Universidade do Minho
Escola de Ciências

Carlos Miguel Fernandes Ferreira

Avaliação e Modelação do Risco de Ocorrência de Cheias em Bacias Hidrográficas

janeiro de 2023

Avaliação e Modelação do Risco de Ocorrência de Cheias em Bacias Hidrográficas
Carlos Miguel Fernandes Ferreira
UMinho |2023

Universidade do Minho
Escola de Ciências

Carlos Miguel Fernandes Ferreira

Avaliação e Modelação do Risco de Ocorrência de Cheias em Bacias Hidrográficas

Dissertação de Mestrado
Mestrado em Ciências e Tecnologias do Ambiente – Remediação Ambiental

Trabalho efetuado sob a orientação do(a)
Professor Doutor Jorge Vieira Pamplona e Professora Doutora Cecília Maria Vasconcelos Costa Castro

março de 2023

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho

Atribuição

CC BY

https://creativecommons.org/licenses/by/4.0/

Agradecimentos

Gostaria de expressar minha sincera gratidão aos meus orientadores Professores Cecília Castro e Jorge Pamplona pelo apoio e predisposição ao longo deste projeto. Foram fundamentais para o meu crescimento académico e profissional, compartilhando generosamente os seus conhecimentos. Além disso, gostaria de agradecer à Doutora Ana Margarida Bento pelo seu acompanhamento e pela sua contribuição para o meu progresso. Foi uma grande fonte de motivação para mim e sinto-me muito grato por ter tido a oportunidade de conferenciar com alguém com o seu conhecimento.

Agradeço ao Professor Zêzere, fundador do Projeto DISASTER, e a sua equipa pelo projeto incrível que desenvolveram e pelo apoio que me deram ao possibilitar a partilha de dados essenciais no desenrolar do presente documento. A partilha de informação é fundamental para o avanço do conhecimento e eu agradeço a todos aqueles que abraçam essa filosofia.

Gostaria de agradecer a minha família por todo o apoio e incentivo durante este projeto, especialmente à minha mãe Maria do Céu por ser a base sobre a qual me construí.

Ao meu amigo e companheiro de trincheira Emanuel Vieira.

Por último, mas não menos importante, o meu agradecimento à minha cadela Maia pela sua companhia afincada durante todo este processo. Espero conseguir compensar-te.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

"As emoções positivas que experimentamos são vividas em função dos nossos objectivos. Só somos tecnicamente felizes, se sentirmos que estamos a progredir - e a simples ideia de progredir implica valor. Pior ainda é o facto de o sentido da vida sem valor positivo não ser simplesmente indiferente. Porque somos vulneráveis e mortais, a dor e ansiedade fazem parte integrante da existência humana. Temos de ter alguma coisa para contrapor ao sofrimento que é intrínseco ao Ser. O sentido tem de ser inerente a um sistema profundo de valores, caso contrário, o horror da existência rapidamente se torna mais importante. E o niilismo, com o seu desespero e desesperança, põe-se à espreita..."

Jordan B. Peterson

Resumo

As cheias são uma das catástrofes naturais mais prejudiciais e frequentes na Europa e um problema especialmente grave em Portugal. Cheia é um fenómeno hidrológico extremo, de frequência variável, natural ou induzido pela ação humana, que consiste no transbordo de um curso de água relativamente ao seu leito ordinário, originando a inundação dos terrenos ribeirinhos (Chow, 1956).

A presente dissertação intitulada de *Avaliação e Modelação do Risco de Ocorrência de Cheias em Bacias Hidrográficas* tem como principal objetivo estudar fatores desencadeadores de eventos de cheia. Para tal, recolheram-se e armazenaram-se dados de ocorrência de cheias, assim como de variáveis hidro-meteorológicas, relativos à bacia hidrográfica do rio Douro.

O rio Douro e os seus afluentes apresentam perfis longitudinais bastante íngremes em algumas secções, observando-se, consequentemente, elevações súbitas dos níveis de água após precipitações intensas. Trata-se de uma preocupação constante, especialmente para as populações ribeirinhas, que são frequentemente afetadas por eventos de cheia ao longo dos anos, com impactos económicos e sociais devastadores.

O tratamento e análise dos dados recolhidos, inicia-se com um estudo univariado das diferentes variáveis consideradas. Utilizam-se diversos procedimentos estatísticos, de forma a compreender a eventual relação de cada uma das variáveis observadas com a ocorrência de cheia, quer individualmente quer de forma global. Para tal, utilizam-se testes exatos de Fisher, testes do qui-quadrado, modelos de regressão logística e florestas aleatórias, explicativos do fenómeno de cheia, ajustados com base nos dados disponíveis. No modelo de regressão logística, há necessidade de se usarem os preditores categorizados porque as suas distribuições empíricas apresentam assimetria positiva muito acentuada, com muitos `outliers`. Neste modelo, os preditores importantes são a precipitação acumulada mensal, em mm, e a descarga de superfície mensal dam^3. O modelo apresenta uma especificidade superior a 90% mas sensibilidade de apenas 33,3%, o que não é de estranhar devido à complexidade do fenómeno em análise. A capacidade discriminatória do modelo de regressão logística, medida pela área abaixo da curva ROC (Receiver Operating Characteristic Curve), AUC (Area Under the Curve), é 76,8% sendo, portanto, aceitável.

O algoritmo de florestas aleatórias é utilizado com as variáveis não categorizadas, pois este não depende das suas distribuições. Com os mesmos preditores, obtém-se com este procedimento uma especificidade superior a 99% e uma sensibilidade de 60%, traduzindo um ótimo desempenho tendo em conta a complexidade do fenómeno e o facto de se estarem a usar apenas dois preditores.

Palavras-chave: Cheias, Fatores Desencadeadores, Bacia do Rio Douro, Regressão Logística, Florestas Aleatórias.

Abstract

Floods are one of the most harmful and frequent natural disasters in Europe and a particularly severe problem in Portugal. Flooding is an extreme hydrological phenomenon of variable frequency, either natural or induced by human action, that consists of the overflow of a watercourse relative to its ordinary bed, resulting in the flooding of riverside lands (Chow, 1956).

The present dissertation entitled *Assessment and Modeling the Risk of Flood Occurences in Hydrographic Basins* aims to study the triggering factors of flood events. To this end, data on flood occurrences, as well as hydro-meteorological variables, were collected and stored for the Douro River basin.

The Douro river and its tributaries have steep longitudinal profiles in some sections, resulting in sudden water level increases after heavy precipitation. This is a constant concern, especially for riverside populations, who are frequently affected by flood events over the years, with devastating economic and social impacts.

The data treatment and analysis begins with a univariate study of the different variables. Various statistical procedures are used to understand the possible relationship of each of the observed factors with flood occurrence, both individually and globally. For this, Fisher's exact tests, chi-square tests, logistic regression models, and explanatory random forests are used to adjust the flood phenomenon based on the available data. In the logistic regression model, the predictors need to be categorized because their empirical distributions have very pronounced positive asymmetry, with many `outliers`. In this model, the important predictors are monthly accumulated precipitation, in mm, and monthly surface discharge dam^3. The model has a specificity higher than 90% but sensitivity of only 33.3%, which is not surprising given the complexity of the phenomenon under analysis. The discriminatory capacity of the logistic regression model, measured by the area under the ROC curve, Receiver Operating Characteristic Curve, AUC Area Under the Curve, is 76.8%, being therefore acceptable.

The random forest algorithm is used with the uncategorized variables because it does not depend on their distributions. With the same predictors, this procedure obtains a specificity higher than 99% and a sensitivity of 60%, translating into an excellent performance given the complexity of the phenomenon and the fact that only two predictors are being used.

Keywords: Floods, Triggers, Douro River Basin, Logistic Regression, Random Forests.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ROC: *Receiver Operating Characteristic*

AUC: *Area Under the Curve*

ANOVA: *Analysis of Variance*

SIG: *Sistema de Informações Geográficas*

SNIRH: *Sistema Nacional de Informações de Recursos Hídricos*

OOB: *Out-of-Bag*

MSE: *Mean Square Error*

LNEC: *Laboratório Nacional de Engenharia Civil*

EM-DAT: *Emergency Events Database*

UNDRR: *United Nations Office for Disaster Risk Reduction*

Introdução

1.1 Enquadramento do Tema e Objetivos

Riscos naturais são fenómenos capazes de produzir dano e estão associados à evolução da Terra, ao longo do tempo, contemplando uma variedade de eventos, tais como cheias e inundações, secas, sismos, incêndios, entre outros. Esta é a denominação utilizada para se fazer referência aos riscos que não podem ser facilmente atribuídos ou relacionados com atividade antrópica. No entanto, atualmente, trata-se de uma tarefa cada vez mais difícil a distinção entre causas dos riscos naturais, causas naturais ou antrópicas (Rebelo, 2003).

Só no ano de 2015, os desastres naturais atingiram mais de 98 milhões de pessoas, com um prejuízo avaliado na ordem dos 66,5 mil milhões de dólares [1]. O relatório do Painel Intergovernamental sobre as Alterações Climáticas de 2012, evidenciou que alguns eventos extremos aumentaram a sua frequência e/ou magnitude, sendo as cheias um desses eventos (Reinman, 2012).

Têm sido diversas as abordagens a estes fenómenos no desenvolvimento de teorias de risco. O estudo de riscos naturais é necessário, seja segundo uma perspetiva mais naturalista, social ou financeira, de modo a compreendê-los e prevê-los para que, através de implementação de medidas eficientes, se possam mitigar as suas ocorrências e/ou efeitos.

Posto isto, a presente monografia tem como principal objetivo identificar e estudar os fatores capazes de desencadear a ocorrência de eventos de cheia, compreendendo, de acordo com o plano de trabalho, as seguintes linhas:

- Colheita, classificação e elaboração de dados de natureza diversa, relacionada com os contextos territoriais;
- Caracterizar qualitativa e quantitativamente a bacia hidrográfica do rio Douro;
- Identificar, explorar e analisar criticamente fatores desencadeadores de eventos de cheia;
- Propor modelos explicativos do fenómeno de cheia.

[1] https://www.undrr.org/

1.2 Uma Revisão de Fatores desencadeadores de Cheia

Ao longo do tempo, as cheias claramente se evidenciam entre as catástrofes naturais que apresentam maiores prejuízos materiais e perda de vidas (Alcoforado et al., 2021). Na Europa, os danos causados pelas cheias e inundações representam um terço das perdas económicas associadas a desastres naturais (Santos et al., 2018), (Garrote et al., 2016). Em Portugal, estas são as catástrofes mais recorrentes, assinalando o segundo lugar em termos de número de mortos, pessoas afetadas e prejuízos materiais registados[2]. Os desastres provocados por cheias têm vindo a aumentar no globo, como consequência da expansão urbana em planícies aluviais e áreas impermeabilizadas, da artificialização dos canais do escoamento dos rios, dos sistemas deficientes de drenagem de águas pluviais nos meios urbanos e da falta de limpeza e desassoreamento dos rios. Constituindo assim, uma constante preocupação para as populações e entidades responsáveis pela sua gestão (Pardal et al., 2016).

As cheias são um desastre hidrometeorológico complexo (Llasat et al., 2005), geralmente causadas por eventos de precipitação no local ou a montante do mesmo, sendo fortemente influenciadas pela localização, intensidade e duração do evento de precipitação (Johnson et al., 2016) e pelas condições antecedentes do local, como humidade, tipo e uso do solo, condições geomorfológicas, questões como a sazonalidade, variabilidade e irregularidade dos regimes hidrológicos; a existência de estruturas de proteção e/ou mitigação (Santos et al., 2015).

Numa bacia hidrográfica, os processos hidrológicos ocorrem sobre o efeito da inter-relação entre a ocupação do solo e o clima. São os elementos climáticos, como por exemplo, a precipitação e temperatura que são essenciais à compreensão da quantidade de água disponível, nessa mesma bacia. Ainda, alterações no uso e cobertura do solo podem impactar o ciclo hidrológico e, consequentemente, a quantidade de água disponível, dadas as alterações na taxa de infiltração, interceção, albedo e evapotranspiração (Nunes, 2011).

O ciclo hidrológico, representado na Figura 1, constitui a base da hidrologia. Este descreve os diversos caminhos através dos quais a água circula e se transforma na natureza, constituindo um sistema de enorme complexidade, uma vez que os processos que integram o ciclo são fenómenos da natureza, ou seja, processos essencialmente aleatórios, que resulta num controlo praticamente nulo sobre estes mesmo processos (Hipólito e Vaz, 2011).

[2] https://www.emdat.be/

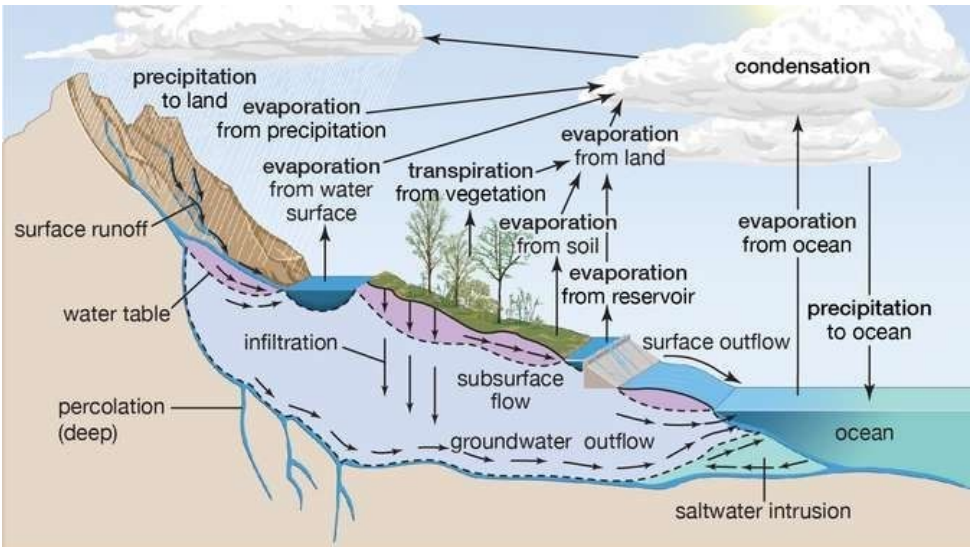

Figura 1: Ciclo Hidrológico, extraído de Encyclopædia Britannica, Inc.

Em Portugal, bem como na região do mediterrâneo, verifica-se um padrão complexo no que toca à variabilidade espacial e sazonal da precipitação, acentuada pela instabilidade na ocorrência de chuvas e intensidade do evento, que acaba por dificultar a tarefa de avaliar a tendência da distribuição da precipitação e o impacto das potenciais alterações climáticas. Variações na quantidade e distribuição inter e intra-anual da precipitação podem surtir em impactos económicos e ambientais significativos. Sendo a precipitação um elemento-chave no ciclo hidrológico, oscilações deste fator, por norma, alteram os regimes fluviais e de recarga de águas subterrâneas, perturbando, assim, a disponibilidade de água na região e, por conseguinte, a produtividade do ecossistema, erosão do solo, ocorrência de incêndios, entre outros (Nunes et al., 2016).

A urbanização de grandes áreas leva a uma considerável redução da infiltração e, por conseguinte, gera-se um aumento do escoamento superficial, o que acaba por criar ou ampliar áreas propensas a cheia. A elevada densidade populacional, de infraestrutura e de atividades socioeconómicas, para além de aumentarem significativamente o risco de ocorrência de cheia, também tendem a intensificar os seus impactes, ou seja, aumentam a vulnerabilidade da região (Diakakis et al., 2020).

Geralmente, na prevenção e correção dos distúrbios provocados por cheias, consideram-se dois tipos de medidas: estruturais, que consistem na execução de obras de engenharia, com o objetivo de controlar as cheias e de reduzir a extensão das áreas inundáveis, de carácter maioritariamente corretivo e não-estruturais, que consistem no planeamento e regulamentação da ocupação de áreas inundáveis, com o objetivo de impedir, por exemplo, a construção de infraestruturas capazes de agravar os efeitos das cheias, com carácter essencialmente preventivo. A adopção de soluções adequadas a cada caso passa pela combinação destes dois tipos de medidas (LNEC, 1990). É fundamental a execução de processos de planeamento eficazes na resolução desta problemática, capaz de desempenhar um papel fundamental à escala nacional, regional e local no sentido de assegurar que o desenvolvimento seja promovido e dirigido de modo sustentável em termos económicos, sociais e

ambientais (Sá et al., 2016).

1.3 Estrutura do Trabalho

Para além da introdução, onde se faz o enquadramento do tema, apresentando os objetivos específicos do trabalho e uma revisão dos principais fatores desencadeadores de cheia, o presente documento contém mais 4 capítulos. No capítulo 2, são apresentadas as fontes de recolha dos dados e diversas metodologias aplicadas. O capítulo 3, apresenta uma breve caracterização do território nacional e da bacia hidrográfica do rio Douro. O capítulo 4, consta de uma análise estatística de dados hidrometeorológicos, entre 1990 e 2021, com o intuito de perceber a sua relação com eventos de cheia. Neste capítulo apresentam-se ainda um modelo de regressão logística e um modelo de florestas aleatórias com o objetivo da explicação do fenómeno. No capítulo 5 apresentam-se conclusões do trabalho, assim como propostas de trabalho futuro. Finalmente, tendo sido utilizados, na análise e tratamento de dados, os software SPSS, versão 25.0 (IBM Statistics) e R, versão 4.2.0., esta monografia conta ainda com um Apêndice com diversos códigos em R utilizados.

Material e Métodos

Neste capítulo apresentam-se as principais fontes de recolha dos dados, assim como uma breve explicação das metodologias utilizadas.

2.1 Material

Na fase inicial deste projeto houve necessidade de se proceder à recolha de dados com informação de cheias, bem como de seus potenciais fatores desencadeadores, tendo como base estudos anteriores e garantir o seu armazenamento digital, à escala da bacia hidrográfica em estudo. Tratou-se de uma fase fundamental para se proceder à análise de risco de cheias.

Os dados sobre cheias tiveram como base o projeto de investigação científica, "DISASTER - Desastres naturais de origem hidro-geomorfológica em Portugal: base de dados SIG para o apoio à decisão no ordenamento do território e planeamento de emergência", financiado pela Fundação para a Ciência e Tecnologia (Zêzere et al., 2022). De facto, este projeto inclui uma base de dados geográfica sobre desastres hidrogeomorfológicos registados em território continental, desde 1865 até à atualidade, constituindo um suporte de extrema importância nos processos de avaliação de risco em Portugal, sendo uma referência para a aplicação de medidas de mitigação destes desastres.

Para além desta informação, recorreu-se ao SNIRH [1], Sistema Nacional de Informação de Recursos Hídricos, para obter informação pluviométrica e hidrométrica. Aqui foram recolhidos dados de precipitação acumulada mensal da estação de Amarante, código 06I/01G, assim como dados de caudal afluente médio mensal e descarga de superfície mensal da estação Albufeira do Carrapatelo (R.E.), código 07I/01A.

É importante referir que a análise ao regime hidrológico está muito condicionada pela disponibilidade de dados.

2.2 Métodos

Nesta secção introduz-se alguma terminologia associada à questão de investigação em análise e faz-se uma breve exposição da metodologia estatística utilizada neste trabalho. A metodologia estatística encontra-se com mais

[1] https://snirh.apambiente.pt/

detalhe em Pestana, 2014; Cox, 1958; Breiman, 1996.

2.2.1 Testes de independência para variáveis categóricas

O estudo de associação entre variáveis qualitativas pode ser efetuado através da construção e análise de tabelas de contingência e pela utilização de testes estatísticos, nomeadamente teste exato de Fisher e testes de Qui-Quadrado.

Estes testes aplicam-se a dados nominais, organizados em classes e permitem testar a independência de duas variáveis ou a homogeneidade de várias distribuições. Nos dois casos a estatística de teste é a mesma, quer se trate de um teste de independência quer se trate de um teste de homogeneidade. A diferença entre estes dois testes reside na questão de investigação.

a) Testes do Qui-Quadrado

Os testes do qui-quadrado baseiam-se numa estatística de teste cuja distribuição é qui-quadrado. Assim, se a variável X assume valores em k categorias e Y assume valores em m categorias, tem-se, num teste de independência:

- H_0: X e Y são independentes.
- H_1: X e Y não são independentes.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(k-1)(m-1)$$

onde:

- O_{ij} = Frequência absoluta observada na célula (i, j)
- E_{ij} = Frequência esperada na categoria (i, j), quando H_0 é verdadeira (isto é independência)

A distribuição da estatística de teste é assintótica, sendo apenas válida se os valores das frequências esperadas em todas as células forem, no mínimo, 1 e, em 80% dos casos, no mínimo 5. Caso estas condições não se verifiquem, deve usar-se o Teste Exato de Fisher.

b) Teste exato de Fisher

O teste exato de Fisher baseia-se numa generalização multivariada da distribuição hipergeométrica. A partir dos valores da soma das linhas e da soma das colunas (distribuições marginais de X e de Y, respetivamente) obtêm-se todas as tabelas, com entradas não negativas, consistentes com esses valores. Para cada uma dessas tabelas são calculadas as probabilidade condicionadas das células usando a distribuição hipergeométrica multivariada.

Para calcular o p-*value* do teste, é necessário ordenar as tabelas de contingência geradas a partir das distribuições marginais de X e Y usando algum critério de dependência. O critério usado é geralmente a diferença entre as frequências observadas e esperadas na tabela original, medida por algum teste de associação, como o teste qui-quadrado. As tabelas que apresentam um desvio igual ou maior da independência são aquelas cujas probabilidades são somadas para calcular o p-*value* . Existem vários critérios que podem ser usados para medir a dependência, como o coeficiente de contingência ou o teste de G-quadrado.

O teste é mais comumente aplicado a matrizes 2×2 sendo computacionalmente pesado para tabelas de maior dimensão, podendo, no entanto, ser utilizado.

2.2.2 Modelos de Regressão Linear Múltipla

Para explicar a eventual relação entre uma variável dependente quantitativa contínua (Y) e um conjunto de variáveis independentes, ou preditores (X), quantitativos ou qualitativos, pode ser possível usar modelos de regressão linear.

A expressão geral para o modelo é:

$$Y_i = B_0 + B_1 X_1 + B_2 X_2 + ... + B_k X_k + \varepsilon_i, \quad i = 1, 2, ..., n$$

Os parâmetros do modelo são obtidos por minimização da soma dos quadrados dos resíduos – método dos mínimos quadrados.

A aplicação deste modelo depende da verificação de vários pressupostos sobre o comportamento dos erros que devem ser Normais, homocedásticos, não correlacionados e com média nula.

2.2.3 Testes para variável resposta contínua

Os testes t de Student e os testes ANOVA são os testes estatísticos mais comumente usados para analisar a relação entre uma variável resposta contínua e uma ou mais variáveis preditoras categóricas. Esses testes baseiam-se em modelos lineares, e pressupõem a normalidade e a homogeneidade dos erros.

Quando as amostras são de pequena dimensão, ou mesmo, sendo de dimensão elevada, as distribuições da resposta, dentro de cada categoria da variável independente, apresentam muitos `outliers` ou são de cauda pesada, esses pressupostos podem não ser válidos, e portanto, é recomendável usar alternativas não paramétricas, como o teste de Wilcoxon-Mann-Whitney ou o teste de Kruskal-Wallis.

a) Teste de Wilcoxon, Mann & Whitney

O teste U de Wilcoxon-Mann-Whitney é uma alternativa não paramétrica ao teste t de Student de comparação das médias de duas populações independentes. Possibilita testar a igualdade de distribuições, com amostras de

pequena dimensão, cujo comportamento se afasta da distribuição Normal.

O teste de U de Wilcoxon-Mann-Whitney baseia-se nas ordens das observações e não nos seus valores. A estatística de teste é dada por

$$U = \min(U_1; U_2);\ U_1 = n_1 \times n_2 + \frac{n_1(n_1+1)}{2} - R_1; U_2 = n_1 \times n_2 - U_1$$

onde,

- $n_1 =$ dimensão da menor amostra
- $n_2 =$ dimensão da maior amostra
- $R_1 =$ soma das ordenações da menor amostra

A estatística de teste, com correção no caso de empates na ordenação dos valores, tem distribuição assintótica Normal.

b) Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é baseado na ordenação dos dados (rank) e testa a hipótese nula de que as médias das ordens são iguais para todos os grupos. O teste de Kruskal-Wallis é mais robusto a violações dos pressupostos de normalidade e homogeneidade dos erros, e é frequentemente usado quando esses pressupostos não são atendidos.

A estatística de teste é calculada como a razão entre a soma das diferenças entre as ordens observadas e esperadas, dividida pela variância das ordens esperadas e tem uma distribuição de qui-quadrado com k-1 graus de liberdade, onde k é o número de grupos.

Quando rejeitamos a hipótese nula no teste de Kruskal-Wallis, é necessário realizar testes post-hoc para identificar quais grupos específicos são responsáveis pela diferença significativa. Alguns testes post-hoc adequados para usar após o teste de Kruskal-Wallis incluem:

- Teste de Dunn: compara cada par de grupos individualmente, e tem menos poder estatístico do que outros testes post-hoc, mas é mais conservador para detetar diferenças significativas;
- Teste de Tukey: compara cada par de grupos e é mais poderoso do que o teste de Dunn, mas também é mais propenso a erros tipo I.

2.2.4 Modelos de Regressão Logística

Num modelo linear geral, tal como os atrás descritos, ANOVA e de Regressão Múltipla, a variável resposta é quantitativa contínua. Existem várias generalizações deste modelo, nomeadamente no que diz respeito ao tipo da

variável dependente. No caso em que a variável resposta, Y, é binária, assumindo apenas valores 0 ou 1, o modelo é de regressão logística.

Num modelo de regressão logística, $E(Y)$ é combinação linear dos p preditores, $X_1, X_2, \ldots, X_p$ através de uma função de ligação g.

$$E(Y|x) = E(Y) = P(Y = 1|x) = \pi(x)$$

Tendo-se:

$$g(x) = g(E(Y)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Da relação anterior vem

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}$$

A importância desta transformação deve-se ao facto de $g(x)$ apresentar algumas propriedades do modelo de regressão linear, mais concretamente a linearidade, continuidade e o facto de $E(Y|x)$ poder assumir valores entre $-\infty$ e $+\infty$.

Um modelo de regressão logística permite classificar um novo caso numa das duas classes da variável resposta. Usualmente, se a probabilidade do caso, estimada pelo modelo, for superior a 0,5, o caso é classificado como 1, caso contrário, o caso é classificado como 0. O valor 0,5 é geralmente designado por *cut-off* e pode ser diferente de 0,5.

Neste tipo de procedimento tem interesse a construção de uma matriz de confusão onde surge a classificação de casos de acordo com o modelo vs. a classificação real (observada), Tabela 1. Nesta tabela a resposta é designada por `endpoint`.

Tabela 1: Exemplo de uma matriz de confusão.

		Estimado	
		`endpoint` = 1	`endpoint` = 0
Observado	`endpoint` = 1	A	B
	`endpoint` = 0	C	D

A sensibilidade do modelo é definida como a probabilidade de prever a ocorrência do evento de interesse entre os indivíduos em que este foi observado, ou seja estimada pela proporção de verdadeiros positivos. A especificidade fornece a proporção de falsos negativos.

$$\text{Sensibilidade} = \frac{A}{A+B}$$

$$\text{Especificidade} = \frac{D}{C+D}$$

a) Curva ROC

A construção de uma curva ROC (*Receiver Operating Characteristic*) permite avaliar a variação da sensibilidade e especificidade para cada valor de *cut-off*. A sensibilidade é apresentada no eixo das ordenadas e (1- especificidade) no eixo das abcissas. Esta análise procura averiguar a capacidade classificativa do modelo. Ao observar o gráfico, verifica-se que o ideal seria encontrar uma área sob a curva próxima de 1.

Utiliza-se a área sobre a curva como o critério para identificar o poder discriminatório de um modelo de regressão logística.

- $ROC = 0.5$, o modelo não faz qualquer discriminação entre os indivíduos com ou sem `endpoint`;
- $0.6 \leq ROC < 0.7$, o modelo apresenta uma discriminação limitada;
- $0.7 \leq ROC < 0.8$, o modelo apresenta uma discriminação aceitável;
- $0.8 \leq ROC < 0.9$, o modelo apresenta uma excelente discriminação;
- $ROC \geq 0.9$, o modelo apresenta uma discriminação quase perfeita.

2.2.5 Árvores de decisão e Florestas Aleatórias

Uma árvore de decisão é um algoritmo de estratificação de um conjunto de dados em conjuntos sucessivamente mais homogéneos, através de observação de regras de decisão baseadas nas características dos dados. Permite visualizar facilmente as regras de decisão e tomar decisões baseadas em dados de forma rápida e eficiente. Quando a variável resposta é qualitativa, fala-se em árvores de classificação, caso seja quantitativa, fala-se em árvores de regressão.

O algoritmo é de partição recursiva. Esta ferramenta é bastante popular, pela sua representação em forma de árvore, por ser bastante intuitiva e de fácil compreensão.

Os nós que constituem as árvores de decisão, refletem os sucessivos testes lógicos aos atributos dos dados, e os ramos representam os resultados desses testes. As folhas da árvore representam as classificações finais. Cada nó interno da árvore representa uma característica do conjunto de dados, e cada ramo representa um valor possível dessa característica. As folhas representam a classificação final do conjunto de dados. A Figura 2 mostra a estrutura de uma árvore de decisão.

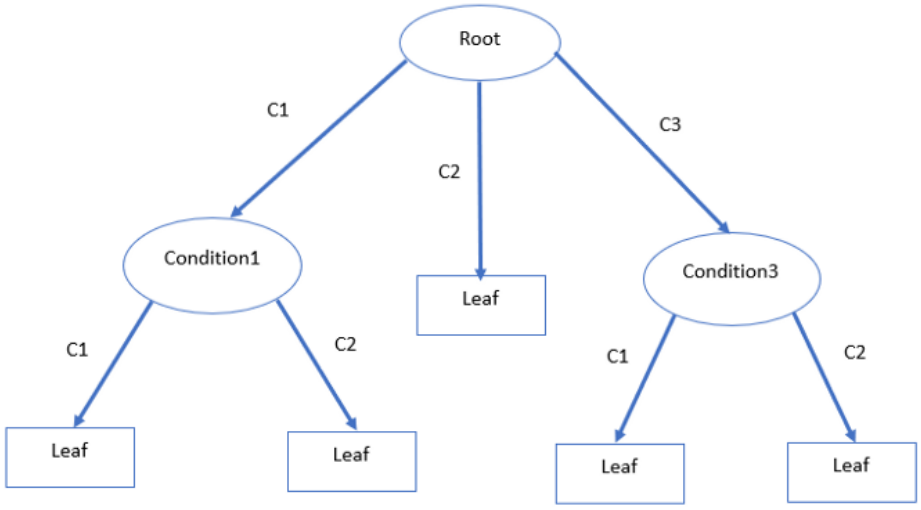

Figura 2: Exemplo da estrutura de uma Árvore de decisão. Extraído de https://regenerativetoday.com/

A escolha do preditor a ser usado em cada nó da árvore, tem por base alguns critérios, tais como impureza, distância ou dependência. O índice de Gini é um indicador comumente usado para medir a impureza dos nós. Mede a probabilidade de um elemento escolhido aleatoriamente ser classificado incorretamente, se for escolhido a partir do conjunto de dados associado ao nó. Quanto menor o índice de Gini, menor é a impureza do nó e melhor é a qualidade da classificação, sendo a sua fórmula de cálculo para o nó O expressa da seguinte forma:

$$G(O) = \sum_{l \neq l^*} p(l;O)p(l^*;O) = 1 - \sum_{l=1}^{L} p(l;O)^2$$

onde,

- L é o número de categorias associado à variável alvo Y;
- O é um nó de uma árvore ou um conjunto de observações;
- l é o índice de categorias para a variável alvo Y;
- $p(l;O)$ a probabilidade empírica de ocorrência da categoria l de Y em O;
- $p(l_*;O)$ a probabilidade efectiva de ocorrência da categoria de l de Y em O.

De forma a evitar o sobre-ajustamento do modelo aos dados, *overfitting*, costuma dividir-se o conjunto de dados em *treino* e *teste*. Os dados de teste não são vistos durante o processo de construção da árvore.

Apesar disto, as árvores de decisão podem ser sensíveis aos dados, ou seja, pequenas variações nos dados de treino podem causar grandes variações na estrutura da árvore gerada, o que pode levar a uma árvore de decisão com *overfitting*, onde a árvore é muito complexa e se ajusta muito bem aos dados de treino, mas não generaliza bem para novos dados.

Para lidar com esse problema, existem outras técnicas como a poda de árvore, que remove ramos da árvore que não contribuem significativamente para a precisão da classificação, e também o uso de florestas aleatórias, que é um conjunto de árvores de decisão geradas a partir de diferentes amostras do conjunto de dados, e a classificação final é feita pela maioria das classificações das árvores. Essas técnicas ajudam a diminuir o *overfitting* e aumentar a estabilidade da árvore.

a) Florestas Aleatórias

As florestas aleatórias (FA) representam uma coleção de árvores de decisão, Figura 3, que seguem regras específicas na determinação do crescimento da árvore, na divisão, na combinação de árvores, no auto-teste e no pós-processamento.

Na determinação do crescimento das árvores, as florestas aleatórias usam um processo chamado "amostragem com substituição" para selecionar uma amostra aleatória do conjunto de dados de treino. Cada árvore é construída a partir de uma amostra diferente, o que ajuda a reduzir o *overfitting*.

Na divisão, as florestas aleatórias usam um processo chamado "amostragem aleatória de atributos" para selecionar aleatoriamente um subconjunto de atributos para testar em cada nó da árvore. Isso ajuda a reduzir a dependência de um único atributo e aumenta a diversidade das árvores.

Na combinação de árvores, as florestas aleatórias usam uma técnica chamada "votação maioritária" para classificar novos exemplos. Cada árvore da floresta classifica o exemplo e a classificação final é a classificação que recebeu mais votos.

O auto-teste é o processo de avaliar o desempenho da floresta num conjunto de dados de teste, para verificar o quão bem a floresta generaliza para dados novos.

O pós-processamento é o processo de otimizar a floresta ajustando parametros como número de árvores, profundidade da árvore, entre outros, para obter o melhor desempenho possível.

Cada árvore de decisão captura diferentes tendências nos dados, uma vez que considera apenas um subconjunto dos mesmo. Um conjunto com Z amostras aleatórias de vetores dos preditores em estudo são aleatoriamente e independentemente selecionados, sendo que para cada uma destas Z amostras constrói-se uma árvore. As árvores selecionadas descrevem uma amostra i.i.d (independentes e identicamente distribuídas) de árvores de uma dada floresta. As árvores geradas são combinadas de modo a obter-se a predição conjunta.

Trata-se de uma abordagem poderosa para a exploração e análise de dados, uma vez que árvores de decisão individuais geralmente apresentam alta variabilidade ou elevado enviesamento. As florestas aleatórias visam mitigar estas questões, para além de oferecer um vasto leque de benefícios, tal como a sua precisão e eficiência em tratar grandes dimensões de dados. Mesmo com presença de dados omissos, proporciona estimativas sobre a importância das variáveis em problemas de classificação, entre outros.

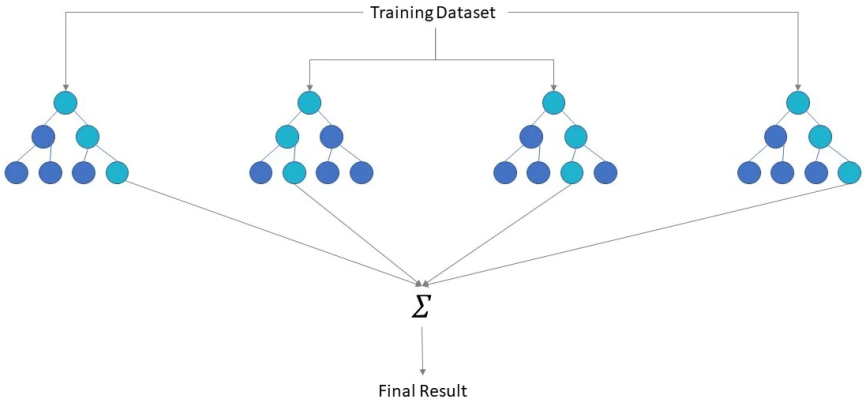

Figura 3: Exemplo de uma Floresta Aleatória. Extraído de https://blent.ai/

Os modelos de florestas aleatórias obedecem a dois parâmetros que condicionam os resultados do modelo, nomeadamente, o número de variáveis submetidas aleatoriamente em cada nó e o número de árvores na floresta.

O número de variáveis submetidas aleatoriamente em cada nó, m, é um hiperparâmetro que controla a complexidade das árvores individuais na floresta. Afeta a diversidade das árvores na floresta, aumentando a precisão do modelo quando aumentado.

O número de árvores na floresta afeta a estabilidade do modelo. Quanto mais árvores na floresta, mais estável o modelo é, mas isso também torna o modelo mais complexo e mais lento.

As florestas aleatórias não precisam de um conjunto de teste para obter uma estimativa imparcial do erro associado, este é estimado internamente, considerando que cada árvore é construída através de uma amostra bootstrap. A técnica de amostragem *bootstrap* consiste em selecionar amostras aleatórias do conjunto de dados de treino, com substituição, e construir uma árvore a partir de cada amostra. Isso ajuda a estimar o erro de generalização do modelo sem precisar de um conjunto de teste.

Cerca de um terço das observações são retiradas da amostra de *bootstrap* (Out of Bag Samples - OOB), que não contribuem para a construção do modelo. Estas amostras OOB são utilizadas para obter uma estimativa imparcial do erro de classificação à medida que se adicionam árvores à floresta e, ainda, obter estimativas para a importância das variáveis para o modelo.

O procedimento de votação conjunta em florestas aleatórias, permite o estabelecimento de uma pontuação para cada variável do modelo. Uma das medidas mais comuns é determinada pela análise do aumento de erro de previsão quando o valor de uma variável num nó de uma árvore é permutado aleatoriamente.

Caracterização Física do Território

3.1 Introdução

O território de Portugal situa-se na zona oeste da Península Ibérica, com uma área aproximada de 88500 km^2. Genericamente, distinguem-se 3 unidades geomorfológicas, das quais o maciço Hespérico, as Orlas Ocidentais e a Bacia Sedimentar do Tejo e do Sado. Sendo que a característica morfológica essencial é a separação pelo rio Tejo de uma região predominantemente montanhosa a norte de uma região quase plana a sul (D. I. Pereira et al., 2014). Nos setores montanhosos, particularmente na Serra da Estrela, para além de temperaturas mais baixas, registam-se elevadas quedas de precipitação que, em termos médios, podem ultrapassar os 2500 mm anuais (Santos e Fragoso, 2013).

Pela sua posição no quadro sinótico global, Portugal apresenta características climáticas claramente mediterrânicas, com um verão seco e quente a opor-se a um inverno tépido e chuvoso. Porém, a distribuição do país em latitude, o maior ou menor afastamento em relação ao oceano e a configuração espacial do relevo justificam comportamentos diferenciados dos elementos climáticos, sendo responsáveis por algumas nuances ao nível regional. Esta variabilidade regional também é consequência da trajetória dos ventos, que influenciam a intensidade e o tipo de precipitação (LNEC, 1990).

No que diz respeito à distribuição anual dos caudais, dos rios portugueses, esta é marcada pela sua forte variabilidade inter-anual, que é frequente em rios mediterrânicos.

Atualmente, em Portugal existem cerca de 260 grandes barragens, definidas de acordo com o estabelecido no Regulamento de Segurança de Barragens em vigor, ou seja, barragens com uma altura superior a 15 m ou barragens com altura compreendida entre os 10 e 15 m em que o volume de armazenamento da albufeira ultrapassa o 1 hm^3. A distribuição das barragens no território é explicada pelas necessidades de usos de água. Em regiões de maior irregularidade de recursos hídricos, em particular no sul e interior do país, onde os aquíferos e os rios não são suficientes para o fornecimento necessário de caudais, a utilização das albufeiras direciona-se maioritariamente para a rega e abastecimento público. No norte do país, onde há uma maior abundância e regularidade de recursos, estas estruturas também têm como fim o aproveitamento hidroelétrico. Embora não evitem a ocorrência de cheias, as barragens apresentam um efeito de atenuação na ocorrência das mesmas[1].

[1] https://apambiente.pt/

A inadequação e insustentabilidade do modo de vida agro-silvo-pastoril, na primeira metade do século XX, a par com os baixos níveis de urbanização e industrialização, levaram a um despovoamento progressivo das áreas rurais, particularmente marcada a partir dos anos sessenta do século passado, que deixa marcas indeléveis no território, no modo como se faz a ocupação do solo e, consequentemente, na paisagem. A emigração para a Europa e as saídas da população activa para os grandes centros urbanos nacionais, mais promissores em termos de oportunidades e qualidade de vida, contribuíram para criar uma dinâmica demográfica regressiva (LNEC, 1990). Neste contexto, tem-se assistido a uma redução dos espaços de uso agrícola, que gradualmente se vão transformando em espaços de matas ou espaços florestais, e um aumento significativo das áreas urbanas, que vão concentrando a população continuamente perdida pelos espaços rurais (M. Pereira et al., 2007).

3.2 Bacia Hidrográfica do Rio Douro

A nascente do rio Douro é situada na serra de Urbion (Sória, Espanha), a cerca de 1700 metros de altitude. Perfaz um curso em torno de 927 km, 322 km em Portugal, acabando por desaguar no oceano atlântico, próximo da cidade do Porto, Figura 4. Trata-se de uma região hidrográfica internacional com uma área total superior a 97 000 km^2, dos quais 18 587 km^2 em território nacional (≈19%), que engloba total ou parcialmente 74 concelhos e cerca de 2 milhões de habitantes (Alcoforado et al., 2021). Apresenta um elevado número de massas de água superficiais, nomeadamente 361 rios, 17 reservatórios, 3 águas de transição e 2 águas costeiras, sendo o terceiro maior rio da Península Ibérica [2].

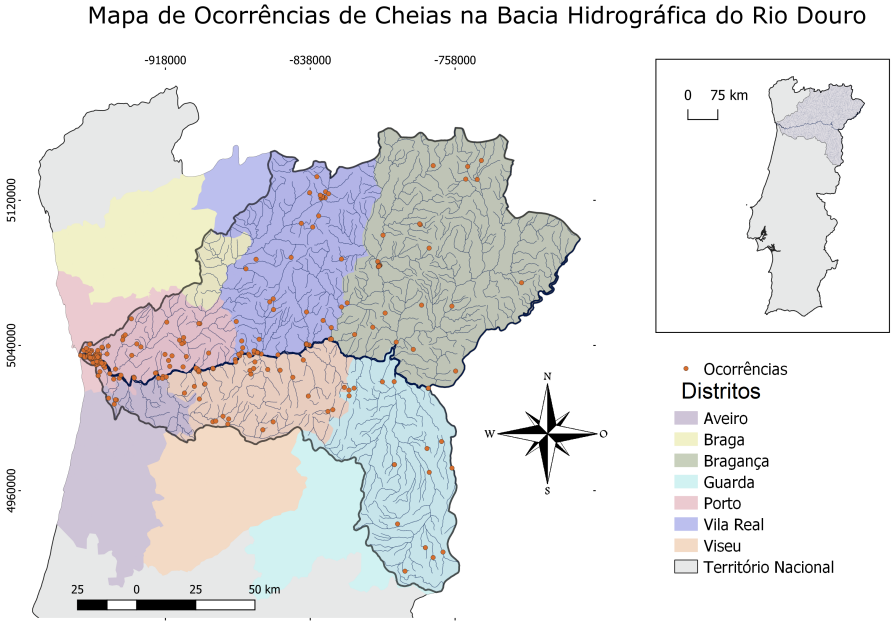

Figura 4: Bacia Hidrográfica do Rio Douro.

[2] apambiente.pt

Em Portugal, o rio Douro desenvolve-se num vale profundo e estreito, e os perfiz longitudinais dos seus afluentes são muito ingremes em algumas secções. Como consequência, os níveis de água tendem a subir de forma mais rápida na sequência de precipitações intensas (Alcoforado et al., 2021; Velhas, 1997). Além disso, e uma vez que a litologia predominante da bacia do Douro é constituída principalmente por rochas metamórficas e graníticas, esta apresenta taxas de infiltração reduzidas, em torno dos 10% (Conceição, 2008). Acabando por sobrecarregar todo o sistema hidrológico, tendo em conta o aumento desproporcional do escoamento.

As variações climáticas, morfológicas e de substrato que caracterizam o território abrangido pela bacia hidrográfica em apreço, permitem que o mesmo atue como suporte para uma elevada diversidade da fauna e flora. Segundo Daveau (Daveau, 1977), a precipitação média anual apresenta uma elevada amplitude de valores registados, variando de 500 mm, na região NE do país, a 2500 mm nas Montanhas a NW, onde flui por exemplo o rio Tâmega. Este contraste de precipitação, deve-se sobretudo ao efeito de relevo da alta cordilheira do noroeste concordante com a linha costeira, que impede que massas de ar húmido atlânticas atinjam o seu lado sotavento (Alcoforado et al., 2021).

Os habitantes da cidade do Porto e da Régua têm estado expostos a eventos de cheias durante séculos, uma vez que estão estabelecidos sobre as margens do rio. Durante o século XVIII, a população do Porto quase duplicou, acabando por apresentar uma maior densidade populacional do que, por exemplo, em Lisboa: 838 habitantes por km^2, face aos 68 habitantes por km^2 (Osswald, 2008).

A tabela 2 expõe as principais características geométricas e de drenagem da bacia hidrográfica do rio Douro.

Tabela 2: Parâmetros Geométricos e de Drenagem da Bacia Hidrográfica do Rio Douro.

Bacia Hidrográfica do Rio Douro	
Área	18600,815 km^2
Perímetro	875,535 km
Raio	76,947 km
Índice de Gravelius	1,811
Fator de Forma	0,411
Densidade de Drenagem	0,56 km^{-1}

O índice de gravelius, $(K_c = \frac{0{,}282P}{\sqrt{A}})$, é um valor que mede a forma da bacia hidrográfica. Traduz a relação entre o perímetro da bacia e o perímetro do círculo de área igual à da bacia e é sempre maior ou igual à unidade, sendo $K_c = 1$ para uma bacia de forma circular. Valores de Kc mais próximos da unidade indicam uma maior propensão para pontas de cheia mais altas na bacia (Hipólito e Vaz, 2011).

O factor de forma, $(K_f = \frac{l}{L} = \frac{A}{L^2})$, é a relação entre a largura média L, e o comprimento da bacia l, definido pelo seu curso de água mais longo. $K_f = 0,411$ indica que a bacia assume um formato irregular, pelo que é menos provável a ocorrência de chuvas intensas que cubram simultaneamente toda a sua extensão.

A densidade de drenagem sobre o terreno, $(\lambda = \frac{1}{A}\sum_i li)$, é também um indicador de tendência para a ocorrência de cheias em bacias hidrográficas. É a relação entre o comprimento total dos cursos de água numa bacia e a área da bacia, expresso em km^{-1}. Trata-se de uma bacia com uma drenagem pobre, $\lambda = 0,56km^{-1}$,

o que significa que a precipitação origina, sobretudo, um escoamento subsuperficial e um escoamento subterrâneo, que se processam de forma mais lenta, originando assim pontas de cheia inferiores (Hipólito e Vaz, 2011).

Resultados e Discussão

Neste capítulo apresentam-se os principais resultados da análise dos dados bem como os modelos ajustados. Procede-se ainda à discussão dos resultados.

4.1 Análise Exploratória dos dados

A análise exploratória das variáveis hidro-meteorológicas baseou-se em 384 registos mensais relativos ao período de Janeiro de 1990 a Dezembro de 2021.

Consideram-se as variáveis *Precipitação Acumulada Mensal*, *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal*.

A precipitação, expressa em milímetros (mm), traduz uma quantidade correspondente a um volume de um litro por um metro quadrado de superfície. Esta quantidade apenas tem significado quando associada a um espaço temporal que, neste caso, é o mês.

A análise descritiva de cada variável inclui, além de representações gráficas da distribuição dos dados, o cálculo de características de localização central, como média e mediana, localização não central, a partir de diversos quantis, dispersão, tal como desvio-padrão e intervalo inter-quartis, assimetria, através de coeficientes de assimetria e curtose, com recurso a coeficientes de achatamento da distribuição dos dados.

Importa ainda efetuar uma análise de respostas omissas e identificar observações aberrantes (`outliers`).

A análise de respostas omissas para as variáveis em estudo encontra-se na Tabela 3.

De notar a falta de disponibilidade de dados para as variáveis hidrométricas, *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal* ambas com elevada percentagem de valores omissos, 26,8% e 40,5% respetivamente.

Tabela 3: Descrição de Valores Omissos para as Variáveis em Estudo.

Processamento de Casos					
		Precipitação Acumulada Mensal (mm)	Caudal Afluente Médio Mensal (m^3/s)	Descarga de Superfície Mensal (dam^3)	Ocorrência de Cheia
N	Válido	351	282	229	384
	Omisso	33	102	155	0

4.1.1 Definição de variáveis categóricas

Define-se a variável *Ocorrência de Cheia* ou *Fator Ocorrência de Cheia*, como uma variável categórica binária (0: "Sem registo de ocorrências na bacia"; 1: "Com registo de pelo menos uma ocorrência na bacia").

Quanto às restantes variáveis em estudo, *Precipitação Acumulada Mensal,* mm , *Caudal Afluente Médio Mensal,* m^3/s, *Descarga de Superfície Mensal,* dam^3 são do tipo quantitativo contínuo. No entanto, por questões metodológicas, de forma a ser possível detetar padrões e obter análises mais interessantes, procedeu-se ainda a uma categorização destas variáveis, tendo como base alguns quantis empíricos. No que se segue, $P50$ representa o quantil 0,5, $P75$ representa o quantil 0,75 e $P90$ representa a diferença entre o quantil 0,9 e o quantil 0,75, $> P90$ representa valores superiores ao quantil 0,9.

$$Precipitação\ Acumulada\ Mensal = \begin{cases} P50,\ se\ Precipitação \leq 49,1\ mm, \\ P90,\ se\ 49,1\ mm < Precipitação \leq 201\ mm \\ > P90,\ se\ Precipitação > 201\ mm \end{cases}$$

$$Caudal\ Afluente\ Médio\ Mensal = \begin{cases} P75,\ se\ Caudal \leq 412\ m^3/s \\ P90,\ se\ 412\ m^3/s < Caudal \leq 906\ m^3/s \\ > P90,\ se\ Caudal > 906\ m^3/s \end{cases}$$

$$Descarga\ de\ Superfície\ Mensal = \begin{cases} Sem\ Descarga \\ Descarga\ inferior\ a\ 2\ milhões\ dam^3 \\ Descarga\ superior\ a\ 2\ milhões\ dam^3 \end{cases}$$

4.1.2 Cheias e Ocorrência de Cheias

A Tabela 4 apresenta a frequência do número de cheias registadas na bacia hidrográfica do Douro, por mês, desde 1990 a 2021. Durante este período, 22 meses registaram a ocorrência de pelo menos uma cheia na bacia hidrográfica, num total de 55 eventos.

Na Figura 5 apresentam-se os valores do número de cheias por mês, de 1990 a 2021.

Note-se que houve um mês, Janeiro de 1996, em que ocorreram 16 cheias. A precipitação acumulada neste mês foi de 394,50 mm, o caudal médio do Carrapatelo era 3.583 m^3/s e descarga à superfície 7.731.420,00 dam^3.

No ano 2001, ocorreram 8 cheias novamente em Janeiro e 4 cheias no mês de Março. Em Janeiro de 2001, a precipitação acumulada foi 429,30 mm e a descarga à superfície 7.947.560,00 dam^3. Em Março de 2001 a

precipitação acumulada foi 662,10 mm e a descarga à superfície 434.680,00 dam^3.

Tabela 4: Número de Cheias Registadas nos 384 meses em estudo.

Número de Cheias Registadas		
N	Frequência Absoluta	Frequência Relativa
0	362	94,3
1	13	3,4
2	4	1
3	2	0,5
4	1	0,3
8	1	0,3
16	1	0,3
Total	384	100

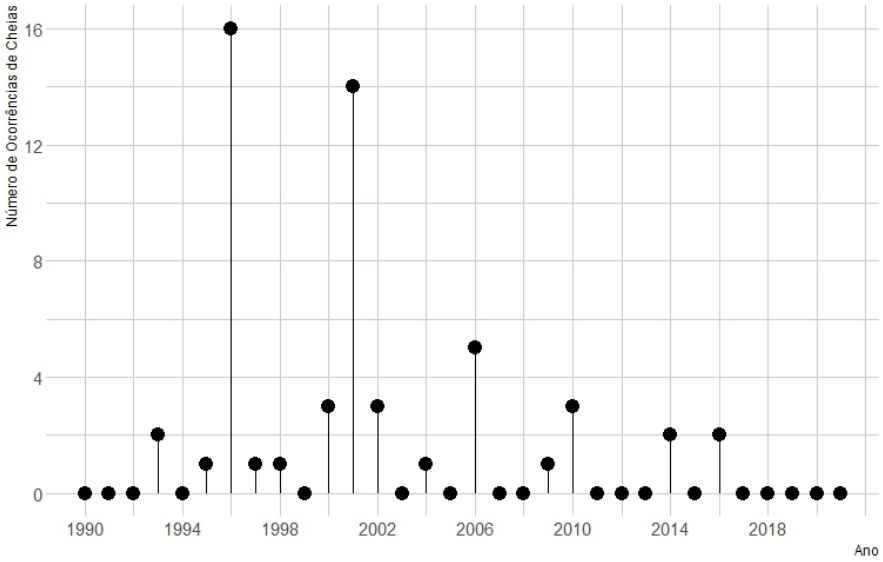

Figura 5: Número de Ocorrências de Cheias por mês, desde 1990 a 2021.

A variável binária *Ocorrência de cheia* definida em 4.1.1, encontra-se correlacionada com o mês (Teste exato de Fisher, p-*value* = 0.0456) com Dezembro, Janeiro e Setembro com uma maior proporção de ocorrência de cheia (12,12%, 12,50% e 15,62%, respetivamente) conforme Figura 6.

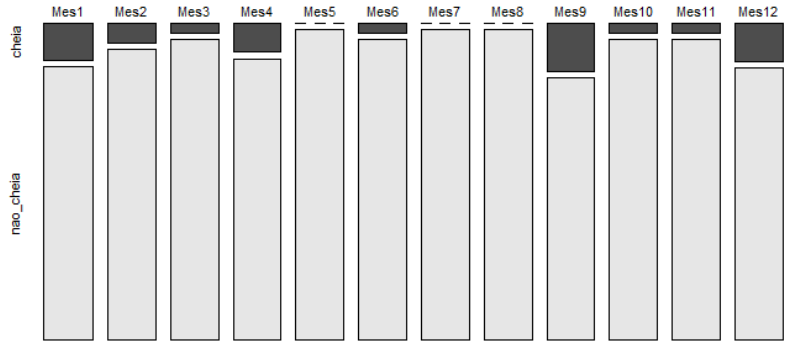

Figura 6: Ocorrência de Cheia vs. Mês.

4.1.3 Precipitação Acumulada Mensal

Na Figura 7 consta a distribuição da variável *Precipitação Acumulada Mensal*. Trata-se de uma distribuição marcadamente assimétrica positiva, indicando grande concentração de observações inferiores à média.

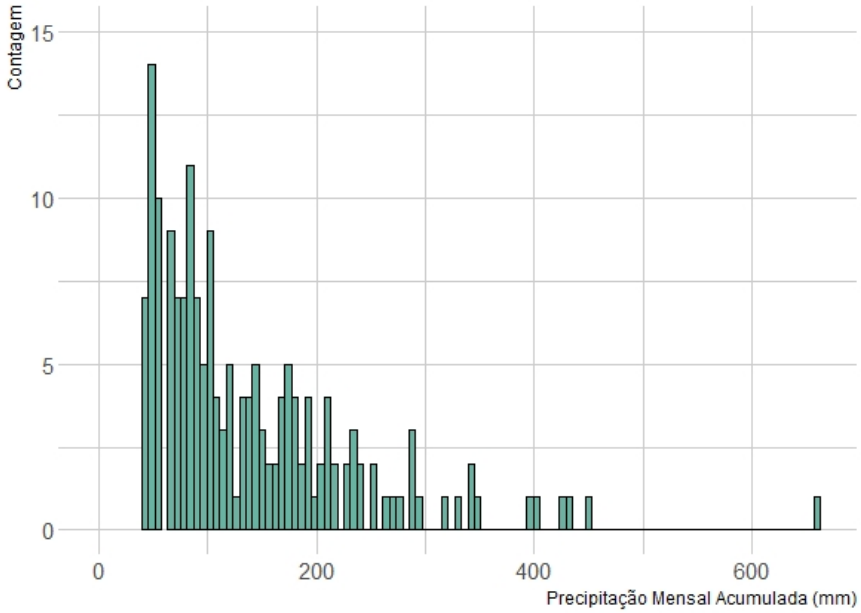

Figura 7: Distribuição da Precipitação Acumulada Mensal.

Os valores das estatísticas sumárias da precipitação acumulada mensal encontram-se na Tabela 5.

Para complementar a caracterização deste conjunto de dados, apresentam-se diagramas de caixa-com-bigodes relativas à distribuição das observações dentro de cada classe de precipitação definida na secção 4.1, Figura 8.

Tabela 5: Estatísticas resumo da Precipitação Acumulada Mensal.

Precipitação Acumulada Mensal (mm)	
Média	79,4
Mediana	49,1
Desvio Padrão	91,6
Mínimo	0
Máximo	662,1
Amplitude	662,1
Amplitude Inter-Quartil	90
Assimetria	2,2
Curtose	6,6

Observa-se a existência de um outlier na classe dos valores superiores ao percentil 90. Esta observação corresponde ao valor de precipitação observado no mês de Março de 2001 em que ocorreram 4 cheias na bacia hidrográfica do Douro, como foi referido anteriormente.

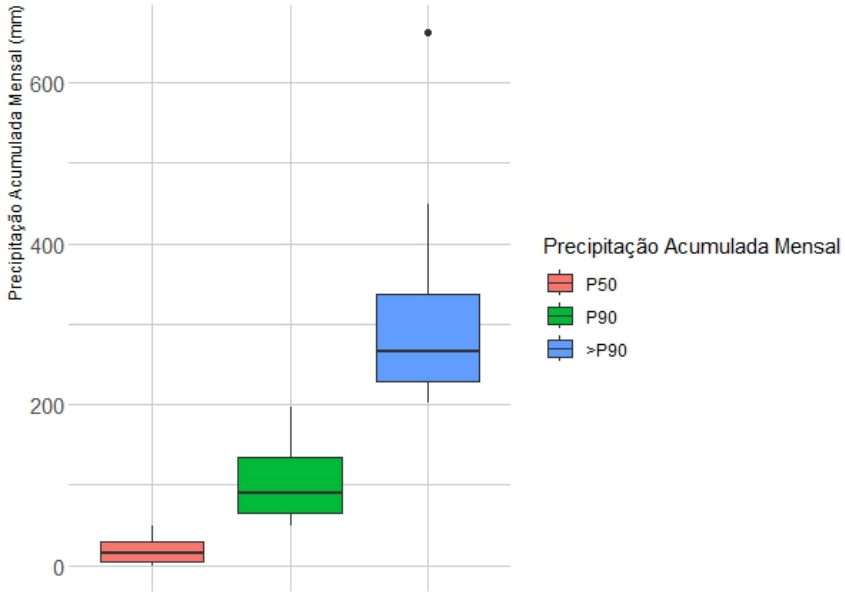

Figura 8: Diagrama de caixa-com-bigodes: Precipitação Acumulada Mensal Categorizada.

Os diagramas de caixa-com-bigodes apresentados na figura 8 revelam apenas 1 outlier, pela forma como foram executados – representação gráfica da distribuição por categoria. Note-se, no entanto, que esta variável apresenta vários `outliers`, mais concretamente 19 observações outlier (correspondente a cerca de 5% dos casos).

a) Análise por Mês

Na Figura 9 encontram-se representadas as distribuições dos valores de precipitação por mês do ano hidrológico, desde 1990 até 2021. Assinalam-se ainda os `outliers` (pontos vermelhos) e as médias observadas em cada mês (pontos pretos).

Os valores mais elevados de precipitação ocorrem de Outubro a Janeiro, com medianas superiores a 100 mm. De Fevereiro a Abril existem ainda observações superiores a 200 mm, embora as medianas estejam agora bastante abaixo de 100 mm. Nestes meses, com exceção de Fevereiro, as distribuições apresentam menos variabilidade que nos meses referidos anteriormente. Os meses de Junho a Agosto apresentam valores de precipitação bastante mais baixos com muito menos variabilidade, especialmente o mês de Julho, o mês mais seco, com uma média de precipitação de apenas 13 mm.

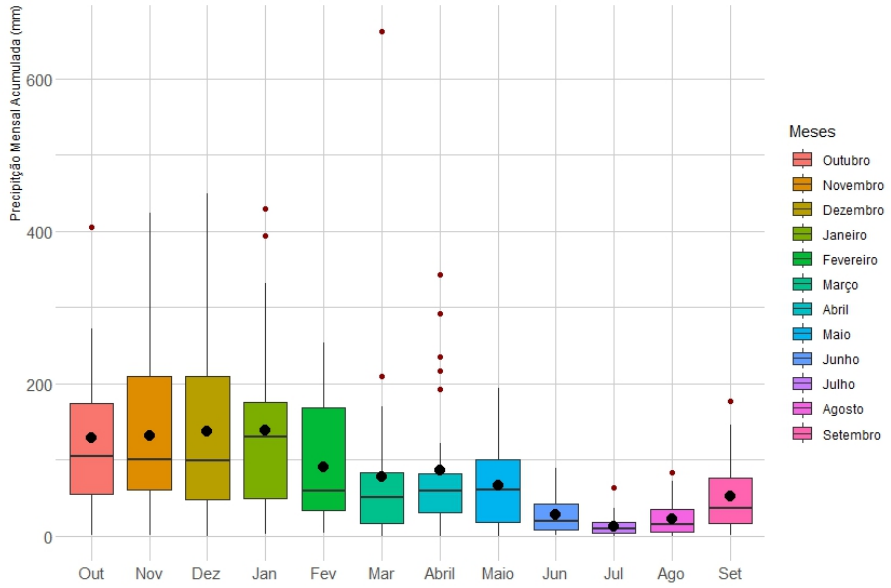

Figura 9: Precipitação Acumulada Mensal em Função do Mês.

b) Análise por Ano

A Figura 10 apresenta a evolução da média mensal de precipitação acumulada ao longo dos anos em estudo.

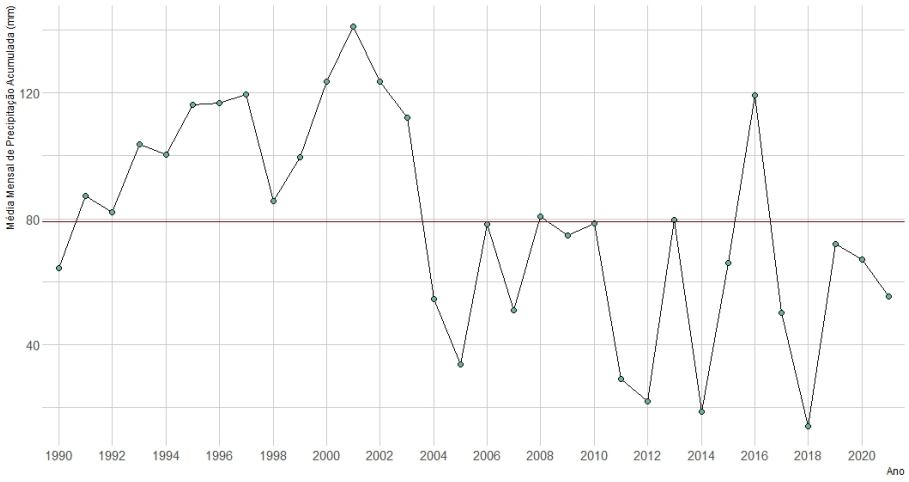

Figura 10: Evolução da Média Mensal de Precipitação Acumulada de 1990 a 2021.

Nesta representação gráfica, a reta horizontal, a vermelho, permite perceber de que forma variam os valores das médias da precipitação acumulada por mês, em cada ano, em torno da sua média global, constante na Tabela 4.

Depois de 13 anos com valores da precipitação acumulada mensal acima da média, seguem-se cerca de 18 anos com valores abaixo, ou iguais, à média global, com exceção de 2016, onde a média das precipitações acumuladas mensais é 120 mm. Neste ano ocorreu uma cheia em Janeiro e outra em Abril.

Na Figura 11 encontram-se representadas as distribuições dos valores de precipitação por ano, desde 1990 a 2021. Assinalam-se ainda os `outliers` para cada ano.

Relativamente ao comportamento dos dados, verificam-se diferenças nas distribuições em cada ano, quer em termos de localização quer em termos de dispersão (executado um teste de Kruskal-Wallis de igualdade de distribuições, obtém-se um p-*value* $= 0{,}002$, muito significativo).

Desde 2010, observa-se um padrão ainda mais complexo no que toca à variabilidade inter-anual de precipitação, mais evidente da variabilidade detetada nos anos anteriores. Como já foi referido, esta variabilidade tem sido justificada pelo enquadramento mediterrânico da bacia hidrográfica (Nunes et al., 2016).

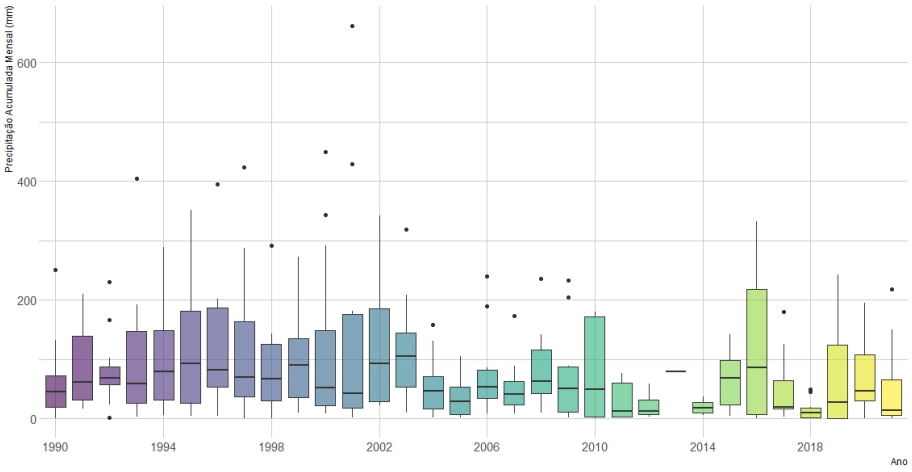

Figura 11: Distribuição da Precipitação Acumulada Mensal por Ano.

c) Análise por Fator Ocorrência de Cheia

Os diagramas de caixa-com-bigodes, Figura 12, representam a distribuição empírica da precipitação mensal acumulada, de acordo com os níveis do *Fator Ocorrência de Cheia*, ou seja, quando não se registou qualquer ocorrência de cheia e quando ocorreu pelo menos uma cheia.

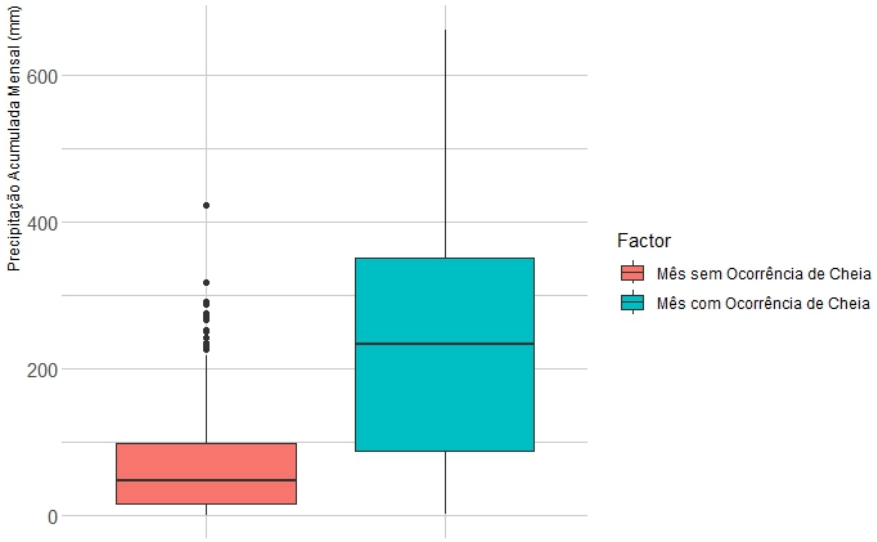

Figura 12: Distribuições da Precipitação Acumulada Mensal por Fator Ocorrência de Cheia.

Nos meses em que ocorreu pelo menos uma cheia, a mediana da precipitação acumulada mensal, 233 mm, é quase 5 vezes superior à relativa aos meses onde não houve registo de cheias (42,5 mm). Para além da diferença clara no centro das distribuições, estas também apresentam diferenças ao nível da dispersão. Nos meses em

que não ocorreu cheia, verifica-se uma amplitude de 423,9 mm e uma amplitude interquartil de 83,08 mm face a 661,10 mm e 307,65 mm, respetivamente, para os meses onde ocorreu pelo menos uma cheia.

De facto, executado um teste de igualdade de distribuições de Wilcoxon-Mann-Whitney, obtém-se uma estatística de teste U = 5426,5, com um p-*value* associado $< 0{,}001$, pelo que as distribuições da precipitação acumulada mensal é diferente nas duas categorias de cheia.

Para além deste facto, importa referir que a variável *Precipitação Acumulada Mensal*, com categorias definidas em 4.1.1, se encontra correlacionada com o *Fator Ocorrência de Cheia*, Tabela 6 e Gráfico 13.

Tabela 6: Precipitação Acumulada Mensal Categorizada vs. Fator Ocorrência de Cheia.

		Precipitação Acumulada Mensal Categorizada			
		P50	P90	>P90	Total
Ocorrência de Cheia	0	171	136	23	330
	1	5	4	12	21
	Total	176	140	35	351

Executado o teste exato de Fisher, obtem-se um p-*value* $< 0{,}001$.

Conclui-se, então, que a variável *Precipitação Acumulada Mensal* é importante na explicação de ocorrência de cheia, sem surpresa. No entanto, a variabilidade observada na distribuição desta variável, no caso em que ocorre cheia, bastante superior à que observa quando não ocorre cheia, é certamente justificada pelo facto desta não ser, por si só, suficiente para explicar o fenómeno.

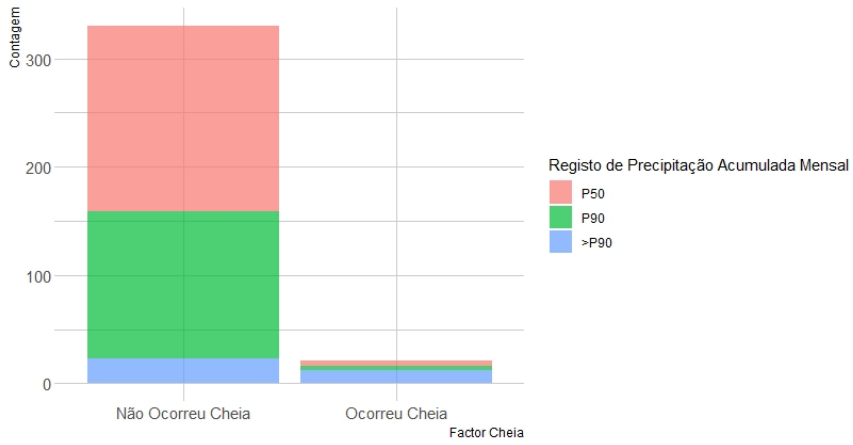

Figura 13: Distribuição da Precipitação Acumulada Mensal Categorizada por Fator Ocorrência de Cheia.

4.1.4 Caudal Afluente Médio Mensal e Descarga de Superfície Mensal

A importância das variáveis *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal* como potenciais preditores da ocorrência de cheia, é justificada pelo que se expõe de seguida.

Em recursos hídricos, o volume de água que atravessa determinada secção transversal de um curso de água

num dado intervalo de tempo designa-se por escoamento. O caudal médio em determinado intervalo de tempo é a razão entre escoamento, expresso em volume, e o intervalo de tempo durante o qual ocorreu, sendo expresso por m^3/s.

O escoamento numa secção transversal de um curso de água é variável ao longo do tempo, podendo ser muito elevado, após precipitações intensas precedidas de períodos chuvosos de tempo, e podendo ser muito baixo ou nulo, após períodos sem precipitação ou com precipitação reduzida (Hipólito e Vaz, 2011).

As estatísticas sumárias para as variáveis *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal* encontram-se na Tabela 7.

Na Figura 14 consta a distribuição da variável *Caudal Afluente Médio Mensal*. Trata-se de uma distribuição assimétrica positiva.

Na Figura 15 e na Figura 16 apresentam-se os diagramas de caixa-com-bigodes para representar a distribuição destas variáveis, em classes definidas a partir de percentis adequados.

Observa-se que metade das observações do caudal afluente médio mensal não ultrapassaram os 500 m^3/s e que 90% das observações são inferiores a 900 m^3/s (ver Figura 15). A classe >P90, com as 10% observações mais elevadas, apresenta uma maior dispersão em relação às restantes. De facto, esta compreende observações no intervalo de 900 m^3/s até 3586 m^3/s, sendo este último um outlier severo, dentro desta classe, registado em Janeiro de 1996, mês onde ocorreram 16 cheias ao longo da bacia hidrográfica do Douro.

Tabela 7: Estatísticas Sumárias: Variáveis Caudal Afluente Médio Mensal e Descarga de Superfície Mensal.

	Regíme Hídrico	
	Caudal Afluente Médio Mensal (m^3/s)	Descarga de Superfície Mensal (dam^3)
Média	373,8	262854
Mediana	220,6	0
Desvio Padrão	413,6	1055163
Mínimo	12,8	0
Máximo	3583	7947560
Amplitude	3570,2	7947560
Amplitude Inter-Quartil	261,1	0
Assimetria	3,2	5,8
Curtose	15,5	36,7

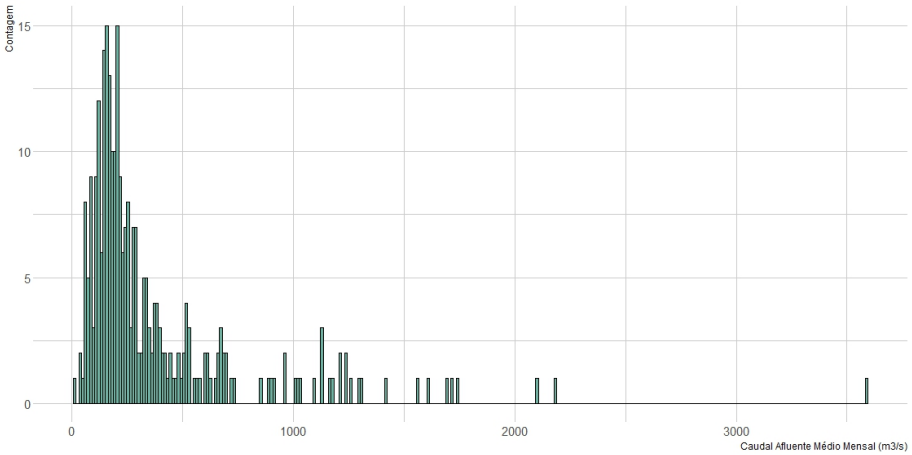

Figura 14: Distribuição do Caudal Afluente Médio Mensal.

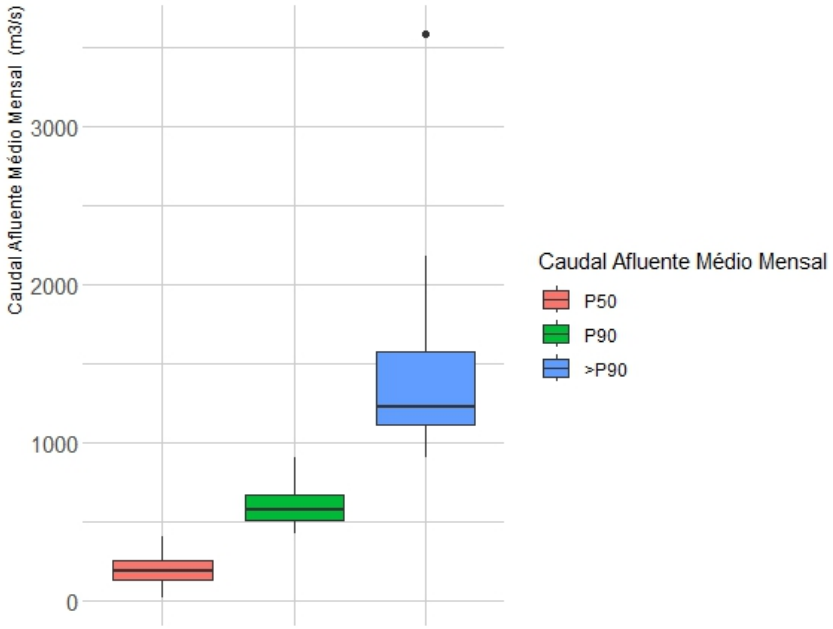

Figura 15: Diagrama de caixa-com-bigodes da Variável Caudal Afluente Médio Mensal.

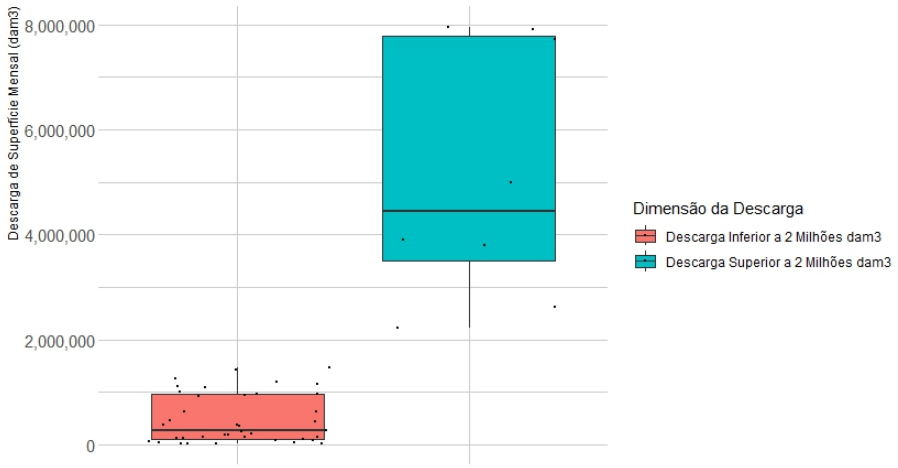

Figura 16: Diagrama de caixa-com-bigodes da Variável Descarga de Superfície Mensal.

Estas variáveis encontram-se, naturalmente, bastante correlacionadas, apresentando um valor do coeficiente de correlação de Spearman positivo muito significativo (0,645), Figura 17

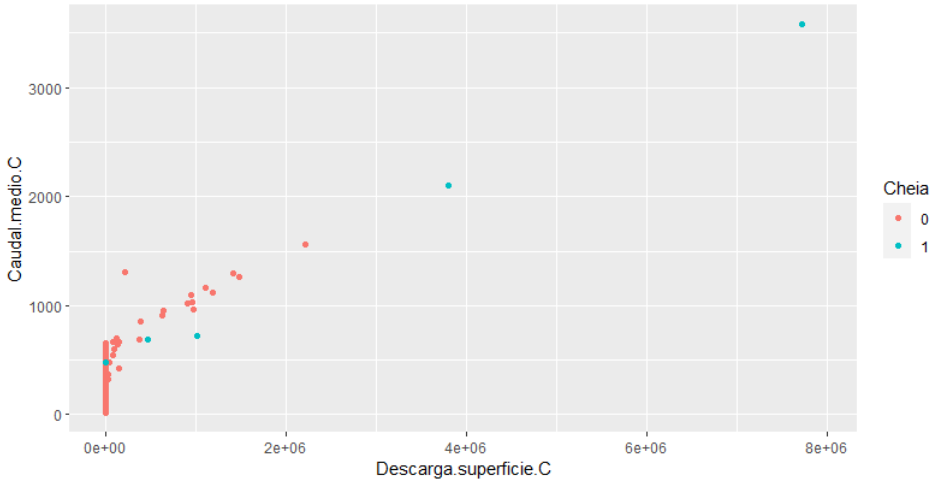

Figura 17: Caudal Afluente Médio vs. Descarga de Superfície.

a) Análise por Ano

A Figura 18 apresenta a evolução das médias dos valores do caudal afluente médio em cada ano, no período de 1990 a 2021 (designado por hidrogama), assim como com os valores da precipitação acumulada por mês correspondentes ao ano.

Na figura 19 encontram-se as distribuições dos valores de caudal por ano, desde 1990 até 2021. Assinalam-se ainda os `outliers`.

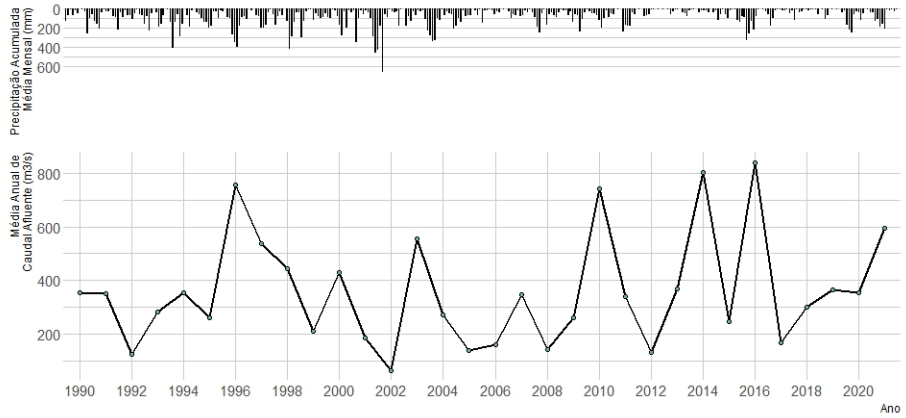

Figura 18: Hidrograma do Caudal Afluente Médio por ano e Precipitação Acumulada Média por mês.

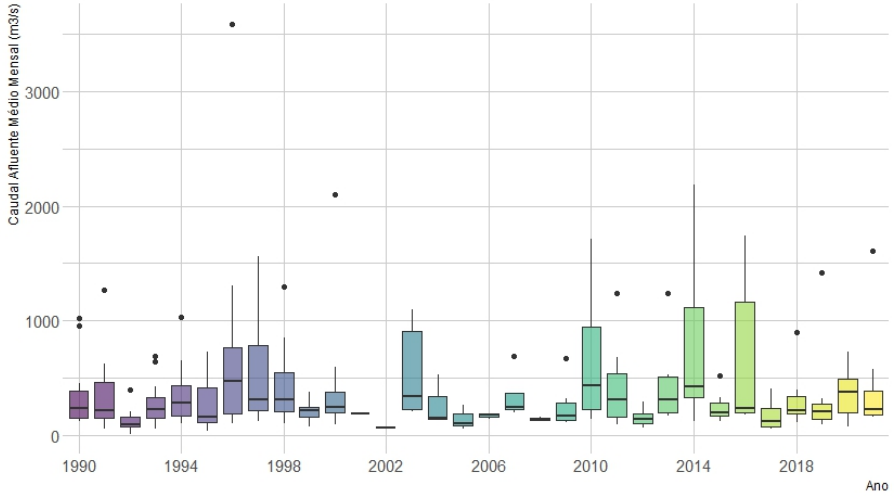

Figura 19: Distribuição de Caudal Afluente Médio Mensal por Ano.

Os valores mensais do caudal médio e da precipitação média encontram-se positivamente associados (com valor do coeficiente de correlação de Spearman 0,421 muito significativo). Genericamente, os valores mais elevados da distribuição dos valores da precipitação estão em linha com os valores mais elevados do caudal. No entanto, nos anos de 1990 e 1991, embora haja um aumento nas médias da precipitação mensal acumulada, este aumento não se reflete no valor médio de caudal, no ano de 1991, que se mantém constante. No período entre 2007 e 2008, dá-se uma queda das médias do caudal, embora os valores médios de precipitação não sofram alterações.

No período entre 2001 e 2005 existem muitas observações omissas no Caudal Afluente Mensal. No entanto, no final do ano de 2001 e ínicio de 2002 apontam os valores de precipitação mais elevados para todo o período em estudo, o que não se reflete nos valores de caudal que registam mínimos históricos. Note-se que nos anos 2001 e 2002 existiram, respetivamente, 15 e 3 eventos de cheia.

b) Análise por Fator Ocorrência de Cheia

Os diagramas de caixa-com-bigodes, Figuras 20 e 21, representam as distribuições observadas de caudal afluente médio mensal e descarga de superfície mensal, na devida ordem, de acordo com os níveis do *Fator Ocorrência de Cheia*.

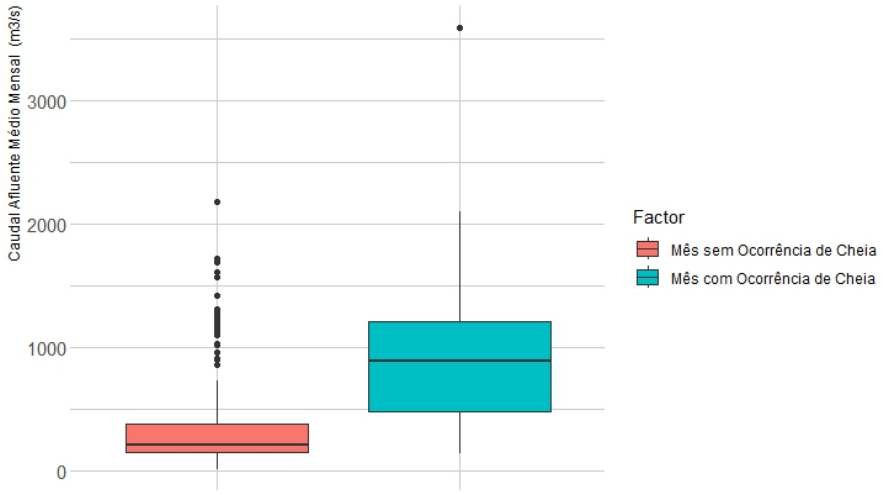

Figura 20: Distribuições de Caudal Afluente Médio Mensal por Fator Ocorrência de Cheia.

Nos meses em que ocorreu pelo menos uma cheia na bacia hidrográfica, a mediana de caudal afluente médio mensal foi 4 vezes superior em relação ao meses sem ocorrência de cheia (889 m^3/s em relação a 212 m^3/s). As distribuições também apresentam diferenças quanto à sua dispersão. Amplitude de 2180 m3/s e amplitude interquartil 233m3/s, para os meses sem registo de ocorrência de cheia, face a uma amplitude de 3444 m3/s e amplitude interquartil 1111m3/s.

Mais de 80% dos meses sem registo de ocorrência de cheia, não foram sujeitos a descarga de superfície. Para estes meses a descarga de maior dimensão é de 3905911 dam^3, sendo este valor um outlier severo. Para os meses com ocorrência de pelo menos uma cheia na bacia hidrográfica, apenas 25% não foram sujeitos a descargas de superfície. Para estes meses, metade das descargas apresentam uma dimensão superior a 1000000 dam^3.

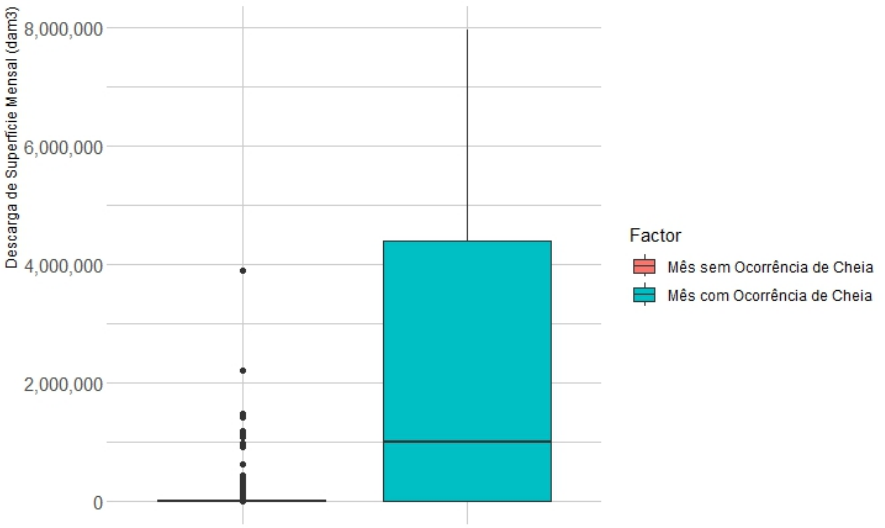

Figura 21: Distribuições de Descarga de Superfície Mensal por Fator Ocorrência de Cheia.

Quase metade dos meses com ocorrência de pelo menos uma cheia registam valores de caudal afluente médio superiores a 900 m^3/s.

Dos 8 meses com descargas de superfície superiores a 2 milhões dam^3, 6 registaram pelo menos uma ocorrência de cheia na bacia.

Executados testes de igualdade de distribuições de Wilcoxon-Mann-Whitney, para as variáveis *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal* nas duas classes de *Fator Ocorrência de Cheia*, conclui-se que estas são significativamente diferentes nos dois casos (valores da estatística de teste U = 2885 e U = 2426 , respetivamente, p-*value* $<$ 0,001).

Considerando a versão categorizada das variáveis *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal*, obtêm-se correlações estatisticamente significativas com a variável associada à ocorrência de cheia, Tabelas 8 e 9 e Figuras 22 e 23, respetivamente (Teste exato de Fisher com p-*value* $<$ 0,001 em ambos os casos).

Conclui-se que *Caudal Afluente Médio Mensal* e *Descarga de Superfície Mensal* são variavéis importantes para a explicação de ocorrência de cheia.

Tabela 8: Caudal Afluente Médio Mensal Categorizado vs. Fator Ocorrência de Cheia.

Caudal Afluente Médio Mensal Categorizado					
		P75	P90	>P90	Total
Ocorrência de Cheia	0	209	38	22	269
	1	3	4	6	13
	Total	212	42	28	282

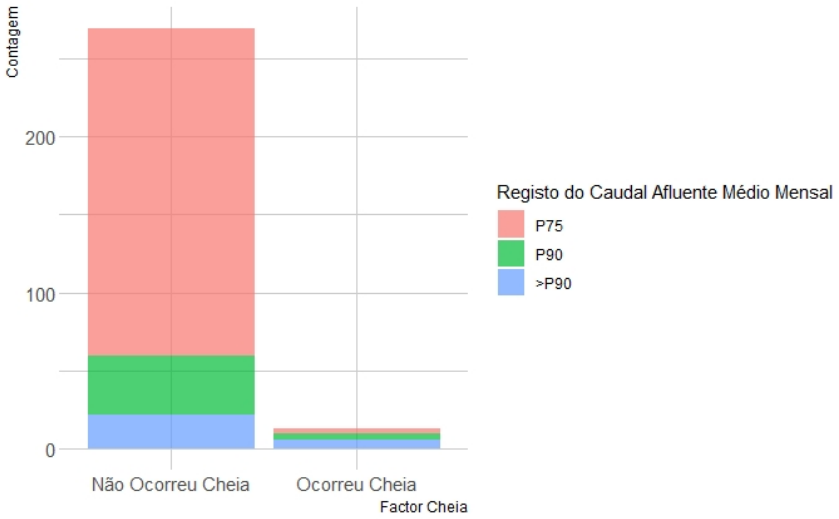

Figura 22: Caudal Afluente Médio Mensal categorizado por Fator Ocorrência de Cheia.

Tabela 9: Descarga de Superfície Mensal categorizada vs. Fator Ocorrência de Cheia.

		Descarga de Superfície Mensal Categorizada			
		Sem Descarga	Descarga <2 milhões dam3	Descarga >2 milhões dam3	Total
Ocorrência de Cheia	0	176	36	2	214
	1	6	3	6	15
	Total	182	39	8	229

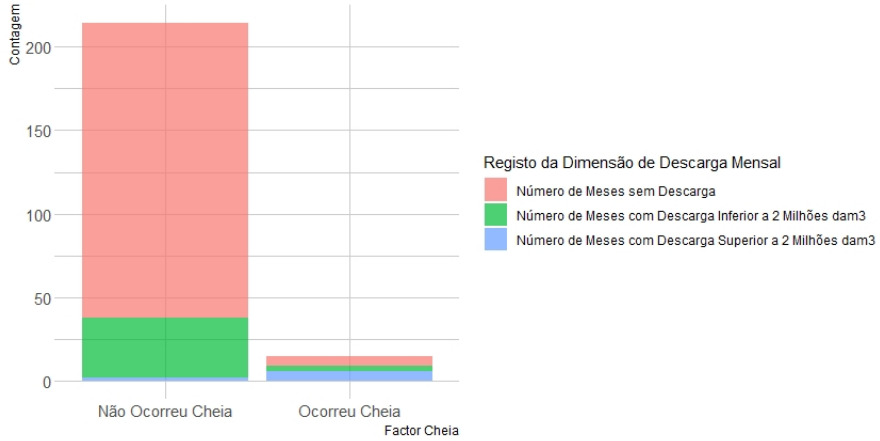

Figura 23: Descarga de Superfície Mensal categorizada em Função por Fator Ocorrência de Cheia.

c) Ocorrência de Cheia e Precipitação

Com o objetivo de investigar a existência de interações significativas entre as variáveis explicativas e ocorrência de cheia, tenta-se agora perceber se as relações entre os valores mensais do caudal médio e da descarga com os

valores da precipitação média são afetadas pela ocorrência de cheia.

As Figuras 24 e 25 apresentam a relação das distribuições de caudal afluente médio mensal e descarga de superfície mensal com a distribuição de precipitação acumulada mensal, respetivamente, considerando os dois níveis do *Fator Ocorrência de Cheia*.

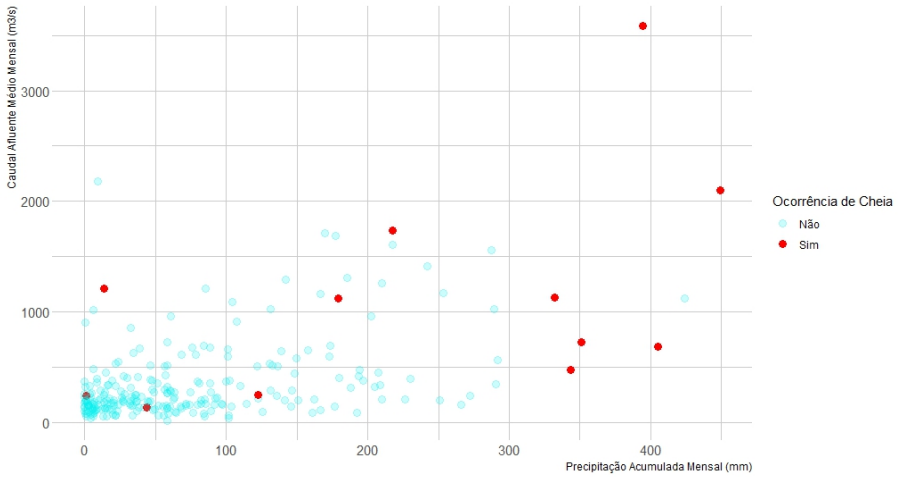

Figura 24: Caudal Afluente Médio Mensal vs. Precipitação Acumulada Mensal por Fator Ocorrência de Cheia.

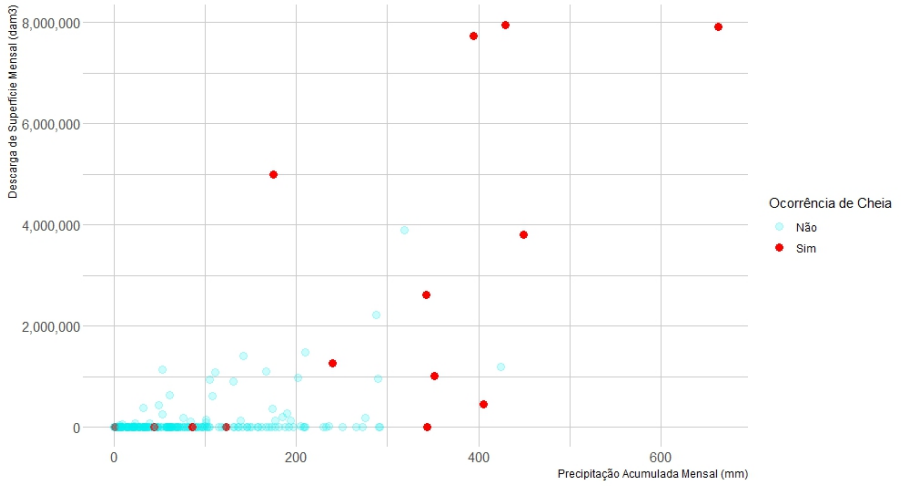

Figura 25: Descarga de Superfície Mensal vs Precipitação Acumulada Mensal por Fator Ocorrência de Cheia.

Quer na relação entre caudal afluente médio mensal e precipitação acumulada mensal, quer na relação entre descarga de superfície mensal e precipitação acumulada mensal, os valores do coeficiente de correlação de Spearman passam de 0,385 e 0,335, respetivamente, no caso em que não há ocorrência de cheia para 0,503 e 0,756, respetivamente, no caso em há ocorrência de cheia. Note-se, no entanto, que o número de observações disponíveis no caso em que houve ocorrência de cheia é muito baixo (tem-se apenas 12 observações no cruzamento de precipitação acumulada mensal com caudal afluente médio mensal e 15 no caso do cruzamento

de precipitação acumulada mensal com descarga de superfície mensal), pelo que estes resultados podem ser inconclusivos. De facto, não se deteta uma associação significativa entre caudal afluente médio mensal e precipitação acumulada mensal no caso da ocorrência de cheia.

4.2 Modelos Propostos

Nesta secção consideram-se os preditores analisados anteriormente em modelos de previsão e explicação de ocorrência de cheia.

De acordo com a análise univariada efetuada nos sub-capítulos anteriores, conclui-se que as variáveis *Precipitação Acumulada Mensal*, *Descarga de Superfície Mensal* e *Caudal Afluente Médio Mensal* são estatisticamente significativas, isto é, influenciam a ocorrência de cheias, de forma independente umas das outras. Quando se incorporam estes preditores num modelo, uma ou mais variáveis podem deixar de ser significativas na explicação do fenómeno, uma vez que o seu efeito é agora condicionado ao efeito das restantes variáveis.

A execução dos modelos que se seguem, em software R, encontra-se disponível no apêndice .1.

4.2.1 Modelo de Regressão Logística

O modelo de regressão logística apresentado é construído com base nos preditores categóricos, uma vez que a versão não categorizada destes preditores apresentam distribuições muito assimétricas positivas.

O processo de construção do modelo de regressão logística inicia-se pela codificação das variáveis categóricas, definindo-se a classe de referência, sendo, neste caso, a última classe, relativa aos valores extremos das variáveis. A Tabela 10 discrimina as variáveis incluídas no modelo com melhor ajustamento aos dados. São apresentados os coeficientes estimados bem como o erro padrão associado. A estatística de Wald e a significância correspondente indicam a significância do preditor e, por último, para efeitos de interpretação, apresentam-se os valores de $\exp(\beta)$.

A partir das estimativas dos coeficientes do modelo são determinadas as estimativas em termos de odds ratio e os respetivos intervalos de confiança, aplicando a função exponencial ao valor do coeficiente e aos extremos de intervalo de confiança associado a cada variável. Sempre que se verificar um valor de odds ratio superior a 1, existe um aumento de risco de ocorrência de cheia, quando comparado com a classe de referência. Pelas estimativas apresentadas, verifica-se que todas as variáveis diminuem o risco de ocorrência de cheia em relação à classe de referência que, em todos os casos, se refere aos valores mais elevados das variáveis em causa, como já foi referido.

Ao avaliar os resultados associados à variável *Precipitação Acumulada Mensal*, constata-se que o risco de ocorrência de cheia tem tendência a diminuir à medida que os valores de precipitação decrescem. Para os meses com registo de precipitação inferior a 49,1 mm (P50), o risco de ocorrência de cheia reduz cerca de 84% quando comparado com os valores 10% mais elevados desta variável (este valor é significativo, p-*value* = 0,025).

Relativamente à variável *Descarga de Superfície Mensal*, quando há descarga, o risco de ocorrência de cheia reduz cerca de 94% quando comparado com descargas superiores a 2 milhões dam^3 (este valor é muito significativo, p-*value* = 0,002).

Tabela 10: Sumário do Modelo de Regressão Logística.

Sumário do Modelo								
							I.C. 95 % exp(β)	
	β	S.E.	Wald	df	Valor de Prova	exp(β)	Inferior	Superior
Precipitação			5,598	2	,061			
Precipitação (P90)	-1,597	,894	3,188	1	,074	,203	,035	1,169
Precipitação (P50)	-1,860	,830	8,029	1	,025	,156	,031	,791
Descarga			9,769	2	,008			
Descarga (<2 M)	-3,275	1,072	9,341	1	,002	,038	,005	,309
Descarga (= 0)	-2,832	1,075	6,937	1	,008	,059	,007	,485
Constante	1,398	,873	2,564	1	,109	4,049		

Utilizando as estimativas dos coeficientes apresentados na Tabela 10, a equação do modelo:

$$\pi(z) = \frac{\exp(z)}{1 + \exp(z)}$$

onde,

$$z = 1,4 - 1,6 \times Precipitação(P90) - 1,9 \times Precipitação(P50) - 3,3 \times Descarga(< 2M) - 2,8 \times Descarga(= 0)$$

O ajustamento do modelo é bom, conforme Tabela 11 com os resultados do teste de Hosmer e Lemeshow (p-*value* = 0,656). Recorde-se que a hipótese nula do teste de Hosmer and Lemeshow é que o modelo de regressão logística se ajusta bem aos dados, ou seja, não há diferença entre a distribuição observada dos eventos e a distribuição prevista pelo modelo de regressão logística.

A tabela de classificação para o modelo proposto, Tabela 12, contém os valores da sensibilidade (taxa de verdadeiros positivos), especificidade (taxa de verdadeiros negativos) e taxa de acerto global do modelo. Os valores obtidos são, respetivamente, 33,3%, 99,1% e 94,7%.

O modelo apresenta um desempenho excelente quanto à classificação de meses sem ocorrência de cheia, pelo que se conclui que os dois preditores utilizados são suficientes para explicar a não ocorrência de cheia. Quanto à previsão do evento de interesse, meses com ocorrência de cheia, embora se trate de um modelo aceitável,

Tabela 11: Teste de Ajustamento de Hosmer e Lemeshow.

Teste de Hosmer e Lemeshow		
Qui-Quadrado	df	Significância
,842	2	,656

Tabela 12: Tabela de Classificação para o Modelo de Regressão Logística Proposto.

Tabela de Classificação				
		Previsto		
		Ocorrência de Cheia		Percentagem
Observado		0	1	Correcta
Ocorrência	0	209	2	99,1
de Cheia	1	10	5	33,3
Percentagem Correcta Global				94,7

apenas classifica corretamente um terço das observações. Note-se que o número de casos de ocorrência de cheia é bastante baixo, o que só por si justifica este valor. No entanto, a inclusão de mais preditores pode incrementar este valor, explicando melhor este fenómeno obtendo-se, consequentemente, uma maior taxa de sucesso.

Finalmente, a análise da curva ROC é uma boa ferramenta para verificar o poder discriminatório do modelo, ilustrando a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 - especificidade) para diferentes limiares de classificação (cut-off), Figura 26.

No caso em estudo, a área sob a curva é de 0,768 o que indica que o modelo têm uma capacidade de discriminação aceitável.

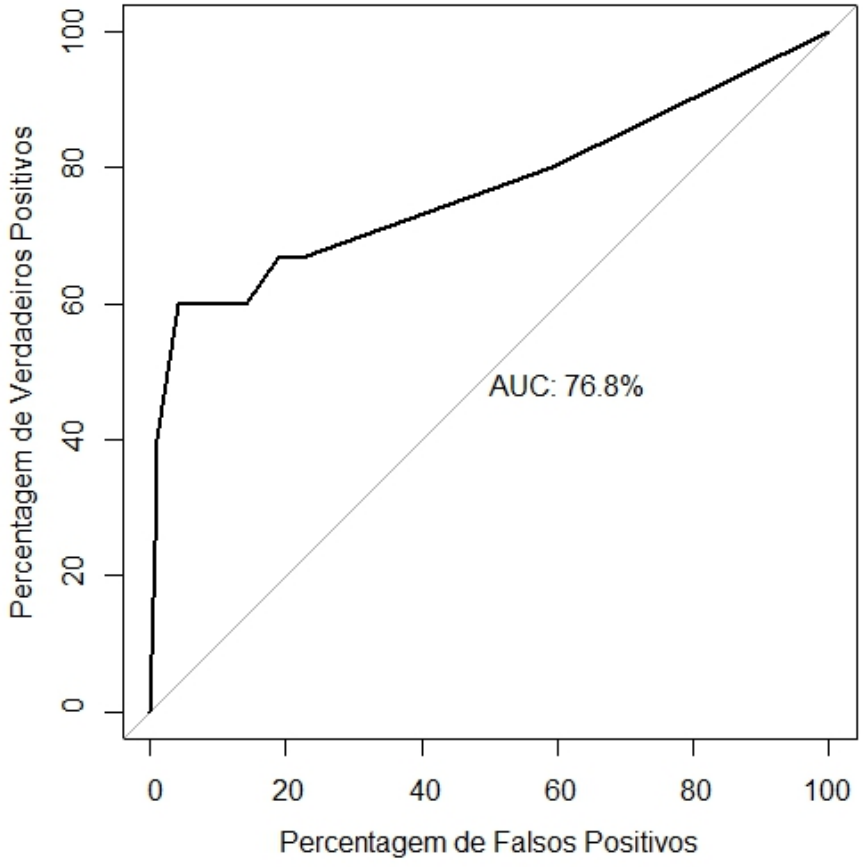

Figura 26: Curva ROC.

4.2.2 Florestas Aleatórias

O algoritmo de floresta aleatória foi utilizado com o objetivo de se tentar perceber se é possível obter melhor classificação de casos de cheia, usando os preditores importantes no modelo de regressão logística, não categorizados. De facto, as árvores de decisão e as florestas aleatórias são menos sensíveis à distribuição dos dados do que os modelos de regressão logística, uma vez que não pressupõem uma distribuição específica para os dados.

Como já foi referido, o algoritmo de floresta aleatória gera múltiplas amostras *bootstrap* e OOB (Out Of Bag) com a finalidade de aferir a qualidade do modelo.

A Figura 27, representa a variação do erro associado às amostras OOB, bem como a variação do erro de classificação de ocorrências e não ocorrências de cheias em função do número de árvores utilizadas no modelo. Verifica-se que não há melhorias significativas na qualidade do modelo após 100 árvores.

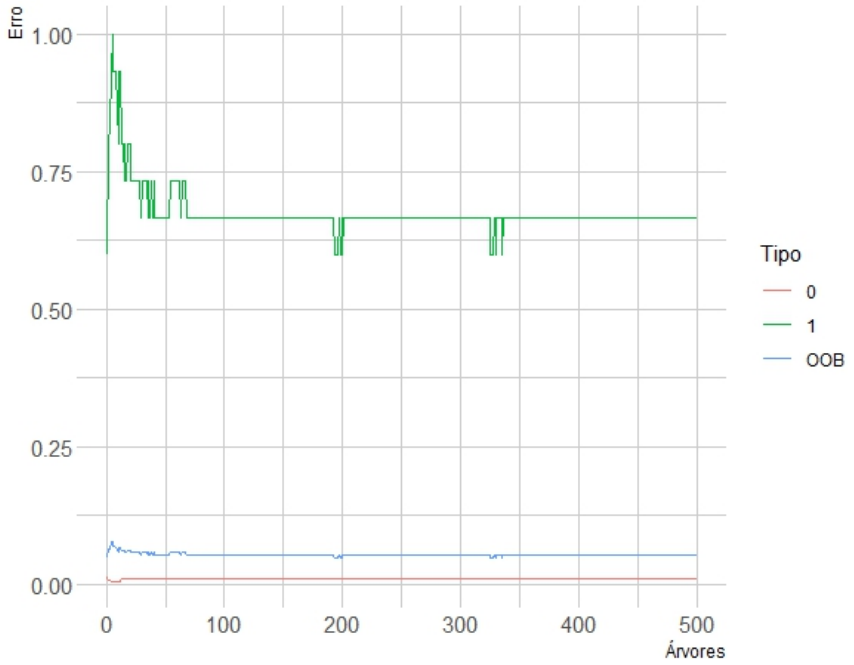

Figura 27: Variação do erro OOB (Out Of Bag) e do erro de classificação para cada evento.

Implementado o método nestas condições, os resultados são bastante melhores.

A importância das variáveis pode ser avaliada de formas diferentes e complementares.

O software R devolve diretamente dois indicadores: incremento de pureza e percentagem de incremento de MSE.

Um valor de incremento de pureza de nó associado a uma variável em num algoritmo de florestas aleatórias indica a capacidade da variável em separar as amostras em classes diferentes. Quanto maior o incremento de

Tabela 13: Tabela de Classificação para o Modelo de Florestas Aleatórias Proposto.

Matriz de Classificação				
		Previsto		
		Ocorrência de Cheia		Percentagem Correcta
Observado		0	1	
Ocorrência de Cheia	0	210	1	99,52
	1	6	9	60,00
Acurácia				96,90

pureza, maior a importância da variável na classificação ou previsão, pelo que a variável *Precipitação Acumulada Mensal* é mais importante na discriminação das duas classes de Cheia, Figura 28. Por outro lado, um valor menor de MSE indica que o modelo está se ajustar melhor aos dados de treino, pelo que quanto menor o incremento de MSE, melhor é o modelo. Neste caso, é também a variável *Precipitação Acumulada Mensal* que tem associada uma maior percentagem de incremento do MSE, Figura 29.

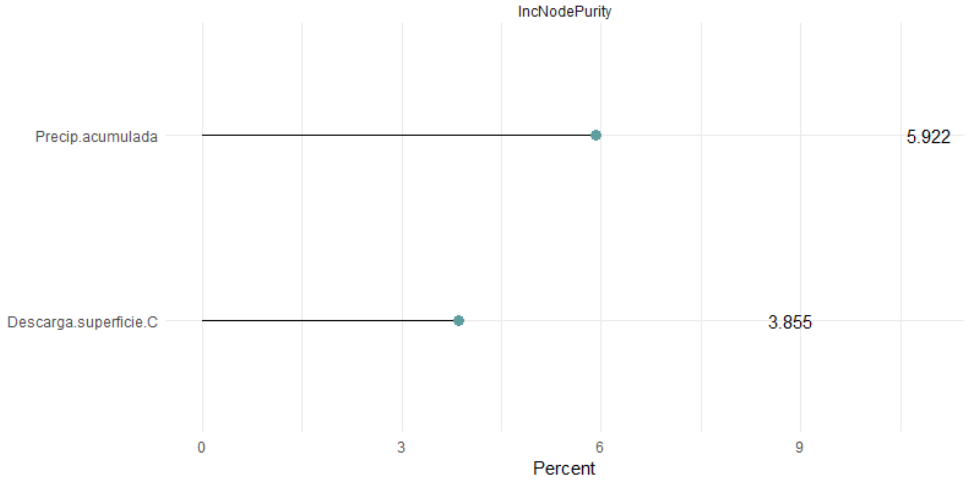

Figura 28: Incremento Pureza Nó.

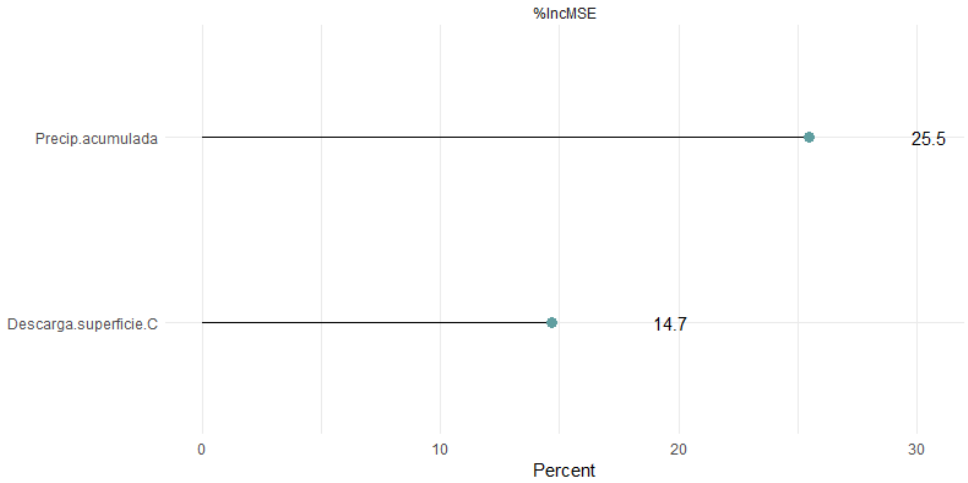

Figura 29: Percentagemn de Incremento de MSE.

Em conclusão, os dois preditores permitem classificar corretamente 60% dos casos em que ocorreu pelo menos uma cheia na bacia hidrográfica do rio Douro e 99,52% dos casos em que não ocorreu qualquer cheia.

A acurácia global 96,9% indica que o modelo possui um ótimo desempenho global, visto que tem uma taxa de acerto de 97% . A especificidade de 99,52% indica que o modelo tem um excelente desempenho na identificação da não ocorrência de cheia, o que atribuí uma enorme credibilidade às previsões de não cheia. De um modo geral, o modelo apresenta muito boas características.

Conclusões

A análise exploratória efetuada, dando especial ênfase à visualização, permitiu não só organizar e resumir o conjunto de dados, como também detetar padrões importantes.

Os resultados da análise permitem concluir que cada uma as variáveis *Precipitação Acumulada Mensal*, *Descarga de Superfície Mensal* e *Caudal Afluente Médio Mensal*, apresenta uma associação estatisticamente significativa com o fenómeno de ocorrência de cheia, sendo relevantes para a previsão de cheias. As associações encontradas entre estas variáveis, estão de acordo e reforçam vários resultados já descritos na literatura da área.

Concluiu-se ainda que a variável *Ocorrência de cheia*, se encontra correlacionada com o mês. De acordo com os dados, Dezembro, Janeiro e Setembro apresentam uma maior frequência de ocorrência de cheia (12,12%, 12,50% e 15,62%, respetivamente).

Foi ainda observado um padrão ainda mais complexo no que toca à variabilidade inter-anual de precipitação desde 2010, bastante mais evidente que nos anos anteriores.

A fim de obter uma explicação de *Ocorrência de cheia* usando os preditores disponíveis, foram ajustados modelos de Regressão Logística com a versão categorizada destes preditores. A versão final do modelo apresentado não contém a variável *Caudal Afluente Médio Mensal*, tendo esta sido descartada por não ser importante quando considerada com as restantes variáveis, estando bastante correlacionada com *Descarga de Superfície Mensal*.

De acordo com o modelo de Regressão Logística final, o decrescimento dos valores de precipitação acumulada mensal reduz significativamente o risco de ocorrência de cheia, validando assim a relação já conhecida e referida na literatura, entre precipitação e ocorrência de cheia. De acordo com este modelo, nos meses com registo de precipitação acumulada inferior a 49,1 mm, o risco de ocorrência de cheia reduz cerca de 84%.

Quanto à variável *Descarga de Superfície Mensal*, os resultados apontam que o risco de cheia diminui com a diminuição das dimensões de descarga. Para meses com descargas inferiores a 2 milhões dam^3, a redução do risco é de cerca de 95%.

Apesar da especificidade do modelo de Regressão Logística final ser bastante elevada, 99,1%, a pouca sensibilidade deste modelo, cerca de 33,3%, pode ser devida a diferentes fatores. Em primeiro lugar, o facto de existir um número bastante elevado observações omissas, havendo apenas 15 casos completos em que ocorreu cheia (o que é manifestamente pouco). Em segundo lugar, existirem apenas dois preditores para explicar um fenómeno tão complexo. Acresce ainda o facto de não se poderem usar as variáveis na sua versão contínua no

ajustamento do modelo de Regressão Logística, por estas variáveis apresentarem elevada assimetria positiva e muitos `outliers`. Para fazer face a este último problema considerou-se a metodologia de Florestas Aleatórias. Com este procedimento, a especificidade passou a ser de 99,5% (há apenas uma não cheia classificada como cheia) e a sensibilidade aumentou para 60% (9 dos 15 casos de cheia são classificados corretamente).

Conclui-se que as variáveis consideradas nos modelos, *Precipitação Acumulada Mensal* e *Descarga de Superfície Mensal*, são importantes na explicação do fenómeno e, apesar da conhecida complexidade do mesmo, conseguem discriminar com precisão elevada a ocorrência de cheia.

5.1 Trabalho Futuro

Existem várias direções para trabalho futuro, devendo-se começar por reforçar a base de dados, de modo a caracterizar mais fielmente a bacia hidrográfica do rio Douro. É ainda fundamental recolher outros dados, relativos a outras variáveis importantes, quer ligadas à precipitação, tais como, localização, intensidade e duração, quer ligadas ao solo, tais como, tipo, uso e humidade, quer relativas à existência de estruturas de proteção, à sazonalidade, etc. Incluir dados piezométricos e relacionados com incêndios poderá ser, também, muito interessante.

O desenvolvimento de uma aplicação para análise rápida e eficiente de dados, contendo diversas metodologias, para uma melhor compreensão dos processos hidrológicos em bacias hidrográficas é também um objetivo futuro.

Bibliografia

Alcoforado, M. J., Silva, L. P., Amorim, I., Fragoso, M., & Garcia, J. C. (2021). Historical floods of the Douro River in Porto, Portugal (1727–1799). *Climatic Change*, *165*(1), 1–20.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Chow, V. T. (1956). Hydrologic studies of floods in the United States. *International Association of Hydrological Sciences*, *42*, 134–170.

Conceição, T. (2008). *Impacto das acções antropogénicas no comportamento sedimentar do rio Douro* (tese de doutoramento). Dissertação de Mestrado, Departamento de engenharia civil da Universidade de Aveiro, Portugal.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Daveau, S. (1977). *Répartition et rythme des précipitations au Portugal*.

Diakakis, M., Damigos, D. G., & Kallioras, A. (2020). Identification of Patterns and Influential Factors on Civil Protection Personnel Opinions and Views on Different Aspects of Flood Risk Management: The Case of Greece. *Sustainability*, *12*(14). https://doi.org/10.3390/su12145585

Garrote, J., Alvarenga, F., & Díez-Herrero, A. (2016). Quantification of flash flood economic risk using ultra-detailed stage–damage functions and 2-D hydraulic models. *Journal of Hydrology*, *541*, 611–625.

Hipólito, J. R., & Vaz, Á. C. (2011). Hidrologia e Recursos Hídricos. 1ª Edição.

Johnson, F., White, C. J., van Dijk, A., Ekstrom, M., Evans, J. P., Jakob, D., Kiem, A. S., Leonard, M., Rouillard, A., & Westra, S. (2016). Natural hazards in Australia: floods. *Climatic Change*, *139*(1), 21–35.

Llasat, M.-C., Barriendos, M., Barrera, A., & Rigo, T. (2005). Floods in Catalonia (NE Spain) since the 14th century. Climatological and meteorological aspects from historical documentary sources and old instrumental records. *Journal of hydrology*, *313*(1-2), 32–47.

LNEC. (1990). As Cheias em Portugal Caracterização das Zonas de Risco, 1° Relatório: Análise Preliminar.

Nunes, A. (2011). Evolução dos caudais no rio Beça: resposta à variabilidade climática ou às mudanças no uso do solo? *Parte: http://hdl. handle. net/10316.2/2645*.

Nunes, A., Moreira, C. O., & Paiva, I. R. (2016). Territórios de água.

Osswald. (2008). *Nascer, Viver e Morrer no Porto Seiscentista* (tese de doutoramento). University of Porto.

Pardal, J., Cunha, L., & Tavares, A. (2016). As cheias na sub-bacia hidrográfica do Rio dos Fornos: pontos críticos e medidas de minimização. *Territórios de Água*, 15–30.

Pereira, D. I., Pereira, P. J. S., Santos, L. J. C., & da Silva, J. M. F. (2014). Unidades Geomorfológicas de Portugal Continental. *Revista Brasileira de Geomorfologia*, *15*(4).

Pereira, M., Figueiredo, N., & Caeiro, D. F. (2007). Considerações sobre a fragmentação territorial e as redes de corredores ecológicos. *Geografia*, *16*(2), 5–24.

Pestana, M. (2014). Análise de dados para ciências sociais a complementaridade do spss 6 a edição Revista, Atualizada e Aumentada Maria Helena Pestana João Nunes Gageiro. *no. September*, 1–2.

Rebelo, F. (2003). *Riscos naturais e acção antrópica: estudos e reflexões*. Imprensa da Universidade de Coimbra/Coimbra University Press.

Reinman, S. L. (2012). Intergovernmental panel on climate change (IPCC). *Reference Reviews*.

Sá, L., Almeida, M., Freire, P., & Tavares, A. (2016). Gestão do Risco de Inundação-Documento de Apoio a Boas Práticas. *ANPC/PNRRC, 44p*.

Santos, M., & Fragoso, M. (2013). Precipitation variability in Northern Portugal: data homogeneity assessment and trends in extreme precipitation indices. *Atmospheric Research*, *131*, 34–45.

Santos, M., Fragoso, M., & Santos, J. A. (2018). Damaging flood severity assessment in Northern Portugal over more than 150 years (1865–2016). *Natural Hazards*, *91*(3), 983–1002.

Santos, M., Santos, J., & Fragoso, M. (2015). Historical damaging flood records for 1871–2011 in Northern Portugal and underlying atmospheric forcings. *Journal of Hydrology*, *530*, 591–603. https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.10.011

Velhas, E. (1997). As cheias na área urbana do porto.: Risco, percepção e ajustamentos. *Territorium*, (4), 49–62.

Zêzere, J., Pereira, S., Tavares, A., Bateira, C., Trigo, R., Quaresma, I., Santos, P., Santos, M., & Verde, J. (2022). *DISASTER database on hydro-geomorphologic disasters in Portugal (Version 1, 2022)*. Zenodo. https://doi.org/10.5281/zenodo.7117037

Apêndice

.1 Códigos R

```r
# BIBLIOTECAS NECESSÁRIAS
library(caret)
library(ggsankey)
library(haven)
library(dplyr)
library(randomForest)
library(pROC)
library(hrbrthemes)
library(ResourceSelection)
library(gridExtra)
library(grid)
#####################################
# IMPORTAR DADOS
Dados_12set_Carlos <- read_sav("Dados_12set_Carlos.sav")
dados=na.omit(dados)
dados$Precip.acumulada=as.numeric(dados$cat.precip.acum)
dados$Cheia=as.factor(dados$Cheia)
dados$catD.descarga.C=as.factor(dados$catD.descarga.C)
str(dados)
#####################################
# MODELO DE REGRESSÃO LOGÍSTICA PROPOSTO
mylogit1 <- glm(Cheia ~ . , data = dados, family = "binomial")
summary(mylogit1)
mylogit1
```

```
###############
#CURVA ROC
par(pty="s")
roc(dados$Cheia, mylogit1$fitted.\emph{value }s, plot=TRUE,
    legacy.axes=TRUE, percent=TRUE,xlab="Percentagem de Falsos Positivos",
    ylab="Percentagem de Verdadeiros Positivos",
    print.auc=TRUE)
###############
# RESULTADOS ESTIMADOS
results_prob <- predict(mylogit1,dados,type='response')
###############
# CUT-OFF
results <- ifelse(results_prob > 0.5,1,0)
###############
# OBSERVADO
answers <- dados$Cheia
###############
# ACURÁCIA
misClasificError <- mean(answers != results)
###############
# MATRIZ DE CONFUSÃO
table(answers, results)
###############
# TESTE DE HOSMER E LEMESHOW
hoslem.test(dados$Cheia ,mylogit1$fitted.\emph{value }s,g=10)
####################################
# MODELO DE FLORESTAS ALEATÓRIAS PROPOSTO
model1 <- train(Cheia ~ .,
                data = dados,
                method = 'rf',
                trControl = trainControl(method = 'cv',number = 5))
model=randomForest(Cheia ~., data=dados, proximity=TRUE )
###############
# ERRO OBB - GRÁFICO
```

```r
oob.error.data=data.frame(
  Trees=rep(1:nrow(model$err.rate),times=3),
  Tipo=rep(c("OOB", "0","1"),each=nrow(model$err.rate)),
  Erro=c(model$err.rate[,"OOB"],
         model$err.rate[,"0"],
         model$err.rate[,"1"]))
ggplot(data=oob.error.data, aes(x=Trees, y =Erro))+
  geom_line(aes(color=Tipo))+
  labs(x="Árvores")+
  theme_ipsum()
###############
# IMPORTÂNCIA DAS VARIÁVEIS
imp <- varImp(model1)
ggplot(imp) +
  labs(title = "Importância das Variáveis", x ="Global", y="Importância")+
  theme_ipsum()
###############
# CLASSIFICAÇÃO
Predmyforest1990 <- predict(model,dados)
t(table(Predmyforest1990, dados$Cheia))
prop.table(t(table(Predmyforest1990, dados$Cheia)), margin=1)
###############
# ACURÁCIA
misClasificError <- mean(dados$Cheia != Predmyforest1990)
1-misClasificError
###############
# RESULTADOS ESTIMADOS
results <- ifelse(Predmyforest1990 == 1,1,0)
results=as.factor(results)
###############
#OBSERVADO
labels <- dados$Cheia
###############
# MATRIZ DE CONFUSÃO
```

```r
confusionMatrix(results,labels)
###################################
# CONSTRUÇÃO DO HIDROGRAMA
media_anual_ = aggregate(dados,list(dados$YEAR), FUN=mean)
M=media_anual_[,c(1,3,4)]
g1= ggplot(data = media_anual_, aes(x=Group.1, y=x)) +
  geom_bar(stat = 'identity', color="black", fill="#69b3a2", width=0.2 ) +
  scale_x_discrete(breaks = seq(1990, 2021, 2)) +
  ylab("Precipitação Acumulada \n Média Anual (mm)") +
  scale_y_reverse()+
  theme_ipsum()+
  theme(axis.title.x    = element_blank(),
        axis.text.x     = element_blank(),
        axis.ticks.x    = element_blank())
g2=ggplot(data=M, aes(x=Group.1, group=1)) +
  geom_line(aes(y=Caudal.medio.C),color="Black",size=0.8)+
  geom_point(aes(y=Caudal.medio.C),shape=21, color="black", fill="#69b3a2", size=2)+
  labs(y="Média Anual de Caudal Afluente (m3/s)", x="Ano")+
  scale_x_discrete(breaks = seq(1990, 2021, 2)) +
  theme_ipsum()
g1 <- ggplot_gtable(ggplot_build(g1))
g2 <- ggplot_gtable(ggplot_build(g2))
maxWidth = unit.pmax(g1$widths[2:3], g2$widths[2:3])
g1$widths[2:3] <- maxWidth
g2$widths[2:3] <- maxWidth
grid.arrange(g1, g2, ncol = 1, heights = c(1, 3))
```