UNIVERSIDADE
LUSÓFONA

Análise de dados de caudal para apoio na tomada de decisão em sistemas de distribuição de água

Trabalho Final de curso

Relatório final

Nome do Aluno: Leandro Pinheiro

Nome da Orientadora: Maria Almeida Silva

Nome da Coorientadora: Dália Loureiro (Laboratório Nacional de Engenharia Civil)

Trabalho Final de Curso | LEI | 28 de Junho de 2024

www.ulusofona.pt

Direitos de cópia

Resumo

A água é um elemento fundamental à vida e, com as alterações climáticas existe cada vez mais escassez deste elemento, sendo necessário controlar o uso dela. Os sistemas de distribuição de água potável devem encontrar-se divididos em zonas de medição controlada (ZMC), sendo que, nestas zonas, todos os pontos de entrada e saída de água devem ser medidos para um controlo adequado das perdas de água. Apesar disso, existem perdas de água por fugas e roturas, ou até mesmo por usos ilícitos da água e por erros de medição. Portanto, propõe-se uma metodologia para processamento dos dados de caudal medidos em ZMC, que permita calcular indicadores que alertem sobre problemas na qualidade destes dados e na operação da rede. Existem diversos artigos e teses sobre perdas e roturas de água e também algumas metodologias para análise de dados de caudal. Contudo, em geral, estas são focadas apenas num problema das séries de caudal, não os analisando de uma forma mais global, tal como se pretende neste estudo. É um trabalho pertinente tendo em conta os dias de hoje, onde as alterações climáticas são cada vez mais visíveis e quando em Portugal se vê cada vez mais escassez de água. É então essencial dispor de dados fiáveis para identificar rapidamente fugas e roturas, solucionando estes problemas, evitando grandes perdas de água e custos adicionais associados.

Este trabalho é realizado em colaboração com o Laboratório Nacional de Engenharia Civil (LNEC) e recorrendo a dados anónimos de caudal obtidos em sistemas reais. Estes dados serão analisados através de métodos implementados em Python e de algumas ferramentas já desenvolvidas em R. Neste trabalho, serão aplicados conhecimentos e competências adquiridos nas unidades curriculares (UC) de Data Science e Probabilidades e Estatística.

Ao longo deste trabalho, iremos aplicar a metodologia proposta a séries temporais de caudal de sistemas de distribuição de água, tentando obter uma série sem falhas, com passo de tempo pré-definido e normalizado entre medições. No final, pretendemos avaliar a qualidade dos dados, propondo indicadores que meçam essa qualidade.

Abstract

Water is fundamental to life and, with climate change, there is an increasing scarcity of this element, making it necessary to control its use. The potable water distribution systems must be divided into District Metered Areas (DMA), in which all water inlet and outlet points must be measured to properly control water losses. Despite that, there are water losses from pipe bursts and leakages, or even from illicit usages of the water and measurement errors, and therefore, we propose a methodology for processing flow data measured in DMA making it possible to calculate indicators that alert about problems in the quality of data and in the network operation. There are multiple articles and theses about leakages and pipe bursts and some methodologies for flow data analysis. Although, in general, these are just focused on one problem in the flow data, not analyzing it in a general way, how it's intended in this study. It's a relevant work considering nowadays climate change is even more noticeable and in Portugal, we see more and more shortage of water. It is therefore essential to have reliable data to quickly identify leakages and pipe bursts, solve these problems and avoid large water losses and additional costs associated.

This work is done with the cooperation of LNEC, using anonymous flow data obtained from real systems. This data will be analyzed with methods implemented in Python and some tools already developed in R. In this work, knowledge and skills acquired in the curricular units of Data Science and Probabilities and Statistics will be applied.

Over this work, we will apply the proposed methodology to the flow time series of water distribution systems, trying to obtain a flawless series, with a pre-defined time step and normalized. In the end, we pretend to evaluate the quality of data, proposing identifiers that measure that quality.

Índice

Lista de Figuras

Lista de Tabelas

1 Identificação do Problema

A água é um bem essencial que afeta todos os seres vivos, não sendo possível a vida na terra sem este recurso, e, portanto, torna-se crucial fazer uma gestão controlada do mesmo. Em Portugal vários sistemas de distribuição de água para consumo humano são divididos por setores, onde todos os pontos de entrada e saída de água são monitorizados em termos de caudal, sendo designados por Zonas de medição e controlo (ZMC) [8]. Estas zonas distribuem a água a nível urbano, e têm várias dimensões, dependendo da urbanização à qual fazem a distribuição das águas. Olhando para os dados disponibilizados pela ERSAR, vemos que em 2021 28,8% da água em Portugal não foi faturada [6]. A água não faturada deve-se a consumo autorizado não faturado, perdas reais (e.g.,a fugas ou roturas), perdas aparentes (e.g., usos ilícitos da água, erros de medição).

É essencial as entidades gestoras da água serem capazes de detetar rapidamente estas fugas ou roturas para evitar elevadas perdas de água e custos adicionais relacionados. Para tal, é necessário usar dados de caudal fiáveis. De modo aos dados de caudal serem fiáveis é necessário implementar uma metodologia que permita criar indicadores para avaliar a sua qualidade. A qualidade destes dados pode ser influenciada por aspetos relativos à cadeia de medição e registo, como seja, condições de instalação dos equipamentos de medição, o próprio equipamento de medição, o modo de registo e comunicação das medições, métodos para processamento e armazenamento dos dados. Problemas ao nível desta cadeia de medição podem comprometer a fiabilidade dos dados de caudal, como falhas nos dados, ou eventos anómalos, comprometendo também a sua aplicabilidade para controlo operacional dos sistemas de distribuição de água. Com base na avaliação da qualidade, podem ser identificados os problemas e assim serem aplicadas vias de resolução. Pretende-se também com base na análise dos dados poder classificar as séries como fiáveis ou não.

Este trabalho tem então dois objetivos:

- Propor uma metodologia, baseada em estudos anteriores, e aplicar métodos para análise exploratória e processamento de dados de caudal.
- Definir indicadores para avaliação da qualidade dos dados de caudal.

2 Viabilidade e Pertinência

A água é um recurso vital que sustenta a vida e que é necessária em inúmeras áreas, tais como na área da saúde, em processos industriais, na cozinha, na agricultura e na produção de energia. Com as alterações climáticas, existe cada vez mais escassez de água e torna-se essencial fazer um uso sustentável da mesma. Assim, colaborando com o LNEC, iremos desenvolver uma metodologia para fazer uma melhor gestão da água.

É importante com esta análise retirar indicadores que avaliem a qualidade dos dados, pois são essenciais para conseguir identificar os problemas, podendo então estes serem solucionados o mais rapidamente possível, melhorando a fiabilidade, e potenciando a sua aplicação para a deteção célere de perdas de água, evitando custos adicionais relacionados.

Estes sistemas de distribuição de água devem ser melhorados ao longo dos anos aplicando novas tecnologias e metodologias que garantam a qualidade e diminuição das perdas de água, podendo então ser dada continuidade nesse âmbito a este trabalho.

Observando o caso de Mafra presente na Série Guias Técnicos 3 [8], observamos que foram implementadas intervenções para controlo de perdas, tais como as técnicas de controlo de perda, técnicas e equipamentos de localização e deteção de fugas, e ainda se usou aplicações computacionais para controlo de perdas de água na rede. Vemos que esse controlo de perdas em termos económicos traduziu-se num ganho de 208 947€ comparativo a um ano em que o custo das perdas seria igual. E num balanço final permitiu diminuir o tempo de interrupção no abastecimento para reparar avarias, diminuir o volume de água perdido nas avarias, minimizar problemas na qualidade da água e ainda retardar eventuais restrições de fornecimento de água.

3 Benchmarking

Para a possível realização deste TFC foi estudado um conjunto de teses e artigos. A tese de D. Loureiro (2010) [2] descreve várias metodologias já existentes que utilizam por base as etapas de análise descritiva de dados, eliminação de eventos anómalos, transformação de dados, combinação de dados e redução de dados. Descreve também cada etapa, falando também das diferentes abordagens e análise e controlo de perdas em sistemas de distribuição de água. H. Alegre (2005) [8] apresenta aplicações reais de algumas das abordagens, podendo observar-se os resultados adquiridos e o quão impactantes são.

Em J. Quevedo (2010) [9], percebemos como validar e reconstruir dados de caudal, identificando o que são falhas de dados e dados em falta, sendo ainda propostas técnicas para preencher as tais falhas, baseadas em modelos de séries temporais. Alterações a estas técnicas, para preenchimento de falhas, são apresentadas em R. Barrela (2015) [12], sendo também aplicadas a séries temporais de caudal. B. Ferreira (2023) [1], através do uso de inteligência artificial identifica roturas em tempo real, permitindo perceber o que são anomalias e como as detetar. M. Silva (2016) [11] também apresenta uma metodologia para deteção de anomalias, baseada em técnicas de clustering e modelos simbólicos.

Os artigos D. Loureiro (2011) [4], S. Mounce (2011) [13] e D. Loureiro (2015) [3] fazem a aplicação de vários métodos, utilizando diferentes tecnologias, a dados reais, demonstrando abordagens que permitem melhorar o controlo das perdas de água, monitorizar os dados e detetar os eventos anómalos.

Para compreender a aplicação dos dados de caudal em sistemas de distribuição de água, foi realizado um estudo dos artigos publicados pela ERSAR (2023) [6] e [7]. No contexto dessas análises, foi observada a percentagem de água não faturada, organizada por ano conforme apresentado em [6]. E também foi realizada uma caracterização detalhada do setor de águas e resíduos, como descrito em [7]. Essa abordagem proporciona uma visão abrangente e realista dos dados relacionados à faturação de água e aos sistemas de distribuição de água em Portugal.

4 Engenharia

4.1 Metodologia

A metodologia a ser desenvolvida será composta pelos seguintes passos:

- Recolha de dados: Onde irão ser recolhidos dados de caudal para serem aplicados e estudados nas etapas seguintes.

- Análise exploratória dos dados: Nesta etapa estudamos os dados, através do cálculo de estatísticas descritivas tais como média, desvio padrão, mediana, máximos, mínimos, quartis, número de valores em falta, entre outros, e apresentamos alguns em gráficos relativos a análises preliminares dos dados.

- Normalização: Este procedimento serve para regularizar o espaço de tempo entre as observações consecutivas das séries temporais de caudal.

- Preenchimento de Falhas: Aquando da medição de dados de caudal, seja por problemas técnicos ou humanos, ocorrem falhas nas medições. De modo a que não sejam descartadas, é preciso identificar o melhor método a implementar para preencher essas falhas e, assim, conseguir analisar as séries.

- Identificação de eventos anómalos: Situações como roturas e fugas de água, ou utilizações anormais da mesma, levam à existência de eventos anómalos nas séries de caudal. Apesar de ser preferível que as entidades gestoras tenham um registo destas ocorrências, isto nem sempre acontece, o que leva a que seja necessário fazer uma identificação *a posteriori* destes eventos.

- A Aplicação dos dados: A partir do momento em que a recolha dos dados é concretizada, iremos começar a aplicar as diversas etapas a estes dados usando as ferramentas em Python e em R.

4.2 Conceitos teóricos

Neste capítulo irão ser descritos vários conceitos teóricos, consolidando os temas, para melhor compreensão dos mesmos nos capítulos seguintes.

4.2.1 Eventos anómalos em séries de caudal

Eventos anómalos são dados que apresentam um comportamento distinto do típico da série de caudal em sistemas de distribuição de água[2]. As anomalias mais comuns estão relacionadas com duplicações de dados, valores negativos, valores demasiado elevados ou baixos ou até períodos sem dados [1]. Para definir os dados que seriam considerados eventos anómalos de cada mês devido a serem valores demasiado elevados ou baixos, utilizou-se a regra de Tukey (1977) [14]. Esta é a regra implementada pelo Python ao fazer um diagrama de caixas. Esta regra consiste em calcular a diferença entre o 3º quartil (Q3) e o 1º quartil (Q1), a qual se designa de amplitude interquartil (IQR). Segundo a regra de Tukey, todos os valores abaixo de Q1-1.5*IQR ou acima de Q3+1.5*IQR são considerados eventos anómalos. Uma vez que o valor de 1.5 pode ser demasiado conservativo, optou-se por utilizar 3, o que significa que se detetam eventos anómalos severos, Tukey (1977) [14].

4.2.2 Normalização dos dados para um passo de tempo regular pré-definido

Os dados obtidos na recolha não têm um passo de tempo regular, isto é, não existe uma medida de tempo certa entre as medições de caudal. De modo a termos dados com um passo de tempo regular, iremos realizar a normalização dos dados. Fazemos isto para reduzir a redundância e dependência da informação, garantindo integridade e eliminando anomalias.

Para isso é criada uma série de caudal a partir da original, com espaço de tempo definido, no nosso caso esse passo de tempo foi de 15 minutos. Todos os valores da série original com medidas de tempo que existam também na série modificada são mantidos. Para os restantes é visto se a o tempo entre os dois pontos mais próximos é menor do que a duração limite de falha. Esta duração limite de falha é um valor pré definido ao realizar a normalização. Se o tempo entre dois pontos for menor do que a duração limite de falha, o valor normalizado y para o instante de tempo x é calculado através da interpolação linear dos dois pontos mais próximos, caso contrário este é deixado sem valor. Este processo encontra-se esquematizado na Figura 1, para o caso em que o tempo entre os dois pontos mais próximos é menor do que a duração limite de falha definida. O valor normalizado é calculado por interpolação linear usando a seguinte fórmula:

$$y = y0 + (y1 - y0)\,((x - x0)/(x1 - x0))$$

Onde:

y0 é o valor de caudal no instante x0,

y1 é o valor de caudal no instante x1.

Sendo a duração entre os dois pontos mais próximos maior do que a duração limite de falha, este ponto é deixado sem valor.

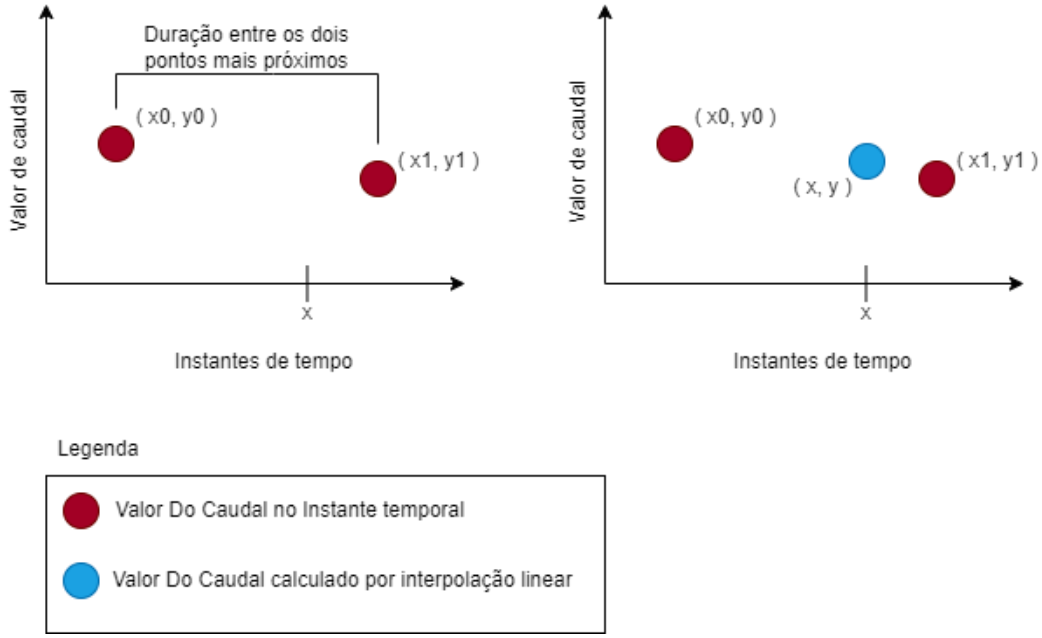


Figura 1 - Cálculo de valores de caudal na etapa de normalização

4.2.3 Preenchimento de falhas

Para o preenchimento de falhas foram utilizados 3 métodos distintos, todos eles desenvolvidos em R.

- Forecast – É uma biblioteca em R que tem ferramentas para reconstrução de dados nas quais foram definidas as sazonalidades a ter em consideração para fazer a reconstrução dos dados (diária, semanal, anual)
- Método baseado na Quevedo [9] [12] – Inicialmente é criado um modelo que representa os dados de fluxo diário de forma agregada, depois é analisada a variação dentro de cada dia, observando como o fluxo varia ao longo das horas do dia, e por fim combinando os dados de fluxo diário com a sua variação é então reconstruída a série.
- Método baseado no modelo TBATS [10] [12] – este modelo começa por transformar a série de modo a estabilizar a sua variância, de seguida são analisadas as tendências, depois são modelados os erros da previsão e por fim é modelada a sazonalidade tendo em consideração vários períodos sazonais (diário, semanal, anual).

4.2.4 Qualidade da medição

Existem vários aspetos que podem comprometer a qualidade da medição tais como:

- Períodos sem medições;
- Dados não normalizados;
- Espaçamento de tempo entre medições;
- Eventos anómalos.

Estes problemas podem ser causados pelo funcionamento anómalo dos medidores de caudal, problemas na aquisição, transmissão, processamento e armazenamento dos dados de caudal.

Todos estes fatores reduzem a qualidade da medição. Então ao longo deste trabalho algumas informações sobre estes problemas foram retiradas como indicadores da qualidade da medição.

5 Solução Desenvolvida

5.1 Introdução

Para este trabalho foi definida uma metodologia que passa por 7 etapas, sendo estas a recolha dos dados, análise exploratória dos dados, normalização, preenchimento de falhas, identificação de eventos anómalos, tomada de decisões e por fim aplicação dos dados. Todas estas etapas encontram-se descritas na secção 4.1, sendo os resultados apresentados na secção 5.3.

Foram então definidas as tecnologias e ferramentas necessárias para desenvolver esta metodologia.

Com este trabalho espera-se ser possível definir indicadores para avaliação da qualidade dos dados de caudal com base na classificação das séries como fiáveis ou não. Durante este trabalho, irão ser aplicados os conhecimentos adquiridos curricularmente bem como se pretende adquirir conhecimentos em temas não lecionados nas unidades curriculares da licenciatura.

Na seguinte hiperligação encontra-se o código do projeto sem os dados, por questões de confidencialidade dos mesmo.

5.2 Tecnologias e Ferramentas Utilizadas

Para o desenvolvimento deste trabalho foi utilizado o Google Colaboratory, onde foi desenvolvido o código em linguagem Python. Esta linguagem foi escolhida por ser considerada uma linguagem de programação útil no contexto de análise de dados e por incluir bibliotecas que permitem o fácil manuseamento dos dados assim como a visualização gráfica. Exemplos dessas bibliotecas são: Pandas, numpy, seaborn e matplotlib.pyplot. Noutras etapas desta metodologia foram também utilizadas algumas ferramentas já existentes desenvolvidas em R.

5.3 Implementação

A metodologia então a ser desenvolvida será constituída pelos seguintes passos:

- Recolha de dados
- Análise exploratória dos dados
- Normalização
- Preenchimento de Falhas
- Identificação de eventos anómalos
- A Aplicação dos dados

5.3.1 Recolha de dados

Foram fornecidos pelo LNEC um conjunto de documentos excel, contendo dados sobre os valores de caudal em instantes de tempo. Estes documentos continham um elevado número de linhas tendo então estes dados sido extraídos com ferramentas em python.

5.3.2 Análise Exploratória

Foram recolhidos os dados de caudal, em m^3/h , de cinco zonas distintas, por motivos de confidencialidade dos dados iremos designar estas por AL, AR, CE, MA, PC, todos estes com início em julho de 2013 e fim em junho de 2016. Diversas análises estatísticas, entre as quais as descritas acima, foram aplicadas a todas as zonas em estudo.

Na análise dos dados de AL, começámos por verificar que não existem dados duplicados e que estão organizados por ordem cronológica. Em média existem medições a cada 5 minutos e 5 segundos, o que representa um valor médio de medições diárias de 288. Esta série de caudal apresenta 2974 medições em falta, espaçadas pelos 3 anos de dados, o que corresponde a menos de 1% dos dados. Na figura 2, podemos observar os valores de caudal em m^3/h ao longo destes 3 anos. Podemos também observar na figura 3 as médias mensais dos dados referente a AL, onde facilmente conseguimos observar um crescimento expectável dos valores de caudal nos períodos de mais calor (Verão). A este fenómeno chamamos de sazonalidade anual. Os diagramas de caixas presentes na figura 8 dos anexos permite observar os valores de caudal a cada mês, sendo que os pontos pretos representam os eventos anómalos segundo a regra de Tukey aplicada [14]. De referir que, apesar da visualização dos diagramas de caixas poder enganar, os valores de caudal nunca são 0, apenas atingem valores inferiores a 1.

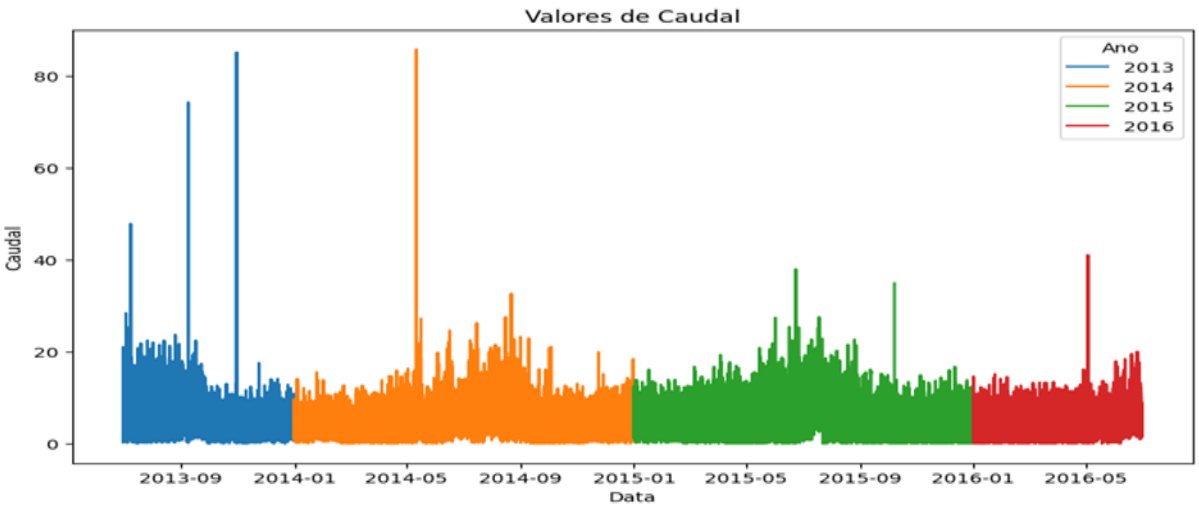


Figura 2 - Valores de caudal em AL

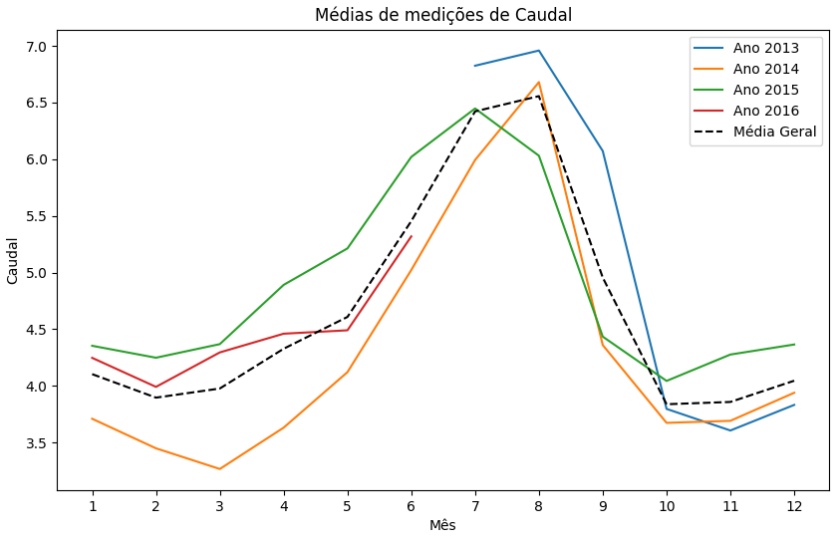


Figura 3 – Médias mensais dos valores de caudal em AL (sazonalidade anual)

Das análises efetuadas às restantes zonas foram ainda tiradas as seguintes observações:

Da análise feita a AR será importante referir que desde setembro de 2014 até setembro de 2015 (exclusive) os valores de caudal foram mais elevados do que nos restantes anos, como se pode observar nas figuras 10 e 11, que se encontram em anexo. Esta foi a zona com menos falhas tendo apenas 177 de um total de 293 188 medições realizadas nesta zona, o que corresponde a menos de 0.07% do total de medições.

Em CE os valores de caudal estão entre 69 e 109, que são valores muito distintos das restantes zonas.

Na análise realizada em MA verificamos que em finais do ano 2015 e inícios do ano 2016 existe uma grande ausência de dados, facilmente observada na figura 22, que se encontra em anexo. Podemos também observar na figura 23 que janeiro de 2016 tem valores muito superiores em relação aos outros anos, tendo sido detetados vários eventos anómalos neste mês, segundo a regra de Tukey aplicada [14]. Além disso, este mês de janeiro apresenta também muitas falhas de dados.

Os valores de caudal de PC foram diferentes dos expectáveis como podemos ver na figura 4 que mostra um gráfico das médias dos valores de caudal em PC. Neste observamos que o ano de 2014 apresenta um comportamento atípico relativamente aos restantes. Para além disto PC também não tem dados referentes a cerca de 33 dias e apresenta ainda, 31 medições em falha.

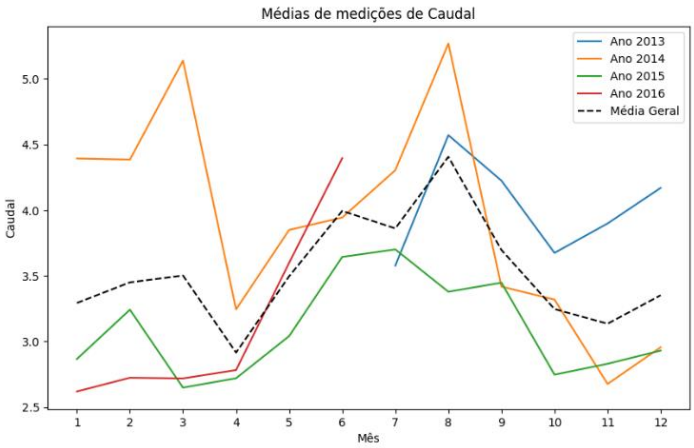


Figura 4 - Médias mensais dos valores de caudal em PC (sazonalidade anual)

Todos os gráficos do resultados obtidos encontram se em anexo, organizados pelas respetivas zonas (AL – Figura 8; AR - Figuras 9 a 11; CE - Figuras 12 a 14; MA - Figuras 15 a 17; PC - Figuras 18 a 19).

Nesta análise conseguimos retirar como indicadores de qualidade a percentagem de registos sem valor de caudal.

5.3.3 Normalização dos dados para um passo de tempo regular pré-definido

Chegando a esta etapa, foi começado o processo descrito em 4.2.2 e foram realizadas 3 normalizações com durações limite de falha diferentes: 15 min, 30 min e 60 min. Foram utilizados os dados de AL como exemplo. Após este processo foi visto que os eventos anómalos já detetados na análise exploratória causavam impacto na série normalizada, o que retirava alguma credibilidade da mesma, tendo sido então tomada a decisão de serem retirados estes valores antes de efetuar a normalização.

Utilizando o método de Tukey para deteção de eventos anómalos severos descrito em 4.2.1, foram retiradas 209 medições à série de caudal de AL, o que equivale a 0.07% dos dados. Das restantes séries de caudal foram retiradas as seguintes percentagens de dados que correspondem a eventos anómalos, AR - 0,04%, CE – 0,01%, MA – 0,48% e em PC – 0,2%.

Observando os resultados destas 3 normalizações percebe-se que o limite de falha de 15 minutos é muito restrito aumentando as falhas de curta duração, o que aumenta a perda de credibilidade da série. Em relação ao limite de falha de 60 minutos, este é demasiado abrangente o que apesar de reduzir o número de falhas, faz com que os valores existentes afastem mais a série da realidade. Foi então definido que a melhor duração de limite de falha para normalização seria a de 30 minutos, aproximando mais os valores normalizados à realidade.

Na Figura 5 podemos observar as diferenças entre a série original e a série normalizada, saltando logo à vista o desaparecimento dos picos existentes na série original. Podendo também ser observado no entanto que se mantém na série normalizada as sazonalidades presentes na série original.

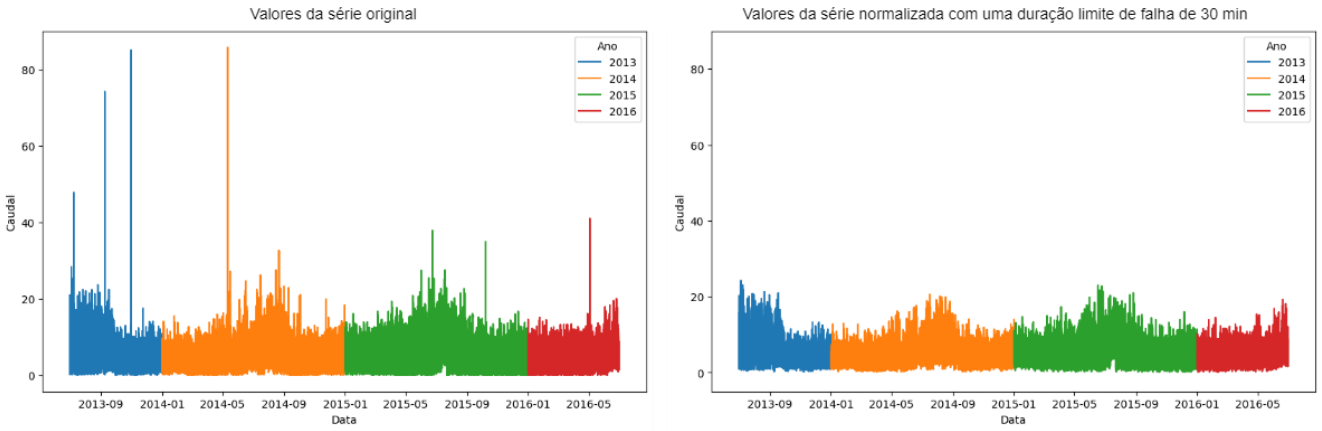


Figura 5 - Comparação dos valores de caudal da série original com a série normalizada

Da normalização realizada às restantes zonas os resultados foram parecidos, tendo também sido observado que a melhor duração limite de falha a aplicar seria de 30 minutos. Em relação aos valores anómalos retirados das séries, em nenhum caso a percentagem desses valores chegou a 1% dos dados, tendo sido o caso mais elevado em MA, em que foram retirados 0,48% dos dados.

5.3.4 Preenchimento de falhas

Estando a normalização feita, foram analisadas as falhas existentes de modo a definir qual o melhor método a aplicar a cada uma para o seu preenchimento. Na figura 6, podemos observar o diagrama de caixas refente à duração das falhas em AL. Neste caso existiam 547 valores em falta o que corresponde a 0,52% da amostra tendo a falha com maior duração 1 dia 3 horas e 45 minutos. Foi também observado que apenas 2,08% das falhas tem uma duração maior do que 1 dia, neste caso não existiram falhas superiores a 2 dias.

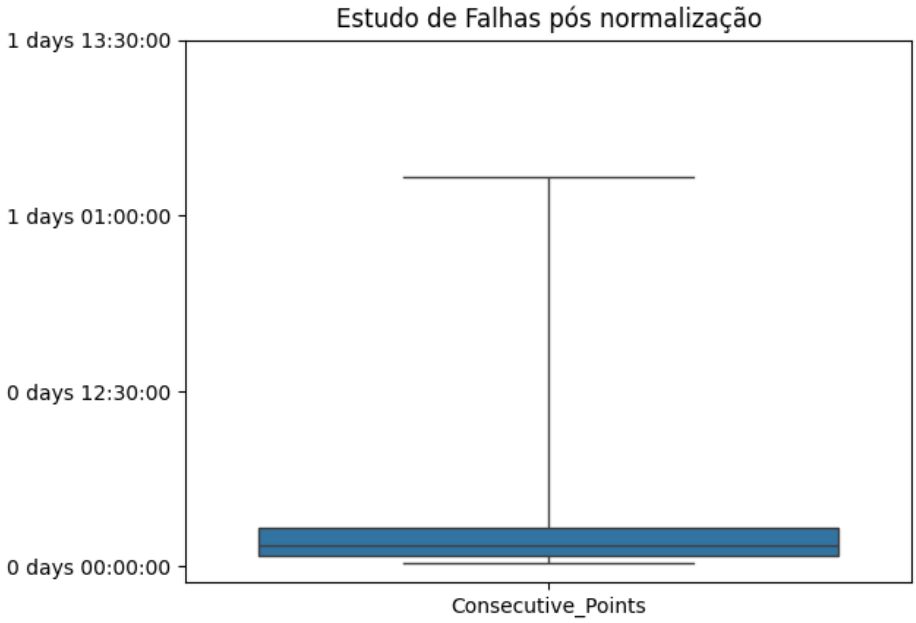


Figura 6 - Diagrama de Caixas referente às falhas existentes após normalização em AL

Na análise feita às restantes séries retiramos as seguintes percentagens de valores em falta: AR – 0,22%, CE – 0,9%, MA – 5,16% e em PC – 3,79%. Apesar da percentagem de valores em falta em PC ser menor do que em MA, foi em PC que ocorreu a falha de maior duração: 29 dias 12h 45min. Esta falha era expectável pois existiam 33 dis sem dados em PC. Apesar disso, 89,81% das falhas foram menores do que 1 dia. No caso de MA, apesar de ter a maior percentagem de falha, 93,37% das suas falhas tiveram duração menor do que 1 dia. A Tabela 1 apresenta a distribuição do número de falhas com duração inferior a 1 dia em MA de acordo com alguns intervalos de duração.

Tabela 1 - Análise temporal de falhas menores que 1 dia em MA

DURAÇÃO DA FALHA	NÚMERO DE FALHAS	% DE FALHAS MENOR QUE 1 DIA
< 5H	140	90,32%
<= 5H E < 10H	11	7,10%
<= 10H E < 15H	4	2,58%
<= 15H E < 20H	0	0%
<= 20H E < 24H	0	0%

Fazendo uma análise mais detalhada às 140 falhas menores do que 5h observamos que a maioria destas (117) são menores do que 2h.

Estando então a análise de falhas concluída passamos então ao preenchimento das falhas, para tal foram utilizados 3 métodos distintos todos desenvolvidos em R: biblioteca forecast, método desenvolvido em R. Barrela [12] com base em J. Quevedo [9] e método desenvolvido em R. Barrela [12] com base no modelo TBATS (De Livera [10]).

O método presente na biblioteca forecast, a qual tem ferramentas para reconstrução de dados, nas quais podemos definir as sazonalidades a ter em consideração, foi aplicado com duas variantes: considerando apenas as sazonalidades diária e semanal e considerando as sazonalidades diária, semanal e anual. Estas variantes foram designadas de Forecast_DSA e Forecast_DS. Os restantes 2 métodos utilizados foram métodos já existentes desenvolvidos por R. Barrela [12] que têm por base os métodos de Quevedo [9] e TBATS [10] respetivamente. Em R. Barrela (2015) [12] vemos que o método mais fiável é o baseado no TBATS mas em termos computacionais este demorou cerca de 25 minutos a reconstruir uma série com 105 216 dados, enquanto que os métodos Forecast_DSA e Forecast_DS demoram apenas alguns segundos, e o método baseado na Quevedo demora cerca de 2 minutos.

Aplicados todos os métodos a todas as séries foi então calculado o erro da estimativa da falha em cada instante de tempo. Uma vez que o valor real não é conhecido para o cálculo do erro, foi considerada a estimativa obtida pelo método baseado no modelo TBATS como a correta com base nos resultados obtidos em R. Barrela [12]. Após o cálculo do erro em todos os instantes em falha, foi calculada a média e a mediana dos erros, tendo sido obtidos os resultados presentes na Tabela 2.

Tabela 2 - Médias e medianas do erro da estimativa do valor em falha relativamente ao método TBATS

	FORECAST_DSA		FORECAST_DS		QUEVEDO	
	Média	**Mediana**	**Média**	**Mediana**	**Média**	**Mediana**
AL	54,59%	31,93%	56,11%	36,60%	0%	0%
AR	39,35%	12,16%	30,20%	13,17%	22,90%	15,63%
CE	26,66%	16,03%	26,19%	12,84%	0%	0%
MA	23,75%	15,17%	22,31%	16,56%	26,55%	18,41%
PC	37,75%	24,48%	53,37%	28,82%	14,11%	10,26%

Desta análise é possível ver que, destas 3 abordagens, o método com erros mais baixos foi o baseado na Quevedo, tendo sido a única exceção a série de MA. A partir destes valores não conseguimos tirar uma conclusão sobre o melhor método de reconstrução de dados, pois apesar do método baseado na Quevedo ter obtido valores ideais em duas das séries, nas restantes ainda obteve erros com alguma significância, sendo então necessário serem realizadas análises mais profundas para perceber o que poderá estar a impactar estas diferenças tão elevadas.

O método baseado no modelo TBATS, devido a falta de dados apenas conseguiu reconstruir completamente a série de AL, sendo que em AR ficaram em falta 7 valores, em CE 854 valores, em MA 4 441 valores e em PC 1 057 valores. Estes resultados são apenas referentes aos momentos que o método baseado no modelo TBATS obteve resultados, por exemplo em CE os 0% do erro de estimativa obtidos pelo método baseado na Quevedo não incluem os dados que não foram reconstruídos pelo método baseado no modelo TBATS.

No caso da série MA, o método baseado no modelo TBATS não conseguiu reconstruir o valor de caudal em 4 441 instantes de tempo, uma vez que o elevado número de falhas próximas impediu o ajuste do modelo TBATS aos dados e, consequentemente, a previsão do caudal nas falhas. Tal não aconteceu com o método baseado na Quevedo, uma vez que este não necessita do mesmo ajuste do modelo, permitindo assim a reconstrução das falhas em todos os instantes de tempo.

No início desta etapa, quando foi feito o estudo de falhas pós normalização, puderem ser retirados como indicadores de qualidade a percentagem de valores omissos, o número de falhas inferiores a 12 horas, o número de falhas entre 12 a 24 horas e o número de falhas superior a 24 horas.

5.3.5 Identificação de eventos anómalos

Na etapa 5.3.3 foram detetados os eventos anómalos como explicado anteriormente. Foi também fornecido pelo LNEC um ficheiro excel com as avarias ocorridas, e então foi feita a correlação entre os eventos anómalos de cada série e as avarias ocorridas de modo a perceber quantos destes eventos podem ser justificados por estas avarias. Desta análise obtivemos que em AL temos que dos 209 eventos anómalos detetados, 27 correspondem a algum período de avaria, ou seja aproximadamente 13% dos seus eventos anómalos. Das restantes séries obtivemos os seguintes resultados, em AR 29% dos seus eventos anómalos são em período de avaria, em CE 4%, em MA 5% e em PC 24%.

Daqui foram retirados como identificadores de qualidade os eventos anómalos detetados sendo estes analisados e registados como eventos anómalos de caudal máximo ou mínimo, demonstrando o número destes eventos e os seus limites.

5.3.6 Qualidade da medição

O objetivo proposto por este trabalho era retirar indicadores que avaliem a qualidade dos dados das séries de modo a facilitar a tomada de decisões. De modo a responder a isto após toda a análise feita, é produzida uma tabela que mostra para cada mês de cada ano os seguintes dados:

- Série Bruta (A série no seu estado original)
    - I.SB1: registos sem caudal (%)
    - I.SB2: eventos anómalos de caudal máximo(LI) (m^3/h)
    - I.SB3: eventos anómalos de caudal máximo(LS) (m^3/h)
    - I.SB4: nº de eventos anómalos de caudal máximo (nº)
    - I.SB5: eventos anómalos de caudal mínimo(LI) (m^3/h)
    - I.SB6: eventos anómalos de caudal mínimo (LS) (m^3/h)
    - I.SB7: nº de eventos anómalos de caudal mínimo (nº)
- Série Normalizada
    - I.SN1: valores omissos (%)
    - I.SN2: falhas com duração inferior a 12 h (n.º)
    - I.SN3: falhas com duração entre 12h e 24 h (n.º)
    - I.SN4: falhas com duração superior a 24 h (n.º)

Aqui quando falamos de eventos anómalos de caudal máximo e mínimo são referentes àqueles que foram definidos como eventos anómalos por estarem acima de Q3+3*IQR e abaixo de Q1-3*IQR respetivamente. Em relação à distinção entre LI e LS, LS representa o evento anómalo com valor de caudal mais elevado, e LI representa o evento anómalo com valor de caudal mais baixo.

Na tabela 3 é mostrado um exemplo ,puramente ilustrativo, destes indicadores para o mês de julho do ano 2013. São utilizadas as nomenclaturas definidas acima para apresentar cada coluna, de modo a facilitar a visualização.

Tabela 3 - Valores para analizar a qualidade de medição

Ano e mês	Série Bruta							Série Normalizada			
	I.SB1	I.SB2	I.SB3	I.SB4	I.SB5	I.SB6	I.SB7	I.SN1	I.SN2	I.SN3	I.SN4
2013-07	0.0007	254	315	5	3	7	2	0.68	2	1	3

Neste conjunto de dados vemos que na série bruta no ano de 2013 no mês de julho, 0.0007% dos registos não tem dados, foram detetados 5 eventos anómalos de caudal máximo, tendo sido destes 5 o valor mais elevado 315 e o menor valor obtido foi de 254, foram também detetados 2 eventos anómalos de caudal mínimo, cujos valores são 3 e 7, após normalização neste mês 0.68% dos dados são omissos, e representam 6 falhas, 2 delas menores que 12h, 1 entre 12h e 24h e 3 superiores a 24h.

Todo este conjunto de dados permite perceber se a série analisada é fiável ou não permitindo então identificar problemas, de modo a estes poderem ser solucionados o mais rapidamente possível, melhorando a fiabilidade, e potenciando a sua aplicação para a deteção célere de perdas de água, evitando custos adicionais relacionados.

5.4 Abrangência

Para o desenvolvimento deste trabalho é importante dizer que todas as cadeiras lecionadas na licenciatura tiveram um impacto. Porém existem cadeiras sem as quais este trabalho seria impossível de realizar cadeiras essas de fundamentos de programação, linguagens de programação I e II, data science e probabilidades e estatística.

Fundamentos de programação assim como linguagens de programação I e II foram essenciais pois foram as UC's onde aprendi as bases de programar assim como as boas técnicas de programação, as quais aplico neste trabalho.

Data science foi a UC onde aprendi a linguagem Python, assim como as várias bibliotecas e aplicações das mesmas, para realizar análises aos dados graficamente e não só.

Probabilidades e estatística é uma UC também essencial para este trabalho sendo esta onde aprendemos a fazer análises estatísticas que aqui já foram feitas e apresentadas graficamente e não só, sem as quais não seria possível fazer uma correta análise dos dados.

6 Método e Planeamento

Para o correto desenvolvimento deste TFC tiveram de ser pensadas as diversas tarefas a ser realizadas ao longo do mesmo. Foram então consideradas as tarefas necessárias para o correto desenvolvimento deste trabalho. Num contexto geral temos as 4 entregas do TFC, o desenvolvimento do presente relatório e a pesquisa bibliográfica necessária para compreender o tema a ser estudado. Num contexto específico, temos as tarefas descritas no capítulo anterior necessárias para o desenvolvimento da metodologia proposta neste TFC. De forma resumida, foram então definidas as seguintes tarefas:

- Pesquisa Bibliográfica.
- Recolha de Dados.
- Análise exploratória dos dados.
- Normalização.
- Preenchimento de falhas.
- Identificação de eventos anómalos.
- Aplicação dos dados.
- Escrita do relatório.
- Entregas do TFC, estas são as entregas definidas para entregas do relatório.

Definidas as tarefas apresenta-se aqui em detalhe o calendário em formato Gantt:

	Nome	Início	Fim
1	Pesquisa Bibliográfica	02-10-2023 8:00	31-05-2024 17:00
2	Recolha de Dados	20-11-2023 9:00	04-12-2023 9:00
3	Análise exploratória dos Dados	30-11-2023 8:00	29-12-2023 17:00
4	Normalização	30-12-2023 8:00	22-03-2024 17:00
5	Preenchimento de Falhas	22-03-2024 8:00	17-05-2024 17:00
6	Identificação de eventos Anómalos	15-05-2024 12:00	14-06-2024 17:00
7	Aplicação dos Dados	01-12-2023 9:00	24-06-2024 9:00
8	Escrita do Relatório	02-10-2023 8:00	28-06-2024 17:00
9	1ª Entrega do TFC	02-10-2023 8:00	02-10-2023 8:00
10	2ª Entrega do TFC	19-01-2024 9:00	19-01-2024 9:00
11	3ª Entrega do TFC	14-04-2024 8:00	14-04-2024 8:00
12	Entrega Final do TFC	28-06-2024 8:00	28-06-2024 8:00

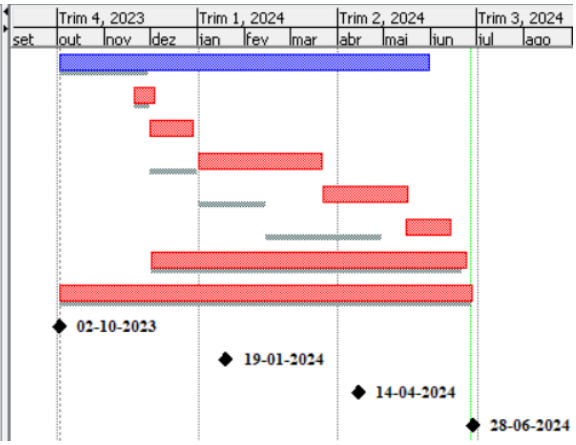


Figura 7 - Diagrama do calendário em formato Gantt (realizado através do software Project Libre)

7 Conclusão e trabalhos futuros

Este trabalho teve desde início um tema muito pertinente e fundamental nos dias de hoje, onde as temperaturas continuam a aumentar assim como a escassez de água, sendo então um dever de todos trabalhar de modo a evitar perdas de água.

Com este trabalho foi possível trabalhar mais nas componentes lecionadas ao longo do curso, aprofundando conhecimentos e ganhando ainda mais alguns. Este trabalho possibilitou também perceber como funcionam os sistemas de distribuição de água. Salienta-se que, uma vez que se trata de um trabalho de investigação, não é possível apresentar um vídeo sobre o mesmo e, sendo utilizados dados confidenciais, apenas é disponibilizado o código no GitHub (sem dados).

Em termos de trabalho futuro, são descritos alguns pontos. No caso do preenchimento de falhas, poderá ser feita uma análise mais detalhada do tipo de falhas existentes e de quais os melhores métodos para cada uma delas. Além disso, uma análise baseada na distinção entre os horários diurnos e noturnos poderá também ser vantajosa. Outro ponto onde poderá haver trabalho futuro é na etapa de deteção de eventos anómalos. Neste trabalho, não houve tempo para aplicar métodos mais avançados de deteção de eventos anómalos, inclusive já foram desenvolvidos métodos em R com base na tese de M. Silva (2016) [11] que permitem fazer uma deteção de eventos anómalos utilizando os padrões de consumo de água. A aplicação de métodos como este possibilitará que sejam retirados mais alguns indicadores de qualidade a ser acrescentados aos mencionados neste trabalho.

Bibliografia

[1] B. Ferreira. Real-time Pipe Burst Location Using Artificial Intelligence Techniques, Tese de Doutoramento, Instituto Superior Técnico, 2023.

[2] D. Loureiro. Metodologias de análise de consumos para a gestão eficiente de sistemas de distribuição de água, Tese de Doutoramento, Instituto Superior Técnico, 2010.

[3] D. Loureiro, C. Amado, A. Martins, D. Vitorino, A. Mamade, S. Coelho. Water distribution systems flow monitoring and anomalous event detection: A practical approach, Urban Water Journal, 2015.

[4] D. Loureiro, H. Alegre, S. T. Coelho, A. Martins and A. Mamade. A new approach to improve water loss control using smart metering data, Water Science & Technology: Water Supply, 2011.

[5] DEISI, Regulamento de Trabalho Final de Curso, 2021.

[6] ERSAR. Relatório Anual dos Serviços de Águas em Portugal (2022) Volume 1 – Caracterizaçãodo setor de águas e resíduos, 2023.

[7] ERSAR. Decisão sobre a definição dos valores de ANF a considerar no cálculo da TRH, para efeitos de repercussão no utilizador final, 2023.

[8] H. Alegre, S. Coelho, M. Almeida, P. Vieira. Controlo de perdas de água em sistemas públicos de adução e distribuição, Série Guias Técnicos, 3, 2005.

[9] J. Quevedo, V. Puig, G. Cembrano, J. Blanch, J. Aguilar, D. Saporta, G. Benito , M. Hedo, A. Molina. Validation and Reconstruction of Flow Meter Data in the Barcelona Water Distribution Network, Control Engineering Practice, 2010.

[10] Livera, A.M., Hyndman, R.J., & Snyder, R. D. (2011), Forecasting time series with complex seasonal patterns using exponential smoothing, Journal of the American Statistical Association, 106(496), 1513-1527.

[11] M. Silva. Modelação da Incerteza e Deteção de Outliers para Melhoria do Diagnóstico de Perdas em Sistemas de Abastecimento de Água, Tese de Mestrado, Instituto Superior Técnico, 2016.

[12] R. Barrela. Data reconstruction of flow time series in water distribuition networks, Tese de Mestrado, Instituto Superior Técnico, 2015.

[13] S. Mounce, R. Mounce, J. Boxall. Novelty detection for time series data analysis in water distribution systems using support vector machines, Journal of Hydroinformatics, 2011.

[14] Tukey, J. W. Exploratory data analysis. Addison-Wesley Publishing Company, 1977

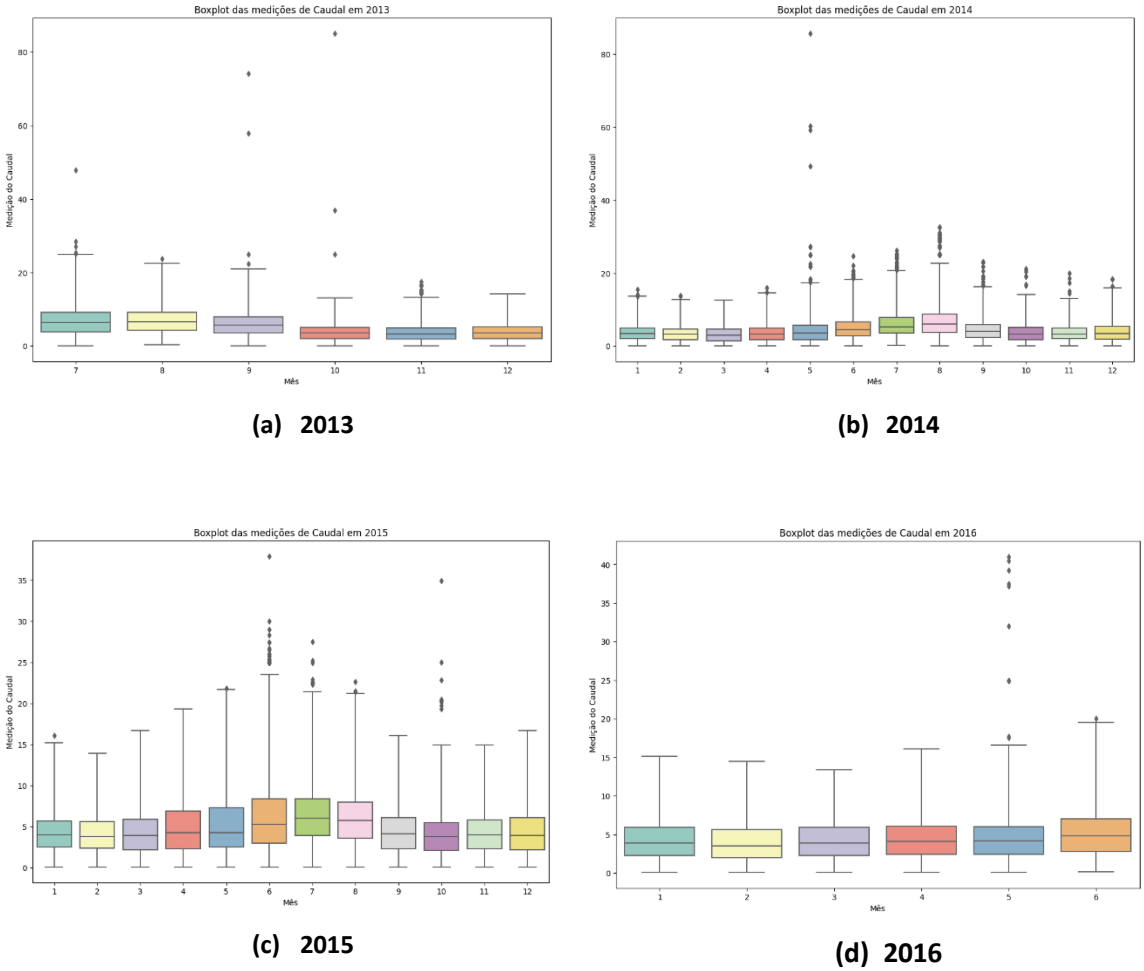
[15] Universidade Lusófona de Humanidades e Tecnologia, www.ulusofona.pt, 2021.

Glossário

DMA	District Metered Areas
ERSAR	Entidade Reguladora dos Serviços de Água e Resíduos
IQR	amplitude inter-quartil
LEI	Licenciatura em Engenharia Informática
LI	Limite inferior
LNEC	Laboratório Nacional de Engenharia Civil
LS	Limite superior
Q1	1º Quartil
Q3	3º Quartil
TFC	Trabalho Final de Curso
UC	Unidade Curricular
ZMC	Zonas de Medição Controlada

Anexos

Análise Exploratória de AL



(a) 2013 **(b) 2014**

(c) 2015 **(d) 2016**

Figura 8 - Diagrama de caixas dos valores de caudal em AL

Análise Exploratória de AR

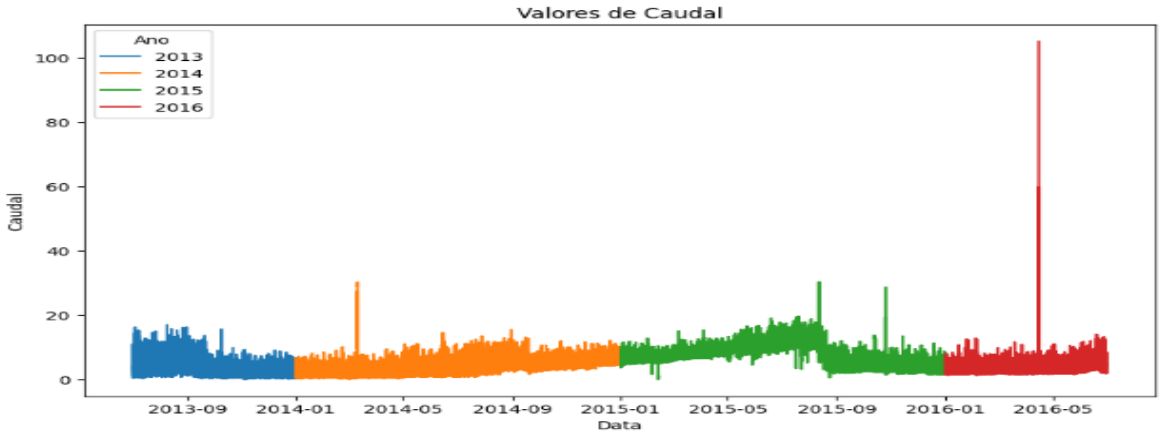


Figura 9 - Valores de caudal em AR

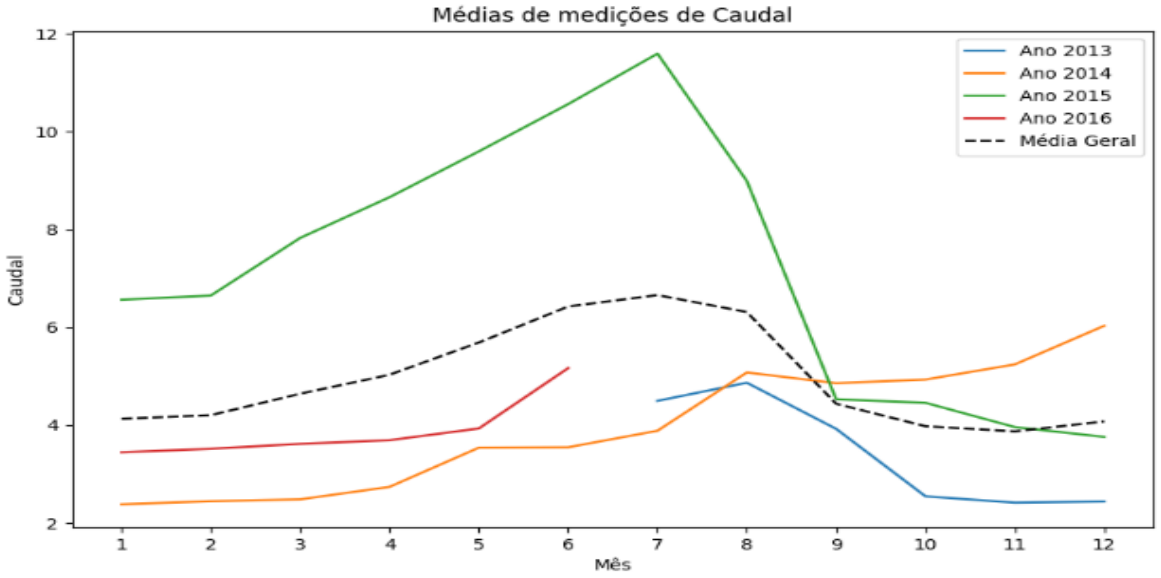
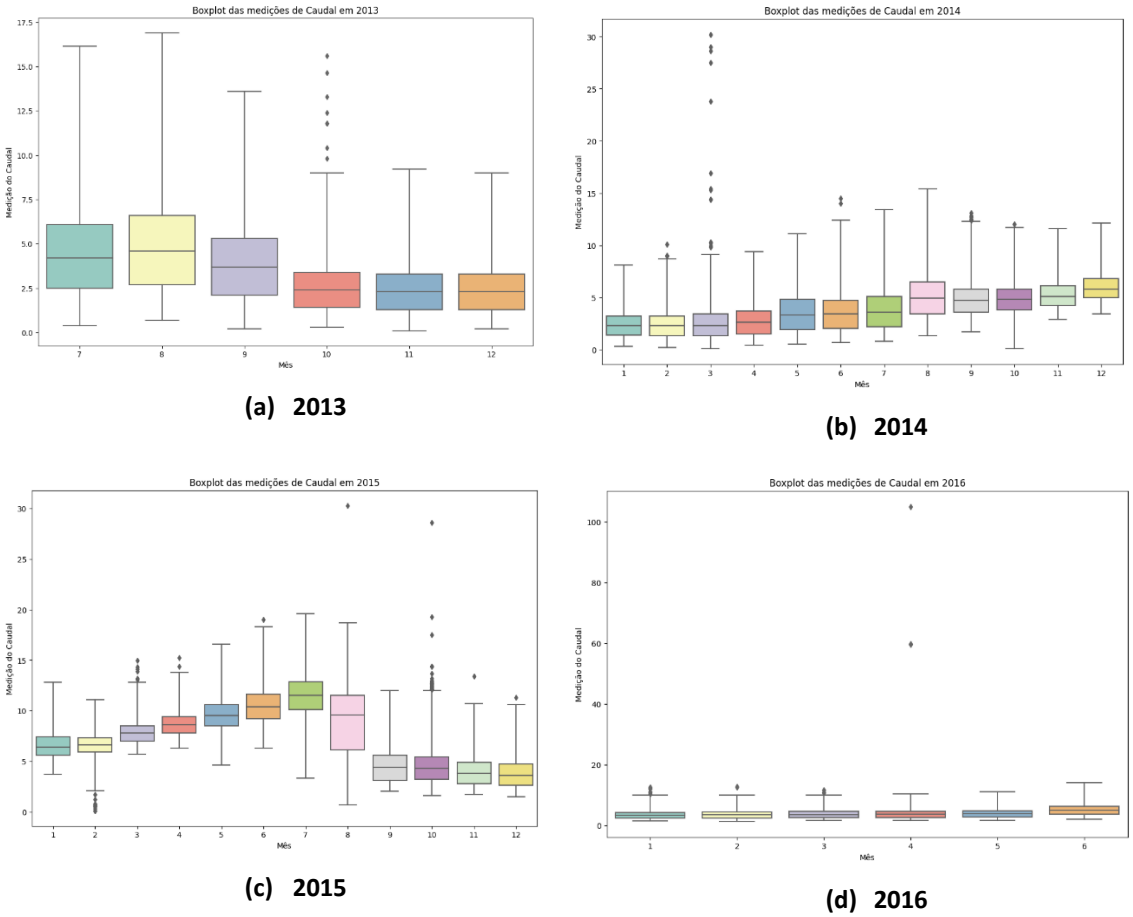


Figura 10 - Médias dos valores de caudal em AR



(a) 2013

(b) 2014

(c) 2015

(d) 2016

Figura 11 - Diagrama de caixas dos valores de caudal em AR

Análise Exploratória de CE

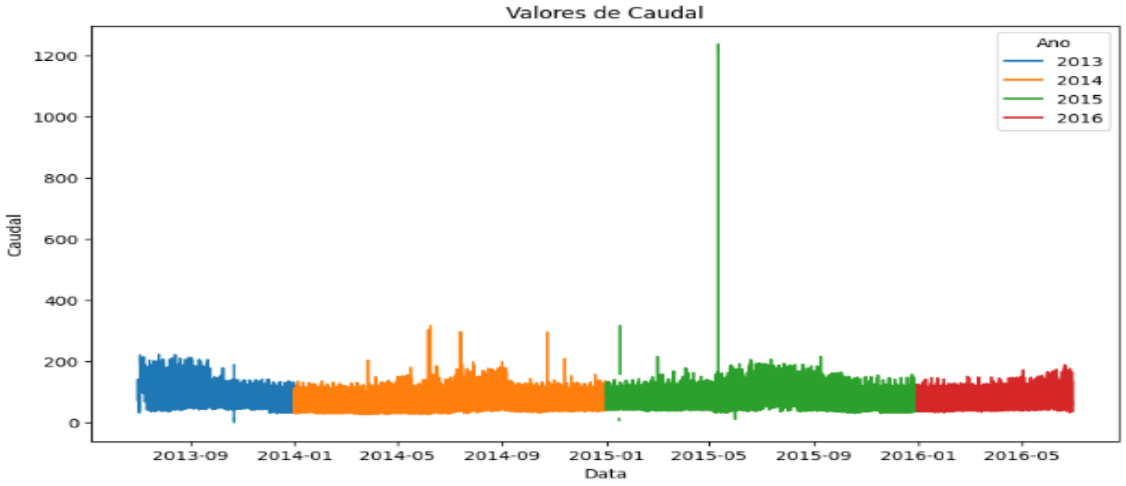


Figura 12 - Valores de caudal em CE

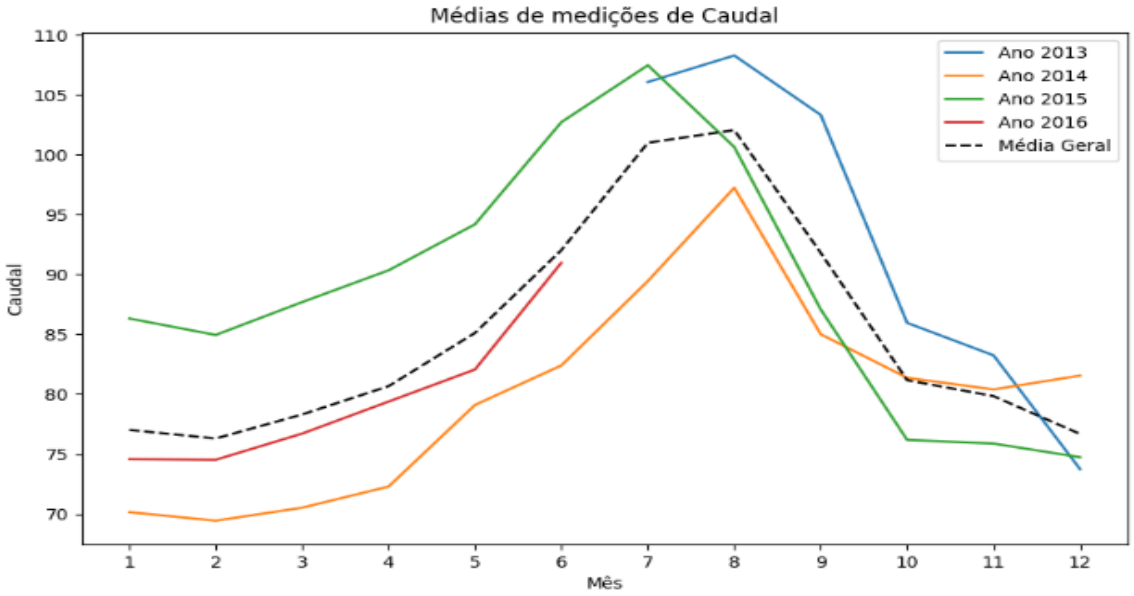
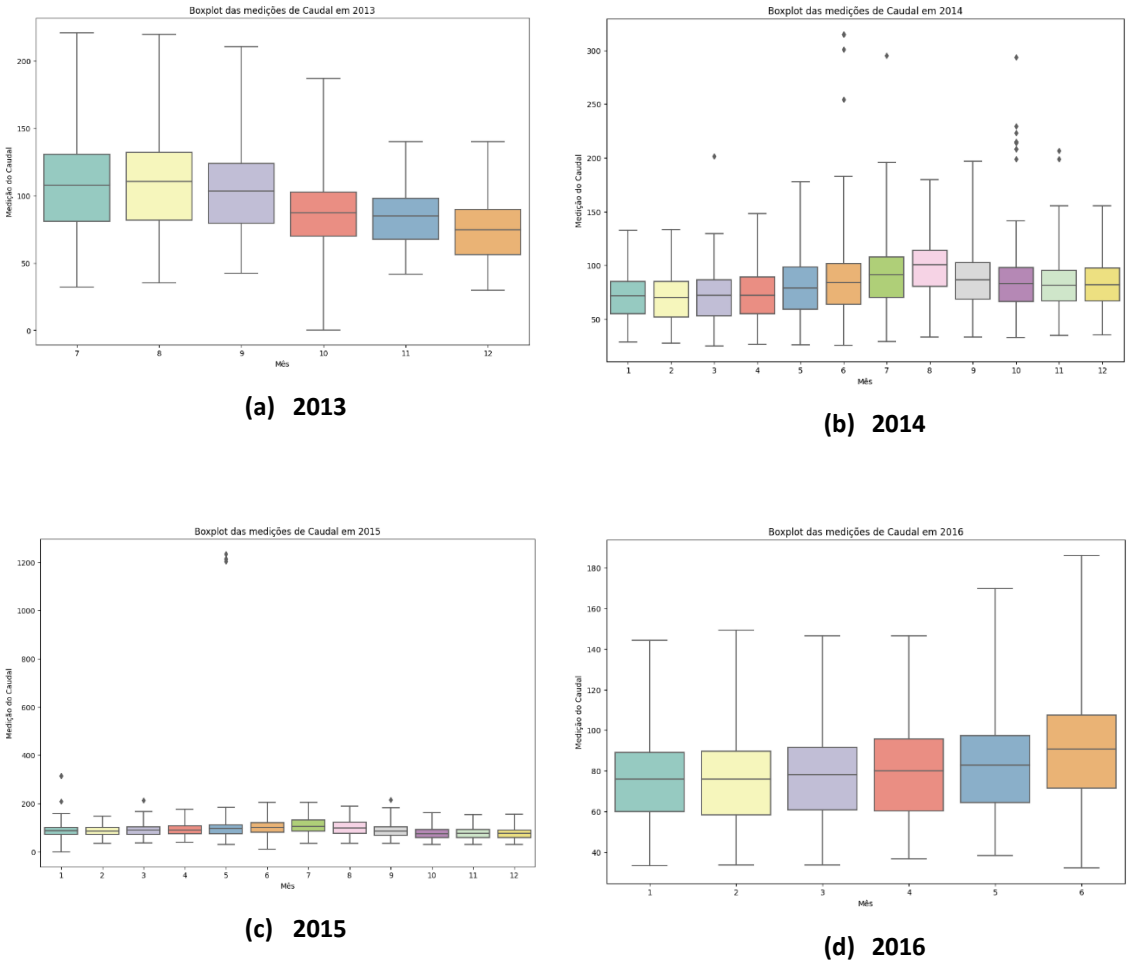


Figura 13 - Médias dos valores de caudal em CE



(a) 2013

(b) 2014

(c) 2015

(d) 2016

Figura 14 - Diagrama de caixas dos valores de caudal em CE

Análise Exploratória de MA

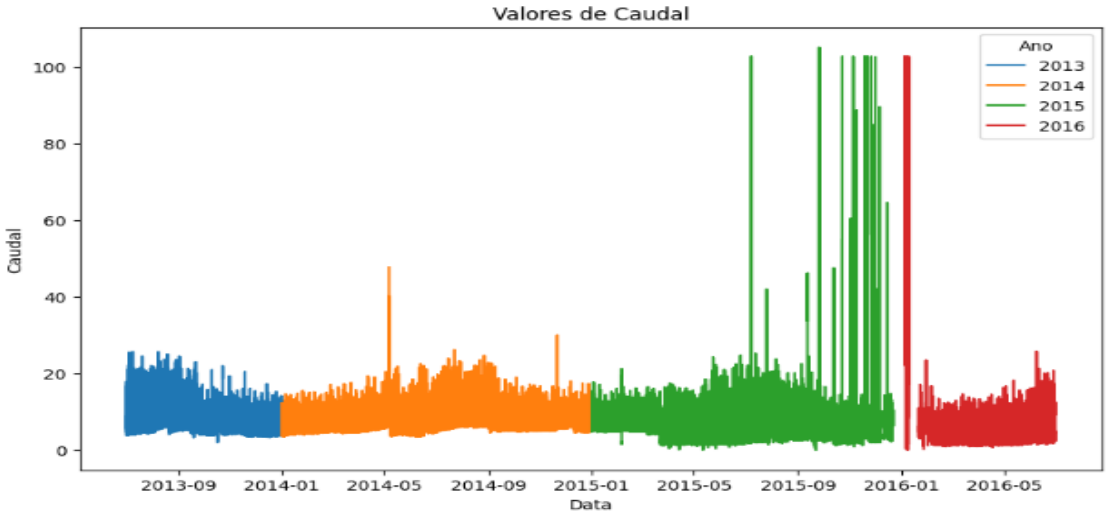


Figura 15 - Valores de caudal em MA

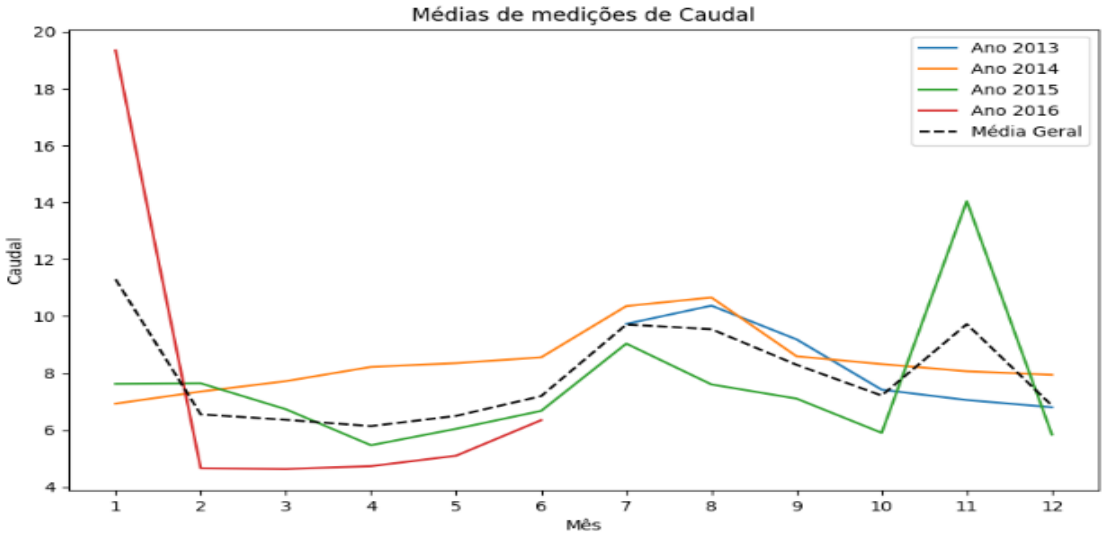
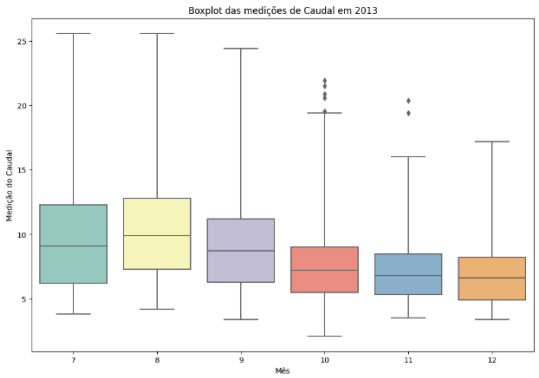
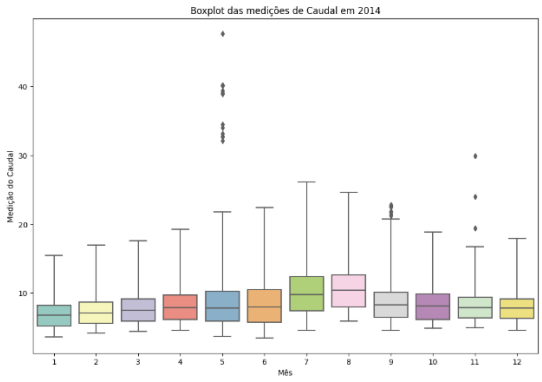


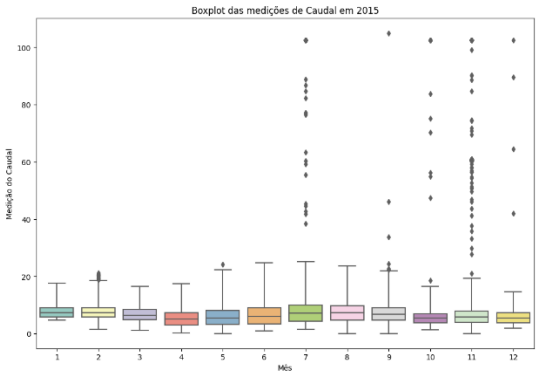
Figura 16 - Médias dos valores de caudal em MA



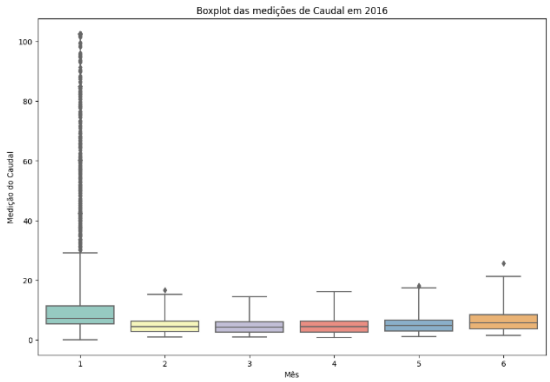
(a) 2013



(b) 2014



(c) 2015



(d) 2016

Figura 17 - Diagrama de caixas dos valores de caudal em MA

Análise Exploratória de PC

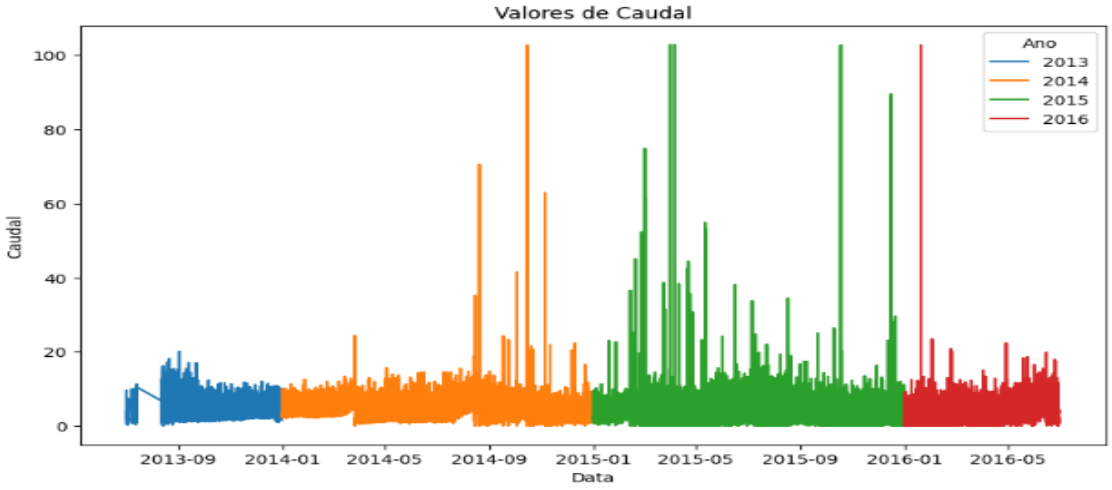
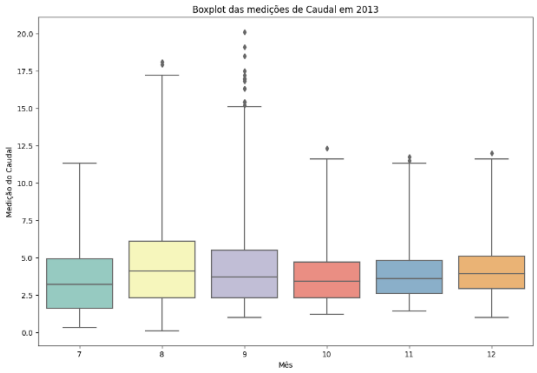
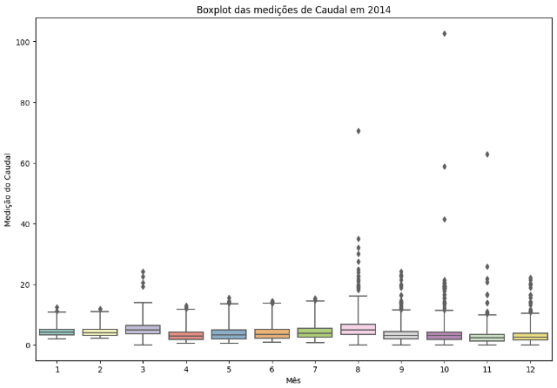


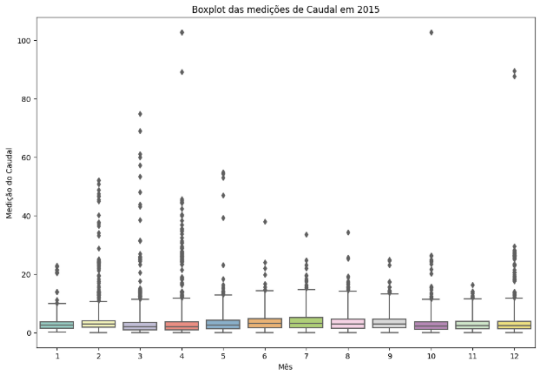
Figura 18 - Valores de caudal em PC



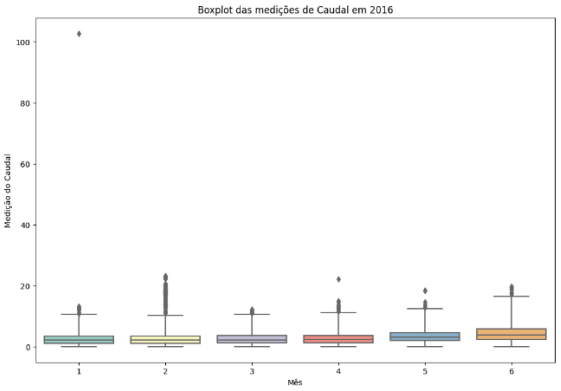
(a) 2013



(b) 2014



(c) 2015



(d) 2016

Figura 19 - Diagrama de caixas dos valores de caudal em PC