UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Previsão de Curto Prazo do Consumo de Energia

Sérgio Manuel da Conceição Duarte

Mestrado Integrado em Engenharia da Energia e do Ambiente

Dissertação orientada por:
Professora Doutora Ana Isabel Lopes Estanqueiro (FCUL)
Mestre António Manuel Vitoriano Couto (LNEG)

2019

“Prediction is very difficult, especially about the future.”

- Niels Bohr

Agradecimentos

À minha mãe, por todo o apoio que sempre me deu.

À Inês, pela motivação e pela paciência que teve comigo nas horas mais difíceis.

A todos aqueles que me acompanharam nas longas horas de trabalho.

Gostaria de agradecer também à Professora Doutora Ana Estanqueiro e ao Mestre António Couto pela oportunidade de desenvolver este trabalho.

Um especial agradecimento ao António pela disponibilidade, orientação e sabedoria partilhada ao longo deste percurso.

Resumo

O combate às alterações climáticas, bem como a redução da dependência energética externa passam pela instalação e exploração em larga escala de novas fontes energética renováveis, endógenas e não poluentes. Contudo, a introdução destas fontes no sistema electroprodutor (SE), com caráter estocástico, confere um nível de incerteza adicional no equilíbrio do mesmo. Neste equilíbrio, é fulcral atuar não só no lado da geração, mas igualmente no lado da procura, em oposição à perspetiva tradicional da gestão dos SEs, em que predomina o paradigma que a oferta deve estar sempre preparada para seguir o consumo, *i.e.*, satisfazer totalmente, a procura, cujo comportamento é, tipicamente, considerado incontrolável e inelástico.

Uma das formas mais consensuais para permitir esta mudança, assenta no conceito de gestão do consumo (*Demand Side Management*), que tem por objetivo flexibilizar o consumo, de modo a que este se adapte a uma produção variável no tempo ou em situações de constrangimento ou de estímulos tarifários. No entanto é necessário ter uma boa previsão do mesmo, de forma a solicitar atempadamente esta resposta do lado do consumo.

Com a necessidade de previsões fidedignas como pano de fundo, na presente dissertação é proposta a implementação e comparação de vários modelos, de previsão a curto prazo (24h), utilizando três métodos diferentes, sendo estes posteriormente comparados com um método de referência (*baseline*). A *baseline* utilizada consiste numa regressão linear simples, utilizando o consumo de energia elétrica verificado no instante *t-24horas* como variável independente. Os três métodos utilizados foram a Regressão Linear Multivariada (MLR), k-vizinhos mais próximos (KNN) e uma Rede Neuronal Artificial (ANN). Recorrendo a uma técnica estatística de agrupamento de dados (k-*medoids*), é ainda feita uma identificação dos perfis diários de consumo presentes na série temporal em análise, a identificar padrões diários, semanais e sazonais. Estes métodos foram aplicados à série de consumo habitacional para Portugal, BTN C, disponibilizada publicamente pela REN, utilizando os valores registados de 2014 a 2018 (inclusive).

No problema em estudo a Rede Neuronal Artificial foi identificada como o melhor método. Foram obtidos MAPE de 5,6%, 4,3% e 4,2% e RMSE de 13,4MW, 11,7MW e 10,7MW para a MLR, KNN e ANN, respetivamente. Comparativamente, a *baseline* conseguiu um MAPE de 7,8% e um RMSE de 19,3 MW.

Num nível mais granular, foram analisados em detalhe os desvios na previsão e identificadas as horas de maior consumo como as mais problemáticas de prever. O mesmo também se verificou ao nível dos meses do ano, onde os meses mais frios demonstraram ser os mais problemáticos, não só pelo o nível de intensidade do valor mas devido à variabilidade que existe nestes meses. Ao nível diário, os dias de transição de regime (sábado e segunda-feira) e o domingo apresentaram erros consideravelmente mais elevados relativamente aos restantes dias da semana.

Com este trabalho, as conclusões retiradas permitem demonstrar a importância e a vantagem da aplicação das metodologias de i) agregação para compreender e caracterizar os diferentes perfis de consumo de energia elétrica e ii) previsão a curto prazo do consumo de energia elétrica com recurso ao método de aprendizagem automática, nomeadamente, Redes Neuronais Artificiais.

Palavras Chave: *Previsão do Consumo, Perfis diários de consumo, Rede Neuronal Artificial, Regressão Linear Multivariada, K-Vizinhos mais próximos.*

Abstract

Clean, endogenous renewable energy sources are the key to stopping (or at least slowing) climate change, as well as reducing external energy dependency. However, the large-scale integration of these stochastic sources introduces an increasing uncertainty in the electrical power system balance. This balance will need to rely not only in generation side management, but also on demand side management, as opposed to the traditional power system management paradigm, which dictates that generation should always be ready to follow demand, whose is deemed uncontrollable.

Strategies such as Demand Side Management have been devised to attenuate this uncertainty. The purpose of this strategy is to provide flexibility for the power system through the electricity consumption according to the available renewable power production, or grid constraints or even tariff incentives. This entails a need for accurate consumption forecasts to enable a proper demand response.

With the need for an accurate forecast as motivation, the present dissertation proposes modeling, through various methods of the electrical load considering a short-term horizon - 24 ahead. The modelling will be done by three different methods: Multiple Linear Regression (MLR), k-nearest neighbors (KNN) and an Artificial Neural Network (ANN). The models created by each method will then be compared against a baseline, a Simple Linear Regression using the load value at $t - 24h$ as the independent variable. The typical load profiles are also evaluated, via a clustering method (k-medoids), in order to identify daily, weekly and seasonal patterns present in the data. These methods were applied to a household load time series for Portugal, BTN C, for the years 2014 through 2018, made publicly available by REN.

At the end of this analysis, the Artificial Neural Network was identified as the best method, among those studied, in the present case study. The errors obtained for each method were a MAPE of 5.6%, 4.3% and 4.2%, and a RMSE of 13.4MW, 11.7MW and 10.7MW for MLR, KNN e ANN, respectively. By comparison, the baseline achieved a MAPE of 7.8% and a RMSE of 19.3 MW.

On a more granular level, forecast error showed that the hours with higher demand were more difficult to accurately predict, along with higher demand months (colder months in this case). Moreover, the regime transition days (Saturdays and Mondays) as well as Sundays are the ones with the biggest errors.

The conclusions drawn from the work developed show the importance and advantages of *i*) typical electrical load profile aggregation analysis and *ii*) using machine learning methods to perform short-term electrical load forecast, specifically Artificial Neural Networks.

Key Words: *Short-term forecast, daily load profiles, Artificial Neural Networks, Multiple Linear Regression, k-nearest neighbors.*

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ANN	*Artificial Neural Network (rede neuronal artificial)*
ANOVA	*Analysis of Variance (Análise de Variância)*
ARIMA	*Auto-Regressive Integrated Moving Average (modelo auto-regressivo de medias móveis integrado)*
ARMA	*Auto-Regressive Moving Average (modelo auto-regressivo de medias móveis)*
AVAC	*Aquecimento, Ventilação e Ar Condicionado*
BTN	*Baixa Tensão Normal*
CO_2	*Dióxido de Carbono*
DBSCAN	*Density-based spatial clustering of applications with noise*
FER	*Fontes de Energia Renováveis*
KNN	*K-Nearest Neighbours (k vizinhos mais próximos)*
MAPE	*Mean Absolute Percent Error (erro médio absolute percentual)*
MDA	*Mediana dos desvios absolutos*
MLR	*Multiple Linear Regression (regressão linear multivariada)*
MSE	*Mean Squared Error (erro quadrático médio)*
REN	*REN - Redes Energéticas Nacionais, SGPS, S.A.*
RMSE	*Root Mean Squared Error (raíz do erro quadrático médio)*
SE	*Sistema Eletroprodutor*
SQ	*Soma de Quadrados*
SVM	*Support Vector Machine (máquina de vetor de suporte)*

1 Introdução

1.1 Enquadramento

A energia elétrica é atualmente um dos pilares do funcionamento de todas as sociedades modernas. Grande parte das atividades do quotidiano dos países desenvolvidos, e de grande parte dos países em desenvolvimento, assentam num fornecimento fiável e ininterrupto de energia elétrica.

No entanto, atualmente, em todo o mundo, a produção de energia elétrica assenta fortemente no uso de combustíveis fósseis (carvão e gás natural, principalmente)[1], o que acarreta vários problemas, quer ambientais, quer económicos. A utilização destes combustíveis na produção de energia elétrica produz cerca de 40% das emissões globais de CO_2[1], que, sendo um gás com efeito de estufa, contribui para um aumento do aquecimento global. Este facto torna a redução das emissões associadas à produção de energia num dos grandes desafios da sociedade atual, sendo crucial a implementação de novas formas de produção de energia elétrica.

Para além da componente ambiental há ainda um fator económico e geopolítico no que toca à utilização de combustíveis fósseis. A importação de matérias primas para a produção de energia elétrica representa uma dependência externa de outras nações e dos mercados internacionais, cuja volatilidade pode afetar muito a economia nacional, e assim penalizar os consumidores.

Com o objetivo de responder a estes desafios, têm vindo a ser exploradas diferentes formas de produção de energia elétrica designadas por fontes de energia renováveis (FER). Estas fontes têm vindo a representar uma fatia cada vez maior dos sistemas electroprodutores (SE), sendo Portugal um dos países que mais tem apostado neste tipo de tecnologia, Figura 1.1. Devido aos incentivos políticos no início de século, fortemente suportados pela maturidade da tecnologia, o crescimento da produção renovável nos últimos anos assentou fortemente na energia eólica. Tendo em consideração o compromisso ambicioso de Portugal para a descarbonização do SE, é expectável um forte crescimento similar na tecnologia solar fotovoltaica.

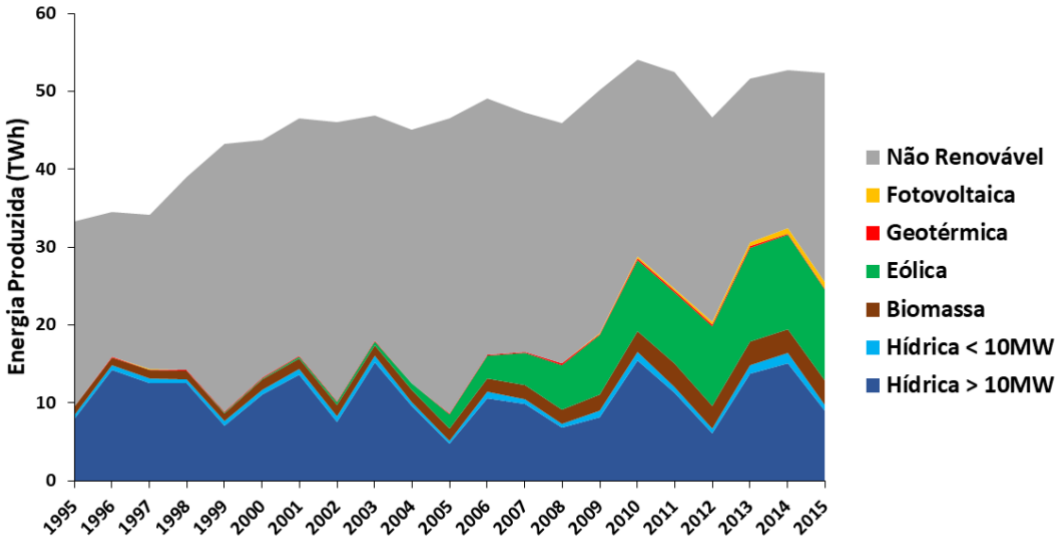

Figura 1.1: Evolução da penetração de fontes renováveis no sistema energético nacional

A crescente percentagem de FER nos SE acarreta novos desafios na sua gestão. O princípio de funcionamento de qualquer sistema electroprodutor dita que a produção tem de igualar o consumo em

qualquer instante do tempo. Atualmente isto é garantido por um portfólio de geração flexível, que segue o perfil de um consumo rígido. O crescimento da penetração de FER na produção de energia introduz um grau de incerteza adicional no equilíbrio do sistema, uma vez que estas, devido à natureza estocástica do seu recurso primário, não oferecem garantia de potência acabando por ter um papel muito passivo no SE. Para colmatar isto, continuamos a recorrer à *i)* geração flexível, usualmente designada por despachável, cuja fonte de energia primária assenta nos combustíveis fósseis ou nas centrais hídricas ou *ii)* consumo de energia através da bombagem. O recurso a estas soluções, encarece os custos gerais do SE apresentando assim um forte impacto socioeconómico. Para mitigar este impacto, e tendo em consideração a segurança e robustez do SE, é necessário que a produção de energia elétrica proveniente das FER seja objeto de previsão, que juntamente com a previsão do consumo permite a elaboração do plano de operações e respetivo escalonamento de funcionamento das centrais despacháveis.

A previsão do consumo possibilita igualmente o desenvolvimento de mecanismos para flexibilizar o consumo, de modo a que este se adapte, *e.g.*, ao perfil de produção das FER ou aos estímulos tarifários – Gestão da Procura (usualmente designado por *Demand Side Managment*), sendo crucial para a gestão otimizada do SE, e para os novos conceitos como cidades inteligentes, comunidades de energia, entre outros, que representam novas formas de gestão do binómio oferta/procura de energia.

1.2 Motivação

Os novos acordos internacionais, como o Acordo de Paris, celebrado na COP21, preveem metas ambiciosas no que toca à redução do aumento da temperatura média da Terra relativamente aos valores pré-industriais (manter este aumento "bem abaixo" do 2°C)[2]. Esta redução apenas será conseguida com uma grande diminuição do consumo de combustíveis fósseis. Neste aspeto, a integração de FER no sistema electroprodutor será de enorme importância para cumprir os objetivos em vista.

Como já foi referido no Capítulo 1.1, alguns recursos renováveis (nomeadamente o recurso eólico) têm uma grande variabilidade associada, o que leva à necessidade de desenvolver soluções que consigam responder às súbitas variações da produção elétrica. Uma solução encontrada foi a chamada Resposta da Procura (usualmente designado por *Demand Response*). Este mecanismo consiste em contrabalançar as variações na produção (devido à redução do recurso) através da redução/mobilização do consumo. Isto permitirá aumentar os níveis de penetração de FER no SE e ajudará a garantir a estabilidade e robustez necessária, sem recurso à alocação de reservas adicionais que amplificam os custos do SE ou à instalação de dispendiosos sistemas de armazenamento. Terá também, como consequência uma redução do trânsito de energia na rede em horas de ponta, o que irá reduzir os custos associados ao dimensionamento desta infraestrutura.

Uma implementação sólida de novos paradigmas de gestão dos SE, como o *Demand Response*, assenta numa boa previsão do consumo de energia quer ao nível do sistema, quer ao nível dos consumidores individuais[3], [4]. Esta necessidade serve de motivação ao trabalho apresentado na presente dissertação. Conjugando esta previsão com uma boa previsão da produção a partir de FER será possível: *i)* a manutenção dos padrões elevados da qualidade do serviço, nomeadamente no que respeita à segurança de abastecimento e robustez do sistema, bem como a otimização do seu desempenho económico; e *ii)* proceder a uma implementação adequada dos conceitos emergentes de flexibilização do consumo, permitindo assim ter um sistema electroprodutor com elevada penetração de FER, tendencialmente 100% renovável.

A execução desta dissertação assenta em três fases principais: 1) pré-processamento dos dados de consumo e variáveis meteorológicas; 2) caracterização dos perfis diários do consumo de energia elétrica; e 3) previsão determinística do consumo de energia para um horizonte temporal de 24 horas, com resolução horária. A caracterização será realizada recorrendo a uma técnica estatística de

agrupamento de dados (c*lustering*). Devido ao grande impacto verificado no consumo, a caracterização será feita tendo em consideração a sazonalidade, dias úteis e fins de semana[5]. Adicionalmente serão estudados os fatores meteorológicos com mais impacto na variabilidade do consumo de energia. Os fatores mais relevantes serão incorporados nas técnicas estatísticas de previsão determinísticas que serão aplicadas nesta dissertação. Estas técnicas serão implementadas utilizando o *software* MATLAB e dados de consumo de energia publicamente disponíveis.

1.3 Organização da Dissertação

No Capítulo 1 é apresentada a evolução da penetração das fontes de energia não despacháveis no sistema electroprodutor nacional. A partir daqui é feita a exposição da motivação e contexto do problema abordado na dissertação.

No Capítulo 2 é exposto o estado da arte relativamente ao tópico desta dissertação focando as metodologias estatísticas de agrupamento de observações mais comuns bem como os algoritmos mais utilizados para previsão do consumo de energia elétrica. São ainda identificadas e discutidas as variáveis endógenas e exogéneas mais relevantes referidas na literatura.

O Capítulo 3 apresenta os dados e a metodologia usada destacando matematicamente os conceitos nos quais se baseiam os métodos utilizados, desde os algoritmos de agrupamento até aos algoritmos de previsão utilizados.

O Capítulo 4 apresenta os resultados decorrentes da aplicação das metodologias ao caso de estudo em análise. Em específico, são expostos e avaliados os resultados obtidos através da metodologia implementada, bem como diagnosticados os desvios encontrados da previsão em relação aos valores observados.

Por fim, no Capítulo 5, apresentam-se as ilações retiradas durante este trabalho bem como possíveis desenvolvimentos futuros.

2 Estado da Arte

2.1 Enquadramento do problema da previsão consumo

O problema da previsão é transversal aos mais diversos setores de atividade, como o financeiro, científico, industrial, político, etc. De forma a enquadrar coerentemente o caso em estudo, uma classificação comum, embora nem sempre consensual deste tipo de problemas, relativamente ao horizonte temporal da previsão, é: muito curto-prazo, curto-prazo, médio-prazo e longo-prazo. Segundo [6], [7], estes diferentes horizontes podem ser definidos conforme apresentado na Tabela 2.1.

Tabela 2.1: Horizontes da previsão de consumo de energia e os seus objetivos

Horizonte da Previsão	Escala Temporal	Objetivo
Muito Curto Prazo	< 1 hora	Previsão do consumo de edifícios no contexto de *Micro-Grids*
Curto Prazo	1 hora ~ 1 semana	Operação do sistema e participação em mercado de energia elétrica
Médio Prazo	1 semana ~ 1 ano	Operação do sistema (aquisição de combustível, manutenção, etc)
Longo Prazo	> 1 ano	Planeamento e manutenção do sistema electroprodutor

O tipo de dados de interesse para este trabalho é conhecido como série temporal. Estes consistem tipicamente numa sequência, organizada cronologicamente, de uma determinada variável de interesse, aqui, o consumo de energia elétrica [8] (Figura 2.1).

A análise deste tipo de dados recai em duas categorias generalizadas: métodos[1] qualitativos e métodos quantitativos. Os métodos qualitativos são resultado de um julgamento subjetivo relativamente à variável de interesse [9]. Apesar esta abordagem ser dada por "*experts*" na área em estudo, normalmente não assenta numa base estatística de análise de dados históricos, quer por estes não estarem disponíveis, quer por não existirem em quantidade que permita essa avaliação [10]. Os métodos quantitativos recorrem aos dados históricos disponíveis para estabelecer o comportamento da variável em estudo e apresentar uma relação estatística formal entre valores passados e presentes. Estes métodos podem ser classificados em univariados no caso de recorrerem apenas a uma variável independente para estimação da variável dependente, e multivariados no caso de recorrerem a mais do que uma variável independente [10]. De acordo com [9], para se obter resultados com elevada precisão (ou melhorar os resultados) a aplicação prática da previsão requer, muitas vezes, uma combinação de métodos qualitativos e quantitativos.

1 Os termos método e modelo são frequentemente empregues sem distinção. Aqui será feita a distinção referida em [9]. Um método é um procedimento destinado a estimar valores futuros com base em registos históricos e presentes, sem que seja necessário ter por base um modelo probabilístico. Um modelo consiste num conjunto de parâmetros, ajustados aos dados em análise, de forma a ser obtida a previsão pretendida.

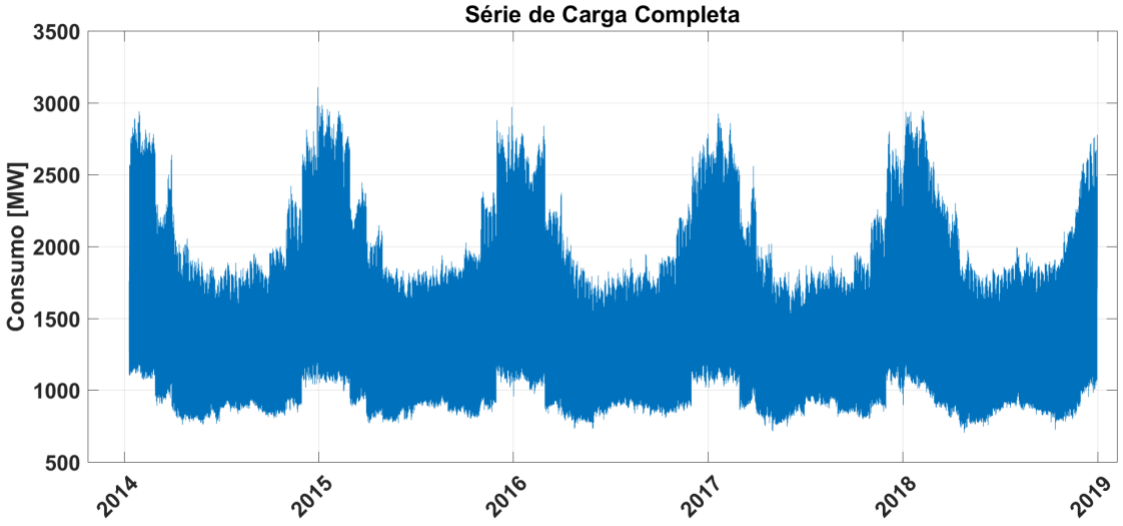

Figura 2.1: Carga BTNC horária do sistema electroprodutor português entre 2014 e 2019

Relativamente ao prazo da previsão, independentemente ser curto, médio ou longo, importa definir dois conceitos: horizonte e intervalo [10]. O horizonte da previsão representa o número de períodos futuros para o qual a previsão deve ser efetuada. O intervalo da previsão é a frequência com que os resultados são apresentados.

2.2 O processo da previsão

O processo de previsão tem como objetivo transformar uma ou mais variáveis independentes (*inputs*) em uma ou mais variáveis dependentes (*outputs*). Este processo é caracterizado por alguns passos chave [10]:

- **Definição do problema**
- **Recolha de dados**
- **Análise descritiva dos dados**
- **Escolha do modelo de previsão**
- **Validação do modelo**
- **Previsão**
- **Avaliação de desempenho**

A identificação do problema consiste na avaliação do prazo de previsão, horizontes e intervalos de previsão bem como dos erros admissíveis nos resultados, entre outros fatores que poderão ser relevantes para o caso em questão.

Na fase de recolha de dados deverão ser a recolhidas as variáveis em estudo (objeto da previsão) e as variáveis independentes necessárias à construção do modelo de previsão.

Para a análise descritiva dos dados é necessário, em primeiro lugar e estando a trabalhar com uma série temporal, ter em conta que observações sucessivas não são acontecimentos independentes [9], e como tal, a ordem das observações deve ser respeitada. Segundo [10], de forma a obter uma maior sensibilidade dos dados em análise, os mesmos deverão ser representados na forma de gráfico. Este procedimento permite identificar anomalias nos dados, tendências e sazonalidades que de outra forma poderiam não ser evidentes.

Após uma análise dos dados, aplica-se um método estatístico de previsão. Esta tarefa consiste na escolha e ajustamento de um ou mais modelos aos dados analisados, isto é, que reproduzam a variável dependente, em função da variável (ou variáveis) independente(s), dentro de uma determinada margem de erro.

Depois de selecionado, o método deve ser validado. Esta validação é feita através da avaliação da performance da previsão. Para este efeito, normalmente, o método é ajustado apenas a uma parte dos dados disponíveis, sendo os restantes utilizados para validação do mesmo.

Uma vez validado, o método é implementado e o seu controlo é feito continuadamente através da medição do erro (*e.g.*, viés) das previsões efetuadas, de modo a verificar a validade continuada do método, e, se necessária, a sua atualização.

2.3 Métodos

Neste subcapítulo apresenta-se uma revisão de literatura relevante que aborda os principais passos no processo da previsão, nomeadamente, Pré-processamento dos dados, Caracterização dos perfis de Consumo, Variáveis independentes utilizadas na previsão, Métodos de previsão e Métricas de erro utilizadas, de modo a *i)* enquadrar o trabalho atual e *ii)* obter alguma sensibilidade quanto aos métodos usualmente utilizados neste tipo de problemas.

2.3.1 Pré-processamento dos dados

Antes de aplicar os métodos estatísticos de previsão, é comum aplicarem-se procedimentos de pré-processamentos aos dados em análise [6], [11], [12]. Os tipos de pré-processamento mais comuns são: a limpeza de dados, integração, transformação e redução dimensional. Estes tratamentos podem ser utilizados em várias combinações ou sozinhos. A limpeza de dados consiste na remoção ou modificação de valores de valores incorretos e na introdução de valores em falta. A integração consiste na combinação de dados de dados de fontes diferente (*e.g.*, carga registada por vários sensores diferentes juntamente com a temperatura registada por outra fonte). A transformação consiste, por exemplo, na normalização dos dados de modo a colocá-los numa escala pré-definida (*e.g.* [0, 1]) ou na transformação de um valor registado de 15 em 15 minutos numa média horária, reduzindo assim o número das observações. A redução dimensional consiste na redução do número de variáveis existentes. Esta redução pode ser feita, por exemplo, através de análise de componentes principais ou da avaliação da correlação entre variáveis e eliminação de variáveis independentes fortemente correlacionadas, por forma a remover possíveis colinearidades presentes nas no espaço destes atributos. Todas estas técnicas pretendem garantir a robustez dos dados, trazendo igualmente benefícios na melhoria da eficiência computacional [12], [13].

Outro tipo de pré-processamento é a decomposição e classificação dos dados por agrupamento (*clustering*) [6], [11]. A decomposição, no contexto da análise do consumo de energia elétrica, refere-se à separação dos efeitos sazonais, semanais e de dias especiais (feriados) [14]. Por outro lado, a classificação em diferentes conjuntos/agrupamentos (*clusters)* refere-se à aglomeração das observações por gau de semelhança (*e.g.* menor distância euclidiana entre observações), e à separação desses conjuntos por dissemelhança, que pode ajudar a identificar perfis de consumo típicos e ainda levar a melhorias nos resultados da previsão [7], [11], [15], [16].

2.3.2 Caracterização dos perfis de Consumo

Os diagramas de consumo de energia elétricas seguem padrões tipicamente bem definidos. Contudo, esses padrões podem variar de acordo com o tipo de consumidor, estação do ano, entre outros fatores. De acordo com [7], [11], [15]–[19], o agrupamento das observações em conjuntos pode ajudar a identificar perfis típicos nos dados de consumo.

Estes perfis típicos podem ser segmentados em duas categorias: 1) tipo de consumidor [20] e 2) tipo de dia [19]. Esta diferenciação deve-se a tipo de dados em análise. Por exemplo, para obter os perfis típicos para os diferentes tipos de consumidor são necessários os registos de consumo individuais (habitacional ou industrial). Este tipo de caracterização do consumo tem especial importância no contexto do mercado liberalizado como forma de criar ofertas específicas aos diferentes segmentos do mercado energético. Prevê-se ainda que num futuro próximo, este tipo de caracterização possa vir a ser importante num contexto de *Smart Grid*, onde todas as habitações terão contadores bidirecionais com registos de contagem individual históricos. Estes registos permitirão fazer previsões do consumo e ajustar o mesmo de acordo com as necessidades do sistema [18], [20].

A obtenção dos perfis típicos para os diferentes dias da semana, para além de ter também valor na criação de ofertas específicas, leva também uma melhor operação do sistema como um todo, permitindo agrupar dias semelhantes e ajustar a produção às necessidades do consumo [19].

Este agrupamento pode ainda trazer benefícios à previsão, uma vez que divide a série temporal em conjuntos com observações mais semelhantes entre si, permitindo um melhor ajuste dos modelos de previsão [15], [16].

As técnicas de agrupamento particionam um conjunto de dados em subconjuntos com base num determinado critério de semelhança (*e.g.*, distância euclidiana entre as observações). Apesar da grande quantidade de algoritmos de agrupamento de dados existentes na literatura atual, não é possível determinar todos os subconjuntos existentes em todas as bases de dados (*i.e.*, os métodos não são perfeitos; a estrutura dos dados reais raramente é conhecida; mesmo em situações em que a estrutura é conhecida, estes conjuntos podem não ser matematicamente determináveis) [20], pelo que pode ser necessário testar vários métodos antes selecionar o mais indicado para o conjunto de dados em estudo. Este número de subconjuntos tem que ser resultante de um equilíbrio entre a capacidade computacional disponível e a representatividade física de cada um dos subconjuntos gerados. Isto que requer uma análise em dois âmbitos *i)* complexidade computacional dos algoritmos utilizados e *ii)* conhecimento do domínio em estudo.

2.3.2.1 Técnicas de Agrupamento (*Clustering)*

Apesar de alguns autores [21], [22] dividirem as técnicas de Agrupamento em mais categorias, estas podem ser divididas em quatro categorias principais (Figura 2.2): Agrupamento Não-Hierárquico (*Partitional* ou *Non-Hierarchical Clustering*), Agrupamento Hierárquico (*Hierarchical Clustering*), Agrupamento por densidade (*Density-based Clustering*) e Agrupamento em grelha (*Grid-based Clustering*) [23], [24].

Figura 2.2: Categorias das técnicas de agrupamento de dados

- **Agrupamento Não-Hierárquico (*Partitional or Non Hierarchical Clustering*)**

Um método de agrupamento diviso reparte um conjunto de *n* observações não classificadas em *k* conjuntos.

Alguns dos algoritmos divisivos mais utilizados são o *k-means* e *k-medoids* (*Partition Around Medoids)* [21], [23]. Em ambos os algoritmos, o número de conjuntos, *k*, tem de ser definido *a priori* pelo utilizador, o que introduz viés na obtenção dos mesmos, impedindo a obtenção dos grupos naturalmente presentes nos dados, sendo até impossível determinar o número ideal dos mesmo em bases de dados muito extensas. No entanto, existem métodos heurísticos que permitem contornar este problema, encontrando uma estimativa para *k*. Por exemplo, [25] refere o Método do Joelho (do inglês, *knee*) também chamado de *Elbow Method* (Método do Cotovelo)[2] [24], [26] onde é calculado valor da função objetivo para vários valores de *k* (neste caso a soma das distâncias euclidianas entre conjuntos). Ao criar um vetor com estes valores é então possível obter o ponto em que a redução do valor dessa função é menor com o aumento do número de conjuntos, ou seja, onde a curva começa a atingir a parte mais "plana" (Figura 2.3).

Após a definição do número de conjuntos, os centros dos mesmos são criados (aleatoriamente para o *k-means* e utilizando observações aleatórias no *k-medoids*) e a partir destes, os dados são divididos

[2] A diferença de nomenclatura encontra-se relacionada com a concavidade da curva da função objetivo, que não tem qualquer influência no resultado do método.

ao longo de várias iterações (onde se vão atribuindo as observações a novos centros de conjuntos) através da minimização de uma função objetivo tipicamente, a distância euclidiana entre as observações dentro de cada conjunto.

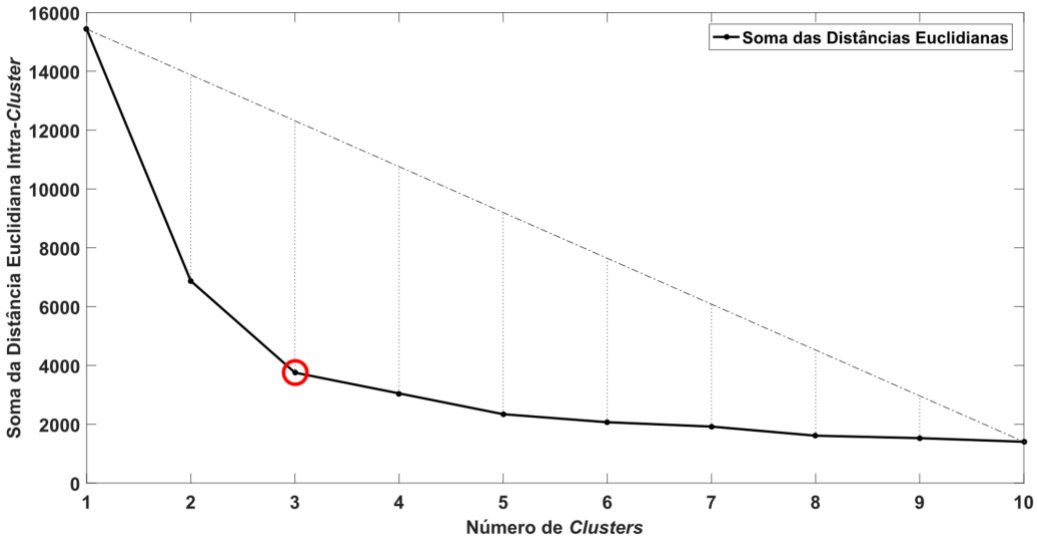

Figura 2.3: Exemplificação do método do Cotovelo

A seleção inicial dos centros dos conjuntos tem impacto no resultado dos mesmos, visto que o algoritmo converge para um ótimo local que pode estar muito afastado do ótimo global. Apesar disto, estes algoritmos convergem rapidamente pelo que não é computacionalmente exigente aplica-los [27].

- **Agrupamento Hierárquico (*Hierarchical Clustering*)**

As técnicas de Agrupamento hierárquicas podem ser classificadas em: aglomerativas e divisivas.

As técnicas divisivas consideram, inicialmente, todo o conjunto de observações como um único conjunto e procedem à divisão sucessiva deste conjunto inicial e dos subsequentes até atingir um resultado satisfatório. A Figura 2.4 apresenta uma representação visual do tipo de particionamento obtido por uma técnica hierárquica. Desta forma, o utilizador podes escolher o nível de agrupamento que melhor se enquadra na análise em estudo. Em contraste, as técnicas aglomerativas consideram cada observação como um conjunto individual, e procedem à sua combinação sucessiva até chegarem ao resultado final [28].

No geral, estas técnicas apresentam como grande limitação a capacidade de puderem reajustar os conjuntos depois de dividirem os dados nas técnicas divisivas ou depois de os juntarem nas técnicas aglomerativas. De modo a colmatar esta limitação, usualmente estes são combinados com outros algoritmos sob a forma de um algoritmo híbrido [21].

As técnicas hierárquicas têm a vantagem de não precisarem da definição do número de conjuntos *a priori*, um dado que pode ser difícil de definir no contexto de um problema real, com dados sobre os quais não temos muita informação. No entanto, devido à complexidade computacional quadrática, este está restringido a conjunto de dados de pequena dimensão [29].

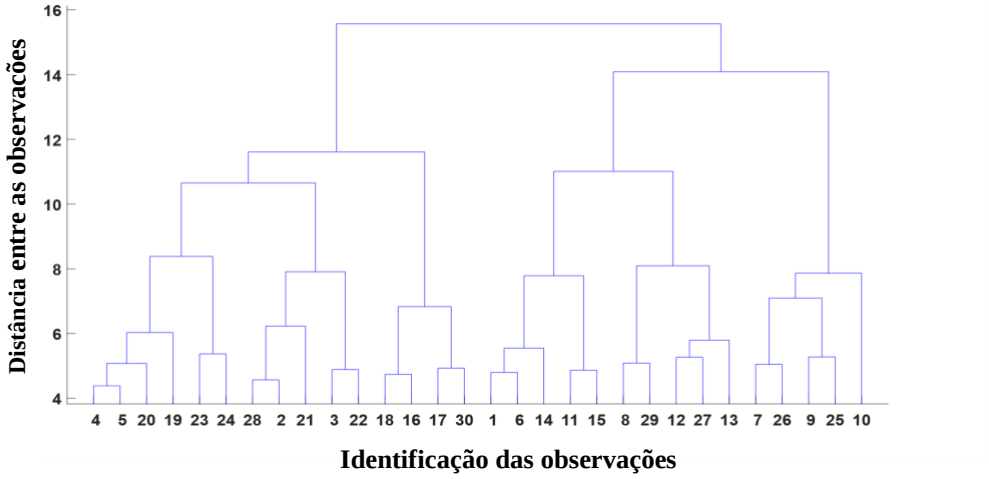

Figura 2.4: Exemplo de um Dendrograma onde está representada uma técnica de agrupamento hierárquica. O eixo dos yy's representa a distância entre as observações que são separadas a cada divisão do dendrograma e o eixo dos xx's representa as observações em análise

- **Agrupamento por densidade (*Density-based Clustering*)**

Uma vez que os métodos de agrupamento não-hierárquicos se baseiam na distância entre observações para identificar os possíveis subconjuntos presentes nos dados em estudo, estes apenas conseguem identificar conjuntos com forma esférica visto que a distância será igual em todas as direções do espaço. A Figura 2.5 ilustra esta limitação com o *k-means*. Como os conjuntos de dados no mundo real podem ter formas arbitrária, estes métodos poderão ter dificuldades em agrupar corretamente dados que tenham distribuições mais complexas. De forma a ultrapassar esta limitação foram criados métodos baseados na densidade de observações numa determinada zona do espaço [24].

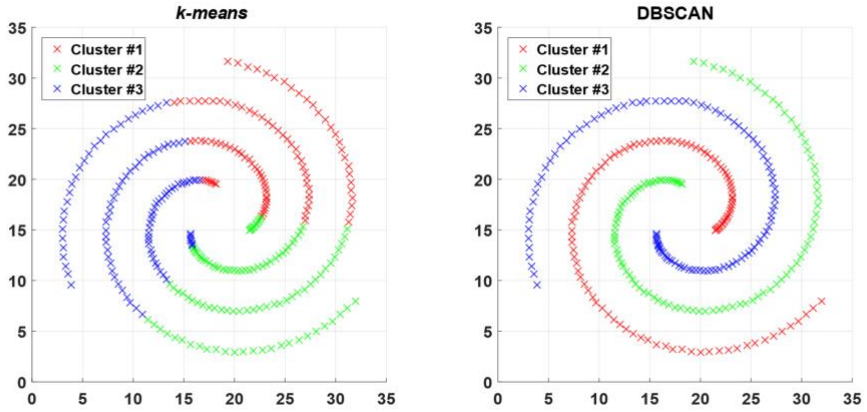

Figura 2.5: Comparação entre os algoritmos kmeans e DBSCAN

De uma forma geral, a ideia destes métodos é continuar a expandir os conjuntos enquanto a densidade de observações na "vizinhança" do ponto em estudo for superior a um determinado valor. Este método pode ser utilizado para filtrar *outliers* e ruído. Na Figura 2.5 observa-se como a estrutura inerente a um conjunto de dados é identificada com um algoritmo de agrupamento por densidade (DBSCAN) enquanto que o *k-means* (técnica de agrupamento não-hierárquica) falha a identificar esta estrutura.

- **Agrupamento em grelha (*Grid-based Clustering*)**

Os métodos de agrupamento em grelha dividem o espaço das observações num número finito de células, dando origem a uma grelha. As operações de agrupamento desenvolvem-se neste espaço quantizado. A grande vantagem desta abordagem é o reduzido esforço computacional, que é tipicamente independente do número de observações e apenas depende do número de dimensões da grelha criada [24]. Um exemplo de um algoritmo de agrupamento em grelha é o *Self Organizing Maps* (SOM). Um SOM é um tipo especial de rede neuronal artificial que mapeia as variáveis em estudo para um espaço bidimensional (*i.e.*, um mapa), organizando as mesmas de acordo com uma dada métrica de semelhança, produzindo assim os agrupamentos finais [30], como demonstrado na Figura 2.6.

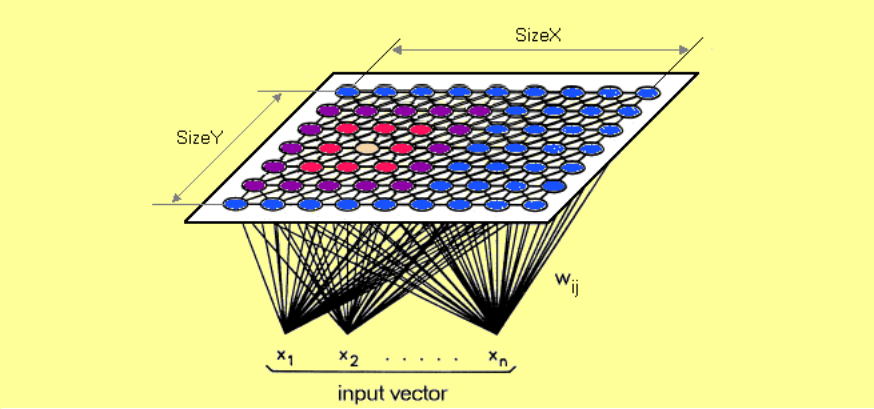

Figura 2.6: Exemplo do funcionamento de um SOM [50], onde $x_{1\ldots n}$ representa as variáveis em estudo e w_{ij} os pesos que as mapeiam para um espaço bidimensional representado por $SizeX \times SizeY$.

2.3.2.2 Avaliação dos Conjuntos Obtidos

Antes de aplicar qualquer técnica de Agrupamento é necessário proceder a alguma análise dos dados em estudo, de modo a determinar se existem de facto conjuntos separáveis nos mesmos, quantos poderemos esperar obter e posteriormente avaliar a qualidade dos mesmo juntamente com o seu significado físico.

Para este efeito, [24] define três tarefas essenciais antes analisarmos os resultados do agrupamento de dados:

- **Avaliar a tendência de agrupamento**: num conjunto de dados deve começar por se avaliar se a estrutura inerente ao mesmo é ou não aleatória. Um método de agrupamento poderá originar alguns conjuntos a partir de um conjunto de observações aleatórias, no entanto, tentar interpretar estes conjuntos não trará nenhum conhecimento adicional acerca dos dados;
- **Determinar o número de conjuntos**: como alguns algoritmos (*e.g. k-means* e *k-medoids*) requerem que o número de conjuntos seja estabelecido *a priori*, é necessário encontrar este valor antes de avaliar os resultados do agrupamento. Este valor, mesmo nos algoritmos que não seja obrigatório como parâmetro de entrada deve ser estimado, de modo validar o conhecimento ganho com a segmentação do conjunto de dados;
- **Validar os conjuntos obtidos:** uma vez aplicada a técnica de agrupamento, é necessário validar os conjuntos obtidos. Para este efeito podem ser utilizado critérios internos e externos [22], [24], [31]. Os critérios internos baseiam-se em métricas obtidas a partir dos dados em si (pode ser útil para proceder a uma avaliação matemática dos conjuntos antes de ser efetuada uma análise à composição dos mesmos; uma

metodologia para tal é definida no Anexo A - Avaliação dos agrupamentos obtidos), enquanto que os critérios externos se baseiam num conhecimento prévio da estrutura dos conjuntos de modo a avaliar os resultados obtidos, nomeadamente, a sua representatividade física.

2.3.3 Variáveis independentes utilizadas na previsão

A revisão de literatura apresentada por [6] identifica quatro tipos de variáveis recorrentes na previsão do consumo de energia: variáveis socioeconómicas (*e.g.* PIB, taxas de crescimento económico, etc.), ambientais (*e.g.* temperatura, humidade, etc.), edifício[3] (*e.g.*ocupação, dimensões, entre outras) e índice temporal (normalmente uma variável associada à data e hora da observação em questão). É também possível efetuar a previsão sem utilizar nenhuma destas variáveis, recorrendo apenas aos valores históricos e assentado a seleção dos pontos escolhidos (atrasos) na correlação existente entre os acontecimentos sucessivos (autocorrelação) [6]. No entanto, como referido em [11], a introdução de variáveis independentes adicionais que mostrem causalidade com a variável dependente podem melhorar os resultados da previsão.

Nas previsões recorrendo as redes neuronais artificiais, a introdução de variáveis meteorológicas (como temperaturas de bolbo seco e ponto de orvalho, humidade relativa, velocidade do vento, entre outras [7], [11], [12], [15], [32]), é prevalente. De acordo com [6], a introdução destas variáveis, depois de avaliada a sua correlação com a carga no sistema, leva a melhorias nos resultados da previsão.

Apesar destes vários estudos referirem a utilização de variáveis meteorológicas na previsão, poucos fazem uma análise quantitativa relativamente às melhorias reais que isto traz. A falta de literatura sobre isto é ainda evidenciada em [33]. Este estudo faz ainda esta mesma comparação, utilizando modelos de Regressão Linear Multivariada onde conclui que a adição da temperatura, humidade, velocidade do vento, precipitação e radiação solar como variáveis independentes leva a uma melhoria global de 0.015% na previsão. O estudo conclui ainda que é necessária a utilização de um método de seleção de atributos relevantes (*Feature Selection*) de forma a melhorar a performance da previsão. Estes métodos, conforme descrito por [34], podem ser classificados em 3 tipos diferentes: filtro (*filter*), embrulho (*wrapper*) e integrado (*embedded*). Os métodos de filtro (*filter*) removem as variáveis menos significativas *a priori*, sendo depois criado um modelo com o resultado daqui obtido. As variáveis são eliminadas com um critério como a correlação, por exemplo. Os métodos embrulho (*wrapper*) envolvem todo o algoritmo de treino no processo de seleção de variáveis, treinando várias iterações (tantas quantas variáveis existirem) do modelo adicionando (ou removendo) variáveis a cada uma delas e avaliando o a *performance* do modelo obtido. As variáveis a originem melhorias no resultado são mantidas e as restantes descartadas para a construção do modelo final. Os métodos integrados (*embedded*) introduzem o processo de seleção das variáveis diretamente no processo de treino, de modo a evitar a procura completa que acontece nos métodos integrados, reduzindo assim a complexidade computacional. Adicionalmente, combinações destes métodos poderão ser criadas, dando origem aos chamados métodos híbridos [35].

Outro estudo onde podemos encontrar uma análise quantitativa da utilização de variáveis meteorológicas na previsão do consumo de energia é em [32], desta vez com recurso a Redes Neuronais Artificiais. Aqui os autores analisam o consumo de um edifício de escritórios e dividem o mesmo em três categorias: aquecimento, ventilação e ar condicionado (AVAC), iluminação e tomadas. Apesar disto, é feita uma previsão para o consumo total onde é obtido um erro médio absoluto percentual

[3] Para previsão do consumo em edifícios.

(MAPE) de 15% sem a utilização de variáveis meteorológicas de 13,6% com a sua utilização, evidenciando assim a melhoria que estas podem trazer à previsão do consumo de energia.

Na literatura atual, é ainda possível verificar a identificação de feriados, vésperas de feriados, dias seguintes a feriados e fins de semana como variáveis relevantes para melhorar a precisão da previsão do consumo de energia [7], [11], [12], [15]. Assim, é possível concluir que o tipo dia tem grande influência semana na carga do sistema.

2.3.4 Métodos de previsão

O desenvolvimento de técnicas de previsão de curto prazo é a segunda mais explorada, logo a seguir à de longo prazo [6]. Isto é explicado pelas necessidades da indústria relativamente a estes tipos de previsão. O planeamento estrutural e de manutenção do sistema electroprodutor requer previsões a longo prazo fiáveis, de modo a permitir que a produção acompanhe o crescimento do consumo. Da mesma forma, e de modo mais crítico, o planeamento operacional do sistema requer também uma previsão de curto prazo fiável, uma vez é necessário gerir a energia que está a ser produzida em cada instante do tempo, de modo a suprir a carga consumida no mesmo.

Uma vez que o objeto de estudo desta dissertação é a previsão de curto prazo do consumo de energia, serão aqui discutidos os métodos mais relevantes para este fim. Segundo [6], os métodos mais utilizados são a regressão linear múltipla e as redes neuronais artificiais (do inglês, *Artificial Neural Networks - ANN*), seguidas pelos métodos de análise autorregressivos de séries temporais (ARMA e ARIMA, com e sem variáveis exógenas e sazonalidade) e pelas máquinas de vetores de suporte (do inglês, *Support Vector Machines – SVM)*. As vantagens e desvantagens destes métodos são apresentados na Tabela 2.2, no entanto, nenhum destes métodos é claramente superior aos outros, pois são empregados em casos de estudo diferentes, com dados dispares na sua representatividade e métricas de erro diferentes, tornando assim difícil uma comparação direta e objetiva. Apesar disto, é de referir que estes métodos apresentam erros de previsão relativamente baixos na generalidade das aplicações. Por exemplo, [36] refere a obtenção de um MAPE de 1,53% e1,97% para os métodos de previsão ARIMA e ANN, com um horizonte de 24h, respetivamente. Os autores de [37] referem erros da mesma ordem para os mesmo métodos, tendo-os testado com dados de vários países obtiveram erros (MAPE) entre 1,82% e 3,67% para o ARIMA e 1,45% a 2,99% para ANN. Quanto às SVM, [38] mostra a obtenção de um MAPE de 7% para SVM e de 13,6% para ANN, no entanto, é de referir que os dados utilizados neste estudo se referem a valores de consumo habitacionais, que têm uma variância maior do que os dados de consumo agregados do sistema como um todo, o que leva a maiores erros na previsão. Este comportamento, é explicado pelo efeito estatístico de alisamento da série temporal associado ao cancelamento natural das flutuações através da agregação de vários consumidores dentro de uma área de controlo minimizando a existência de valores extremos no conjunto de dados.

Uma das desvantagens dos métodos estatísticos (*e.g.*, Regressão Linear Múltipla, Análise Séries Temporais) reside na incapacidade para lidar com a ocorrência de eventos com padrões distintos dentro da própria série temporal, nomeadamente, a distinção entre os ciclos semanais, de fim de semana e de dias especiais (*e.g.*, feriados). Já os métodos baseados em ANN têm capacidade de aceitar estas características como variável independente e de modelar relações não lineares implícitas entre o consumo e as variáveis que a afetam [7].

Tabela 2.2: Características dos métodos de previsão mais comuns [6]

Método de previsão	**Vantagens**	**Desvantagens**
Regressão Linear	• Simplicidade; • Pode utilizar apenas uma (Simples) ou várias (Múltipla) variáveis independentes;	• Apenas captura relações entre variáveis linearmente correlacionadas; • Sensível a valores extremos (*outliers*)*;* • Variáveis usadas na previsão têm de ser linearmente independentes;
Análise de Séries Temporais (Box-Jenkins)	• Adaptável, existem muitas versões do método e já foi amplamente estudado; • Capaz de lidar com sazonalidade e não-estacionariedade; • Requer apenas os dados históricos da série;	• Improvável que tenha bom desempenho na previsão a longo prazo; • É computacionalmente exigente estimar os parâmetros do modelo; • Requer um sólido conhecimento da estatística inerente à série temporal;
Redes Neuronais Artificias	• Não é necessário conhecer a relação entre as variáveis dependentes e independentes; • Capaz de lidar eficazmente com relações não lineares; • Capaz de lidar com a presença de ruído no conjunto de dados sem que isso afete significativamente o resultado da previsão.	• Não resultam num modelo matemático com significado físico; • É computacionalmente exigente treinar a rede neuronal; • Necessita de grande quantidade de dados históricos de variáveis independentes;
Máquina de Vector de Suporte	• Ajustamento do parâmetro de regularização da função objectivo ajuda a evitar o sobreajustamento aos dados de treino (*over-fitting*)*;* • Problema de otimização convexo (não há mínimos locais); • Utilização do *"kernel trick",* que mapeia o espaço das variáveis para um espaço vetorial não linear, permitindo capturar relações não lineares de forma mais eficiente;	• É difícil escolher uma "boa" função *kernel*; • Computacionalmente exigente para conjunto de dados grandes;

2.3.5 Métricas de erro utilizadas

O ponto mais importante do estudo das técnicas de previsão é a avaliação dos resultados. Na literatura, esta avaliação é feita através da determinação do erro entre os valores previstos pelos modelos e os valores reais. No entanto, este erro pode ser apresentado sob várias formas, sendo este um dos fatores que torna a comparação dos diferentes métodos mais difícil [6].

Os tipos mais comuns de avaliação dos métodos de previsão são: o viés, o erro médio absoluto percentual (MAPE), erro quadrático médio (RMSE), correlação de Pearson [6], [12].

$$viés = \left(\frac{1}{N}\sum_{t}\frac{y_{prev,t} - y_{obs,t}}{y_{obs,t}}\right) \times 100 \tag{2.1}$$

$$MAPE = \left(\frac{1}{N}\sum_{t}\left|\frac{\left(y_{prev,t} - y_{obs,t}\right)}{y_{obs,t}}\right|\right) \times 100 \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t}(y_{prev,t} - y_{obs,t})^2} \tag{2.3}$$

$$r = \frac{\sum\ (y_{prev} - \mu_{y_{prev}})(y_{obs} - \mu_{y_{obs}})}{\sqrt{\sum\ (y_{prev} - \mu_{y_{prev}})^2}\sqrt{\sum\ (y_{obs} - \mu_{y_{obs}})^2}} \tag{2.4}$$

$y_{prev,t}$ e $y_{obs,t}$ representam o valor do consumo de energia elétrica no instante *t* previsto e observado, respetivamente. $\mu_{y_{prev}}$ e $\mu_{y_{obs}}$ representam o valor médio do consumo de energia elétrica previsto e observado, respetivamente. *N* representa o número de registos da série.

3 Dados e metodologia

No presente capítulo será feita uma descrição i) dos dados utilizados ii) das metodologias desenvolvidas e aplicadas nesta dissertação. O capítulo divide-se da seguinte forma:

- Descrição dos dados utilizados;
- Descrição dos pré-tratamentos efetuados;
- Descrição da metodologia de agrupamento;
- Descrição das metodologias de previsão;
- Apresentação do *pipeline* criado para ingestão e processamento dos dados e ajuste e validação dos modelos criados.

3.1 Tipologia de dados

De modo a efetuar a previsão da variável em estudo (consumo de energia elétrica) é necessário, em primeiro lugar, obter dados históricos relativamente à mesma e às variáveis meteorológicas relevantes para a estimar.

Os dados de consumo utilizados são disponibilizados publicamente pela REN [8]. Este conjunto de dados encontra-se divido entre os diferentes perfis de consumo tipo identificados pela REN, de acordo com os termos estabelecidos no n.º 4 do artigo 272.º do Regulamento de Relações Comerciais do setor elétrico (ERSE) [39]. Os perfis de consumo tipo[4] disponibilizados são: clientes finais de média tensão (MT)[5], clientes finais de baixa tensão especial[6] (BTE), clientes finais de baixa tensão normal (BTN) desagregados nas classes BTN A, BTN B e BTN C[7] e, o consumo para iluminação pública (IP) [39]. Estes perfis de consumo tipo, ou iniciais, têm como objetivo a aplicação a todos os clientes finais que não tenham equipamentos de medição com registos em períodos de 15 minutos. Na presente dissertação será apenas estudada a carga[8] referente a um destes perfis, BTN C, que se refere, maioritariamente, aos consumidores domésticos [40]. Esta escolha deve-se ao potencial que este tipo de consumidores oferece para os novos paradigmas de gestão do consumo. Foram considerados 5 anos de dados históricos de consumo (01/01/2014 - 31/12/2018), com resolução horária.

Na Tabela 3.1 encontram-se representados os valores de consumo para o ano de 2014, por classe de consumidor. Uma vez que não foi possível obter os valores anuais de consumo para todos os anos em estudo, assumiu-se o valor de 2014 para os restantes anos em análise.

4 Os perfis encontram-se normalizados pelo valor total anual do respetivo consumo tipo e são inferidos através da aquisição de dados de sistemas de telecontagem sendo posteriormente submetidos para aprovação pelos operadores (transmissão e distribuição) de rede.

5 Destina-se à indústria de componentes automóveis, metalúrgica, moldes, vitrificação, grande hotelaria, etc. Tensão entre fases cujo valor eficaz é superior a 1kV e igual ou inferior a 45kV.

6 Destina-se a clientes residenciais, lojas, escritórios e pequenas empresas. Potências contratadas iguais ou inferiores a 41,4kVA e uma potência mínima contratada de 1,15kVA.

7 Características dos clientes do perfil: BTN A - maioritariamente empresas com potência contratada elevada, BTN B - maioritariamente hotelaria e alguns domésticos com elevado consumo anual e BTN C - maioritariamente domésticos

8 8 Por simplicidade, na presente dissertação optou-se por utilizar, maioritariamente, a designação "carga" na apresentação dos resultados em detrimento de "consumo de energia elétrica".

Tabela 3.1: Características e consumo anual de energia em 2014 dos diferentes perfis de consumo tipo analisados

Tipo de perfil	Características dos clientes	Consumo anual em 2014 (GWh)	Peso no consumo total em 2014 (%)
BTE	Consumo comercial/industrial	335.4	0.7
BTN A	Maioritariamente empresas com potência contratada elevada	4982.8	10.8
BTN B	Maioritariamente hotelaria e alguns domésticos com elevado consumo anual	610.8	1.3
BTN C	Maioritariamente domésticos	12530.5	27.2
IP	Iluminação pública	1477.9	3.2
MT	Indústria	13935.0	30.2

As variáveis meteorológicas consideradas nesta dissertação (temperatura de bolbo seco, temperatura de ponto de orvalho, pressão atmosférica, nebulosidade e velocidade do vento) foram obtidas do *National Center for Atmospheric Research*, provenientes do conjunto de reanálises ERA-5 (ds630.1), no mesmo período mencionado para os dados de consumo [41]. Estes dados apresentam uma resolução temporal de uma hora e espacial de 0.25º (aproximadamente 25 km). Nesse sentido, foram obtidos os dados para todos os pontos espaciais sobre Portugal.

3.1.1 Pré-Tratamento

O primeiro passo do pré-processamento dos dados em análise é a avaliação da integridade do conjunto de dados. Este passo consiste na procura de valores em falta e de *outliers* e na sua consequente remoção.

Relativamente ao processo de normalização, optou-se por a utilizar a normalização por *min-max*, à qual será atribuída a escala [-1, 1] e é feita da seguinte forma:

$$x_t{}' = 2 * \frac{(x_t - \min(x))}{(\max(x) - \min(x))} - 1 \tag{3.1}$$

Onde:

- $x_t{}' - valor\ normalizado\ para\ o\ instante\ t$
- $x_t - valor\ não\ normalizado\ para\ o\ instante\ t$
- $\min(x) - valor\ mínimo\ da\ amostra$
- $\max(x) - valor\ máximo\ da\ amostra$

A normalização, feita desta forma, ajuda a comparar valores com unidades diferentes, permitindo *i)* manter a mesma distribuição dos dados e *ii)* também, uma convergência mais rápida de durante a fase de treino de alguns métodos de previsão (*e.g.*, descida gradiente) que serão aplicados nesta dissertação. Este passo é crucial em técnicas como as redes neuronais que recorrem a pesos para estabelecer a relação entre as variáveis dependentes/independentes [42].

3.2 Determinação dos perfis diários de consumo

De modo a caracterizar os perfis de consumo presentes nos dados em análise será utilizada uma técnica de agrupamento, de modo a identificar de uma forma independente, a sazonalidade dos dados, fins de semana, dias especiais, etc. Os diferentes perfis obtidos serão analisados detalhadamente procedendo-se a identificação das variáveis endógenas que permitam compreender cada perfil. Essas variáveis serão usadas na fase de previsão do consumo.

De modo a agrupar os perfis diários de consumo semelhantes, transformam-se os registos de consumo de energia numa matriz $X_{t,h}$, tal que:

$$X_{t,h} = \begin{bmatrix} Z_{1,1} & Z_{1,2} & Z_{1,h} & \dots & Z_{1,24} \\ Z_{2,1} & Z_{2,2} & Z_{2,h} & \dots & Z_{2,24} \\ Z_{3,1} & Z_{3,2} & Z_{3,h} & \cdots & Z_{3,24} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{t,1} & Z_{t,2} & Z_{t,h} & \cdots & Z_{t,24} \end{bmatrix} \quad (3.2)$$

Onde $Z_{t,h}$ representa o consumo Z, registada à *h* do dia *t*. Os dados são então agrupados num número *k* de clusters (C_i), onde $C_i \subseteq X$ e $i \in \{1,2,\dots,k\}$. Como descrito na secção 2.3.2.1, existem vários algoritmos de segmentação. Nesta dissertação optou-se pela aplicação do algoritmo *k-medoids* uma vez que este método será menos sensível a valores extremos, uma vez que utiliza quantis em vez de valores médios, como no caso do algoritmo *k-means*. Este algoritmo reparte as diferentes observações num número *k* de *clusters* (C_i) definido pelo utilizador [22], [43]. Neste caso em concreto, este agrupamento resulta em vários subconjuntos de consumos diários que servirão para identificar os perfis. A seleção do *k* será feita pelo método do Cotovelo, descrito em 2.3.2.1. No Anexo A é apresenta-se uma validação do número ótimo de agrupamentos identificado.

3.3 Métodos de Previsão

Na presente dissertação pretende-se avaliar o desempenho de diferentes métodos estatísticos na previsão de energia. Desta forma, após a revisão da literatura relevante ao tema, foram selecionados 3 métodos a compara: 1) Regressão Linear Multivariada, MLR; 2) *k*-vizinhos mais próximos, *KNN*; e 3) Rede Neuronal Artificial, ANN. Estes foram com um método de referência (*baseline)*, de modo a obtermos termos de comparação. Não sendo o objetivo de estudo da presente dissertação, a configuração dos parâmetros de cada um dos métodos estudada de forma aprofundada tendo sido adotadas as práticas mais comuns – os *rules of thumb* – presentes na literatura. Esta *baseline* consistirá numa regressão linear simples, com o consumo verificado nas 24h anteriores à hora a prever. De seguida apresenta-se uma breve descrição das características destes métodos.

3.3.1 Regressão Linear Multivariada

O objetivo da regressão linear multivariada[9][44] é estabelecer uma relação entre um conjunto de variáveis independentes, $x_1, x_2, \dots, x_p \in \boldsymbol{X}$, e uma variável dependente, $\boldsymbol{Y_x}$. Esta relação será então aproximada por um modelo com a forma:

$$Y_x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon_x \quad (3.3)$$

Esta equação pode ser definida matricialmente da seguinte como:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (3.4)$$

onde $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$, $\boldsymbol{\beta} = [\beta_0, \dots, \beta_p]$ e $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

[9] A regressão linear simples acontece no caso em que apenas temos uma variável independente, $\boldsymbol{X} = x_1$

De forma a estimar os coeficientes $\boldsymbol{\beta}$ da regressão, são assumidos alguns pressupostos [44]:

- $\boldsymbol{Y_x}$ é uma variável aleatória, dependente linearmente de um vetor não aleatório $\boldsymbol{x} \in \boldsymbol{X}$;
- $\bar{y}_x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ é um valor médio de $\boldsymbol{Y_x}$;
- β_0 representa o valor médio de $\boldsymbol{Y_x}$ quando $\boldsymbol{x} = 0$;
- $\beta_1, \beta_2, \ldots, \beta_p$ representam a variação de $\boldsymbol{Y_x}$ por variação unitária de $\boldsymbol{x}$;
- ε_x é o erro aleatório do modelo, causado por efeitos desconhecidos além dos descritos por $\boldsymbol{x}$, com uma distribuição de probabilidade Gaussiana (*i.e.*, independentes e identicamente distribuídas);

Estes coeficientes $\boldsymbol{\beta}$ podem ser estimados através dos Mínimos Quadrados, onde se minimiza a equação:

$$\begin{aligned} SQ(\beta_0, \beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} (\varepsilon_i^2) \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \end{aligned} \tag{3.5}$$

Ou, matricialmente:

$$SQ(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\boldsymbol{Y} - \boldsymbol{X\beta}) \tag{3.6}$$

Os valores de $\boldsymbol{\beta}$ que minimizam esta equação resultam das derivadas parciais de SQ em ordem a cada um destes coeficientes de forma a obter o estimador $\widehat{\boldsymbol{\beta}}$:

$$\frac{\partial SQ(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \Leftrightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{3.7}$$

Uma vez encontrados os valores dos estimadores $\widehat{\boldsymbol{\beta}}$ é possível calcular os valores $\widehat{\boldsymbol{Y}}$ como:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{3.8}$$

de onde se obtém os resíduos do ajustamento $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. Desta expressão surge que:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} = \boldsymbol{HY} \tag{3.9}$$

Onde $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$representa a matriz que projeta $\boldsymbol{Y}$ para o espaço das colunas de $\boldsymbol{X}$. Assim temos também que:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{MY} \tag{3.10}$$

Com estas expressões podemos então provar que a soma dos quadrados total é igual à soma dos quadrados devida à regressão mais a soma dos quadrados devida ao erro:

$$\begin{gathered} \boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H} + \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^T \big((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} + \boldsymbol{HY}\big) \\ = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} + \boldsymbol{Y}^T \boldsymbol{HY} = \boldsymbol{Y}^T \boldsymbol{MY} + \boldsymbol{Y}^T \boldsymbol{HY} \\ = \boldsymbol{Y}^T \boldsymbol{M}^T \boldsymbol{MY} + \boldsymbol{Y}^T \boldsymbol{H}^T \boldsymbol{HY} = (\boldsymbol{YM})^T \boldsymbol{MY} + (\boldsymbol{YM})^T \boldsymbol{HY} \\ = \boldsymbol{e}^T \boldsymbol{e} + \widehat{\boldsymbol{Y}}^T \widehat{\boldsymbol{Y}} \end{gathered} \tag{3.11}$$

Uma forma de normalizar as somas de quadrados é centrar os valores em 0:

$$\begin{gathered} \boldsymbol{Y}^T\boldsymbol{Y} - n\bar{Y}^2 = \boldsymbol{e}^T\boldsymbol{e} + \hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}} - n\bar{Y}^2 \Longrightarrow \\ \Longrightarrow \sum_{i=1}^{n} {Y_i}^2 - n\bar{Y}^2 = \sum_{i=1}^{n} {e_i}^2 + \sum_{i=1}^{n} {\hat{Y}_i}^2 - n\bar{Y}^2 \Longrightarrow \\ \Longrightarrow \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} {e_i}^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \Longrightarrow \\ \Longrightarrow SS_{Tot} = SS_E + SS_{Reg} \end{gathered} \tag{3.12}$$

Nesta equação, SS_{Reg} representa a variabilidade explicada pelos estimadores obtidos com base nas variáveis independentes escolhidas, enquanto que SS_E se refere à variabilidade residual presente nos dados, que não pode ser explicada pelo modelo.

Para avaliar o modelo de regressão obtido pode recorrer-se ao coeficiente de determinação, que é dado pela seguinte expressão:

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_E}{SS_{Tot}} \tag{3.13}$$

que varia entre 0 e 1 e indica a variabilidade dos dados explicada pelo modelo criado.

Um método mais robusto de avaliação do modelo pode ser feito através da análise de variância (ANOVA) aos coeficientes encontrados [44]. Com o que foi exposto até aqui, é possível construir a seguinte tabela para esta análise:

Tabela 3.2: Tabela de ANOVA para a Regressão Linear Múltipla

Fonte da Variação	Graus de Liberdade	Soma de Quadrados	Média dos Quadrados	F
Modelo	p	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$MS_{Reg} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{p}$	$\frac{MS_{Reg}}{MS_E}$
Erro	n-p-1	$\sum_{i=1}^{n} {e_i}^2$	$MS_E = \frac{\sum_{i=1}^{n} {e_i}^2}{n-p-1}$	-
Total	n-1	$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$MS_{Tot} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$	-

A hipótese nula a testar sob estas condições é

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1: \exists i,\ \beta_i \neq 0, i = 1, \ldots, p$$

A hipótese alternativa formulada é bilateral [44], que pode ser rejeitada se:

- O *p-value*, dado por $P\left(F > F_{p,n-p-1}\right)$, seja superior ao um nível de significância pré-estabelecido;
- O valor de $F = \frac{MS_{Reg}}{MS_E}$ seja inferior ao valor de $F_{crítico}$ determinado por uma distribuição Fischer de Snedecor [44], ao nível significância escolhido, com os graus de liberdade do modelo e do erro.

3.3.2 *K* Vizinhos Mais Próximos

O algoritmo *k vizinhos mais próximos* (do inglês *K Nearest Neighbours*, *KNN*) pode ser usado tanto em problemas de classificação como de regressão, aqui, por vezes designado como previsão por analogia. Das principais vantagens deste algoritmo face aos restantes algoritmos em análise, destacam-se: *i*) é relativamente simples de compreender e implementar, sem deixar conseguir resultados robusto; *ii)* não necessita de fase de treino realizando a sua previsão com base em valores históricos observados

e *iii*) é uma abordagem não paramétrica, que não implica quaisquer pressupostos acerca da distribuição das variáveis a prever [45], [46]. A intuição por detrás do KNN é a de que se um evento $\boldsymbol{Y_t}$ aconteceu no passado devido a $x_{1t}, x_{2t}, \dots, x_{n^{10}t} \in \boldsymbol{X_t}$, então, se $\boldsymbol{X_{t+h}}$ tomar valores iguais a um qualquer $\boldsymbol{X_i} \in \boldsymbol{X_t}$ já registado, podemos atribuir $\boldsymbol{Y_{t+h}}$ a um valor igual ao $\boldsymbol{Y_i}$ correspondente.

Na realidade é muito pouco provável que ocorram eventos exatamente iguais, por isso é necessário definir um método pelo qual seja possível avaliar a semelhança entre dois eventos $\boldsymbol{p}$ e $\boldsymbol{q}$. Para este fim é comum utilizar a distância euclidiana entre as duas observações, sendo esta distância dada pela seguinte fórmula:

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{p_i} - \boldsymbol{q_i})^2} \tag{3.14}$$

O algoritmo, aplicado desta forma, pode ser vulnerável à existência de valores extremos (*outliers)*, uma vez que só o ponto mais próximo de $\boldsymbol{X_{t+h}}$ é considerado. De forma a resolver isto, pode diluir-se esta dependência numa só observação ao considerar-se não apenas uma, mas uma janela das $\boldsymbol{k}$ observações mais próximas do $\boldsymbol{X_{t+h}}$. Daqui aparece a nomenclatura do algoritmo $\boldsymbol{k}$ *vizinhos mais próximos*.

Este algoritmo é mais comummente utilizado para classificação, onde se escolhe a classe mais frequente encontrada na janela dos *k vizinhos mais próximos*. Como o objeto de estudo da presente dissertação é a previsão, será aqui descrita a aplicação deste algoritmo à tarefa de regressão. Sob estas condições, a seguinte toma a seguinte forma:

Tabela 3.3: Pseudocódigo do algoritmo do k vizinhos mais próximos[11]

Início
Variáveis: Conjunto de dados históricos ($\boldsymbol{X}, \boldsymbol{Y}$), *k*, exemplo em estudo ($\boldsymbol{x_{t+h}}$) **Para** i **em** Conjunto de Treino: $D_i \leftarrow$ Calcular distância entre $\boldsymbol{X_i}$ e $\boldsymbol{x_{t+h}}$ **Fim Para** $\boldsymbol{Y^{NN}} \leftarrow k$ observações $\boldsymbol{Y}$ correspondentes a $\boldsymbol{X}$ para a qual D_i é mínimo $\mathbb{M}_k(\boldsymbol{x_{t+h}}) = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{Y_j^{NN}}$
Fim

O valor de *k* influencia o resultado da previsão $\mathbb{M}_k(\boldsymbol{x_{t+h}})$, por isso é necessário existir um critério que auxilie objetivamente esta escolha. Para isto podemos recorrer a uma abordagem semelhante ao método do Cotovelo, descrito para o algoritmo de agrupamento *k-medoids*.

A análise do número ótimos de vizinhos é apresentada no Anexo B. Na presente dissertação foi utilizado um valor de k=10.

10 *n* representa o número total de variáveis independentes.
11 A implementação deste algoritmo em MATLAB pode ser encontrada no Anexo B.

3.3.3 Rede Neuronal Artificial

As redes neuronais têm vindo a ser amplamente utilizadas devido à sua grande capacidade generalização. Esta capacidade é bastante útil na resolução de padrões complexos como os que existem nas séries temporais (como o consumo de energia), sem que seja necessário decompor os processos estocásticos e modelar todas as suas componentes. A relação complexa entre as variáveis independentes e a variável dependente é obtida pelo processo de treino [11]. O algoritmo de aprendizagem mais utilizado no treino das redes neuronais é denominado de *retro propagação* (do inglês, *Backpropagation*). Este processo é uma generalização da descida do gradiente do erro para uma rede com vários níveis de parâmetros a otimizar.

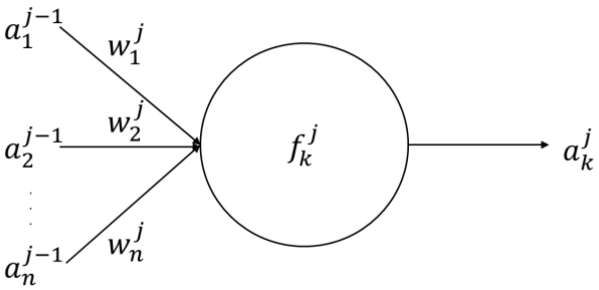

*Figura 3.1: Nodo de uma rede neuronal artificial onde $\boldsymbol{a^{j-1}}$ representa o resultado da camada anterior; $\boldsymbol{w^j}$ representa os pesos da rede entre a camada **j-1** e **j**; $\boldsymbol{f_k^j}$ representa a função de ativação no nodo **k** da camada **j** e $\boldsymbol{a_k^j}$ representa o resultado desse mesmo nodo*

Uma rede neuronal artificial é uma estrutura de unidades computacionais simples (Figura 3.1), interconectadas de forma complexa entre si (Figura 3.2). Estas unidades, também designadas de nodos ou neurónios, são as unidades fundamentais da ANN e são estas que lhe conferem a capacidade aprendizagem [47].

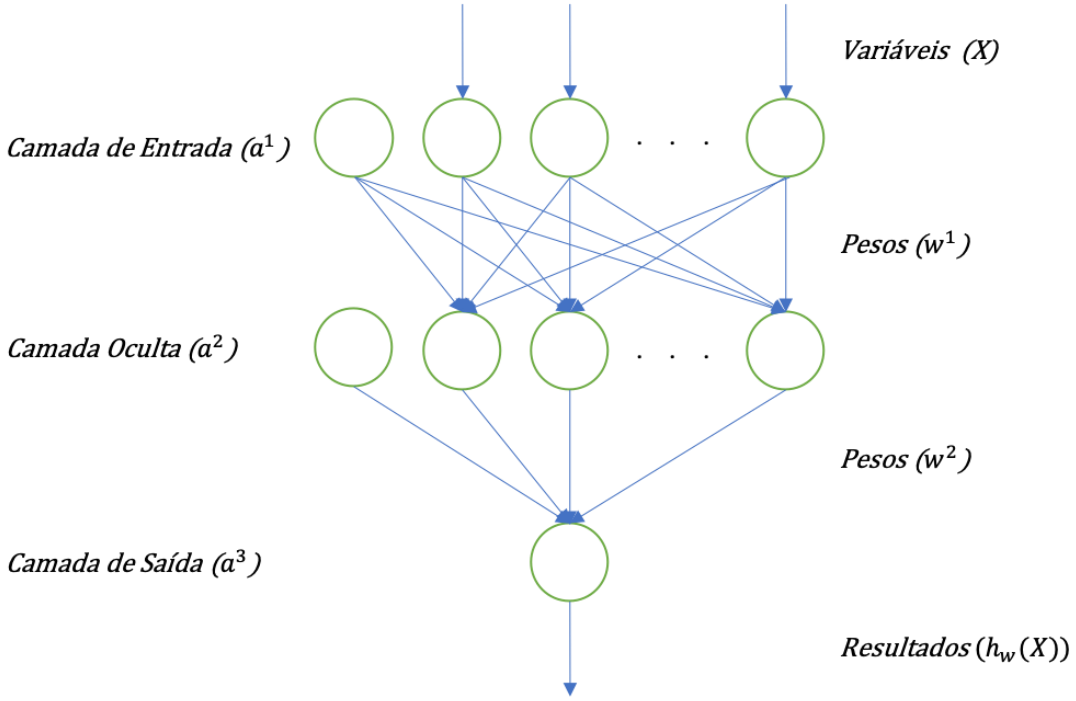

Figura 3.2: Estrutura de uma Rede Neuronal Artificial com uma camada oculta

Estes nodos encontram-se interligados de modo a formar uma rede. A rede pode ter uma forma arbitrária, definida pelo utilizador, em termos de profundidade (nº de camadas ocultas – *hidden layers*) e de dimensão das camadas ocultas (nº de nodos em cada uma). A dimensão das camadas de entrada (*input*) e de saída (*output*) são definidas pelo contexto do problema em análise.

Cada um destes nodos apresentados na Figura 3.1 pode ser descrito matematicamente da seguinte forma:

$$a_k^j = f_k^j\left(\sum_{k=0}^{n}\left(w_k^j a_k^{j-1}\right)\right) \tag{3.15}$$

Em cada um destes nodos existe uma função de ativação que é alimentada pela soma ponderada (pelos pesos *w* da rede) dos resultados nodos da camada precedente. Assim, o resultado de cada um dos nodos é transmitido para a camada seguinte.

O objetivo da rede neuronal é conseguir atingir os registos observados da variável dependente através de uma combinação das variáveis independentes utilizadas como dados de entrada. Para isto a rede tem de ser treinada, *i.e.*, os pesos $\boldsymbol{W}$ têm de ser calculados para atingir o objetivo desejado [48]. A rede neuronal pode ser descrita como uma função $h_{\boldsymbol{w}}(\boldsymbol{X})$ parametrizada pelos pesos $\boldsymbol{W}$ [47]. Desta forma, é possível definir uma função custo para o ajuste desta função, que representa uma medida do erro total entre o resultado obtido pela rede e os valores reais de $\boldsymbol{y}$, tal que

$$J(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y_i} - h_{\boldsymbol{w}}(\boldsymbol{X_i}))^2 \tag{3.16}$$

Ou, na forma matricial:

$$J(\boldsymbol{W}) = \frac{1}{2}\left(\boldsymbol{y} - h_{\boldsymbol{w}}(\boldsymbol{X})\right)^T(\boldsymbol{y} - h_{\boldsymbol{w}}(\boldsymbol{X})) \tag{3.17}$$

Os parâmetros $\boldsymbol{W}$ da rede são os únicos valores possíveis de ajustar de forma a reduzir o valor desta função custo. Como a função $J(\boldsymbol{W})$ é composta pelas funções contínuas e diferenciáveis em cada um dos nodos, é ela mesmo contínua e diferenciável relativamente a todos os pesos $\boldsymbol{W}$ que compõem a rede. Podemos assim minimizar a função, utilizando o método da descida do gradiente, para o qual é necessário calcular o gradiente da função em ordem aos pesos da rede:

$$\nabla J = \left(\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \dots, \frac{\partial J}{\partial w_l}\right) \tag{3.18}$$

onde *l* refere-se ao número total de pesos existentes em toda a rede.

Uma vez calculado o gradiente, é possível minimizar a função atualizando os pesos da rede de acordo com a seguinte expressão:

$$\Delta w_i = -\alpha\frac{\partial J}{\partial w_i}, com\ i = 1,2,\dots,l \tag{3.19}$$

Onde α é a taxa de aprendizagem, que define o tamanho do "passo" dado no sentido contrário ao gradiente. A escolha de um valor reduzido da taxa de aprendizagem leva a muitas iterações até a convergência. Em alguns casos, esta escolha permite identificar apenas um mínimo local da função objetivo. Por outro lado, a escolha de uma taxa de aprendizagem alta requer menos tempo computacional, mas pode resultar numa superação do valor ideal. Assim, este parâmetro visa atingir um compromisso entre o "passo" dado e o tempo computacional de modo a evitar a divergência do método.

Algoritmo – Treino da rede neuronal[12]:

- Inicializar os pesos da rede, w_i, aleatoriamente
- Calcular $h_{\boldsymbol{w}}(\boldsymbol{X})$ com os pesos gerados (*Eq. 3.16*)
- Calcular o erro $J(\boldsymbol{W})$
- Calcular o gradiente ∇J
- Atualizar os pesos: $w_i^{novo} = w_i^{antigo} + \Delta w_i$
- Calcular $h_{\boldsymbol{w}}(\boldsymbol{X})$ com os pesos atualizados
- Parar quando o algoritmo atingir a condição de paragem (*e.g.:* $\Delta J(\boldsymbol{W}) \sim 0$)

[12] As configurações da rede utilizadas, implementadas em MATLAB, são apresentadas no Anexo C.

3.4 Síntese da metodologia de processamento para previsão do consumo de energia

Uma representação esquemática dos principais passos da metodologia aplicada nesta dissertação é apresentada na Figura 3.3.

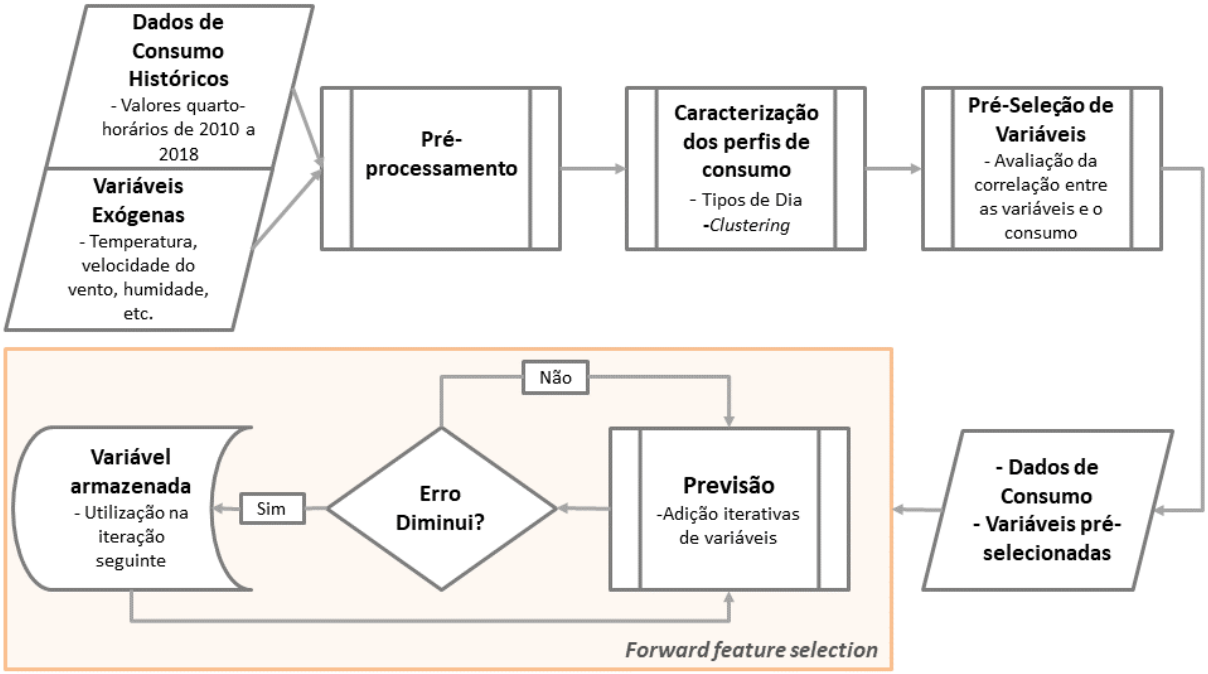

Figura 3.3: Pipeline de processamento dos dados de consumo e variáveis independentes

A metodologia começa com o pré-processamento dos dados procedendo-se à remoção de valores extremos/anómalos, de valores em falta[13] e normalização. De seguida são identificados estatisticamente os perfis diários de consumo por meio da técnica de agrupamento não supervisionada *k-medoids*. Uma vez obtidos e caracterizados estes perfis, é feita a avaliação das variáveis independentes, relativamente à sua correlação com o consumo de energia, visando a identificação das variáveis potencialmente mais uteis na tarefa de previsão. Estes dados, serão introduzidos num algoritmo iterativo de seleção aditiva de variáveis (*forward feature selection*), e avaliado o erro quadrático médio (MSE[14]). Este procedimento foi aplicado *i)* de forma individual aos três métodos descritos anteriormente e *ii)* ao período entre 01/01/2014 e 31/12/2017 (período de calibração das metodologias), sendo o ano de 2018 usado para validação das metodologias.

[13] No caso concreto em estudo não foram encontrados valores em falta nem valores anómalos.
[14] O MSE é definido como o quadrado do RMSE apresentado na secção 2.3.5

4 Apresentação e Discussão dos Resultados

No presente capítulo apresenta-se uma análise dos resultados obtidos com base na metodologia exposta na seção anterior. Primeiramente serão analisados e discutidos os perfis diários do consumo de energia elétrica. Em seguida serão analisados e comparados os resultados da previsão após aplicação dos três métodos selecionados nesta dissertação.

4.1 Identificação e caracterização dos perfis diários de consumo de energia elétrica

4.1.1 Identificação do número ótimo de agrupamentos

Apesar da divisão em diferentes perfis de consumo ser feita pela REN, esta divisão tem apenas por base o nível de potência contratada. No entanto, através da análise da série temporal associada a cada um dos perfis tipo, é possível identificar diferentes padrões de consumo, a diferentes escalas (*e.g.* diários, semanais, sazonais, etc.). A identificação destes padrões pode ser importante na identificação de variáveis relevantes para o ajuste de modelos de previsão bem como permitir identificar *a priori* os comportamentos típicos de consumo. Com a identificação destes perfis pretende agregar-se dias com perfil de consumo que sejam semelhantes entre si, de modo a obter agrupamentos com baixa variabilidade relativamente ao conjunto original.

Uma inspeção visual da série em estudo sugere a existência de padrões distintos no diagrama de consumo do perfil BTN C. Na Figura 4.1(a) verifica-se que estes padrões coincidem com as estações do ano, com um consumo mais elevado nas estações frias e mais reduzido nas estações quentes. Em termos semanais é possível observar um padrão bem definido que se repete ao longo do ano. A Figura 4.1(b) demonstra a o padrão semanal existente ao longo do mês de Janeiro de 2014. Nos dias correspondentes ao fim de semana (Sábado e Domingo) existe menos procura de energia, e os restantes dias apresentam um perfil diário bastante semelhantes entre si.

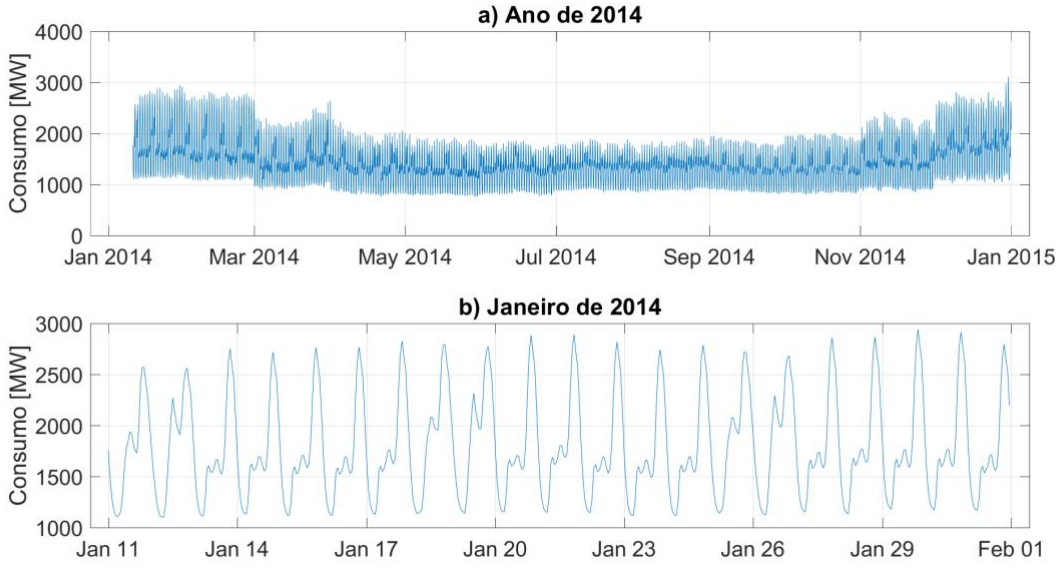

Figura 4.1: Diagrama de carga referente ao ano de 2014. (a) representa a sazonalidade do consumo ao longo do ano. (b) mostra os padrões diários e semanais no mês de Janeiro de 2014.

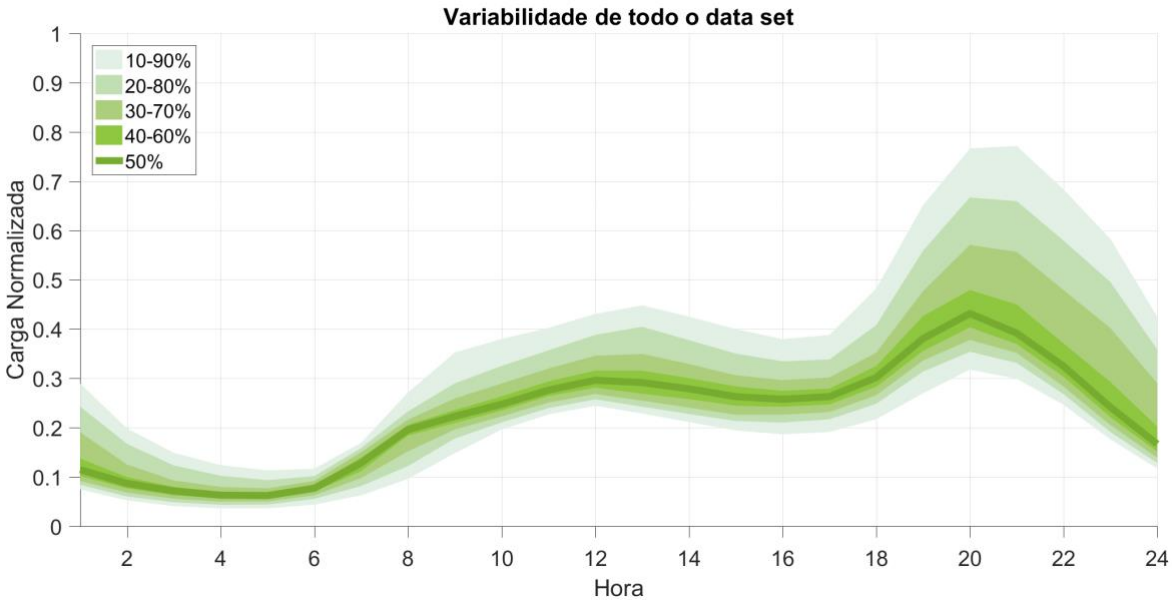

Figura 4.2: Percentis da carga BTN C relativos aos 4 anos de observações

A variabilidade presente na série é reforçada pelo gráfico apresentado na Figura 4.2, onde estão representados os percentis diários do perfil de consumo BTN C durante o período em análise, onde é possível observar uma elevada variabilidade dos perfis, principalmente nas horas de ponta da noite (entre as 18h e as 22h).

Na presente análise, recorrendo a série temporal do perfil tipo BTN organizada de acordo com a matriz apresentada na equação 3.1, utilizou-se a técnica de agrupamento *k-medoids* para identificar estatisticamente os padrões diários típicos de consumo de energia elétrica para este tipo de perfil. O número ótimo de agrupamentos, i.e., *k* foi identificado pelo método do cotovelo, representado na Figura 4.3.

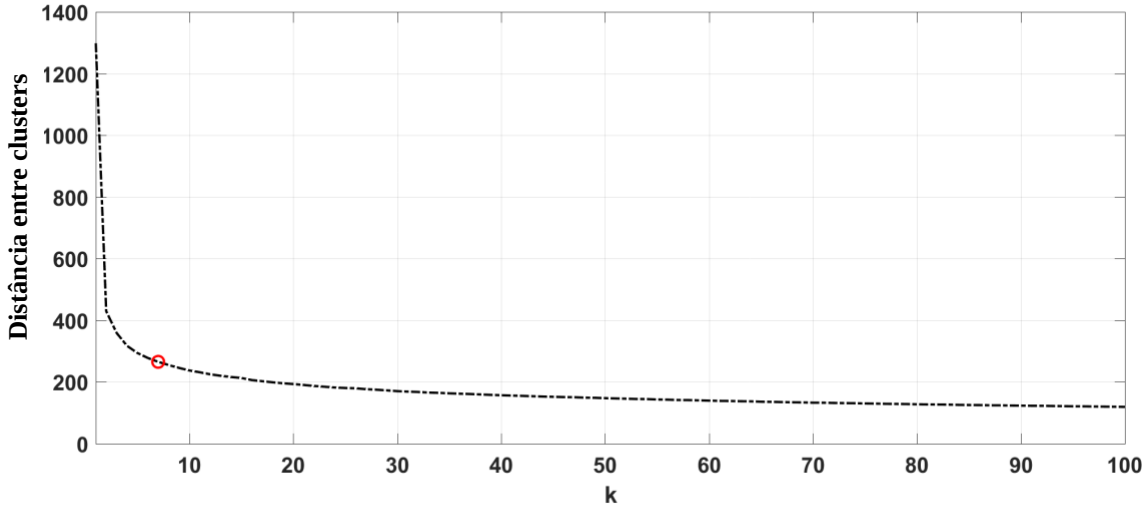

Figura 4.3: Método do cotovelo para seleção do k

De acordo com a Figura 4.3 o número ótimo de agrupamentos (*k*) é sete. A adequabilidade deste número de agrupamentos é reforçada no Anexo A. Este valor aparenta ser razoável, tendo em conta o contexto do problema. De modo a compreender melhor os fenómenos capturados, de seguida será feita uma análise detalhada à composição dos agrupamentos criados.

4.1.2 Caraterização dos perfis diários de consumo de energia elétrica

Os agrupamentos encontrados por estas técnicas podem não refletir os resultados um dado processo físico específico. No entanto, comportamentos predominantes em determinados grupos de observações serão capturados. Neste sentido, será aqui feita a análise aos agrupamentos obtidos, de modo a validar a sua qualidade e significado físico.

Na Figura 4.4 estão representados os centróides de cada agrupamento, *i.e.*, o perfil diário de consumo de energia elétrica que representa a mediana de cada agrupamento obtido. Aqui é possível verificar as distinções encontradas, quer em termos de magnitude, quer em termos de distribuição do consumo de energia ao longo do dia.

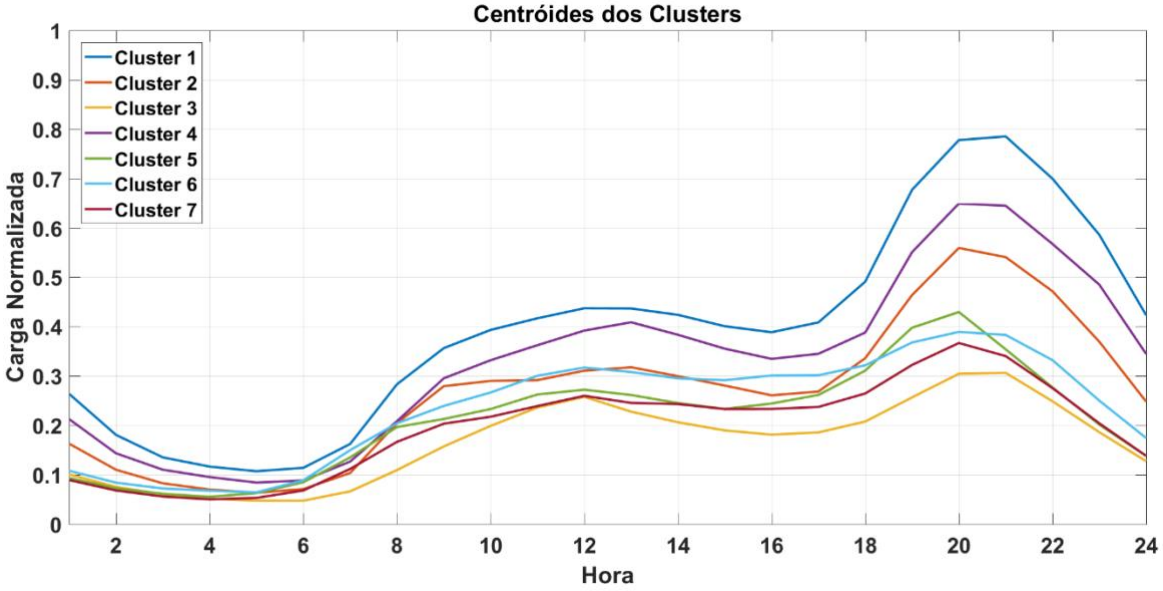

Figura 4.4: Perfis representativos dos clusters criados

As Figura 4.5 - Figura 4.8 permitem analisar a constituição destes conjuntos em detalhe nomeadamente, a sua variabilidade e as observações atribuídas a cada um deles. Importa referir que nas figuras que se seguem optou-se por apresentar o agrupamento por similaridade entre si e não a ordenação obtida através da aplicação do algoritmo.

Como se verifica na Figura 4.2, o conjunto apresenta bastante variabilidade. A redução desta variabilidade é bastante acentuada nos perfis encontrados. Os *clusters* 1 e 4 (Figura 4.5) encontram-se associados a um elevado consumo em comparação com os restantes agrupamentos e à amplitude diária mais elevada encontrada dentro dos conjuntos obtidos. Esta característica é típica dos dias de semana (*cluster* 1) e de fim de semana (*cluster* 4) de inverno, como identificado abaixo. Esta variabilidade e alto consumo, uma vez que são foram registados no inverno, dever-se-ão, maioritariamente, a 1) aparelhos de aquecimento elétricos e 2) necessidades de iluminação. Destaca-se ainda uma menor dispersão dos perfis no caso do agrupamento 4, em comparação ao agrupamento 1, sugerindo que o agrupamento 1 contém perfis de consumo mais distintos entre si.

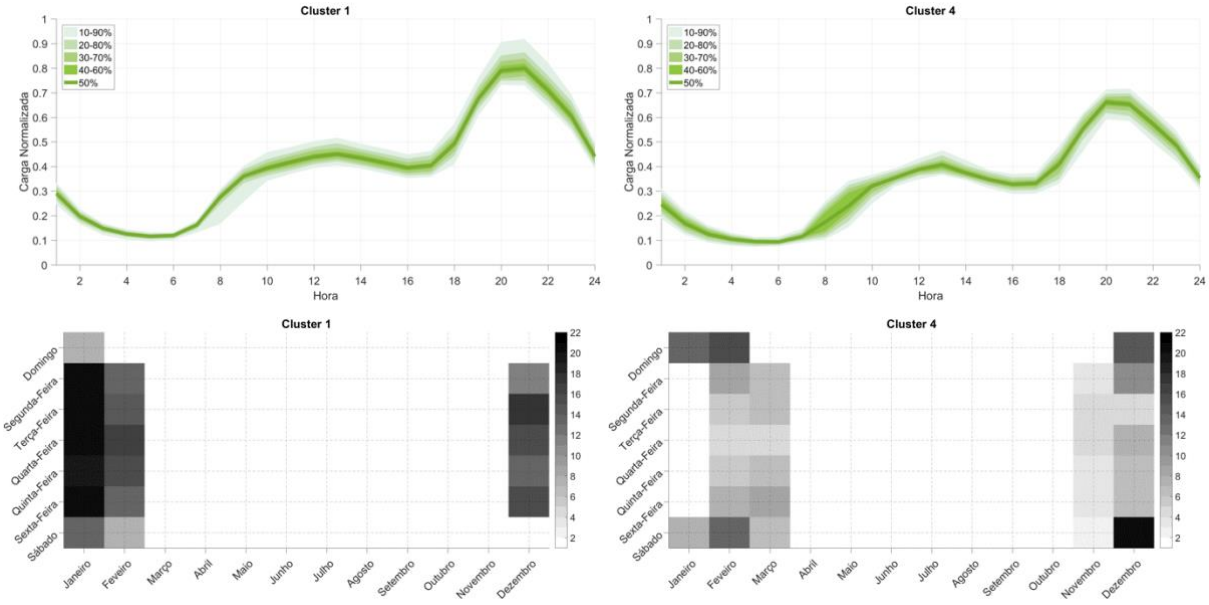

Figura 4.5: Percentis da carga BTN C e calendário de ocorrências dos agrupamentos: 1 (lado esquerdo) e 4 (lado direito).

Os *clusters* 2 e 5, representados na Figura 4.6, representam, maioritariamente a primavera e o outono, ou seja, as estações do ano de associadas às transições das condições atmosféricas. É um agrupamento natural, visto serem estações mais amenas, com consumos mais reduzidos do que no inverno. No *cluster* 2, a maior maioria das observações identificadas representam Março e Novembro, os meses imediatamente seguinte e anterior ao inverno, respetivamente, pelo que que é natural verificar aqui um consumo mais elevado associado às necessidades de aquecimento e iluminação. Este conjunto também apresenta alguma dispersão, sobretudo entre as 06-10h, que pode ser associada ao agrupamento simultâneo de dias de semana e de fim de semana, onde é expectável um comportamento distinto entre os consumidores deste tipo de perfil. Já o *cluster* 5 apresenta um consumo mais reduzido, com um pico acentuado por volta das 20h, quando existem maiores necessidades de consumo de energia elétrica quando os consumidores chegam a casa.

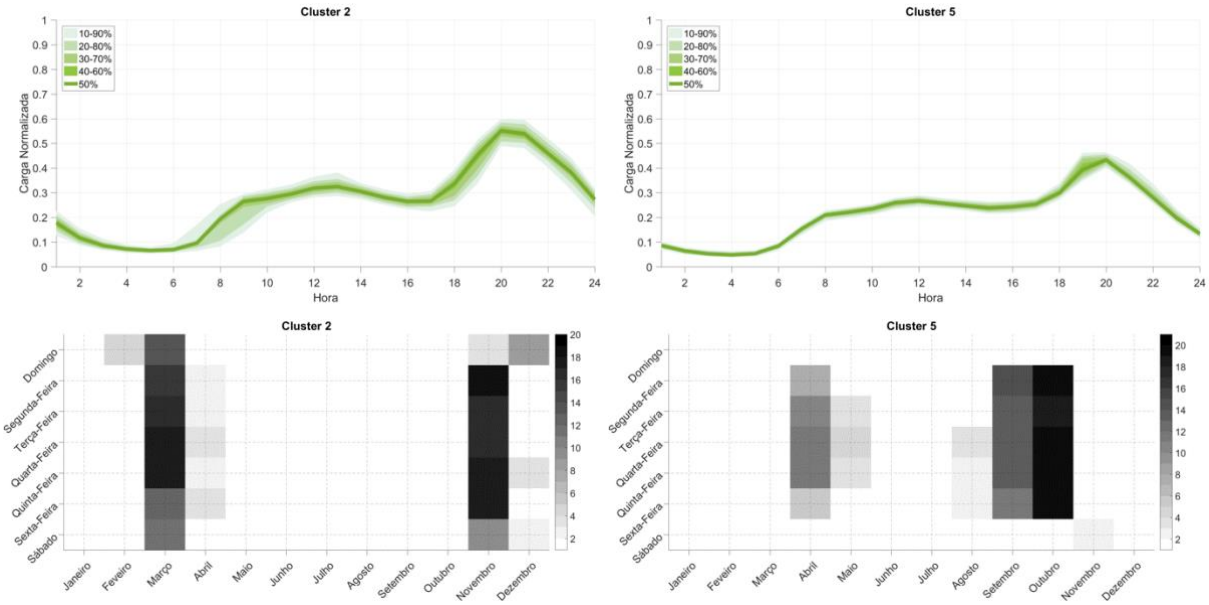

Figura 4.6: Percentis da carga BTN C e calendário de ocorrências dos agrupamentos: 2 (lado esquerdo) e 5 (lado direito).

Os *clusters* 6 e 7, representados na Figura 4.7, representam a carga registada nos dias de semana nos meses mais amenos em Portugal onde o consumo é mais reduzido, uma vez que as necessidades de aquecimento e de iluminação são mais reduzidas e nem toda a população dispõe de equipamento para arrefecimento. A diferença mais significativa entre estes dois perfis reside na magnitude do consumo. Destaca-se ainda que o agrupamento 6 apresenta alguma frequência de ocorrência nos meses de Março, Abril e Novembro durante os fins de semana. Este comportamento pode ser parcialmente explicado pela ocorrência de períodos onde as necessidades de aquecimento eram reduzidas e por isso o perfil de consumo é similar ao observado durante os meses mais amenos durante a semana. É ainda notório que apenas existe um pico de consumo à noite nestes dois agrupamentos, contrariamente ao verificado no agrupamentos 3 (Figura 4.8), que, embora com uma magnitude mais reduzida, apresenta dois picos e corresponde, maioritariamente, aos fins de semana sensivelmente no mesmo período que os agrupamentos 6 e 7. Este comportamento pode ser parcialmente explicado pelo facto de os consumidores inseridos neste tipo de perfil se encontrarem em casa e efetuarem consumos associados às atividades domésticas (e.g., máquina lavar roupa). Através da análise detalhada dos resultados, é ainda possível verificar que o agrupamento 3 encontra-se associado aos feriados durante os dias de semana no período entre abril e agosto.

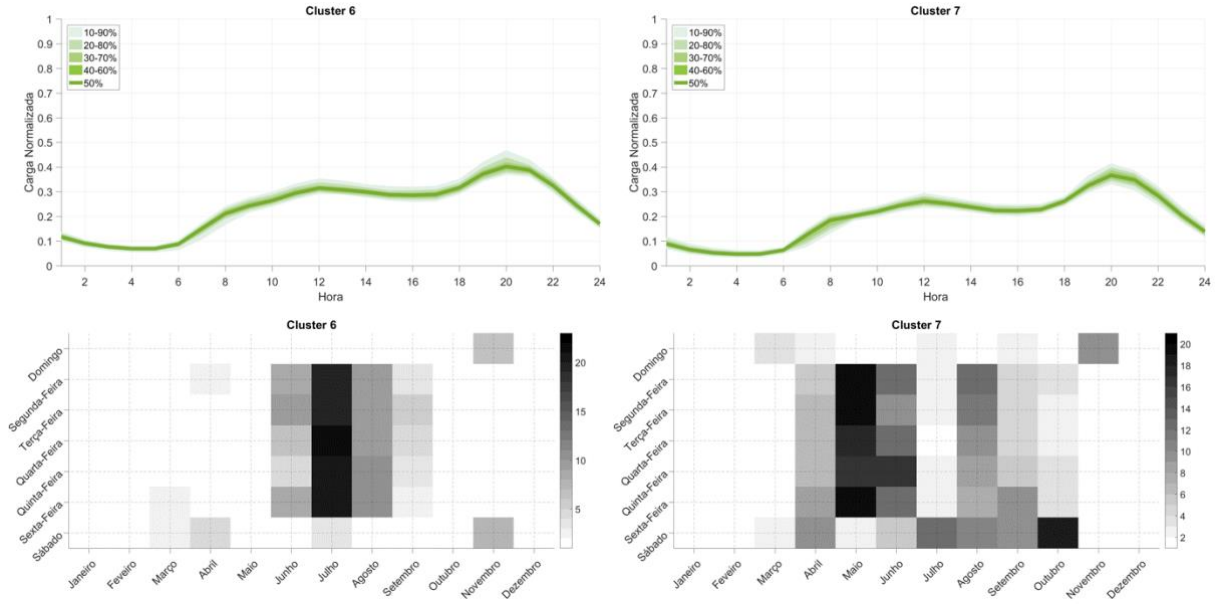

Figura 4.7: Percentis da carga BTN C e calendário de ocorrências dos agrupamentos: 6 (lado esquerdo) e 7 (lado direito).

É de notar que os todos os agrupamentos apresentam um número muito próximo de observações (Figura 4.9). A reduzida dispersão dos perfis anteriormente apresentados e o tipo de dias atribuídos a cada conjunto, demonstra que o método *k-medoids* conseguiu capturar o significado físico da separação efetuada. Assim, esta caracterização permitiu encontrar 1) uma distinção clara entre os consumos em dias de semana e dias de fim de semana; 2) perfis apresentam uma forte dependência do dia e mês em análise e 3) um padrão sazonal com consumos bastante distintos entre os meses quentes e os meses frios. Tendo isto em conta, os resultados apresentados e discutidos anteriormente encontram-se alinhados com a revisão de literatura efetuada, e, consequentemente, será pertinente introduzir informação sobre dias da semana e úteis, o mês em análise, bem como os índices dos agrupamentos obtidos nos modelos de previsão.

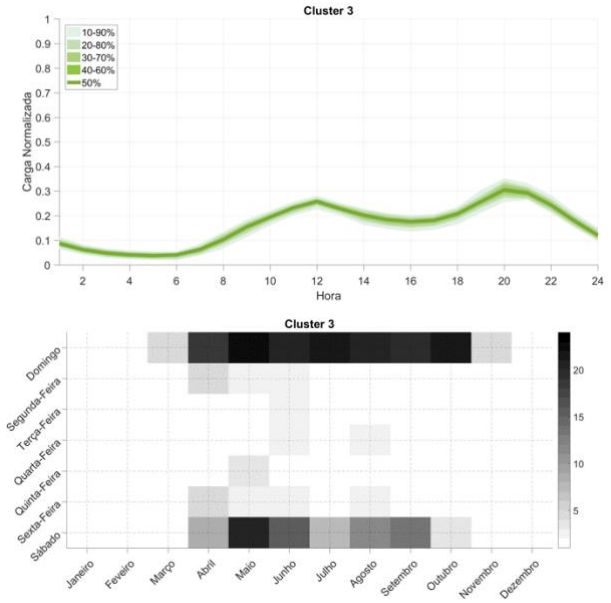

Figura 4.8: Percentis da carga BTN C e calendário de ocorrências do agrupamento 3

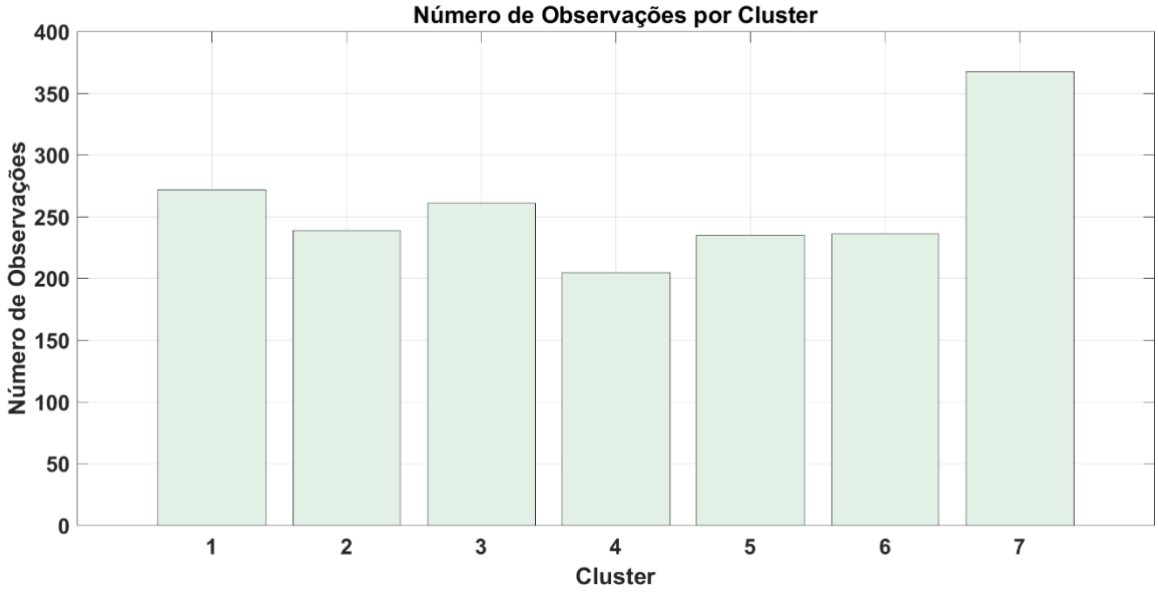

Figura 4.9: Número total de observações atribuídas a cada cluster

4.2 Previsão do consumo de energia elétrica

Nesta secção são apresentados os resultados das metodologias de previsão para um horizonte temporal de 24 horas.

4.2.1 Análise de variáveis endógenas e exógenas

O modelo inicial desenhado, antes de ser efetuada a seleção de atributos relevantes, está divido em 2 tipos de variáveis: variáveis endógenas, que se referem a valores extraídos da série temporal (dia útil, hora, dia da semana, mês e consumo elétrico de horas anteriores) e variáveis exógenas que se referem a variáveis meteorológicas (velocidade do vento, temperatura de bolbo seco, temperatura de ponto de orvalho, nebulosidade e pressão atmosférica).

As variáveis endógenas, que foram referidas na revisão bibliográfica, foram selecionadas por dois motivos adicionais, tendo um deles já sido exposto no subcapítulo anterior. As horas de atraso, para além das 24h anteriores, que estão na base do problema, foram selecionadas devido à forte correlação existente relativamente às horas anteriores, Figura 4.10.

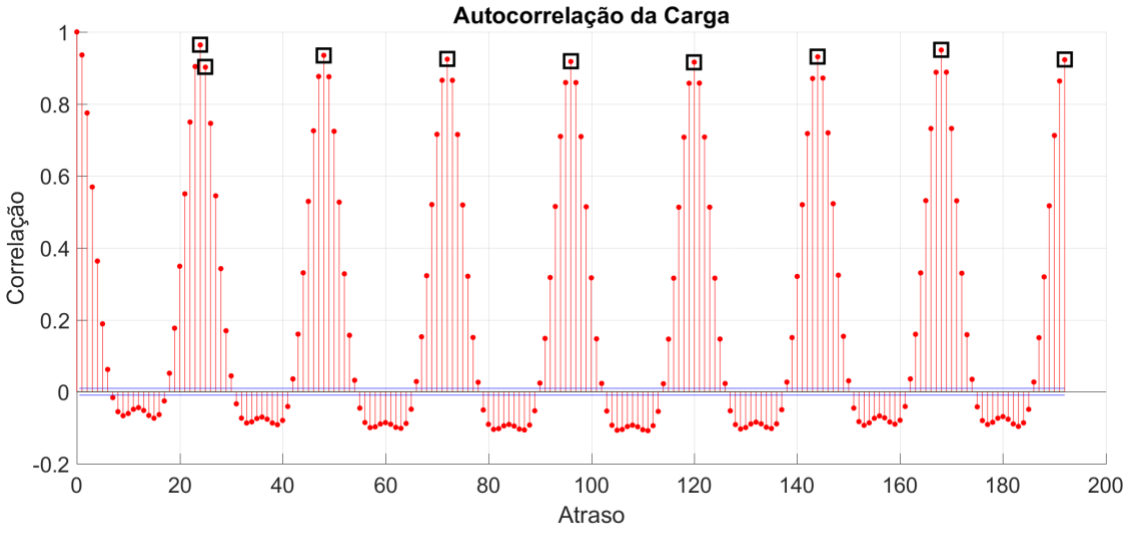

Figura 4.10: Autocorrelação da série temporal do consumo de energia elétrica do perfil tipo BTN C

Efetivamente, é possível verificar na Figura 4.10 que as mesmas horas dos 8 dias anteriores apresentam uma correlação superior a 0.9 (quadrados pretos) em relação ao consumo de energia atual. Desta forma, todos estes pontos foram introduzidos como variáveis independentes no modelo, *i.e.*, o modelo foi alimentado com informação dos desfasamentos temporais com correlação superior a 0.9.

Na Figura 4.11, apresentam-se as grelhas de correlação horária entre as séries temporais do consumo de energia elétrica do perfil em análise e as variáveis exógenas mais relevantes identificadas na revisão da literatura efetuada.

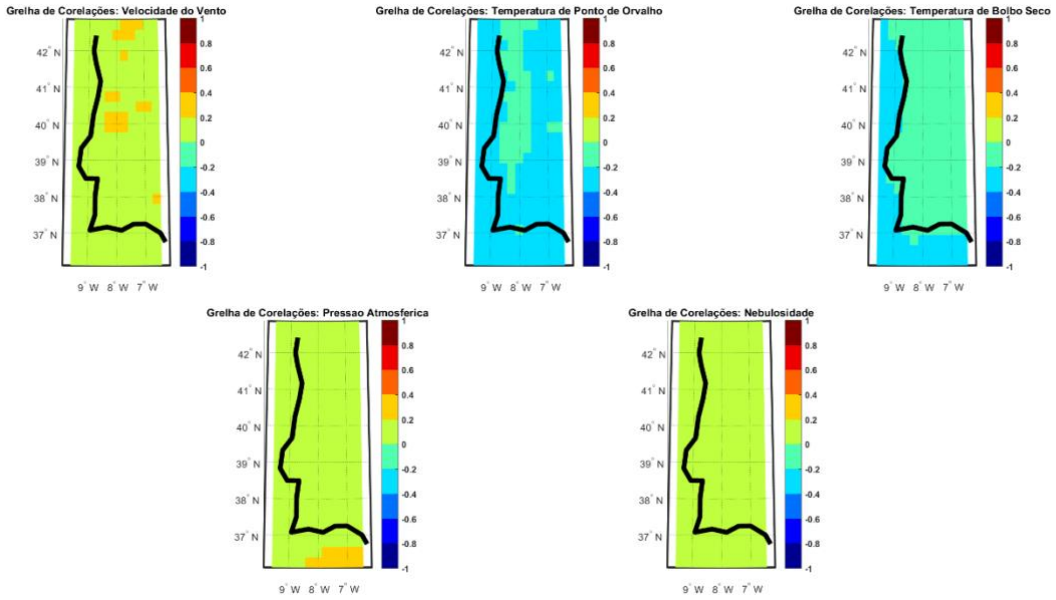

Figura 4.11: Grelhas de correlação espacial das variáveis meteorológicas relativamente ao consumo de energia elétrica do perfil BTN C

Na Figura 4.11 é possível verificar, em média, uma correlação bastante reduzida. Contudo, em alguns pontos das grelhas das variáveis da velocidade do vento, temperaturas de ponto de orvalho e bolbo seco e pressão atmosférica, os valores da correlação são mais significativos sugerindo alguma relação com o consumo de energia. A influência da temperatura (de bolbo seco) é ainda sustentada pela análise dos perfis típicos, onde agrupamentos distintos foram identificados para as épocas quentes e frias do ano. Assim, a baixa correlação aqui verificada pode dever-se a uma dependência não linear. Nesse sentido, decidiu adicionar-se os pontos da grelha mais correlacionados (em módulo) destas cinco variáveis, normalizados de acordo com a expressão 3.1, e permitir que o algoritmo de seleção de atributos relevantes avaliasse a sua influência no resultado final da previsão.

De modo a fazer a previsão operacional com um horizonte temporal até 24h, estas variáveis foram desfasadas neste mesmo número de horas de acordo com a Figura 4.12.

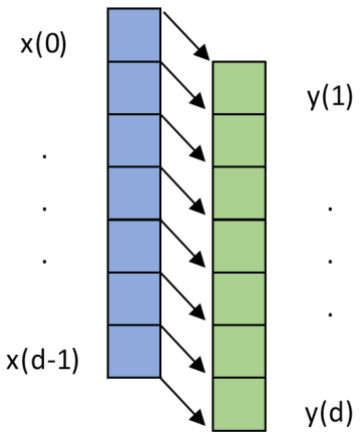

Figura 4.12: Desfasamento das variáveis exógenas (coluna azul) em relação ao valor do consumo (coluna verde) no dia d.

A matriz de variáveis a ser introduzida no algoritmo de seleção de atributos relevantes assume assim a seguinte forma:

Tabela 4.1: Matriz de variáveis selecionadas para o algoritmo de feature selection

Dia Util	Hora	Dia da Semana	Mes	h-24	h-25	h-48	h-72	h-96	h-120	h-144	h-168	h-192	Velocidade do Vento	Temperatura de Bolbo Seco	Temperatura de Ponto de Orvalho	Nebulosidade	Pressao Atmosferica	Cluster Index
1	0	7	1	185	252	189	189	196	177	200	196	172	-0,779	-0,147	0,028	-0,758	0,125	4
1	1	7	1	141	185	143	142	147	134	159	152	131	-0,800	-0,147	0,028	-0,589	0,109	4
1	2	7	1	116	141	119	118	120	113	131	123	110	-0,762	-0,148	0,041	-0,695	0,083	4
1	3	7	1	106	116	107	108	109	102	112	106	99	-0,738	-0,149	0,062	0,835	0,064	4
1	4	7	1	102	106	104	104	104	99	102	98	95	-0,764	-0,148	0,079	0,549	0,043	4

4.2.2 Seleção de atributos relevantes

O algoritmo de seleção de atributos relevantes foi aplicado no período entre 2014 a 2017 para identificar as variáveis que permitem a melhor performance dos três diferentes métodos de previsão do consumo de energia em análise.

A carga verificada nas 24h antes da hora a prever[15] foi o ponto de partida para todos os métodos, sendo em seguida adicionadas as restantes variáveis, uma a uma, aleatoriamente, e avaliando o erro quadrático médio. Como medida adicional, para a MLR, sendo uma técnica estatística, foi calcula uma tabela de ANOVA de modo a validar as variáveis aqui selecionadas. Estes resultados são apresentados no Anexo D. Na Tabela 4.2 são apresentadas as variáveis identificadas pelo algoritmo implementado para cada um dos métodos:

Tabela 4.2: Variáveis selecionadas para cada método indicadas com "1" (0 representa as variáveis descartadas)

Variável	Baseline	MLR	KNN	ANN
Dia Útil	0	1	0	1
Hora	0	1	1	1
Dia da Semana	0	1	1	1
Mês	0	1	1	1
Carga(h-24)	1	1	1	1
Carga(h-25)	0	1	0	1
Carga(h-48)	0	1	1	0
Carga(h-72)	0	1	1	1
Carga(h-96)	0	0	0	1
Carga(h-120)	0	1	1	1
Carga(h-144)	0	1	1	0
Carga(h-168)	0	1	1	1
Carga(h-192)	0	1	1	1
Velocidade do Vento	0	1	1	0
Temperatura de Bolbo Seco	0	1	0	1
Temperatura de Ponto de Orvalho	0	1	1	0
Nebulosidade	0	1	1	1
Pressão Atmosférica	0	1	0	1
Índice do Cluster	0	1	0	1

Como se pode verificar, existem variáveis transversais a todos os métodos, que se pressupõe aqui serem as que capturam a maior variabilidade da série, ou seja, as mais relevantes para o problema

[15] *h-n* refere-se à hora a prever, *h*, menos *n* horas.

em estudo. É de notar que estas são maioritariamente endógenas, pelo que a série pode ser robustamente modelada pelos seus valores passados, tendo as restantes variáveis ganhos marginais em termos de melhoria da previsão. No entanto, do ponto de vista da operacionalização do sistema, estes ganhos marginais podem representar melhorias na identificação de feriados ou outros períodos com consumos que se afastam da norma, pelo que pode existir um interesse prático em obtê-los. É ainda importante referir que os índices dos *clusters* foram selecionados pela ANN e KNN, revelando que houve um ganho de informação proveniente da introdução deste vetor, o que reforça a separação dos conjuntos criados no passo anterior. Os métodos de previsão baseados em redes neuronais artificiais (ANN) e *K* vizinhos mais próximo (KNN) são os que requerem menos variáveis. Por outro lado, o método MLR é o necessita de mais variáveis. Este resultado pode ser explicado pela capacidade do método ANN para lidar com fenómenos não-lineares sendo por isso necessário menos variáveis para obter uma elevada performance na previsão do consumo.

4.2.3 Previsão de referência (*baseline*)

Tendo como objeto de estudo a previsão de curto prazo do consumo de energia, torna-se necessário criar um modelo de referência (*baseline*), com o qual comparar o modelo e métodos desenvolvidos na presente dissertação. O modelo *baseline* escolhido é baseado numa regressão linear simples tendo como variável independente apenas o consumo registado nas 24h anteriores à hora em estudo. O pressuposto deste modelo simples é não necessitar de nenhum conhecimento sobre o problema em causa e avaliar se as variáveis adicionadas ao modelo final têm efeito benéfico na previsão. Na presente dissertação, a regressão linear simples que minimiza a soma de erros quadráticos, deu origem ao seguinte modelo:

$$\hat{y}_t = 0.0043 - 1.2008 y_{t-24} \tag{4.1}$$

4.2.4 Resultados da previsão

Os métodos de previsão foram avaliados de acordo com as métricas de erro apresentados na secção 2.3.5, nomeadamente, RMSE, Correlação, MAPE e viés. Estas métricas foram também avaliadas para o método de referência (*baseline)* de modo a conseguir avaliar a melhoria relativa de cada um dos métodos. De modo a testar os diferentes métodos treinados, o ano de 2018 completo) foi retirado do conjunto de treino/calibração e utilizado apenas para validação do modelo.

Na Figura 4.13 está representada a carga verificada uma semana de junho juntamente com as previsões efetuadas pelos diferentes métodos. É notório o bom ajuste da ANN (a verde) e do KNN (a roxo) relativamente à carga verificada (a azul), principalmente como evidenciado no dia 10 de junho, que corresponde a um feriado.

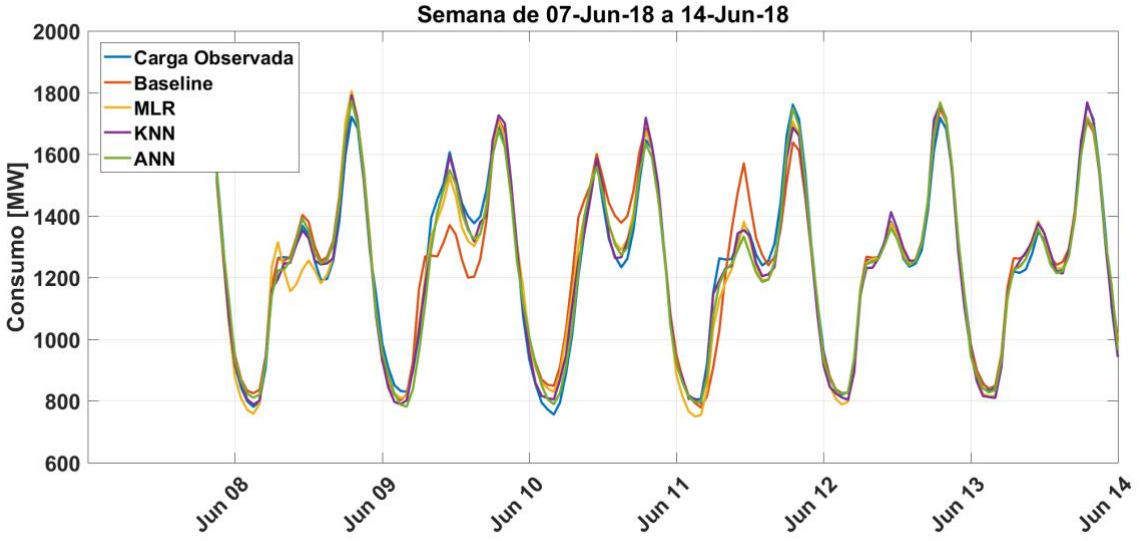

Figura 4.13: Resultados dos diferentes métodos para uma semana de junho de 2018

De modo a evidenciar este desempenho, a Figura 4.14 apresenta uma semana de setembro, com, onde são mais evidentes as dificuldades na previsão nos dias 10 (segunda-feira) e 16 (sábado). Esta dificuldade deve-se à transição de regime semana/fim-de-semana e vice-versa. Verifica-se que a Rede Neuronal, e aqui acompanhada também pela Regressão Linear Multivariada e pelo K-Vizinhos mais próximos, segue de perto os valores observados durante os dias 18 a 21. Gráficos adicionais de ajuste dos modelos à carga registada em diferentes semanas do ano são apresentados no Anexo E.

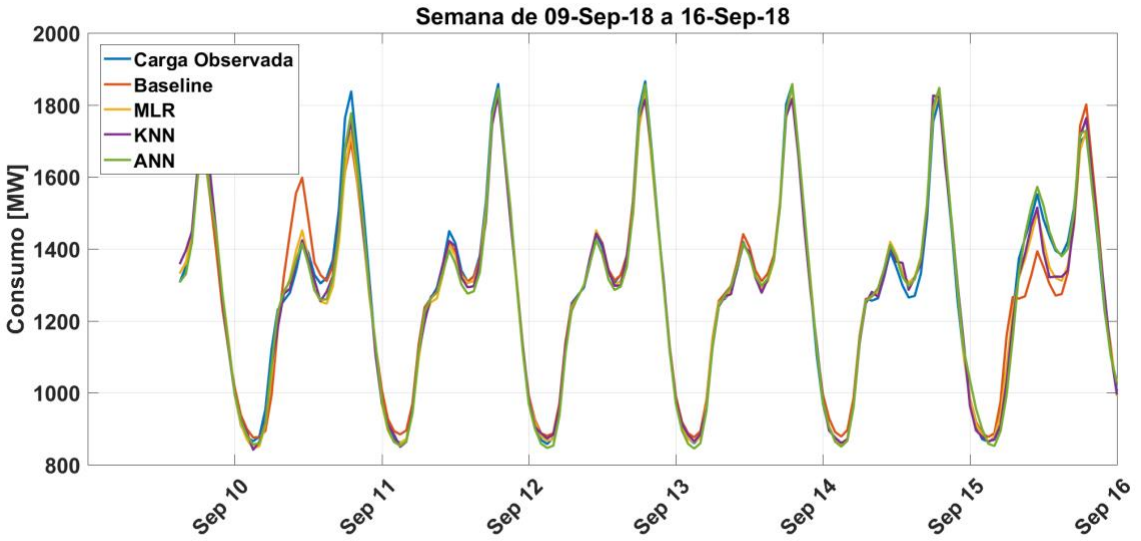

Figura 4.14: Resultados dos diferentes métodos para uma semana de setembro de 2018

De modo a obter uma maior confiança nos resultados obtidos, foram calculadas, para o ano de validação, métricas totais, horárias, por dia da semana e por mês do ano. Na Tabela 4.3 são apresentadas as métricas de obtidas para cada um dos métodos.

Tabela 4.3: Métricas totais obtidas para cada um dos métodos, no conjunto de teste (ano de 2018)

	Baseline	**MLR**	**KNN**	**ANN**
RMSE [MW]	103,7734	72,0839	68,9790	58,5489
Correlação	0,9689	0,9851	0,9864	0,9902
MAPE [%]	4,4484	3,3285	3,1532	2,8016
Viés [%]	0,46598	0,26803	0,11486	0,02472

O bom ajuste verificado para a Rede Neuronal é corroborado pelos resultados obtidos para as métricas utilizadas (Tabela 4.3). Apesar disto, é possível verificar que todos os métodos conseguiram melhorias significativas relativamente à *baseline*. De modo a obter uma maior sensibilidade às melhorias conseguidas, estas métricas foram avaliadas em maior detalhe, em termos horários, por dia da semana e mensal.

4.2.4.1 Análise dos desvios horários

Para além das métricas apresentadas anteriormente, para os erros horários foi ainda avaliada a variabilidade dos erros relativamente à hora do dia para a qual foi efetuada a previsão de modo a tentar perceber quais as horas mais "problemáticas" de prever.

Na Figura 4.15 é apresentada a dispersão horária do erro absoluto percentual.

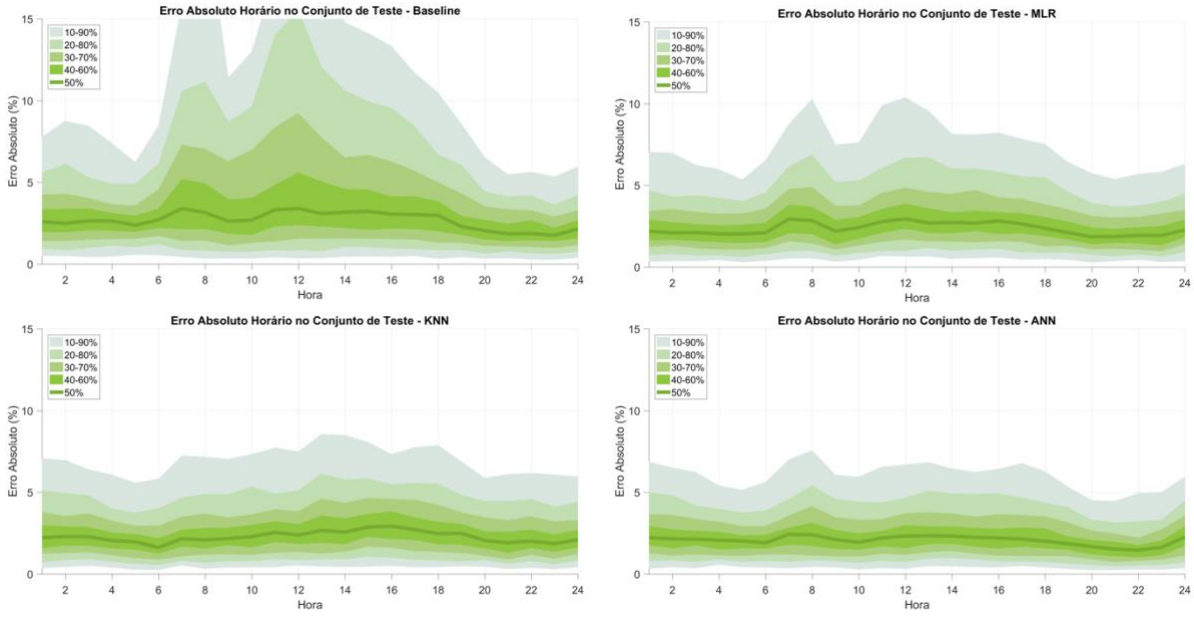

Figura 4.15: Perfil diário do Erro Absoluto Percentual Horário para todos os métodos avaliados

Pela Figura 4.15 é possível verificar que i) o método de referência (baseline) tem erros muito superiores a todos os outros métodos; ii) os métodos MLR como o KNN apresentam alguns erros elevados às horas de ponta de cheia da manhã e da tarde; e iii) a ANN apresenta um comportamento muito mais uniforme do que todos os outros métodos. Isto indica que a ANN não é tão influenciável por valores extremos (horas de ponta de cheia) como os restantes métodos.

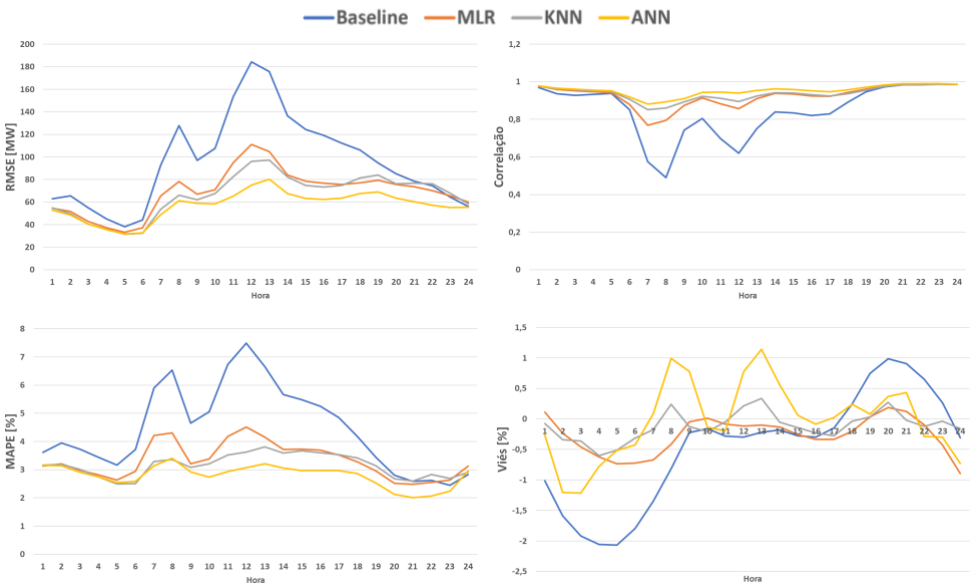

Figura 4.16: RMSE, Correlação, MAPE e Viés horários durante 2018

Na análise das restantes métricas (Figura 4.16) é igualmente evidente a problemática da previsão das horas de ponta enquanto que, simultaneamente evidenciam as melhorias de cada um dos métodos relativamente à *baseline* e entre si.

4.2.4.2 Análise dos desvios diários

Na Figura 4.17, apresentam-se os valores das métricas para cada dia da semana.

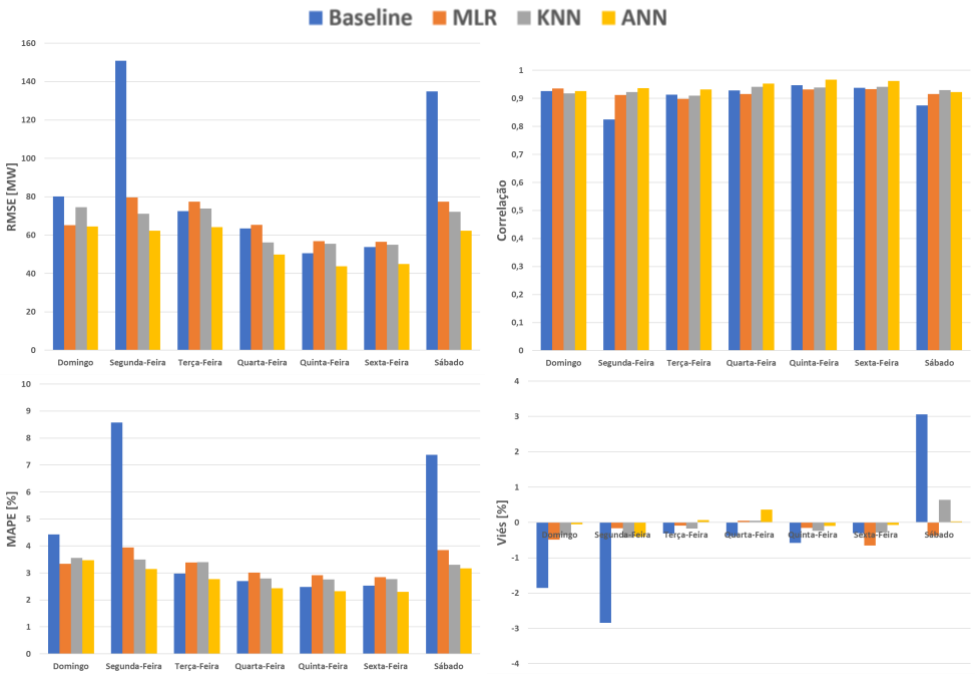

Figura 4.17: RMSE, Correlação, MAPE e Viés por dia da semana para o conjunto de teste

A avaliação dos erros por dia da semana evidencia a incapacidade do método de referência (*baseline*) de identificar os fins de semana. Verifica-se que os erros são mais elevados aos sábados, domingos e segundas-feiras. Isto acontece porque este modelo persiste a carga verificada no dia anterior, fazendo com que estes dias de transição sejam os mais problemáticos. No entanto, é possível verificar que o método MLR apresenta uma performance mais reduzida que o método de referência nos últimos dias da semana (quinta e sexta feira). O mesmo comportamento foi observado para o método KNN em algumas métricas estatísticas. É útil notar que a ANN apresenta consistentemente melhorias relativamente a todos os outros métodos, em média, para todos os dias da semana. No entanto, é ainda visível que a segunda-feira é o dia em que todos os métodos apresentam maiores dificuldades, devido à mudança súbita do regime de consumo.

4.2.4.3 Análise dos desvios mensais

Na Figura 4.18, apresentam-se os valores das métricas para cada mês do ano.

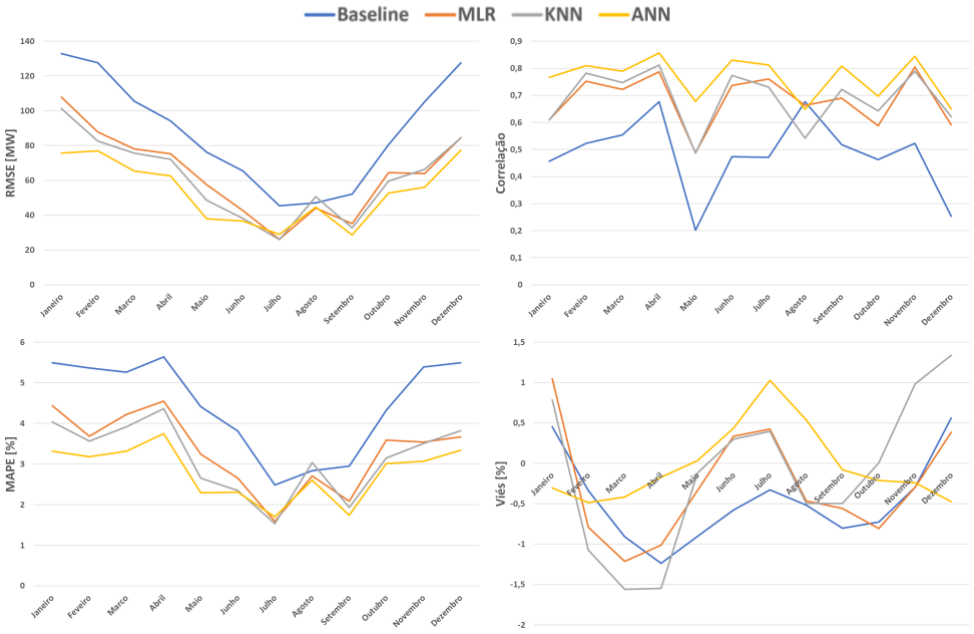

Figura 4.18: RMSE, Correlação, MAPE e Viés mensais para o conjunto de teste

A avaliação dos erros médio mensais indica que os meses de maior consumo são mais problemáticos de prever do que os que apresentam menor consumo. Este resultado pode ser associado à elevada variabilidade do consumo que existe nestes meses bem como a magnitude dos valores existentes. No mês de agosto é possível verificar igualmente um aumento dos valores do MAPE e RMSE, em comparação com os meses adjacentes. Este comportamento pode ser parcialmente explicado pela época estival onde o consumo de energia neste tipo de perfil é bastante alterado, face ao restante período de dados de treino/histórico, apresentando maior dificuldade na sua previsão

4.2.4.1 Análise dos desvios por agrupamento (*cluster*) dos perfis diários de consumo

Na Tabela 4.4 apresentam-se os erros obtidos relativamente a cada um dos *clusters* criados, de modo a perceber existe sistematicidade no erro de acordo com agrupamento em análise.

Tabela 4.4: Comparação do MAPE [%] relativo a cada um dos clusters obtidos para todos os métodos

Método	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Baseline	4,16	5,74	6,40	5,60	3,05	3,10	3,97
MLR	3,47	4,10	3,31	3,97	3,08	2,46	3,23
KNN	3,35	4,00	2,84	3,92	2,81	2,35	3,02
ANN	2,81	3,44	2,83	3,36	2,22	2,32	2,79

De acordo com a tabela anterior é possível identificar performance distinta de acordo com o agrupamento em análise. No entanto, as variações no método KNN e ANN são inferiores a 4%. Para estes métodos, o agrupamento 2, associado aos dias da semana nos meses de transição (março e novembro), é o que apresenta os desvios mais significativos na previsão. Por outro lado, o agrupamento 6, que foi associado aos dias de semana nos meses mais amenos em Portugal onde o consumo é mais reduzido, é o que apresenta, em média entre os algoritmos analisados, os desvios mais reduzidos. Destaca-se ainda o elevado erro observado no agrupamento 3 para o método usado como *baseline*. Este agrupamento encontra-se associado maioritariamente aos fins de semana e feriados, e, consequentemente, o perfil de consumo nas 24 horas anteriores é significativamente diferente.

Os valores apresentados na tabela anterior indicam igualmente que os erros mais reduzidos foram observados nos agrupamentos com baixa variabilidade no perfil diário. Isto reforça o valor que a caracterização destes perfis típicos de consumo de energia elétrica pode ter para melhorar a fiabilidade da previsão.

No Anexo F, são apresentados gráficos adicionais que demonstram a amplitude diária do erro absoluto percentual para cada um dos *clusters*.

4.2.4.2 Análise das distribuições dos erros da previsão

Apesar dos gráficos e métricas obtidos demonstrarem ajustes satisfatórios aos dados em estudo, podem existir problemas mais subtis com os modelos que não sejam bem capturados apenas com estes elementos. Como verificado no capítulo 0, a presença de viés foi prevalente em todos os métodos. Isto indica que há uma direção tendencial para os erros da previsão. Todos os métodos apresentaram um viés negativo, indicando que, em média, as previsões se encontram abaixo dos valores observados. Vários fenómenos podem influenciar este erro, desde as variáveis selecionadas aos métodos escolhidos para a previsão. No entanto, no presente problema, o candidato mais provável será o da omissão de variáveis independentes. Dois fatores pesam nesta conclusão:

1. O consumo de energia é um problema extremamente complexo e difícil de modelar, pelo que é virtualmente impossível encontrar todas as variáveis que explicam o consumo, especialmente à escala dos consumos domésticos;
2. Como se verifica na Figura 4.10, para além dos atrasos utilizados como variáveis independentes (mesma hora dos 8 dias anteriores) os atrasos das (cerca de) 4 horas imediatamente anteriores à hora atual também apresentam uma correlação relativamente elevada. Estes atrasos não foram utilizados, uma vez que, tendo em atenção a operacionalização da previsão, à data, não é prático obter estes valores em tempo quase real. Esta omissão conduz a uma perda do poder preditivo dos modelos, uma vez que informação importante e bastante correlacionada pode estar em falta.

Para complementar a análise anterior dos erros das previsões, e de modo a avaliar estes efeitos nos valores previstos para o período de validação, foram avaliadas as distribuições dos erros da previsão para todos os métodos.

A. Baseline

Na Figura 4.19 apresentam-se os gráficos dos diagnósticos efetuados ao modelo criado para a *baseline*.

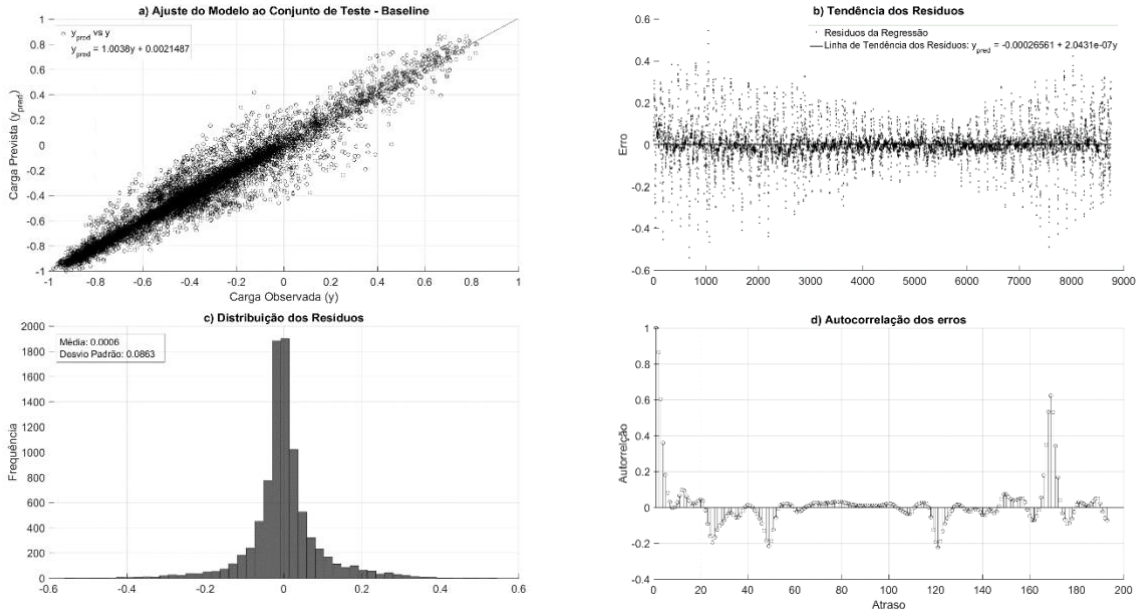

*Figura 4.19: Análise detalhada os resultados da previsão obtida com o método **baseline**. a) regressão entre os valores observados e previstos; b) tendência observada nos resíduos do conjunto de teste; c) distribuição dos resíduos; d) autocorrealação nos resíduos.*

A Figura 4.19 a) representa o ajuste da regressão. Aqui é possível visualizar uma dispersão considerável ao longo de toda a reta, indicando a existência de desvios significativos ao longo de toda a distribuição do consumo. As subfiguras b), c) e d) referem-se aos resíduos da regressão. Na primeira verificamos que existe uma tendência positiva (ainda que muito ligeira) nos resíduos que os afasta da média zero. Isto pode ser problemático a longo prazo, pelo que seria necessário reajustar frequentemente o modelo após a sua operacionalização. Pelos picos verificados na dispersão dos erros da mesma subfigura, é notório o padrão semanal existente (168 horas de atraso), que é corroborado pela Figura 4.19 d), onde está representada a autocorrelação. Aqui verificamos o pico referido anteriormente nas 168h. Na Figura 4.19 c) verificamos que os resíduos seguem uma distribuição *aproximadamente* normal, notando-se, no entanto, um ligeiro desvio para o lado positivo da distribuição. Isto deve-se ao viés presente no modelo.

B. Regressão Linear Múltipla

A Figura 4.20 apresenta os gráficos dos diagnósticos efetuados ao modelo criado para a MLR. Na Figura 4.20 b), os picos semanais já não são evidentes e na subfigura d) é de notar o pico das 168h está bastante atenuado. Isto deve-se ao facto de as variáveis adicionais terem sido introduzidas no modelo, o que representa um ganho de informação adicional, diminuindo assim os erros sistemáticos obtidos com o método de referência.

Verifica-se que distribuição dos erros está agora centrada, demonstrando que o viés diminuiu bastante da *baseline* para este modelo.

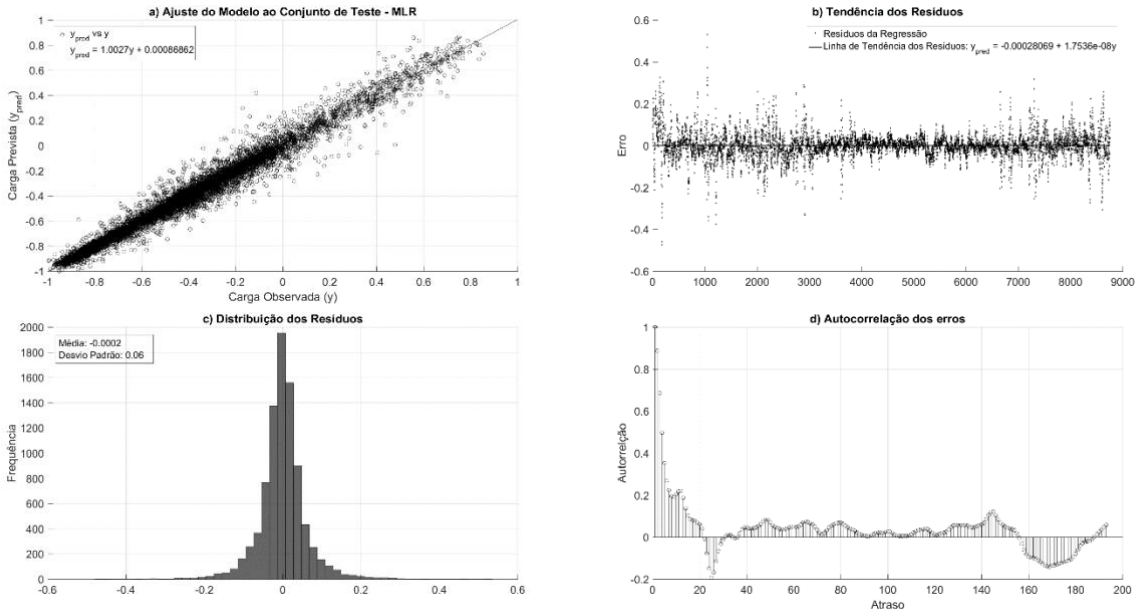

*Figura 4.20: Análise detalhada os resultados da previsão obtida com a **Regressão Linear Múltipla**. a) regressão entre os valores observados e previstos; b) tendência observada nos resíduos do conjunto de teste; c) distribuição dos resíduos; d) autocorrealação presente nos resíduos.*

C. *K* Vizinhos Mais Próximos

Na Figura 4.21 apresentam-se os gráficos dos diagnósticos efetuados à regressão obtida pelo método KNN para o conjunto de validação.

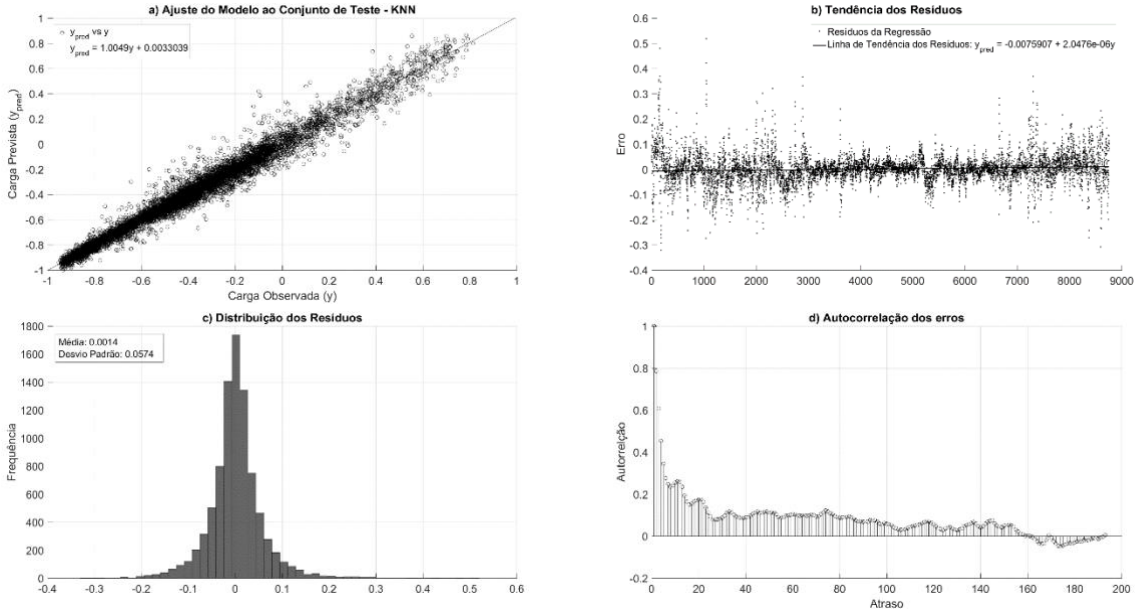

*Figura 4.21: Análise detalhada os resultados da previsão obtida com o **k vizinhos mais próximos**. a) regressão entre os valores observados e previstos; b) tendência observada nos resíduos do conjunto de teste; c) distribuição dos resíduos; d) autocorrealação presente nos resíduos.*

No KNN nota-se uma menor dispersão na Figura 4.21 a), o que indica um melhor ajuste ao conjunto de teste. Nota-se que a autocorrelação se mantém também aqui, e os resíduos apresentam uma

tendência positiva. A distribuição apresenta um ligeiro desvio positivo, no entanto aparenta estar muito próxima da *normal*.

D. Rede Neuronal Artificial

Na Figura 4.22 apresentam-se os gráficos dos diagnósticos efetuados à regressão obtida pelo método ANN para o conjunto de validação.

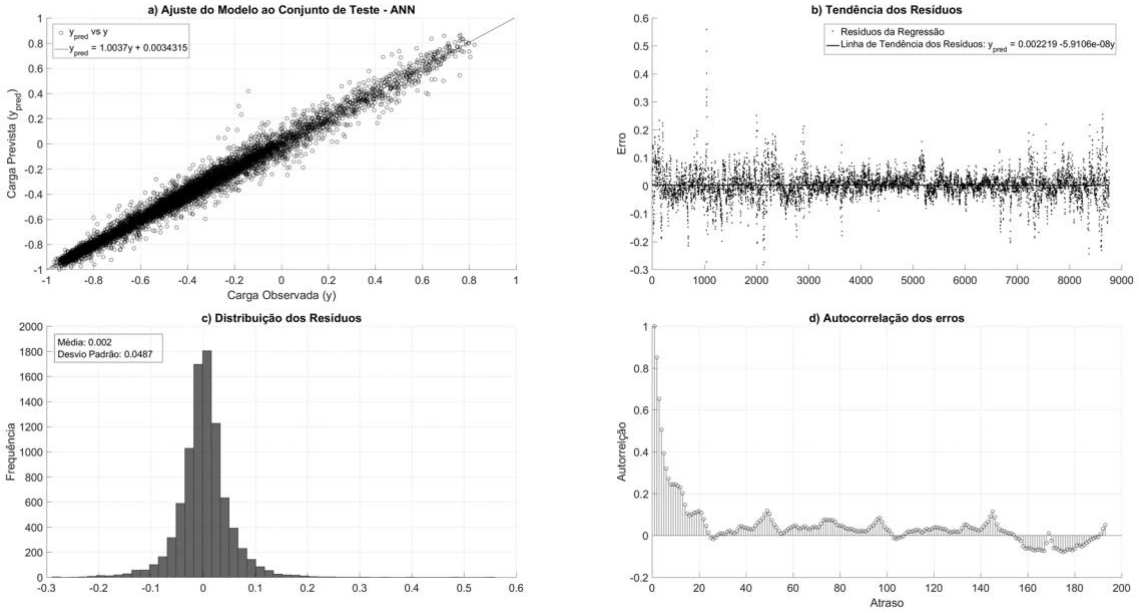

*Figura 4.22: Análise detalhada os resultados da previsão obtida com **Rede Neuronal Artificial**. a) regressão entre os valores observados e previstos; b) tendência observada nos resíduos do conjunto de teste; c) distribuição dos resíduos; d) autocorrealação presente nos resíduos.*

A regressão feita pela rede neuronal apresentou os melhores resultados. Isto pode ser verificado também pela dispersão apresentada na Figura 4.22 a). É ainda útil notar que os problemas destacados nos métodos analisados anteriormente (autocorrelação e tendência positiva dos resíduos, sendo esta segunda muito ligeira, apresentando agora os resíduos uma distribuição perfeitamente centrada) destacados se encontram também presentes evidenciados neste método.

5 Conclusão e desenvolvimentos futuros

O crescimento da penetração de fontes de energia não despacháveis no SE acarreta desafios ao nível operacional, nomeadamente, a gestão da produção para suprir as necessidades do consumo. Com vista a tentar mitigar este impacto, os novos paradigmas de gestão do binómio oferta/procura (e.g., gestão da procura) tentam atuar igualmente no lado da procura, de modo a conseguir encontrar o equilíbrio do sistema. Para isso, no entanto, são necessárias estimativas confiáveis do consumo de energia elétrica. Nesse sentido, a presente dissertação implementa e analisa a performance de três metodologias de previsão para um horizonte temporal de 24 horas.

De modo a atingir os objetivos definidos na introdução desta dissertação, foram analisados dados associados a consumo domésticos, disponibilizados pelo operador do sistema electroprodutor nacional. Estes dados correspondem aos perfis de consumo tipo para clientes finais de baixa tensão normal (BTN) classe C. A motivação para o uso desta classe assenta no potencial que este tipo de consumo oferece para os novos paradigmas como a gestão do consumo e ao seu peso no consumo total de energia elétrica em Portugal. Foram ainda utilizadas variáveis meteorológicas para avaliar a sua influência na previsão do consumo de energia.

A primeira parte do trabalho desenvolvido passou pela aquisição, tratamento e transformação dos dados brutos. Este processamento passou pela verificação da integridade, conversão de formato e extração de atributos relevantes intrínsecos à série temporal da carga. Destes atributos identificados destacam-se os valores passados do consumo de energia elétrica, que apresentam uma forte correlação com o consumo atual. Posteriormente, foram caracterizados os perfis diários típicos do consumo, onde foram identificadas e caracterizadas a frequência de ocorrência de cada perfil de consumos de acordo com os dias da semana e meses do ano. Este procedimento permitiu concluir que existe uma forte dependência dos perfis de consumo destes dois parâmetros sendo relevante a sua introdução como variável de entrada no modelo de previsão. Assim, a identificação da sazonalidade foi introduzida através do mês a que a previsão diz respeito. Foram também introduzidos os índices dos conjuntos encontrados. Para além destas variáveis, foram ainda consideradas variáveis meteorológicas, cuja correlação com este perfil de consumo provou ser reduzida.

Com base na informação assimilada nos passos anteriores, juntamente com a informação identificada na revisão de literatura aplicou-se, de forma individual, um algoritmo de seleção de atributos relevantes (*feature selection*) aos três métodos de previsão analisados nesta dissertação: Regressão Linear Multivariada, *K* Vizinhos mais Próximos e Rede Neuronal Artificial. Este procedimento foi efetuado para o período de teste/calibração (2014-2017) e visou identificar nas variáveis disponíveis para alimentar os métodos de previsão, aquelas que permitem a melhor performance dos métodos de previsão.

Com base nas métricas definidas (Erro Médio Absoluto Percentual - MAPE, Raiz do Erro Quadrático Médio – RMSE, correlação e viés) os três métodos de previsão foram avaliados e comparados a um método de referência (*baseline*) para o período de validação (ano de 2018). Os resultados permitem observar uma melhoria significativa dos três métodos de previsão relativamente ao método de referência utilizado. Todos os métodos aqui utilizados foram identificados como boas opções para este tipo de problemas. Contudo os resultados obtidos demonstram que a Rede Neuronal Artificial possibilitou a obtenção das previsões mais fiáveis de entre os métodos avaliados. Esta demonstrou ter uma boa capacidade de identificação de fins de semana e feriados, provando assim ser útil no problema da previsão do consumo de energia elétrica a curto prazo. A metodologia *K* Vizinhos mais próximos apresentou igualmente uma elevada performance nas métricas estatísticas analisadas. Relativamente ao método baseado em regressões lineares múltiplas, em média, este apresenta uma performance superior

ao método de referência. Contudo, uma análise mais detalhada aos resultados deste método permite identificar que em determinados dias (quinta e sexta feiras), o MAPE e RMSE do método de referência apresenta valores mais reduzidos. Foi ainda possível identificar o sábado, domingo e segunda-feira como os dias mais problemáticos para a previsão do consumo de energia, sendo que este sábado e segunda-feira são dias de transição de regimes de consumo (semana para fim de semana e vice-versa) e que o domingo tem um comportamento típico, com um consumo mais reduzido e um diagrama de carga mais achatado (picos mais reduzidos). Quanto aos meses do ano, registam-se maiores dificuldades na previsão nos períodos frios, uma vez que isto conduz a um aumento do consumo doméstico e uma elevada variabilidade, originando um aumento nos erros obtidos.

Como trabalho futuro, seria interessante tentar criar modelos para cada um dos perfis típicos identificados neste trabalho (usualmente designada na literatura como *regime-switching approach*). A baixa variabilidade dentro de cada um destes perfis poderia trazer benefícios na previsão. O maior desafio aqui seria classificar os dias *a priori* de modo a selecionar o modelo correto a utilizar. Este estudo incidiria sobre a avaliação da possibilidade de identificação dos agrupamentos (*clusters*) através das variáveis independentes utilizadas. De forma a aprofundar esta temática, trabalhos futuros deveriam analisar igualmente a aplicação destas metodologias aos diferentes perfis de consumo disponibilizados pelo operador do sistema elétrico nacional. Este mesmo agrupamento pode também ser realizado ao nível dos consumidores, com a disseminação dos *smart meters*. Este agrupamento teria em vista segmentar consumidores de acordo com os seus perfis de consumo. Do ponto de vista da criação *microgrids* seria vantajosa esta segmentação, de modo distribuir de forma equilibrada as cargas pela rede.

6 Bibliografia

[1] L. Adkins, P. Benoit, M. Gray, C. Hood, G. Kamiya, e C. Lee, «Energy, Climate Change and Enviroment 2016 Insights», 2016.

[2] UNFCCC. Conference of the Parties (COP), «Paris Climate Change Conference-November 2015, COP 21», 2015.

[3] N. O'Connell, P. Pinson, H. Madsen, e M. O'Malley, «Benefits and challenges of electrical demand response : A critical review», *Renew. Sustain. Energy Rev.*, vol. 39, pp. 686–699, 2014.

[4] J. G. Jetcheva, M. Majidpour, e W. P. Chen, «Neural network model ensembles for building-level electricity load forecasts», *Energy Build.*, vol. 84, pp. 214–223, 2014.

[5] E. A. Feinberg e D. Genethliou, «Load Forecasting», em *Applied Mathematics for Restructured Electric Power Systems*, 1.a ed., J. H. Chow, F. F. Wu, e J. A. Momoh, Eds. New York: Springer US, 2006, pp. 269–285.

[6] C. Kuster, Y. Rezgui, e M. Mourshed, «Electrical load forecasting models: A critical systematic review», *Sustain. Cities Soc.*, vol. 35, n. August, pp. 257–270, 2017.

[7] Y. H. Hsiao, «Household Electricity Demand Forecast Based on Context Information and User Daily Schedule Analysis From Meter Data», *IEEE Trans. Ind. Informatics*, vol. 11, n. 1, pp. 33–43, 2015.

[8] REN, «SIMEE - Perfis Consumo», 2017. [Em linha]. Disponível em: http://www.mercado.ren.pt/PT/Electr/InfoMercado/Consumo/Paginas/PerfisConsumo.aspx. [Acedido: 25-Nov-2017].

[9] C. Chatfield, *Time-Series Forecasting*. Florida: Chapman &C Hall/CRC No, 2000.

[10] D. C. Montgomery, C. L. Jennings, e M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, 2nd ed. New Jersey: JohnWiley & Sons, Inc., 2015.

[11] M. Q. Raza e A. Khosravi, «A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings», *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, 2015.

[12] K. Amasyali e N. M. El-gohary, «A review of data-driven building energy consumption prediction studies», *Renew. Sustain. Energy Rev.*, vol. 81, n. 2018, pp. 1192–1205, 2018.

[13] The Pennsylvania State University, «5.2 Smoothing Time Series | STAT 510», 2018. [Em linha]. Disponível em: https://onlinecourses.science.psu.edu/stat510/node/70. [Acedido: 01-Fev-2018].

[14] The Pennsylvania State University, «5.1 Decomposition Models | STAT 510», 2018. [Em linha]. Disponível em: https://onlinecourses.science.psu.edu/stat510/node/69. [Acedido: 01-Fev-2018].

[15] I. P. Panapakidis, «Clustering based day-ahead and hour-ahead bus load forecasting models», *Int. J. Electr. Power Energy Syst.*, vol. 80, pp. 171–178, 2016.

[16] M. Mordjaoui, S. Haddad, A. Medoued, e A. Laouafi, «Electric load forecasting by using dynamic neural network», *Int. J. Hydrogen Energy*, vol. 42, n. 28, pp. 17655–17663, 2017.

[17] T. Hong e S. Fan, «Probabilistic electric load forecasting: A tutorial review», *Int. J. Forecast.*, vol. 32, n. 3, pp. 914–938, 2016.

[18] S. Ramos, I. Praca, Z. Vale, T. M. Sousa, e V. Faria, «Load profiling tool to support smart grid operation scenarios», *2014 IEEE PES T&D Conf. Expo.*, pp. 1–5, 2014.

[19] A. Tanwar, E. Crisostomi, P. Ferraro, M. Raugi, M. Tucci, e G. Giunta, «Clustering Analysis of the Electrical Load in European Countries», *"Metodi e tecniche innovative per l'Integrazione di Sistemi per l'energia Elettrica e Termica (MISET)*. pp. 1–8, 2015.

[20] S. Ramos, J. M. Duarte, F. J. Duarte, Z. Vale, e P. Faria, «A data mining framework for electric load profiling», *IEEE PES Conf. Innov. Smart Grid Technol. Lat. Am. (ISGT LA)*, pp. 1–6, 2013.

[21] S. Aghabozorgi, A. Seyed Shirkhorshidi, e T. Ying Wah, «Time-series clustering - A decade review», *Inf. Syst.*, vol. 53, pp. 16–38, 2015.

[22] O. Maimon e L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. New York: Springer Science+Business Media, LLC 2005, 2010, 2010.

[23] M. Halkidi, Y. Batistakis, e M. Vazirgiannis, «Clustering algorithms and validity measures», *Proc. Thirteen. Int. Conf. Sci. Stat. Database Manag. SSDBM 2001*, n. May 2014, pp. 3–22, 2001.

[24] J. Han, M. Kamber, e J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, Massachussets: Morgan Kaufmann Publishers, 2012.

[25] V. Satopää, J. Albrecht, D. Irwin, e B. Raghavan, «Finding a "kneedle" in a haystack: Detecting knee points in system behavior», *Proc. - Int. Conf. Distrib. Comput. Syst.*, pp. 166–171, 2011.

[26] J. Yang *et al.*, «k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement», *Energy Build.*, vol. 146, pp. 27–37, 2017.

[27] P. P. S. Bradley, U. Fayyad, e C. Reina, «Scaling clustering algorithms to large databases», *Knowl. Discov. Data Min.*, pp. 9–15, 1998.

[28] The MathWorks Inc., «Hierarchical Clustering - MATLAB & Simulink», 2019. [Em linha]. Disponível em: https://www.mathworks.com/help/stats/hierarchical-clustering.html. [Acedido: 20-Set-2019].

[29] X. Wang, K. Smith, e R. Hyndman, «Characteristic-based clustering for time series data», *Data Min. Knowl. Discov.*, vol. 13, n. 3, pp. 335–364, 2006.

[30] X. Liu, G. Xiong, X. Liu, R. Anand, X. Shang, e J. Cao, «Smart cities, urban sensing, and big data: mining geo-location in social networks», em *Big Data and Smart Service Systems*, X. Liu, R. Anand, G. Xiong, X. Shang, X. Liu, e J. Cao, Eds. Academic Press, 2017, pp. 59–84.

[31] F. J. F. Duarte, A. L. N. Fred, e M. F. C. Rodrigues, «Weighted Evidence Accumulation Clustering Using Subsampling», *Pattern Recognit. Intell. Syst.*, pp. 22–28, 2006.

[32] E. Vinagre, L. Gomes, e Z. Vale, «Electrical energy consumption forecast using external facility data», *Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015*, n. 12004, pp. 659–664, 2016.

[33] K. Chapagain e S. Kittipiyakul, «Performance analysis of short-term electricity demand with atmospheric variables», *Energies*, vol. 11, n. 4, pp. 2015–2018, 2018.

[34] G. Chandrashekar e F. Sahin, «A survey on feature selection methods», *Comput. Electr. Eng.*, vol. 40, n. 1, pp. 16–28, 2014.

[35] L. Chuang, C. Ke, e C. Yang, «A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification», em *International MultiConference of Engineers and Computer Scientists*, 2008, pp. 146–150.

[36] L. Friedrich e A. Afshari, «Short-term Forecasting of the Abu Dhabi Electricity Load Using Multiple Weather Variables», *Energy Procedia*, vol. 75, pp. 3014–3026, 2015.

[37] G. Dudek, «Neural networks for pattern-based short-term load forecasting: A comparative study», *Neurocomputing*, vol. 205, pp. 64–74, 2016.

[38] E. Vinagre, T. Pinto, S. Ramos, Z. Vale, e J. M. Corchado, «Electrical Energy Consumption Forecast Using Support Vector Machines», em *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, 2017, pp. 171–175.

[39] ERSE, «Regulamento de Relações Comerciais do setor elétrico», *Entidade Reguladora dos Serviços Energéticos*, 2018. .

[40] EDP Distribuição, «Atualização dos perfis de consumo , de produção e de autoconsumo para o ano de 2018 Documento Metodológico (artigo 272 . º do Regulamento de Relações Comerciais)», 2019.

[41] UCAR, «NCAR's RDA», 2019. [Em linha]. Disponível em: https://rda.ucar.edu/. [Acedido: 10-Jan-2018].

[42] S. Raschka, «About Feature Scaling and Normalization», 2018. [Em linha]. Disponível em: http://sebastianraschka.com/Articles/2014_about_feature_scaling.html#feature-scaling---standardization. [Acedido: 03-Fev-2018].

[43] L. Jin *et al.*, «Comparison of Clustering Techniques for Residential Energy Behavior using Smart Meter Data», *Int. Conf. Artif. Intell. Smart Grids Build.*, n. Chicco 2012, pp. 260–266, 2017.

[44] J. Gomes, «Regressão Linear», *Apontamentos de Estatísitca Aplicada - DEIO, FCUL*. pp. 1–74, 2011.

[45] J. D. Kelleher, B. Mac Namee, e A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics*. Cambridge, Massachusetts: MIT Press, 2015.

[46] X. Wu *et al.*, «Top 10 algorithms in data mining», *Knowl. Inf. Syst.*, vol. 14, n. 1, pp. 1–37, 2008.

[47] P. Cortez e J. Neves, «Redes Neuronais Artificias», *Escola de Engenharia, Universidade do Minho*. p. 52, 2000.

[48] H. Chen, C. a. Canizares, e A. Singh, «ANN-based short-term load forecasting in electricity markets», *2001 IEEE Power Eng. Soc. Winter Meet. Conf. Proc. (Cat. No.01CH37194)*, vol. 2, n. C, pp. 411–415, 2001.

[49] P. J. Rousseeuw, «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis», *J. Comput. Appl. Math.*, 1987.

[50] A. Ralhan, «Self Organizing Maps - Towards Data Science», *Self Organizing Maps*, 2018. [Em linha]. Disponível em: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4. [Acedido: 20-Set-2019].

Anexos

Anexo A – Avaliação dos agrupamentos obtidos

A.1 - Descrição do método da Silhueta

Uma vez determinada a composição de cada agrupamento (*cluster*), é necessário avaliar esta de uma forma mais rigorosa. Uma vez que não temos classes com a as quais possamos comparar os resultados obtidos, esta avaliação torna-se mais difícil. Uma forma intuitiva de abordar este problema consiste em relacionar a semelhança das observações atribuídas a cada cluster com a semelhança das mesmas relativamente a observações atribuídas a clusters diferentes. Mais formalmente, como referido em [49], podemos recorrer à técnica da silhueta, que nos permite representar graficamente semelhança dentro de cada cluster.

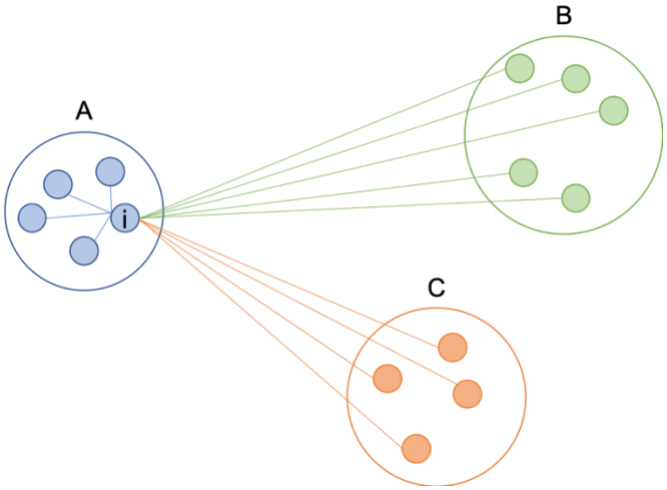

Figura A. 1: Distâncias entre i e as restantes observaçõe

De modo a tornar explicação mais clara, definamos o conjunto dos conjuntos formados pelo método de *clustering*, $C = \{A, B, C\}$. A silhueta de cada observação é dada por:

$$s_i = \frac{b_i - a_i}{max\{a_i,\ b_i\}} \tag{A.1}$$

Onde a_i é a dissemelhança média do elemento *i* do conjunto A em relação a todos os elementos do mesmo conjunto. Esta medida representa a semelhança da observação *i* aos restantes elementos atribuídos ao seu conjunto. Por sua vez, é necessário calcular a dissemelhança da observação *i* às observações atribuídas a outros conjuntos, b_i. Este valor é dado pela dissemelhança mínima às observações que não estão contidas no conjunto A, $min\{d(i,j)\}, \{j \in C \mid j \notin A\}$. Estas distâncias estão representadas na Figura A. 1.

É possível afirmar que há um bom agrupamento caso o valor de s_i esteja próximo de 1. Isto acontece se a_i for muito menor que b_i, isto é, se a dissemelhança entre *i* e os elementos do conjunto a que foi atribuído pelo algoritmo seja muito menor que a dissemelhança entre *i* e todas as observações que foram atribuídas a agrupamentos diferentes.

Uma vez calculados os valores de *s* para todas as observações, podemos representá-las num gráfico, como representado na Figura A. 2.

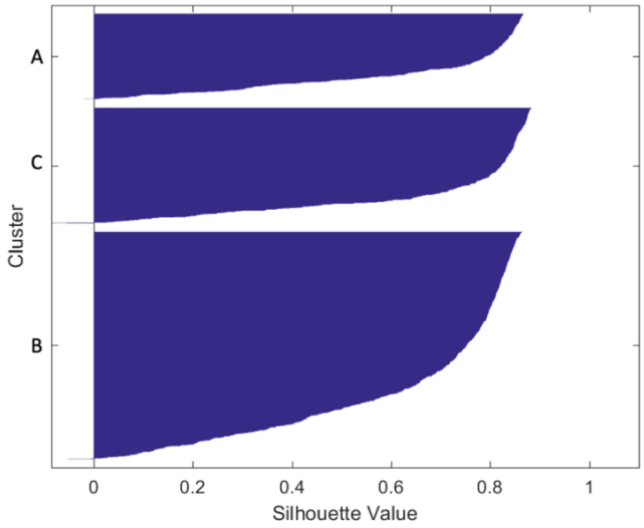

Figura A. 2: Exemplo de um gráfico de silhueta

A.2 - Validação do k obtido através deste método

A mostra o gráfico da silhueta dos agrupamentos obtidos.

Figura A. 3: Silhueta dos agrupamentos obtidos

Aqui as silhuetas com valores mais elevados encontram-se afastadas dos restantes conjuntos, indicando que estão corretamente separadas. À medida que o valor da silhueta se aproxima de zero, a distância da observação a agrupamentos vizinhos aproxima-se da fronteira de decisão. Um valor negativo indica que a observação pode conter alguma informação que pode ser associada a outro agrupamento.

Anexo B – Método KNN

B.1 – Implementação do algoritmo em MATLAB

De seguida apresenta-se o código criado, em Matlab, para a previsão de acordo com o método KNN:

```
function [y_pred] = knnRegressor(X_stored, y_stored, X_query, k)
%% KNN regression
% The function takes in 3 arguments:
%   - X_stored: past values of the independent variables in the
database;
%   - X_query: values of X for which the corresponding y is to be
%   predicted;
%   - k:  number of closest observations (neighbors) over which to
compute
%   the prediction of y (y_pred).
%
% y_pred is computed as the average over the k closest corresponding
% y_stored values

y_pred = zeros(size(X_query, 1), 1);
[idx, dist] = knnsearch(X_stored, X_query, 'k', k, 'Distance',
'euclidean');

for i = 1:size(idx,1)
    y_pred(i) = mean(y_stored(idx(i,:)));
end

end
```

B.2 – Seleção do número ótimo de vizinhos

A Figura B. 1, apresentada a análise feita para auxiliar a escolha do número ótimos de vizinhos para o conjunto de dados em análise. Através da análise gráfica optou-se por utilizar *k* igual a 10. Apesar de o valor selecionado não ser exatamente o "cotovelo" da curva permitiu obter bons resultados sem "alisar" demasiado o resultado da previsão.

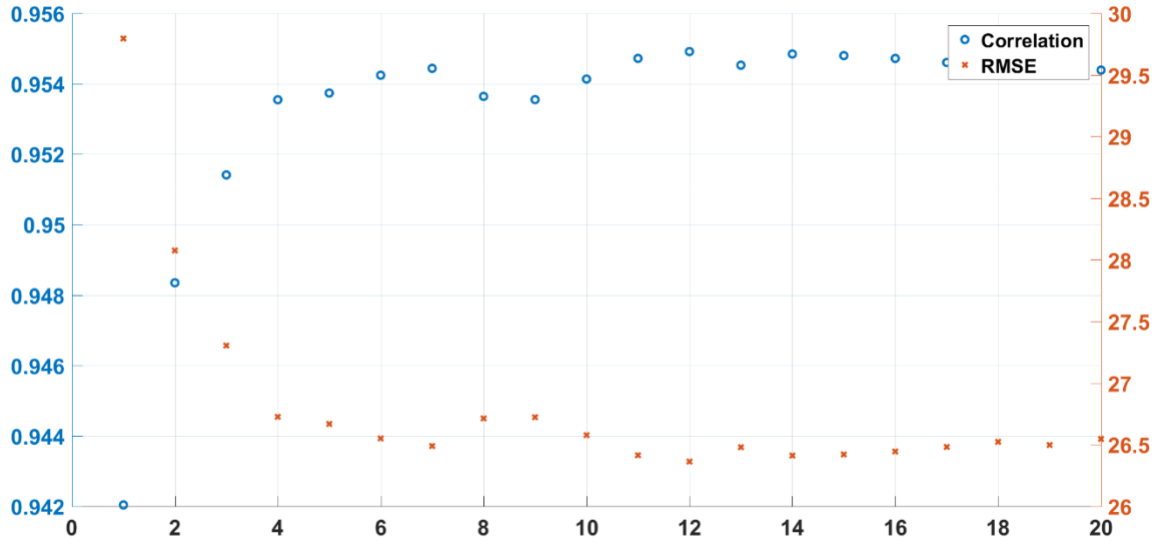

Figura B. 1: Análise de sensibilidade ao valor k feita para o método k vizinhos mais próximos. No eixo das abcissas é apresentado o número de k testados.

Anexo C - Método ANN

C.1 – Implementação do algoritmo em MATLAB

Código da função criada para gerar a Rede Neuronal Artificial, com os parâmetros utilizados:

```
function [ypred, NN, tr] = ANN(Xtrain, ytrain, Xtest, net_size)

% netsize can be a scalar or an array

net = feedforwardnet(net_size);
net.divideFcn = 'dividerand';
net.trainParam.epochs = 1500;
net.trainParam.goal = 0;
% net.performParam.regularization = 0.01; % não aplicável na
%arquitetura utilizada
net.trainFcn = 'trainlm';
% net.trainParam.lr = 0.01; % não aplicável na arquitetura utilizada

for i = 1:length(net.layers)

    if i == length(net.layers) % camada de output com função linear
        net.layers{i}.transferFcn = 'purelin';
    else
        net.layers{i}.transferFcn = 'tansig'; % hidden layers têm
    end                                       % tangents hiperbólicas
                                              % como função de
ativação
end

[NN, tr] = train(net,Xtrain',ytrain');

ypred = NN(Xtest')';

end
```

C.2 – Seleção da configuração mais adequada

Na Figura C. 1, apresentada a arquitetura da rede neuronal artificial obtida após vários testes de sensibilidade. A arquitetura consiste em 13 neurónios na camada de entrada, 10 na camada escondida e um na camada de saída. A função de ativação na camada escondida é uma tangente hiperbólica. Na camada de saída a melhor performance foi obtida recorrendo a uma função de ativação linear.

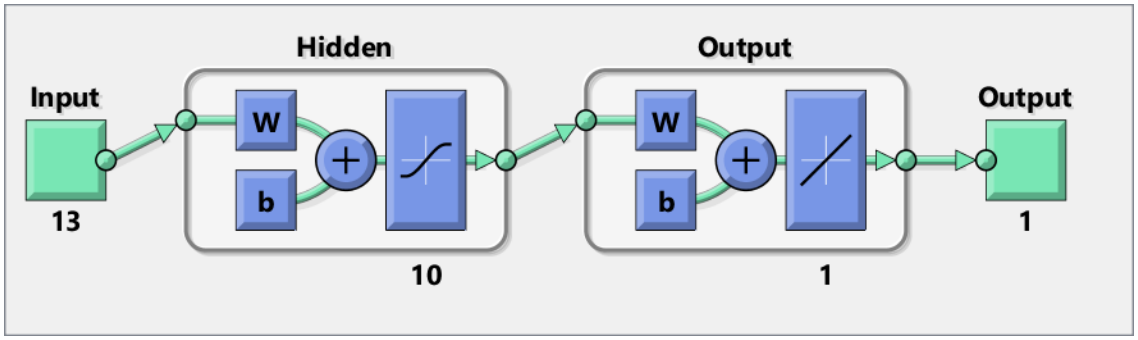

Figura C. 1: Arquitetura da Rede Neuronal Artificial utilizada

Anexo D – Teste t aos coeficientes do modelo de MLR

Na Tabela D. 1 apresenta-se os parâmetros estatísticos:

Tabela D. 1: Teste t aos parâmetros individuais do modelo

	Coeficiente	Soma de Erros	Estatística de teste *t*	p-value
Intercept	-1,5895	0,0023	-696,7642	0,0000
Dia_Util	0,0127	0,0007	17,3566	0,0000
Hora	0,0006	0,0001	8,6208	0,0000
Dia_da_Semana	0,0025	0,0002	14,4854	0,0000
Mes	0,0021	0,0001	18,3161	0,0000
h-24	0,0006	0,0000	126,7578	0,0000
h-25	0,0000	0,0000	-1,8992	0,0575
h-72	0,0000	0,0000	-2,8887	0,0039
h-120	0,0001	0,0000	18,7457	0,0000
h-168	0,0000	0,0000	-6,2099	0,0000
h-192	0,0001	0,0000	15,9751	0,0000
Velocidade do Vento	0,0005	0,0000	115,2578	0,0000
Temperatura de Bolbo Seco	-0,0003	0,0000	-84,2690	0,0000
Temperatura de Ponto de Orvalho	-0,0063	0,0014	-4,5863	0,0000
Nebulosidade	0,0101	0,0027	3,8039	0,0001
Pressao Atmosferica	-0,0286	0,0023	-12,6347	0,0000
Cluster_Index	-0,0057	0,0004	-13,3086	0,0000

Na Tabela D. 1 verifica-se que duas variáveis poderão não ser relevantes para o modelo: *h-25* e *Cluster Index*. Todos os restantes parâmetros são considerados relevantes, a um nível de significância de 5% (pValue < 0,05 para todos eles). O coeficiente de determinação (R_2) do modelo é de 0,9675.

Anexo E - Gráficos adicionais da previsão

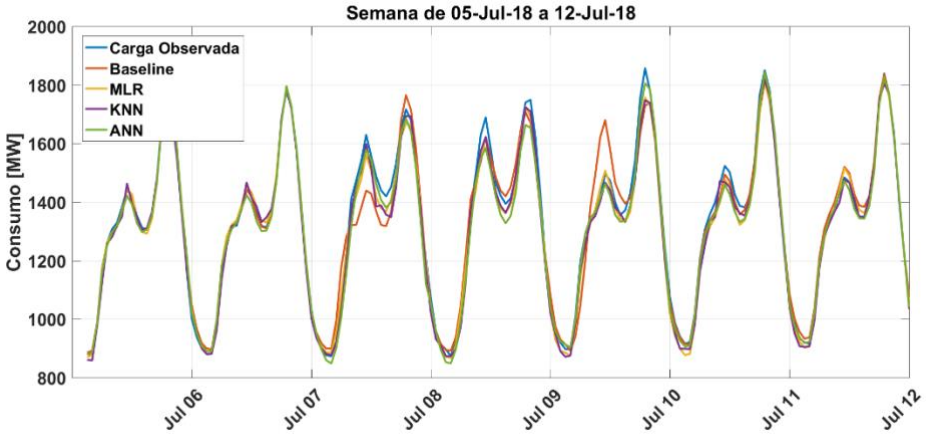

Figura E. 1: Ajuste dos modelos, semana de 5 de julho

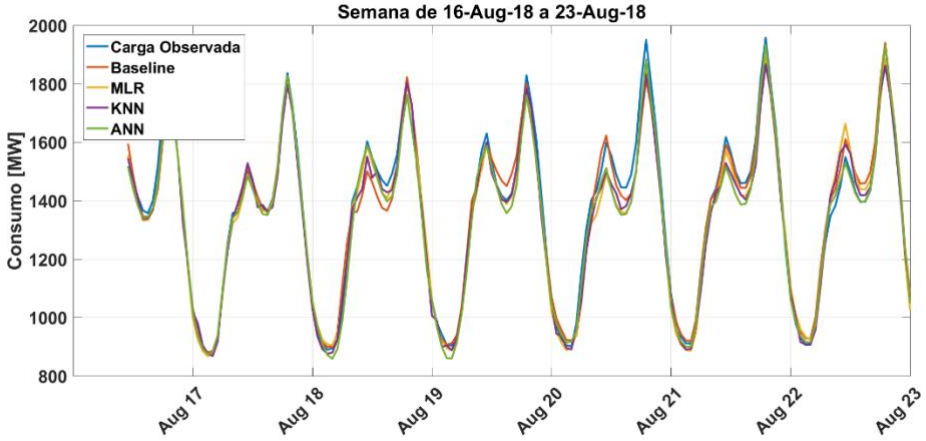

Figura E. 2: Ajuste dos modelos, semana de 16 de agosto

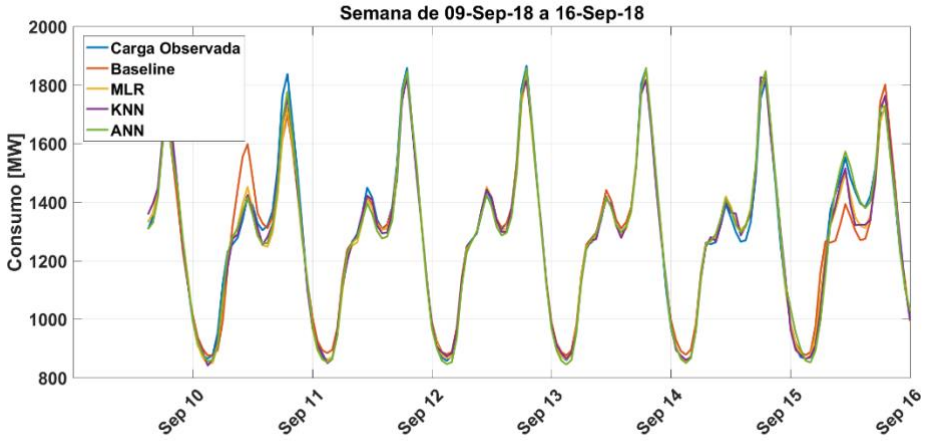

Figura E. 3: Ajuste dos modelos, semana de 16 de agosto

Anexo F - Desvios horários por cluster

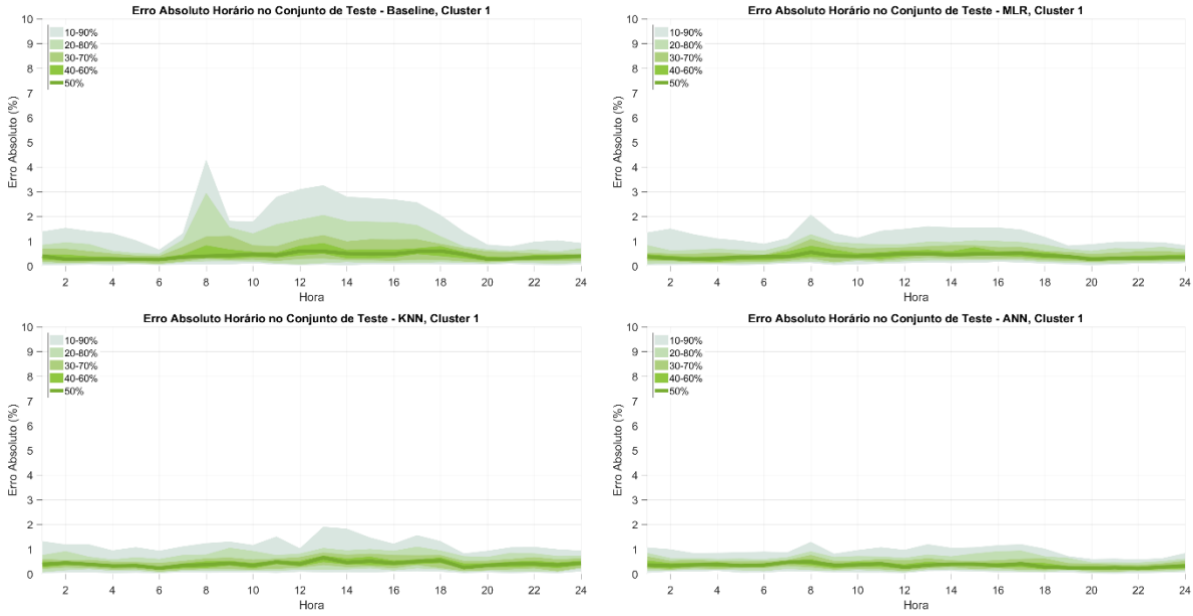

Figura F. 1: Erros horário médio, desagregado para o cluster 1, para todos os métodos

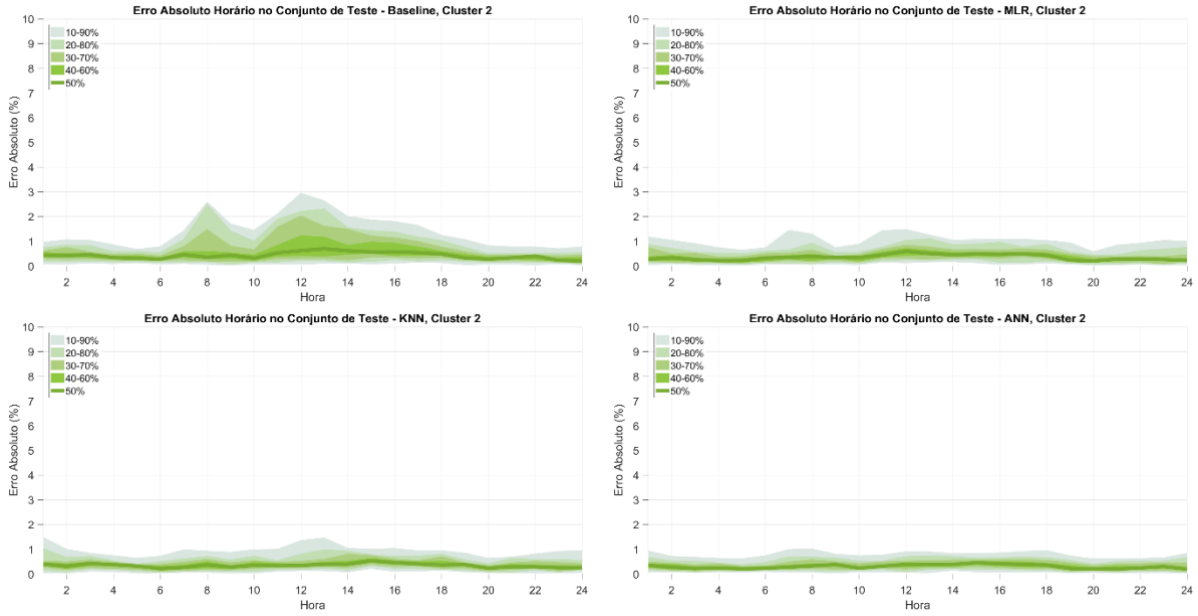

Figura F. 2: Erros horário médio, desagregado para o cluster 2, para todos os métodos

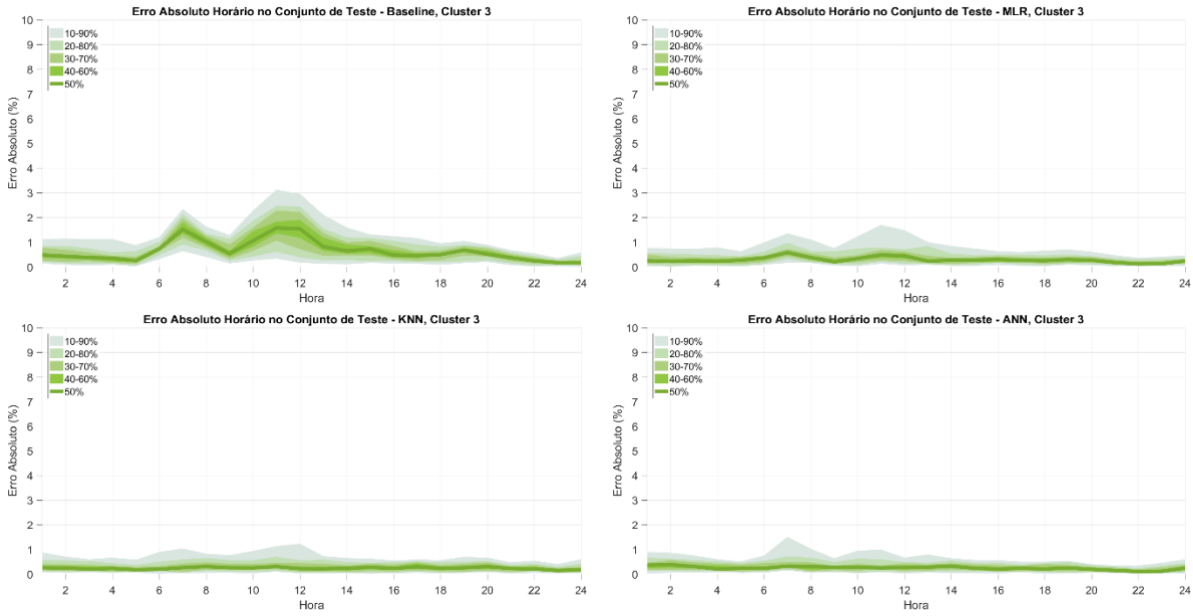

Figura F. 4: Erros horário médio, desagregado para o cluster 3, para todos os métodos

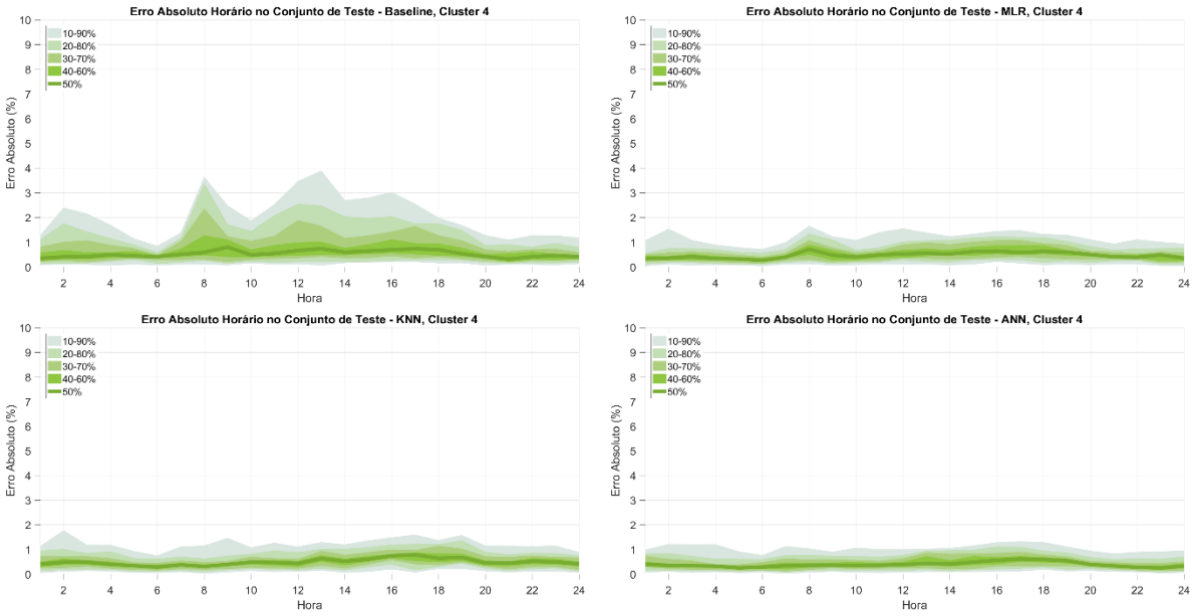

Figura F. 3: Erros horário médio, desagregado para o cluster 4, para todos os métodos

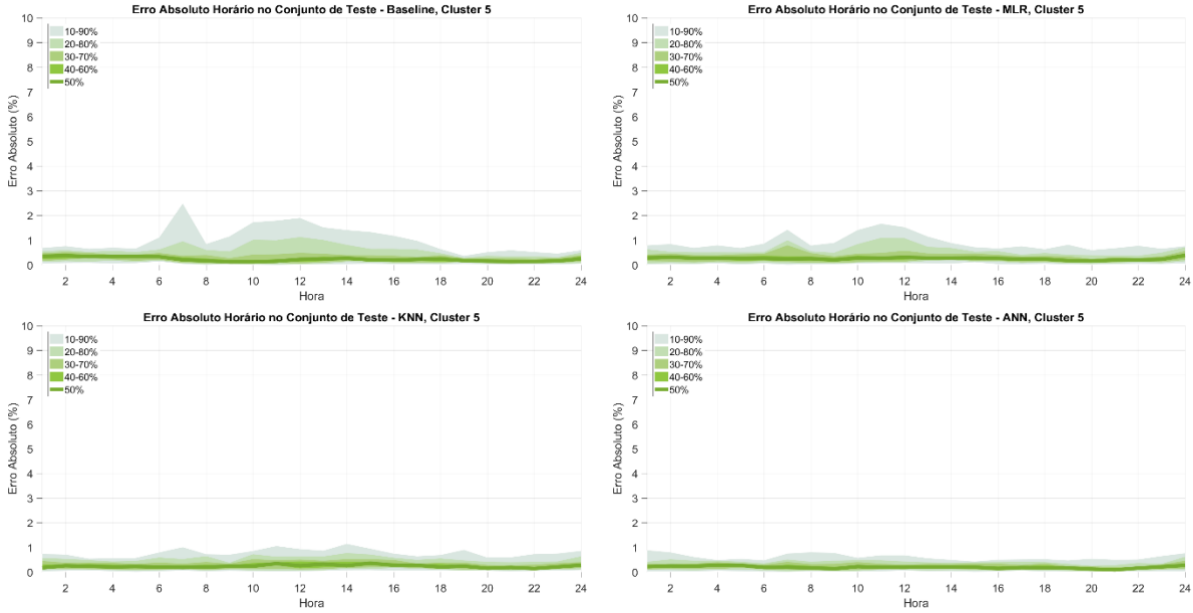

Figura F. 5: Erros horário médio, desagregado para o cluster 5, para todos os métodos

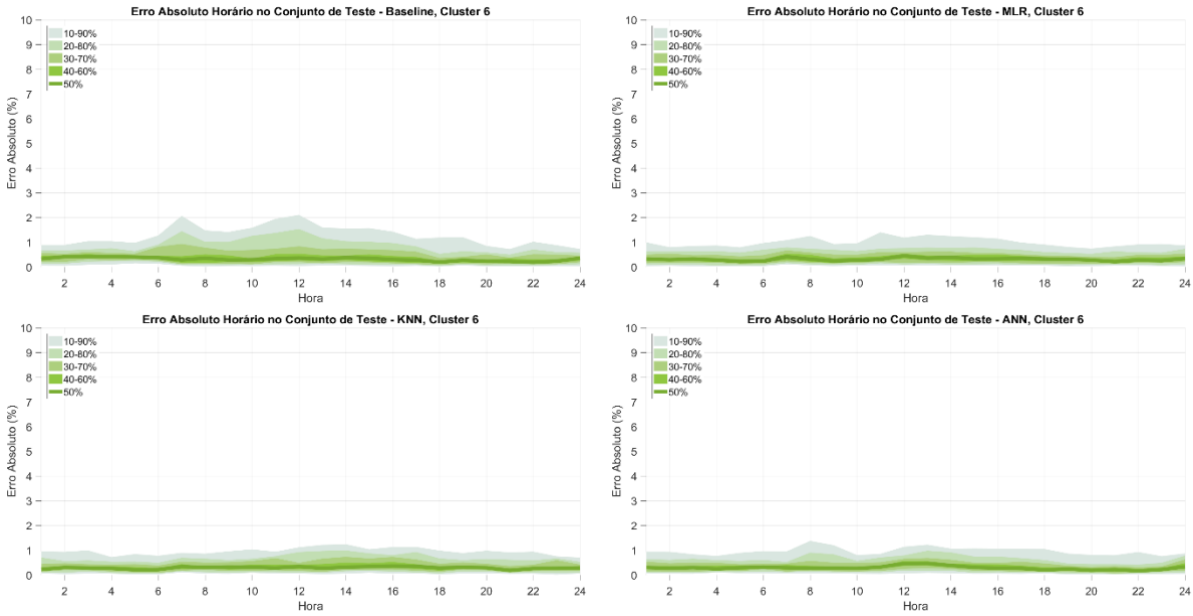

Figura F. 6: Erros horário médio, desagregado para o cluster 6, para todos os métodos

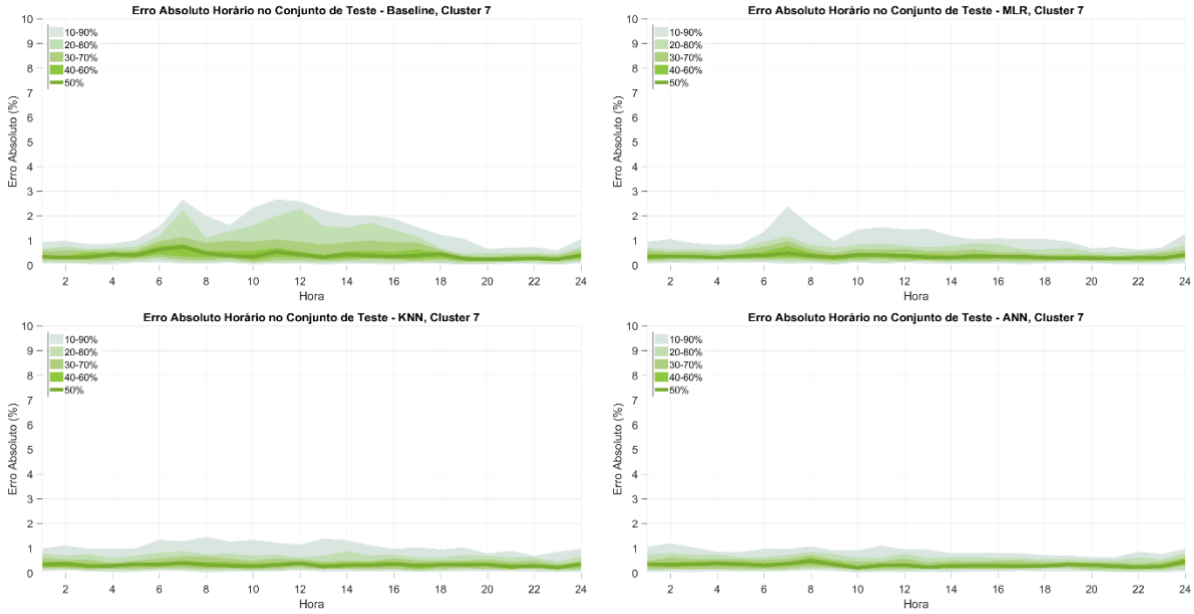

Figura F. 7: Erros horário médio, desagregado para o cluster 7, para todos os métodos