireless, LLC Eugene, Oregon USA, 2008.

[22] Holland, J. H. “Adaptation in Natural and Artificial Systems”, Ann Arbor, MI: University of Michigan Press, 1975.

[23] Michalewicz Z. “Genetic Algorithms + Data Structures = Evolution Programs (3ed.)”, pp. 13-105, 1996.

[24] http://lmarti.com/wp-content/uploads/2014/09/02-elements-of-eas.pdf, acedido em: Dezembro de 2015.

[25] http://www.cse.msu.edu/~jain/*Clustering*_Jain_Dubes.pdf vol. 3 e 4, acedido em: Junho de 2016.

[26] http://web.itu.edu.tr/sgunduz/courses/verimaden/paper/validity_survey.pdf, acedido em: Junho de 2016.

[27] http://www.cs.kent.edu/~jin/DM08/ClusterValidation.pdf, acedido em: Junho de 2016.

[28] http://maxwellsci.com/print/rjaset/v6-3299-3303.pdf, acedido em: Julho de 2016.

[29] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766793/, acedido em: Agosto de 2016.

[30] “Data *Clustering* Algorithms and Applications”, editado por C. Aggarwal, Charu e K. Reddy, Chandan; 2014.

[31] http://datawarehouse4u.info/ETL-process.html, acedido em: Agosto de 2016.

[32] http://www.mathworks.com/help/stats/kmeans.html, acedido em: Março de 2016.

[33] https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf, acedido em: Maio de 2016.

[34] http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf, acedido em: Maio de 2016.

[35] https://theses.lib.vt.edu/theses/available/etd-12062005-153906/unrestricted/Proposal-

Face.pdf capitulo 2, acedido em: Agosto de 2016.

[36] https://bib.dbvis.de/uploadedFiles/155.pdf, acedido em: Agosto de 2016.

[37] http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf vol. 6 e 7, , acedido em: Agosto de 2016.

[38] http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/2-mean-and-standard-deviation, acedido em: Setembro de 2016.

[39] https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php, acedido em: Setembro de 2016.

[40] https://algobeans.com/2015/11/30/k-means-*clustering*-laymans-tutorial/, acedido em: Fevereiro de 2016.

[41] pdfs.semanticscholar.org/3ffe/8f8a7b0d00297e0cd74d20b5d936349d6cbc.pdf, acedido em: Setembro de 2016.

[42] https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/, acedido em: Setembro de 2016.

[43] http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm/, acedido em: Fevereiro de 2016.

Anexos

Anexo A

Exemplo Numérico do *K-Means* [43]

A Tabela 5 apresenta a informação relativa a um exemplo numérico, que explica o princípio de funcionamento do algoritmo *K-Means*. Os 4 objetos, considerados como dados de treino, são constituídos por 2 atributos: índice do peso e pH.

Objecto	**Atributo 1** (**X**): Índice Do Peso	**Atributo 2** (**Y**): pH
Medicamento A	1	1
Medicamento B	2	1
Medicamento C	4	3
Medicamento D	5	4

Tabela 5 – Conjunto de dados de exemplo numérico.

Definindo *K* = 2, sabe-se que cada objeto pode pertencer a 2 grupos de medicamentos (*cluster* 1 e *cluster* 2), portanto, de modo a determinar a que *cluster* é que cada objeto pertence, tendo como suporte as características de cada um, utiliza-se o algoritmo *K-Means*.

Cada medicamento representa um ponto com 2 atributos (X, Y), os quais podem ser representados como coordenadas, tal como se pode verificar na Figura 44.

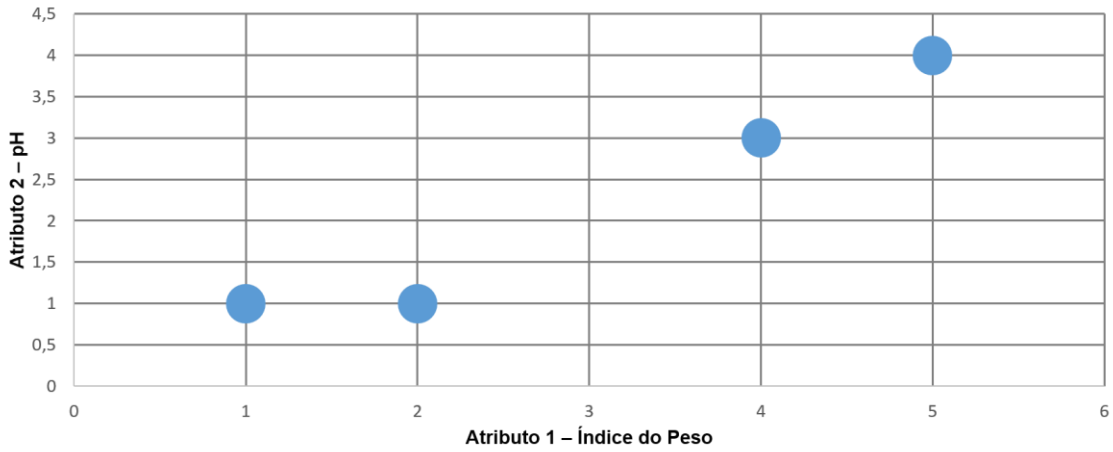

Figura 44 – Representação dos elementos de dados.

Em seguida é apresentada uma descrição de cada iteração percorrida pelo algoritmo *K-Means*.

1. Valor inicial dos *centroids*: admitindo que os medicamentos A e B foram escolhidos aleatoriamente como *centroids* iniciais, têm-se as seguintes coordenadas dos mesmos:

   $c_1 = (1,1)$ e $c_2 = (2,1)$, os quais se apresentam ilustrados, a vermelho, na Figura 45.

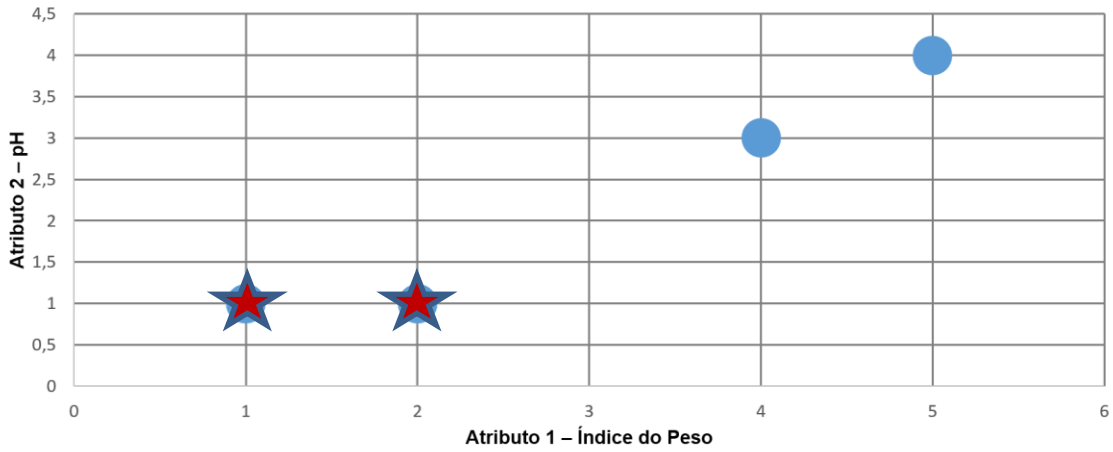

Figura 45 – Iteração 0 do algoritmo *K-Means*.

2. Distância objetos – *centroids*: de modo a calcular a distância de cada objeto aos *centroids* de cada *cluster*, utiliza-se a distância euclidiana, sendo D^0, a matriz resultante, relativa à iteração 0.

$$\text{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) \rightarrow \text{Cluster 1} \\ c_2 = (2,1) \rightarrow \text{Cluster 2} \end{matrix}$$

$$\begin{matrix} \text{A} & \text{B} & \text{C} & \text{D} \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} \text{X} \\ \text{Y} \end{matrix}$$

   Cada coluna da matriz anterior descreve as coordenadas de cada medicamento $(\text{A}, \text{B}, \text{C e D})$. As linhas 1 e 2 de D^0 correspondem à distância de cada objeto aos *centroids* 1 e 2, respectivamente. A título de exemplo, de modo a obter a distância do $\text{medicamento D} = (5,4)$ ao primeiro *centroid* $c_1 = (1,1)$, calcula-se $\sqrt{(5-1)^2 + (4-1)^2} = 5 = \text{D}^0(\text{linha } 1, \text{coluna } 4)$. O cálculo da distância deste objecto ao segundo *centroid* $c_2 = (2,1)$, é dada por $\text{D}^0(2,4) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$.

3. Agrupamento de objetos: cada objeto é atribuído a um determinado *cluster*, sendo a distância (mínima) aos *centroids* de cada um, o critério de atribuição utilizado. Como tal, e observando a matriz de distâncias D^0, o medicamento A é atribuído ao *cluster* 1 e os restantes ao *cluster* 2.
   G^0 é a matriz de grupos resultante da atribuição descrita no parágrafo anterior.

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \rightarrow \text{Cluster 1} \\ \rightarrow \text{Cluster 2} \end{matrix}$$
$$\begin{matrix} A & B & C & D \end{matrix}$$

Um elemento de G^0 é "1" se, e só se, este tiver sido atribuído a esse *cluster*.

4. Iteração 1, atualização da localização dos *centroids*: sabendo a constituição de cada *cluster*, o próximo passo é a atualização da localização dos *centroids* de cada um, tendo em conta o agrupamento realizado anteriormente. O *cluster* 1 é constituído apenas por um objeto, mantendo-se por isso inalterada a localização do respetivo *centroid* $c_1 = (1,1) = \text{coordenadas do medicamento A}$.

   O novo posicionamento do *centroid* do *cluster* 2 é obtido através do cálculo da média das coordenadas dos seus 3 objetos constituintes $(B, C \text{ e } D)$, sendo $c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right) \approx (3.67, 2.67)$. A Figura 46 apresenta as localizações atualizadas dos *centroids*.

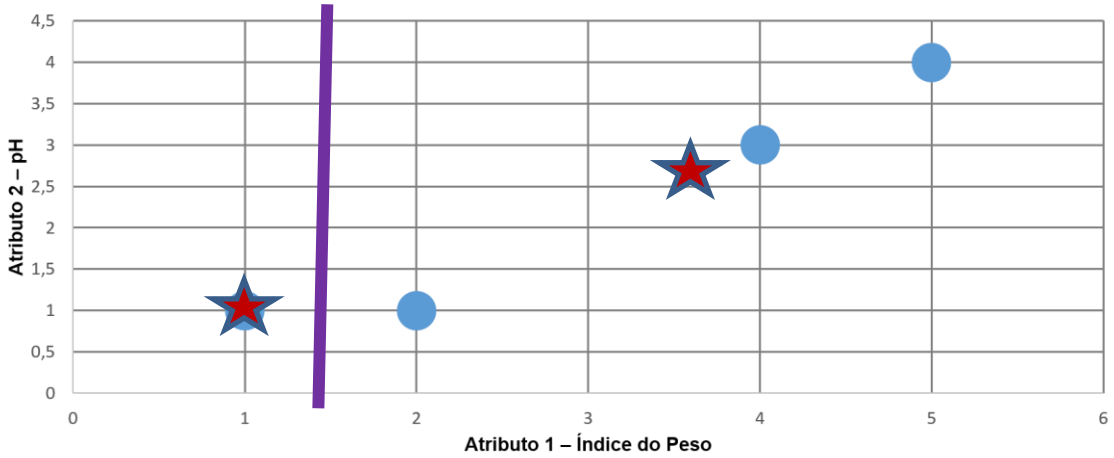

Figura 46 – Iteração 1 do algoritmo *K-Means*.

5. Iteração 1, distância objetos - *centroids*: tal como foi realizado no passo 2, a presente etapa foca-se no cálculo das distâncias de cada objeto às novas posições dos *centroids* de cada *cluster*. Sendo a matriz de distâncias D^1, o resultado desses cálculos.

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} c_1 = (1,1) \rightarrow \text{Cluster 1} \\ c_2 = (3.67, 2.67) \rightarrow \text{Cluster 2} \end{matrix}$$
$$\begin{matrix} A & B & C & D \end{matrix}$$
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Note-se que a primeira linha da matriz D^1 não sofreu alterações, devido ao facto da localização do *centroid* c_1 ter-se mantido inalterada.

6. Iteração 1, agrupamento de objetos: sendo esta etapa idêntica ao passo 3, a atribuição dos objetos é realizada com base na distância mínima de cada um, aos respetivos *centroids*.

Após a análise da nova matriz de distâncias D^1, o medicamento B é atribuído ao *cluster* 1, enquanto os restantes mantêm-se inalterados. Resultando, assim, uma nova matriz de grupos G^1.

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \rightarrow \text{Cluster 1} \\ \rightarrow \text{Cluster 2} \end{matrix}$$
$$\text{A} \quad \text{B} \quad \text{C} \quad \text{D}$$

7. Iteração 2, atualização da localização dos *centroids*: repetindo o passo 4 e tendo em conta o agrupamento realizado no passo anterior, as novas coordenadas de ambos os *centroids* são dadas por: $c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5, 1)$ e $c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5, 3.5)$. A Figura 47 ilustra o posicionamento de ambos os *centroids*.

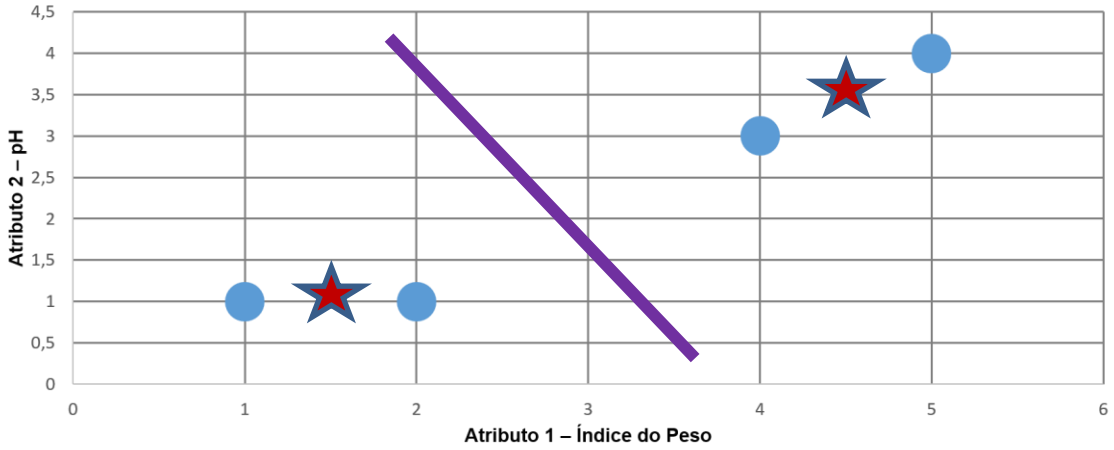

Figura 47 – Iteração 2 do algoritmo *K-Means*.

8. Iteração 2, distância objetos – *centroids*: repetindo, mais uma vez, o passo 2, obtém-se a nova matriz de distâncias D^2.

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{matrix} c_1 = (1.5, 1) \rightarrow \text{Cluster 1} \\ c_2 = (4.5, 3.5) \rightarrow \text{Cluster 2} \end{matrix}$$
$$\begin{matrix} \text{A} & \text{B} & \text{C} & \text{D} \end{matrix}$$
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} \text{X} \\ \text{Y} \end{matrix}$$

9. Iteração 2, agrupamento de objetos: repetindo, mais uma vez, o passo 3, a atribuição de objetos é realizada com base na distância mínima, sendo G^2 a matriz de grupos resultante.

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \rightarrow \text{Cluster 1} \\ \rightarrow \text{Cluster 2} \end{matrix}$$
$$\text{A} \quad \text{B} \quad \text{C} \quad \text{D}$$

Tendo sido atingida a igualdade $G^2 = G^1$, conclui-se que os objetos de ambos os *clusters* se mantiveram imóveis, o que significa que o algoritmo convergiu, isto é, alcançou a estabilidade, não sendo por isso necessária a continuação do algoritmo. A Tabela 6 apresenta o resultado da aplicação do *K-Means*, no presente exemplo numérico.

Objecto	**Atributo 1** (**X**): Índice Do Peso	**Atributo 2** (**Y**): pH	***Cluster*s Resultantes**
Medicamento A	1	1	1
Medicamento B	2	1	1
Medicamento C	4	3	2
Medicamento D	5	4	2

Tabela 6 – Resultado da aplicação do *K-Means*.

O algoritmo termina com $c_1 = (1.5, 1)$ e $c_2 = (4.5, 3.5)$, sendo que os medicamentos A e B pertencem ao primeiro *cluster* e C e D ao segundo.