

Mestrado em Sistemas de Informação Geográfica e Ordenamento do Território

Estudo e análise preditiva sobre as fugas na rede de distribuição de água de Vila Nova de Gaia

André de Oliveira

M

2020



André de Oliveira

**Estudo e análise preditiva sobre as fugas na rede de distribuição
de água de Vila Nova de Gaia**

Dissertação realizada no âmbito do Mestrado em Sistemas de Informação Geográfica e
Ordenamento do Território, orientada pelo
Professor Doutor Ricardo Baptista
e coorientada Professor Doutor Alberto Gomes

Faculdade de Letras da Universidade do Porto
outubro de 2020

Estudo e análise preditiva sobre as fugas na rede de distribuição de água de Vila Nova de Gaia

André de Oliveira

Dissertação realizada no âmbito do Mestrado em Sistemas de Informação Geográfica e
Ordenamento do Território, orientada pelo Professor Doutor Ricardo Baptista e coorientada pelo
Professor Doutor Alberto Gomes

Membros do Júri

Professor Doutor
Faculdade - Universidade ...

Professor Doutor
Faculdade - Universidade ...

Professor Doutor
Faculdade - Universidade ...

Classificação obtida: valores

Para todos os que caminharam comigo

AGRADECIMENTOS

O meu percurso académico nos últimos dois anos e o desenvolvimento desta dissertação não teria sido possível sem um conjunto de pessoas, que de uma forma ou de outra, deram o seu contributo. Deixo aqui o meu obrigado a essas pessoas.

Em primeiro lugar agradecer à empresa Águas de Gaia, que me proporcionou todas as condições para o desenvolvimento desta dissertação. Um especial agradecimento ao Dr. José Claudio que me incentivou e entusiasmou para o desenvolvimento desta temática.

Aos meus orientadores, Professor Ricardo Baptista e Professor Alberto Gomes, que ao longo destes meses acompanharam e contribuíram para a conclusão desta meta. Uma palavra ainda de apreço ao Professor Alberto Gomes, não só pelo acompanhamento nos últimos meses com a dissertação, mas também ao longo dos últimos 2 anos, tanto a nível académico como a nível extracurricular, sempre com boa disposição.

Aos meus colegas, Jorge Jesus e João Guerreiro, pelo companheirismo, entusiasmo e troca de ideias. Sem vocês, os resultados que aqui apresento certamente não seria igual.

Ao meu irmão, que desde que me lembro sempre esteve disponível para me ajudar, qualquer que fosse o problema que tinha em mãos.

Aos meus pais, por todo o contributo que tiveram ao longo da minha vida, e sempre me incentivaram, e proporcionaram todas as condições para seguir os meus interesses.

Por fim, e mais importante, à Rita que sempre me apoiou nos momentos mais difíceis e tornou qualquer bom momento, ainda melhor.

Muito obrigado a todos. Espero um dia conseguir retribuir tudo o que me deram.

RESUMO

Vila Nova de Gaia, segundo os censos de 2011, é o 3º município com mais habitantes em Portugal, no entanto, é apenas o 11º com maior densidade populacional. Com uma área de 168,5 km², acaba por se traduzir numa rede de distribuição de água extensa. Com o envelhecimento da rede, é natural que surjam fugas ao longo do seu percurso. As fugas de água implicam impactos negativos a nível económico e ambiental.

A distribuição de água no município é assegurada pela empresa Águas de Gaia, EM SA, que já possui modelos de deteção de fugas. No entanto, estes modelos incidem em fugas que já ocorreram. É essencial o desenvolvimento de modelos capazes de identificar áreas vulneráveis na rede e prever possíveis fugas. Existindo este tipo de informação, as entidades com poder de decisão e responsáveis pela manutenção, têm em seu poder dados que podem auxiliar na monitorização e dando poder de decisão relativamente aos investimentos a fazer relativamente à manutenção das infraestruturas.

Este estudo, propõe-se a avaliar quais os fatores que possam ter influência sobre fugas na rede de distribuição de água, e com base neles, desenvolver um modelo preditivo. O estudo terá como base as fugas ocorridas nos anos 2017, 2018 e 2019. A metodologia proposta baseia-se na identificação de possíveis fatores que possam afetar de forma negativa a rede de distribuição e fazer a avaliação destes sobre as fugas ocorridas. A partir da informação aí extraída, serão desenvolvidos modelos de *Machine Learning*, seguindo os métodos de classificação e regressão *Decision Trees* e *Random Forest*. Os dados recolhidos, bem como a informação e os modelos gerados foram estruturados e integrados numa base de dados geoespacial, tendo como base tecnologia SIG (Sistemas de Informação Geográfica), possibilitando a representação espacial das fugas e das condicionantes e o desenvolvimento de cartografia de vulnerabilidade.

Os resultados obtidos neste estudo atestam a importância do SIG e da metodologia no contributo à tomada de decisão. Permite perceber quais os fatores de predisposição do território que possam desencadear fugas numa rede de distribuição de água.

PALAVRAS CHAVE: Sistemas de Informação Geográfica; Rede de Distribuição de Água; Fugas; Modelos Preditivos; *Machine Learning*.

ABSTRACT

Vila Nova de Gaia, according to 2011 census, is the 3rd municipality with most population in Portugal, however, its only in the 11st regarding population density. With an area of 168,5 km², this translates into an extensive water network distribution. With an aging network, it's only natural to occur leaks along the way. Water leaks translates into negative economic and environmental impact.

Water distribution on the municipality is assured by the company Águas de Gaia, EM SA, which already possesses several leak detections models. However, these models focus on present leaks. It is crucial to develop tools to identify vulnerable areas in the network and predict possible leaks. With this information, decision-making entities and those responsible for the network maintenance, have in their hands data that helps monitoring the network but also power to their decisions when investing in infrastructural maintenance.

This study proposes to evaluate which factors have influence in water leakages on the distribution network, and based on them, develop a predictive model. The study will have under its foundation the leaks that occurred in the years 2017, 2018 and 2019. The proposed methodology is based on the recognition of the factors that might negatively affect the distribution network and evaluate this information with the occurred leaks. With the extracted information it will be developed Machine Learning models, following classification and regression methods with Decision Trees and Random Forests.

With the collected data and the generated information and models were integrated in a spatial geodatabase, based on GIS (Geographic Information Systems) technology, allowing to represent spatially the leaks, the conditioning factors and vulnerability cartography.

The obtained results in this study attest the importance of GIS and the used methodology on the contribution to decision-making. It allows to understand the predisposition factors of the land that might initiate a leak on a water distribution network.

Keywords: Geographic Information Systems; Water Network Distribution; Leaks; Predictive Models; Machine Learning.

ÍNDICE

AGRADECIMENTOS	vi
RESUMO	vii
ABSTRACT	viii
ÍNDICE	ix
LISTA DE FIGURAS	xi
CAPÍTULO 1	1
INTRODUÇÃO	1
1.1 Enquadramento e Problemática da investigação	1
1.2 Objetivos	3
1.3 Contribuição Científica Enquadramento e Problemática da investigação	4
1.4 Estrutura da dissertação	5
CAPÍTULO 2	6
ESTADO DO CONHECIMENTO ATUAL	6
2.1 Vulnerabilidades nas Redes de Distribuição de água	6
2.2 Modelos e técnicas de <i>Machine Learning</i>	12
2.2.1 Decision Trees	18
2.2.2 Random Forest	20
CAPÍTULO 3	24
MATERIAIS UTILIZADOS	24
3.1 Área de estudo	25
3.2 Tratamento de dados	27
3.3 Desenvolvimento dos modelos preditivos	33
CAPÍTULO 4	35
PARÂMETROS DE VULNERABILIDADE GEOGRÁFICA	35
4.1 Dados Contínuos	35
4.1.1 Elevação (altitude)	35
4.1.2 Declive	38
4.1.3 pH	41

4.1.4 Densidade populacional	44
4.1.5 Consumo de água	46
4.1.6 Ocupação do solo	49
4.1.7 Geologia	52
4.2 Distância a objetos	55
4.2.1 Rede hidrográfica	55
4.2.2 Rede viária	58
4.2.3 Rede ferroviária	61
4.3 Matriz de <i>Pearson</i>	63
CAPÍTULO 5	65
RESULTADOS	65
5.1 Avarias observadas – Densidade Kernel	65
5.2 Avarias observadas – Análise <i>Hotspot</i>	67
5.3 Previsão de suscetibilidade – <i>Random Forest Regressor</i>	68
5.4 Previsão de fugas – <i>Decision Tree Classifier</i>	75
CAPÍTULO 6	78
CONSIDERAÇÕES FINAIS	78
REFERÊNCIAS BIBLIOGRÁFICAS	80
ANEXOS	84

LISTA DE FIGURAS

Figura 1- Diferentes tipos de Tubagens e Acessórios numa Rede de Distribuição de água. Fonte: AGEM.....	7
Figura 2- Erosão Hídrica provocada no interior da tubagem. Adaptado: TLV.....	7
Figura 3- Corrosão no Material de distribuição de água. Fonte: Allied Corrosion Industries Inc.....	7
Figura 4- Ciclo de Vida de uma tubagem. Adaptado: Rogers & Grigg, 2009.	10
Figura 5- Modelação de fugas num período de 24 horas. Adaptado de: R Liemberger & Farley.....	11
Figura 6- Esquema metodológico para a aplicação de modelos ML. Adaptado de: T.M. Mitchell, 1997.	12
Figura 7 - Produção Científica com os termos "Geographic Information Systems" e "Machine Learning". Fonte: Scopus.....	14
Figura 8 - Nuvem de palavras a partir dos títulos de documentos científicos com os termos “Geographic Information Systems” e “Machine Learning”. Fonte: Scopus.....	15
Figura 9 - Diferentes Domínios e Técnicas de Machine Learning. Adaptado de: Arroyo, 2017.	16
Figura 10 - Dois possíveis resultados de clusterização a partir aprendizagem não-supervisionada. Fonte: Dey, 2016.	17
Figura 11 - Interação num modelo de aprendizagem reforçada. Adaptado: Busoniu, Babuska, Schutter, & Ernst, 2010.	18
Figura 12 - Árvore de Decisão para a classificação de animais.	19
Figura 13 - Fluxo de métodos ensemble. Adaptado: Géron, 2017.	21
Figura 14 - Esquema Preditivo de um modelo Random Forest. Adaptado: Bakshi, 2020.....	23
Figura 15 - Enquadramento Geográfico de Vila Nova de Gaia. Fonte: AGEM e CAOP.....	26
Figura 16 - Processo de Georreferenciação de Moradas.....	27
Figura 17 - Tipos de avaria durante o período de estudo. Fonte: AGEM.	28
Figura 18 - Ilustração de uma RDA e as suas infraestruturas. Fonte: CM Sesimbra (2017).	28
Figura 19 - Amostra da Rede de Distribuição de Água. Fonte: AGEM	29
Figura 20 - Isolamento de roturas. Fonte: AGEM.....	29
Figura 21 - Número de avarias por tubagem calculado a partir da proximidade das ocorrências. Fonte: AGEM.....	30
Figura 22 - Somatório de avarias associadas à RDA.....	31
Figura 23 - Conversão de tubagens em formato de linha para ponto central.	32
Figura 24 - Conversão de tubagens em formato de linha para <i>end-points</i>	32

Figura 25 - Associação dos valores das variáveis independentes à variável dependente. Adaptado de: ESRI, Inc.....	32
Figura 26 - Diferentes plataformas e bibliotecas utilizados para o desenvolvimento dos modelos preditivos.	33
Figura 27 - Diferentes parâmetros de vulnerabilidade selecionados e processos para o desenvolvimento dos modelos preditivos.	34
Figura 28 - a) Triangulação de dados topográficos. b) Diferença entre um MDS e um MDT. Fonte: ESRI.	35
Figura 29 - Relação entre avarias observadas e variável elevação.	36
Figura 30 - Carta Hipsométrica de Vila Nova de Gaia.....	37
Figura 31 - Dados de entrada para o cálculo do declive. Fonte: ESRI.	38
Figura 32 - Exemplo de declive plano na faixa litoral, combinação da carta de declives com ortofotomapa.	39
Figura 33 - Exemplos de declives elevado ao longo do rio Douro, nas freguesias do Olival e Crestuma numa combinação da carta de declives com ortofotomapa.....	39
Figura 34 - Relação entre avarias observadas e variável declive.	39
Figura 35 - Carta de Declives de Vila Nova de Gaia.....	40
Figura 36 - Aplicação Survey123 (a) e recolha de dados do solo com um medidor pH (b).	41
Figura 37 - Pixeis sem informação de pH nos dados provenientes da ESDAC.	42
Figura 38 - Relação entre avarias observadas e o pH do solo.	42
Figura 39 - Variação do pH do solo em Vila Nova de Gaia.....	43
Figura 40 - Relação entre avarias observadas e densidade populacional.....	44
Figura 41 - Densidade Populacional, 2011, em Vila Nova de Gaia.	45
Figura 42 - Esquema metodológico para a espacialização do consumo de água.	47
Figura 43 - Relação entre as avarias observadas e volume de água consumido.....	47
Figura 44 - Representação do volume de água consumido em 2019 em Vila Nova de Gaia.....	48
Figura 45 - Percentagem de avarias observadas por classe de ocupação do solo.....	50
Figura 46 - Carta de Ocupação do Solo de Vila Nova de Gaia.	51
Figura 47 - Regras topológicas utilizadas para a vectorização das cartas geológicas.	52
Figura 48 . Percentagem de avarias por unidade geológica.	53
Figura 49 - Carta Geológica de Vila Nova de Gaia.....	54
Figura 50 - Cálculo da distância euclidiana. Fonte: ESRI.	55
Figura 51 - Relação entre fugas observadas e distância à rede hidrográfica.	56
Figura 52 - Carta Hidrológica de Vila Nova de Gaia.	57

Figura 53 - Relação entre fugas observadas e distância à rede viária.	59
Figura 54 - Rede Viária em Vila Nova de Gaia.	60
Figura 55 - Relação entre as fugas observadas e distância à rede ferroviária.	61
Figura 56 - Rede Ferroviária em Vila Nova de Gaia.	62
Figura 57 - Avarias detetadas no período 2017-2019 e respetiva densidade Kernel.	66
Figura 58 - Análise de Clusters de ocorrência de fugas em 2017 e 2019.	68
Figura 59 - Resultados produzidos pelos diferentes modelos RFR.	72
Figura 60 - Contabilização das infraestruturas com uma suscetibilidade prevista superior a 0.7.	74
Figura 61 - <i>Decision Tree</i> resultante do modelo F.	76
Figura 62 - Previsão de fugas a partir do modelo <i>Decision Tree</i>	77
 Tabela 1 - Balanço Hídrico Standard. Fonte: ERSAR.....	 3
Tabela 2 - Parâmetros utilizados em modelos preditivos para fugas de água. Adaptado de: Jenkins, Gokhale, & McDonald, 2014.....	9
Tabela 3 – Taxa de ocupação do solo por freguesia. Fonte: DGT.....	50
Tabela 4 - Matriz de Pearson para os parâmetros de vulnerabilidade,	64
Tabela 5 - Especificidade sobre os principais modelos RFR desenvolvidos.	69
Tabela 6 -Performance inicial e final dos modelos RFR.	70

CAPÍTULO 1

INTRODUÇÃO

1.1 Enquadramento e Problemática da investigação

"Globalmente, perdem-se, por ano, 179.722.877 metros cúbicos de água, o que corresponde a um prejuízo de 85,5 milhões de euros, em 258 municípios", (DECO, 2020)

O relatório divulgado pela DECO em março de 2020, revela que as perdas de água nos municípios portugueses são muito significativas o que se traduz em impactos negativos económicos, mas também ambientais. A entidade refere ainda que os principais fatores para as perdas de água, são provocadas nos momentos de armazenamento, transporte e distribuição de água. Segundo os dados da Entidade Reguladora dos Serviços de Água e Resíduos (ERSAR), o consumo médio diário de água, por pessoa, é de 192 litros (dados para 2018). Para colocar estes valores em perspetiva, a água perdida anualmente, permitiria servir sensivelmente 2,5 milhões de pessoas, o que corresponde, segundo os últimos censos, a 1/4 da população portuguesa.

A água é um dos bens mais essenciais à vida humana e restantes seres vivos no planeta, no entanto, universalmente é um dos bens mais mal geridos. Esta má-gestão acaba por ser um dos principais fatores para a escassez de água, a nível global, e que em Portugal acaba por acentuar-se em períodos de seca que são cada vez mais frequentes. As fugas de água representam mais impactos negativos, para além dos mencionados, na medida em que as fugas são prejudiciais para as estruturas da rede de distribuição, o que acaba por envolver elevados custos desde o processo da sua identificação até à resolução da mesma. Este, implica a realização de trabalhos no solo e, consequentemente, exerce impactos negativos para o meio ambiente. O mundo está a enfrentar uma crise de água e que se prevê que se intensifique nos próximos anos (D. J. P. E. Rogers, 2014)

Os municípios portugueses, e concretamente o município de Vila Nova de Gaia, área de estudo na presente dissertação, não são exceção à crise que se enfrenta no âmbito da água. É essencial a manutenção da rede de distribuição de água dos municípios, mas também o desenvolvimento de meios e métodos que permitam reduzir as perdas de água. Atualmente, os métodos mais eficazes para a identificação de fugas de água envolvem métodos e equipamentos acústicos, que são mais eficazes em redes de distribuição compostas por tubos de metal. No entanto, nem todas as redes são compostas por estes materiais e a identificação de fugas por este método é moroso. É essencial o desenvolvimento de técnicas que permitam a identificação rápida

e eficiente de fugas que estão a ocorrer. É igualmente importante conhecer a rede de modo a que seja possível o desenvolvimento de modelos que permitam perceber quais as tubagens mais suscetíveis a fugas de águas, para que as entidades de tomada de decisão possam ter informação prévia e melhorar a qualidade de decisão relativamente aos locais onde deverá ser aplicado maior investimento na manutenção da rede (Burn et al., 1999; Martini et al., 2018).

Todos os anos, as entidades encarregues da distribuição de água, necessitam de realizar o balanço hídrico. O balanço hídrico não segue um formato universal, havendo até diferentes formatos e definições dentro do próprio país. O ERSAR adotou, no entanto, o formato mais generalizado (Roland Liemberger & Farley, 2004), conforme segue na Figura 1. A Associação Portuguesa de Distribuição e Drenagem de Águas (APDA), refere que seguindo estas definições, é possível obter então uma estimativa relativa às perdas de água. Para este processo é essencial saber o volume total de entrada no sistema, numa zona bem definida. Através de contadores, é realizado o cálculo de água consumida, seja ela faturada ou não. É também essencial ter valores estimados para a água não contabilizada. A água utilizada no combate a incêndios é um exemplo de água não faturada e estimada. O seguinte passo, pressupõe a estimativa de perdas aparentes. Estas podem estar associadas a erros no fornecimento de leituras por parte dos clientes. Com os dados obtidos até esta fase, é possível calcular as perdas de água através da diferença entre a água de entrada no sistema e consumo autorizado. Numa fase final, são adotados os métodos disponíveis mais indicados para o cálculo das diferentes categorias nas perdas reais. Este estudo pode ser feito através da análise dos caudais noturnos, dados de medição zonada, duração das roturas, entre outros. O presente estudo irá debruçar-se sobre as fugas na rede de distribuição, principalmente sobre as condutas de distribuição e nos ramais de ligação.

Água entrada no sistema [m³/ano]	Consumo autorizado [m³/ano]	Consumo autorizado faturado [m³/ano]	Consumo faturado medido [m³/ano]	Água faturada [m³/ano]
			Consumo faturado não medido [m³/ano]	
		Consumo autorizado não faturado [m³/ano]	Consumo não faturado medido [m³/ano]	Água não faturada (perdas comerciais) [m³/ano]
			Consumo não faturado não medido [m³/ano]	
	Perdas de água [m³/ano]	Perdas aparentes [m³/ano]	Uso não autorizado [m³/ano]	
			Perdas de Água por erros de medição [m³/ano]	
		Perdas reais [m³/ano]	Perdas reais nas condutas de água bruta e no tratamento (quando aplicável) [m³/ano]	
			Fugas nas condutas de adução e/ou distribuição [m³/ano]	
			Fugas e extravasamentos nos reservatórios de adução e/ou distribuição [m³/ano]	
			Fugas nos ramais de ligação (a montante do ponto de medição) [m³/ano]	

Tabela 1 - Balanço Hídrico Standard. Fonte: ERSAR

1.2 Objetivos

O desenvolvimento deste estudo tem como grande objetivo obter resposta para as seguintes questões:

- 1) Quais as infraestruturas na rede de distribuição de água de VNG que potencialmente podem ter fugas não identificadas?
- 2) Com base na predisposição geográfica, i.e., considerando um conjunto de variáveis representativas, quais são os locais mais suscetíveis à ocorrência de fugas de água no município?

Com este objetivo, serão desenvolvidas várias análises tendo por base o histórico de fugas ocorridas nos anos 2017, 2018 e 2019 e analisar os padrões geo-espaciais das mesmas. Para isso, após realizar uma análise da bibliografia sobre o assunto, foram aferidas as principais variáveis identificadas como desencadeantes de fugas de água. A partir da conjugação desta informação, é proposto o desenvolvimento de um modelo preditivo para fugas de água, utilizando algoritmos de *Machine Learning* e a representação cartográfica dos resultados. Os resultados obtidos a partir deste modelo serão comparados, para efeitos de validação, com o modelo estatístico *Hotspot Analysis*. De forma a encontrar resposta para as questões de partida, foram delineados os seguintes objetivos:

- 1) Realizar uma análise estatística e cartográfica entre as fugas de água ocorridas e diferentes variáveis disponíveis e selecionadas;
- 2) Desenvolver de modelos preditivos dinâmicos e de suscetibilidade a fugas de água na distribuição de água de VNG;
- 3) Criar uma base cartográfica temática com os resultados obtidos na modelação efetuada.

A análise estatística e cartográfica realizou-se em duas fases: i) análise de padrões de distribuição de fugas ocorridas, ii) cruzamento de dados entre as variáveis e o inventário de fugas ocorridas no período temporal em análise. Esta fase final implicou o desenvolvimento de regressões lineares para interpretar o comportamento da rede face aos diferentes parâmetros inerentes a cada uma das variáveis.

A produção do modelo preditivo de fugas baseia-se na aplicação do modelo *Decision Trees* (Árvore de Decisões), que consiste na observação de um evento (neste caso, as fugas de água na tubagem), numa zona de treino (*training areas*) e na análise estatística de cada variável associada a esse evento. A partir desta observação, é desenvolvido uma previsão para a totalidade da rede do município.

A partir do modelo desenvolvido, foi atribuído a cada tubo na rede de distribuição um valor entre 0 e 1 (em que 1 apresenta uma suscetibilidade a fugas extremamente elevado) e a partir desses valores foi concebida cartografia que permite identificar as áreas do município onde existe uma suscetibilidade a fugas de água muito elevada.

O objetivo final será aumentar o leque de ferramentas que as entidades decisoras têm aos seu dispor, neste caso, a empresa Águas de Gaia, EM.

1.3 Contribuição Científica Enquadramento e Problemática da investigação

O desenvolvimento desta dissertação é fundamentado pela crença que os métodos desenvolvidos e resultados obtidos acrescentem valor às ferramentas já existentes para gestão de uma rede de distribuição de água. As entidades não deverão apenas utilizar métodos já existentes, devendo procurar inovar e desenvolver novas instrumentos de apoio à tomada de decisão, especialmente na gestão de um recurso crucial e escasso como é a água. Atualmente, os SIG apresentam-se como uma plataforma de dados tabulares e geo-espaciais, capaz de dar apoio a equipas de terreno e de elaboração de estatísticas sobre a rede. No entanto, o seu potencial de análise espacial, em concreto sobre redes de distribuição de água, ainda se encontra por explorar. Já existe um grande leque de produção científica utilizando modelos preditivos em áreas como a previsão de

movimentos de vertente, incêndios, entre outros (referências). Contudo, em matéria de gestão de redes de água, particularmente na previsão de fugas, a produção científica ainda é relativamente escassa, quando comparando com outras áreas do conhecimento científico. O presente estudo propõe-se selecionar os algoritmos de machine learning que melhor se adequam à problemática e aplicá-los a uma área de estudo – Vila Nova de Gaia.

Este estudo pode elucidar outros sobre quais são as variáveis principais a ter para obter resultados positivos e interpretar o comportamento das mesmas com o evento (avaria) em análise. Atualmente, os municípios portugueses são obrigados a desenvolver cartografia de risco para perigos como incêndios, inundações, entre outros (referências). Com o aumento da procura e demanda de água, e o aumento de custos associados à sua utilização, as medidas de contenção à sua extração e a racionalização na sua utilização poderão começar a surgir no futuro próximo. À semelhança da cartografia anteriormente mencionada, poderá fazer sentido os municípios/entidades gestoras das redes de abastecimento desenvolverem cartografia de suscetibilidade de fugas para reduzir o volume de água perdido anualmente. Assim, os métodos utilizados neste estudo poderão ser um ponto de partida para essa cartografia.

1.4 Estrutura da dissertação

A presente dissertação encontra-se dividida em cinco capítulos. O primeiro capítulo dedica-se à caracterização da problemática de estudo e a uma abordagem aos métodos que se pretende aplicar para atingir os objetivos desejados. O capítulo seguinte refere-se a toda a revisão literária do assunto investigado e encontra-se dividido em duas componentes. Numa primeira, é feita uma descrição da problemática, como ocorre o desgaste ao longo do tempo na infraestrutura de uma RDA e quais os parâmetros que contribuem para a sua vulnerabilidade. É uma fase capital, na medida em que é obtido o conhecimento de quais as variáveis a utilizar para a concretização dos objetivos. A segunda é referente à revisão dos modelos de ML, quais os diferentes domínios e subdomínios. Esta é também uma fase crucial, uma vez que são identificadas as características dos principais algoritmos e quais os que melhor se adequam à problemática. A partir deste ponto existe informação suficiente para fazer uma escolha ponderada sobre quais os algoritmos a utilizar e quais os cuidados a ter.

Os capítulos seguintes são referentes à parte também chamada de “prática”. O terceiro capítulo aborda os métodos e técnicas utilizados. São descritos os procedimentos de tratamento dos dados e de que forma os parâmetros de vulnerabilidade se relacionam com o território e com o tema. O capítulo seguinte dedica-se ao desenvolvimento do modelo e à apresentação de resultados. Por fim, no capítulo final são referidas as contrariedades encontradas e são elaboradas as considerações finais e apresentadas sugestões para futuros estudos.

CAPÍTULO 2

ESTADO DO CONHECIMENTO ATUAL

2.1 Vulnerabilidades nas Redes de Distribuição de água

A problemática das perdas de água em sistemas de distribuição é uma temática complexa onde entram os mais diversos ramos da ciência. Existem vários estudos de engenharia hidráulica que abordam o problema com diferentes metodologias (Moser et al., 2015), trabalhando com características individuais de cada tubagem, tendo acesso a uma variedade de dados como a velocidade e pressão, gerados a partir de contadores. Existem também metodologias que abordam a problemática com uma visão geoespacial como no campo da Detecção Remota (Agapiou et al., 2014), onde são aplicados algoritmos para a identificação de água no subsolo. Neste estudo, é proposto o desenvolvimento de um modelo utilizando ferramentas de Machine Learning com uma forte componente geoespacial, ou seja, utilizando diversas variáveis que descrevem fisicamente o território e que possam ter influência sobre a rede de distribuição.

Seguindo esta linha, este capítulo encontra-se dividido em diferentes partes. A primeira parte, é referente à revisão bibliográfica, nomeadamente sobre os parâmetros identificados como influências negativas na conservação da RDA. A segunda parte aborda os conceitos teóricos sobre as técnicas e métodos utilizados para o desenvolvimento deste estudo. Este inclui ainda a seleção de parâmetros e será feita uma abordagem sobre cada um, tentando perceber de que forma a sua atuação influencia a degradação da rede de distribuição de água. Ao longo da revisão bibliográfica, serão identificados os principais autores da temática de estudo e as variáveis utilizadas, bem como o raciocínio subjacente à sua utilização.

A OMS, Organização Mundial de Saúde, (Farley et al., 2001) define as fugas de água como um componente de perdas de água numa rede, e compreende as perdas físicas de tubos, juntas e acessórios mas também o transbordo nos reservatórios. Estas perdas de água podem ser severas e podem não ser detetadas durante meses, até mesmo anos. As maiores perdas são aquelas associadas a uma rutura numa tubagem principal, enquanto que as perdas mais pequenas podem estar associadas nas ligações de tubagem/ramais/acessórios que vão “pingando” pequenas quantidades de água. O volume total de água perdida vai depender na pressão existente no local da fuga, mas também do tempo que demora a ser detetada.

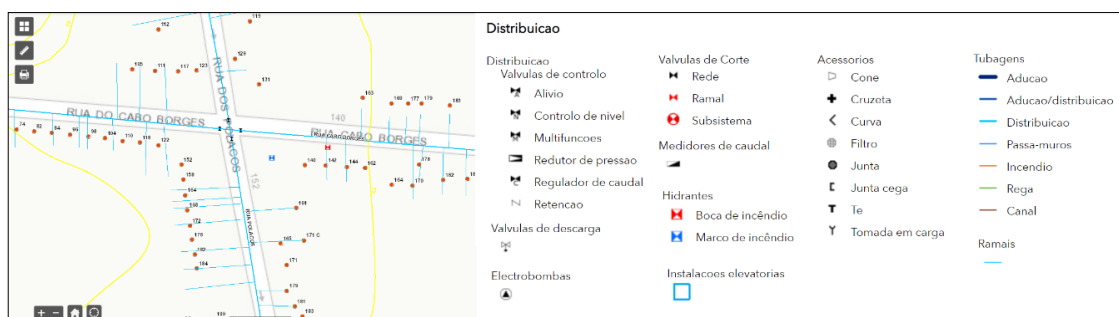


Figura 1- Diferentes tipos de Tubagens e Acessórios numa Rede de Distribuição de água. Fonte: AGEM.

A OMS refere ainda que existem dois mecanismos que são geralmente responsáveis por fugas que se intensificam ao longo do tempo. Estes dois mecanismos estão associados ao desgaste provocado pela circulação da água (a erosão hídrica), bem como a corrosão do material da tubagem. Um sistema com vários pontos de ligação, válvulas, contadores e hidrantes, apresenta várias oportunidades para a ocorrência de ruturas. Estão identificados vários condicionantes, para a existência de fugas (referência), desde logo as características da rede como a idade da tubagem e a qualidade do material utilizado.

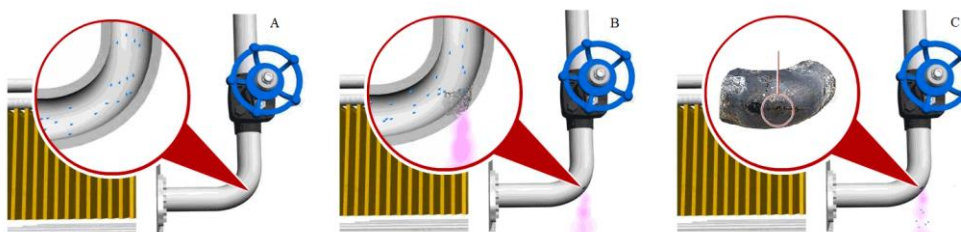


Figura 2- Erosão Hídrica provocada no interior da tubagem. Adaptado: TLV.



Figura 3- Corrosão no Material de distribuição de água. Fonte: Allied Corrosion Industries Inc.

Em 2014, os autores Jenkins, Gokhale e McDonald efetuam um levantamento sobre os parâmetros de vulnerabilidade mais comuns e utilizados pelos principais autores da temática de investigação de modelos preditivos para fugas numa RDA. Os autores demonstram o potencial no desenvolvimento de modelos preditivos a partir do histórico de fugas ocorridas numa área de estudo quando combinado com a análise espacial. O grande objetivo no desenvolvimento neste tipo de modelos é identificar as tubagens com elevado risco de falha. Estes modelos vão ao encontro daquilo que este estudo pretende alcançar: identificar as tubagens na rede de distribuição de água mais suscetíveis e priorizar a sua substituição/manutenção. Quando se combina este tipo informação obtida a partir dos modelos com a opinião de especialistas, as entidades gestores têm em sua posse ferramentas que melhoram a tomada de decisão no momento de reparação, substituição e reabilitação de equipamento na rede. É ainda referido na literatura (referências) que são poucas as entidades que investem no desenvolvimento de modelos preditivos, devido à complexidade do seu desenvolvimento e o tipo de dados necessários. É necessário, à partida, dados que, por vezes, são impossíveis de obter para toda a rede, ou até mesmo para uma área restrita, como o tipo de material e data de instalação (Berardi et al., 2008). No entanto, ultimamente, os modelos tornaram-se mais sofisticados, e aliando-se à análise espacial, não estão dependentes da existência de bases dados sobre o material que compõem a rede. Para o desenvolvimento de modelos como redes neurais e modelos de regressão, são necessárias mais variáveis explicativas de forma a treinar e testar um modelo e prever possíveis fugas. Estas variáveis podem estar associadas à localização geográfica da tubagem, como por exemplo o tipo de solo onde se localizam ou a distância a que se encontram de um determinado objeto, como vias ferroviárias.

A Tabela 2 condensa os parâmetros utilizados em 16 publicações diferentes realizadas por vários autores. Desde logo, é possível verificar uma grande variedade de parâmetros utilizados e que nem todos foram sempre usados pela maioria dos autores. Isto acontece porque para cada modelo, deverá ser feita uma interpretação da área de estudo e perceber quais os parâmetros que eventualmente terão influência no território, como é por exemplo o caso do Índice de Congelamento (*Freezing Index*), que não faz sentido a sua aplicabilidade numa área de estudo que não regista temperaturas negativas. Existem ainda outros parâmetros, que apesar de serem considerados cruciais, a não existência de uma base dados com essa informação, impossibilita a sua utilização na modelação.

Parâmetros	Autores															
	(1) LeGat & Eisenbeis 2000	(2) Mailhot et al. 2000;	(3) Park & Loganathan 2002	(4) Vanrenterghem-Raven 2007	(5) Rogers & Grigg 2009	(6) Berardi et al. 2008	(7) Wang et al. 2009	(8) Wood & Lence 2009	(9) Carrión et al. 2010	(10) Debón et al. 2010	(11) Kleiner & Rajani 2010	(12) Park et al. 2011	(13) Xu et al. 2011	(14) Kleiner & Rajani 2012	(15) Martins et al. 2013	(16) Toubou et al. 2013
Material	X	X		X	X	X	X	X	X	X	X				X	
Diâmetro	X		X	X	X	X	X	X	X	X	X		X		X	X
Idade	X	X		X	X	X	X	X	X	X	X		X	X	X	X
Extensão	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Pressão	X				X			X		X		X				
Tipo de Solo	X	X			X			X		X						
Proteções na Tubagem								X			X					
Uso do Solo		X		X	X			X				X				
Trafego	X			X				X	X	X						
Debaixo de Água				X												
Infraestrutura a sup.				X	X			X		X						
Gravidade Operacional	X															
Tipo de Montagem	X											X				
Pontos de Fornecimento						X						X				
Escoamento								X								
Profundidade							X	X								
Falha nas válvulas											X			X		
Índice de Congelamento											X			X		
Defeito Hídrico											X			X		
Recência														X		
Clustering														X		

Tabela 2 - Parâmetros utilizados em modelos preditivos para fugas de água. Adaptado de: Jenkins, Gokhale, & McDonald, 2014

A grande maioria dos autores, tendo à sua disposição os dados, utilizou e considerou a idade das tubagens, como um fator de grande importância para a ocorrência de fugas. Como qualquer objeto, o desgaste acumulado ao longo do tempo, devido à sua utilização, vai fragilizando os materiais e componentes dos objetos e estes acabam por se tornar suscetíveis a avarias. No caso da infraestrutura de RDA, o ciclo de vida caracteriza-se por três fases distintas.

A primeira fase, representa a cauterização e que pode ter um elevado risco de avarias. No entanto, estas avarias são identificadas ainda quando tubagem não está operacional. Está frequentemente associado a material defeituoso ou instalação feita de maneira incorreta. Estando a sua instalação estabilizada, o ciclo de vida de uma tubagem entra na sua segunda fase- a fase de uso. É neste período que, idealmente, a tubagem passa a esmagadora maioria do seu tempo de vida. O risco de avarias nesta fase é muito baixo, podendo, no entanto, existir variação no caso de fenómenos extremos como por exemplo um terremoto. A terceira fase, a de desgaste, dá-se quando a tubagem começa a atingir o limite de vida útil, e o risco de avarias e a sua frequência vai aumentado. É nesta fase, que as entidades devem identificar as tubagens mais degradadas e proceder à sua manutenção ou substituição (P. D. Rogers & Grigg, 2009).

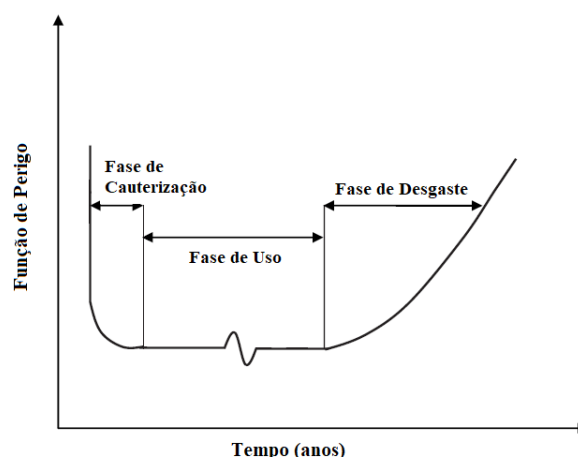


Figura 4- Ciclo de Vida de uma tubagem. Adaptado: Rogers & Grigg, 2009.

As características da tubagem desempenham um papel importante na relação com as avarias na rede. O tipo de material da tubagem, a sua extensão, ou o seu diâmetro são parâmetros que devem ser considerados, quando possível, em modelos preditivos de fugas. Saber em que fase se encontra uma tubagem no seu ciclo de vida, é um dado importante para prever a sua suscetibilidade a fugas. O tipo de material vai estar sempre associado ao tipo de solo onde este se encontra. É por isso impossível o desenvolvimento de modelos generalizados para prever fugas. Enquanto que numa zona A, o tipo de material X pode ser o mais indicado, numa zona B, esse mesmo material poderia ter uma elevada suscetibilidade a fugas. É por isso essencial, que as entidades conheçam o território no momento de instalação da rede, de forma a tomar decisões acertadas na escolha do material (Kutyłowska, 2016).

Conhecer a pressão a que se encontra a água circulante é um dado importante para as entidades gestoras. Por vezes, a pressão não está regularizada de forma correta e, se tiver uma pressão mais elevada do que o recomendável, poderá originar fugas na rede. Uma rede densa, é por norma positivo, uma vez que desta forma a água vai sendo distribuída por diferentes ramais, diminuindo a pressão na tubagem principal. É possível perceber o impacto que a pressão tem, quando é feita a comparação entre as fugas, a pressão e o consumo ao longo de um dia (Figura 5). Assim é possível verificar que nas horas da madrugada, quando o consumo é baixo, a pressão sobre o escoamento aumenta e consequentemente o número de fugas aumenta (Roland Liemberger & Farley, 2004). Conhecer a pressão das tubagens em tempo real, é também crucial para a identificação de fugas. Se forem registadas anomalias nos valores de pressão de uma tubagem, isso pode estar associado a uma possível fuga. Tendo esta informação em tempo-real, pode ser procedido à verificação *in situ*. Na eventualidade de existir fuga, as perdas de água são reduzidas drasticamente. A elevação e declive acabam por estar indiretamente relacionados com

as características de instalação bem como a pressão. Quando ao longo de uma tubagem, o escoamento é feito a jusante, verifica-se uma relação inversa, se a elevação vai diminuindo ao longo da tubagem, a pressão aumenta devido à gravidade (Wannapop et al., 2018).

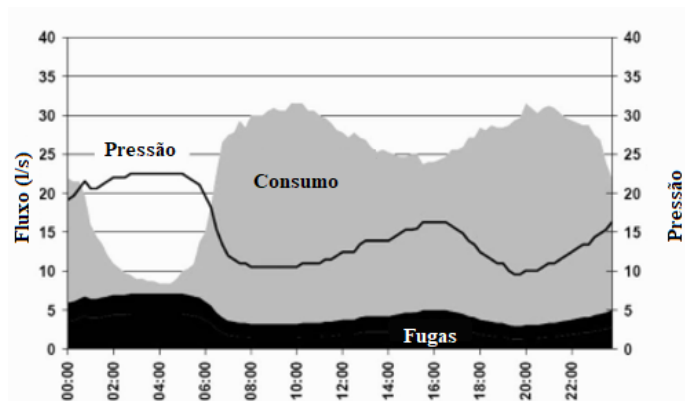


Figura 5- Modelação de fugas num período de 24 horas. Adaptado de: R Liemberger & Farley.

Em 2014, foi publicado um estudo sobre avarias na RDA numa cidade espanhola, em que se realizou uma análise sobre a relação entre os diferentes tipos de vias da rede viária e fuga na RDA. O tráfego foi classificado em três categorias diferentes: passeios pedonais, vias com tráfego normal e vias com elevado tráfego. Neste estudo foi possível verificar que o número de avarias nas duas primeiras categorias foi 64% e 42%, respetivamente, inferior relativamente às fugas ocorridas em vias com elevado tráfego. A influência da rede viária e de maquinaria pesada (rede ferroviária) vai sempre depender da proximidade a que a infraestrutura se localiza, mas também a que profundidade está localizada a tubagem. Acabam por ser três variáveis que se encontram fortemente relacionadas. As infraestruturas subterrâneas estão preparadas para suportar perigos naturais, no entanto, a sua degradação é natural ao longo do tempo, especialmente quando está constantemente sobre o stress de vibrações no solo provocado pelo tráfego de viaturas (García-Mora et al., 2015).

O uso do solo desempenha também um papel na problemática de fugas na RDA, tanto na sua ocorrência como na resolução do problema. A localização de uma fuga sobre um tipo de solo específico pode ter agravantes económicas e estruturais, como é o caso de fugas ao longo de uma via principal, sob um edifício residencial ou sob serviços cruciais para a população, como é o caso de escolas ou hospitais. O custo associado à reparação de fugas nestas localizações poderá não valer a pena o investimento para a sua reparação. Em modelos para a tomada de decisão sobre a ordem de reparação, o uso do solo acaba então por ter um papel preponderante. Uma fuga sob a rede ferroviária pode ter um custo 50% superior ao expectável, e uma fuga ocorrida num espaço

de serviços sociais pode ser superior, sendo no entanto difícil quantificar, uma vez que cada ocorrência é um caso específico (P. D. Rogers & Grigg, 2009). A RDA torna-se então mais suscetível a avarias uma vez que a sua manutenção e reparação apresentam elevados custos neste tipo de solo. O uso do solo acaba também por estar relacionado com o volume de água consumido. O volume de água consumida será naturalmente superior em zonas residenciais ou industriais do que em florestas. Como já fora analisado (Figura 5), um consumo de água reduzido pode ser prejudicial para a rede, estando, no entanto, dependente da pressão exercida na rede.

2.2 Modelos e técnicas de *Machine Learning*

Alan Turing, figura central do filme “*Imitation Game*”, é considerado o pai da ciência da computação teórica e foi no artigo publicado em 1950, “*Computing Machinery and Intelligence*”, que defendeu pela primeira vez que as máquinas podiam aprender e atuar com base na aprendizagem (Turing, 1950). Isto veio a acontecer nas décadas que o seguiram e constitui uma realidade atual, até mesmo em grandes empresas multinacionais como a Google ou Amazon que utilizam técnicas de *deep learning* e *machine learning* (ML, Aprendizagem Automática) para a recolha de grande quantidade de dados e eficiência na suas tarefas (Mayers, 2019). Quando é desenvolvido um programa, o desenvolvedor fornece instruções passo a passo que o programa obedece. Os programas manipulam a informação e estrutura de dados e fornecem resultados a partir dos cálculos que efetua. Não importa o número de vezes que este programa é utilizado que ele nunca irá aprender a operar autonomamente. No entanto, quando se pretende que o programa aprenda a realizar uma tarefa, abordamos a sua programação de maneira diferente (Sharma, 2020).

No desenvolvimento de ML, é necessário fornecer ao programa dados com respostas expectáveis, normalmente chamados de dados treino (*training data*). O programa aprende com base nos dados inseridos e gera um modelo, que deverá ser treinado. É com esse modelo que consegue prever resultados expectáveis com novos dados e respostas desconhecidas (T. M. Mitchell, 1997).

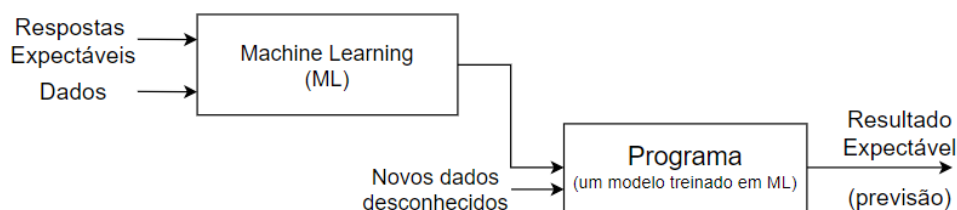


Figura 6- Esquema metodológico para a aplicação de modelos ML. Adaptado de: T.M. Mitchell, 1997.

Os modelos de ML, possuem características muito próprias e como tal, apresentam duas faces. Apesar de ser uma ferramenta poderosa, tem consigo vantagens e desvantagens (Data Flair, 2019).

Vantagens:

- 1) Identifica facilmente tendências e padrões: Modelos de ML conseguem rever grandes quantidades de dados e descobrir tendência/padrões específicos que não seriam identificados por humanos;
- 2) Não necessita de intervenção humana: em alguns modelos não existe a necessidade de monitorizar o projeto a cada fase, uma vez que se dá a possibilidade da máquina aprender e desenvolver os seus próprios algoritmos;
- 3) Melhoria constante: os modelos de ML estão em constante melhoria à medida que vão tendo mais experiência. Isto melhora a sua precisão e consequentemente a tomada de melhores decisões;
- 4) Capacidade em lidar com dados multidimensionais e multivariados: são modelos robustos em ambientes dinâmicos e incertos;
- 5) Aplicação ampla: é possível aplicar modelos de ML em diferentes áreas do conhecimento como a medicina, economia, marketing, ou neste caso, na geografia.

Desvantagens:

- 1) Aquisição de dados: para um modelo de ML ser fiável, necessita um enorme volume de dados, e de boa qualidade. Frequentemente a existência ou não-existência de dados de qualidade é um *deal breaker* para o desenvolvimento de modelos ML;
- 2) Tempo e recursos: os modelos de ML precisam de tempo para aprenderem a desenvolver ao ponto de serem capaz de desempenhar o seu propósito com fiabilidade. Também necessitam de um esforço computacional por parte da máquina onde são desenvolvidos;
- 3) Interpretação de resultados: outro grande desafio passa pela habilidade por parte do utilizador em interpretar os resultados gerados de forma correta.
- 4) Elevada suscetibilidade a erros: apesar da autonomia destes modelos, estão suscetíveis a erros. São especialmente frequentes quando os dados de treino são em pouca quantidade e aviesados. Isto resulta num modelo tendencioso, a prever resultados incorretos e que não podem ser detetados durante largos períodos.

Técnicas de ML têm se tornado ferramentas comuns e aceites para o desenvolvimento de grandes *data sets*, tipicamente para o uso na previsão e classificação de elementos. A necessidade de análises espaciais para perceber as relações de causa/efeito tem sido demonstrado pelas várias

investigações levadas a cabo por geógrafos e matemáticos. Esta necessidade acabou por dar origem à criação do campo dos SIG. A introdução do conhecimento geoespacial, representa um conhecimento mais profundo do que simplesmente a descrição de um local. Por exemplo, saber que a localização de um animal em particular pode estar relacionada com um tipo de vegetação é um conhecimento menos profundo e detalhado do que saber que a existência de um animal y está dependente da existência da vegetação x . Este segundo caso, pode ainda envolver elementos como as características geomorfológicas e tipo de solo. Este tipo de conhecimento carrega muito mais significado em termos de previsão e explicação (Whigham et al., 1992).

Modelos de ML têm sido cada vez mais adotados na comunidade científica juntamente com ferramentas de SIG. Uma pesquisa na base de dados científica *Scopus*, acaba por mostrar isso mesmo. Deste modo, realizou-se uma pesquisa, a 12 de Agosto de 2020, utilizando os termos “*Geographic Information Systems/GIS*” e “*Machine Learning*”, para perceber as características dos artigos utilizando estes dois termos (Figura 7). Com o auxílio do módulo de programação em R, *bibliometrix*, foi possível perceber de forma eficaz a evolução da produção científica utilizada com estes dois termos. A primeira publicação a unir estes dois termos surgiu em 1991, num artigo intitulado “*Machine Induction of Geospatial Knowledge*”, por Whigham, P.A. e McKay, R. Este artigo faz uma análise sobre o potencial da aplicação de modelos ML em análises espaciais, como já fora observado no ponto anterior. Desde então foram feitas 1368 novas publicações (presente na base de dados bibliométrica da Scopus). Entre 1992 e 2013, assistiu-se a um aumento gradual no número de publicações por ano. Foi a partir do ano 2014 que se começou a verificar um aumento exponencial de publicações, tendo atingido o seu pico 2019 com perto de 350 artigos. Tendo em conta que na altura da realização deste estudo, ainda decorria o ano de 2020, é possível perspetivar que o número de publicações no presente ano atingir valores próximos a 2019.

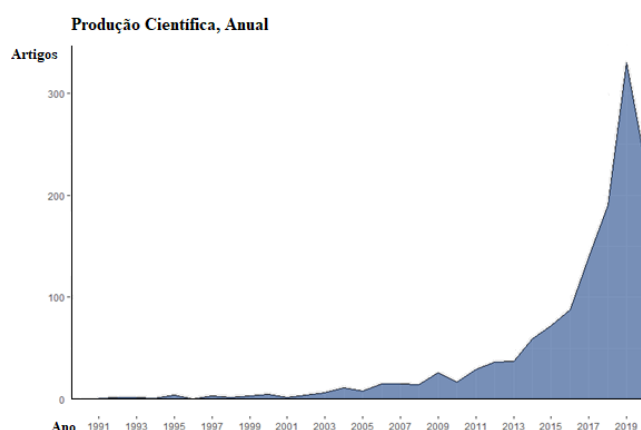


Figura 7 - Produção Científica com os termos “*Geographic Information Systems*” e “*Machine Learning*”. Fonte: Scopus.

networks province multiple techniques support flood model ensembles forests basin prediction reservoir monitoring



Dans le langage **ML**, existent différents langages modules et génériques. Nous avons

Dentro do ramo de ML, existem diferentes domínios, modelos e técnicas. Neste capítulo, será feita uma abordagem sobre os três principais domínios de ML: modelos de aprendizagem supervisionada, aprendizagem não-supervisionada e aprendizagem por reforço (Arroyo, 2017). Nos dois primeiros, existem ainda subdomínios com características próprias e com diferentes objetivos. Realizou-se uma abordagem sobre os principais domínios e, por fim, uma análise mais detalhada sobre diferentes modelos de aprendizagem supervisionada, tendo em conta que os modelos aplicados neste estudo são baseados neste domínio, uma vez que os resultados que se pretende obter, enquadram-se nas características deste.

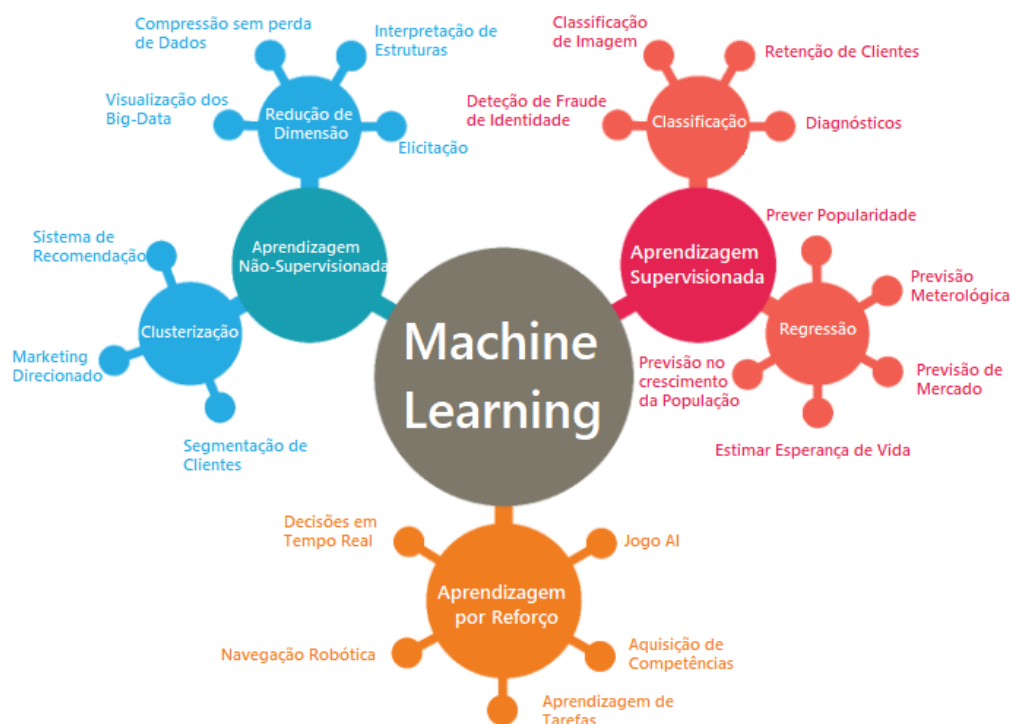


Figura 9 - Diferentes Domínios e Técnicas de Machine Learning. Adaptado de: Arroyo, 2017.

A aprendizagem supervisionada é a metodologia mais importante em ML (Cunningham et al., 2008). Este tipo de abordagem pressupõe que os dados de entrada estejam corretos com base em experiências passadas e que já se conheça o resultado, no momento de desenvolvimento e treino do modelo. O modelo consegue então, com base nessas experiências passadas, “devolver” um resultado, quando apresentado um novo conjunto de dados desconhecidos (Stuart & Peter, 2003). As variáveis podem ser discretas ou contínuas, resultando num problema de classificação ou regressão. Nesta dissertação pretende-se, através de diferentes modelos, prever uma possível avaria em cada tubagem da RDA, mas também avaliar a suscetibilidade de cada tubagem. São duas abordagens distintas que irão produzir resultados distintos. No primeiro caso, irá tentar prever um evento, ou seja, a ocorrência ou não de uma fuga, enquanto que no segundo caso o resultado obtido será uma probabilidade de fuga num futuro próximo.

A aprendizagem não supervisionada é um conjunto de técnicas onde o modelo é treinado com dados onde não são indicados os resultados ou qualquer desfecho desejado. Isto quer dizer que os dados de introdução para treino são idênticos aos dados introduzidos após o modelo estar treinado. O objetivo que se pretende com este tipo de aprendizagem é a identificação de

subestruturas nas amostras, algo útil para a organização e *clusterização* de dados. A avaliação da qualidade destes modelos é difícil, uma vez que não existe nenhuma unidade de medida nos dados de treino para a separação dos clusters (Shalev-Shwartz & Ben-David, 2014).

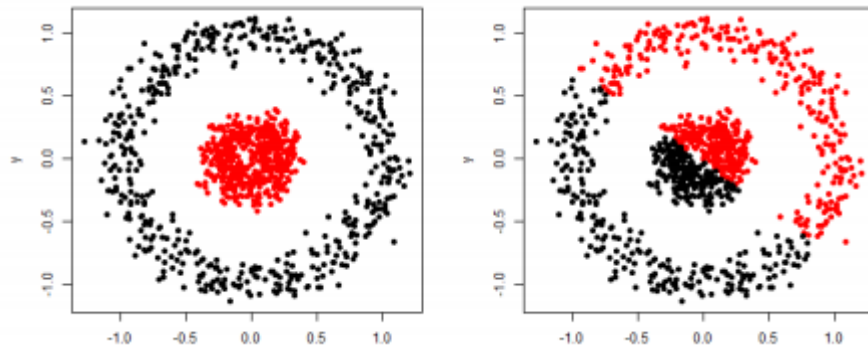


Figura 10 - Dois possíveis resultados de clusterização a partir aprendizagem não-supervisionada. Fonte: Dey, 2016.

Em modelos de aprendizagem por reforço, existe um controlador que interage com um processo através de três sinais: um *state signal*, que descreve o estado do processo, um *action signal*, que permite ao controlador influenciar o processo e um *scalar reward signal*, que devolve ao controlador *feedback* relativo ao seu desempenho imediato. O comportamento do controlador é ditado através das suas políticas e função que deriva a situação para ações. O processo comportamental é descrito pelas suas dinâmicas que determinam como o estado se altera com base nas ações do controlador. O desenvolvimento destes modelos são altamente complexos e representam elevados custos monetários. Este tipo de modelo é frequentemente utilizado no campo da robótica, mas também no *online* em jogos, vulgarmente apelidados de *bots* (Busoniu et al., 2010).

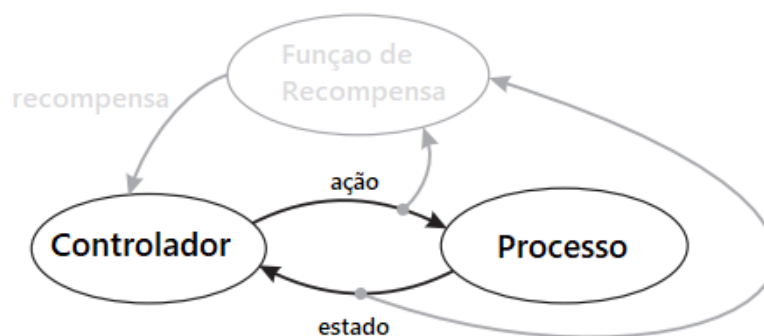


Figura 11 - Interação num modelo de aprendizagem reforçada. Adaptado: Busoniu, Babuska, Schutter, & Ernst, 2010.

Como já fora observado, as técnicas que serão utilizadas no desenvolvimento deste trabalho são técnicas de aprendizagem supervisionadas. Nesta fase final do capítulo, serão discutidos os diferentes algoritmos, em específico aqueles que serão adotados no desenvolvimento da investigação.

2.2.1 Decision Trees

Uma árvore de decisão é uma representação da decisão tomada para determinar uma classe a um dado momento (Utgoff, 1989). São testes não-paramétricos usado para classificação e regressão. O objetivo é criar um modelo que prevê um valor de um objeto, simplesmente obedecendo regras a partir de dados independentes a esse objeto. É uma ferramenta de apoio para a tomada de decisão, mas também para prever eventos, custos e utilidade (em termos económicos). O nome tem origem no formato do seu fluxo, que se assemelha a uma árvore. O processo tem início na raiz da árvore (por norma representada ao contrário), que se vai ramificando graças a diferentes conjuntos de regras. Na Figura 12, observamos um exemplo simples de uma árvore de decisão. A representa a totalidade da população ou amostra e esta é dividida a partir da primeira regra e dois ou mais conjuntos. A primeira regra está, por norma, associada à variável que melhor divide o conjunto de dados, neste caso para a classificação de animais a primeira pergunta que deverá ser feita é se o animal tem penas. Cada ramo da árvore, representa uma divisão do conjunto de dados. O primeiro conjunto serão os animais com penas e o segundo conjunto animais sem penas. Apenas com esta questão não é ainda possível classificar o animal e por isso surgem outras questões (nó de decisão). Para o conjunto de dados em exemplo, é possível classificar o animal, isto acontece quando já não é possível dividir os dados em novos

conjuntos e não vale a pena colocar mais questões (sub-nós). É aplicada a designação de “folha” quando não é possível dividir o conjunto e é atribuída uma classificação ou valor.

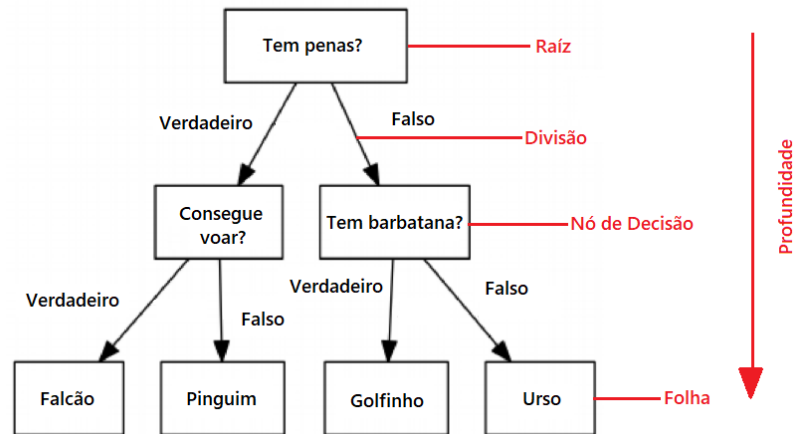


Figura 12 - Árvore de Decisão para a classificação de animais.

No desenvolvimento de um modelo de árvore, com uma *dataset* com várias variáveis, decidir a posição de cada uma delas na árvore pode ser uma tarefa complexa. Se for utilizada uma ordem aleatória, os resultados podem não ser positivos e a precisão do modelo pode diminuir. Para resolver este problema, foram desenvolvidas soluções por alguns investigadores, nomeadamente na utilização de critérios como: entropia, ganho de informação, Índice *Gini*, Chi-quadrado, Redução na Variância e o Rácio de ganhos. Estes critérios são aplicados a cada uma das variáveis e aquelas que obtiverem melhores classificações, são colocados na raiz da árvore. No caso do modelo desenvolvido neste estudo, foi utilizado o índice de *Gini*, em que se avalia a função de custo para a divisão no conjunto de dados. Favorece grandes partições e é fácil de implementar. Este índice funciona quando a variável dependente é categórica e binária (Wang & Suen, 1984). Acaba por se ajustar ao que se pretende alcançar, ou seja, tentar prever a fuga em cada tubagem, sendo neste caso duas categorias, “com fuga” e “sem fuga”.

$$Gini = 1 - \sum_{i=1}^n (p_i)^2$$

Onde:

p = à probabilidade de um elemento ser classificado como uma classe distinta.

São diversas as vantagens na utilização de árvores de decisão, nomeadamente:

- 1) Os resultados são simples de perceber e interpretar. É possível visualizar a árvore;
- 2) Não há necessidade de normalização de dados;
- 3) O custo da utilização da árvore é logarítmico no número de dados utilizados para treinar a árvore;
- 4) É capaz de lidar dados numéricos e categóricos;
- 5) É um modelo *white-box* (Gupta, 2017). Ou seja, em cada observação é possível explicar facilmente os resultados através de lógica booleana. Em contraste, redes neurais utilizam modelos *black-box* onde os resultados são mais difíceis de interpretar;
- 6) Possibilidade de utilizar testes estatísticos para validar o modelo, o que aumenta a sua fiabilidade;
- 7) O modelo tem uma boa performance, mesmo existindo dados incertos.

As desvantagens na utilização de árvores de decisão também são várias, tais como:

- 1) Utilizadores podem gerar árvores demasiado complexas que não generalizam os dados de forma eficaz, isto é designado como *overfitting*. É necessário estabelecer alguns parâmetros no momento de treino do modelo, como o número mínimo de amostras por folha e profundidade máxima para evitar este tipo de problemas;
- 2) As árvores podem ser instáveis. Pequenas variações nos dados podem resultar numa árvore completamente diferente. Este problema pode ser mitigado através de conjuntos (*ensemble*, ver secção seguinte);
- 3) Existem conceitos que são difíceis de aprender porque as árvores não são capazes de as exprimir facilmente, como é o caso de problemas de paridade ou multiplexador;
- 4) Utilizadores podem gerar árvores tendenciosas se algumas variáveis independentes dominarem. É necessário balancear a *data set* antes de aplicar a uma árvore.

2.2.2 Random Forest

O algoritmo de *Random Forest* (Floresta Aleatória) foi também aplicado no desenvolvimento deste estudo e trata-se de um método *ensemble*. Este tipo de métodos combina vários algoritmos (Figura 13) para obter uma melhor performance preditiva do que se obteria utilizando apenas um algoritmo (Delibasic et al., 2013). O autor Robi Polikar (2006) identifica cinco principais razões para a utilização de métodos *ensemble*. Começa por identificar razões

estatísticas, dado que uma boa performance em dados de treino, nem sempre se traduzem numa boa generalização e a performance em novos, nem sempre é boa. Quando se lida com um grande volume de dados um classificador/regressor pode não ser suficiente para lidar com um *dataset* tão grande. O autor, para ilustrar este problema dá ainda um exemplo sobre as tubagens de distribuição de gás que abrangem grandes extensões. Nestas redes, as condições materiais e geográficas onde se encontram as infraestruturas, numa determinada área de estudo, podem ser bastante diferenciadas em diferentes localizações. Faz por isso sentido usar diferentes algoritmos para adaptar a cada caso. O oposto também se verifica, quando existem poucos dados. Encontrar algoritmos que se adequem e representem aos dados de treino é crucial para obter um bom modelo, no entanto, é difícil alcançar isto quando os dados são escassos. A utilização de métodos *ensemble* provam-se eficazes para resolver este problema. “Dividir e conquistar”, uma tática comum nos modelos de ML, como por exemplo as árvores de decisão, mas também os métodos de *ensemble*. Por vezes os problemas são demasiado complexos para serem resolvidos com apenas um classificador e por isso, faz sentido limitar e separar os dados. A última razão que autor identifica é a fusão de dados, para quando existem diversas fontes de dados, a sua natureza também é diferente. Um único classificador, não consegue aprender quando existem diversas fontes. É por isso necessário saber interpretar o cenário e usar vários classificadores para que o modelo consiga digerir as várias fontes de dados e ser capaz de interpretar os dados e gerir informação (M Gashler et al., 2008).

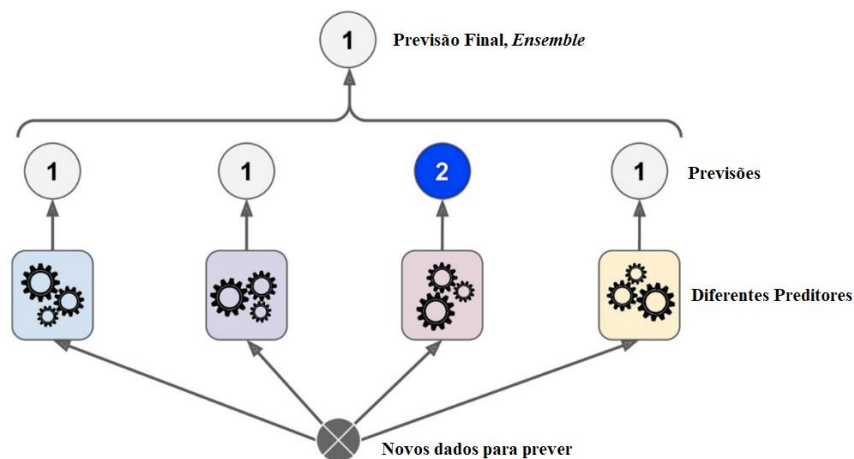


Figura 13 - Fluxo de métodos ensemble. Adaptado: Géron, 2017.

Face a estes motivos apresentados, faz todo o sentido considerar um método de *ensemble* para além da classificação da árvore de decisões. Vale a pena relembrar as características do território e do equipamento, uma vez que a característica da tubagem a norte de VNG pode ter

características diferentes da tubagem existente a sul. Isto é algo que também se verificou aquando a revisão da literatura relativamente às fugas de água.

A título de exemplo, o material que deverá ser utilizado está relacionado com o tipo de solo onde este vai ser aplicado. É então levantada a seguinte questão: Faz sentido utilizar o mesmo classificador para estas duas situações? Outra razão, também levantada por Robi Polikar reside no elevado volume de dados, dando até um exemplo bastante semelhante ao nosso estudo de caso. Tendo em conta o volume de *features* que serão analisadas, faz sentido tentar utilizar diferentes algoritmos.

Por fim, com a árvore de decisões será utilizado um classificador para tentar identificar um evento (com ou sem fuga), enquanto que no modelo *ensemble* a usar, irá atribuir-se um valor de suscetibilidade para cada tubagem.

A escolha sobre o método *ensemble* a ser utilizado acabou por recair sobre a floresta aleatória (*random forest*, *RF*). Os modelos RF derivam de árvores de decisão, onde, em vez de utilizar uma árvore, utilizam-se várias e o resultado é a média/moda de resultados das diferentes árvores utilizadas. Estes modelos são utilizados tanto para a classificação como para regressão e corrigem um dos pontos negativos identificados na utilização das árvores de decisão: o *overfitting*. Para evitar isto, os modelos *ensemble* utilizam meta-algoritmos desenhados para melhorar a estabilidade e previsão do modelo ML, o que é apelidado de *bagging* (ou *bootstrap aggregating*). Esta técnica ajuda o algoritmo a identificar onde deve começar a crescer a árvore (*root*). Outra característica importante no RF, é a avaliação sobre a importância das variáveis (M. Gashler et al., 2008).

O desenvolvimento de um modelo de florestas aleatórias é algo mais complexo e onde o desenvolvedor tem mais controlo. Existem parâmetros que devem ser ajustados durante o desenvolvimento, de modo a calibrar o modelo para se obter a melhor capacidade preditiva possível. Os dois principais parâmetros são o número de estimadores e o número máximo de *features*. O primeiro, como o nome indica, refere-se ao número máximo que uma floresta deverá ter. Por norma, a qualidade do modelo é proporcional ao número de estimadores. No entanto, o tempo de processamento é superior e dependente da capacidade da máquina onde é desenvolvido o modelo. O segundo parâmetro refere-se ao número de *features* que se deve agregar por conjunto, no momento de dividir em subconjuntos (nó de decisão). Neste caso, quanto menor o número, melhores resultados. Pode, no entanto, gerar um enviesamento na análise. Genericamente, para problemas de regressão não deverá ser atribuído nenhum valor. O mesmo se aplica no momento de definir a profundidade máxima de cada árvore. No desenvolvimento do modelo deve-se

procurar otimizar os parâmetros adequados à máquina onde estes são processados, uma vez que os consumos de memória RAM são elevados e em algumas situações a máquina não consegue concluir o processamento.

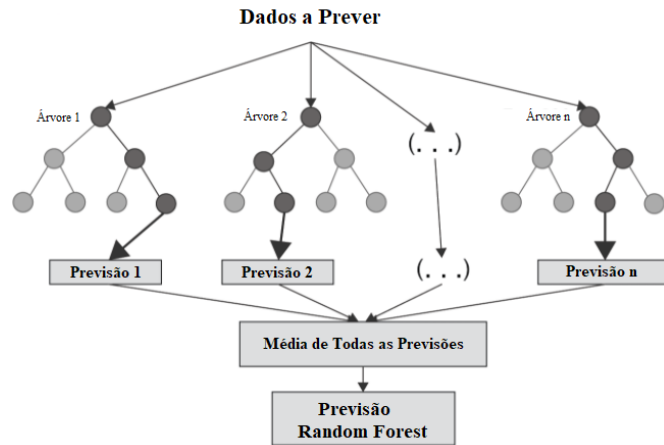


Figura 14 - Esquema Preditivo de um modelo Random Forest. Adaptado: Bakshi, 2020.

CAPÍTULO 3

MATERIAIS UTILIZADOS

Para o desenvolvimento dos modelos preditivos, foi necessário a recolha de parâmetros de vulnerabilidade diferentes. Os parâmetros escolhidos acabaram por ser uma conjugação de variáveis identificadas como relevantes na revisão da literatura estudada, com os dados disponíveis. Os diversos estudos sobre a temática (Breiman, 2001), identificam que os dados relativos à idade da rede de distribuição são os mais flagrantes para a ocorrência de fugas. Uma tubagem com vários anos, implica maior desgaste na sua infraestrutura. No entanto, atualmente, essa informação não se encontra disponível para a área de estudo. O desenvolvimento da rede é um processo gradual e não existe informação sobre o ano de instalação de cada uma das tubagens. Não obstante, foi mantida a ambição de previsões com recursos *open-source*, mas também com os dados disponibilizados pela empresa AGEM.

A recolha e tratamento dos dados foi porventura o passo mais crucial e mais moroso ao longo desta investigação. Foi necessário perceber quais as variáveis que se ajustavam à investigação, realizar a sua recolha e fazer a sua adaptação para a metodologia a adotar. A base de dados da rede de distribuição da AGEM, apesar de não conter dados relativamente à idade das estruturas, é bastante completa no sentido em que a sua georreferenciação está completa, existem dados sobre o material das tubagens e existem dados relativamente ao consumo de cada número de polícia. No entanto, acaba por estar numa situação em que é rica em dados, mas pobre em informação. Os dados são apenas registos isolados de uma determinada característica de um objeto e apenas quando são relacionados com outros dados, é gerada informação e é permitida a sua interpretação (Vaz & Alves, s.d.).

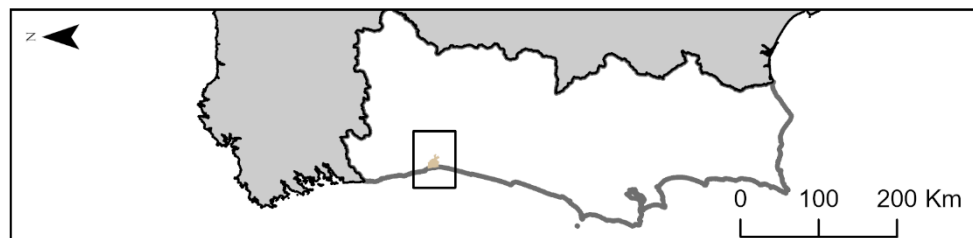
Ao longo deste capítulo, inicialmente, será realizado uma caracterização geral da área de estudo. De seguida, serão descritos os métodos e técnicas utilizados para o tratamento de dados, nomeadamente da variável dependente (as avarias) e as variáveis independentes (parâmetros de vulnerabilidade). Ao longo dessa narração, será feita uma descrição específica do território a partir de cada parâmetro de vulnerabilidade e perceber a correlação com a variável dependente.

3.1 Área de estudo

O conhecimento sobre o espaço físico que nos rodeia é fundamental para a implementação com sucesso de qualquer medida ou política. O município de Vila Nova de Gaia (Figura 15) localiza-se na região Norte (NUT II), inserido na Área Metropolitana do Porto (NUT III). Encontra-se no litoral de Portugal, e as sua margem ocidental encontra-se delimitada pelo Oceano Atlântico. A norte e nordeste, partilha fronteira com os municípios do Porto e Gondomar, ambos separados de VNG pelo Rio Douro. Diretamente a sul e sudeste, partilha a fronteira com os municípios de Espinho e Santa Maria da Feira.

Vila Nova de Gaia é um município com uma área de 168,5km² e, segundo os censos de 2011, uma população de 302 mil habitantes, a terceira maior em Portugal. É um município que atualmente, e após a reorganização administrativa do território em 2013 reorganizou-se em 15 freguesias. No entanto, este estudo procurou manter a norma que a empresa AGEM utiliza para fins estatísticos, continuando a usar as 24 freguesias, prévias à reorganização. Estas freguesias possuem em média uma área de 7km², sendo Pedroso, Sandim e Grijó as freguesias com maior área (19,4km²; 13,5km² e 11,3km², respetivamente). São Pedro da Afurada e Seixezelo são as freguesias com menor área, com 1km² e 1,5km², respetivamente.

Relativamente à RDA, o município possui (até à data da realização deste estudo) uma extensão de rede (todo o tipo de infraestrutura) com mais de 14 mil km, distribuídos por 36 subsistemas de distribuição. Estes subsistemas estão, por norma, servidos pelo menos, por um reservatório, havendo no total 63, repartidos pelo município. Estes reservatórios e subsistemas, encontram-se conectados ao longo do território por infraestruturas de adução. Na Figura 15 é ainda possível conferir as estatísticas referentes à extensão de cada freguesia, bem como a extensão de infraestruturas de distribuição de água.



Freguesia	Área (km2)	Extensão (km)
Arcozelo	8,5	982
Avintes	8,8	662
Canelas	6,9	604
Canidelo	8,9	982
Crestuma	4,7	238
Grijó	11,3	903
Gulpilhares	5,5	559
Lever	7,8	334

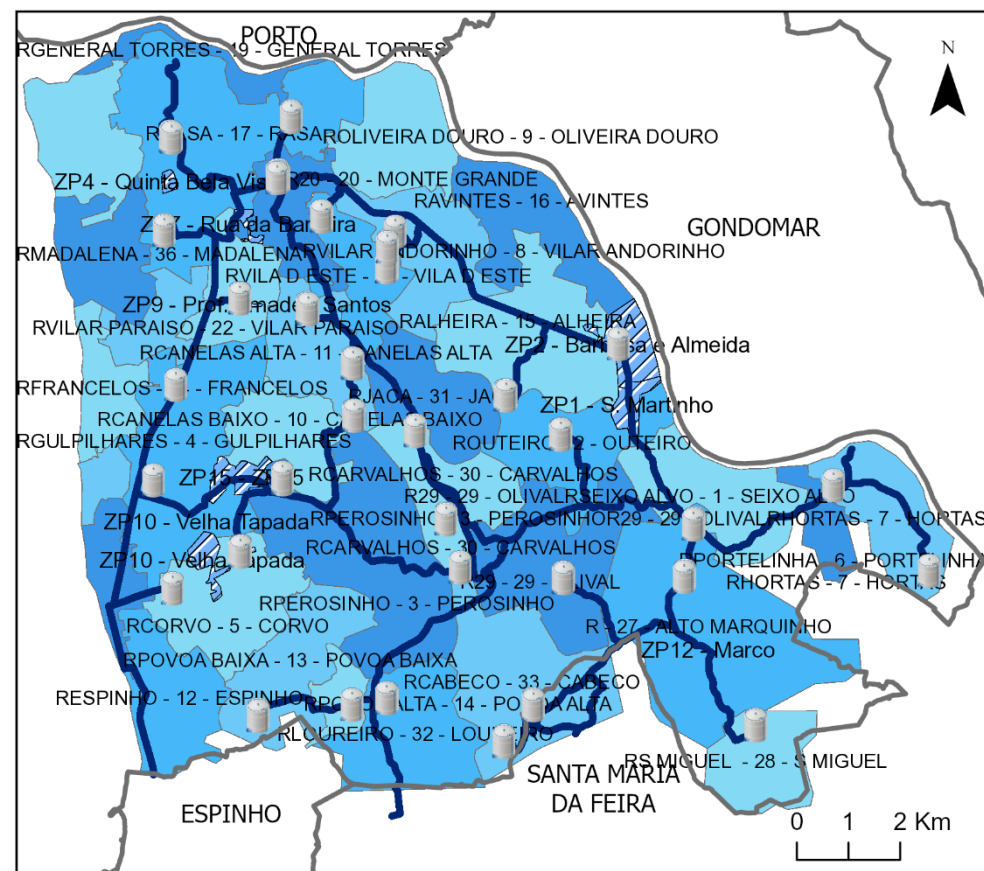
Freguesia	Área (km2)	Extensão (km)
Madalena	4,7	495
Mafamude	5,3	864
Olival	8,0	528
Oliveira Do Douro	7,5	703
Pedroso	19,4	1406
Perosinho	5,3	408
Sandim	13,5	593
São Félix Da Marinha	7,9	710

Distribuição

- Zonas Picagem Distribuição
- Celulas de reservatórios
- Adução
- Subsistemas de Distribuição

Limites Administrativos

- Freguesias - Vila Nova de Gaia
- Portugal
- Espanha



Freguesia	Área (km2)	Extensão (km)
São Pedro Da Afurada	1,0	111
Seixezelo	1,5	107
Sermonde	1,7	108
Serzedo	6,7	519
Valadares	5,1	458
Vila Nova De Gaia (Santa Marinha)	5,9	839
Vilar De Andorinho	7,1	555
Vilar Do Paraíso	5,3	488

Figura 15 - Enquadramento Geográfico de Vila Nova de Gaia. Fonte: AGEM e CAOP.

3.2 Tratamento de dados

Numa primeira fase, foi necessária a georreferenciação das avarias identificadas no período de estudo, e a sua associação com a rede de distribuição de água. Este passo inicial revelou-se um processo mais complexo do que o expectável. O histórico de avarias existentes na AGEM são aquelas em que se procedeu à sua reparação. Estas avarias não se encontravam georreferenciadas, nem se encontravam numa base de dados SQL associadas a uma tubagem específica. O histórico de ocorrências encontrava-se guardado em ficheiros .csv diferentes, separados por ano e com estruturas diferentes. No entanto, em todos os ficheiros havia informação crucial para proceder à georreferenciação de cada avaria: a morada e freguesia onde ocorreu a reparação. A partir destes dois campos, foi possível proceder à geocodificação das moradas, acrescentando informação sobre o município (VNG), distrito (Porto) e país (Portugal) para aumentar a precisão da ferramenta de geocodificação. Para a geocodificação foi utilizado o serviço API “ArcGIS World Geocoding Service”. Tendo os pontos sido gerados automaticamente, procedeu-se a uma validação manual para corrigir moradas que tenham sido localizadas no sítio errado.

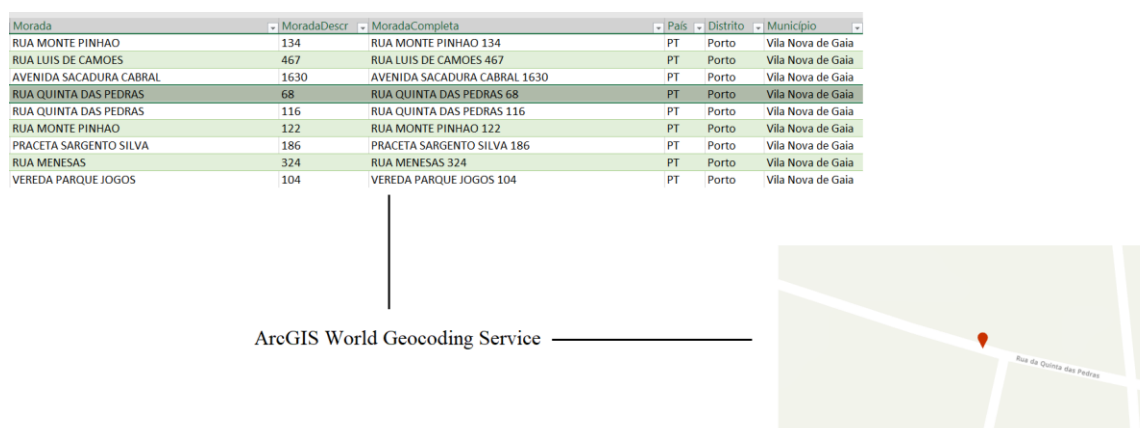


Figura 16 - Processo de Georreferenciação de Moradas.

Nesta fase surgiu um problema nos dados, uma vez que as avarias identificadas podem referir-se a qualquer tipo de avaria na RDA. Como já fora referido, a infraestrutura da RDA é composta por vários tipos de tubagem com diferentes funções, ramais que fazem a ligação dos clientes à rede, válvulas para regular a pressão e outros acessórios. O somatório de todas as *features*, de todas estas categorias ultrapassariam a centena de milhares e comparativamente com um número de avarias muito reduzido, seria impossível desenvolver um modelo capaz de fazer a previsão para tantas *features* (e com características diferentes) com o histórico existente. O

histórico de avarias (reparações) existente é referente aos anos de 2017, 2018 e 2019, contabilizando mais de 1700 pontos. No entanto, conforme a problemática já descrita, as entidades estimam que o número de avarias não corresponde ao das reparações, e os serviços suspeitam de mais avarias não identificadas, algo que os modelos a serem desenvolvidos poderão dar resposta. A Figura 17, dá-nos a informação que a grande maioria (82%) das avarias ocorreram em ramais, ou seja, nas linhas de ligação entre os clientes e as condutas principais. Nestas últimas, ocorreram 15% das fugas e as restantes avarias estão associados a equipamentos e acessórios onde têm uma representação pouco significativa.

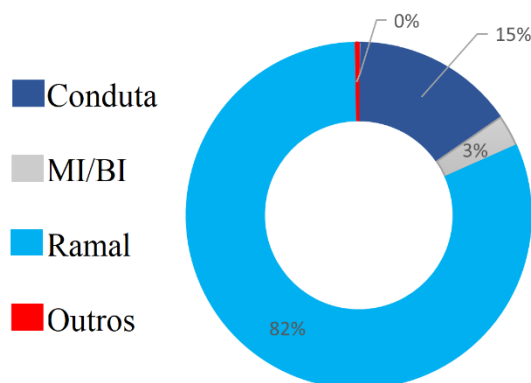


Figura 17 - Tipos de avaria durante o período de estudo. Fonte: AGEM.

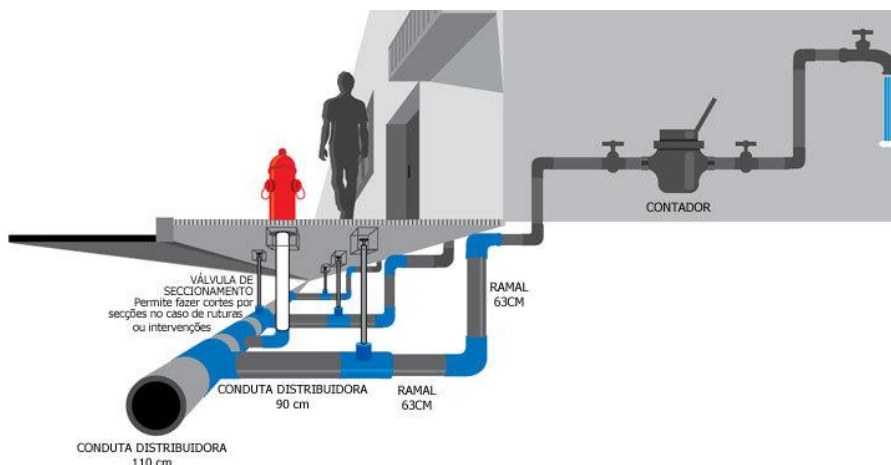


Figura 18 - Ilustração de uma RDA e as suas infraestruturas. Fonte: CM Sesimbra (2017).

Para resolver o problema identificado foi precisar pensar no objetivo inicial desta investigação. Com a utilização de ferramentas de SIG e ML, pretende-se obter uma representação espacial de possíveis avarias não identificadas e realizar uma carta de suscetibilidade. Por isso, fez sentido associar as avarias a um único tipo de equipamento. Tendo em conta que é nos ramais

que se localizam a maioria das avarias, faria sentido abordar o problema a partir dos mesmos. No entanto, o número de *features* nos ramais é demasiado elevado para se obter um modelo fiável (sensivelmente, 1 ramal por cada número de polícia). Acresce que os ramais possuem extensões bastante reduzidas e sem características diferenciadas. Os ramais vão estar sempre associados à tubagem (conduta) principal, e esta sim, tem características próprias como o tipo de material, o número de clientes que serve, consumos, etc. Assim, estabelece-se uma relação de um para muitos (uma tubagem principal faz a ligação com x ramais) (Figura 19), significando que o número de *features* referentes à tubagem é muito menor e irá garantir maior fiabilidade nos modelos a desenvolver. Existe ainda o fator de que uma avaria num ramal irá ter impacto noutros ramais servidos pela mesma tubagem (Figura 20).



Figura 19 - Amostra da Rede de Distribuição de Água. Fonte: AGEM

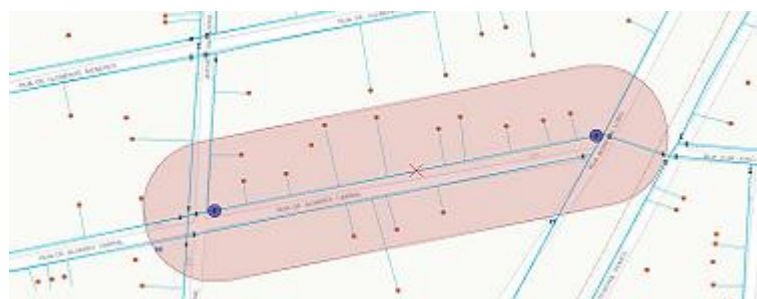


Figura 20 - Isolamento de roturas. Fonte: AGEM

Estabelecido que o modelo irá lidar com as tubagens da RDA, levantam-se duas questões adicionais:

1. O resultado do *geocode* não garante que o ponto gerado intersecte com uma linha que representa a tubagem e não existe no .csv original uma chave estrangeira que identifique a *feature* onde esta ocorreu. Como deverá ser feita a ligação entre os pontos de avaria e a tubagem?

2. Sabendo que a tubagem é representada por uma linha contínua, como deverá ser feita a associação de cada tubagem a cada uma das variáveis independentes, sabendo da possibilidade dos valores serem diferentes ao longo da tubagem?

Para abordar ambas as questões, foi necessária alguma agilidade e adaptação para dar uma resposta. No primeiro caso, não havendo informação digitalizada sobre a *feature* onde ocorreu a reparação, foi necessária uma solução automática para todos os casos. Recorrer aos trabalhadores da AGEM para identificar a tubagem em causa não era algo exequível, especialmente quando se está a lidar com milhares de reparações. Foi por isso realizado um cálculo de proximidade entre os pontos geocodificados e as tubagens da RDA. O resultado deste cálculo são duas colunas na tabela dos pontos de avaria em que uma é referente ao ID da tubagem e a outra é referente à distância entre as duas *features*. Com a primeira coluna foi então criada um relacionamento de um para muitos (uma tubagem pode ter tido várias avarias) entre as duas *feature class*. A segunda coluna acabou por ter também a sua utilidade, na medida em que quando se registavam elevadas distâncias, poderiam revelar indícios da morada ter sido geocodificada de forma incorreta. Com esta abordagem, é agora possível identificar as tubagens onde não ocorreram quaisquer reparações nos três anos em causa, as que tiveram avarias e quantas avarias ocorreram.

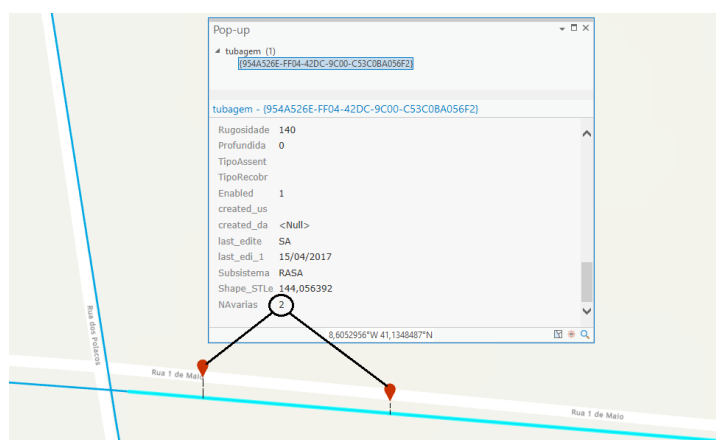


Figura 21 - Número de avarias por tubagem calculado a partir da proximidade das ocorrências. Fonte: AGEM.

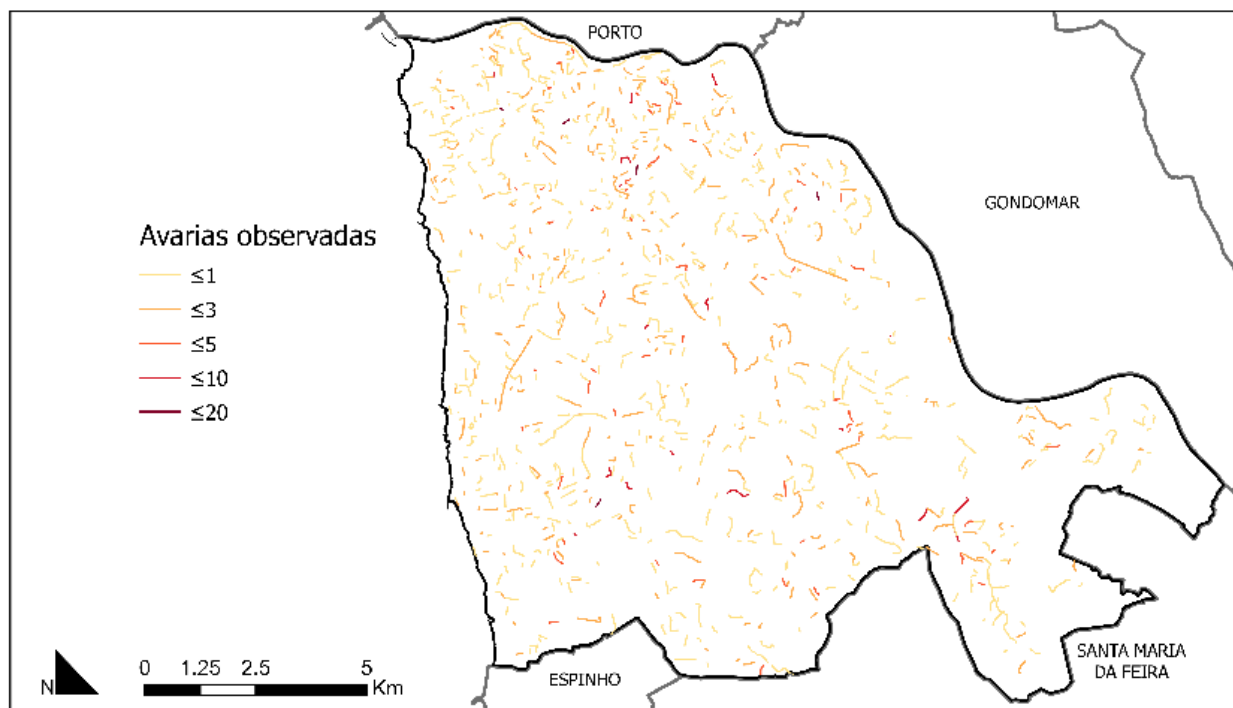


Figura 22 - Somatório de avarias associadas à RDA.

Para responder à segunda questão levantada, foram testadas duas abordagens. Numa primeira, foi utilizada a ferramenta de desenho de pontos ao longo de linhas (neste caso as tubagens). Aqui, resultaria um ponto por cada linha, localizado a meio do segmento (Figura 23). A cada um destes pontos foi associado o valor respetivo das variáveis independentes através de ferramentas de análise espacial: união por localização no caso de dados vetoriais e extração de valores por pontos no caso de dados em formato *raster* (Figura 25).

Este procedimento acaba por levantar outra questão, em especial quando se está a lidar com dados contínuos, como por exemplo a altitude. A altitude é um elemento que pode ter grandes variações ao longo de uma tubagem e por isso, o valor de elevação a montante da tubagem pode ser completamente diferente a jusante. Foi por isso testada uma variante do procedimento adotado anteriormente, só que neste caso, foram gerados dois pontos por cada tubagem: um a montante e outro a jusante (Figura 24). Foi então feita a recolha dos dados referentes às variantes independentes e associados aos *end-points* das tubagens. Esta estratégia acabou por se revelar inviável, uma vez que, como pode ser observado na Figura 24, as extensões das tubagens são frequentemente demasiado curtas, o que resulta numa duplicação de features, sem o acrescento de informação, algo que iria diminuir a qualidade dos modelos. Este procedimento foi ainda aproveitado para testar a hipótese da variação de uma variável ser um possível parâmetro de vulnerabilidade para fugas na RDA. Isto foi testado utilizando a variável elevação, calculando a diferença entre a elevação a montante e a jusante. Os resultados obtidos não foram positivos, na

medida em que as extensões da grande maioria das tubagens são demasiado insignificantes para se verificar uma diferença relevante/conclusiva. Foi então utilizado o primeiro método testado, utilizando o ponto central de cada tubagem, evitando assim a duplicação de dados e criação de ruído no modelo.

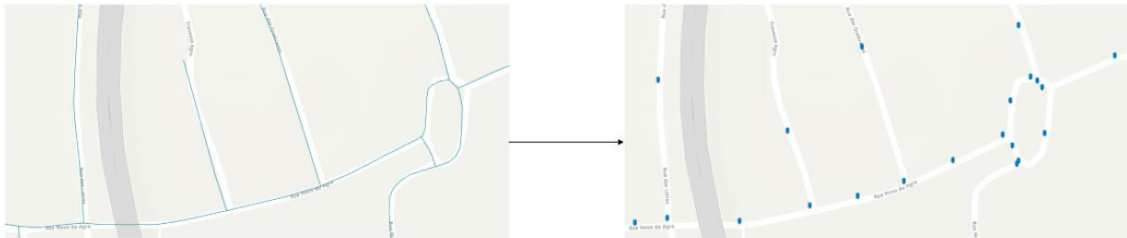


Figura 23 - Conversão de tubagens em formato de linha para ponto central.

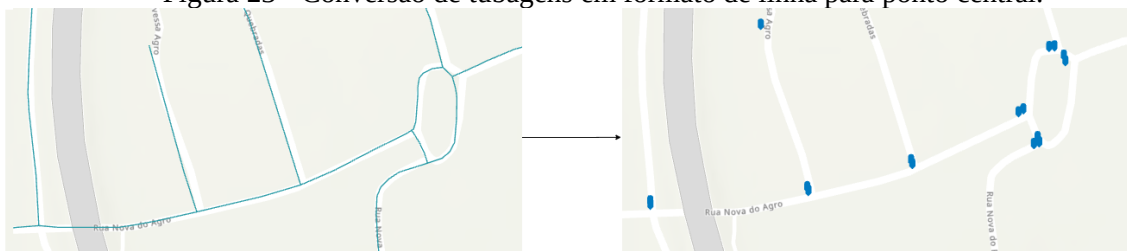


Figura 24 - Conversão de tubagens em formato de linha para *end-points*.

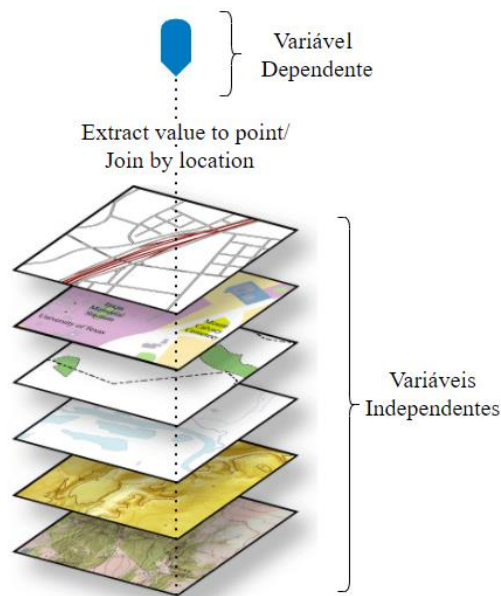


Figura 25 - Associação dos valores das variáveis independentes à variável dependente.
Adaptado de: ESRI, Inc.

3.3 Desenvolvimento dos modelos preditivos

A relação dos diferentes parâmetros de vulnerabilidade com a RDA foi desenvolvida com recurso ao software *ArcGIS Pro*, com o qual se organizou os dados numa *geodatabase*. Estes foram trabalhados recorrendo a diferentes plataformas e bibliotecas (

Figura 26). Os modelos foram concebidos com recurso à linguagem de programação *python*. Para isso, recorreu-se à plataforma *Anaconda* para a gestão dos módulos e bibliotecas, e todo o código desenvolvido elaborou-se nos *Jupyter Notebooks*. A biblioteca *scikit-learn* foi crucial para o desenvolvimento destes modelos, uma vez que é uma biblioteca com os algoritmos de *machine learning*. A biblioteca *pandas* permitiu a leitura das *data frame* com os dados de treino e dados a prever. Foi ainda utilizado a biblioteca *NumPy* para facilitar a leitura de alguns resultados, dentro do *notebook*, ao longo do código desenvolvido. As bibliotecas *seaborn* e *matplotlib* foram utilizadas para a criação de gráficos. Por fim, os resultados de previsão foram novamente importados para ambiente SIG para a sua especialização. Vale a pena salientar que todo este processo foi realizado com recurso a ferramentas *open-source*, com a exceção do software *ArcGIS Pro*. No entanto, os processos aqui realizados poderiam ser replicados em software *open-source* como o *QGIS*.



Figura 26 - Diferentes plataformas e bibliotecas utilizados para o desenvolvimento dos modelos preditivos.

Os dados utilizados para o desenvolvimento das variáveis explicativas, proveem de diversas fontes. A grande maioria deles são dados *open-source* e os restantes foram disponibilidades pela empresa Águas de Gaia. As variáveis explicativas podem ser divididas em três categorias. Numa primeira, é referente aos atributos da própria rede e da tubagem. A segunda categoria é referente à distância a certos objetos que podem influenciar o comportamento das infraestruturas. Por fim, as variáveis podem ser categorizadas por dados de fenómenos contínuos. Ao contrário da categoria a anterior (que de certa forma foram representados continuamente), estes não têm

qualquer relação com objetos.

Para o desenvolvimento dos dados, foram seleccionadas áreas treino para treinar e preparar o modelo. Foram desenvolvidas várias variações e ajustes nos parâmetros e verificar com os dados teste a qualidade dos modelos que estavam a ser produzidos.

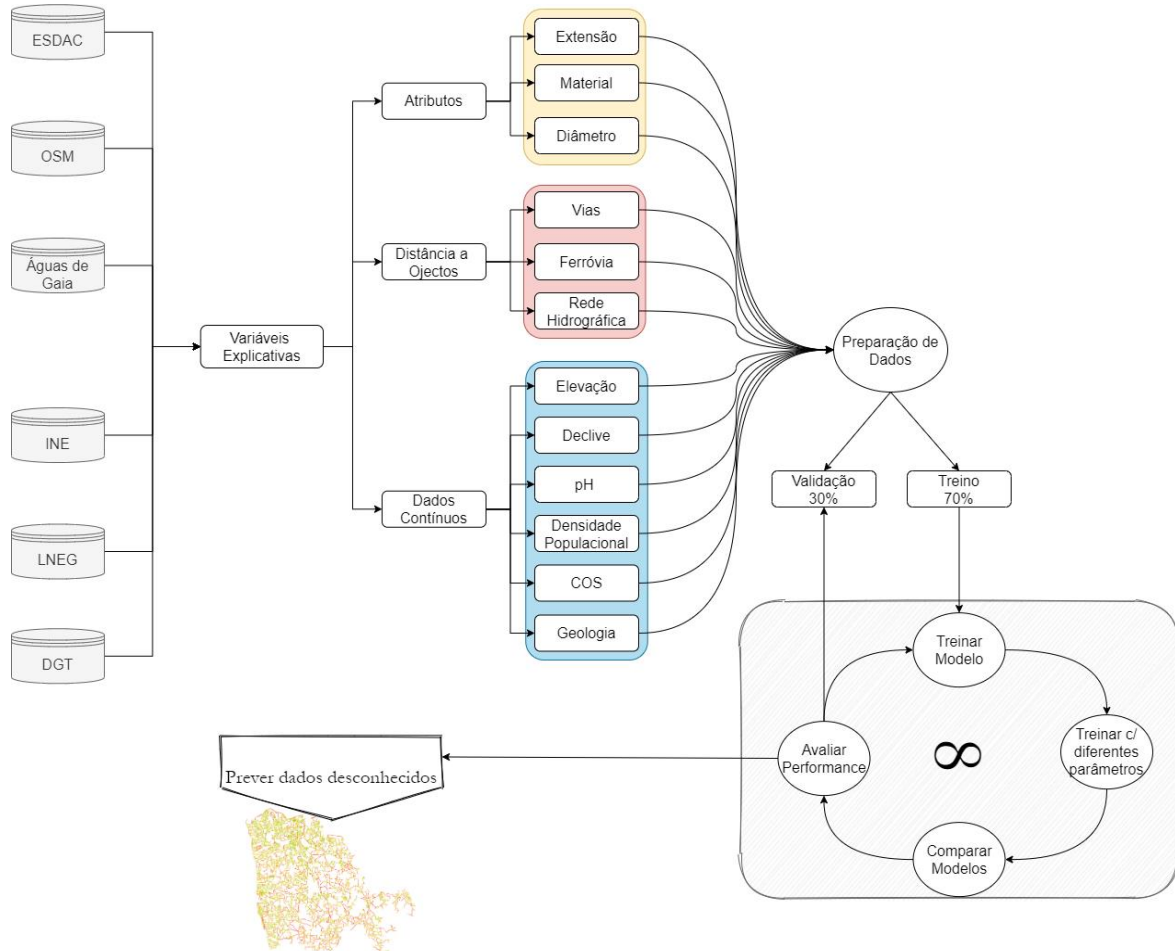


Figura 27 - Diferentes parâmetros de vulnerabilidade seleccionados e processos para o desenvolvimento dos modelos preditivos.

CAPÍTULO 4

PARÂMETROS DE VULNERABILIDADE GEOGRÁFICA

4.1 Dados Contínuos

4.1.1 Elevação (altitude)

A orografia é identificada como um dos principais fatores relacionados com fugas de água. Os diferentes valores de elevação podem causar condicionamentos no momento de instalação das tubagens (Jenkins et al., 2015).

Para o desenvolvimento da carta de elevação, foram utilizados dados topográficos, especificamente curvas de nível a 5 metros, e marcas topográficas (pontos), fornecidos pela AGEM. A partir destes dados, foi gerada uma triangulação com o método *Delaunay*. Este método assegura que nenhum vértice se encontra dentro de um circuncírculo. Tendo em conta que os dados utilizados, o resultado é obtido é um Modelo Digital de Terreno (MDT), visto que não são contemplados os elementos à superfície (MDS). Esta questão acaba por ser relevante, tendo em conta a temática em estudo, uma vez que a rede se encontra no interior do solo. Os dados de entrada e saída encontram-se em formato vetorial, é por isso pertinente converter para formato *raster*, visto estarmos a lidar com dados contínuos.

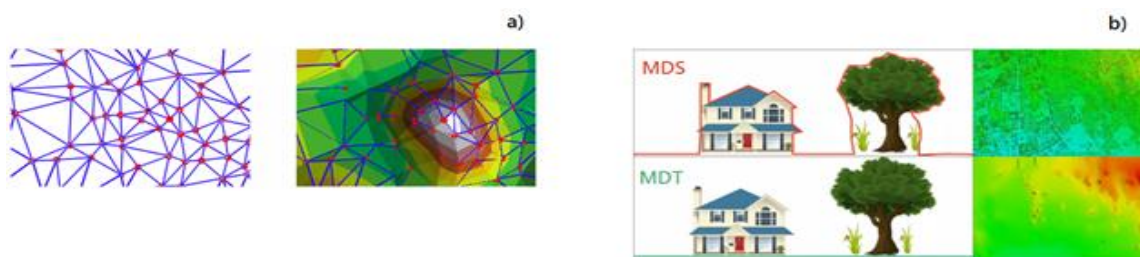


Figura 28 - a) Triangulação de dados topográficos. b) Diferença entre um MDS e um MDT. Fonte: ESRI.

A elevação do município de Vila Nova de Gaia é relativamente baixa, uma vez que o ponto mais elevado não ultrapassa os 261 metros, e se localiza entre as vilas de Sandim e Seixezelo. No mapa hipsométrico (Figura 30), é possível verificar que a elevação no município é escalonada. Esta situação torna-se mais flagrante quando observamos a reclassificação da

elevação, onde existe um conjunto de elevações alinhadas seguindo a direção NNO/SSE.

Sensivelmente 20% do município encontra-se a uma altitude abaixo dos 50 metros. Estas áreas são essencialmente áreas litorais, mas também as áreas que marginam Rio Douro, no sentido Este-Norte. Na faixa litoral, a altitude permanece abaixo dos 50 metros, em média, por 2km no sentido interior do município. A faixa junto ao rio Douro é mais reduzida, mas bastante mais variável que na situação anteriormente referida. É ainda possível encontrar áreas abaixo dos 50 metros, nas margens dos Rios Uima e Febros.

Cerca de 60% do território do município situa-se entre os 50 e 150 metros, sendo que acabam por se encontrar entre as margens dos rios/oceano e os pontos mais altos, nomeadamente a Serra de Canelas e o Monte da Virgem. Apenas 15% do território encontra-se a uma elevação superior a 150 metros e sensivelmente 5% estão a mais de 200 metros.

Efetuada uma análise entre as fugas ocorridas e a elevação, verificamos que existe uma correlação negativa. Uma regressão linear, comprova que quanto menor a elevação, menos fugas ocorrem. É, no entanto, importante referir que esta análise é redutora. A elevação do município é bastante reduzida e concentrada entre os 50-150metros, é por isso natural que a maior parte das fugas ocorram nestas áreas. Acumulando a este facto, como mais à frente irá ser analisado, a rede de distribuição concentra-se essencialmente nestas áreas. Cerca de 65% das fugas ocorreram neste intervalo de valores de elevação.

Efetuada uma análise entre as fugas ocorridas e a elevação, verificamos que existe uma correlação negativa. Uma regressão linear, mostra que quanto menor a elevação, menos fugas ocorrem. No entanto, esta não é uma tendência muito forte visto que o declive da reta não é elevado. É possível ainda identificar um elevado número de avarias num intervalo de elevação. Aproximadamente 1/3 das avarias encontra-se entre os 68 metros e 104 metros de elevação.

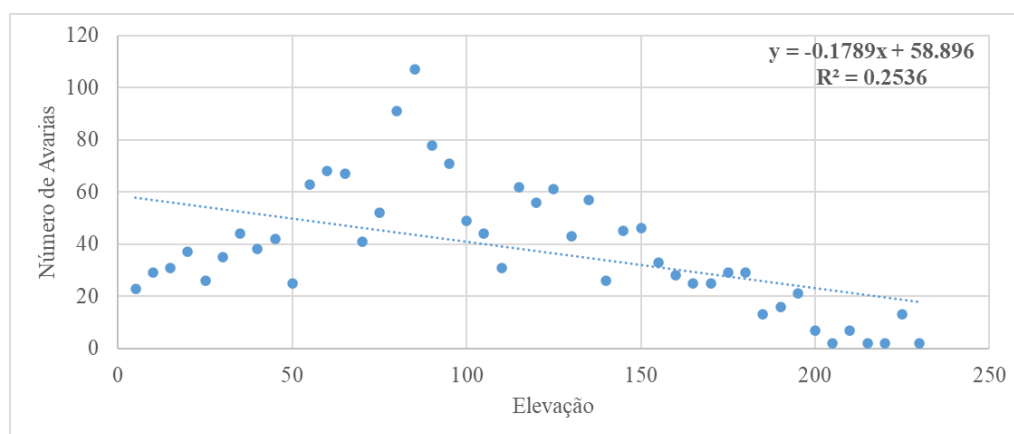


Figura 29 - Relação entre avarias observadas e variável elevação.

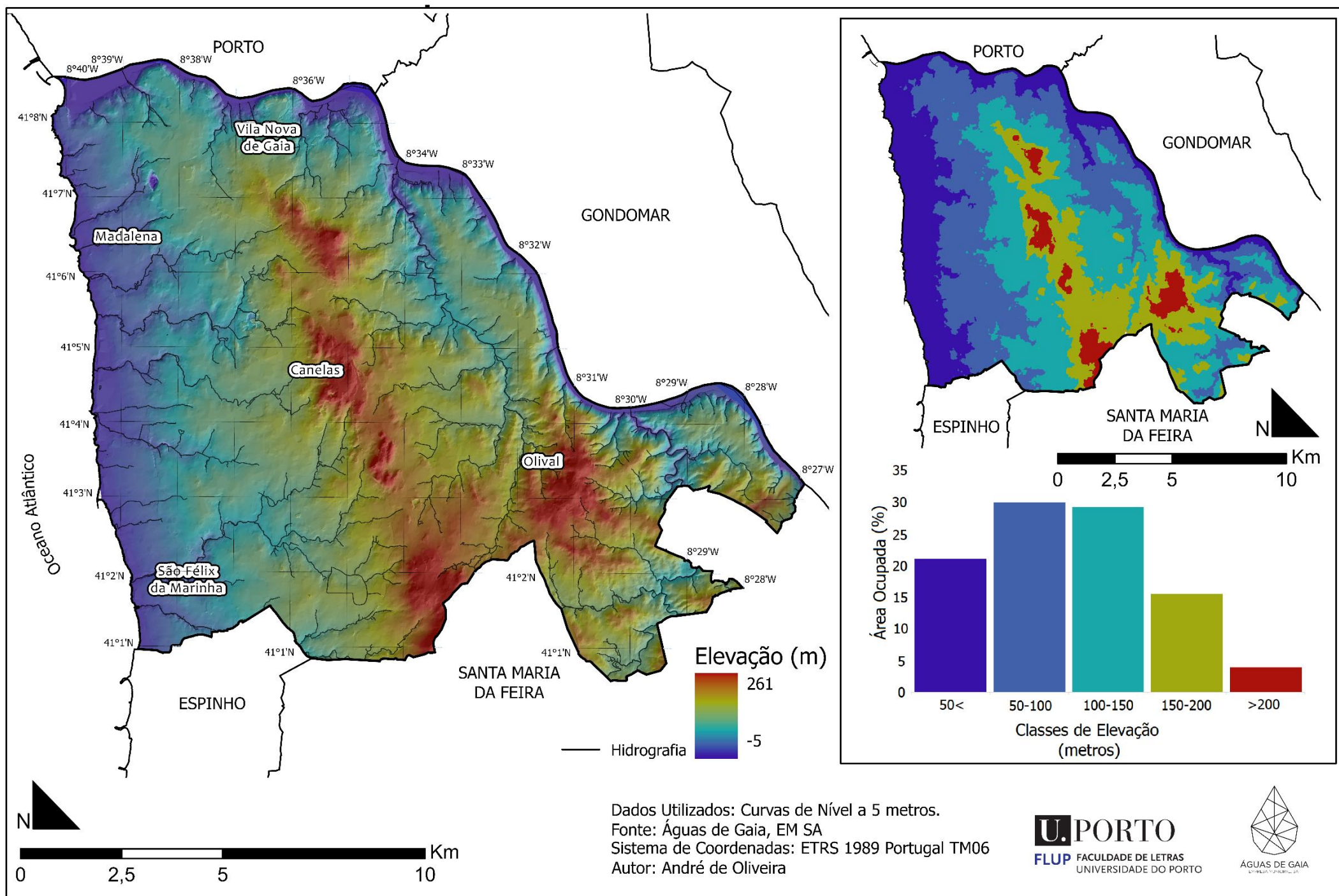


Figura 30 - Carta Hipsométrica de Vila Nova de Gaia.

4.1.2 Declive

Para gerar a carta de declives, foi necessário utilizar os resultados obtidos no desenvolvimento da carta de elevação. Os declives são calculados com base nos valores de elevação. Para cada pixel, o valor de declive é calculado com base na diferença máxima de elevação entre os seus vizinhos. Os declives podem ser calculados em graus ou em percentagem. No caso deste estudo, os valores serão apresentados em graus, seguindo a norma utilizada em planos municipais.

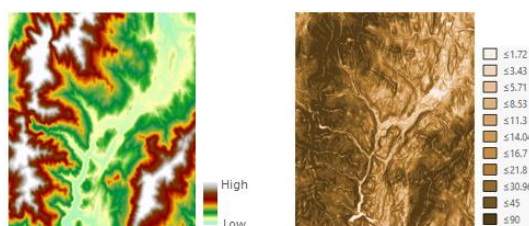


Figura 31 - Dados de entrada para o cálculo do declive. Fonte: ESRI.

O declive apresenta-se como um possível fator de desgaste na rede de distribuição, principalmente quando uma tubagem, ao longo da sua extensão, apresenta uma grande variação de declives. Pode-se considerar que o município de Vila Nova de Gaia é, na grande maioria da sua área, plano. É possível delimitar no município duas áreas distintas relativamente aos valores de declive. Uma primeira área, na faixa litoral do município, até 5km para o interior, no sentido Oeste-Este, onde a grande percentagem da área apresenta declives inferiores a 5°. Existem casos pontuais onde o declive chega a ultrapassar os 15°, no entanto, estes estão relacionados com os rios e ribeiras que aí encontram, e a pedreira da Quinta do Moinho. Esta relação acaba por ser bastante semelhante àquela que se verificou na análise à elevação do Município.

A faixa ao longo do rio Douro, bem como a zona mais oriental do município, são aquelas onde se encontra valores de declive superiores. Grande parte do território coberto pelas freguesias de Lever, Crestuma, Sandim e Olival apresenta declives superiores a 5°. Na encosta do rio Douro, chega-se a registar valores de declive na casa dos 30°. São, no entanto, áreas bastante reduzidas em termos de tamanho.

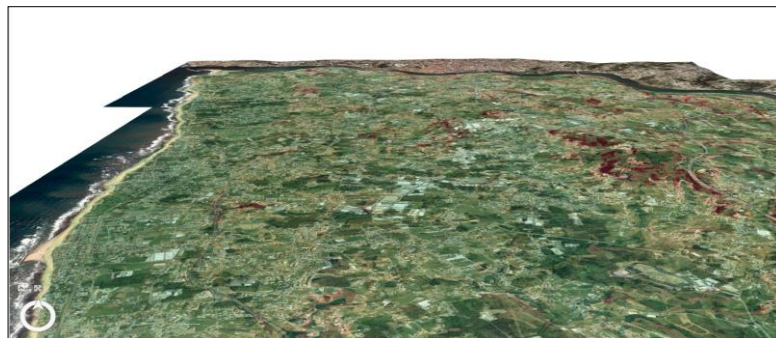


Figura 32 - Exemplo de declive plano na faixa litoral, combinação da carta de declives com ortofotomapa.

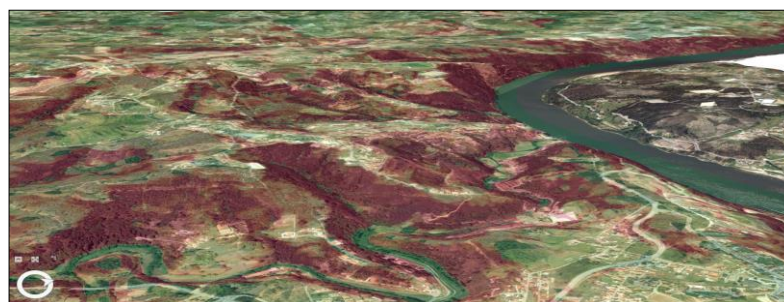


Figura 33 - Exemplos de declives elevado ao longo do rio Douro, nas freguesias do Olival e Crestuma numa combinação da carta de declives com ortofotomapa.

Relativamente à relação entre o declive e as observações relativas ao número de fugas ocorridas, o resultado não é completamente semelhante à análise isolada dos declives. Apesar de haver uma relação negativa, o número de fugas reduz quando aumenta o valor de declive, a sua distribuição não é de todo igual. O valor de declive que regista o maior número de fugas, é de 4°, com cerca de 462 fugas. Sensivelmente 72% das fugas ocorreram a declives iguais ou inferiores a 5°. Cerca de ¼ das fugas ocorreram com um declive superior a 5°, mas inferior a 15°. Apesar de perto de 10% do território ter um declive superior a 15°, não mais que 2% das fugas ocorreram dentro destes valores.

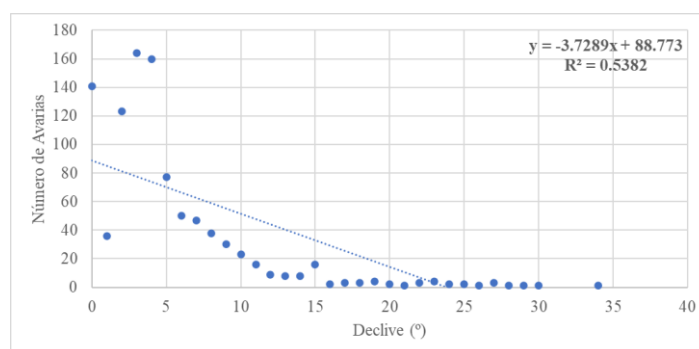


Figura 34 - Relação entre avarias observadas e variável declive.

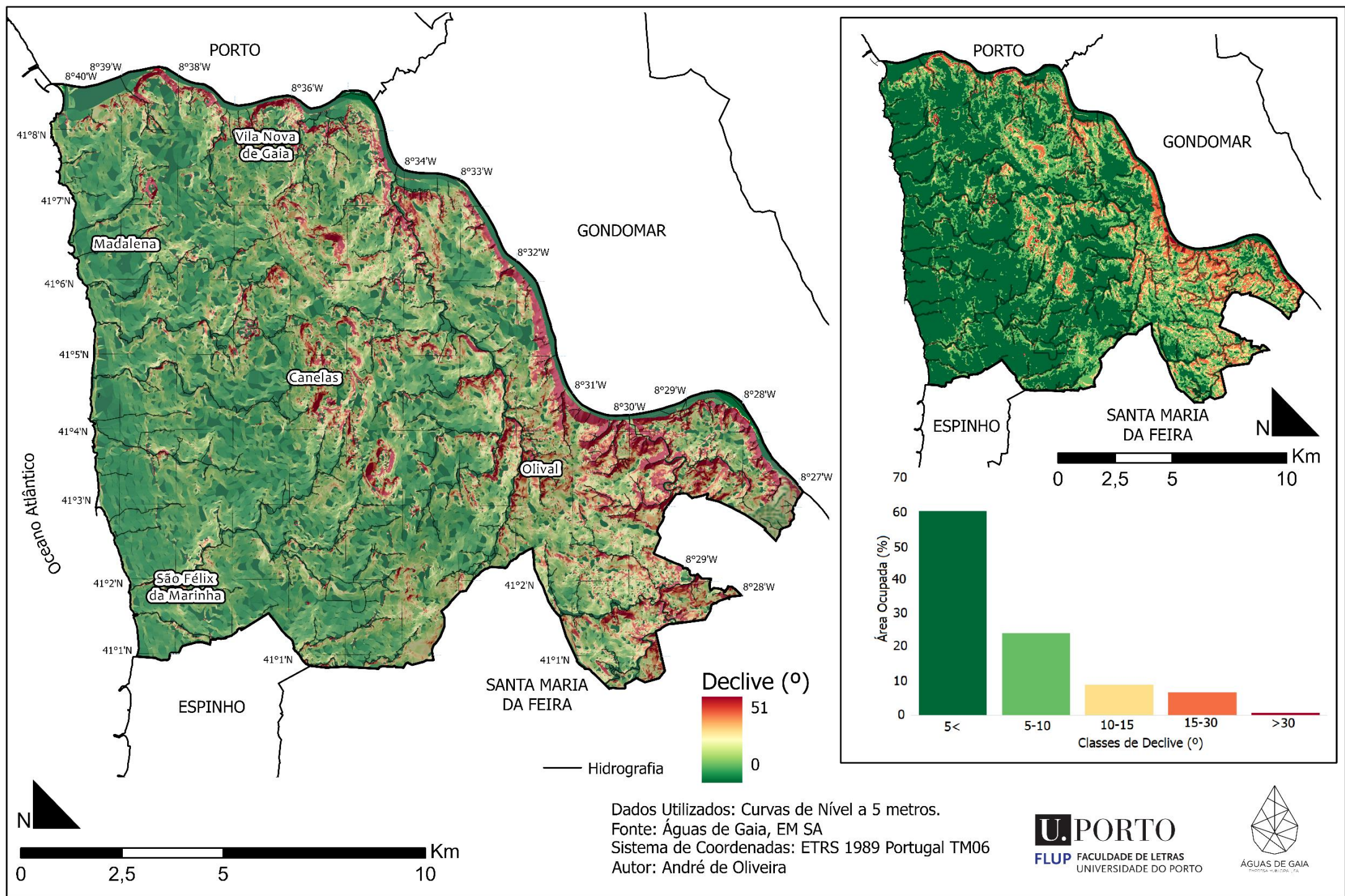


Figura 35 - Carta de Declives de Vila Nova de Gaia.

4.1.3 pH

O pH no solo é identificado pelos especialistas como um forte contribuinte para o desgaste na rede, e consequentemente como agente promotor de fugas. Isto acontece, especialmente, em redes mais antigas em que as tubagens são compostos por materiais ferrosos. Um pH baixo, traduz-se em acidez que corrói mais rapidamente os materiais ferrosos (referência). Numa fase inicial para a recolha de dados deste parâmetro, foi efetuado trabalho de campo para a recolha de pH do solo utilizando um medidor para tal efeito (modelo). Foi desenvolvido um formulário geo-espacial com recurso à aplicação *Survey123 Connect*, bem como uma quadrícula para definir os pontos de recolha. No entanto, foi possível perceber que os dados que foram recolhidos não eram fiáveis e que o método utilizado não estava a ser o mais indicado. Para usar dados mais fidedignos, implicaria maior investimento no equipamento utilizado, bem como a recolha de amostras do solo para análise laboratorial.



Figura 36 - Aplicação Survey123 (a) e recolha de dados do solo com um medidor pH (b).

Para contornar esta contrariedade, foi requerido o acesso a dados, à agência da UE, *European Soil Database & Properties* (ESDAC). A agência, concedeu prontamente diversos dados sobre o solo, dos países da comunidade europeia, entre os quais o pH do solo em Cloreto de cálcio (CaCl_2). O cálculo do pH em Cloreto de cálcio, é o método padronizado para o cálculo do pH do solo (2018). (NSW Agriculture; Acid Soil ACTION, 2000) Tratando-se de dados continentais, foi necessário o tratamento de dados, ou seja, a delimitação pela área de estudo e a projeção para o sistema de coordenadas em estudo. Foi ainda necessário um tratamento adicional nestes dados. Os dados referentes, ao município de Gaia encontravam-se incompletos, sem

informação para várias áreas. Foi então aplicado o cálculo “*focal statistics*”. Este cálculo permite atribuir um valor aos pixels sem informação, sendo feito o cálculo da média, tendo em conta os valores vizinhos. Os dados acabam por perder alguma fiabilidade acumulado também pelo facto da escala dos dados (pixels de 500m).

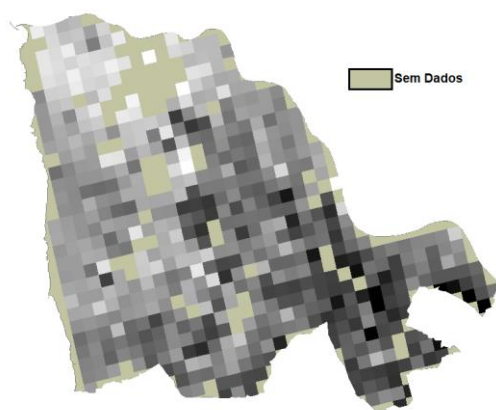


Figura 37 - Pixels sem informação de pH nos dados provenientes da ESDAC.

Observando os resultados obtidos, verificamos que o solo da AE é ácido, variando entre um pH de 4.1 e 5.4. Perto de 50% do solo tem um pH compreendido 4.7 e 4.9. É a norte do município que se encontram solos menos ácidos. Ao relacionar os dados relativos às fugas e ao pH do solo onde se encontram, verifica-se que estes seguem uma relação não-linear. Esta observação acaba por fazer sentido, quando temos em conta que uma percentagem esmagadora da rede, é composta por PVC, um material que não corrosivo.

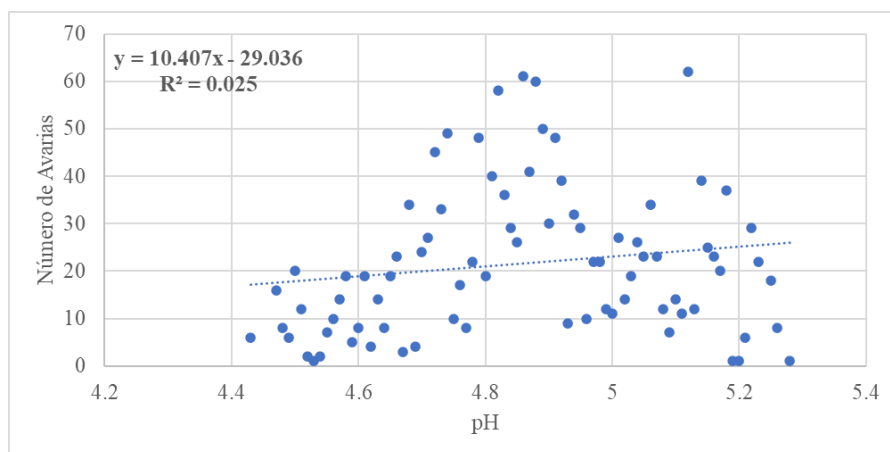


Figura 38 - Relação entre avarias observadas e o pH do solo.

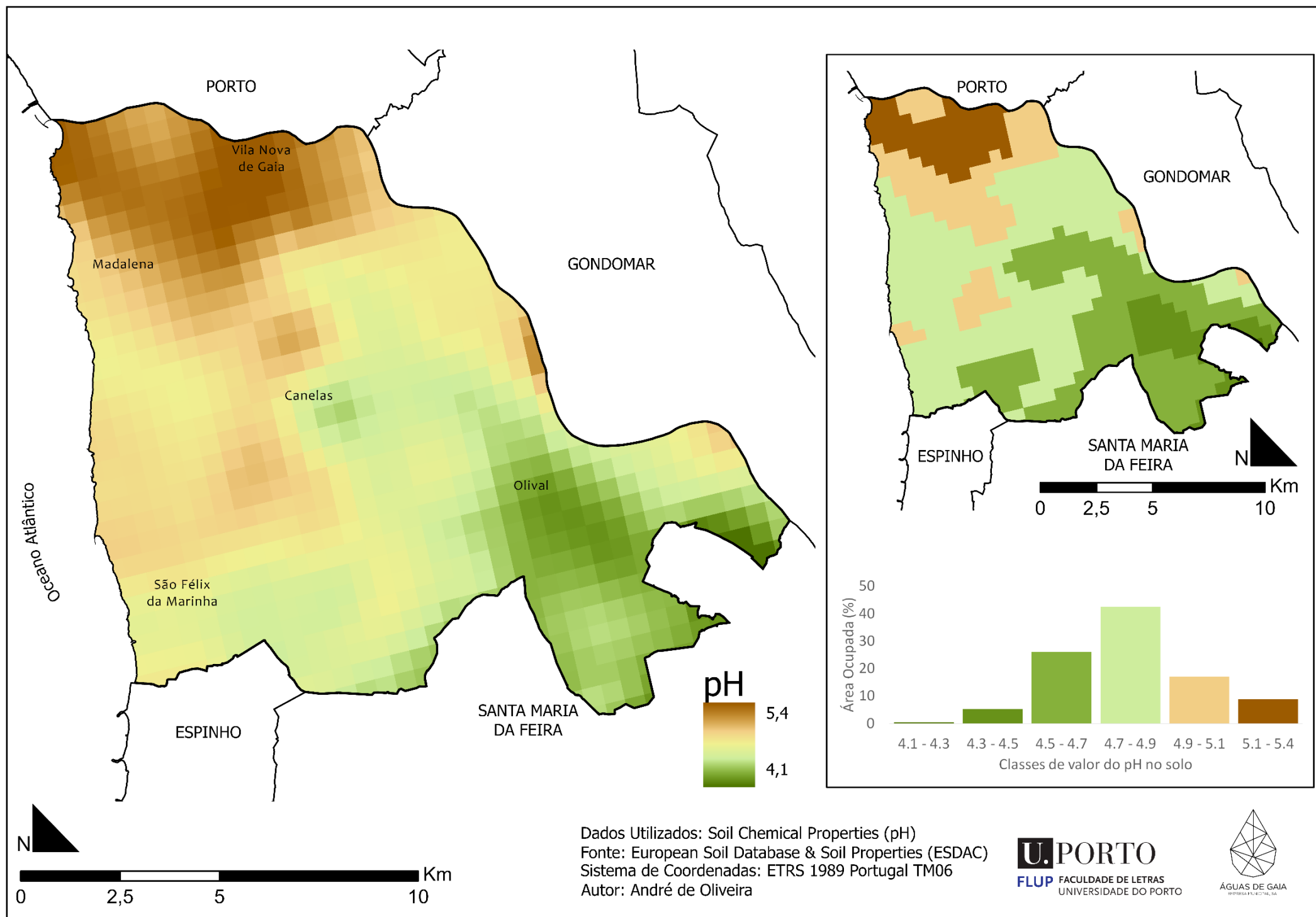


Figura 39 – Variação do pH do solo em Vila Nova de Gaia.

4.1.4 Densidade populacional

Para a análise da densidade populacional, foram recolhidos os dados dos Censos de 2011. A escala de dados utilizada foram as subsecções estatísticas. Considerou-se a escala mais adequada para associar valores de densidade a uma fuga local. Ao trabalhar à escala de freguesias, ou mesmo de secção, os valores da densidade populacional poderiam ser mais baixos do que na verdade se encontram nas subsecções. Os casos mais evidentes, são aqueles nas freguesias com uma componente rural muito forte.

A densidade populacional é frequentemente um fator de desgaste na rede de distribuição. A densidade populacional pode ter dois comportamentos opostos. Por um lado, como já fora analisado no estado de arte, uma densidade mais forte poderá traduzir-se num consumo superior, aumentando a distribuição em direcções diferentes e diminuindo a pressão na rede. No entanto, uma densidade superior implica por norma uma maior circulação de pessoas e viaturas, que como já se verificou, poderá ter impacto no desgaste da rede. É possível verificar que existe uma grande concentração populacional nas freguesias de Santa Marinha e Mafamude (figura 41). A densidade nestas freguesias, chega a ser 14 vezes superior àquelas verificadas nas freguesias com menor população como Sandim, Lever e Crestuma. A densidade populacional é superior nas freguesias a Norte e litoral, diminuindo à medida que nos deslocamos para as freguesias interiores. Áreas com elevada densidade populacional acabam por ter uma grande vantagem para os órgãos que fazem a gestão de redes. A participação da comunidade é um trunfo no sentido em que podem reportar situações de avarias, quando disponibilizadas meios para tal.

O número de fugas ocorridas é superior em territórios de baixa densidade. Cerca de 70% das fugas ocorreram em territórios com uma densidade populacional inferior a 5000 habitantes por km². Apesar de ainda haver um elevado número de fugas entre os 5000 e 10000 hab/km², este número parece estabilizar até densidades na casa dos 30000. O número de fugas em subsecções com densidades populacionais mais elevadas, foi bastante reduzido, Apenas 1% das fugas ocorreram subsecções com mais de 30mil hab/km².

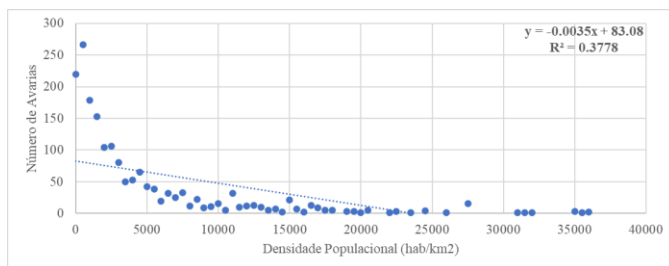
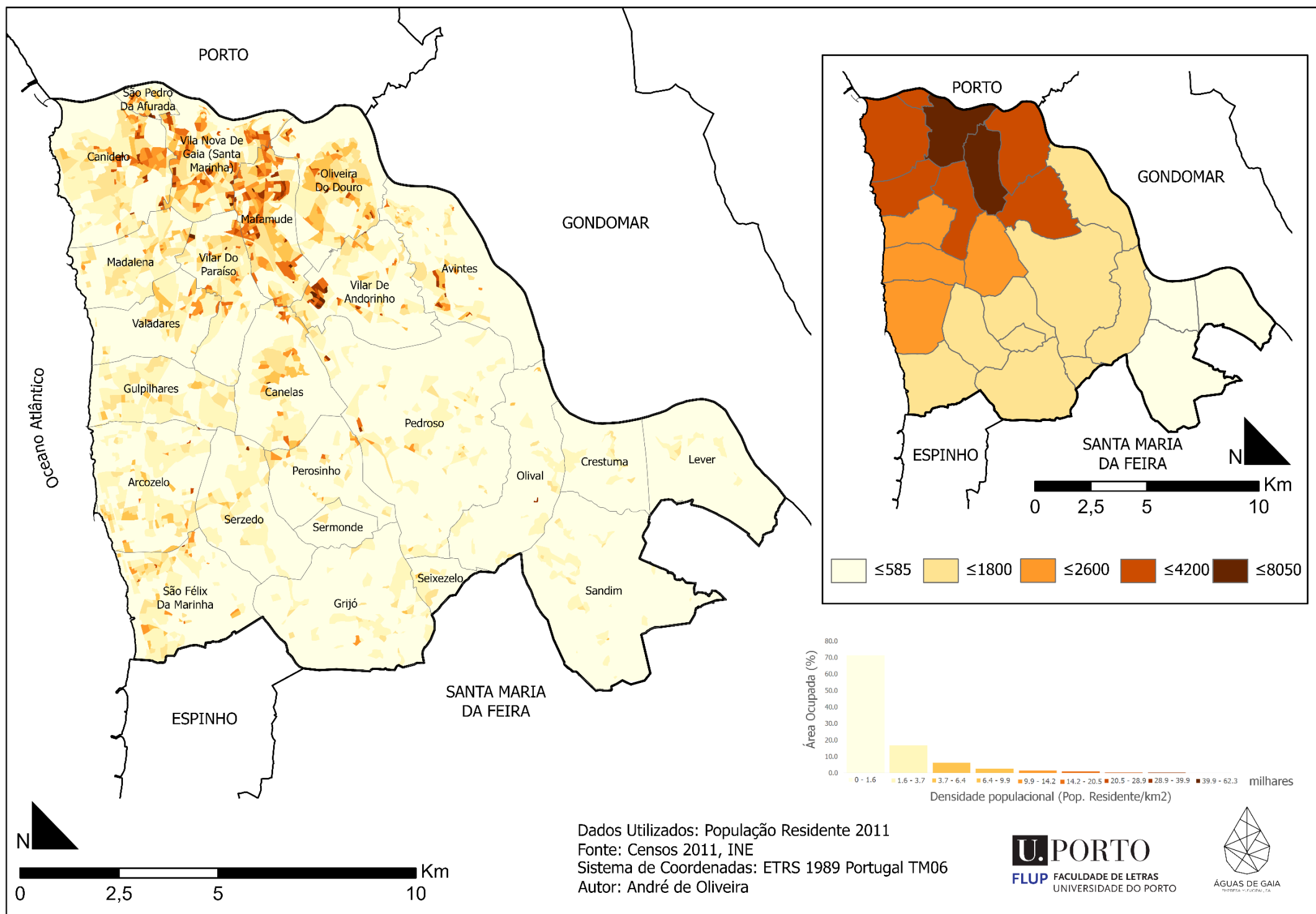


Figura 40 - Relação entre avarias observadas e densidade populacional.



4.1.5 Consumo de água

A obtenção de dados relativos ao consumo de água para o município de VNG, implicou várias consultas numa base de dados SQL. Os dados originais estão associados à faturação mensal para cada cliente. O primeiro passo, foi realizar o somatório do consumo por cliente. De seguida, foi necessário associar cada unidade de instalação a um edifício, o que implica uma relação 1-para-muitos. Foi calculado o somatório de todas as instalações, agrupados por edifício. Tendo estes dados tabulares, foi criado o relacionamento com os dados geométricos. Neste caso, trata-se de dados em formato de ponto. Para o objetivo deste estudo, foi desenvolvida uma grelha hexagonal de 30000 m². Para cada hexágono, foi calculado o somatório de todos os consumos (pontos), e desta forma, foi possível ter o consumo de forma contínua para todo o município e associar cada avaria a esses consumos.

Relativamente ao consumo (Figura 44), é possível verificar que as áreas de elevado consumo, principalmente a norte do município, estando relacionadas com as áreas urbanas do município de VNG. Os consumos mais elevados verificam-se nas freguesias de Canidelo, Santa Marinha, Mafamude e Oliveira do Douro. Para além destas áreas, o consumo de água está também fortemente litoralizado, onde se verificam os maiores consumos, enquanto que os consumos nas áreas interiores do município, como as freguesias de Crestuma, Lever e Sandim, são relativamente baixos. O consumo de água está, logicamente, associado ao número de população residente. Porém, é ainda possível verificar na Figura 44, áreas em que o consumo é elevado, mas estas acabam por estar associados à atividade industrial. A freguesia de Pedroso, tal como anteriormente analisado, apesar de ser a freguesia com maior extensão, acaba por ser das freguesias com menor consumo, e com valores muito inferiores aos verificados nas freguesias a norte. A análise sobre este parâmetro acaba por ter fortes relações com outros, e acaba por ser uma ferramenta interessante para analisar o território.

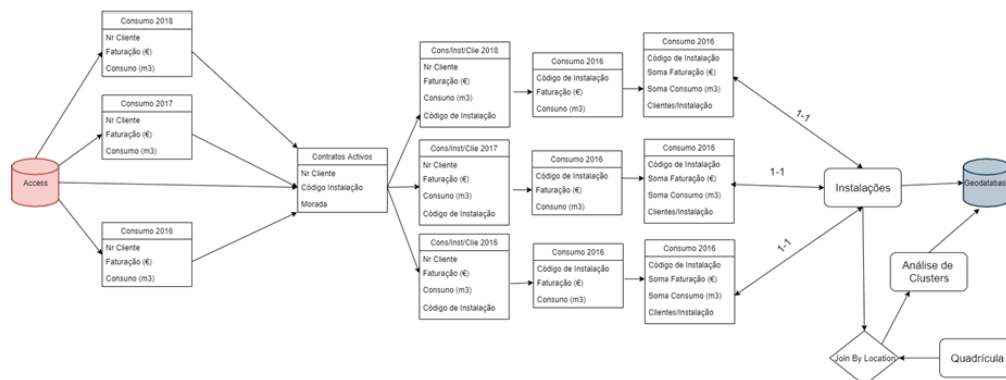


Figura 42 - Esquema metodológico para a espacialização do consumo de água.

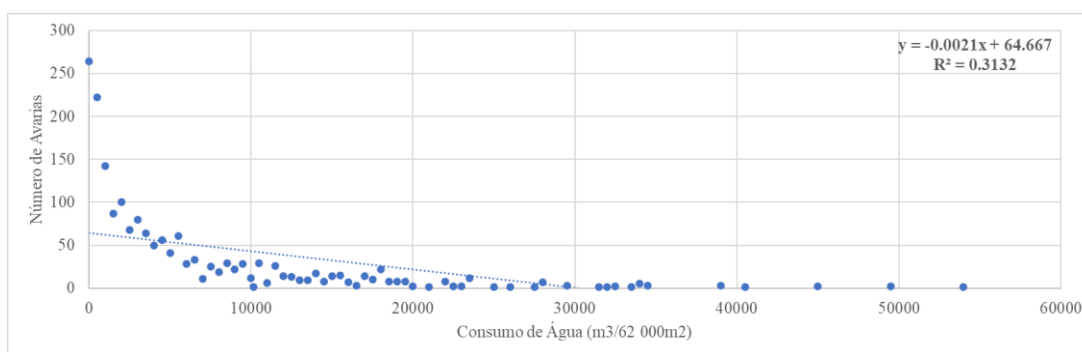


Figura 43 - Relação entre as avarias observadas e volume de água consumido.

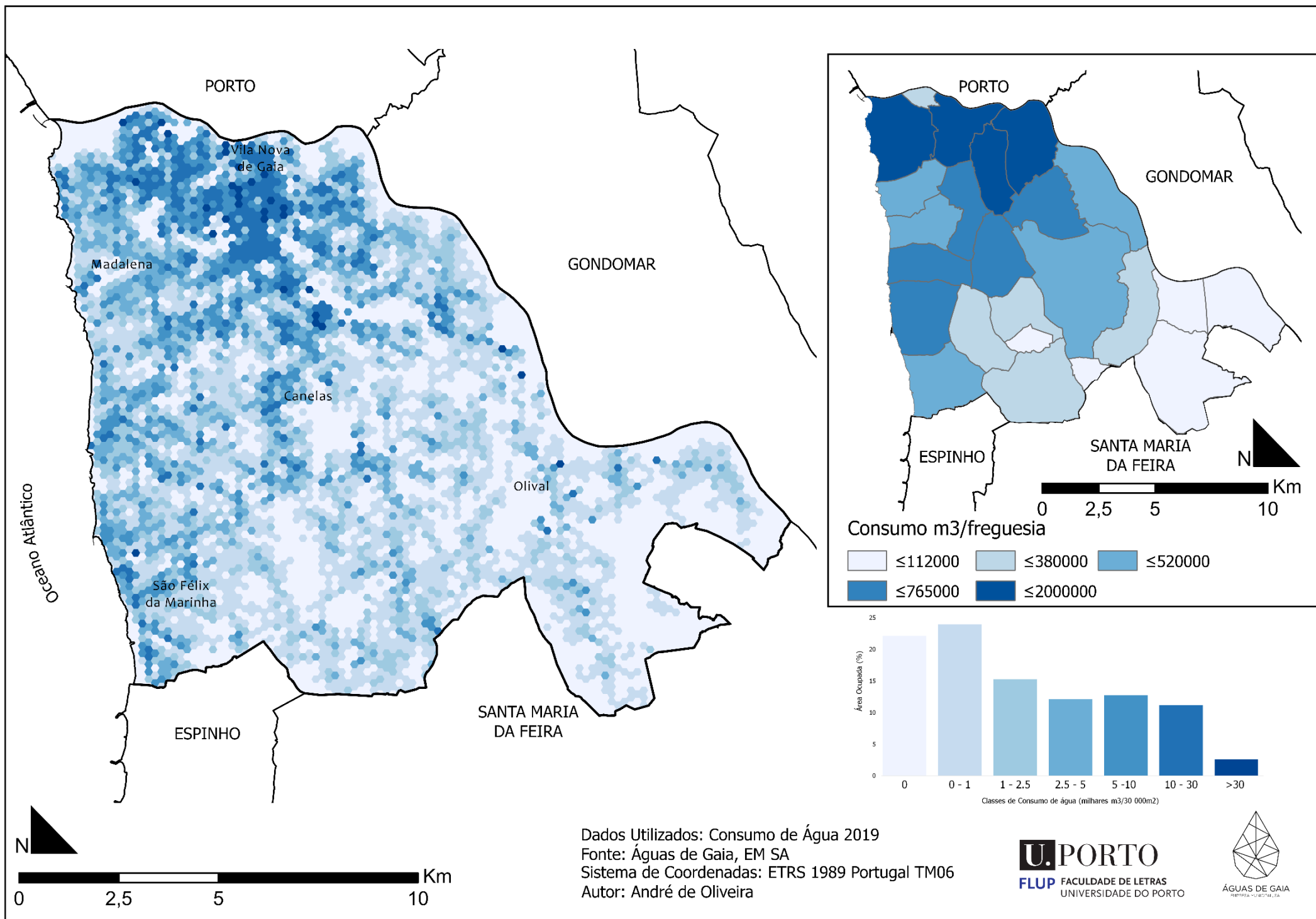


Figura 44 - Representação do volume de água consumido em 2019 em Vila Nova de Gaia.

4.1.6 Ocupação do solo

A forma como o território é utilizado apresenta-se também como um potencial fator para impacto negativo na RDA. O uso do solo acaba também por estar indiretamente relacionado com diferentes variáveis como é o caso da densidade populacional e consumo. Para o tratamento desta variável foi utilizado a Carta de Ocupação do Solo relativa ao ano de 2018, disponibilizada pela Direção Geral do Território (DGT). Os dados relativos ao uso do solo foram trabalhados em duas escalas diferentes. Numa primeira fase, para caracterizar o território foram utilizadas 8 “mega-escalas”:

1. Territórios artificializados;
2. Agricultura;
3. Pastagens;
4. Superfícies agro-florestais;
5. Florestas;
6. Matos;
7. Espaços descobertos ou com pouca vegetação;
8. Massas de águas superficiais.

Pela Figura 46, verifica-se que perto de $\frac{3}{4}$ do território de VNG é Território Artificializado e Florestas. Entrando no detalhe de subclasses, observamos que perto de $\frac{1}{4}$ do território é ocupado por tecido urbano descontínuo. Dentro dos territórios artificializados, as classes seguintes com maior representação são as da rede viária e espaços associados, bem como o tecido urbano contínuo predominantemente vertical, com cerca de 3,5% do território. As florestas de eucalipto é outra subclasse com uma grande ocupação do território, com cerca de 22%. Dentro da classe de florestas, existem ainda as florestas de pinheiro bravo e de outras folhosas, com cerca de 5% e 3% de representação do território. As restantes sub-classes de floresta acabam por não ser muito representativas no município. O solo caracteristicamente agrícola também representa uma forte componente no uso do solo de VNG, com uma ocupação de perto de 20%. São as culturas temporárias de sequeiro e regadio que predominam no solo agrícola.

Freguesias	Territórios Artificializados	Agricultura	Pastagens	Florestas	Matos	Espaços descobertos ou com pouca	Massas de água superficiais
Arcozelo	58.6	16	1.6	16.5	4.3	2.9	0
Avintes	37	21	0	27.1	1.9	0	13.1
Canelas	56.2	7.6	1.2	32	3.1	0	0
Canidelo	65.8	8.5	0	7.6	5.4	5.7	7
Crestuma	17.7	12.6	0	61.7	2	0.4	5.6
Grijó	34.6	22.8	1.4	39.1	2.1	0	0
Gulpilhares	54.9	17	0	15.1	8.3	4.8	0
Lever	16.4	11.8	0.2	59	2	0	10.6
Madalena	58.2	15.1	0.6	15	6.8	4.1	0.1
Mafamude	83.6	3.7	0.5	6.3	6	0	0
Olival	30.4	20.1	0.1	44.6	3.1	0.1	1.6
Oliveiro do Douro	61.2	14	0	11.7	3.2	0.6	9.2
Pedroso	35	22.2	0.6	37.9	4.4	0	0
Perosinho	39.7	20.1	0.3	36.9	3.1	0	0
S. Félix da Marinha	46.3	16.9	6.6	27.3	0.9	2	0
S. Pedro da Afurada	63.1	0	0	12	0	0	24.9
Sandim	25.2	18.9	0.4	54.2	1.4	0	0
Santa Marinha	80.1	3.8	0	4.5	4.7	0	7
Seixezelo	45.1	25.8	0	27.7	1.5	0	0
Sermonde	38.2	29.5	4.3	26.1	2	0	0
Serzedo	38	22	1	37	2	0	0
Valadares	51.5	24.3	0.8	19.6	0.8	3	0
Vilar do Andorinho	45	19.8	0.2	30.7	4.3	0	0
Vilar do Paraíso	55.7	20.1	0.3	22.5	1.5	0	0

Tabela 3 – Taxa de ocupação do solo por freguesia. Fonte: DGT.

A distribuição de fugas ocorridas acaba por ir de encontro a esta análise efetuada. Cerca de 83% das fugas ocorreram em três classes de solo: tecido edificado descontínuo (54% das fugas), tecido edificado contínuo predominantemente vertical (15) e tecido edificado contínuo predominantemente horizontal (13%). Nas florestas, que apesar de ocupar uma grande percentagem do território, o número de fugas aqui identificadas é muito reduzido. Isto acontece pela fraca densidade de rede neste tipo de território. Vale ainda a pena salientar, o número de fugas ocorridas em território associado a indústrias, com cerca de 5% das fugas a ocorrerem nestes espaços.

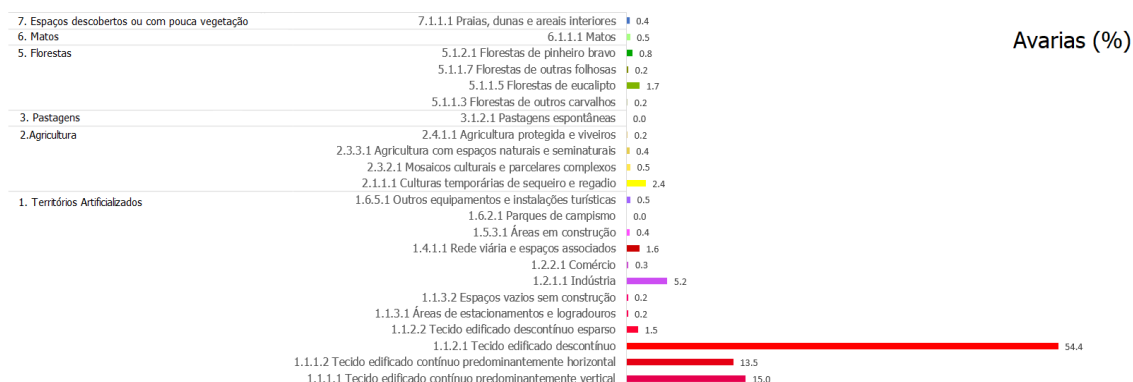


Figura 45 - Percentagem de avarias observadas por classe de ocupação do solo.

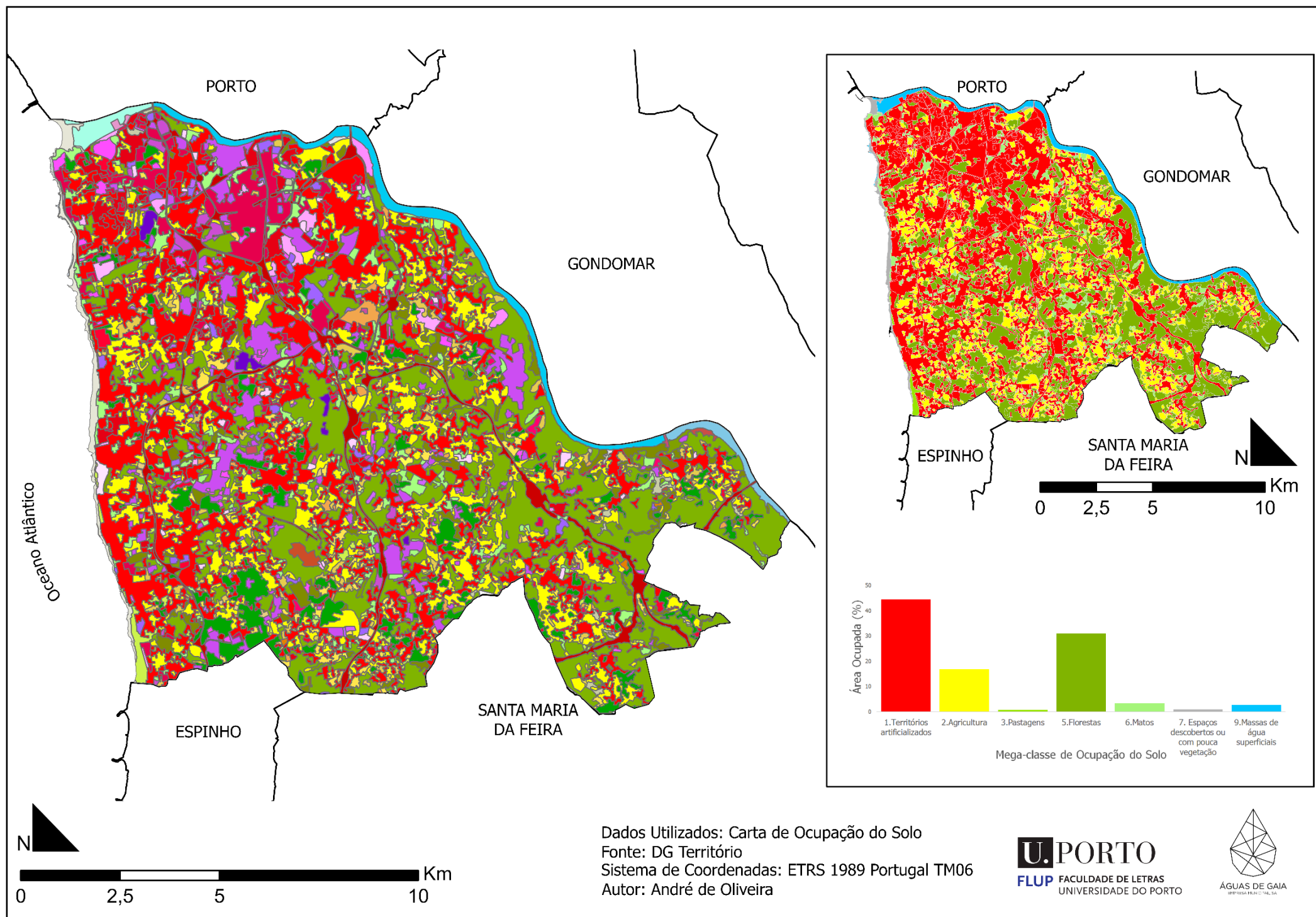


Figura 46 - Carta de Ocupação do Solo de Vila Nova de Gaia.

4.1.7 Geologia

A preparação dos dados referentes ao parâmetro de vulnerabilidade de geologia foi elaborada com o recurso às cartas geológicas de Portugal, disponibilizadas pelo Laboratório Nacional de Energia e Geologia (LNEG). A área de estudo está dividida em três distintas cartas (folhas 09-C, 13-A e 13-B, de 1957, 1962 e 1963, respetivamente), que por sua vez estavam em formato de imagem. Foi necessário proceder à vectorização destas imagens, sendo criado uma *feature data set* com os diferentes domínios (classes existentes). Vale a pena mencionar que algumas classes geológicas não foram incluídas devido à sua representação espacial muito reduzida, tendo sido agregadas com outros e utilizado uma classificação de nível superior. Na *data set*, foram ainda definidas as regras topológicas para assegurar uma vectorização correta e sem erros (Figura 47). As regras utilizadas foram: i) não pode haver sobreposição de polígonos. Isto iria causar erros no momento de cálculo da extensão de cada classe, e poderia haver a sobreposição de classes diferentes, e ii) não pode haver espaços vazios, ou seja, sem classificação geológica.



Figura 47 - Regras topológicas utilizadas para a vectorização das cartas geológicas.

O território em estudo é composto essencialmente (60%) por rocha granítica, havendo, no entanto, várias faixas de rochas metasedimentares. A mais evidente, é a que vai do extremo noroeste do município, e percorre o município, com orientação noroeste-sudeste, terminando no município de Santa Maria da Feira. O litoral do município é caracterizado por aluviões, areias e cascalheiras de praia. Ao longo desta faixa litoral, o seu interior é ocupado por vários depósitos de natureza marinha e fluvial, variando entre os 30 e 130 metros.

Numa análise referente às unidades litológicas e as fugas observadas, é possível perceber que o gráfico resultante é bastante semelhante ao que representa a percentagem de área ocupada por cada unidade. Ainda assim, a proporção de fugas ocorridas em solo de depósitos e formações areno-pelítica de cobertura é consideravelmente superior às áreas ocupadas. Este poderá ser um indício do impacto da litologia possa ter sobre a infraestrutura de RDA. Nesta áreas, dada a sua natureza porosa em que dominam as areias, a acumulação e a permanência de água, de certeza que contribuirá para um maior desgaste da tubagem e respetivos acessórios.

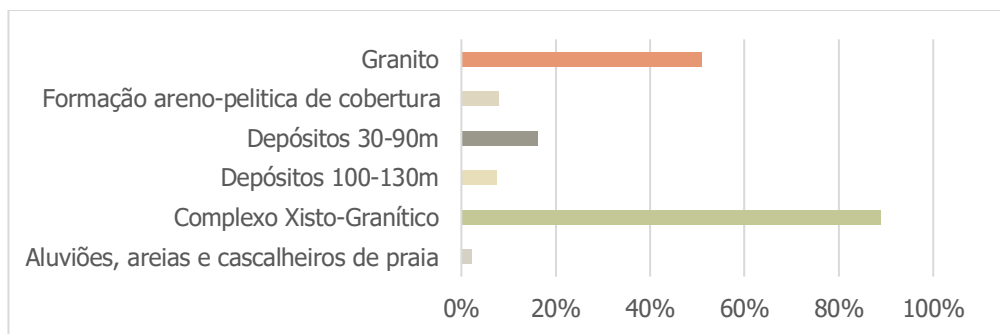


Figura 48 . Percentagem de avarias por unidade geológica.

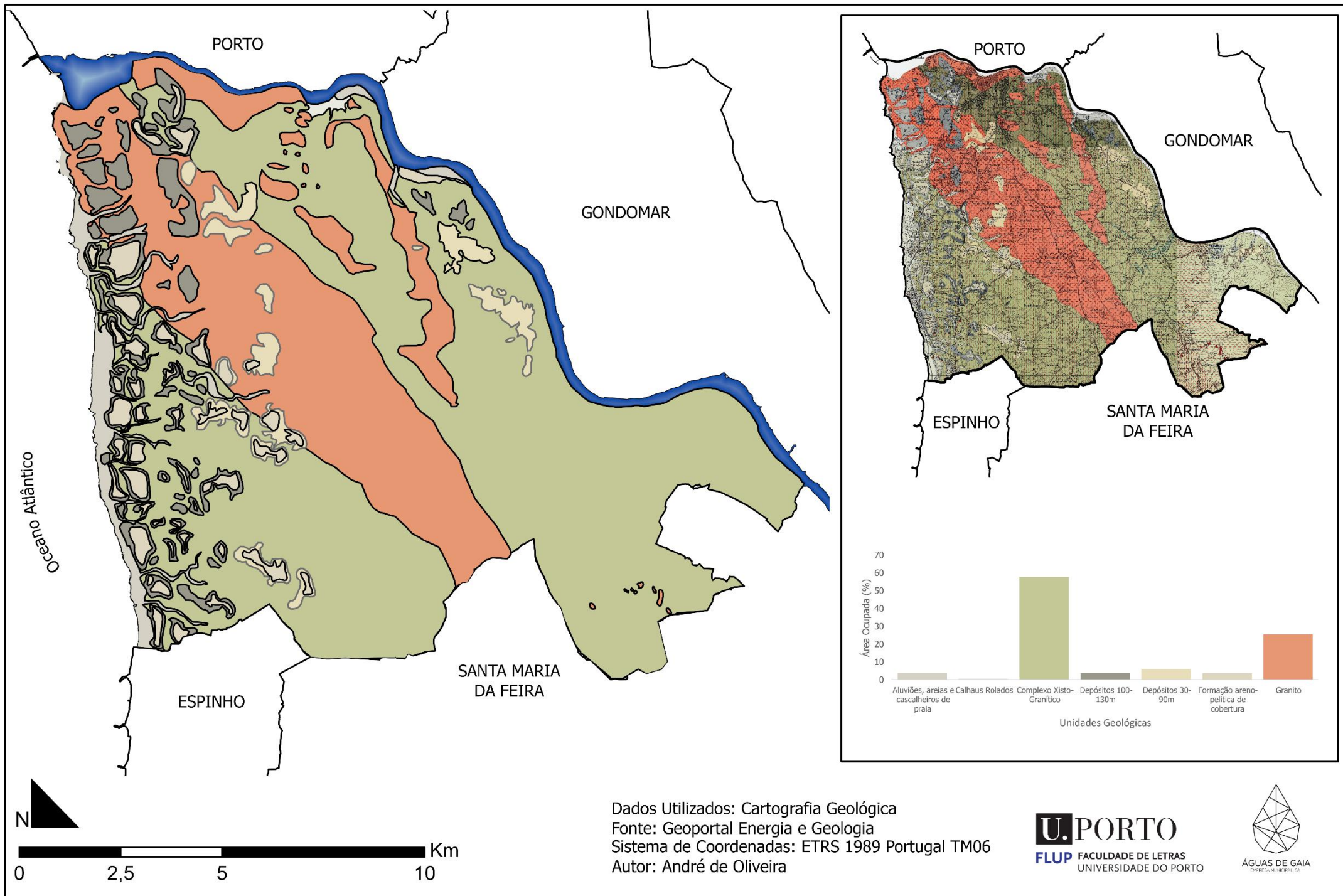


Figura 49 - Carta Geológica de Vila Nova de Gaia.

4.2 Distância a objetos

4.2.1 Rede hidrográfica

Para entender a relação das fugas ocorridas com a rede hidrográfica, optou-se por calcular a distância de qualquer ponto do município à rede hidrográfica. Para isso, utilizou-se o cálculo da distância euclidiana (Figura 50), onde é calculado a distância do centro de um pixel à fonte dos dados (neste caso a *shapefile* da rede hidrográfica), em linha reta. Foram utilizados os corpos de água do município de Vila Nova de Gaia, mas também dos municípios adjacentes, visto a existência de situações em que uma área do município está mais perto de um corpo de água pertencente a um município adjacente.

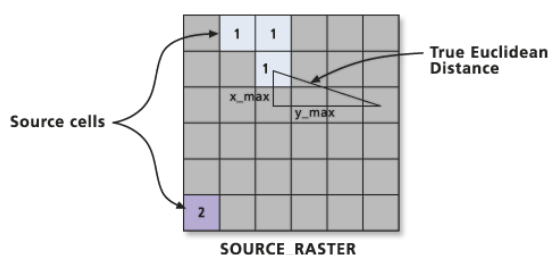


Figura 50 - Cálculo da distância euclidiana. Fonte: ESRI.

A rede hidrográfica do município é bastante densa, mas, apresenta duas realidades distintas. O município é delimitado a Norte-Este pelo rio Douro, o segundo rio mais extenso de Portugal e o rio com maior caudal médio (referência), no entanto, o interior do município é caracterizado por vários rios e ribeiras com caudal e extensão muito reduzidos, quando comparadas com o rio Douro. Acaba por haver, uma grande densidade, quando temos em conta todos os cursos de água. A forte densidade é comprovada pelo cálculo de distâncias de um determinado ponto ao curso de água mais próximo. Nenhum ponto do município se encontra a uma distância superior a 1km, e mesmo a 500 metros, apenas sensivelmente 5% do território se encontra a essa distância ou superior. O cálculo revela que 60% do território se encontra a uma distância compreendida entre 100 e 500 metros, sendo que que $\frac{1}{3}$ situa-se a uma distância de 100-250 metros. Esta proximidade com cursos de água fica evidente, quando verificamos que perto de $\frac{1}{3}$ do território se encontra a menos de 100 metros.

Ao longo dos limites do município, o rio Douro estende-se ao longo de 27km, até ao Oceano Atlântico, onde vai desaguar. Pertencem ainda à bacia hidrográfica do Douro dois dos principais rios do município, o rio Febros e o rio Uima. O rio Uima percorre o município ao longo de 8km, a partir da freguesia de Sandim, até à freguesia de Crestuma, onde se torna um afluente do rio Douro. Já o rio Febros, com uma maior extensão, percorre o município ao longo 16km, desde a freguesia

de Seixezelo (onde está localizada a sua nascente), até à freguesia de Avintes, onde também se torna afluente do Rio Douro. O relevo ao longo destes dois rios é acidentado, correndo ao longo de vales encaixados. No município existe ainda o rio Valadarinhos com uma extensão de 4,7km. A rede hidrográfica no município é completada por várias ribeiras, sendo as ribeiras de Espírito Santo, Francelos e Granja as que possuem maior extensão.

A relação com as fugas ocorridas, é semelhante àquela verificada com a elevação e declive, havendo um elevado número de fugas quando a distância é reduzida, e um decréscimo à medida que a distância aos cursos de água é maior. O número de fugas a mais de 500 metros do curso de água mais próxima ronda os 10% (sendo que o território ocupado por estas características é inferior a 5%). É a 30 metros do corpo de água mais próxima, que se verifica o maior número de fugas, com cerca de 124. A uma distância inferior aos 100 metros, foram identificadas cerca de 537 fugas o que corresponde 21% das fugas. Mais de metade das fugas foram identificadas a uma distância entre 150 e 400 metros.

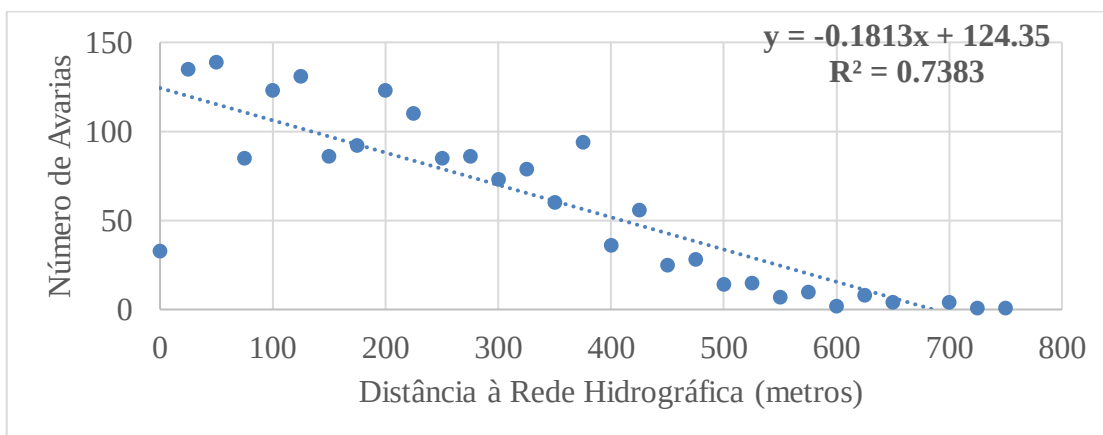


Figura 51 - Relação entre fugas observadas e distância à rede hidrográfica.

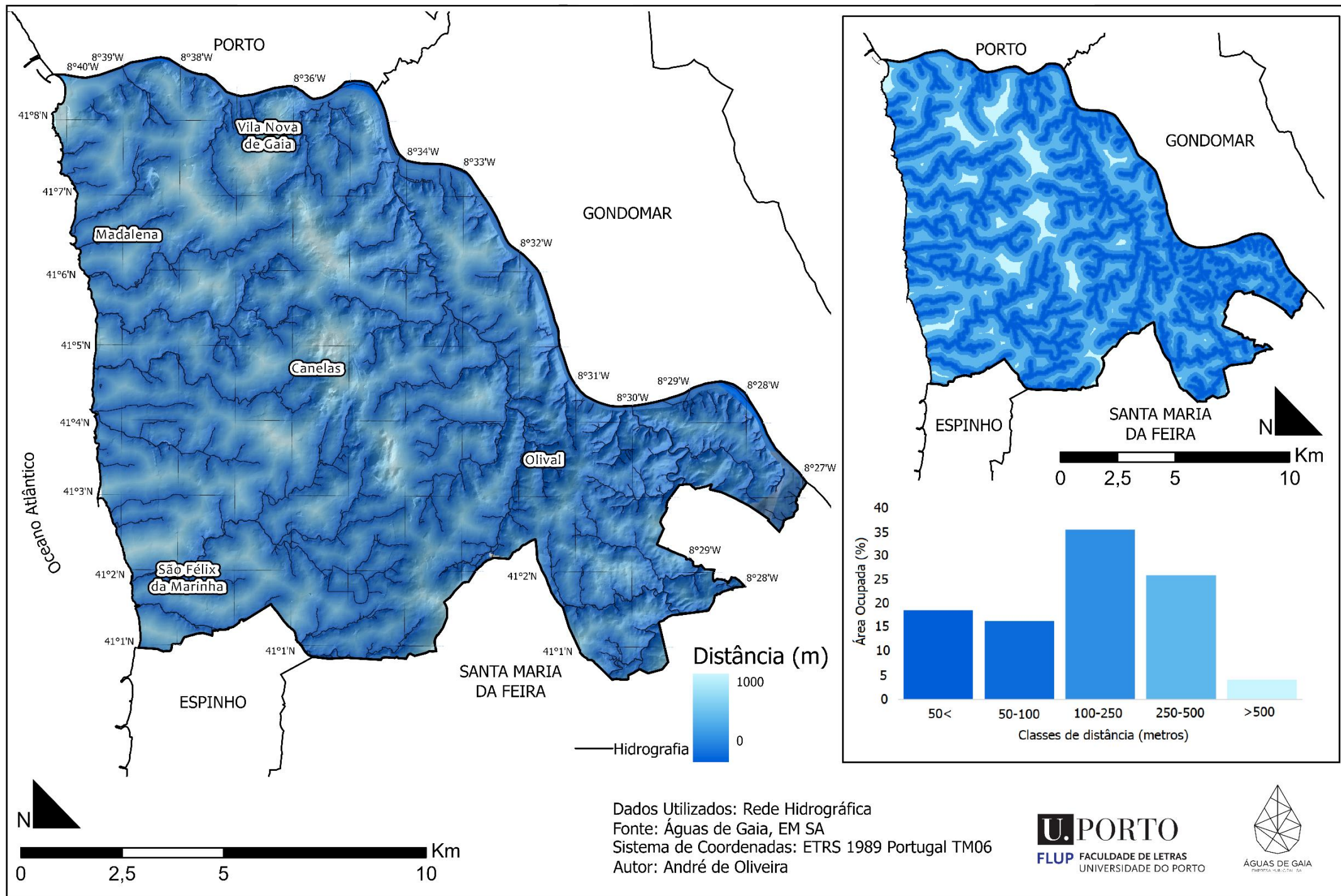


Figura 52 - Carta Hidrológica de Vila Nova de Gaia.

4.2.2 Rede viária

O parâmetro de vulnerabilidade referente à rede viária foi tratado de forma semelhante à distância à rede hidrográfica. No entanto, ao contrário da rede hidrográfica, não foram considerados todos os troços de via existentes no município. Neste caso, utilizou-se apenas aqueles com elevada afluência e onde são praticadas velocidades mais elevadas pelas viaturas, e que resulta em maior vibração no solo e, consequentemente, um fator causador de fuga através do desgaste da rede. Foi então realizado o cálculo da distância euclidiana para toda a cobertura do município, tendo em conta as vias classificadas como Autoestradas, Itinerários Principais e Estradas Nacionais.

É na zona Norte do município, nas freguesias de Santa Marinha e Mafamude que se verifica uma grande densidade de vias com as características referidas. É nesta zona que existe um grande fluxo de viaturas, muito graças ao eixo urbano do município, mas também por ser onde é feita a ligação com o município do Porto. Nestas duas freguesias, é muito baixa a percentagem de áreas que se encontram a mais de 350 metros de uma destas vias. Na zona central do município existe também uma grande proximidade às vias primárias, graças à A1 e à EN1 que se prolongam até à área inicialmente mencionada. A zona Sul do município está dividida por estas duas vias, onde existe duas áreas, diretamente a Oeste e a Este, onde se verificam as maiores distâncias às vias no município. A Sul, junto ao litoral verifica-se de novo uma proximidade às vias, graças à EN109 e A29. Do “lado oposto”, acontece algo semelhante, por sua vez graças às vias EN222, A41 e A32. Fazendo uma análise por classes, utilizando as quebras naturais, verificamos que mais de $\frac{1}{3}$ do território encontra-se a menos de 340 metros da via mais próxima. Perto de 40% do território encontra-se a uma distância superior a 750 metros de uma via primária.

A relação entre a distância à rede viária e as fugas observadas, é uma relação inversamente proporcional. À medida que a distância à rede viária, diminuí, o número de fugas também diminui. Existe uma forte concentração, cerca de 60%, de fugas ocorridas a menos de 500 metros da rede viária. A relação diminui bastante a partir desta distância, onde apenas 20% das fugas ocorreram nos 500 metros seguintes. Vale , no entanto, a pena salientar, o número de fugas ocorridas a uma distância entre os 2000 e 2500 metros da rede viária.

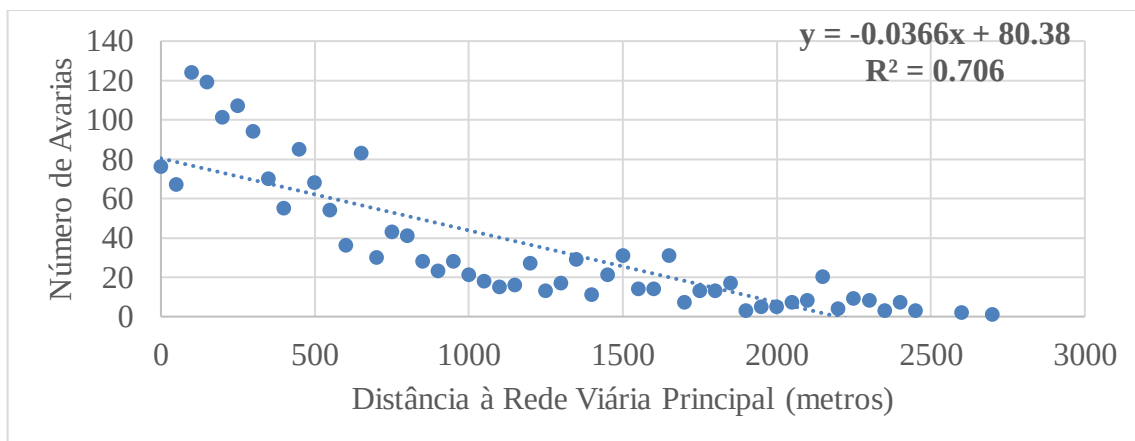


Figura 53 - Relação entre fugas observadas e distância à rede viária.

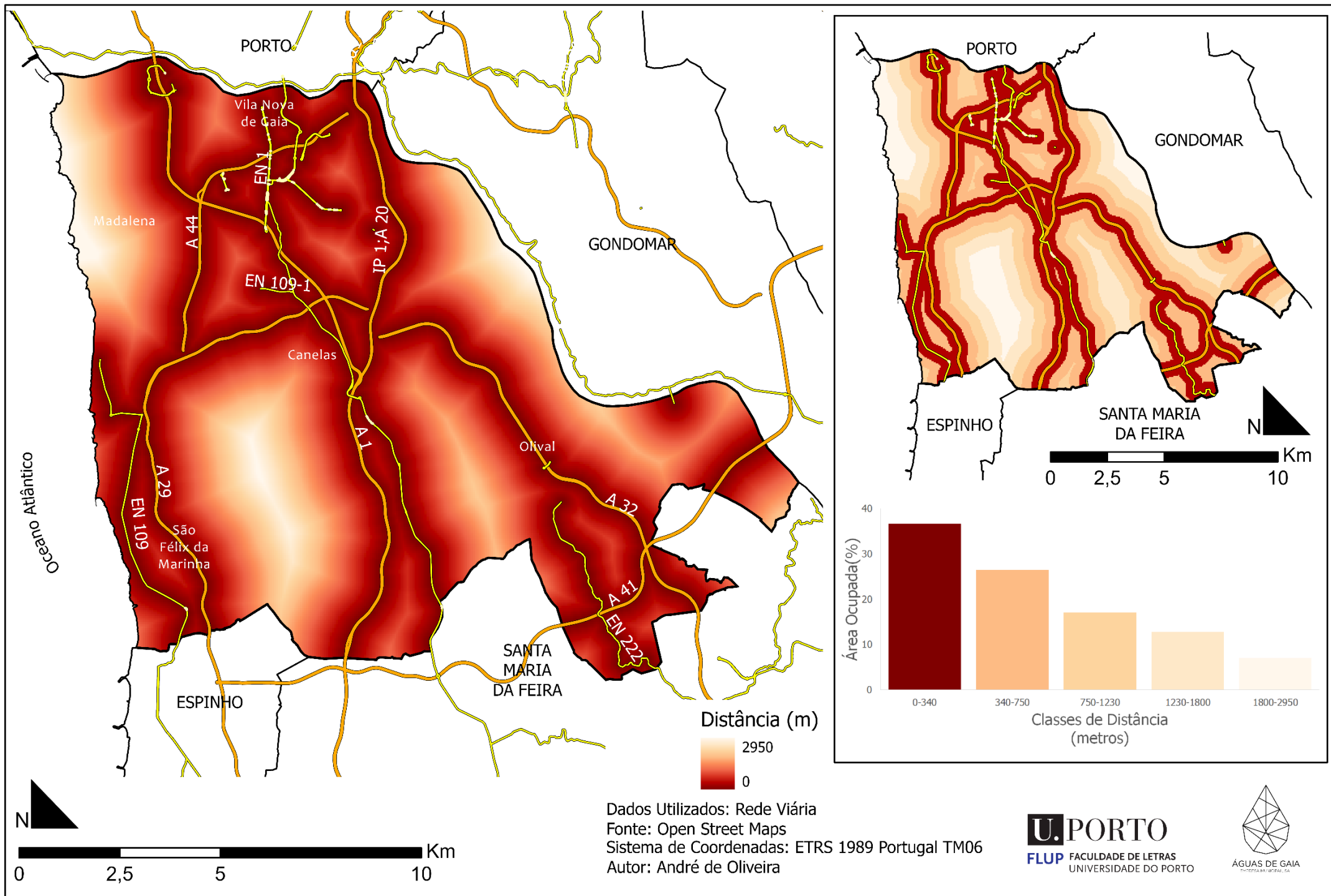


Figura 54 - Rede Viária em Vila Nova de Gaia.

4.2.3 Rede ferroviária

O parâmetro da rede ferroviária foi trabalhado de forma bastante semelhante aos últimos dois. Ao contrário da rede viária, foi considerado toda a rede ferroviária que, em boa verdade, não é extensa no município de Vila Nova de Gaia. Foi efetuado o cálculo da distância euclidiana para todo o município.

Na área de estudo, existe apenas uma linha ferroviária ativa, a linha do Norte, que faz a ligação Porto-Lisboa. A velocidade média dos comboios que circulam nesta linha, no município de Vila Nova de Gaia é de 80km/h. Foi também considerada a linha do Metro do Porto, que no município, faz o seu percurso à superfície. A linha é, no entanto, bastante reduzida aqui, possuindo uma extensão de apenas 3km. A proximidade relativa à rede, segue uma distribuição linear, onde existe uma proximidade bastante próxima (e lógica) nas freguesias de Mafamude, Santa Marinha, Canidelo, Madalena, Valadares, Gulpilhares, Arcozelo e São Félix da Marinha. As restantes freguesias apresentam uma distância cada vez maior, sendo superior nas freguesias interiores.

Efetuando a análise relativamente ao número de fugas, existe uma relação praticamente inexistente. o número de fugas à medida que a distância à ferrovia aumenta, é quase constante, à semelhança do que foi analisado no parágrafo anterior. É ainda importante ressaltar a existência de *outliers*, a distâncias muito reduzidas à linha ferroviária.

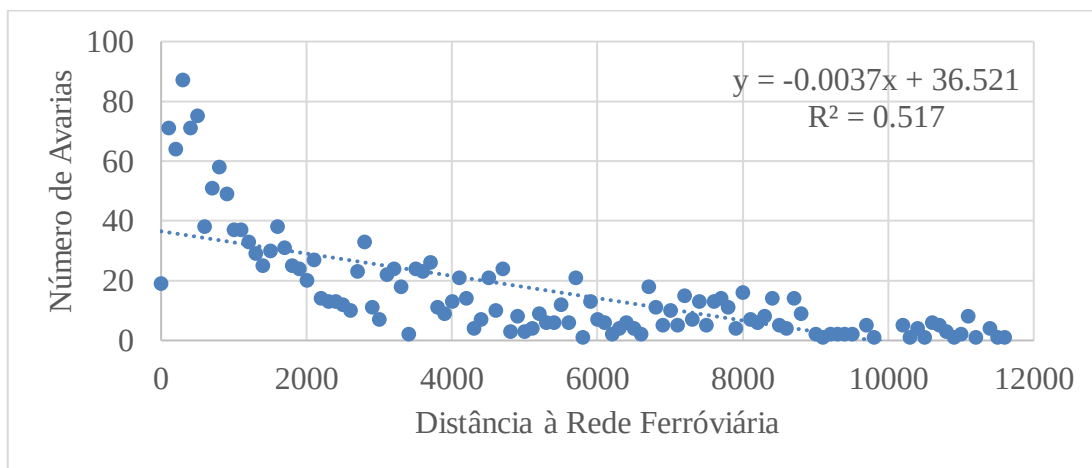


Figura 55 - Relação entre as fugas observadas e distância à rede ferroviária.

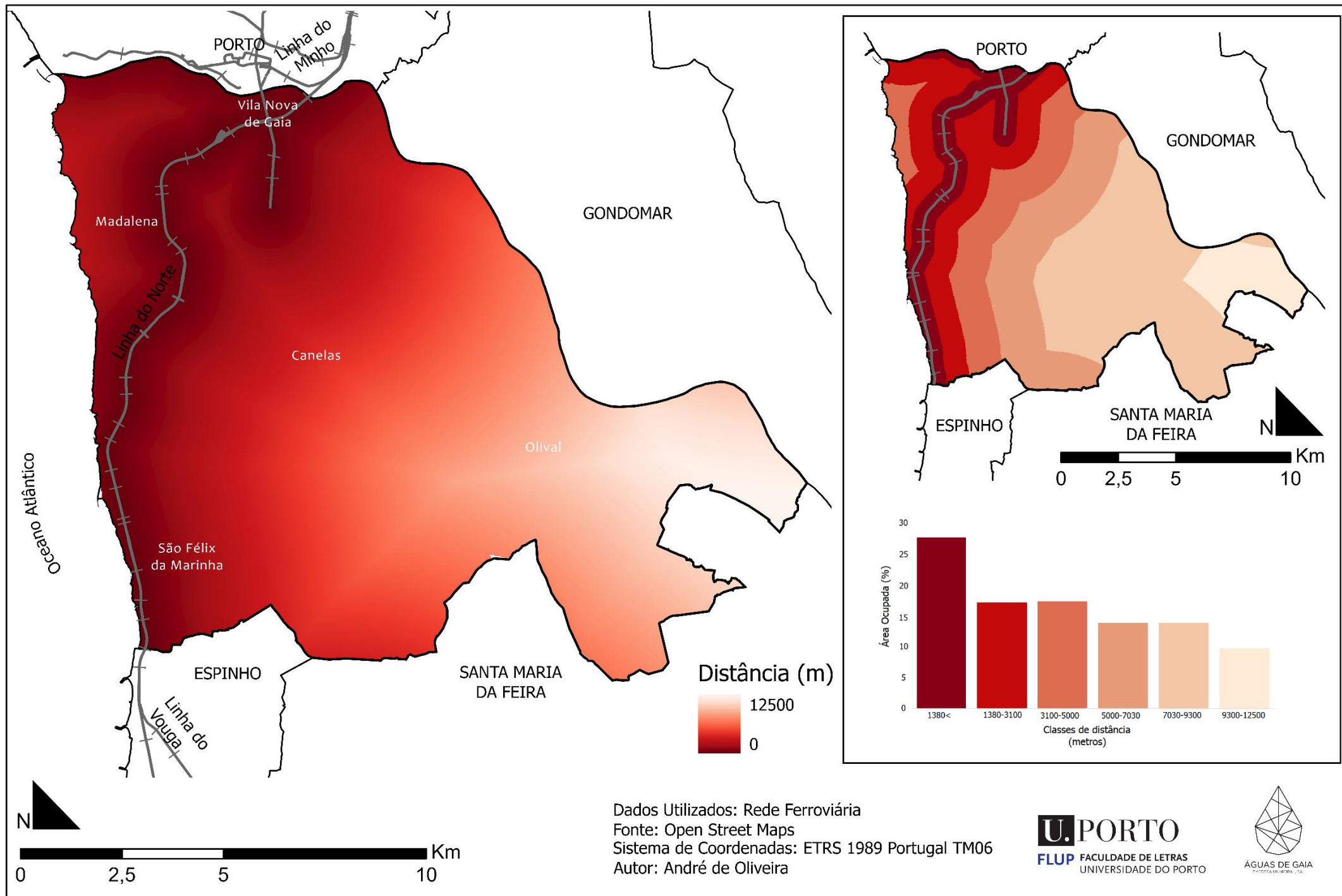


Figura 56 - Rede Ferroviária em Vila Nova de Gaia.

4.3 Matriz de Pearson

Para além da análise sobre a relação dos parâmetros de vulnerabilidade com as avarias, faz todo o sentido fazer uma análise entre os parâmetros. Para isso, foi efetuado o cálculo entre as variáveis, através do coeficiente de correlação de *Pearson* (p), regularmente utilizado em estatística descritiva (University of South Australia, s.d.). Os valores variam entre -1 e 1, onde os extremos revelam correlações perfeitas (negativa e positiva, respetivamente). Um valor igual a 0, expressa que não existe uma correlação entre as variáveis (Kent State University, 2020).

$$p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{cov(X, Y)}{\sqrt{var(X) * var(Y)}}$$

Onde:

$x_1 \dots x_n$ e $y_1 \dots y_n$ são valores medidos de ambas as variáveis e $\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i$;

$\bar{y} = \frac{1}{n} * \sum_{i=1}^n y_i$ são as médias aritméticas de ambas as variáveis.

Na matriz de Pearson resultante () é possível verificar que a grande maioria das variáveis têm correlações muito pouco significativas, sendo até nulas em vários os casos. Este resultado acaba por ser positivo, como já fora observado na revisão literária. Variáveis pouco correlacionadas resultam em menos “ruído” no desenvolvimento do modelo sendo assim mais robusto.

É, no entanto, possível verificar algumas correlações fortes, principalmente entre as variáveis “*dummy*”. Como é o caso do território artificializado e o espaço agrícola ou das áreas de granito ou complexo xisto-granítico. Estas correlações são naturalmente negativas, uma área de território artificializado nunca pode ser também agrícola, ou vice-versa. Esta correlação não é perfeita devido à diferença no número de registos de cada uma.

É ainda possível verificar alguns valores de correlação interessantes. Vale a pena salientar a correlação de 0,5 entre a elevação e a distância à rede ferroviária. Esta correlação acaba por ser lógica, uma vez que a rede ferroviária está localizada em territórios de baixa elevação. Da mesma forma, é a correlação entre a densidade populacional e o consumo de água. Acaba por ser uma correlação linear, onde quanto maior a densidade populacional, maior o consumo de água.

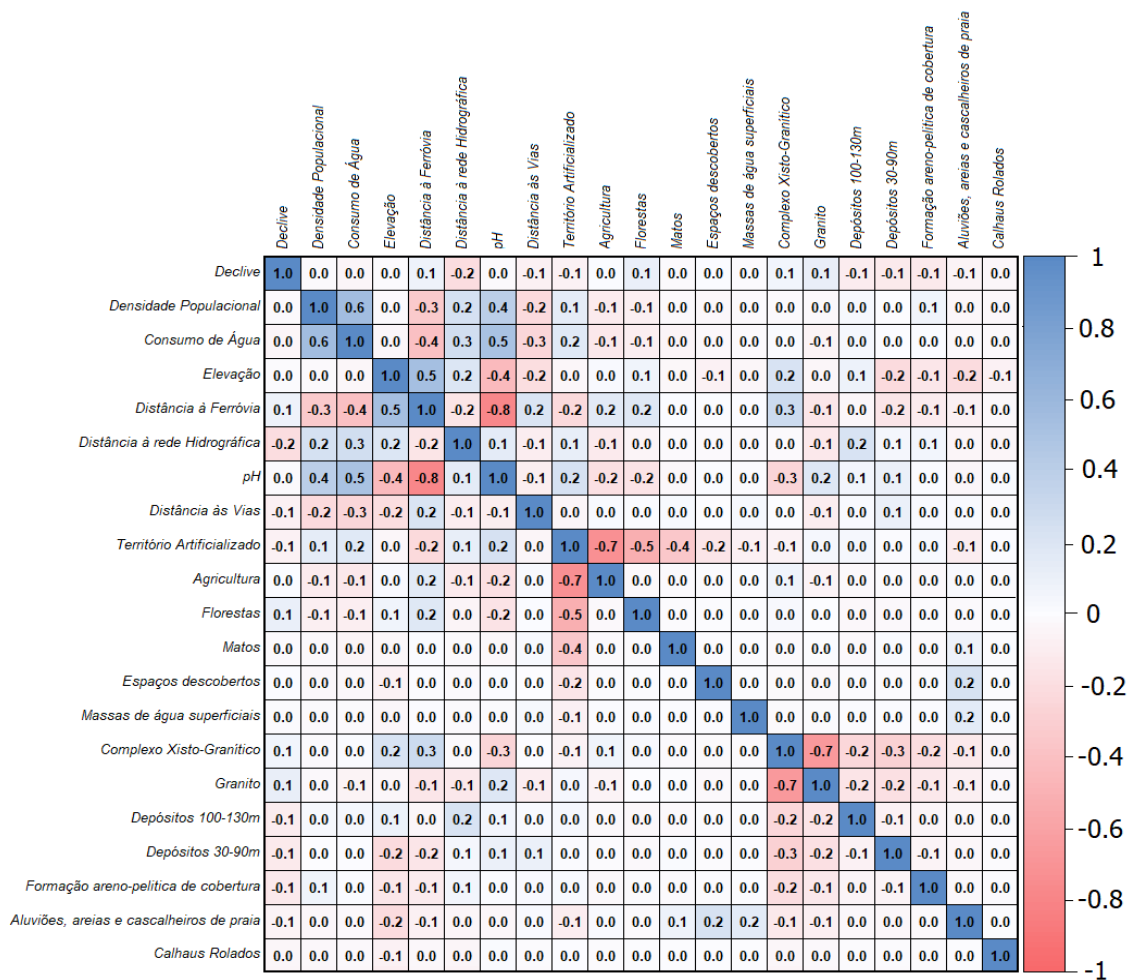


Tabela 4 - Matriz de Pearson para os parâmetros de vulnerabilidade,

CAPÍTULO 5

RESULTADOS

5.1 Avarias observadas – Densidade Kernel

Para iniciar a análise sobre os resultados obtidos, faz todo o sentido fazer uma análise da localização das fugas, antes da aplicação do modelo preditivo. A Figura 57 é o resultado do processo de *geocode* realizado numa primeira fase do desenvolvimento do estudo. Para melhor interpretação dos resultados, foi efetuado o cálculo da densidade Kernel. Para este cálculo, foi considerando o tamanho dos *píxeis* de saída, uma vez que estes influenciam a visualização. *píxeis* com maior área resultam em áreas mais suaves, enquanto que *píxeis* com menor área, apesar de menos suavizados, produzem resultados mais detalhados. Foi essencial encontrar um balanço entre estas duas características, utilizando então um *píxel* de 50m². A densidade Kernel calcula a densidade de *features* na envolvente das próprias *features* (neste caso, as avarias).

$$Densidade = \frac{1}{(raio)^2} \sum_{i=1}^n * [pop_i (1 - \frac{dist_i}{raio})^2]$$

Onde:

i = É os dados input para a ferramenta, neste caso as avarias.

pop = É um valor de peso associado a cada ponto. Neste caso, como cada ponto representa uma avaria, o valor é igual a 1.

pop = É a distância de cada *píxel* às avarias (i).

É possível identificar imediatamente 3 áreas de elevada densidade. Uma primeira, está fortemente relacionada com o centro urbano de VNG (Mafamude e Santa Marinha), seguindo um percurso até à zona ribeirinha norte do município. É nesta área que existe o valor máximo de densidade e é também a área de maior extensão. As outras duas áreas, apesar de mais diminutas, são rapidamente detetadas na cartografia resultante. Nos limites de freguesia de São Félix da Marinha, Arcozelo e Serzedo é possível identificar uma concentração de avarias. Por fim, é nos limites das freguesias de Olival e Pedroso onde é possível identificar várias avarias, numa área relativamente reduzida.

Apesar destas áreas de densidade de avarias, é interessante observar uma grande dispersão de avarias ao longo do município. Apesar do número extensão total de rede nas freguesias de Crestuma e Sandim ser relativamente baixa, ainda é possível encontrar um valor de avarias considerável. Estes dados acabam por coincidir com a literatura previamente estudada, onde locais com poucos pontos de distribuição, acabam por sobrecarregar a rede em termos de pressão o que se traduz em fugas na RDA. As zonas com reduzidas fugas/nenhumas estão associadas a espaços florestais e agrícolas.

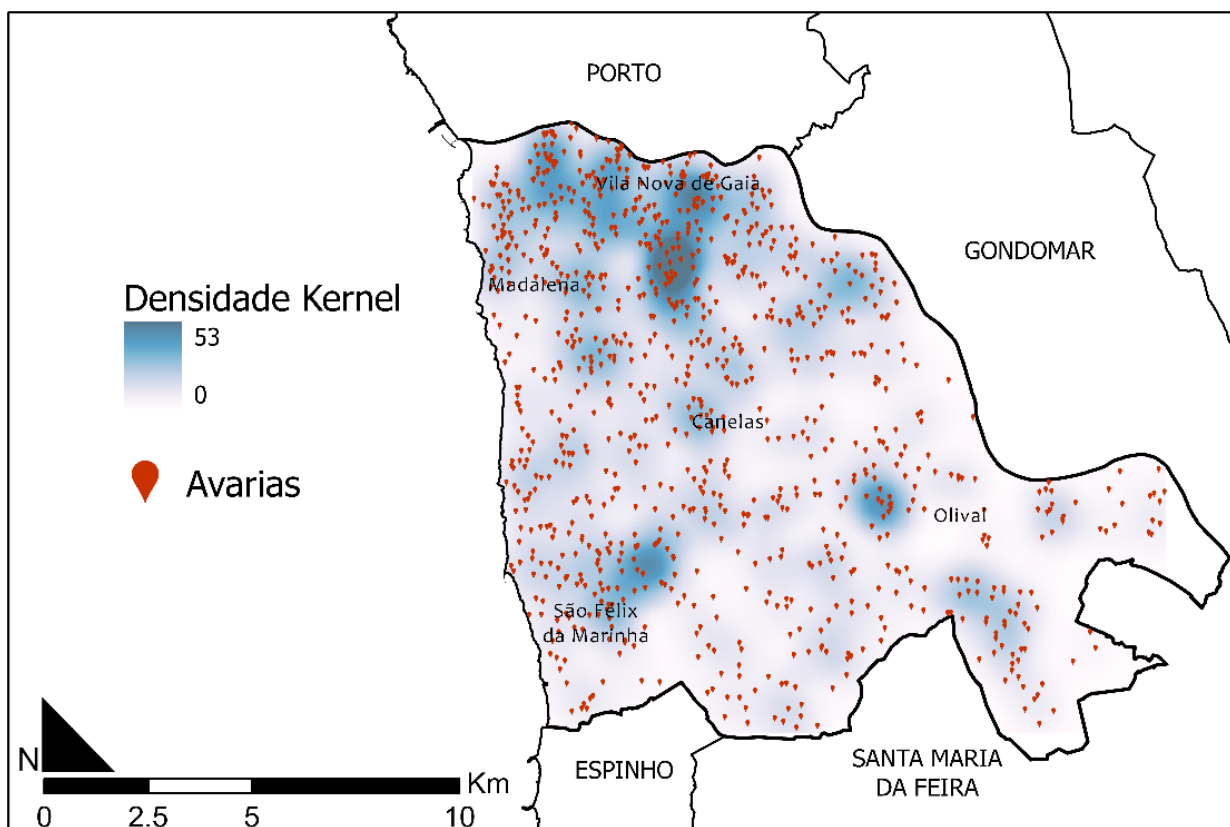


Figura 57 - Avarias detetadas no período 2017-2019 e respetiva densidade Kernel.

5.2 Avarias observadas – Análise *Hotspot*

Com o objetivo de complementar a análise anteriormente efetuada, foi realizado o cálculo de clusters, com recurso à ferramenta “*Hotspot Analysis*”. Para efeitos de análise de resultados, foi desenvolvida uma malha em que cada unidade equivale a 50 hectares. A cada uma dessas unidades, foi associado o número de fugas aí ocorridas. O método utilizado para o parâmetro de distância foi uma distância fixa de 1000km, tendo em luz as extensões das tubagens da rede. Foi ainda realizado o cálculo do índice de *Moran I* para a análise de resultados estatísticos. Este índice analisa a autocorrelação espacial entre um conjunto de dados ("Pearson Correlation," 2020). Neste caso, o valor obtido relativo ao *z – score* (15,13), fornece a informação que os clusters têm menos de 1% de probabilidade de serem gerados aleatoriamente.

É evidente as semelhanças entre o resultado obtido (Figura 58) e a densidade de fugas (Figura 57), especialmente nas áreas com várias fugas ocorridas. Novamente, o centro urbano do município é onde está localizado o maior *hot spot* de fugas. Novamente, é possível identificar de igual forma *hot spots* na zona sudoeste do município, a norte da freguesia de São Félix da Marinha e entre os limites administrativos de Pedroso e Olival. A vantagem da utilização desta ferramenta é que, para além de uma delimitação mais concreta entre as áreas com elevado número de fugas, permite uma melhor interpretação do “espaço vazio”. Enquanto que nos dados relativos à densidade, verificou-se que existia fugas em praticamente toda a extensão do município, a análise de clusters permite interpretar as fugas em espaços onde o número de ocorrências é reduzido, através dos *cold spots*. Os principais *cold spots* identificados estão localizados principalmente nas margens este e sudeste do município, ou seja, nas freguesias de Sandim, Lever e Crestuma. É também interessante verificar a proximidade entre um *cold spot* e um *hot spot* a este e oeste, respetivamente, da freguesia de Olival. É ainda possível verificar mais *cold spots* em casos pontuais no município, sendo ainda relevante salientar a margem ribeirinha na freguesia de Avintes.

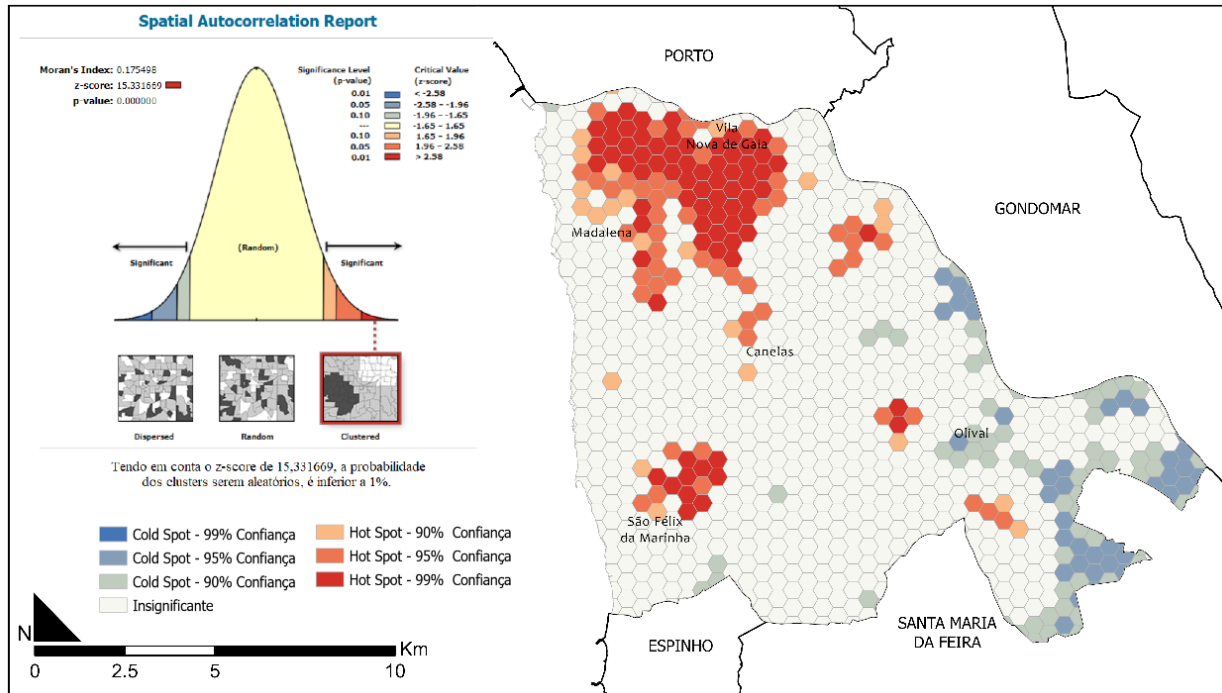


Figura 58 - Análise de Clusters de ocorrência de fugas em 2017 e 2019.

5.3 Previsão de suscetibilidade – *Random Forest Regressor*

O desenvolvimento de um modelo RFR, pressupõe a utilização de dados de treino, e neste caso a delimitação das áreas de treino (AT), onde se conhece os resultados, para prever para dados desconhecidos. Assim, o objetivo será detetar infraestruturas suscetíveis a avarias, com base nos parâmetros disponíveis e utilizados. A grande problemática em estudo, são as fugas não detetadas e isso acarreta a dificuldade de delimitar áreas de treino 100% fiáveis, onde é possível afirmar que todas as fugas ocorridas foram detetadas, especialmente trabalho num período de observações tão curto (2017-2019). Foi por isso necessário a adoção de estratégias capazes de superar esta adversidade.

As delimitações das áreas de treino tiveram muito em conta os resultados obtidos na Figura 57 e Figura 58. O princípio utilizado foi desenvolver o modelo com dados capazes de caracterizar o município com todas as particularidades associadas ao longo da sua extensão. Isto teve em conta, para além das fugas identificadas, mas também os diferentes parâmetros analisados no capítulo anterior. As AT definidas contabilizam cerca de 10% das tubagens do total do município. Os resultados produzidos pelo modelo vão sempre depender dos dados de treino e dos ajustes feitos. Ao longo das inúmeras adaptações efetuadas no modelo, serão apresentados três modelos resultantes, que acabam por se multiplicar. Como já fora referido, o desenvolvimento deste estudo

está de certa forma débil, uma vez que não existe informação relativa à idade da tubagem. É por isso impossível prever uma data para a ocorrência de uma avaria. Estes diferentes modelos conferem diferentes cenários, que poderão avaliar o risco de avarias no presente, num futuro próximo ou avarias num futuro distante. Os resultados produzidos pelos modelos podem ser facilmente manipulados pelos dados de entrada, sendo essencial que se realizem várias adaptações e comparações para que o modelo seja robusto e capaz de produzir resultados fiáveis e úteis para os órgãos de tomada de decisão. A tabela abaixo demonstra as diferenças nos dados de entrada nos diferentes modelos desenvolvidos.

	Áreas de Treino	Totalidade dos Parâmetros	Seleção de Parâmetros	+10% de Avarias fora da AT	+20% de Avarias fora da AT	+100% de Avarias fora da AT
Modelo A	✓	✓	×	×	×	×
Modelo B	✓	×	✓	×	×	×
Modelo C	✓	✓	×	✓	×	×
Modelo D	✓	×	✓	✓	×	×
Modelo E	✓	✓	×	×	✓	×
Modelo F	✓	×	✓	×	✓	×
Modelo G	✓	✓	×	×	×	✓
Modelo H	✓	×	✓	×	×	✓

Tabela 5 - Especificidade sobre os principais modelos RFR desenvolvidos.

Apesar de todos os modelos terem o mesmo núcleo de dados, dados das AT, eles acabam por ter características que os distinguem. Tal como referido no parágrafo anterior, existem quatro principais modelos, que estão diretamente relacionados com os dados de entrada. Dentro destes modelos é ainda feita a distinção pelos parâmetros utilizados. Enquanto que numa primeira versão são utilizados todos os parâmetros, na segunda versão são removidos os parâmetros identificados como baixa valia. Apesar de em alguns modelos se verificar melhorias na performance do modelo, isso não se verificou em todos os casos. Tal como referido, a diferença entre as principais versões dos modelos, prende-se com os dados de entrada. Nos primeiros modelos são utilizados apenas os dados das áreas de treino, enquanto que nos seguintes são acrescentados dados relativos a tubagem onde se observou fugas. Estas foram escolhidas aleatoriamente e acrescentadas aos modelos. Nas primeiras versões, foram acrescentadas 10% das fugas ocorridas, para além daquelas dentro das áreas de treino, na seguinte foram adicionadas 20%, enquanto que na versão

final, foram utilizadas 100% das tubagens onde se verificou fugas.

Na biblioteca *sklearn*, onde foram desenvolvidos os modelos, o algoritmo “*Out of bag Score*” permite avaliar a performance do modelo (Li et al., 2007). Foi efetuado este cálculo para cada um dos modelos desenvolvidos, tanto no momento inicial como no final. Após os resultados iniciais, os parâmetros são ajustados e testados várias vezes, até se atingir um resultado satisfatório. Em todos os modelos, verificou-se uma melhoria na performance após o ajuste dos parâmetros, sendo nos Modelos E e F que se verificaram as melhorias mais significativas. O resultado deste algoritmo acaba por estar com uma forte relação com o número de avarias adicionadas nos dados de entrada uma vez que nestes modelos a performance melhorou significativamente.

Tal como referido, os Modelos A e B, C e D, ..., são semelhantes sendo que a principal diferença está na remoção de variáveis pouco explicativas. No entanto, não se verificou melhorias em metade dos modelos, sendo que as melhorias na outra metade não são muito significativas.

Modelos	Performance Inicial	Performance Final
Modelo A	0.59	0.62
Modelo B	0.60	0.62
Modelo C	0.70	0.76
Modelo D	0.72	0.75
Modelo E	0.77	0.83
Modelo F	0.77	0.85
Modelo G	0.93	0.94
Modelo H	0.93	0.94

Tabela 6 -Performance inicial e final dos modelos RFR.

Ao comparar os resultados obtidos pelos diferentes modelos (Figura 59), verificamos que, apesar de alguma semelhança em determinadas áreas, os resultados são visivelmente diferentes. Outro padrão que é possível observar, é que à medida que o número de avarias adicionada ao modelo aumenta, a previsão de suscetibilidade também aumenta consideravelmente. Isto faz sentido, tendo em conta a proporção entre o total de *features* utilizadas e *features* com avarias. Os resultados obtidos acabam por ser altamente sobrevalorizados. Apesar de uma elevada performance do modelo, este acaba por não ser muito útil para a tomada de decisão. Não existe um padrão nem é possível identificar áreas com elevada suscetibilidade, e por isso, os modelos G e H não são úteis. Do lado oposto do espectro, os modelos A e B, tendo apenas as várias observadas

nas áreas de treino, produz resultados sobrevalorizados e incapaz de identificar tubagens, e consequentemente áreas, onde a suscetibilidade é elevada. Segundo estes modelos, perto de 80% das tubagens têm uma suscetibilidade igual ou inferior a 0.2. São resultados que dificilmente representam a realidade e que possuem as piores avaliações de performance. Por estes motivos, também não vale a pena aprofundar a análise sobre o resultado destes dois modelos.

Os modelos que produziram os resultados mais interessantes foram os modelos C, D, E e F, ou seja, os modelos com dados das áreas de treino, mais 10%/20% das avarias ocorridas fora destas áreas. Existe uma clara distinção entre tubagens com diferentes níveis de suscetibilidade sendo possível delimitar áreas onde existe uma concentração de tubagens elevada suscetibilidade. Existe, no entanto, ainda diferenças entre os modelos C/D e E/F. O primeiro conjunto acaba por ter uma maior percentagem de *features* nas classes com valores mais baixos, sendo que perto de 60% se encontra abaixo dos 0.5, enquanto que no segundo conjunto estes valores aproximam-se mais dos 50%. Esta diferença é “compensada” principalmente nas classes onde o valor de suscetibilidade encontra-se entre 0.6 e 0.7. Enquanto nos Modelos C e D estas classes representam sensivelmente 10%, nos modelos E e F estes rodam $\frac{1}{4}$ do total.

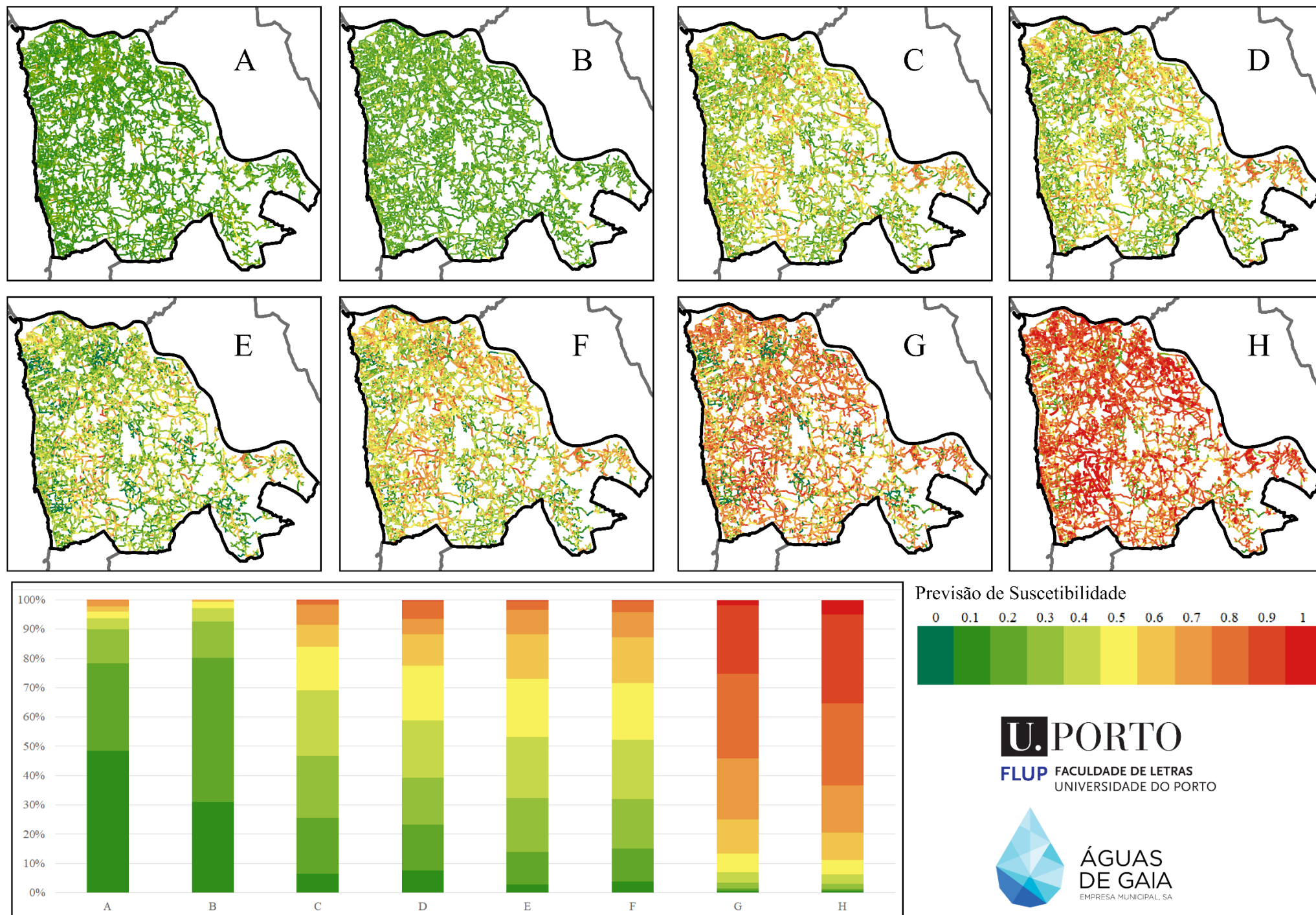


Figura 59 - Resultados produzidos pelos diferentes modelos RFR.

O modelo escolhido para uma análise mais profunda e, como sugestão de aplicação, foi o modelo F, aquele desenvolvido com os dados de entrada das áreas de estudo acrescidos 20% das fugas observadas, selecionadas aleatoriamente. Este modelo para além de ter tido uma performance satisfatória, possui uma distribuição ideal pelas diferentes classes de suscetibilidade. Para uma melhor interpretação dos resultados, os mesmos foram representados numa quadrícula de 25 hectares. Voltou-se a utilizar o conceito inicial de converter as tubagens em pontos, no seu ponto central, para associar os valores à quadrícula. O que cada célula representa, é o somatório do número de tubagens com um valor de suscetibilidade previsto seja superior a 0.7. Esta forma de representar os resultados é uma técnica para delimitar áreas mais urgentes para vistorias e manutenção das infraestruturas.

Assim, é no sector oriental do município, nas freguesias de Crestuma e Lever, que se verificam a maior acumulação de tubagens com elevada suscetibilidade, sendo identificadas 26 tubagens numa única célula. Esta área é aquela que desperta mais atenção por este motivo, mas também porque é uma das áreas identificada em todos os modelos como de elevada suscetibilidade, mesmo nos modelos com valores altamente sobrevalorizados. Curiosamente, esta área foi designada como *cold spot* no momento de analisar a distribuição das fugas ocorridas. Este poderá ser um indício de uma área com várias fugas por identificar e que estejam a produzir elevadas perdas de água e monetárias, diariamente. Ao contrário desta, a faixa norte, a partir da freguesia de Avintes, revela-se também uma área de interesse onde possui uma malha regular com várias tubagens identificadas como suscetíveis, no entanto, foi uma área com forte incidência nas fugas ocorridas no período de estudo. Da mesma forma, a área dos limites administrativos, a oeste, da freguesia de Serzedo acaba por também estar identificada com várias tubagens prevista elevada suscetibilidade. A cobertura desta área acaba por ter, no entanto, uma extensão superior àquela observada na análise de *hotspots*.

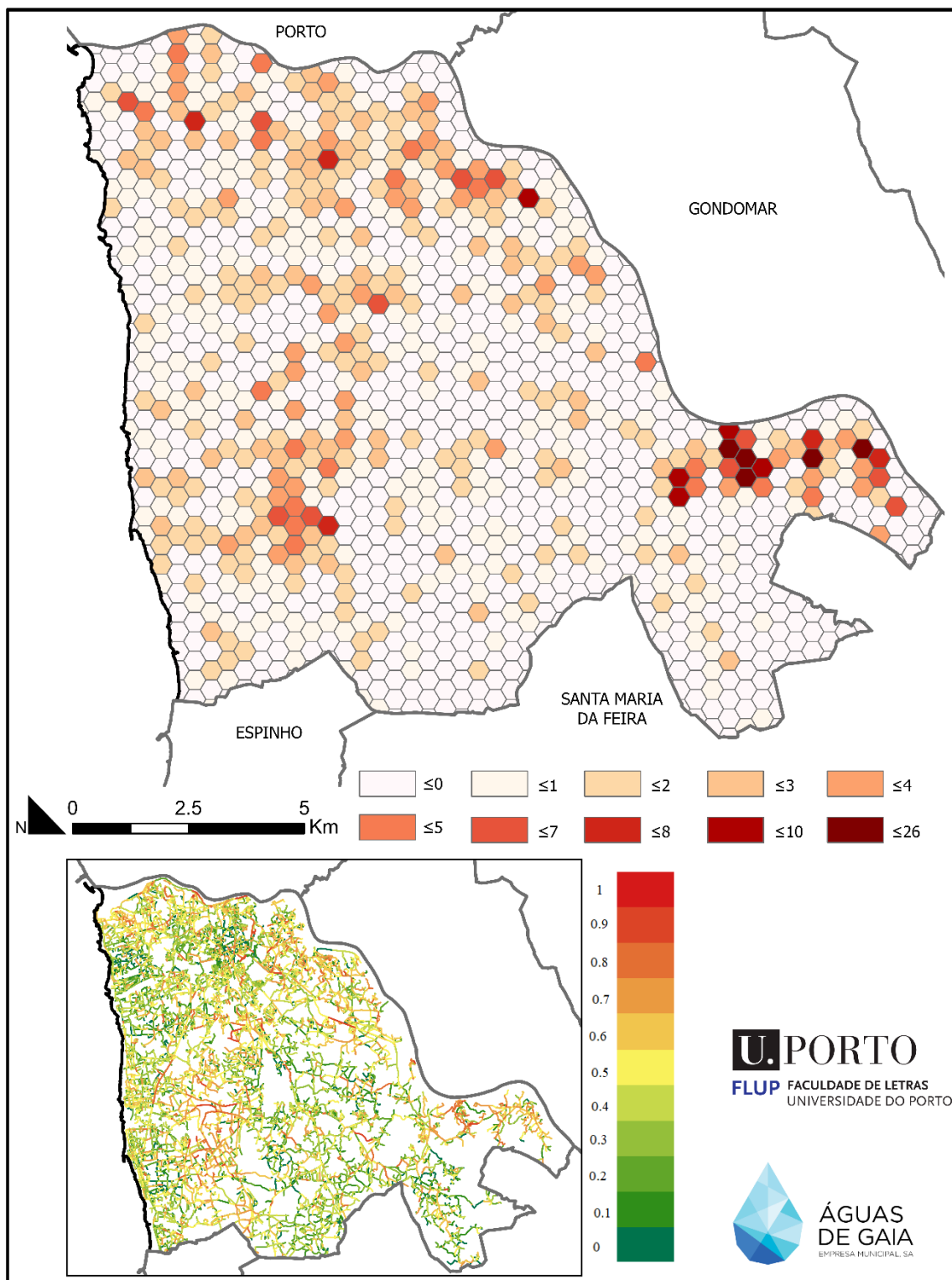


Figura 60 - Contabilização das infraestruturas com uma suscetibilidade prevista superior a 0.7.

5.4 Previsão de fugas – *Decision Tree Classifier*

O modelo de DT, teve como base o modelo de dados desenvolvido e selecionado do RFR, ou seja, o modelo F. Apesar do RFR ser uma variante de DT, este foi desenvolvido posteriormente. E neste caso procura-se realizar uma classificação. Neste tipo de estudo, como referido no Capítulo 2, prever e classificar um evento é muito mais complexo de que efetuar uma regressão. Prever uma fuga numa tubagem, com base naquelas observadas num período de três anos, é uma tarefa difícil e que poderá produzir resultados poucos fiáveis e certamente longe de uma certeza. Houve, porém, o interesse para desenvolver este tipo de modelo para identificar infraestrutura que possa ter fuga no presente, e que não esteja identificada. Uma das vantagens de utilizar este tipo de modelo é que permitem ao utilizador interpretar e visualizar a *decision tree* desenvolvida. Enquanto que nos modelos RFR, são dezenas ou até centenas de árvores utilizadas, neste caso, o modelo DT procura utilizar a árvore mais genérica e que melhor se adapta a todos os casos na área de estudo. É assim possível ter uma perceção sobre a ordem das variáveis e quais os valores de “corte” em cada classe. Para o desenvolvimento deste modelo, foi definido que o número mínimo de *features* em cada nó seria de 250. Apesar de não ser o valor que se traduz numa maior performance do modelo, é um valor que permite uma melhor interpretação da árvore produzida. Valores inferiores iriam traduzir-se numa performance mais elevada, no entanto iriam existir centenas de critérios e a legibilidade da árvore seria não-existente. Assim, com estes parâmetros definidos, a performance situou-se em cerca de 0.88. O que isto quer dizer é que a partir dos dados entrada, em que 70% serviram para efeitos de treino e os restantes 30% para efeitos de teste, cerca de 88% das *features* de treino foram previstas corretamente.

Figura 61 é possível ver a árvore resultante, onde o parâmetro da distância da rede ferroviária foi a variável escolhida de corte inicial, seguido pela elevação e pH. Ao longo da árvore, todos os parâmetros inseridos foram considerados para a classificação da infraestrutura. No primeiro recorte, o modelo classifica 5 tubagens pelo simples facto das mesmas se encontrarem a uma distância da ferrovia superior a 9599 metros. Isto acaba por representar um dos principais problemas deste tipo de modelo que passa pela generalização dos dados. Isto torna-se ainda mais débil tendo em conta a extensão da área de estudo.

O modelo previu fugas em 18% das tubagens de toda a rede. Na Figura 62 é possível verificar algumas semelhanças com os resultados observados no modelo de RFR, em especial nas áreas das freguesias de Crestuma e Lever. No entanto, ao contrário do modelo RFR, não existe uma concentração de previsão de avarias ao longo da faixa norte do município, nem na zona urbana do município. É, no entanto, possível identificar uma concentração de avarias prevista nas

freguesias de Canidelo e São Pedro da Afurada. As áreas interiores das freguesias de Arcozelo e S. Félix da Marinha também não estão identificadas como no modelo de RFR, mas é possível verificar na faixa litoral destes municípios uma concentração de tubagens com avarias previstas.

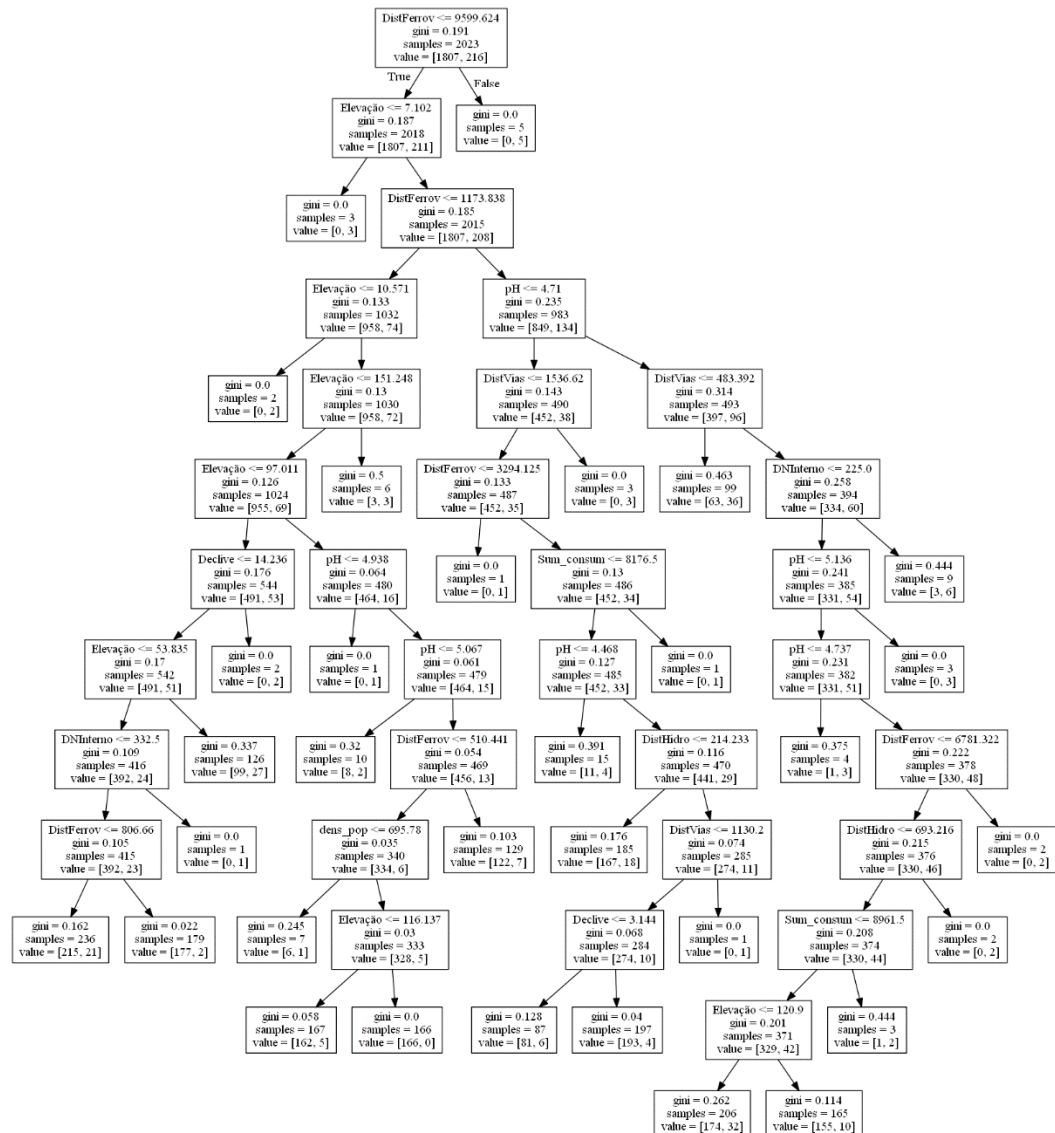


Figura 61 - Decision Tree resultante do modelo F.

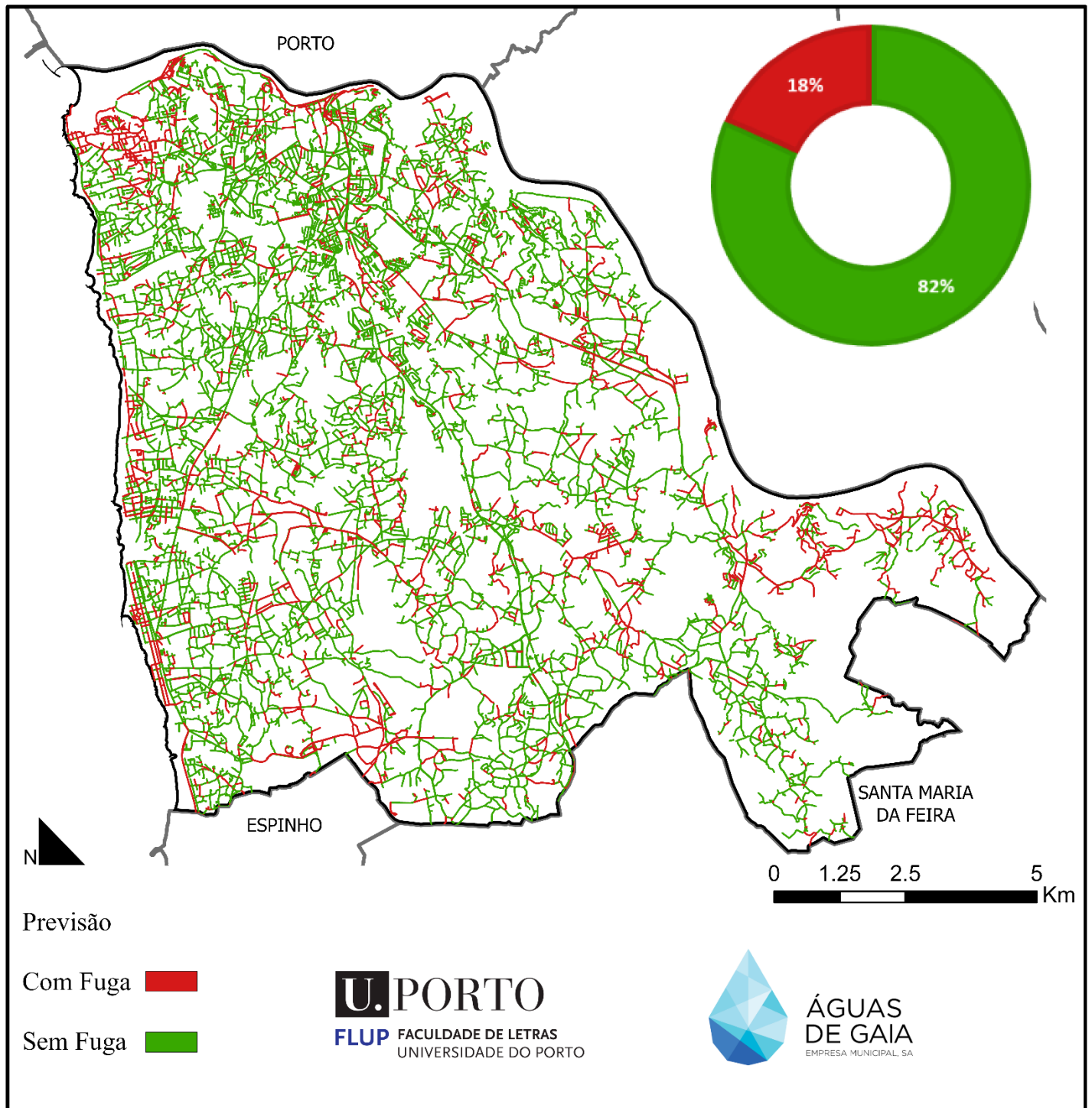


Figura 62 - Previsão de fugas a partir do modelo *Decision Tree*.

CAPÍTULO 6

CONSIDERAÇÕES FINAIS

O desenvolvimento de ferramentas de apoio aos órgãos de tomada de decisão é essencial para o bom funcionamento das infraestruturas e uma redução de custos e redução de perdas de água. Os modelos desenvolvidos ao longo desta investigação permitem identificar áreas e priorizar infraestruturas para a sua manutenção. Estes modelos acabam por ter uma forte componente geográfica, no entanto, é preciso ter em mente que a manutenção deste tipo de infraestruturas pertence essencialmente aos domínios como a Engenharia Civil e Engenharia Hidráulica. Os desenvolvimentos finais destes modelos não deverão depender de um único domínio da ciência, sendo essencial a coordenação de visões multidisciplinares.

A Águas de Gaia, tal como muitas outras entidades gestoras da rede de distribuição de água, tem utilizado até aqui quadrículas sobre a sua área de atuação, e a partir dessas quadrículas, têm procurado fugas no campo com recurso a equipamento especializado. Estas quadrículas têm como objetivo organizar o trabalho a realizar e dar apoio a registar um histórico para existir a informação sobre onde já foi realizado trabalho de deteção. No entanto, não existe informação adicional associado a estas quadrículas. Com os resultados obtidos ao longo desta investigação, e com a espacialização das fugas também em quadrículas, é possível definir áreas prioritárias para a realização de trabalho de campo para a deteção de fugas. Estes modelos transformam-se em ferramentas de trabalho com potencial para rentabilizar o tempo e recursos humanos.

Os resultados produzidos pelo *Decision Trees* e *Random Forest Regressor*, com propósitos diferentes, obtiveram resultados diferentes. A decisão de utilizar o modelo DT porventura não se justifica neste tipo de temática, uma vez que existe uma grande dificuldade em classificar a ocorrência de um evento, especialmente quando tratando-se de infraestrutura que se vai deteriorando ao longo do tempo. Por sua vez, o RFR permite prever a suscetibilidade da infraestrutura o que acaba por fazer muito mais sentido de um ponto vista de manutenção de equipamento. Nas diferentes modelos RFR produzidos é possível identificar áreas comuns com elevado valor de suscetibilidade. As áreas mais suscetíveis estão localizadas na zona norte do município, nas freguesias de Crestuma e Lever e a área entre a freguesia de Serzedo e o litoral do município.

Tendo em conta as lacunas identificadas ao longo do presente (ausência de dados sobre idade da infraestrutura e pressão em cada tubagem), estes modelos estão prontos a serem melhorados a qualquer momento e, numa perspetiva futura, e na eventualidade da recolha de dados em falta, ou na aplicabilidade dos modelos noutras áreas onde existam esses dados, os modelos podem ser adaptados e até alterar o seu comportamento. Eventualmente, em vez de prever a suscetibilidade, com recurso à idade, e tendo em conta o tempo estimado de vida de cada um dos equipamentos e como estes se comportaram face aos parâmetros de vulnerabilidade, seria possível prever um período para a ocorrência de fugas. Este tipo de melhorias, aumentava ainda o poder de decisão uma vez iria abrir uma perspetiva temporal e possibilitar o agendamento de manutenções rotina para evitar fugas e consequentemente, reduzir as perdas de água.

REFERÊNCIAS **BIBLIOGRÁFICAS**

- Allied Corrosion Industries Inc. (s.d.). *Basic Theory of Metallic Corrosion*. Obtido em Agosto de 2020, de [www.alliedcorrosion.com](https://www.alliedcorrosion.com/cathodic-protection-basics/basic-theory-of-metallic-corrosion/): <https://www.alliedcorrosion.com/cathodic-protection-basics/basic-theory-of-metallic-corrosion/>
- Arroyo, A. (2017). *linkedin*. Obtido de Business Intelligence and its relationship with the Big Data, Data Analytics and Data Science: <https://www.linkedin.com/pulse/business-intelligence-its-relationship-big-data-geekstyle/>
- Bakshi, C. (Junho de 2020). *Random Forest Regression*. Obtido em Agosto de 2020, de [gitconnected](https://levelup.gitconnected.com/random-forest-regression-209c0f354c84): <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>
- Berardi, L., Giustolisi, O., Kapelan, Z., & Savi, D. (2008). *Development of pipe deterioration models for water distribution systems using EPR*.
- Bhatia, N. (Junho de 2019). *What is Out of Bag (OOB) score in Random Forest?* Obtido de Towards Data Science: <https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710>
- Breiman, L. (2001). Machine Learning. *Random Forests*, pp. 5-32.
- Busoniu, L., Babuska, R., Schutter, B., & Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*.
- Chowdary, D. (Maio de 2020). *Medium*. Obtido em Junho de 2020, de Decision Trees Explained With a Practical Example: <https://medium.com/towards-artificial-intelligence/decision-trees-explained-with-a-practical-example-fe47872d3b53>
- CM Sesimbra. (Março de 2017). *Autarquia tem apostado na renovação da rede de água em todo o concelho*. Obtido em Outubro de 2020, de Sesimbra:

- https://www.sesimbra.pt/pages/1077?news_id=5287
- Cord, M., & Cunningham, P. (2008). *Machine Learning Techniques for Multimedia - Case Studies on Organization and Retrieval*.
- Data Flair. (2019). *Advantages and Disadvantages of Machine Learning Language*. Obtido em Maio de 2020, de Data Flair: <https://data-flair.training/blogs/advantages-and-disadvantages-of-machine-learning/>
- Dey, S. (2016). *Simply Data Science*. Obtido em Agosto de 2020, de Comparing Spectral partitioning / clustering (with Normalized Graph Laplacian) with KMeans Clustering in R: <https://sandipanweb.wordpress.com/2016/07/30/comparing-spectral-clustering-with-normalized-graph-laplacian-with-kmeans-clustering/>
- ERSAR. (s.d.). *Factos e Números*. Obtido em Julho de 2020, de ERSAR: <http://www.ersar.pt/pt/setor/factos-e-numeros>
- ESRI. (2020). *How Slope works*. Obtido em Maio de 2020, de pro.arcgis: How Slope works
- ESRI. (2020). *Understanding Euclidean distance analysis*. Obtido de desktop.arcgis: Understanding Euclidean distance analysis
- ESRI. (s.d.). *ArcGIS for Desktop*. Obtido em Setembro de 2020, de What are map projections?: <https://desktop.arcgis.com/en/arcmap/10.3/guide-books/map-projections/what-are-map-projections.htm#>
- ESRI. (s.d.). *What is a TIN surface?* Obtido em Maio de 2020, de desktop.arcgis: <https://desktop.arcgis.com/en/arcmap/latest/manage-data/tin/fundamentals-of-tin-surfaces.htm>
- Farley, Malcolm, World Health Organization. (2001). *Leakage management and control : a best practice training manual*.
- García-Mora, B., Bebón, A., Santamaría, C., & Carrión, A. (2014). *Modelling the failure risk for water supply networks with interval-censored data*.
- Gashler, M., Giraud-Carrier, C., & Matinez, T. (2008). *The Seventh International Conference on*

- Machine Learning and Applications. *Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous*, (pp. 900-905).
- Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.
- Gupta, P. (Maio de 2017). *Decision Trees in Machine Learning*. Obtido em Maio de 2020, de Towards Data Science: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
- Jenkins, L., Gokhale, S., & McDonald, M. (2014). *Comparison of Pipeline Failure Prediction Models for Water Distribution Networks with Uncertain and Limited Data*.
- Kam Ho, T. (1995). IEEE. *Random Decision Tree*.
- Kutylowska, M. (2014). *Modelling of Failure Rate of Water-pipe Networks*.
- Liemberger, R. (2004). *Developing a Non-Revenue Water Reduction Strategy, Part 1: Investigating and Assessing Water Losses*.
- Liemberger, R., & Farley, M. (s.d.). *Developing a Non-Revenue Water Reduction Strategy, Part 1: Investigating and Assessing Water Losses*.
- Mayers, J. (Junho de 2019). *Why AI is making tech giants like Google and Amazon even harder to beat*. Obtido em Setembro de 2020, de PitchBook: <https://pitchbook.com/news/articles/why-ai-is-making-tech-giants-like-google-and-amazon-even-harder-to-beat>
- Mitchell, T. (1997). *Machine Learning*.
- Mulaka, M. M. (Setembro de 2012). Malawi Medical Journal. *A guide to appropriate use of Correlation coefficient in medical research*.
- NSW Agriculture; Acid Soil ACTION. (2000). LEAFLET NO.2. *Understanding Soil pH*.
- Polikar, R. (2006). IEEE Circuits and Systems Magazine. *Ensemble Based Systems in Decision Making*, pp. 21-45.
- R, W., Jearsiripongkul, T., & Jiamjiroch, K. (2018). *Elevation Effect in Urban Water Distribution*

Model.

Rogers, P., & Grigg, N. (2009). *Failure Assessment Modeling to Prioritize Water Pipe.*

Russel, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach.*

scikit-learn. (s.d.). *Decision Trees*. Obtido em Fevereiro de 2020, de scikit-learn: <https://scikit-learn.org/stable/modules/tree.html#id2>

Sharma, D. (2020). *Deep Learning without Tears - A simple Introduction.*

Shwartz, S., & David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.*

Statistical Consulting - Kent State University. (2020). *Pearson Correlation*. Em S. C.-K. University, *SPSS Tutorials.*

TLV. (s.d.). *Erosion in Steam and Condensate Piping*. Obtido em Agosto de 2020, de TLV: <https://www.tlv.com/global/TI/steam-theory/piping-erosion.html>

Turing, A. (1950). *Computing Machinery and Intelligence.*

University of South Australia. (s.d.). *Research Methodologies and Statistics - Descriptive statistics.*
Obtido de University of South Australia:
<https://lo.unisa.edu.au/mod/book/view.php?id=646432>

Utgoff, P. (1989). Kluwer Academic Publishers. *Incremental Induction of Decision Trees*, pp. 161-186.

Wahid, A. (2017). *Big data and machine learning for Businesses*. Obtido em Julho de 2020, de Data & Analytics.

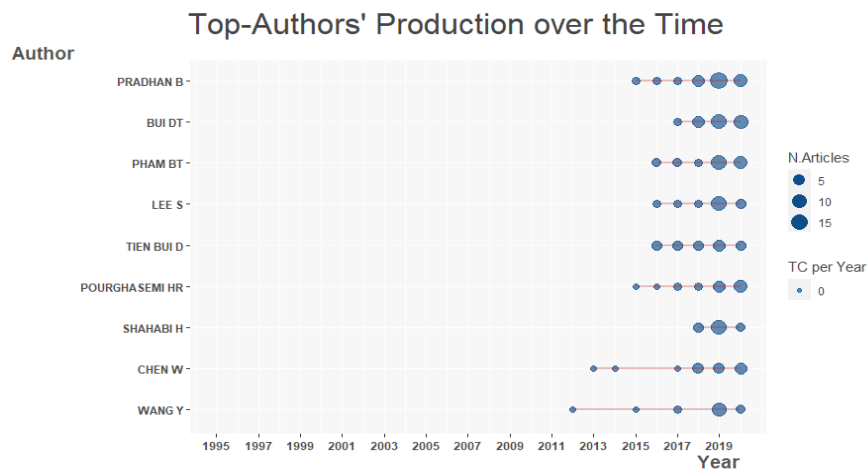
Whigham, P., & McKay, R. (1992). *Machine Induction of Geospatial Knowledge.*

ANEXOS

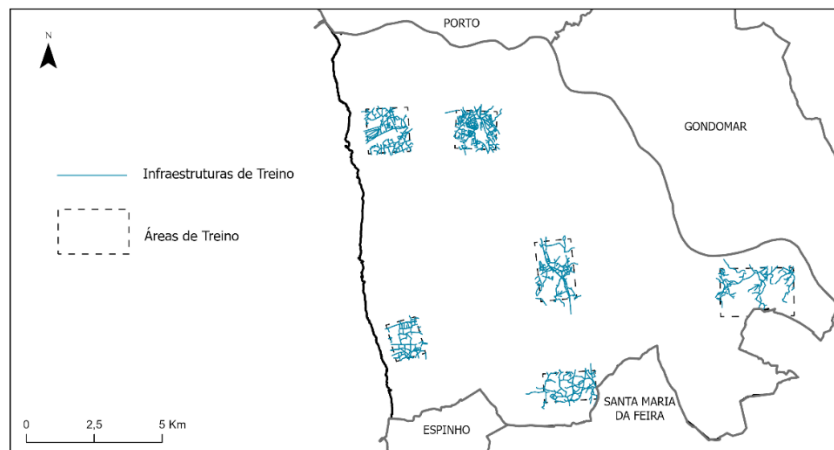
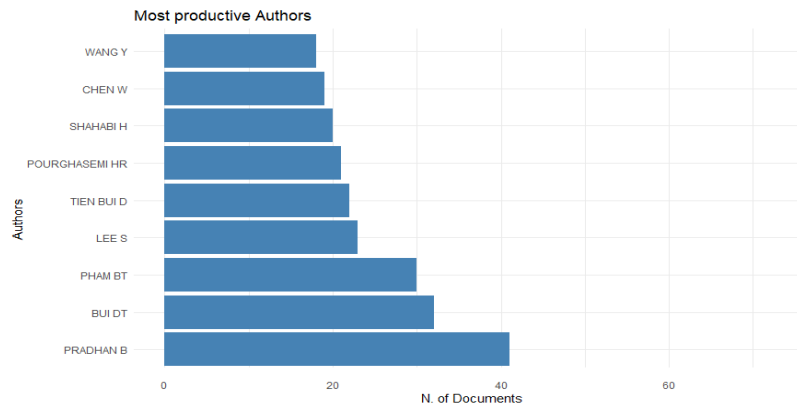
Anexo 1 - Código python para obter os títulos dos artigos com os termos "GIS" e "Machine Learning"	84
Anexo 2 - Principais autores com publicações com os termos "GIS" e "Machine Learning", e sua produção anual.	84
Anexo 3 - Áreas de treino utilizadas para o desenvolvimento dos modelos preditivos.	85
Anexo 4 - Descrição estatística dos parâmetros utilizados.	85
Anexo 5 - Relação entre os parâmetros de vulnerabilidade.	86
Anexo 6 - Importância das variáveis.	87

```
f = open("C:/gis_machinlearning.bib", "r")
searchlines = f.readlines()
f.close()
for i, line in enumerate(searchlines):
    if "title" in line:
        for l in searchlines[i:i+1]: print l,
        print
```

Anexo 1 - Código python para obter os títulos dos artigos com os termos "GIS" e "Machine Learning"



Anexo 2 - Principais autores com publicações com os termos "GIS" e "Machine Learning", e sua produção anual.

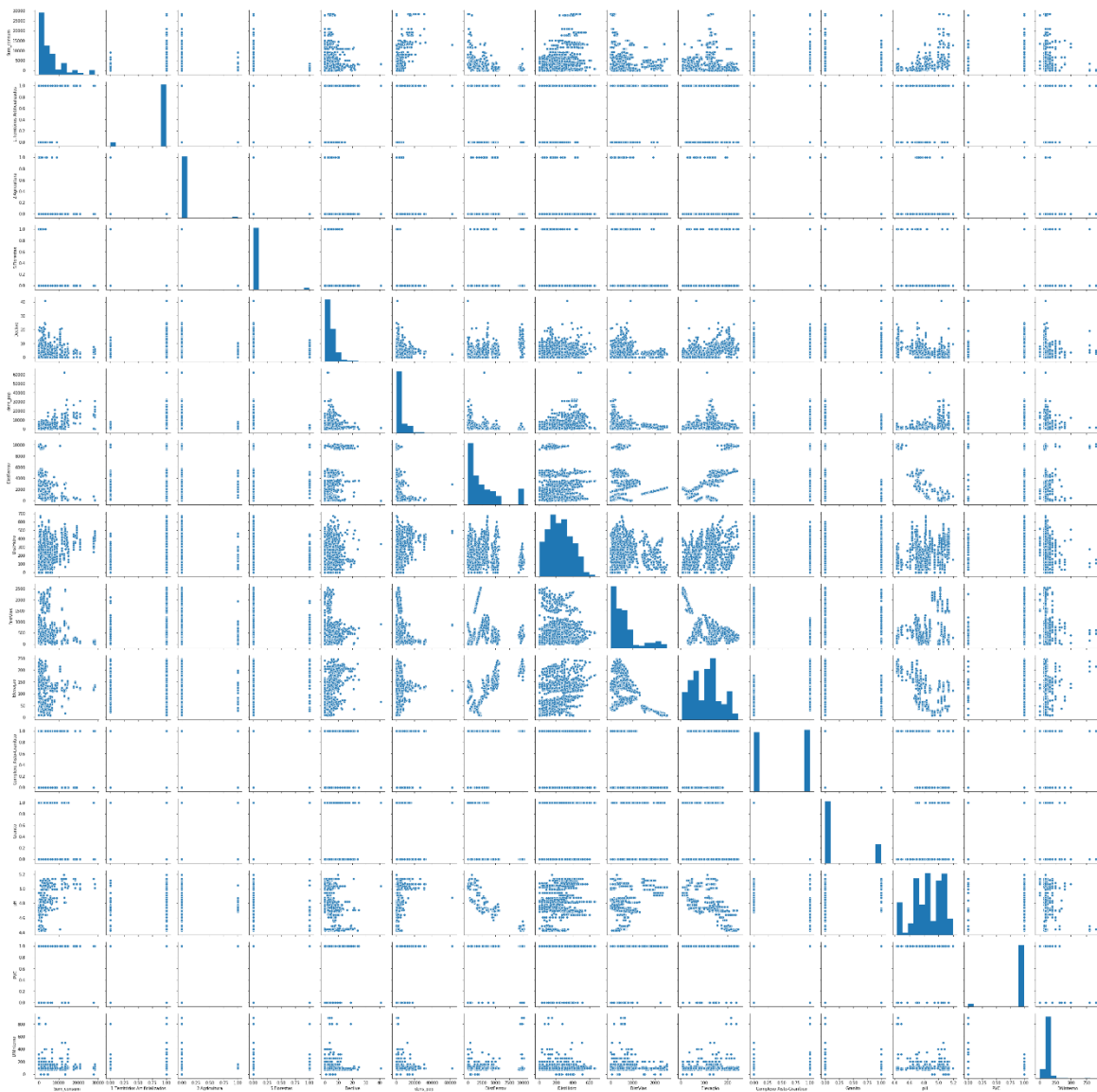


Anexo 3 - Áreas de treino utilizadas para o desenvolvimento dos modelos preditivos.

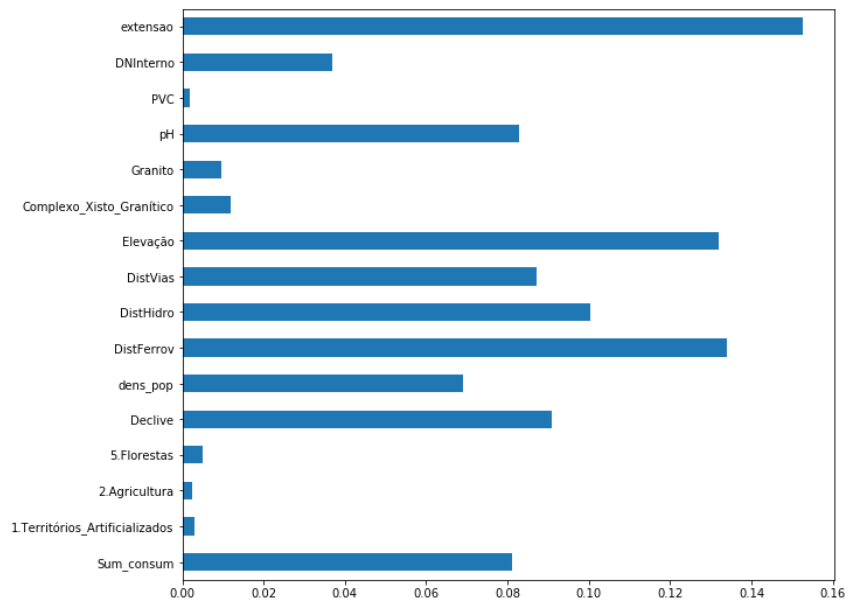
	PKEY	NAvarias	Sum_consum	1.Territórios_Artificializados	2.Agricultura	5.Florestas	Declive	dens_pop	DistFerrov
count	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000
mean	11176.650173	0.113841	7393.703460	0.914187	0.03737	0.034256	0.571197	5397.222311	2786.527169
std	7054.673866	0.317673	10306.507582	0.280136	0.18970	0.181918	186.102949	7320.847178	2964.585790
min	3.000000	0.000000	-97.000000	0.000000	0.00000	0.000000	-9999.000000	0.000000	-9999.000000
25%	4506.750000	0.000000	953.000000	1.000000	0.00000	0.000000	2.048925	980.198975	480.000000
50%	10733.500000	0.000000	3132.000000	1.000000	0.00000	0.000000	3.465951	2552.125000	1111.821899
75%	17587.750000	0.000000	10611.000000	1.000000	0.00000	0.000000	5.270106	6252.740234	5180.888184
max	22954.000000	1.000000	83821.000000	1.000000	1.00000	1.000000	27.053013	47431.101560	11072.885740

	DistHidro	DistVias	Elevação	Complexo_Xisto_Granítico	Granito	pH	PVC	DNInterno	extensao
count	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000	2890.000000
mean	297.270393	546.670651	100.278646	0.589965	0.239100	4.902047	0.943945	119.254325	74.309351
std	180.824623	548.142483	195.703533	0.491925	0.426608	0.232273	0.230068	88.748106	123.182869
min	0.000000	0.000000	-9999.000000	0.000000	0.000000	4.427779	0.000000	0.000000	0.028862
25%	150.742645	141.421356	46.713651	0.000000	0.000000	4.723488	1.000000	90.000000	3.277957
50%	268.328156	350.000000	114.299488	1.000000	0.000000	4.890849	1.000000	90.000000	25.729650
75%	424.264069	812.399460	140.000000	1.000000	0.000000	5.118401	1.000000	110.000000	103.446751
max	891.964111	2358.495361	225.000000	1.000000	1.000000	5.277678	1.000000	900.000000	2146.350098

Anexo 4 - Descrição estatística dos parâmetros utilizados.



Anexo 5 - Relação entre os parâmetros de vulnerabilidade.



Anexo 6 - Importância das variáveis.