Sérgio Filipe Martins Casca

Licenciado em Engenharia Informática

Perfis e Estimativa de Consumo de Serviços em Estabelecimentos Hoteleiros

Dissertação para obtenção do Grau de
Mestre em Engenharia Informática

Orientador: João Moura Pires, Prof. Auxiliar,
Universidade Nova de Lisboa

	Júri:
Presidente:	Prof. Nuno Preguiça
Arguente:	Prof.ª Maribel Alves Santos
Vogal:	Prof. João Moura Pires

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Agosto, 2015

Perfis e Estimativa de Consumo de Serviços em Estabelecimentos Hoteleiros

À minha família

AGRADECIMENTOS

Ao Professor João Moura Pires, por todo o apoio e aconselhamento.

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."
George E. P. Box, Box e Draper (1987)

RESUMO

Desde meados de 1960 que a atividade hoteleira tem vindo a ganhar relevância no sector económico. O consumo de recursos no exercício da atividade hoteleira é hoje em dia encarado com uma perspetiva diferente, principalmente devido ao crescente peso relativo dos custos de operação durante as últimas décadas e ao sucessivo aumento de taxas e tributações de consumo por parte de fornecedores e autoridades governamentais.

O projeto Soltgest é uma solução de gestão de eficiência de consumo de serviços tais como a água, o gás e a eletricidade, dirigido ao sector hoteleiro. Desenvolvida pelo grupo empresarial Altran e em funcionamento desde 2010 em onze hotéis localizados na Região Autónoma da Madeira, a solução é composta por uma plataforma aplicacional e por sistemas de sensores instalados em estabelecimentos hoteleiros, destinados ao registo de consumo de serviços. O projeto foi ainda enriquecido com a informação das condições meteorológicas no local através de estações meteorológicas instaladas em algumas unidades hoteleiras. As leituras dos sensores de consumo de serviços e das estações meteorológicas são registadas a um ritmo horário e enviadas para a plataforma Soltgest que por sua vez possibilita a monitorização das mesmas. A plataforma Soltgest permite ao utilizador a possibilidade de indicar o número de hóspedes presentes na unidade hoteleira e os respetivos quartos ocupados e de comparar o consumo de serviços registado com a ocupação presente. A análise da informação recolhida é no entanto limitada a sumários estatísticos sobre diferentes perspetivas.

Nesta dissertação é apresentado o estudo da aplicação de técnicas de data mining aos dados recolhidos pela solução Soltgest, com o intuito de analisar a possível existência de padrões de consumo de serviços nas unidades hoteleiras. Foram aplicadas técnicas não-supervisionadas de agrupamentos para a identificação de perfis de consumo por unidade hoteleira e aplicadas técnicas supervisionadas de regressão com o intuito de estimar o consumo de cada serviço por estabelecimento.

Neste estudo foi possível identificar um perfil de consumo de serviços comum às unidades hoteleiras analisadas e gerados modelos preditivos capazes de estimar o consumo de serviços com um erro reduzido.

Abstract

Since the mid-1960s the hotel business is gaining importance in the economic business sector. The utilities' consumptions in the practice of hotel business is nowadays seen with a different perspective, mainly due to the increasing operating costs over the past decades and to the increasing rates and taxes applied by the suppliers and government authorities.

The Soltgest project is an utilities' consumption efficiency management solution aimed at the hotel sector. The solution was developed by Altran business group and is operating in eleven hotels located at the Autonomous Region of Madeira since 2010. The solution comprises an application platform and sensor systems installed at the hotels' establishments for measuring utilities' consumptions. The project was further enriched with local weather conditions data through weather stations installed at several hotels. The utilities' consumptions sensors and the weather stations report at an hourly rate and send the data gathered to the Soltgest platform. The Soltgest platform allows the user to specify the number of guests present at the hotel establishment along with the respective number of rooms occupied for comparison against the utilities consumptions. The analysis of the collected information is however limited to statistical summaries in different perspectives.

This dissertation presents the study of data mining techniques with the data collected by the Soltgest solution in order to identify possible utilities' consumption patterns. Unsupervised learning techniques were applied to profile each hotel establishment given their utilities' consumptions and supervised learning techniques were applied to estimate the utilities' consumptions in each hotel unit.

In this study it was possible to identify a common utilities' consumptions profile between hotel establishments and predictive models capable of estimating utilities' consumption with a small error rate.

Conteúdo

Lista de Figuras

LISTA DE TABELAS

GLOSSÁRIO

ACIF-CCIM Associação Comercial e Industrial do Funchal - Câmara do Comércio e Indústria da Madeira.

AVAC Aquecimento, Ventilação e Ar Condicionado.

EEM Empresa de Eletricidade da Madeira.

IGEO Informação Geográfica, Governo de Portugal.

RAM Região Autónoma da Madeira, Portugal.

SGIDI Sistema de Gestão de Investigação, Desenvolvimento e Inovação.

CAPÍTULO 1

INTRODUÇÃO

A atividade turística tem vindo a ganhar relevância no sector económico desde o início dos anos 60 tendo promovido o desenvolvimento e consequente aumento de qualidade de vida, Nissan et al. (2011).

Como em todas as atividades económicas, a gestão eficiente das despesas correntes é um dos pilares fundamentais do sucesso e uma preocupação constante e presente dos empresários. Por exemplo, o programa *Energy Star* incentiva e ajuda as empresas e os empresários a definir uma gestão mais eficiente do consumo energético, promovendo a redução do consumo através de medidas inteligentes e ajustadas ao negócio, resultando em benefícios financeiros. Em 2007 foram realizados dois estudos pelo programa *Energy Star*, EnergyStar (2007a); EnergyStar (2007b) exclusivamente direcionados ao sector turístico, onde reportaram que as despesas relacionadas unicamente com o consumo energético representavam mais de 6% do total das custos de operação nas unidades hoteleiras. O serviço foi igualmente considerado como o custo de operação com maior crescimento nos últimos anos.

Com a sociedade turística a exigir constantemente novos padrões mínimos de qualidade nos estabelecimentos hoteleiros e constante subida dos tarifários nos serviços de eletricidade, água e gás, acrescidos de normas regulamentadoras instauradas por agências governamentais com fim à proteção e preservação do meio ambiente, a gestão eficiente do consumo de serviços no sector hoteleiro está a ganhar preponderância. Os gestores das unidades hoteleiras tem vindo a adotar distintas práticas, consideradas recentes, visando economizar o consumo de serviços. Entre as práticas mais populares destacam-se, por exemplo, as ações de formação para a sensibilização ao consumo de serviços por parte dos funcionários e a colocação de cartões de boas-vindas nos quartos dos estabelecimentos hoteleiros com mensagens apelativas à redução do consumo de serviços em prol do meio ambiente. No entanto, uma das práticas mais recentes e com maior impacto na otimização

da gestão energética nas unidades hoteleiras assenta na adoção de soluções que promovem a instalação de sistemas de sensores para monitorizar o consumos de serviços. Um grupo de unidades hoteleiras na Região Autónoma da Madeira adotou uma solução de monitorização de consumos denominada Soltgest.

1.1 Projeto Soltgest

Direcionado exclusivamente para o sector hoteleiro, o projeto Soltgest é uma solução que visa otimizar a eficiência da gestão de consumo de serviços em estabelecimentos hoteleiros ao monitorizar e reportar o consumo de serviços. Promovido pela ACIF-CCIM (Associação Comercial e Industrial do Funchal - Câmara do Comércio e Indústria da Madeira) e co-financiado pelo Mecanismo Financeiro do Espaço Económico Europeu através do fundo *EEA Grants - Iceland Liechtenstein Norway 2009-2014* e pelo Governo da RAM (Região Autónoma da Madeira), o projeto Soltgest foi executado pelo grupo *Altran* em parceria com a companhia ALH Consultores.

O projeto Soltgest apresenta-se como uma solução para a gestão eficiente do consumo de serviços no sector hoteleiro. O principal contributo do projeto é proporcionar uma gestão mais eficiente, através da monitorização de consumos, que consequentemente leve a uma redução dos custos de operação. É uma solução em curso desde 2010 em onze unidades hoteleiras localizadas na RAM. A arquitetura do projeto assenta em duas camadas: uma camada de sensorial, onde sistemas de sensores foram instalados nos estabelecimentos de hotelaria a fim de registar os consumos de eletricidade, água e gás; uma camada aplicacional, que disponibiliza uma plataforma que monitoriza e reporta as leituras dos sistemas de sensores e os respetivos custos financeiros. A filial holandesa do grupo *Altran* executou a camada aplicacional e a companhia parceira ALH Consultores a instalação da camada sensorial.

A informação recolhida pela plataforma Soltgest, proveniente a um ritmo horário dos sistemas de sensores instalados nos estabelecimentos hoteleiros, provém de dois tipos de sistemas distintos: sensores de consumo de serviços, responsáveis pela medição dos consumos de eletricidade, água e gás; estações meteorológicas, responsáveis por registar as condições meteorológicas no local. Os sensores de consumo de serviços foram instalados em setores comuns dos estabelecimentos, como em restaurantes e *spa's*, e em unidades específicas mas gerais a toda a unidade hoteleira como em sistemas AVAC (Aquecimento, Ventilação e Ar Condicionado), caldeiras, refrigeradores, entre outras.

Por sua vez, a plataforma Soltgest permite monitorizar os registos dos consumos de cada serviço em cada unidade hoteleira. Ao serviço de monitorização foi providenciado aos gestores das unidades hoteleiras a possibilidade de adicionar o número de hóspedes presentes e o respetivo número de quartos de hóspedes ocupados por dia na unidade hoteleira. Além da monitorização, a plataforma permite a elaboração de relatórios com sumários estatísticos da informação recolhida.

Em 2013, a filial portuguesa do grupo *Altran* adquiriu interesse na aplicação Soltgest via o departamento de investigação SGIDI (Sistema de Gestão de Investigação, Desenvolvimento e Inovação). A dimensão do sector empresarial hoteleiro, a escassez de soluções de gestão de eficiência de consumos de serviços dedicados à atividade e o volume e a heterogeneidade de informação já recolhida pela plataforma atraiu o grupo português que possuía uma visão inovadora para o projeto. A visão do grupo visava aplicar técnicas de *data mining* aos dados recolhidos pela plataforma Soltgest a fim de obter, integrar e adicionar mais valor à informação reportada pela aplicação.

1.2 Motivação

Os avanços tecnológicos ocorridos durante a última década têm favorecido os sistemas sensoriais ao permitir a recolha de maiores volumes de informação, de diversas fontes, em formatos distintos e a ritmos cada vez mais intensos. Ao mesmo tempo, surge igualmente a possibilidade de processar e analisar em períodos de tempo úteis os elevados volumes de dados recolhidos. Por consequência, a área de científica de *data mining*, que permite identificar e validar possíveis padrões através de análises analíticas a grandes quantidades de dados, tem vindo a ganhar foco no mundo empresarial.

O domínio do projeto é uma área de investigação de interesse. Além da importância que atividade turística possui como atividade económica, os estudos relacionados com o estudo de consumos de serviços assentam, na sua generalidade, na perspetiva do fornecedor do serviço. Por sua vez, o projeto Soltgest promove o estudo perante a perspetiva do consumidor num contexto económico-social específico onde relativamente poucos estudos foram realizados.

Por último, os possíveis contributos que o estudo pode oferecer à gestão de recursos e à otimização do consumo de serviços promovem diretamente benefícios à proteção e preservação do meio ambiente.

1.3 Descrição do Problema

Desde meados de 2010 que a plataforma Soltgest reúne a um ritmo horário as leituras registadas pelos sistemas de sensores, de consumo de serviços e de estações meteorológicas, em cada unidade hoteleira. A adesão desde pronto à funcionalidade de registo dos valores de ocupação na plataforma Soltgest permitiu adicionar aos recursos disponíveis a informação relativa à ocupação nas unidades hoteleiras.

Dado o volume e a heterogeneidade dos dados recolhidos é proposto por este estudo a aplicação de técnicas de *data mining* com o intuito de verificar a existência de padrões de consumo de serviços. Por cada estabelecimento hoteleiro pretende-se identificar perfis de consumo de serviços e elaborar modelos capazes de estimar os respetivos consumos de eletricidade, água e gás.

1.4 Abordagem

Para o desenvolvimento do estudo foi adotada a metodologia *CRISP-DM*, *Cross-Industry Standard Process for Data Mining*, uma das metodologias mais utilizadas para conduzir projetos de *data mining*, direcionada e recomendada para projetos empresariais, Azevedo (2008); Kurgan e Musilek (2006). Composta por seis etapas principais distintas, o ciclo da metodologia é apresentado na Figura 1.1 (IBM (2012), p.4).

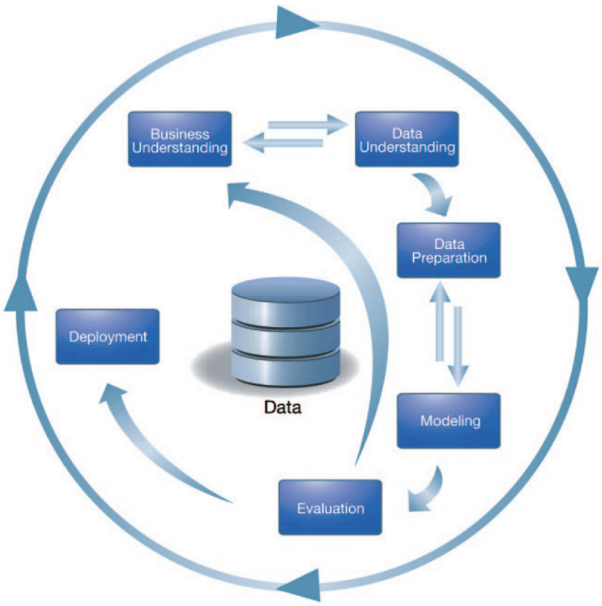


Figura 1.1: Ciclo da metodologia CRISP-DM (Source: IBM, 2012)

Na Figura 1.1 os dados são o centro do estudo e ao seu redor estão definidas as seis etapas que compõem a metodologia. As setas entre as etapas representam o trajeto mais usual que um projeto de *data mining* adota. No entanto, a metodologia defende que o trajeto do estudo seja flexível e que deva existir sempre a possibilidade de o projeto retornar ou avançar para uma qualquer etapa. A ideologia é reforçada pelo círculo exterior que pretende também ilustrar que a metodologia não é um processo fechado e que é possível iterar ou alternar entre etapas se assim for benéfico e adequado ao projeto.

A etapa inicial da metodologia CRISP-DM é a etapa de *Business Understanding* dedicada à compreensão do modelo de negócio do projeto, do domínio ou contexto em que o projeto está definido, dos objetivos pretendidos perante a perspetiva empresarial e, por fim, à definição de metas ou possíveis contributos ao projeto em questão.

Após todo o envolvimento ao projeto segue uma análise preliminar dos recursos disponíveis, a etapa *Data Understanding*. Esta etapa implica: a recolha de todos os conjuntos dados; a realização de análises discriminativas do volume de dados e das suas características; a realização de análises exploratórias; o registo de ocorrências relevantes; a caracterização da qualidade dos dados; a elaboração de um documento de qualidade de dados de forma a sumarizar as análises efetuadas.

A etapa *Data Preparation* é usualmente a etapa mais morosa em todo o processo, atingindo por vezes cerca de 50% ou até mesmo 70% do tempo e do esforço de trabalho de um projeto. Nesta etapa são elaborados e definidos o(s) conjunto(s) de dados que são utilizados para modelar os algoritmos de aprendizagem. A etapa compreende funções tais como a definição de subconjuntos de dados, a fusão de conjuntos de dados, a definição de novos atributos, classificação de dados, o tratamento de valores considerados anormais e a separação do conjunto de dado em sub-conjuntos para efeitos de modelação, validação e teste.

Durante a etapa de *Modelling* são aplicadas diversas técnicas de aprendizagem ao(s) conjunto(s) de dados obtidos durante a etapa de *Data Preparation*. Ao longo desta etapa são também definidas as parametrizações das técnicas de aprendizagem que melhores resultados produzem. O trajeto cíclico entre a etapa de *Data Preparation* e *Modelling* é bastante comum pois a transformação, seleção ,exclusão e adição de variáveis, entre muitas outras operações, são práticas executadas muitas vezes em paralelo com a modelação dos algoritmos de aprendizagem.

A etapa *Evaluation* analisa o trajeto e as opções tomadas durante as etapas anteriores e avalia os resultados obtidos perante os objetivos inicialmente propostos. Esta etapa determina se o estudo realizado cumpre os requisitos inicialmente propostos. No caso de não respeitar os objetivos inicialmente traçados são analisadas as opções de prosseguir o estudo retornando a etapas anteriores da metodologia ou de concluir o estudo com os resultados obtidos.

A etapa *Implementation* é a último etapa da metodologia e tem o propósito de encerrar o estudo realizado. Esta etapa inclui, usualmente, a função de produzir um relatório final reportando todos os processos realizados durante o estudo. No entanto, pode incluir também a função de implementar e integrar os protótipos produzidos durante o estudo do projeto para efeitos de produção, IBM (2011).

1.5 Contribuições

A informação disponibilizada pela plataforma Soltgest inclui os registos dos sistemas de sensores e os registos de ocupação em cada estabelecimento hoteleiro. Como tal, com focos no aumento da eficiência na gestão do consumo de serviços e por consequência na redução dos custos de operação, as principais contribuições do estudo desenvolvido durante a dissertação foram:

- Análise da qualidade dos dados recolhidos pela plataforma Soltgest
- Introdução de informação relevante, como os períodos de ocorrência de eventos sociais e culturais influentes na atividade turística
- Introdução de informação relativa às condições meteorológicas no local provenientes de fontes externas

- Elaboração de perfis de consumo de serviços por hóspede em cada unidade hoteleira
- Geração de modelos para a estimativa do consumo de cada serviço por hóspede e em cada unidade hoteleira

1.6 Estrutura do Documento

O primeiro capítulo é o capítulo introdutório da dissertação. Uma descrição inicial do projeto Soltgest e o seu interesse como investigação científica são apresentados nas Secções 1.1 e 1.2. Na Secção 1.3 é apresentada uma descrição sucinta do problema proposto. A abordagem ao problema e as contribuições resultantes do estudo de investigação são apresentadas nas Secções 1.4 e 1.5 respetivamente.

O documento foi estruturado com base na metodologia CRISP-DM seguida durante o estudo analítico desta dissertação. O Capítulo 2 segue a etapa *Business understanding* e introduz o contexto do projeto Soltgest. No Capítulo 3, etapa de *Data Understanding*, é apresentada a informação disponível e realizada a análise exploratória aos dados. O Capítulo 4 apresenta o processo de preparação dos conjuntos de dados para análise, etapa *Data Preparation*. Os Capítulo 5 e 6 foram destinados, respetivamente, ao estudo da identificação de perfis de consumo de serviços em cada unidade hoteleira e ao estudo da análise preditiva de consumo de cada serviço por unidade hoteleira. Ambos os Capítulos 5 e 6 englobam duas etapas da metodologia CRISP-DM, a etapa de *Modelling* e a etapa de *Evaluation*, dado que em cada um são apresentados os processos de modelação, validação e interpretação dos algoritmos estudados. Nesta dissertação, a secção relativa ao estado-de-arte e à apresentação de algoritmos e técnicas de data mining, que é usualmente apresentada antes do estudo analítico, foi repartida por objetivo proposto e apresentada respetivamente nas primeiras secções dos Capítulos 5 e 6. Ao apresentar a secção relativa aos algoritmos e às técnicas estudadas apenas no início dos Capítulos 5 e 6 a relação entre a secção teórica e prática tornou-se mais clara e evidente. O Capitulo 7 apresenta a conclusão do estudo analítico e uma opinião sobre possíveis trabalhos futuros.

CAPÍTULO 2

INTERPRETAÇÃO DO PROJETO

A etapa inicial da metodologia CRISP-DM é a etapa de *Business Understanding* dedicada à compreensão do modelo de negócio do projeto e do contexto em que se insere.

Neste capítulo é apresentado o contexto em que projeto Soltgest foi desenvolvido nas primeiras cinco secções. Na sexta e última secção é então apresentado em maior detalhe a estrutura do projeto Soltgest, a arquitetura do sistema, a plataforma *web* e uma análise de reconhecimento de mercado de soluções concorrentes.

2.1 Contexto Geográfico e Demográfico

O projeto Soltgest foi executado na RAM, um arquipélago localizado no Oceano Atlântico a cerca de mil quilómetros de Portugal Continental, Figura 2.1. O arquipélago é composto por dois sub-arquipélagos, o sub-arquipélago da Madeira e o sub-arquipélago das ilhas desertas. O sub-arquipélago da Madeira inclui as ilha da Madeira, de Porto Santo, e cinco pequenas ilhas não habitadas enquanto que o sub-arquipélago das Ilhas Desertas inclui dois grupos de ilhas também não habitadas. A RAM possui perto de 801,12km^2 de área terrestre e a ilha da Madeira é a maior das ilhas do arquipélago, totalizando perto de 92% de toda a área de terra do arquipélago, aproximadamente 736.7km^2 e com um comprimento máximo de 57km de largura. A ilha de Porto Santo encontra-se a nordeste da ilha da Madeira, a aproximadamente 39km de distância e possui uma área terrestre de 42.2km^2. A capital da região autónoma é a cidade do Funchal localizada na região sul da ilha da Madeira, com as coordenadas 32^o 39′ N 16^o 55′O.

Em 2011, a RAM possuía 267 mil habitantes, 333.65 habitantes por km^2, e a sua capital, a cidade de Funchal, acomodava praticamente 40% de toda a população. A população da região é caracterizada por ser uma população "madura"onde os jovens com menos de 25 anos correspondem menos de 30% do total da população enquanto que 13% da população

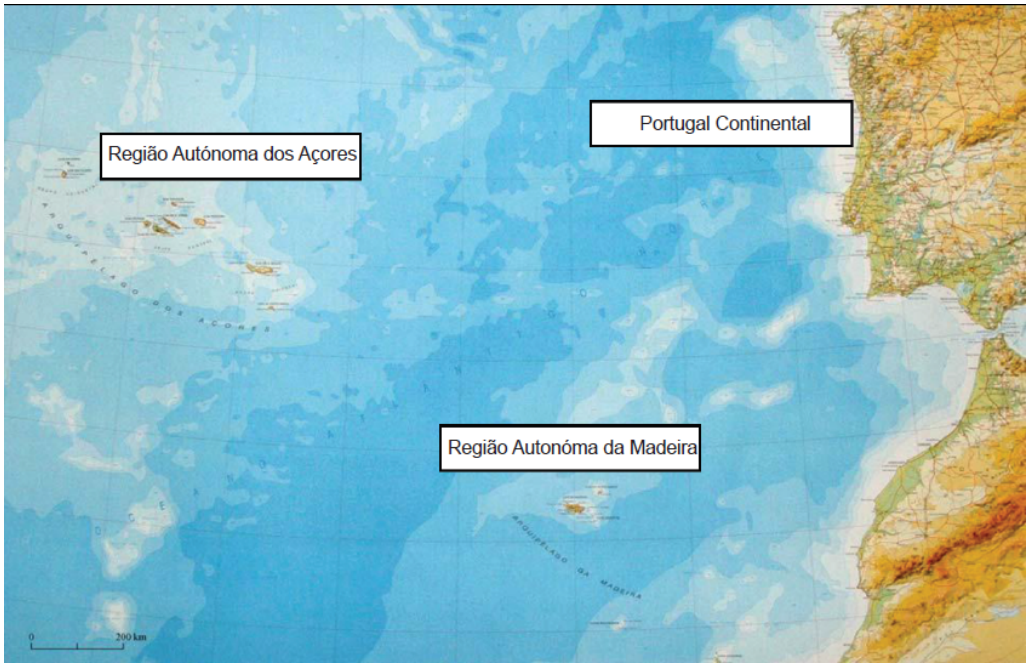
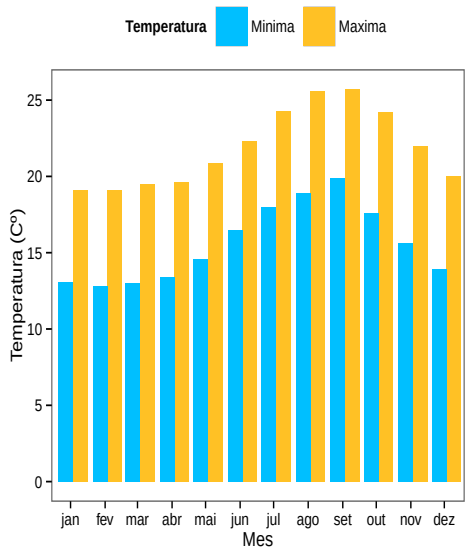


Figura 2.1: Localização Geográfica da RAM, Fonte: IGEO, 2003

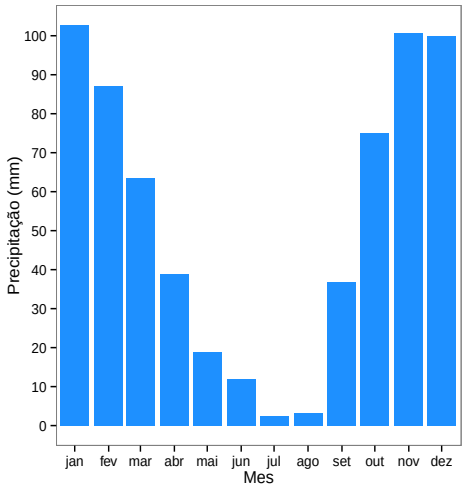
corresponde a idosos com mais de 65 anos de idade. O grupo etário adulto, entre os 25 e os 65 anos de idade, engloba 57% da população, Governo da Madeira (2012b).

2.2 Contexto Climático

O clima da região é caracterizado por um clima Mediterrâneo perante o sistema de classificação climática *Köppen*, "Observed and projected climate shifts 19012100 depicted by world maps of the Koppen-Geiger climate classification", mais especificamente, por um *"clima subtropical com verão seco"*.A definição de *"verão seco"* é caracterizado pela presença de pelo menos um mês de Verão com menos de um terço de precipitação ocorrida durante o mês mais húmido de Inverno e com um valor total de precipitação menor que 30 mm. O clima do arquipélago é também influenciado pela "Corrente do Golfo", provocando temperaturas amenas e ambientes mais húmidos durante todo o ano. A ilha da Madeira possui uma cordilheira central que divide a zona Norte e Sul da ilha e causa diferenças climáticas entre as zonas. A elevada altitude da cordilheira, entre outros fatores, provoca o abrandamento da deslocação das nuvens para a zona Sul promovendo uma maior concentração da precipitação da região na zona Norte e e centro. Com o intuito de ilustrar sucintamente as condições meteorológicas na ilha da Madeira é apresentado na Figura 2.2 os valores médios de temperatura e precipitação por mês ao longo de 30 anos, mais precisamente entre 1961 e 1990, fonte World Meteorological Organization (2015).



(a) Temperatura média por mês na ilha da Madeira



(b) Precipitação média por mês na ilha da Madeira

Figura 2.2: Sumário estatístico das condições meteorológicas na ilha da Madeira entre 1961 e 1990

2.3 Contexto Turístico

A atividade turística tem sido uma atividade economicamente próspera a nível mundial. Em 2012 foi alcançado um novo recorde mundial de viajantes onde foram registados perto de mil milhões de viajantes e um crescimento de 4% relativo ao ano de 2011, Turismo (2014). O continente Europeu tem acompanhado o crescimento da atividade turística a nível mundial com uma pequena variância percentual. A região Oeste da Europa onde Portugal e por consequência a RAM estão incluídos, registou em 2010 o maior índice de atividade turística por cada 100 habitantes e foi a região que apresentou a melhor estimativa de atividade turística por cada 100 habitantes para 2030, com uma previsão de taxa de crescimento a rondar os 41%, United Nations World Tourism Organization (UNWTO) (2011). No entanto, a crise financeira mundial em 2008 afetou a maioria dos mercados económicos e atividades comerciais e houve um período de declínio da atividade turística a nível mundial, especialmente em 2009, como ilustrado na Figura 2.3. Apesar de a atividade turística internacional e nacional ter recuperado em 2010, a RAM foi fortemente afetada apresentando um declínio do número de turistas por dois anos consecutivos. Mesmo tendo recuperado rapidamente da crise económica, Portugal sofreu um revés no crescimento do número de turistas em 2012, tanto a nível nacional como regional, mais especificamente na RAM.

A Figura 2.4 apresenta a distribuição dos turistas na RAM, por ano e nacionalidade, desde 2007 a 2012. Perto de 98% dos turistas eram provenientes do continente Europeu e perto 70% exclusivamente dos países de Portugal, Reino Unido, Alemanha e França,

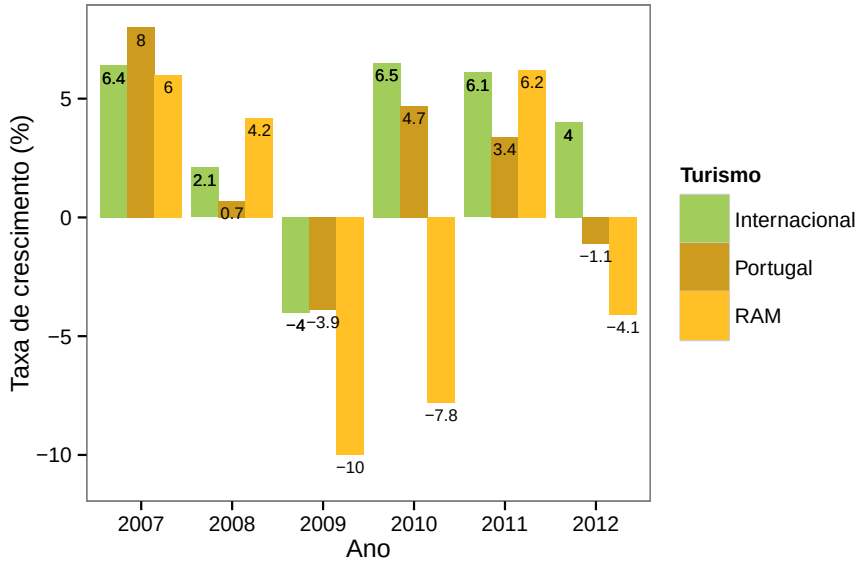


Figura 2.3: Taxa relativa de crescimento da atividade turística a nível internacional

Casca (2013). Apesar do decréscimo de turistas na RAM no ano de 2012 é de salientar o decréscimo do volume turístico proveniente de Portugal Continental desde 2010, um dos países com maior relevo no turismo da RAM. O decréscimo de turistas de Portugal continental é coincidente com os primeiros anos de crise financeira em que o país ainda hoje incorre. Em 2012, o volume de turistas provenientes de Portugal Continental e Reino Unido apresentaram o valor mais baixo desde 2007. No entanto países como a Alemanha, França e os restantes países do continente Europeu apresentaram uma tendência crescente desde 2010.

A RAM possui uma enorme atração turística devido à sua paisagem sendo igualmente conhecida pelos seus eventos sócio-culturais que influenciam diretamente a atividade turística. Certos eventos tais como o *Carnaval da Madeira*, a *Festa da Flor*, o *Campeonato Europeu de Rally*, o *Festival Summer Wine Festival*, a véspera e a noite de *Ano Novo*, entre outros, influenciam milhares de turistas a viajar para a região para participar nos eventos, Governo da Madeira (2011); Governo da Madeira (2012d). A atividade turística na RAM apresenta uma tendência sazonal relativa ao número de turistas presentes na ilha. Na Figura 2.5 é apresentada a média de turistas por mês, a cada período de 5 anos, entre 1977 e 2012. O volume de turistas na RAM, em especial nas últimas duas décadas, apresenta dois períodos anuais de maior afluência, nos meses de Março e Abril e no mês de Agosto, assim como uma tendência comum entre as diversas épocas.

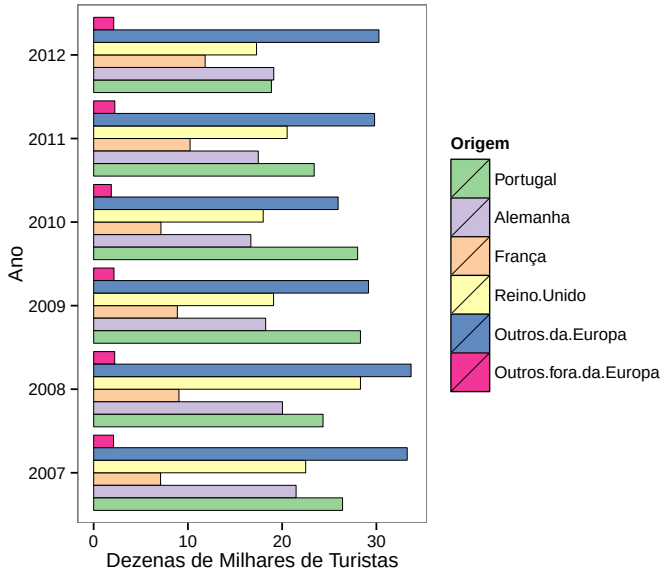


Figura 2.4: Distribuição dos turistas na RAM pela sua nacionalidade

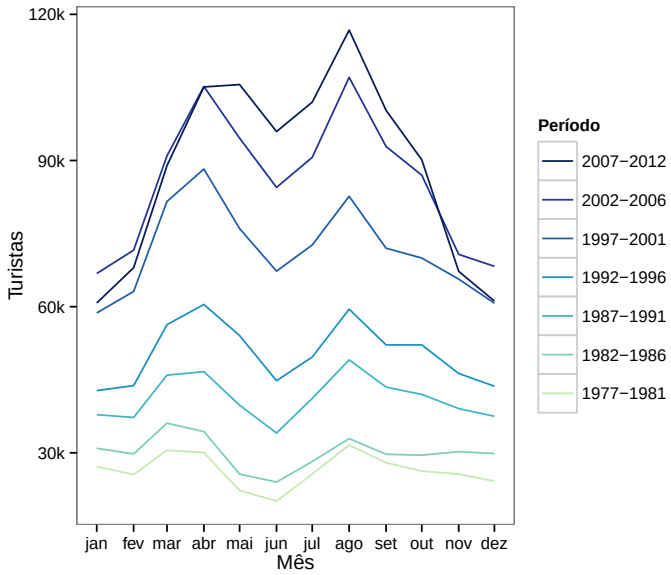


Figura 2.5: Distribuição dos turistas na RAM pela sua nacionalidade

2.4 Contexto Económico

A atividade turística tem um impacto direto na economia de uma região e diversos estudos tem sido elaborados procurando demonstrar a importância da atividade turística no crescimento económico. Em Leitao (2011) é apresentada uma correlação positiva entre o crescimento do turismo e o crescimento da economia em Portugal durante o período de 1995 a 2008 com base no número de turistas, no produto interno bruto, no investimento realizado no setor de restauração e no índice de preço de consumidor. Em Nissan et al. (2011) é obtida a mesma conclusão na grande maioria dos países europeus e nos Estados Unidos da América.

Em 2012 a contribuição direta da atividade turística no produto interno bruto em Portugal foi de 9.5%, World Travel & Tourism Council (2013). Na RAM o sector de turismo representou em 2010 cerca de 7.4% do valor acrescentado bruto na região, sustentado por mais de 2000 empresas, representando 10% do número total de postos de trabalho na região, Governo da Madeira (2012a); Governo da Madeira (2012c). Além do impacto económico do turismo na geração de capital e de emprego, a prosperidade da atividade tem tido também impacto a nível sócio-cultural. A título de exemplo, tem contribuído na conservação da herança cultural, onde o turismo tem sido um dos maiores estímulos para o efeito, na melhoria das condições de vida local dado que o turismo implica muitas vezes a melhoria das condições de saneamento básico da região, entre outros. O meio ambiente tem sido também indiretamente preservado através da conservação de áreas naturais importantes, de locais arqueológicos, históricos e da melhoria de certas infra-estruturas que usualmente levam à redução de problemas de poluição, Fonseca (2006).

O suporte à prosperidade da atividade turística tem também implicado um maior volume de despesas. No setor de hotelaria os custos de operações têm vindo a aumentar, especialmente no setor dos serviços de eletricidade, água e gás. Em Liu et al. (2012), apesar do estudo ter sido realizado numa realidade distinta dada a localização, foi relatado uma forte noção de causalidade preditiva entre o consumo energético e o número de turistas na região, através do teste de *Granger* que procura estimar valores de uma série temporal com base em outra(s) série(s) temporal.

A Empresa de Eletricidade da Madeira, EEM, é no momento presente, a única fornecedora do serviço de energia elétrica na RAM. A EEM tem como funções a produção, transporte e distribuição de energia elétrica. Desde 2008 que as tarifas aplicadas à energia elétrica na RAM têm tido um crescimento superior ao crescimento harmonizado do preço ao consumidor no setor de alojamento na região. De igual forma, o crescimento relativo do preço do gás têm vindo a aumentar mais do que o índice harmonizado dos preços ao consumidor. Por exemplo, em 2010, a tarifa da distribuição de gás propano canalizado aumentou 20 vezes mais do que o índice do preço harmonizado ao consumidor. As tarifas de consumo de água são, em todo o Portugal, estabelecidas em cada município e não foi possível identificar a evolução dos tarifários, Casca (2013).

2.5 Contexto Hoteleiro no projeto Solgest

Em 2009, onze dos 60 estabelecimentos hoteleiros de diferentes categorias na RAM fizeram parte do projeto Soltgest, Governo da Madeira (2012a), estando na sua generalidade localizados na zona oeste da ilha da Madeira. A Figura 2.6 apresenta as localizações geográficas dos onze estabelecimentos, numerados para efeitos de referência, que participaram no projeto Soltgest. Os estabelecimentos hoteleiros foram identificados a azul quando não possuem estações meteorológicas instaladas, índices 4 e 7 até 10, e identificados a verde aquando da presença de estações meteorológicas, restantes índices. É também apresentado na Figura 2.6 a localização geográfica de estações meteorológicas, independentes ao projeto e consideradas relevantes, que disponibilizam gratuitamente as leituras relativas às condições meteorológicas no local onde se encontram. As estações meteorológicas independentes foram consideradas relevantes com base na proximidade aos estabelecimentos hoteleiros e com base no volume de dados existente durante o período comum entre o histórico de dados das estações meteorologistas e os do projeto Soltgest.

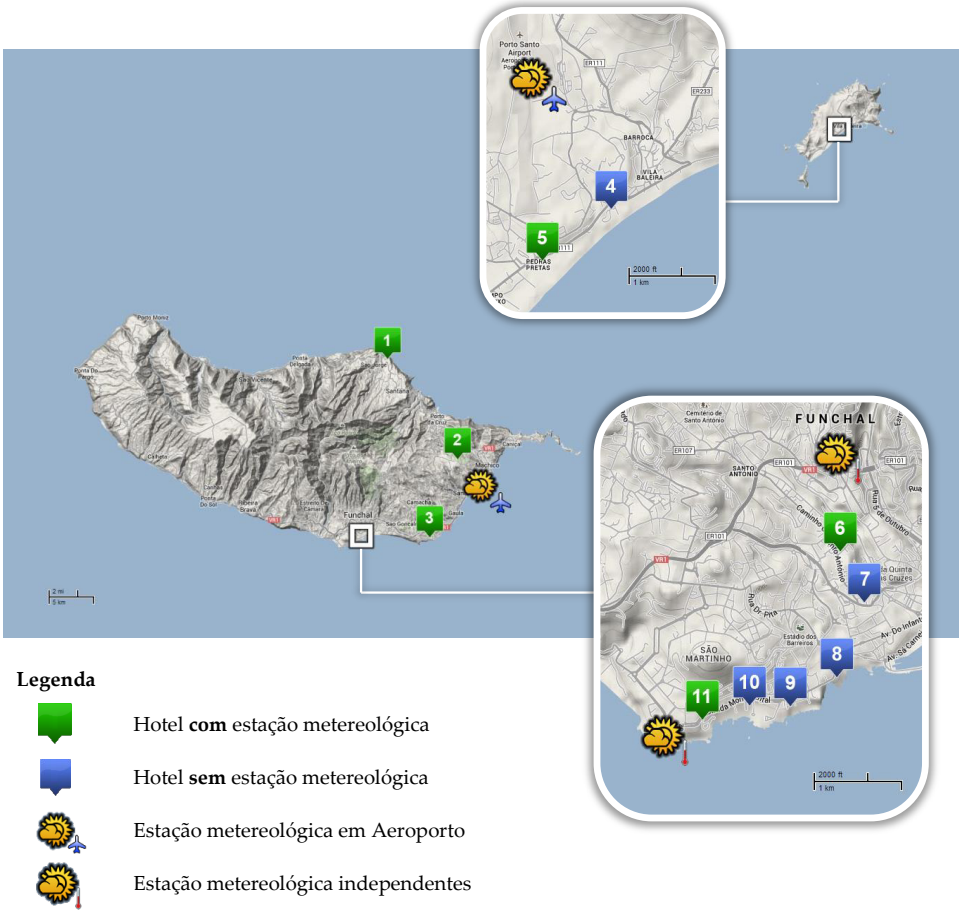


Figura 2.6: Localização geográfica dos estabelecimentos hoteleiros do projeto Soltgest e estações meteorológicas independentes

O maior aglomerado de unidades hoteleiras encontra-se situado próximo da capital da RAM, a cidade do Funchal, localizado na região mais a sul da ilha. Dois dos estabelecimentos, unidades 4 e 5, encontram-se localizados na ilha de Porto Santo. As estações meteorológicas não foram instaladas em todos os estabelecimentos, no entanto, os que não possuem instaladas estão geograficamente próximos de outros estabelecimentos que instalaram e/ou de estações meteorológicas independentes, num raio de aproximadamente de dois quilómetros.

Com o intuito de validar e colmatar possíveis falhas nos conjuntos de dados provenientes das estações meteorológicas instaladas nos estabelecimentos hoteleiros, foram identificadas estações meteorológicas externas ao projeto. Na RAM existem diversas estações meteorológicas que distribuem livremente as leituras das condições meteorológicas no local através de serviços *web*. As estações meteorológicas independentes ao projeto podem ser classificadas em duas categorias: as estações meteorológicas do grupo ANA (Aeroportos de Portugal) e as estações meteorológicas de empresas privadas ou de particulares. As estações meteorológicas do grupo ANA encontram-se localizadas no Aeroporto Internacional da Madeira e no Aeroporto do Porto Santo. Por sua vez, as restantes estações meteorológicas independentes encontram-se espalhadas pela região e podem ser da responsabilidade de uma empresa ou até mesmo de um indivíduo particular.

Na Tabela 2.7 são apresentadas as unidades hoteleiras pelos seus respetivos índices como ilustradas na Figura 2.6. A tabela descreve o nome de cada unidade hoteleira, a sua dimensão, categoria em estrelas, a distância às estações meteorológicas mais próximas por topologia e um sumário do número de sensores por serviço. Como demonstrado, os sistemas de sensores instalados diferem entre cada estabelecimento hoteleiro.

O Hotel Quinta do Furão é a unidade hoteleira mais afastada de todas as restantes e encontra-se na zona Norte da ilha da Madeira que possui um clima de maior humidade e precipitação. É igualmente o único estabelecimento que não possui qualquer sensor para a medição do consumo de água. O número de hotéis com 4 e 5 estrelas é equilibrado, no entanto, apenas um hotel de 5 estrelas, o Hotel Quinta das Vistas, possui uma estação meteorológica instalada. À exceção do Hotel Quinta do Furão, todas as restantes unidades hoteleiras encontram-se a menos de 10km de um outro estabelecimento com estação meteorológica instalada ou de estações meteorológicas independentes. O foco do sistema de sensores do projeto Soltgest refletiu-se no consumo de eletricidade, evidente no número de sensores de consumo eletricidade instalados, sempre acima de 6, em prol dos de consumo de água e gás,na sua maioria com menos de 4 e 2 sensores respetivamente.

ID	Hotel	Dimensão	Estrelas	Distância a Estações Metereológicas			Sensores		
				Hotel	Aeroporto	Independete	Electricidade	Água	Gás
1	Quinta do Furão	45	4	-	-	-	9	0	1
2	Enotel Golf do Santo da Serra	68	4	-	1 < 5km	-	7	2	1
3	Alpino Atlântico	27	4	6 11 < 5km	1 < 5km	2 < 5km	6	2	2
4	Torre Praia	66	4	5 < 1km	1 < 2km	-	8	2	0
5	Porto Santo	99	4	-	1 < 2km	-	9	3	0
6	Quinta das Vistas	71	5	11 < 2km	1 < 10km	2 < 2km	8	4	1
7	Quintinha de S.João	42	5	6 11 < 2km	1 < 10km	2 < 2km	6	2	1
8	Reid's Palace	163	5	6 11 < 2km	1 < 10km	2 < 2km	7	2	6
9	CS Madeira	300	5	6 11 < 2km	1 < 10km	2 < 2km	8	4	2
10	Meliã Madeira Mare	220	5	6 11 < 2km	1 < 10km	2 < 2km	6	2	1
11	Baía Azul	215	4	6 < 2km	1 < 10km	2 < 2km	7	4	2

Figura 2.7: Sumário das unidades Hoteleiras do projeto SOLTGEST

2.6 Sistema Soltgest

O projeto Soltgest foi concebido sobre o conceito de processamento centralizado onde, a cada hora, uma unidade central é responsável por receber as leituras dos sistemas de sensores instalados nas unidades hoteleiras como ilustrado na Figura 2.8. A unidade central é responsável por armazenar as leituras dos sistemas de sensores e de enviar os dados para a plataforma Soltgest. Os sistemas de sensores foram configurados para reportar as leituras à unidade central com uma periodicidade horária.

2.6.1 Sistemas de Sensores

Como referido na Secção 2.5, a cardinalidade dos sistemas de sensores instalados por serviço em cada estabelecimento hoteleiro difere. Os sistemas instalados registam o consumo de eletricidade, água e gás. As estações meteorológicas não foram instaladas em todos os estabelecimentos hoteleiros, Tabela 2.7. Nos estabelecimentos onde foram instaladas

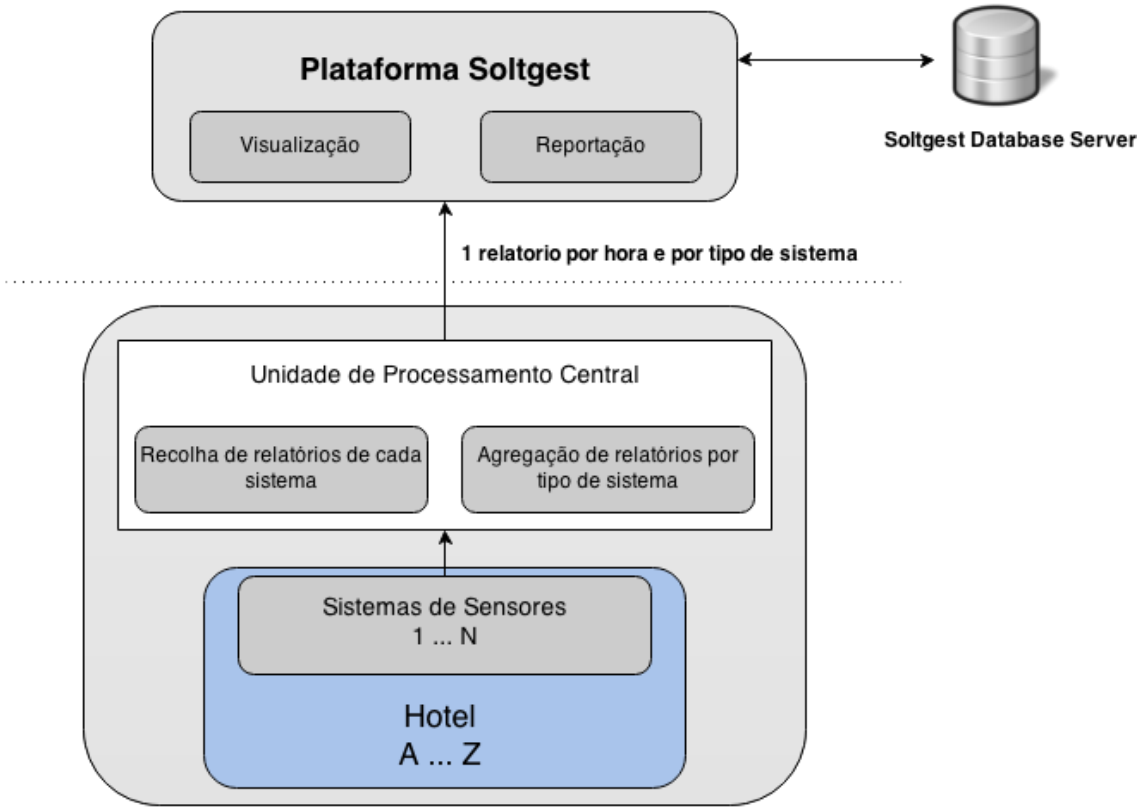


Figura 2.8: Arquitetura da plataforma Soltgest

estações meteorológicas, foi colocada uma estação no interior do estabelecimento e outra no exterior.

2.6.2 Aplicação Soltgest

A plataforma Soltgest disponibiliza uma aplicação *web* que permite monitorizar e analisar as leituras recolhidas pelos sistemas de sensores. Os utilizadores tem a possibilidade de verificar as leituras registadas pelos sensores e realizar análises descritivas ou sumários estatísticos do histórico de leituras. Adicionalmente, a aplicação permite cruzar a informação entre as leituras dos sensores de consumo de serviços, as leituras das estações meteorológicas e os registos de ocupação. Uma recente extensão à aplicação permitiu a inclusão das tarifas aplicadas por cada fornecedor a cada serviço (eletricidade, água e gás) permitindo aos utilizadores visualizar o custo económico proveniente do consumo de cada serviço. A Figura 2.9 exemplifica a *interface* da aplicação *web*.

2.6.3 Soluções Concorrentes

Existem diversas soluções de gestão de eficiência de consumo de serviços oferecidas por fornecedores de serviços como a *Chevron*, por companhias de automatismo e sistemas como a *BigFoot*, *CyberMetrics (FaciliWorks)*, *Icetel Engineering (Domotel)*, entre outras, e

Figura 2.9: Interface da plataforma Soltgest

por companhias de tecnologias de informação como a *IBM (BM® TRIRIGA® Energy Optimization)*. No entanto, o público alvo das atuais soluções são as companhias do setor do comércio, dos transportes, de comunicações e da administração pública com grandes infraestruturas. A generalidade das funcionalidades apresentadas incluem: a monitorização dos sistemas de sensores (alguns em tempo real); elaboração de relatórios; a inclusão de tarifas de cada serviço; sistemas de recomendação de como melhorar a eficiência e/ou reduzir os custos relacionados com o consumo de serviços; a possibilidade de controlar alguns equipamentos tais como os sistemas de AVAC de forma a facilitar a gestão e o controle dos consumos dos mesmos. O número de soluções para a gestão de consumos de serviços para o setor de alojamento é mais escasso e foram identificadas as seguintes soluções:*Evolve Guest Controls* da companhia *Evolve*; *EcoCentral Virtual Engineer* de *Telkonet*; um módulo em particular de um produto de *Business Intelligence* desenvolvido pela *Schneider Electric*.

A *interface* de uma das soluções da concorrência, a aplicação *EcoCentral Virtual Engineer*, é apresentada na Figura 2.10. O contributo principal apresentado pelas aplicações concorrentes tem sido a diminuição dos custos de operação por quarto de hóspedes e, como tal, o foco dos sistemas de sensores tem sido nas próprias habitações ao invés das infraestruturas. Na sua maioria, apresentam também mecanismos de deteção de mal-funcionamento de equipamentos, monitorização em tempo real e, no caso da solução *EcoCentral Virtual Engineer*, mecanismos de controle sobre os equipamentos.

No entanto, nenhuma das soluções apresenta, no momento presente, o uso de técnicas de data mining ao fim de inferir padrões de consumo através da análise analítica dos dados recolhidos.

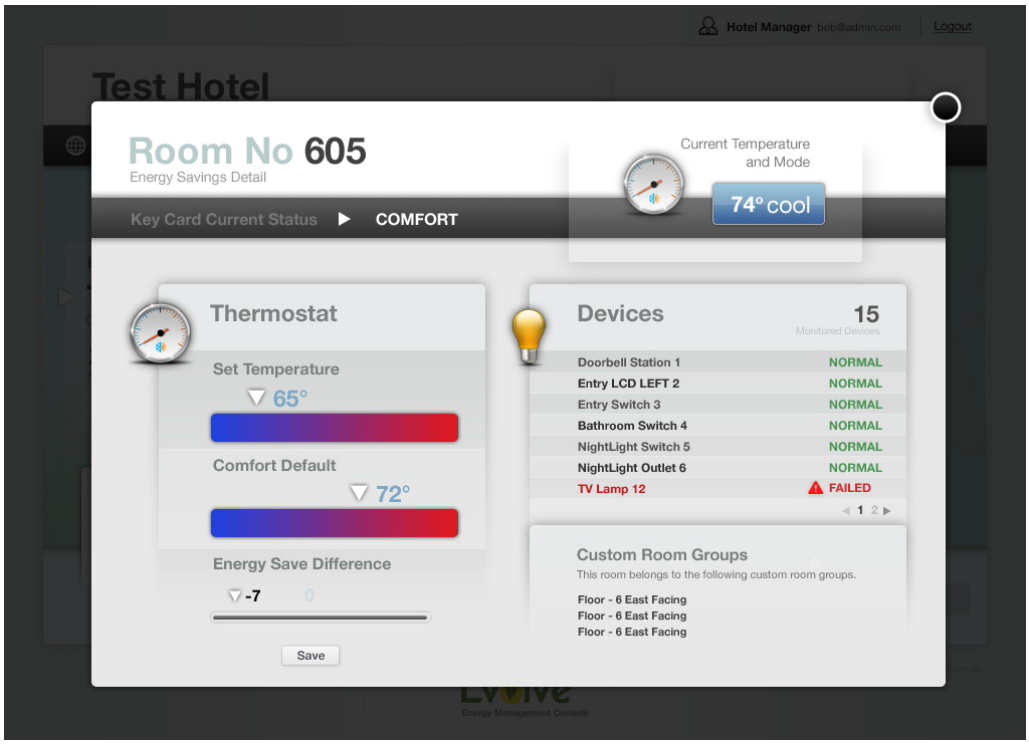


Figura 2.10: *EcoCentral Virtual Engineer* de *Evolve*,(Source: Evolve Website, 2013)

CAPÍTULO 3

INTERPRETAÇÃO DOS DADOS

Neste capítulo são apresentados os conjuntos de dados do projeto Soltgest e os conjuntos de dados recolhidos de fontes externas. Os conjuntos são discriminados consoante os seus atributos e unidades de medida, caracterizados pelo seu volume, analisados perante instâncias em falta e sujeitos a uma análise exploratória.

O capítulo está estruturado em três secções. A Secção 3.1 discrimina e caracteriza as variáveis dos conjuntos de dados quanto à sua semântica, unidade de medida e ritmo de registo. A Secção 3.2 apresenta o volume de dados, o período da informação e o número de observações em falta em cada um dos conjuntos de dados. A Secção 3.3 apresenta uma análise exploratória aos conjuntos de dados.

3.1 Descrição dos Conjuntos de Dados

O conjunto de dados para o projeto Soltgest pode ser discriminado em três categorias: os registos de ocupação em cada unidade hoteleira; as leituras de consumo de serviços; as leituras de condições meteorológicas.

3.1.1 Registos de Ocupação

Os registos de ocupação provêm da utilização da aplicação *web* Soltgest por parte dos funcionários em cada unidade hoteleira. Os registos inseridos indicam o número diário de hóspedes que realizaram *chek-in* em cada unidade hoteleira e o número de quartos de hóspedes que foram respetivamente ocupados. A dimensão do estabelecimento (em termos de número total de quartos) é de igual forma registada. Os atributos são especificados na Tabela 3.1.

Tabela 3.1: Discriminação dos atributos dos registos de ocupação

Registo	Unidade	Granularidade
Número Hóspedes	Hóspedes	Diária
Número de Quartos Alugados	Quartos	Diária
Dimensão	Quartos	Diária

3.1.2 Sensores de Consumo de Serviços

Os sistemas de sensores instalados em cada unidade hoteleira têm como função a leitura horária dos consumos de eletricidade, água e gás, ilustrado na Tabela 3.2.

Tabela 3.2: Discriminação dos atributos dos sistemas de sensores

Serviço	Unidade	Granularidade
Electricidade (Energia Activa)	*Watts*	Horária
Água	m^3	Horária
Gás	m^3	Horária

Em cada estabelecimento hoteleiro foram instalados diferentes sistemas de sensores. As Figuras 3.1 e 3.2 apresentam a disposição dos sensores consoante unidade hoteleira, local de instalação e tipo de sensor.

Figura 3.1: Sensores de eletricidade instalados por estabelecimento hoteleiro

Os sensores de consumo de eletricidade instalados registam dois tipos de leituras horárias distintas, a leitura instantânea e a leitura diferencial. As duas categorias foram nomeadamente designadas por *Total* e *Geral* no sistema Soltgest. A distinção entre as categorias foi indicada por um especialista no setor de Energia na Altran PT, que apontou as leituras dos sensores designados por *Total* ao valor consumido no instante em que a leitura era registada. Por sua vez, os sensores designados como *Geral* foram associados ao diferencial de consumo de eletricidade entre o consumo no momento da leitura e o consumo da leitura precedente. Como tal, as leituras do sensor *Geral Energia Ativa*

correspondem ao consumo de eletricidade ocorrido durante uma hora e efetivamente, o consumo pelo qual o consumidor será tarifado pelo seu fornecedor de energia elétrica.

O foco dos sensores de consumo de serviços colocados em setores mais específicos, referenciados em tom de laranja na Figura 3.1, é no entanto desconhecido, i.e., a única informação disponível relativa a estes sensores é a sua designação na base de dados e apesar de ser de certa forma representativa é igualmente ambígua. Aquando da comparação de sensores, instalados nos setores mais específicos, com a mesma designação mas entre estabelecimentos diferentes, o seu significado e a abrangência das suas leituras, quer a nível de infraestrutura quer a nível de equipamento, podem ser idênticas, semelhantes, ou totalmente distintas. De igual forma, não existe qualquer informação sobre o número de unidades elétricas a ser monitorizadas em cada sensor dos setores mais específicos. Por exemplo, a designação de *Restaurante*, *Frigorífica* e *Cozinha* apresentados na Figura 3.1 são apenas alguns exemplos que caracterizam os pontos anteriormente referidos. Adicionalmente, não foi possível concluir se as leituras dos sensores instalados nos setores mais específicos estão contabilizados nas leituras dos sensores de consumo geral. Como tal, foi assumido que o sensor *Geral Energia Ativa*, destinado à monitorização do consumo de eletricidade a cada hora no estabelecimento hoteleiro, já incluí o consumo registado pelos sensores dos setores mais específicos.

Figura 3.2: Sensores de água e gás instalados por estabelecimento hoteleiro

A heterogeneidade e falta de especificação abrangeu igualmente os sensores de consumo de água e gás, Figura 3.2. Os sensores de água possuem um sensor denominado de *Geral Água* que foi igualmente considerado como o sensor que regista o diferencial de água consumida ao longo de cada hora na unidade hoteleira. Por sua vez, os sensores de gás são na sua totalidade apontados a setores mais específicos.

Dada a incerteza da representatividade dos sensores em setores mais específicos e a alguns obstáculos na comunicação com a entidade reguladora do projeto na RAM, limitou-se o estudo ao sub-conjunto dos sensores que monitorizam o consumo geral em cada unidade hoteleira. Dada a ausência de sensores com a designação de *Geral* para o serviço de Gás e de forma a não excluir o consumo do serviço de gás do conjunto de dados, foi

assumido que o sensor denominado por *Caldeira* representaria o diferencial de consumo de gás por hora nos estabelecimentos. A representação foi atribuída ao sensor *Caldeira* dado o seu volume e à ordem de valores em comparação com os restantes sensores. Em suma, foram apenas considerados para análise os sensores *Geral Energia Ativa*, *Geral Água* e *Caldeira* como os sensores que registam, por hora, o diferencial de consumo de eletricidade, água e gás em cada unidade hoteleira.

3.1.3 Estações Meteorológicas

Os dados disponíveis relativos às condições meteorológicas na RAM são provenientes de duas fontes distintas, da plataforma Soltgest e de um agente externo. A primeira é referente às leituras recolhidas pelas estações meteorológicas instaladas nos estabelecimentos hoteleiros e a segunda referente a um serviço *web* que disponibiliza leituras de estações meteorológicas em diversos locais do planeta.

Os locais de instalação das estações meteorológicas nos estabelecimentos hoteleiros ao abrigo do projeto Soltgest, interiores ou exteriores, são desconhecidos. Como tal, é difícil precisar em que local estão as condições meteorológicas a ser recolhidas, particularmente em que local a unidade interior foi instalada.

De forma a poder validar os dados recolhidos pelas estações meteorológicas instaladas ou de forma a colmatar possíveis falhas nas leituras de condições meteorológicas, recorreu-se a um agente externo, o *website wunderground.com*, que disponibiliza um serviço gratuito para visualizar o histórico das condições meteorológicas em diversos locais do planeta. As leituras disponibilizadas por este serviço foram por sua vez disponibilizados por companhias, públicas ou privadas, ou por indivíduos particulares com estações meteorológicas. Como referido na Secção 2.5, após diversas análises exploratórias a conjuntos de dados provenientes de diversas estações, optou-se pelos dados provenientes das estações meteorológicas do grupo ANA dada a sua credibilidade e importância que a monitorização rigorosa das condições meteorológicas têm para o seu funcionamento. Outros fatores tiveram peso na escolha do grupo ANA tais como o volume de dados disponível, a abrangência a todo o período do projeto Soltgest, a ausência de falhas nas leituras recolhidas e à periodicidade a que são efetuadas as leituras. Para a recolha dos dados registados pelo grupo ANA através do serviço *web*, recorreu-se à implementação de um *script* que iterativamente recolheu os dados diários entre 2010 e 2013. A periodicidade das leituras dos dados recolhidos pelas estações meteorológicas do projeto Soltgest é horária e a periodicidade das leituras provenientes das estações meteorológicas do grupo ANA é de meia hora. As condições monitorizadas pelas estações meteorológicas são apresentadas na Tabela 3.3.

Tabela 3.3: Discriminação dos atributos das estações meteorológicas

Estação	Condição	Unidade	Granularidade
Interior	Temperatura ambiente	^oC	Horária
Interior	Humidade relativa	*Percentagem*	Horária
Interior	Pressão absoluta	$inHg$	Horária
Exterior / Aeroporto (ANA)	Temperatura ambiente	^oC	Horária/30 min.
Exterior / Aeroporto (ANA)	Humidade relativa	*Percentagem*	Horária/30 min.
Exterior / Aeroporto (ANA)	Precipitação	*inches*	Horária/30 min.
Exterior / Aeroporto (ANA)	Velocidade do vento	km/h	Horária/30 min.
Exterior / Aeroporto (ANA)	Direção do vento	^{o} [B.1]	Horária/30 min.

3.1.4 Eventos Socio-Culturais

As datas dos eventos socio-culturais com maior impacto no turismo e relevantes ao projeto Soltgest foram recolhidos de documentos oficiais produzidos pela Direção Geral do Turismo na RAM e disponibilizados *online*, Governo da Madeira (2011); Governo da Madeira (2012d). As datas são apresentadas no Apêndice A.

3.2 Volume e Periodicidade dos Conjuntos de Dados

A plataforma Soltgest iniciou a sua atividade em Julho de 2010, período em que iniciou a recolha de informação proveniente dos sistemas de sensores instalados nas unidades hoteleiras. No entanto, verificou-se que o volume de dados existente por tipo de conjunto de dados e por estabelecimento hoteleiro não é idêntico.

3.2.1 Registos de Ocupação

O processo diário de registo de ocupação na aplicação *web* do projeto Soltgest implica o registo do número de hóspedes presentes no estabelecimento, i.e., que pernoitaram na noite anterior no estabelecimento, o registo do respetivo número de quartos que foram ocupados e o número total de quartos que a unidade hoteleira possuía como disponíveis a hospedar (dimensão). A Figura 3.3 identifica os períodos em que foram registados os valores de ocupação em cada estabelecimento hoteleiro, ordenado de forma descendente por volume de dados. As unidades hoteleiras Q.Furão Hotel, Torre Praia Hotel e Hotel do Santo possuem pouco mais que os primeiros registos de ocupação.

A Tabela 3.4 complementa a Figura 3.3 com um sumário descritivo do volume dos dados, onde é apresentado a primeira e última data dos registos em cada estabelecimento. A janela temporal comum entre todas as unidades hoteleiras, i.e., o domínio temporal do conjunto de dados, foi definida através da data do registo mais antigo, 21 de Setembro de 2010, e da data do registo mais recente entre todos os estabelecimentos, 3 de Setembro de 2013.

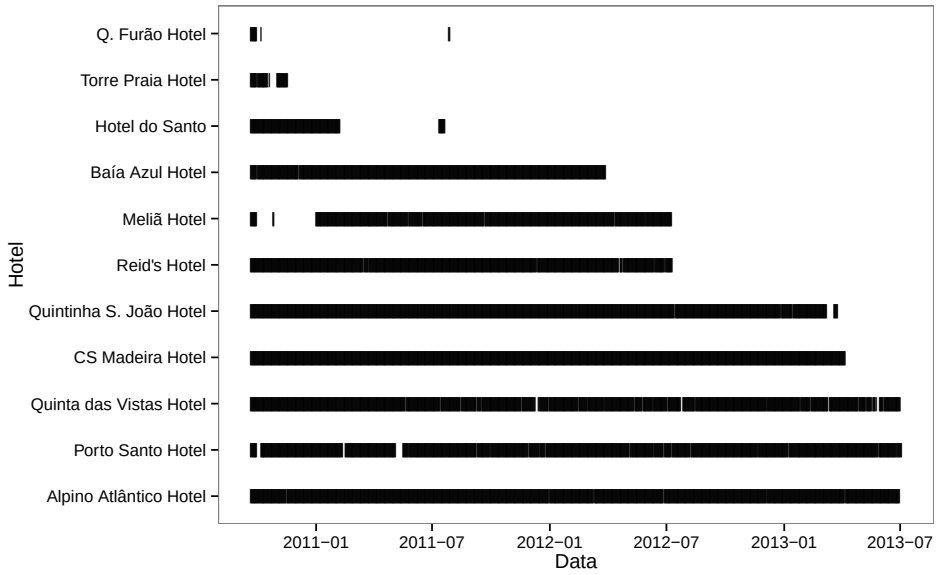


Figura 3.3: Periodicidade dos registos de ocupação

Tabela 3.4: Sumário dos registos de ocupação no conjunto de dados

Hoteís	**n**	**#NA**	**Min**	**Max**
Q. Furão Hotel	13	1004	2010-09-21	2011-07-29
Torre Praia Hotel	44	973	2010-09-21	2010-11-17
Hotel do Santo	150	867	2010-09-21	2011-07-21
Baía Azul Hotel	553	464	2010-09-21	2012-03-28
Meliã Hotel	562	455	2010-09-21	2012-07-09
Reid's Hotel	651	366	2010-09-21	2012-07-10
Quintinha S. João Hotel	903	114	2010-09-21	2013-03-25
CS Madeira Hotel	929	88	2010-09-21	2013-04-06
Quinta das Vistas Hotel	977	40	2010-09-21	2013-07-01
Porto Santo Hotel	981	36	2010-09-21	2013-07-03
Alpino Atlântico Hotel	1008	9	2010-09-21	2013-06-30

A data do primeiro registo de ocupação efetuado é idêntica em todos os estabelecimentos hoteleiros, no entanto, o número de registos e a data do último registo varia entre estabelecimentos. Com base nos objetivos propostos na Secção 1.5, não existem dados suficientes de ocupação nos Hotéis Q.Furão e Torre da Praia para prosseguir com a análise de consumo de serviços em função do número de hóspedes. O conjunto de dados referente ao Hotel do Santo possui menos de meio ano de registos e visto não completar uma sazonalidade anual não é suficientemente expressivo dada a sazonalidade anual inerente à atividade turística.

Os Hotéis Baía Azul, Meliã e Reid's contêm quase de 2 anos de registos e os restantes estabelecimentos praticamente 3 anos.

3.2.2 Sensores de Consumo

No seguimento da Secção 3.2.1 os dados relativos aos Hotéis Q.Furão, Torre Praia e Hotel do Santo foram excluídos de análise. O domínio do intervalo temporal considerado relevante foi de 21 de Setembro de 2010 a 3 de Setembro de 2013 visto ser o período em que existe informação relativa à ocupação nas unidades hoteleiras. A Figura 3.4 apresenta o sumário dos registos dos sistemas de sensores de consumo serviços. Em Apêndice B a Tabela B.1 respetiva.

O Hotel Porto Santo não instalou qualquer sensor para a monitorização do consumo de gás e possui um grande número de falhas no sensor de monitorização de eletricidade. Mesmo considerando a hipótese de que os registos de consumo de eletricidade no Hotel Porto Santo fossem contínuos, estes corresponderiam no máximo a um período total de 28 dias. Com base no possível volume de dados e após observar a não-continuidade dos registos de consumo de eletricidade, os dados relativos ao Hotel Porto Santo não foram alvo de estudo.

À exceção dos sensores do Hotel Meliã e do sensor de gás no Hotel Reid's, todos os restantes hotéis possuem registos a partir de 21 de Setembro de 2010 à uma hora da manhã, o primeiro data do domínio temporal considerado. De notar que o sensor de gás no Hotel Reid's e o sensor de eletricidade no Hotel Meliã apresentaram o primeiro registo substancialmente mais tarde que os restantes serviços na mesma unidade hoteleira.

Em termos de volume total de dados é possível distinguir dois grupos de hotéis, sendo que primeiro grupo possui registos até meados do ano de 2013 onde se incluem os Hotéis Quintinha S.João, o Hotel Baía Azul e o Hotel Reid's, ignorando o sensor de consumo geral de gás no último hotel. O segundo grupo, constituído pelos Hotéis Quinta das Vistas, Alpino Atlântico, CS Madeira e Meliã apresentam registos até finais do ano de 2012.

3.2.3 Condições Meteorológicas

No seguimento das Secções 3.2.1 e 3.2.2, os dados relativos aos Hotéis Q.Furão, Torre Praia, Hotel do Santo e Porto Santo foram excluídos de análise. O domínio do intervalo temporal considerado relevante foi de 21 de Setembro de 2010 a 3 de Setembro de 2013 visto ser o período em que existe informação sobre a ocupação nas unidades hoteleiras.

O volume de dados proveniente das estações metrológicas interiores e exteriores é apresentado na Figura 3.5. Em Apêndice B a Tabela B.3 respetiva.

O volume de dados é distinto entre estabelecimentos hoteleiros, no entanto, o número de registos entre as unidades interiores e exteriores em cada estabelecimento é idêntico.

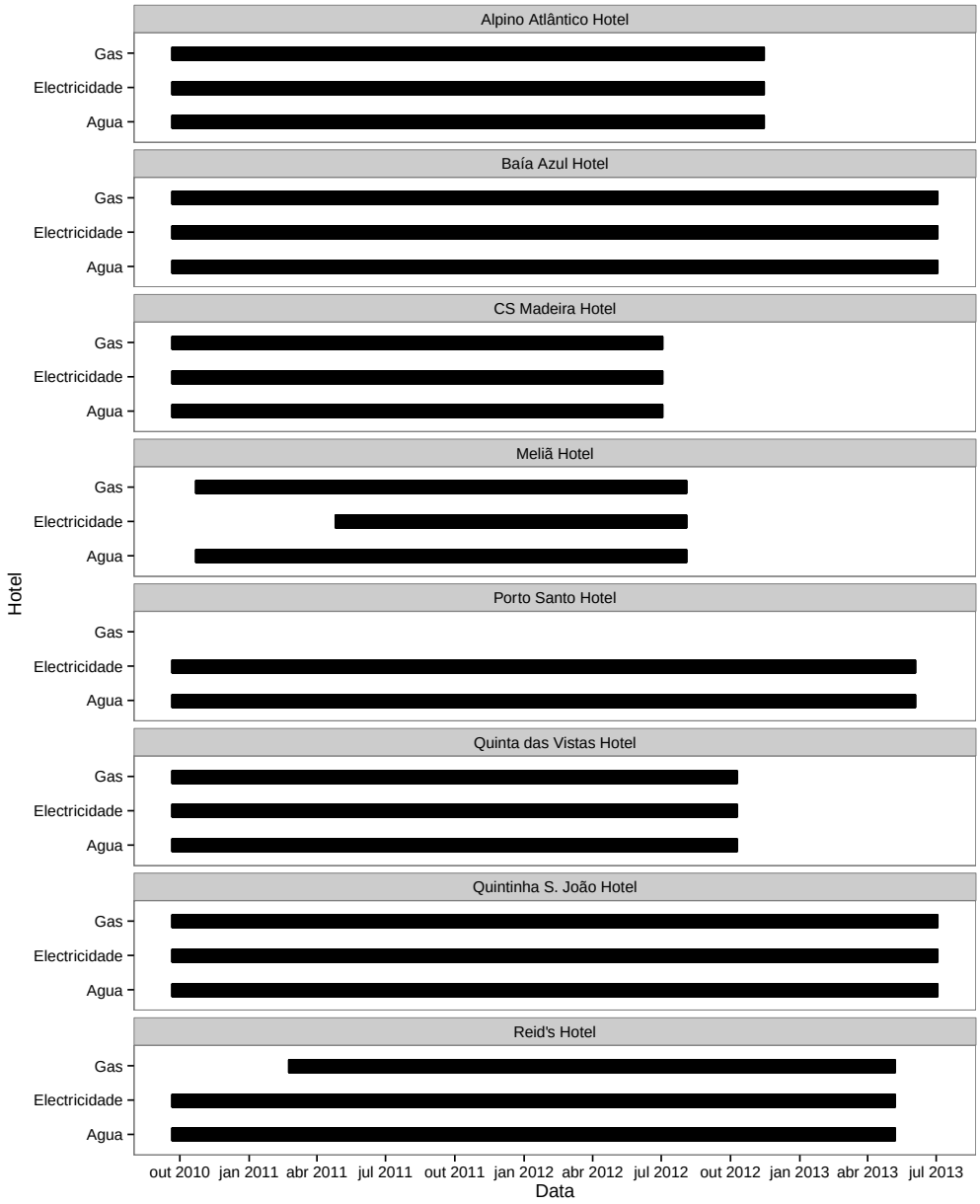


Figura 3.4: Periodicidade dos registos dos sensores de consumo de serviços

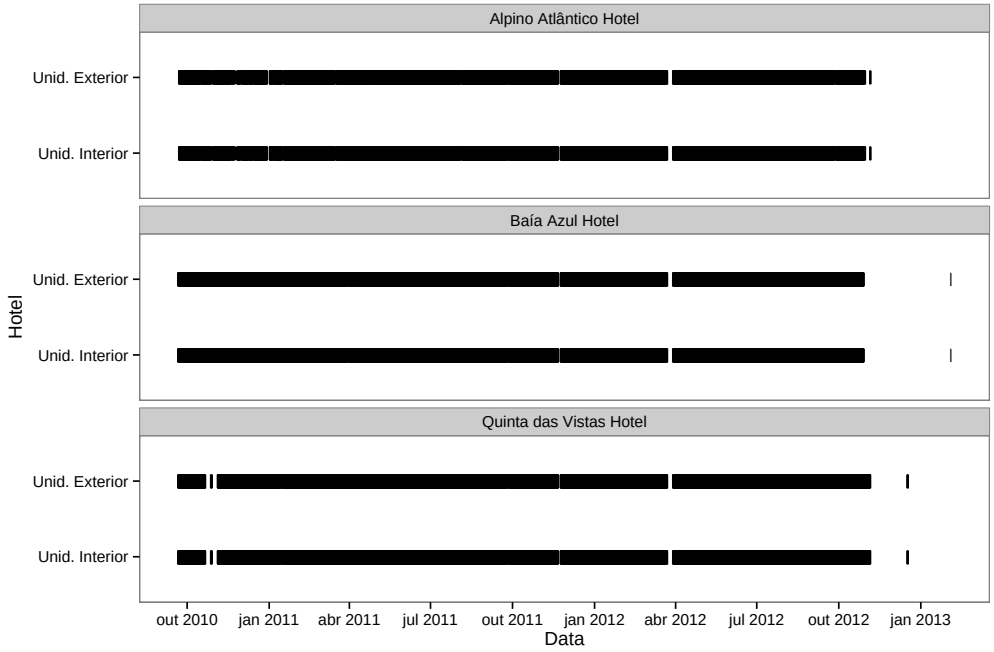


Figura 3.5: Periodicidade das estações meteorológicas

3.3 Análise Exploratória

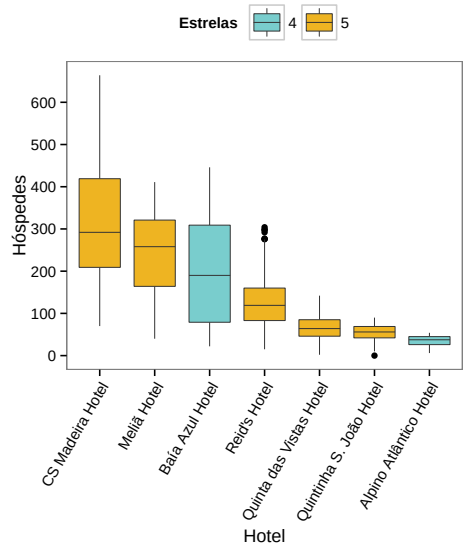
Nesta secção são apresentadas sucintamente as principais análises exploratórias realizadas com os conjuntos de dados disponíveis.

3.3.1 Registos de Ocupação

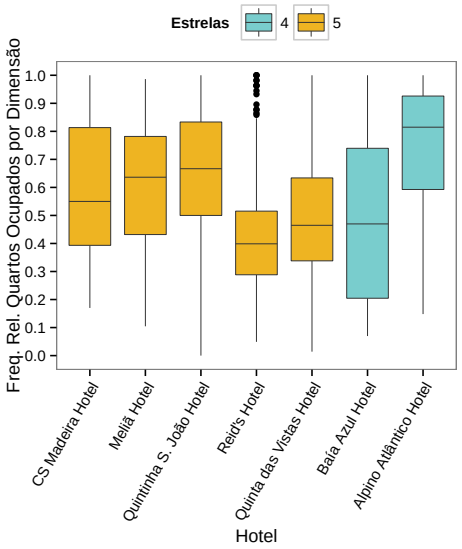
A dimensão de cada estabelecimento é registada diariamente na plataforma Soltgest juntamente com o número de hóspedes e número de quartos ocupados. No entanto o valor de dimensão registado não é constante durante todo o período temporal, assumindo por vezes valores muito semelhantes ou valores semelhantes multiplicados por diferentes potências de 10. De forma a eliminar incoerências, foi assumido que o número total de quartos era idêntico à moda da distribuição de valores e esse valor validado através do *website* da própria unidade hoteleira ou através de outros *websites* apropriados tais como o *Booking.com* e o *TripAdvisor.com*.

Como o número de quartos ocupados e o número hóspedes presentes na unidade hoteleira são inseridos manualmente, foram encontrados alguns registos cujo número de quartos ocupados era superior à dimensão do Hotel. As instâncias que apresentaram incoerência entre o número de quartos ocupados e a dimensão da unidade hoteleira foram excluídas de análise.

Na Figura 3.6(a) é apresentado um boxplot da distribuição do número de hóspedes em cada estabelecimento e o respetivo número de estrelas de cada unidade hoteleira. O volume de hóspedes é significativamente diferente entre as unidades hoteleiras. Dado



(a) Boxplot do número de hóspedes em cada estabelecimento



(b) Boxplot da frequência relativa entre o Número de Quartos Ocupados e a Dimensão de cada estabelecimento

Figura 3.6: Distribuição dos valores de ocupação por estabelecimento hoteleiro

que não existe informação relativa à capacidade máxima de hóspedes em cada unidade hoteleira, foi apresentado na Figura 3.6(b) a frequência relativa entre o número de quartos ocupados e a dimensão do hotel. O Hotel Alpino Atlântico, de 4 estrelas, possui o menor volume de hóspedes entre todas os estabelecimentos e é igualmente a unidade hoteleira com a maior taxa de ocupação. O Hotel Reid's por sua vez é a única unidade hoteleira onde a taxa de ocupação é menor que 50% em cerca de 75% das ocasiões.

A topologia mais usual de um quarto de hóspedes numa unidade hoteleira dedicada ao turismo é a de dois hóspedes por quarto. Para analisar como se distribuem os hóspedes em função do número de quartos alugados é apresentada a Figura 3.7.

Em cada estabelecimento hoteleiro procurou-se estimar uma relação linear entre o número de quartos alugados e o número de hóspedes presentes no estabelecimento. Para o efeito foi aplicado um algoritmo de regressão linear, apresentado no Capítulo 6 Secção 6.1, para analisar essa mesma relação. Cada observação foi igualmente identificada pela respetiva estação do ano. A maioria das unidades hoteleiras apresenta um coeficiente de correlação, r^2, muito próximo de 1, o que indica, de um modo geral, um ajustamento adequado do algoritmo ao conjunto de dados. Os coeficientes de relação, X, apresentam-se muito próximos mas sempre inferiores a 2, o que indica que por cada unidade de quartos ocupados o número de hóspedes presentes acresce aproximadamente 2 unidades. O erro médio absoluto $|e|$ indica por sua vez o quanto o modelo se pode enganar em média na sua estimativa de hóspedes presentes no hotel. Denotam-se no entanto os Hotéis CS Madeira, o de maior dimensão, e Alpino Atlântico, o de menor dimensão, apresentam os coeficientes mais baixos. Os hotéis de maiores dimensões apresentam um erro médio

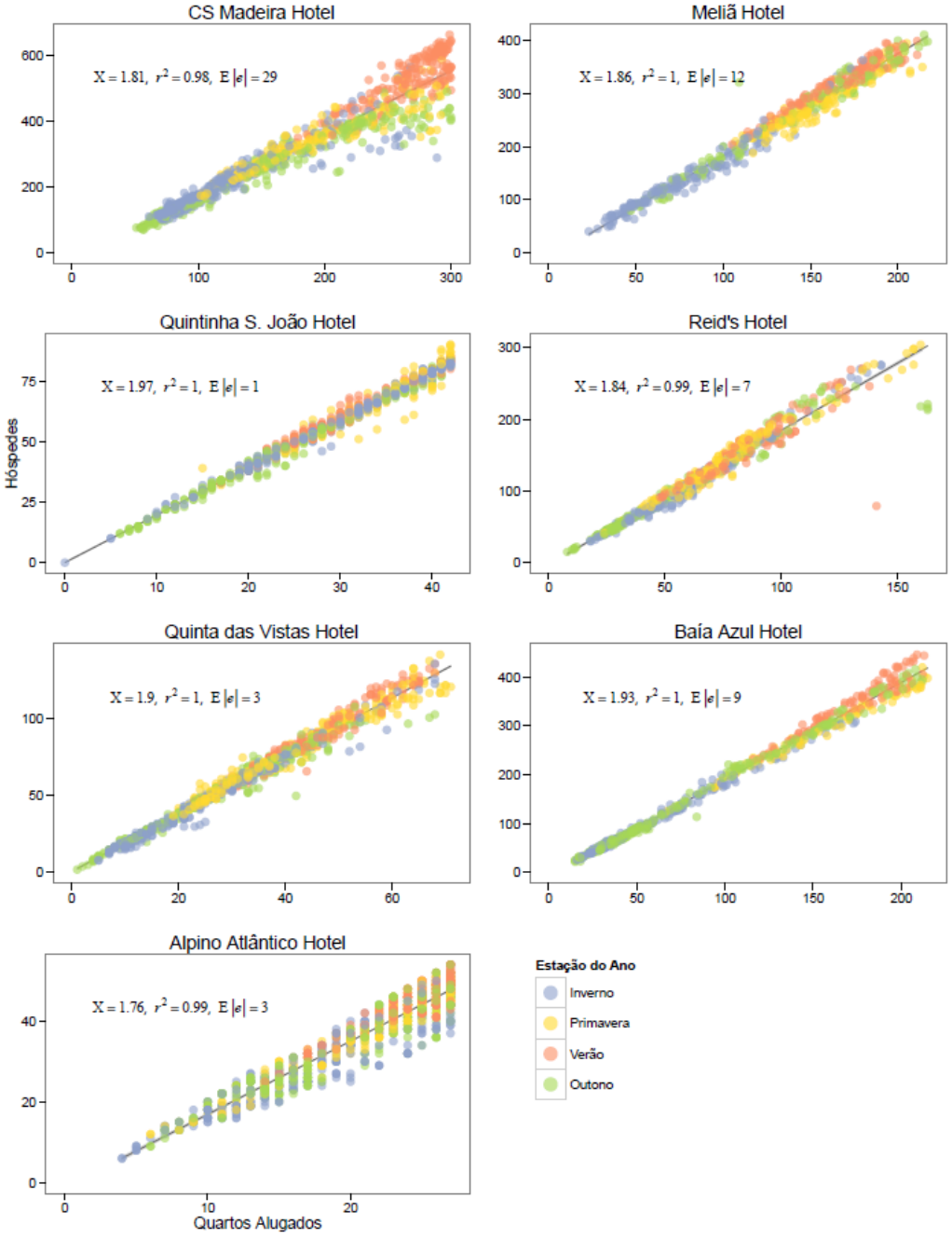


Figura 3.7: Relação entre o número de quartos alugados e número de hóspedes presentes em cada estabelecimento

absoluto superior aos dos hotéis de menores dimensões. Esta diferença pode ser, por exp., indicativo de que os hotéis de maiores dimensões possuem mais quartos individuais ou triplos que os de menor dimensão ou de terem mais camas extras disponíveis para as habitações duplas. A Figura 3.7 sugere também que os períodos de maior ocupação surgem durante as estações de Primavera e Verão na maioria dos estabelecimentos. O Hotel Quintinha de S.João aparenta ter uma distribuição de níveis de ocupação mais nivelados ao longo das estações do ano.

De forma a visualizar a distribuição da taxa de ocupação pelas diferentes estações do ano foi elaborada a Figura 3.8. A figura apresenta a frequência relativa entre o número de quartos ocupados e a dimensão do hotel ao longo das quatro estações do ano. No entanto, foram selecionados para a figura apenas os hotéis com 2 anos completos de histórico de forma a não considerar a anos incompletos. Os hotéis com apenas um ano completo de histórico são apresentados em Apendix B.

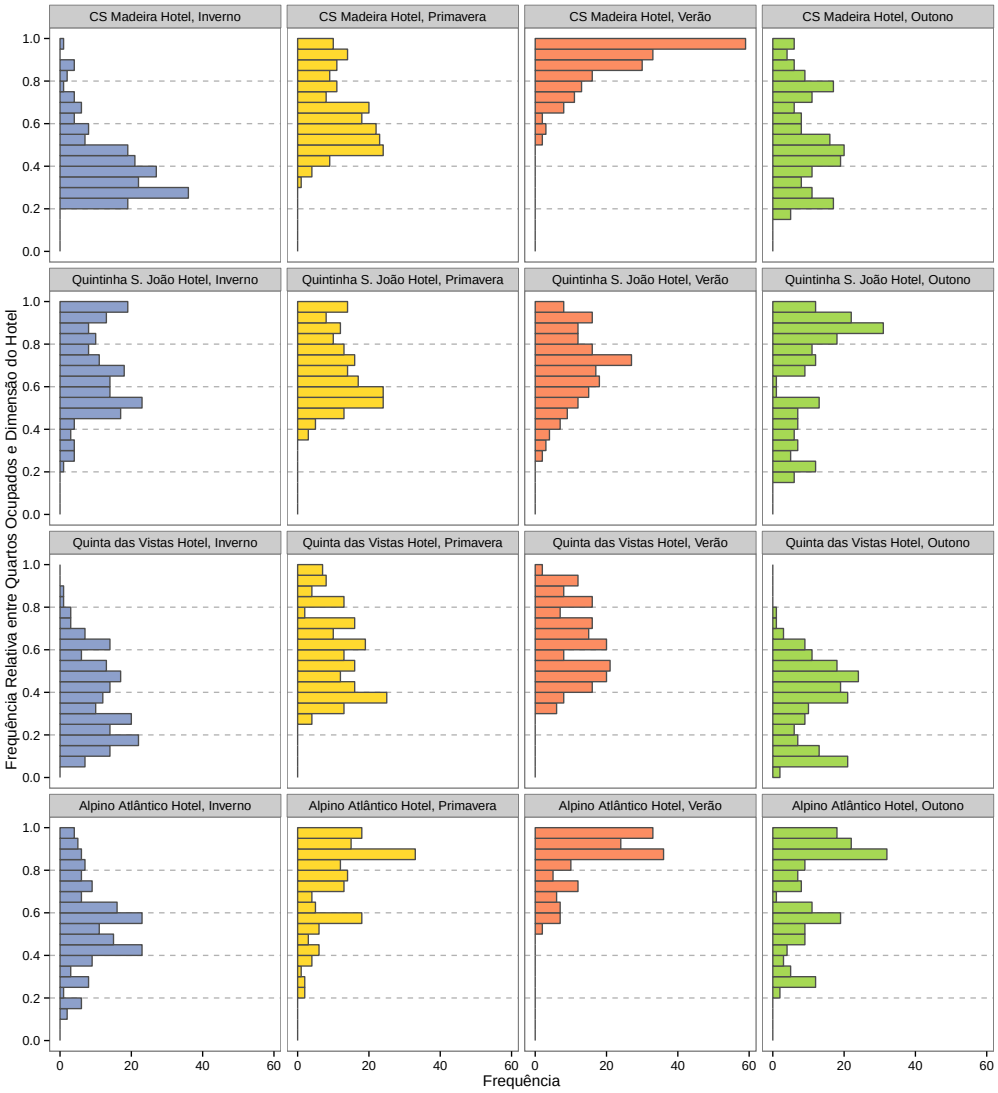


Figura 3.8: Distribuição da Frequência Relativa entre o número de quartos ocupados e a dimensão do hotel por cada estabelecimento e estação do ano

Na generalidade, os valores de ocupação atingem os valores mais altos durante as estações de Primavera e Verão. Os hotéis Quintinha de S.João e Alpino Atlântico, de menores dimensões, demonstram um maior equilíbrio da taxa de ocupação entre as distintas estações do ano. Os de maior dimensão, como por exemplo o hotel CS Madeira, apresentam uma maior diferenciação na taxa de ocupação entre diferentes estações.

A Figura 3.9 apresenta a distribuição da taxa de ocupação entre os anos 2011 e 2012 para os hotéis com 2 anos completos de histórico. É também apresentado por uma linha a tracejado o valor para o qual a taxa de ocupação é de 50%. A diferença entre os anos não revelou ser significativa e a variância na taxa de ocupação apresentou ser semelhante. À exceção do Hotel Quinta das Vistas, todos os restantes hotéis tiveram uma taxa de ocupação superior a 50% em mais de 50% dos dias de cada ano. Neste sentido, denota-se os Hotéis Alpino Atlântico e Quintinha de S.João que apresentam 75% dos valores de

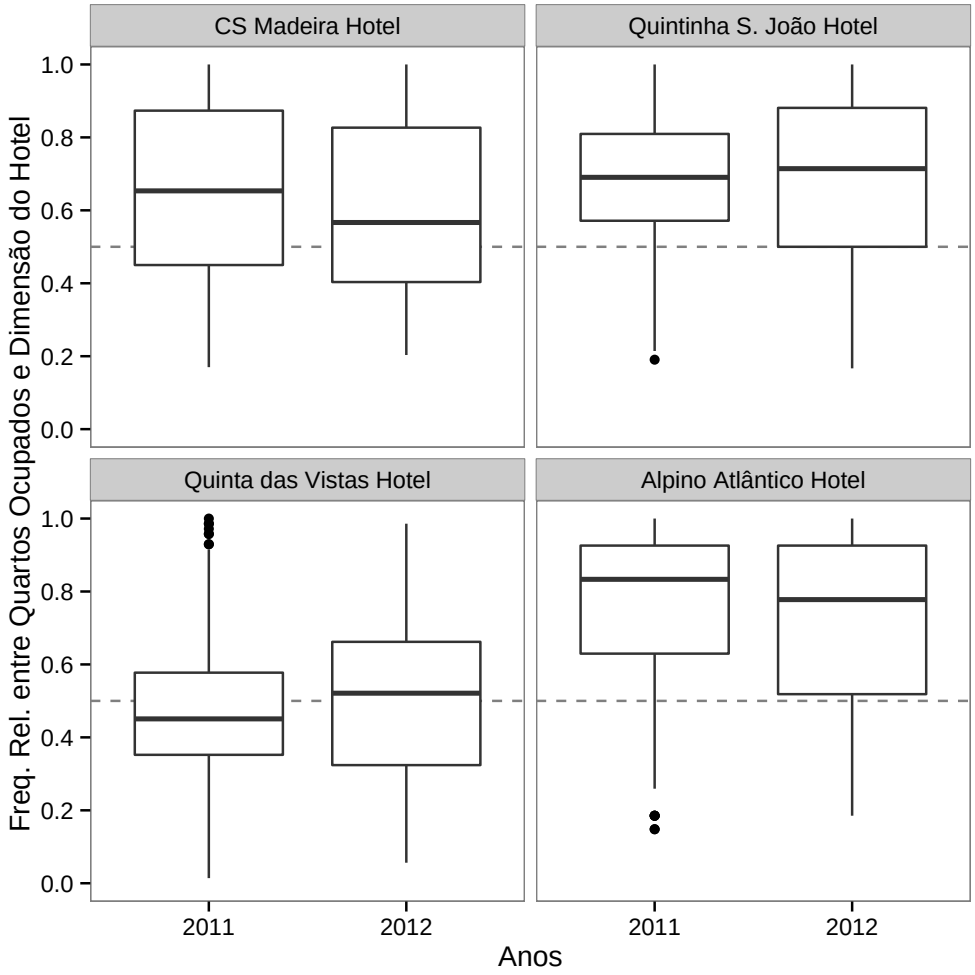


Figura 3.9: Distribuição da Frequência Relativa entre o número de quartos ocupados e a dimensão do hotel por cada estabelecimento e ano

ocupação superiores à metade da dimensão do hotel em 2011 e em 2012.

Verificou-se também que não existe grande variabilidade na taxa de ocupação entre os diferentes dias da semana durante as diferentes estações do ano, ilustrado na Figura B.4 no Apêndice B.

De um modo geral os hotéis apresentam distintos comportamentos relativamente à sua ocupação ao longo do período em questão. No entanto, a variabilidade da ocupação entre estações de ano é de um modo geral concordante entre todos os hotéis e o comportamento entre os anos 2011 e 2012 (entre os quais a informação é disponível) apresenta similaridades.

3.3.2 Condições Meteorológicas

O foco do estudo relativamente às condições meteorológicas que potencialmente influenciassem o consumo de serviços por hóspede recaiu sobre os valores de temperatura

ambiente e os de precipitação.

Infelizmente, a sensibilidade das estações meteorológicas disponíveis ao registo da precipitação ocorrente será provavelmente pouco elevada pois as três estações meteorológicas em análise, Hotéis Alpino Atlântico e Quinta das Vistas e Aeroporto do Funchal, registaram em média um diferencial horário de precipitação de zero em 99.56% nas suas leituras. Como indicado na Figura 2.2(b), a RAM é uma região onde ocorrem períodos de precipitação ao longo de quase todo o ano. Por outro lado, é também referido na Secção 2.2 que a precipitação na ilha da Madeira concentra-se especialmente na zona norte da ilha devido à morfologia do relevo. Porém, os Hotéis e o Aeroporto do Funchal e as suas estações meteorológicas encontram-se nas zonas Sul e Este da ilha da Madeira. Com base no valor constante de zero que a variável de precipitação apresenta, foi assumido não incorporar os valores de precipitação para análise de consumo de serviços.

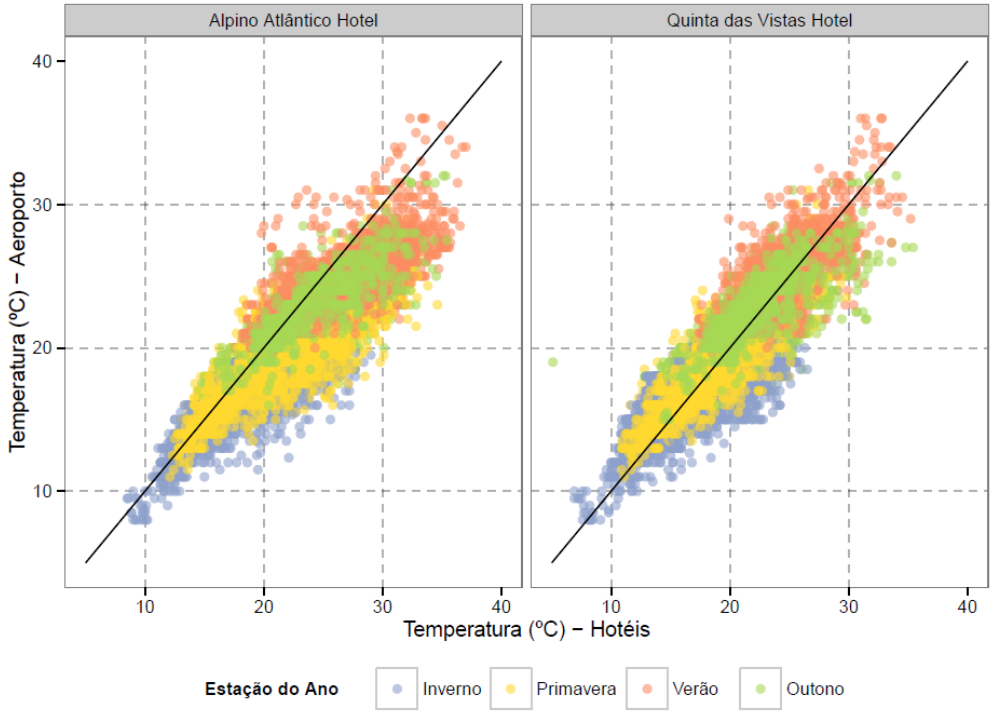


Figura 3.10: Distribuição de valores do sensor de Água no Hotel Reid's

Durante a análise exploratória à variável de temperatura ambiente verificou-se a existência de valores extremos nas leituras das estações meteorológicas instaladas nas unidades hoteleiras. Dos valores extremos referidos, encontram-se 8 leituras horárias da estação do Hotel Quinta das Vistas com valores de temperatura inferiores a -20^{0}C e por esta razão foram excluídas de análise. Como referido na Secção 3.1.3, os valores recolhidos de fontes externas têm o propósito de validar os valores recolhidos pelas estações meteorológicas instaladas nos estabelecimentos hoteleiros e de colmatar alguma falha nas leituras dos mesmos. A Figura 3.10 ilustra a distribuição dos valores de temperatura registados

pelas estações meteorológicas nas unidades hoteleiras perante os valores de temperatura registados pela estação meteorológica instalada no Aeroporto do Funchal. Na Figura 3.10 foi também assinalada uma linha reta a negro a exemplificar um hipotético alinhamento da distribuição das observações num cenário onde a relação entre ambas as variáveis fosse simétrica.

Apesar de as distribuições serem semelhantes, existe alguma diferença nos valores de temperatura registados pela estação no aeroporto e pelas estações instaladas nos estabelecimentos hoteleiros, especialmente com a leituras da estação do Hotel Alpino Atlântico. Para analisar melhor a diferença de valores, foi analisada a diferença entre pares de observações dos valores de temperatura das estações meteorológicas das unidades hoteleiras e a do aeroporto, Figura 3.11. Verificou-se que a diferença entre pares de observações atinge os 12ºC e os 10ªC para o Hotel Alpino Atlântico e o Hotel Quinta das Vistas respetivamente e que de facto existe uma maior variância na diferença de valores com o Hotel Alpino Atlântico que com o Hotel Quinta das Vistas.

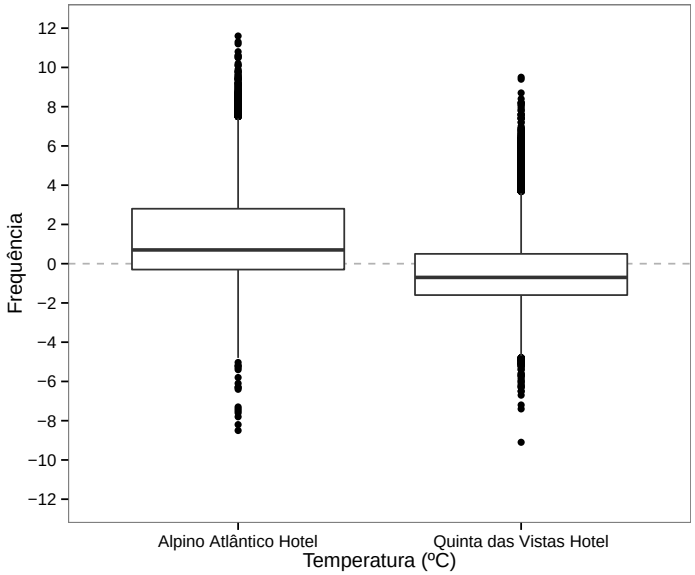


Figura 3.11: Distribuição da diferença entre pares de observações entre as estação meteorológicas nos Hotéis e a estação do Aeroporto

A diferença observada na Figura 3.11 entre pares de observações aliada ao facto de as estações meteorológicas encontrarem-se num raio de distância inferior ou igual a 10km, foi considerado que os dados das estações meteorológicas instaladas nas unidades hoteleiras poderiam não refletir os valores mais precisos das condições meteorológicas no local. Em termos comparativos é também de assinalar que o conjunto de dados proveniente do Aeroporto do Funchal era isento de falhas nas leituras.

Apesar de o propósito inicial dos dados de condições meteorológicas provenientes de fontes externas ter sido validar e colmatar possíveis falhas nos dados já disponíveis, devido a diversos fatores previamente referidos, foram adotados os valores provenientes

da estação do Aeroporto do Funchal como os valores verdadeiros de temperatura ambiente no local.

3.3.3 Sensores de Consumo

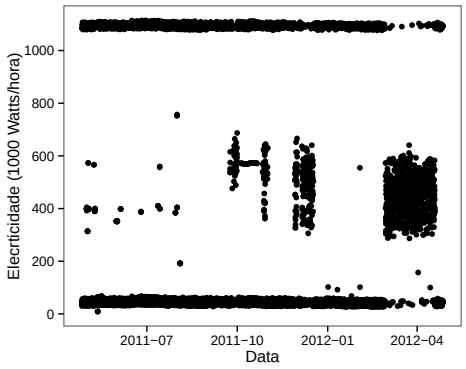
A análise exploratória aos sensores de consumo de serviços identificou algumas situações que provocaram a exclusão de alguns valores e por sua vez, a exclusão de alguns estabelecimentos de análise.

3.3.3.1 Meliã Hotel

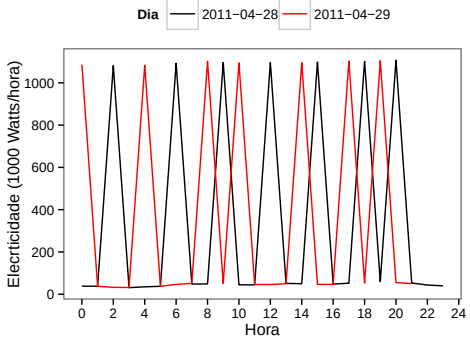
O primeiro registo do sensor de Eletricidade no Hotel Meliã ocorreu 6 meses após o primeiro registo de consumo de Água ou Gás. O sumário estatístico ao sensor de Eletricidade, Tabela 3.5, revelou valores extremos muito elevados e dispersos ao longo de todo período. No entanto a mediana e o primeiro quartil apresentavam valores mais adequados a um cenário real do consumo de eletricidade no estabelecimento.

Tabela 3.5: Sumário estatístico do sensor de Eletricidade no Hotel Meliã

Variavel	**Unidade**	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**
Electricidade	Watts/hora	0	33024	49408	17634993413.4	586464	9385699680128



(a) Distribuição de valores durante todo o período



(b) Tendência da distribuição de valores ao longo de dois dias

Figura 3.12: Distribuição de de um subconjunto filtrado de valores do sensor de Eletricidade no Hotel Meliã

De forma a procurar utilizar alguns dos valores de consumo de eletricidade, foram excluídos de análise as observações de valor superior ou inferior à mediana mais/menos três vezes o intervalo entre quartis. A primeira filtragem de observações resultou em 8233 instâncias de 11176, o que poderia se traduzir em 343 dias de 465 supondo a continuidade das observações. A Figura 3.12(a) demonstra a dispersão dos valores filtrados ao longo de todo o período e a Figura 3.12(b) demonstra a tendência horária da distribuição de valores ao longo de dois dias de forma a ilustrar a variação das observações a nível horário.

Na Figura 3.12(a) a dispersão dos valores na sua totalidade permite categorizar o consumo em dois grupos, um grupo com um consumo inferior a 100k Watts/hora e outro com um consumo superior a 1000k Watts/hora. Em outros momentos, dispersos ao longo do período e com maior frequência ao fim do mesmo, o consumo assume valores intermédios entre ambos os grupos. Na Figura 3.12(b) é possível observar que o consumo aparenta apresentar um padrão relativamente à variação entre valores tão distintos. À primeira vista o padrão aparenta divergir de valores de duas em duas horas, no entanto, o período intercalar entre a variação de valores não é constante. Foi possível estabelecer uma fronteira entre os diferentes valores de consumo aos 55k Watts/h. A Figura 3.13 possibilita verificar que a frequência de observações por hora do dia em cada grupo e ao longo de todo o período é bastante proporcional. No entanto, a Tabela 3.6 ilustra o quanto os valores diferem entre os diferentes níveis de valor.

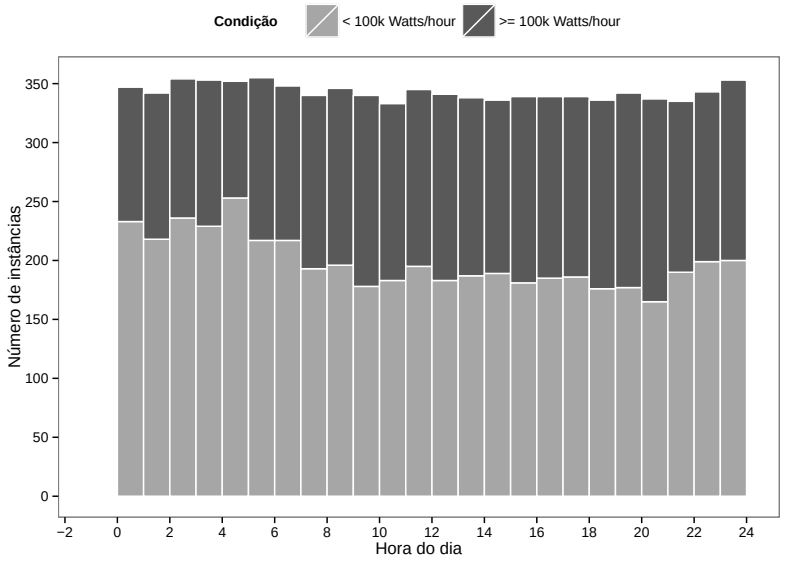


Figura 3.13: Frequência de um subconjunto filtrado de valores do sensor de Eletricidade no Hotel Meliã por nível e por hora

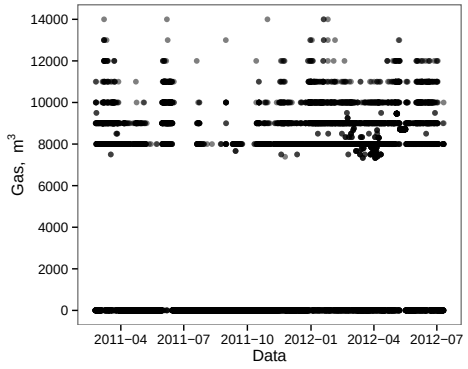
Não possuindo a capacidade de interpretar a distribuição de valores de consumo de eletricidade no Hotel Meliã, os valores de consumo de eletricidade foram desconsiderados e como em pontos anteriores, o conjunto de dados referente ao Hotel Meliã foi desconsiderado.

Tabela 3.6: Sumário estatístico de um subconjunto filtrado do sensor de Eletricidade no Hotel Meliã e repartido por níveis de valor

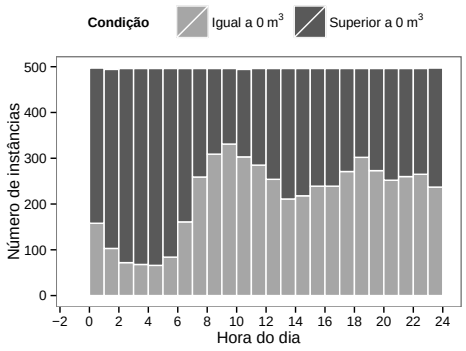
Variavel	Unidade	Níveis	n	Min	$\tilde{x}$	$\bar{x}$	Max
Electricidade	Watts/hour	< 100k Watts/hour	4766	39424	46336	45446.2	91904
		>= 100k Watts/hour	3467	508288	1090944	868413.1	1114368
		all	8233	44928	54016	392005.9	1114368

3.3.3.2 Reid's Hotel

A análise exploratória às leituras dos sensores de consumo do Hotel Reid's revelou um consumo de Gás atípico, Figura 3.18. Ao longo de todo o período, Figura 3.14(a), o consumo de Gás variou, com algumas exceções, entre um consumo de zero m^3 e um conjunto limitado de 7 valores distintos a rondar os 10k m^3. Para procurar perceber as leituras cujo consumo é de zero m^3 analisou-se a frequência ao longo do dia, Figura 3.14(b). A frequência com que as leituras registaram zero m^3 ao longo do dia apresentam um comportamento diferenciado, no entanto, 56% dos registos de leituras de gás equivalem a zero m^3.



(a) Distribuição de valores durante todo o período



(b) Frequência de valores por nível e por hora

Figura 3.14: Distribuição de valores do sensor de Gás no Hotel Redids por hora

Não possuindo a capacidade de interpretar a distribuição de valores de consumo de gás no Hotel Reid's, os valores de consumo de gás foram desconsiderados e como em pontos anteriores, o conjunto de dados referente ao Hotel Reid's foi igualmente desconsiderado.

3.3.3.3 Baía Azul Hotel

O sensor de Água no Hotel Baía Azul registou um valor constante entre *2011-05-29 14:00:00* e *2012-03-18 07:00:00*. Excluindo as observações de valor constante de análise é possível manter 69% do volume dos dados de leituras sensoriais. No entanto, a conjunção do conjunto de dados filtrado com os dados de ocupação do Hotel Baía Azul faz com que apenas 24% do volume dos dados de leituras de consumo de serviço possam ser utilizadas.

Supondo a melhor hipótese, de que as observações são temporalmente contínuas, o conjunto representa 240 dias o que corresponde a pouco mais que meio período anual.

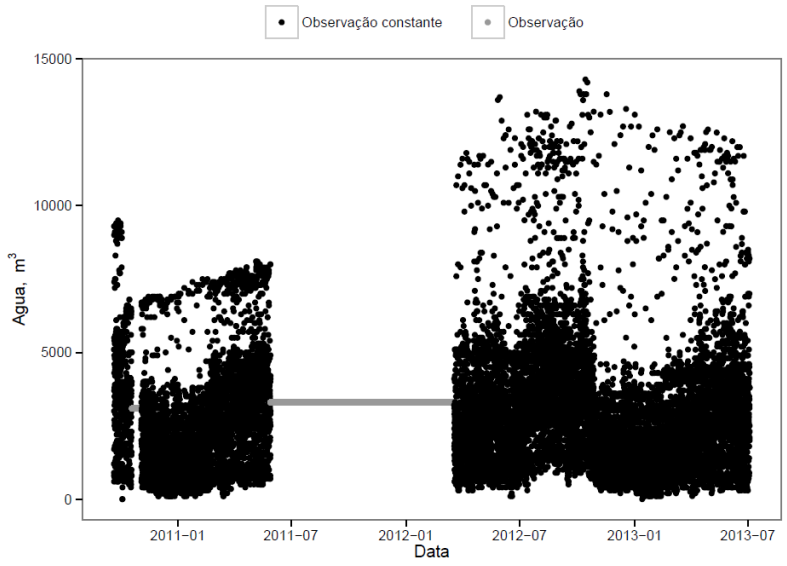


Figura 3.15: Distribuição de valores do sensor de Água no Hotel Baía Azul

Com base no escasso volume de dados o conjunto de dados referente ao Hotel Baía Azul foi excluídos de análise.

3.3.3.4 Alpino Atlântico Hotel

As distribuições de valores das leituras dos sensores de consumo são, sem exceção, assimétricas positivas. Como tal, foi aplicada a escala logarítmica, $log(n+1)$, para a visualização da distribuição de valores das variáveis. Na Figura 3.16 são apresentadas as distribuições de frequência dos sensores de consumo instalados na unidade hoteleira, por estação do ano, juntamente com a distribuição de frequência dos valores de temperatura ambiente.

A distribuição de valores do sensor de água é bastante similar entre as diferentes estações do ano. Denota-se no entanto um pico de frequência de valores nas leituras cujo consumo é igual a zero. O consumo de eletricidade é distinto entre as estações do ano, sendo os valores de consumo mais elevados durante as estações de Verão e de Inverno. A distribuição do consumo de Gás apresenta alguns valores extremos mínimos, no entanto, a variação de consumo entre estações do ano é praticamente inexistente. A distribuição dos valores de temperatura respeitam por sua vez a distribuição esperada, valores mais baixos durante o Inverno, valores mais elevados durante o Verão e uma maior variância durante a Primavera e o Outono.

De forma a visualizar-se a relação entre o consumo dos serviços e o número de hóspede presentes na unidade hoteleira foi contabilizado o total do consumo de cada serviço por dia e o valor de temperatura média do respetivo dia, Figura 3.17.

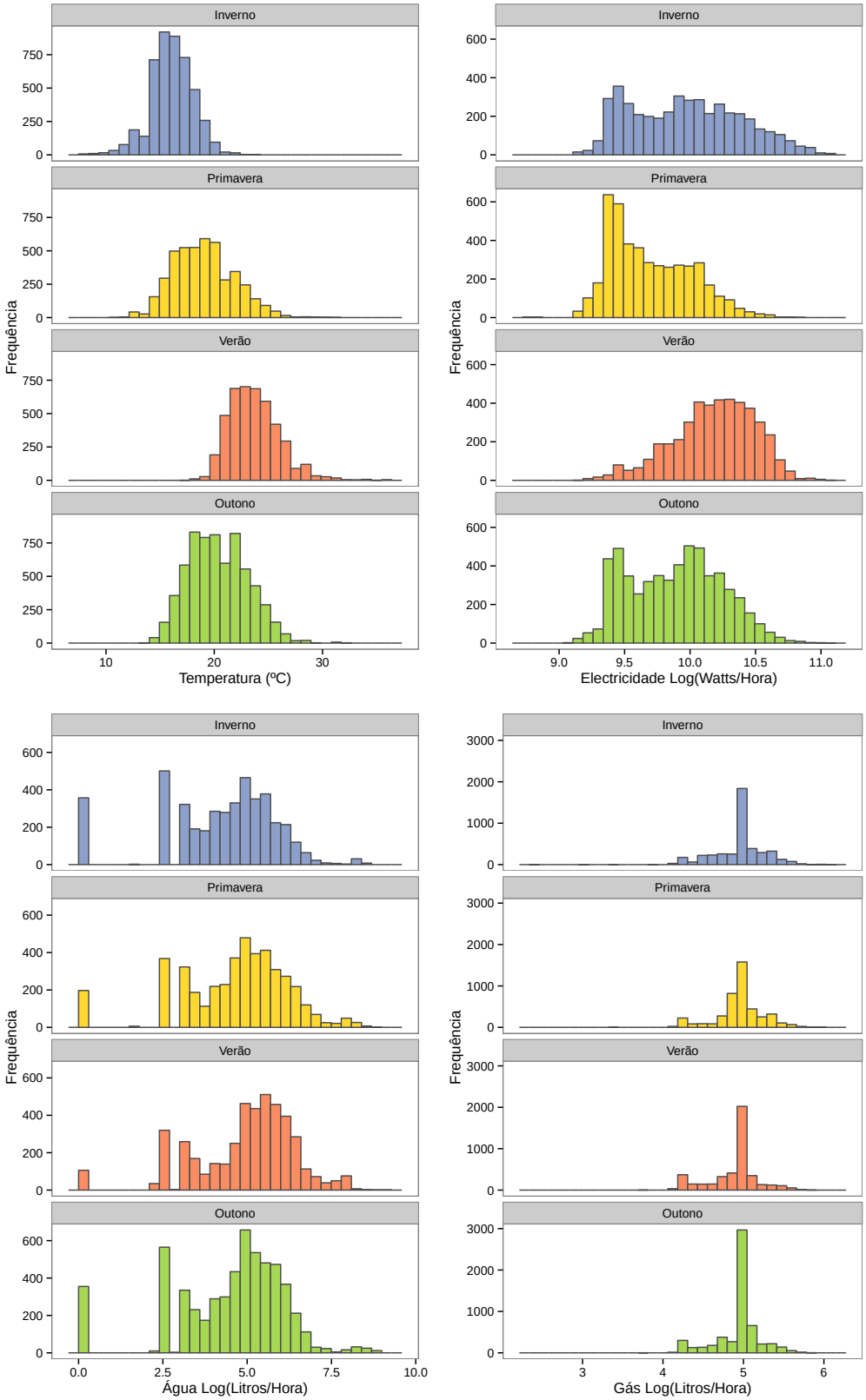


Figura 3.16: Distribuição de valores à escala logarítmica das leituras dos sensores de consumo do Hotel Alpino Atlântico e as leituras de temperatura ambiente por estação do ano

Figura 3.17: Distribuição de valores à escala logarítmica das leituras dos sensores no Hotel Alpino Atlântico por estação do ano

O consumo de eletricidade possui uma correlação moderada com os valores de temperatura ambiente sendo que a correlação entre as estações de Verão e de Outono são moderadamente positivas e no Inverno fracamente negativas. O consumo de eletricidade está também relacionado com o número de hóspedes na unidade hoteleira durante o Outono.

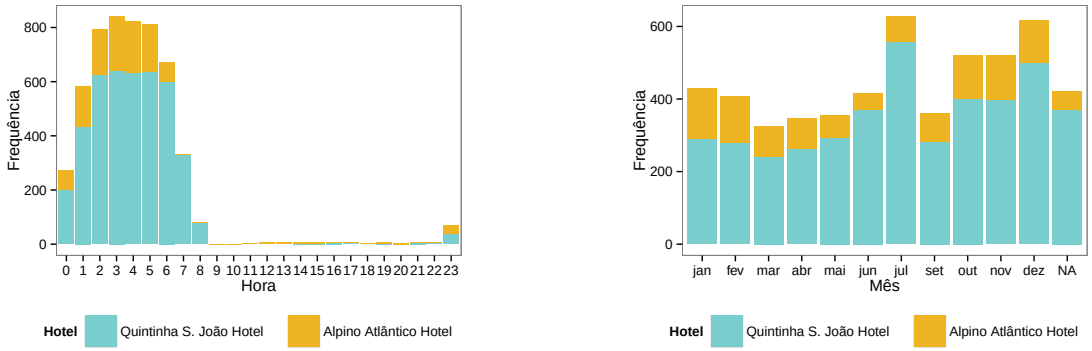
De um modo geral, o consumo de Água encontra-se moderadamente relacionado com o número de hóspedes, especialmente nos meses de Inverno e Primavera, e moderadamente relacionado com o consumo de Gás durante o Inverno.

O consumo de gás encontra-se de certa forma relacionado com o número de hóspedes presentes na unidade mas apenas durante o Inverno. Por sua vez o número de hóspedes encontra-se moderadamente relacionado com os valores de temperatura ambiente, algo que já se suspeitava aquando da visualização da Figura 3.7 que demonstrava valores mais elevados de ocupação durante as estações de Verão e valores mais baixos durante o Inverno.

3.3.3.5 Resumo da Análise Exploratória

A análise exploratória às unidades hoteleiras Quinta das Vistas e Quinta de S.João foi bastante semelhante à análise da unidade Alpino Atlântico. A distribuição de valores entre as diferentes estações do ano foram bastante semelhantes, porém no Hotel Quinta de S.João registou-se uma pequena variância no consumo de gás entre estações, algo que não se verificou nos outros dois hotéis. As relações entre os consumos totais diários, os valores médios de temperatura ambiente e o número de hóspedes foi igualmente semelhante, no entanto, os fatores de correlações entre o consumo de eletricidade e as restantes variáveis registaram valores moderadamente mais elevados, negativamente e positivamente. A síntese da análise exploratória dos Hotéis Quinta das Vistas e Quinta de S.João encontram-se em Apêndice B .

Algo que ocorreu com alguma frequência com os conjuntos de dados do projeto Soltgest foi a identificação de leituras de consumo de água de valor nulo. A Figura 3.18 apresenta a frequência do sensor de Água aquando de valor zero nos Hotéis Alpino Atlântico e Quinta de S.João. Em ambos os Hotéis, a maioria das leituras que registam um valor nulo ocorrem durante o período noturno, respetivamente entre as 23 e as 9 horas, Figura 3.18(a). Este comportamento verificou-se durante todo o período com ligeiras distinções entre os diferentes meses do ano, Figura B.10.



(a) Frequência por Hora

(b) Frequência por Mês

Figura 3.18: Frequência de valores do sensor de Água aquando de valor zero nos Hotéis Alpino Atlântico e Quinta de S.João

No Hotel Quinta das Vistas destaca-se a forte correlação entre o consumo de Água e Eletricidade, com alguma distinção entre as diferentes estações do ano. Destaca-se igualmente o elevado número de valores idênticos a zero no consumo de Gás durante, e quase exclusivamente, o Verão.

No Hotel Quinta das Vistas ocorrem diversas leituras de consumo de gás de valor nulo. Os períodos em que foram registadas essas mesmas leituras não são distinguíveis a nível horário ou por diferentes dias da semana, no entanto, os momentos em que as leituras ocorreram são separáveis por mês ao longo de todo o período mas sem qualquer relação

aparente, Figura 3.19.

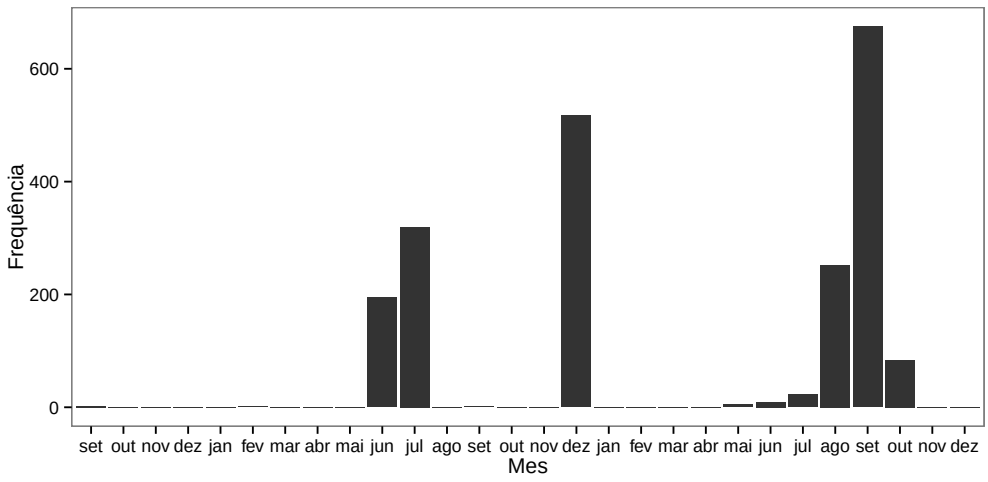


Figura 3.19: Frequência de valores por Mês do sensor de Gás aquando de valor zero no Hotel Quinta das Vistas

Em suma, os hotéis que prosseguiram o estudo após a análise exploratória foram os Hotéis Alpino Atlântico, Quinta das Vistas e Quinta de S.João.

CAPÍTULO 4

PREPARAÇÃO DOS DADOS

Neste capítulo foram elaborados e definidos os conjuntos de dados que foram utilizados para modelar os algoritmos de aprendizagem. A etapa compreendeu funções tais como a definição de subconjuntos de dados, a fusão de conjuntos de dados, a definição de novos atributos, o tratamento de valores considerados anormais e a separação dos conjuntos de dados em sub-conjuntos para efeitos de treino, validação e teste.

O capítulo está dividido em quatro secções. A Secção 4.1 apresenta sucintamente os conjuntos de dados disponíveis para análise e a sua origem. As Secções 4.2 e 4.3 apresentam o processo de preparação dos conjuntos de dados, sob uma granularidade diária e sob uma granularidade horária respetivamente. A Secção 4.4 secção refere como foram definidos os conjuntos de treino, validação e teste.

4.1 Contexto

A análise de dados incide sobre os dados recolhidos pela plataforma Soltgest para três hotéis na RAM, o Hotel Alpino Atlântico, o Hotel Quinta das Vistas e o Hotel Quinta de S. João. Os dados recolhidos contêm leituras horárias de sistemas de sensores consumo de serviços, i.e., leituras de consumo geral de água, eletricidade e gás às quais foram adicionadas as leituras horárias das condições meteorológicas no local, registadas pela estação meteorológica instalada no aeroporto do Funchal. Os conjuntos de dados possuem também os registos diários do número de hóspedes e de quartos de hóspedes que foram respetivamente ocupados em cada unidade hoteleira.

4.2 Granularidade Diária

A caracterização dos consumos de serviços por hóspede, i.e., a identificação de perfis de consumo de serviços para cada unidade hoteleira teve por base a transformação dos conjuntos de dados, dos sistemas de sensores e da estação meteorológica, de uma granularidade horária para uma granularidade diária, idêntica à granularidade da informação relativa à ocupação nos estabelecimentos.

4.2.1 Definição do conjunto de dados

A informação relativa ao número de hóspedes e respetivo número de quartos ocupados é definida por uma granularidade diária, no entanto, todos os restantes dados assentam sobre uma granularidade horária. De forma a representar o comportamento das leituras horárias, de consumos de serviços e condições meteorológicas, perante uma granularidade diária, optou-se por utilizar sumários estatísticos de cada janela temporal diária. Para tal, as janelas temporais que não possuíam 24 leituras horárias foram consideradas inválidas dada a impossibilidade de realizar sumários estatísticos comparáveis entre conjuntos. A Figura 4.1 ilustra os conjuntos de dados presentes na plataforma Soltgest das três unidades hoteleiras em questão. Para cada unidade hoteleira são apresentados os dados referentes aos sensores de consumo e aos registos de ocupação.

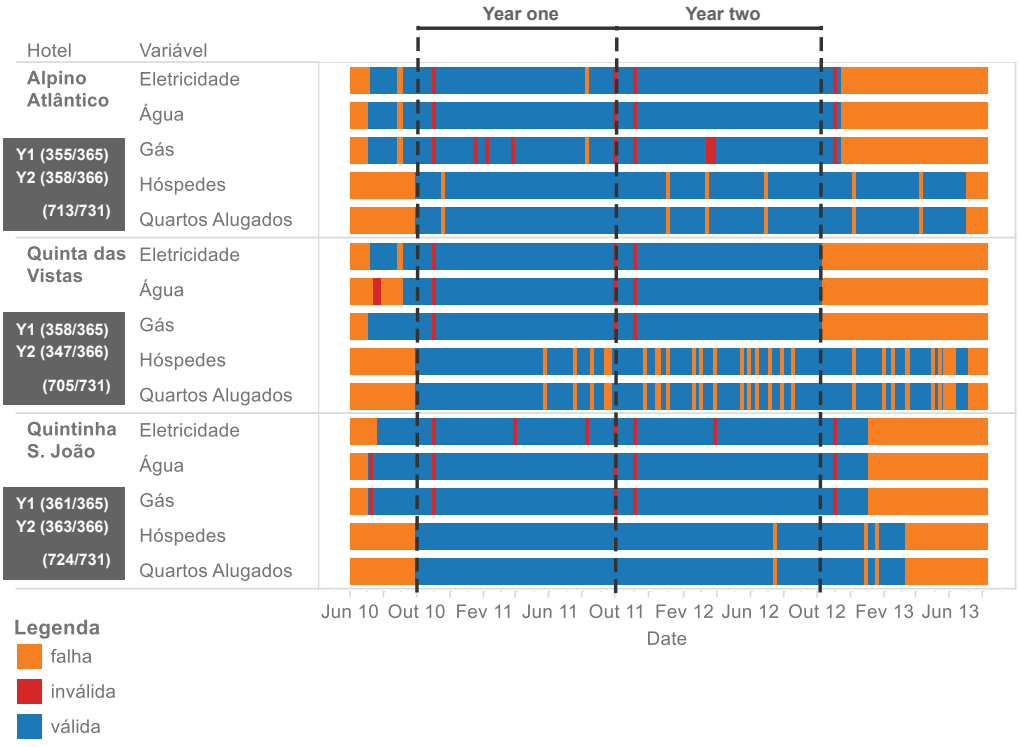


Figura 4.1: Volume de Dados

Na sua totalidade, os conjuntos de dados das três unidades hoteleiras estão compreendidos no período temporal entre Junho de 2010 e Junho de 2013. Cada observação representa uma janela temporal de 24 horas e encontram-se categorizadas por cor, ilustrando a distinção entre janelas temporais válidas, inválidas ou em falta, sendo que a última refere-se às janelas temporais sem qualquer observação, i.e., aos dias para os quais não houve qualquer registo de leituras para o respetivo sensor. Por sua vez as observações dos registos de ocupação são apenas distinguidas pela sua presença.

De forma a captar o maior volume de dados, o mais contínuo possível, com janelas temporais idênticas entre unidades hoteleiras e de forma a obter, no mínimo, os comportamentos sazonais anuais já observados, foram definidos três conjuntos de dados distintos.

Os dois primeiros conjuntos definem duas janelas temporais anuais distintas, identificadas como *Y1* e *Y2* na Figura 4.1, compreendendo os períodos entre Outubro 2010 a Outubro 2011 e Outubro 2011 a Outubro 2012 respetivamente. O terceiro conjunto de dados recolhido para cada unidade hoteleira, designado por *Y1&2*, foi definido como a união dos dois subconjuntos previamente referidos, *Y1* e *Y2* de cada unidade hoteleira, constituindo assim um conjunto de dados com um período temporal praticamente contínuo de dois anos.

A cada um dos novos conjuntos de dados definidos foram apenas incluídas as observações consideradas válidas. A dimensão de cada conjunto de dados resultante e por unidade hoteleira é apresentado na primeira coluna da Figura 4.1, onde é possível verificar que os conjuntos de dados definidos possuem praticamente a mesma dimensão que a dimensão ótima num cenário onde todas as observações seriam válidas.

Após a definição dos novos conjuntos de dados e visando a caracterização de consumo de serviços por hóspede, os valores das observações dos sensores de consumo de serviço foram por sua vez divididos pelo número de hóspedes presentes diariamente em cada unidade hoteleira, de forma a que, os registos de consumo de serviços passassem a ser observados como consumo de serviços por hóspede.

A primeira análise às leituras registadas pelos sensores de consumo revelou diversas leituras de valor zero na recolha dos valores mínimos diários em cada sensor, especialmente nos responsáveis pelo consumo de água e gás. Por sua vez, os valores máximos diários apresentavam um elevado número de valores extremos. Com o intuito de representar o comportamento das observações dos sensores de consumo em cada janela temporal diária foram utilizados sumários estatísticos compostos exclusivamente pelo percentil 25, valor médio e percentil 75 de forma a obter uma maior robustez aos valores extremos registados pelos valores mínimos e máximos.

Aos conjuntos de dados compostos pelos registos dos sensores de consumo e registos de ocupação, sob granularidade diária, foram adicionados os registos da estação meteorológica. A cada janela temporal diária do conjunto de dados das condições meteorológicas foram aplicados os mesmos sumários estatísticos, percentil 25, valor médio e percentil 75, de forma representar o comportamento do valor das observações ao longo de cada dia.

Tabela 4.1: Variáveis dos Conjuntos de Dados

Tipo	Designação	Unidade	Variável
Numérico Discreto	Hóspedes	Hóspede	H
	Quartos Alugados	Quarto	Q
Numérico Contínuo	Eletricidade	Watts p/Dia p/Hóspede	E P25
			E M
			E P75
	Gás	Litros p/Dia p/Hóspede	G P25
			G M
			G P75
	Água	Litros p/Dia p/Hóspede	A P25
			A M
			A P75
	Temperature	Graus Celsius p/Dia	T P25
			T M
			T P75

Os conjuntos de dados constituídos com granularidade diária encontram-se descritos na Tabela 4.4.

4.2.2 Pre-Processamento de dados

A maioria dos algoritmos de agrupamento assentam sobre a noção de similaridade geométrica ou estatística para a identificação de padrões em conjuntos de dados. Uma das medidas mais usais para a definição de similaridade entre observações baseia-se na distância Euclidiana entre pontos. No entanto, dada a perspetiva espacial da distância Euclidiana, os resultados obtidos por estes algoritmos estão fortemente relacionado com a ordem de grandeza e unidade de medida de cada variável do conjunto de dados. De forma a adaptar os valores das variáveis dos conjuntos de dados aos algoritmos de agrupamentos são aplicadas técnicas de escalonamento e a escolha da técnica mais adequada a aplicar é ainda hoje objeto de estudo.

4.2.2.1 Escalonamento

No estudo Milligan e Cooper (1987) foram testadas e validadas diversas técnicas de escalonamento entre as quais as técnicas *Range*, *MinMax* e *Z-Score* (Equações 4.1,4.9,4.10). Durante o estudo foram gerados 864 conjuntos de dados com diferentes características, de forma exclusiva e combinada, tais como: diferentes números de agrupamentos; diferentes distâncias entre os valores médios de cada agrupamento; diferentes distâncias entre o valor médio de cada agrupamento a todos os valores do agrupamento respetivo; a inclusão de valores extremos; a inclusão de ruído; a deformação da estrutura dos agrupamentos. Para a identificação de agrupamentos nos conjuntos de dados gerados foram utilizados algoritmos de agrupamento hierárquicos com diferentes métodos de agrupamento e com base na distância Euclidiana como medida de similaridade entre pontos. O estudo

foi validado perante a verdadeira estrutura de agrupamentos nos conjuntos de dados gerados versus os agrupamentos obtidos pelos algoritmos de agrupamento. Conclui-se em Milligan e Cooper (1987) que o escalonamento por *Range*, Equação 4.1, foi a técnica de escalamento que possibilitou a obtenção melhores resultados e a mais robusta perante as diversidades dos conjuntos de dados.

$$Range = \frac{(x)}{max(x) - min(x)} \tag{4.1}$$

O estudo foi estendido em Steinley (2004) para algoritmos de agrupamento por partição, mais precisamente para a família de algoritmos *K-Means*, procurando o mesmo método de estudo que em Milligan e Cooper (1987). A opinião relativa à recomendação do uso da técnica de escalonamento por *Range* foi reforçada ao apresentar os melhores resultados e uma maior robustez perante uma diversidade de conjuntos de dados com diferentes características.

No entanto, os estudos Milligan e Cooper (1987) e Steinley (2004) foram realizados com conjuntos de dados gerados com variáveis uniformemente distribuídas, às quais foram adicionadas as diversas condições de erro referidas. Ambos os estudos foram igualmente validados perante uma certeza, isto é, perante a verdadeira estrutura de agrupamentos em cada conjunto de dados gerado. No âmbito da Dissertação e do projeto *Soltgest* pretende-se determinar uma possível estrutura de agrupamentos nos conjuntos de dados mas sem a presença de uma verdadeira estrutura para validação. Como tal, a identificação de uma possível estrutura nos conjuntos de dados baseou-se nos princípios de similaridade e dissimilariedade, dentro e entre agrupamentos respetivamente, e a sua semântica analisada no Capítulo 5.

De forma a identificar a técnica de escalonamento a adotar para os conjuntos de dados do projeto Soltgest foi utilizado o índice estatístico *Calinski and Harabasz* Caliński e Harabasz (1974), Equação 4.2, comparado em Milligan e Cooper (1985) com 30 outros índices estatísticos, onde revelou ser o índice que melhor indica o número mais adequado de agrupamentos presentes nos conjuntos de dados. Por sua vez, o índice tem a particularidade de relativizar a dispersão entre agrupamentos e a similaridade nos agrupamentos sendo assim livre de escala tornando-o comparável entre diferentes técnicas de escalonamento.

$$CalinskiandHarabasz = \frac{trace(B)}{trace(W)} * \frac{N-k}{k-1} \tag{4.2}$$

A função *trace* representa o somatório dos valores da diagonal principal de uma matriz. Considerando k como o número de agrupamentos definidos num conjuntos de dados e N o número de instâncias do conjunto, as matrizes B e W representam, respetivamente, a matriz de dispersão dos centros de massa dos agrupamentos ao centro de massa do conjunto de dados e a dispersão das instâncias aos centros de massa dos agrupamentos definidos, Equações 4.4, 4.5, 4.3.

$$B = \sum_{l=1}^{k} N_l(\overline{x_l} - \overline{x})(\overline{x_l} - \overline{x})^T \tag{4.3}$$

$$W = \sum_{l=1}^{k} W_l \tag{4.4}$$

$$W_l = \sum_{x_i \in C_l} (x_i - \overline{x_l})(x_i - \overline{x_l})^T \tag{4.5}$$

A função *trace(W)* é igualmente utilizada para identificação do número adequado de agrupamentos num conjunto de dados. Conhecido por *TraceW* ou *WSS*(*Within Sum Squares*), Equação 4.6, o índice representa o somatório da dispersão em cada agrupamento definido (k agrupamentos definidos), sendo que a dispersão em cada agrupamento é definida pelo somatório da distância, por variável, de cada observação ao centro de massa do agrupamento respetivo. Na presença de n variáveis, o valor de *TraceW* é definido pelo somatório das distâncias obtidas por variável.

$$TraceW = WSS = trace(W) = \sum_{l=1}^{k} trace(W_l) \tag{4.6}$$

$$trace(W_l) = \sum_{p=1}^{n} \sum_{x_i \in C_l} (x_{ip} - \overline{x_{lp}})^2 \tag{4.7}$$

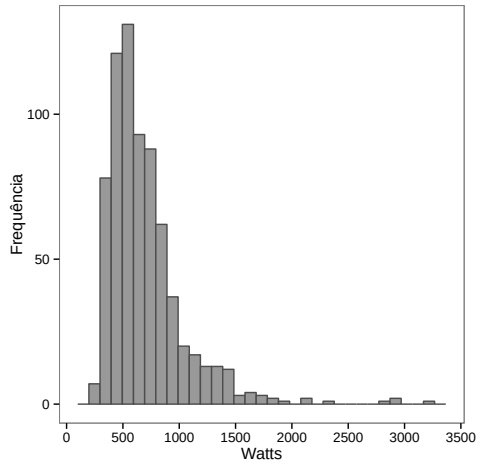
Ao considerarmos apenas um agrupamento, nomeadamente $k = 1$ obtém-se a dispersão total do conjunto de dados ao seu centro de massa, usualmente referido como *TSS*(*Total Sum Squares*), Equação 4.8.

$$TSS = \sum_{p=1}^{n} \sum_{i=1}^{N} (x_{ip} - \overline{x_p})^2 \tag{4.8}$$

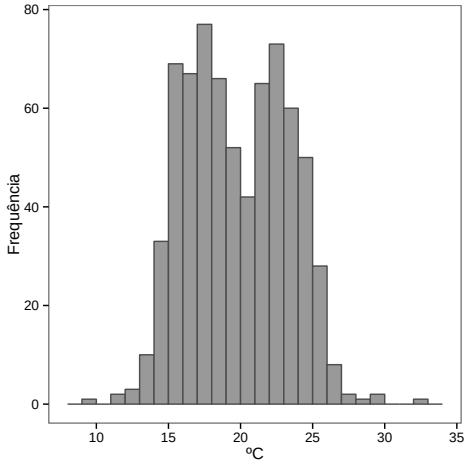
Pela definição da distância Euclidiana, a utilização de diferentes técnicas de escalonamento reflete-se nos diferentes valores que as variáveis assumem, e por sua vez, no valor da dispersão total do conjunto de dados. A consequente dispersão entre agrupamentos após a modelação de algoritmos de agrupamento não é equiparável entre diferentes técnicas de escalonamento. Para possibilitar a comparação entre diferentes técnicas de escalonamento foi necessário definir um termo de comparação que se adapte ao processo de escalonamento. O índice *Calinski and Harabasz* permite essa comparação visto que a razão calculada é livre de escala, possibilitando a comparação entre técnicas de escalonamento.

Quanto maior o valor de *Calinski and Harabasz* maior é similaridade dos agrupamentos e maior a dispersão entre agrupamentos.

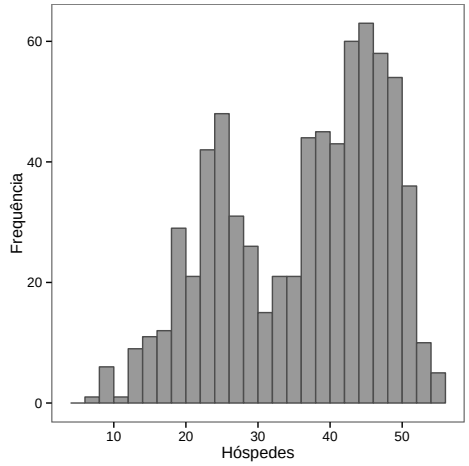
Antes da aplicação de técnicas de escalonamento foram analisadas as distribuições de valores das variáveis dos conjuntos de dados. Os conjuntos de dados definidos na Secção 4.2.1 possuem variáveis com ordens de grandeza e unidades de medida diferentes entre si.



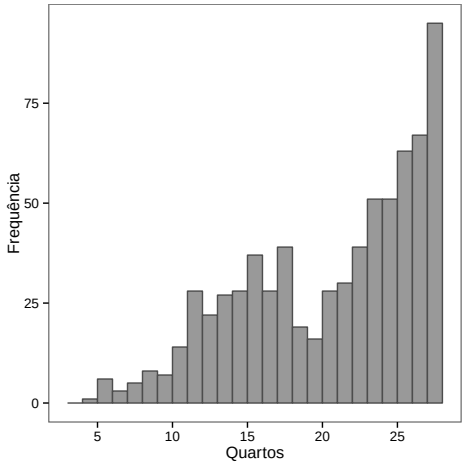
(a) Valores originais de E M



(b) Valores originais de T M



(c) Valores originais de H

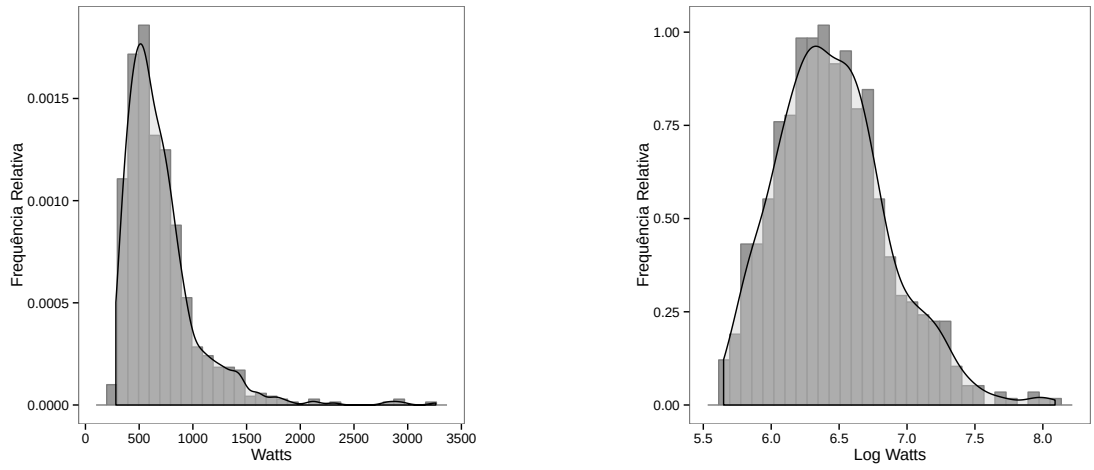


(d) Valores originais de Q

Figura 4.2: Distribuição de frequência de um sub-grupo de variáveis do conjunto de dados Alpino *Y1&2*

A Figura 4.2 apresenta as diferentes distribuições e ordens de grandeza de um grupo selecionado de variáveis do conjunto de dados Alpino *Y1&2*. A distribuição de valores da variável *E M*, consumo médio de eletricidade por hóspede e por dia, é representativa da distribuição de frequência das restantes variáveis de registo de consumos, i.e, tanto dos registos de consumo médio de água e gás como dos percentis recolhidos. Da mesma forma, a distribuição de valores do valor médio de temperatura *T M* é representativa da distribuição de frequência dos percentis de temperatura.

As variáveis *H* e *Q*, referentes ao número de hóspedes e quartos ocupados, possuem uma distribuição tendencialmente assimétrica negativa. Por sua vez, a distribuição de valores da variável *T M* aproxima-se de uma distribuição simétrica. A frequência relativa da variável *E M* é apresentada na Figura 4.2(a) sobreposta por uma função de densidade ajustada à distribuição. A distribuição de valores da variável *E M* e restantes possuem uma assimetria positiva bastante acentuada, Figura 4.2(a), apresentando diversos valores extremos, um cenário refletido em todas as variáveis relativas aos sensores de consumo. Os algoritmos de agrupamento são especialmente sensíveis à presença de valores extremos nos conjuntos de dados dada a perspetiva espacial da distância Euclidiana. De forma a mitigar esta situação foi aplicada a escala logarítmica às variáveis dos sensores de consumo, i.e. à variável *E M* e consequentes, de forma a suavizar a dispersão dos valores extremos. A transformação logarítmica da variável *E M* é apresentada na Figura 5.1(a) onde é possível verificar que com a transformação logarítmica, Figura 4.3(b), a distribuição de valores ficou mais equilibrada e com uma menor dispersão espacial dos valores observados.



(a) Valores originais de E M

(b) Valores logarítmicos de E M

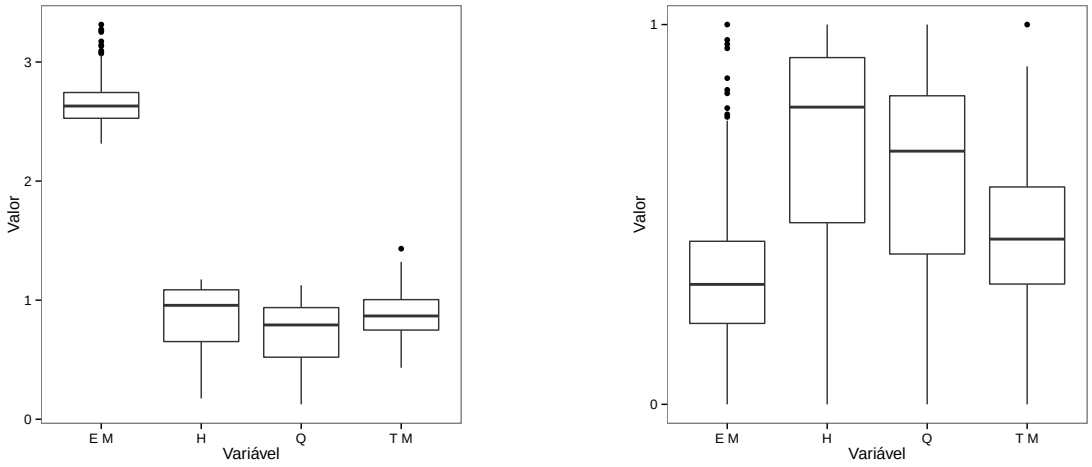
Figura 4.3: Valores originais e logarítmicos do consumo médio diário de eletricidade no conjunto de dados Alpino *Y1&2*

Os métodos de escalonamento estudados foram os métodos *Range*, *Min-Max*, *Z-score* e um último definido por um escalonamento com origem na mediana e com escala definida pelo intervalo entre quartis.

O método de escalonamento por *Range* produz um mapeamento linear dos valores das observações com base nos valores mínimos e máximos de cada variável. Este mapeamento permite escalar o valor das observações para um intervalo quase semelhante entre variáveis, mas não permite a uniformização do valor médio ou da variância entre variáveis, tornando-se menos interpretável. O método de escalonamento *MinMax*,
Equação 4.9, produz um mapeamento linear semelhante ao do método *Range* mas promove o mapeamento para um intervalo entre *[0,1]*, garantindo pelo menos uma instância em cada uma das extremidades. Da mesma forma, o escalonamento por *MinMax* não permite uniformizar o valor médio ou variância entre variáveis.

$$MinMax = \frac{(x - min(x))}{max(x) - min(x)} \tag{4.9}$$

A escalonamento das variáveis por *Range* é apresentada na Figura 4.4(a) e por *Min-Max* na Figura 4.4(b).



(a) Escalonamento por *Range*

(b) Escalonamento por *Min-Max*

Figura 4.4: Escalonamento do conjunto de dados Alpino *Y1&2*

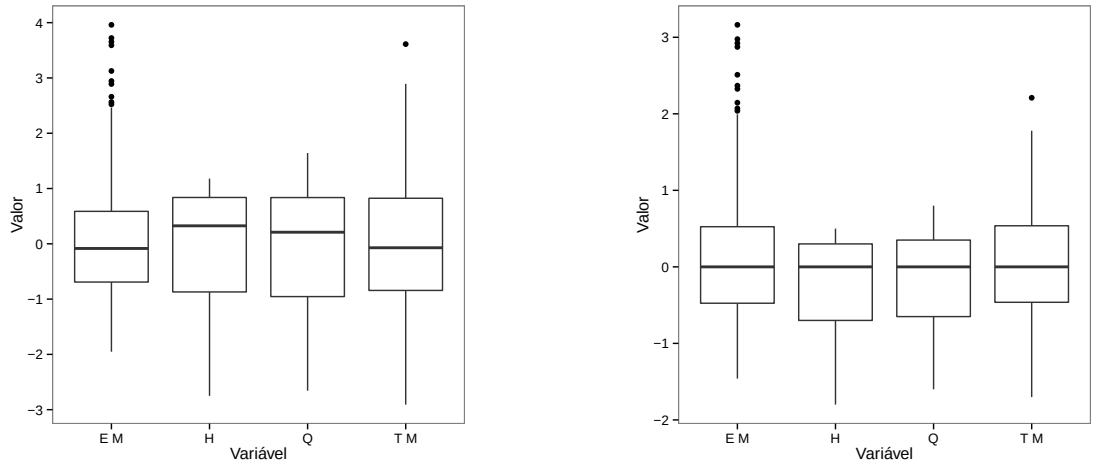
O método *Z-Score*, Equação 4.10, é o método mais clássico das técnicas de escalonamento e produz uma uniformização do valor médio e do desvio padrão entre variáveis, transformando-os para os valores 0 e 1 respetivamente. No entanto o método não garante que o domínio das observações entre variáveis seja idêntico. O escalamento por *Z-score* assume uma distribuição Gaussiana das variáveis e dada a distribuição de valores das observações dos sensores de consumo, optou-se também por estudar uma variante à técnica *Z-score*, apresentada na Equação 4.11, onde são utilizadas medidas mais robustas como origem e escala. Na Equação 4.11 é utilizada a mediana como medida de tendência central ou origem e uma escala definida pelo intervalo entre quartis.

Desta forma, é esperado que a mediana de cada variável assuma o valor zero e que os percentis 25 e 75 assumam os valores entre $[\frac{-1}{2}, \frac{1}{2}]$, consoante a distribuição de valores da variável.

$$Z - Score = \frac{(x - \mu)}{\sigma} \tag{4.10}$$

$$Median/IQR = \frac{(x - median)}{IQR(x)} \tag{4.11}$$

Ao grupo selecionado de variáveis apresentadas na Figura 4.2 foram aplicadas as técnicas de escalonamento *Z-score* e *Median/IQR*, e a distribuição de valores escalados apresentados na Figura 4.5. A variante *Median/IQR* possibilitou obter uma distribuição mais uniforme entre os valores das variáveis.



(a) Escalonamento por *Z-Score*

(b) Escalonamento por Mediana e Intervalo entre Quartis

Figura 4.5: Escalonamento do conjunto de dados Alpino *Y1&2*

Para validar as técnicas de escalonamento apresentadas foram aplicados algoritmos de agrupamentos por partição e hierárquicos aos conjuntos de dados *Y1&2* de cada unidade hoteleira, apresentados mais detalhadamente no Capítulo 5 visto não ser o foco deste capítulo. Na Tabela C.2 são apresentados os valores do índice *Calinski and Harabasz* obtidos com cada um dos modelos gerados com o conjunto de dados *Y1&2* da unidade hoteleira Alpino Atlântico.

Como referido, as técnicas de escalonamento *Range* e *MinMax* e as técnicas *Z-Score* e *Median/IQR* produzem mapeamentos de valores semelhantes entre si e aquando da aplicação dos algoritmos de agrupamento perante o índice *Calinski and Harabasz* registaram-se valores idênticos de similaridade entre agrupamentos. As técnicas de escalonamento *Range* ou *Min-Max* foram as técnicas mais promissoras a adotar perante os valores mais elevados de *Calinski and Harabasz*, independentemente do número de agrupamentos parametrizados. Os melhores valores do índice *Calinski and Harabasz*, assinalados a negrito, foram na

Tabela 4.2: Validação de técnicas de escalonamento por *Calinski and Harabasz* para o conjunto de dados Alpino *Y1&2*

Agrup.		Hierarchical Clustering						Partitional Clustering	
	K	Single	Complete	Average	Centroid	Ward.D2	Diana	**K-Means**	PAM
Range & Min-Max	2	15.8	297.5	580.7	48.1	647.2	**662.6**	**662.6**	629.6
	3	11.4	218.5	294.2	25.7	501.2	423.0	**529.4**	523.7
	4	8.0	346.4	211.0	18.1	434.0	449.0	**456.1**	418.2
	5	6.4	342.5	239.3	180.9	363.9	353.4	381.4	**388.7**
	6	5.5	286.8	212.7	145.2	326.9	313.3	**359.7**	344.3
Z-Score & Median/IQR	2	15.3	410.4	30.2	3.5	427.1	453.8	454.0	438.6
	3	8.8	263.4	16.3	3.1	348.8	287.2	396.0	386.2
	4	6.3	187.8	154.8	2.8	286.1	317.1	325.2	319.5
	5	5.6	214.7	118.3	2.7	248.1	258.3	275.7	279.5
	6	4.6	179.0	96.0	2.5	219.7	232.9	253.0	243.6

sua maioria obtidos com algoritmo *K-Means*. A diferença de valores do índice *Calinski and Harabasz* entre alguns algoritmos é substancial. Contudo essa análise será realizada no Capítulo 5. Os resultados foram comuns entre as diferentes unidades hoteleiras, apresentados em Apêndice C, e por consenso nos resultados obtidos foi adotada a técnica de escalonamento *Range* para todos conjuntos de dados do projeto Soltgest.

Após uma análise mais pormenorizada aos valores das variáveis após o escalonamento por *Range* verificou-se que, mesmo aplicando previamente a escala logarítmica, era presente a existência de diversos valores extremos. Como referido, dada a sensibilidade dos algoritmos de agrupamento a estes valores extremos, foi tomada a opção de retirar as instâncias cujos valores em qualquer uma das variáveis fosse superior ou inferior a $3 * IQR$ (intervalo entre quartis respetivo), valor este usualmente considerado como a fronteira para os valores caracterizados como extremos. De forma a manter o equilíbrio e a capacidade de comparação entre os conjuntos de dados *Y1&2* e os conjuntos *Y1* e *Y2* foi tomada a decisão de excluir apenas as instâncias cujos valores em qualquer uma das variáveis fosse superior ou inferior a $3 * IQR$ no conjunto de dados *Y1&2*, garantindo assim que $Y1\&2 = Y1 \bigcup Y2$ e que apenas os valores considerados extremos no conjunto dos dois períodos anuais fossem excluídos.

Após o pré-processamento dos conjuntos de dados, o volume corresponde a cada um é apresentado na Tabela 4.3.

Tabela 4.3: Volume dos conjuntos de dados originais e após pré-processamento

Alpino Atlântico	Y1	Y2	Y1&2
Conjunto Original	354	358	713
Conjunto Pré-Processado	339	356	695
Volume Preservado	95.8%	99.0%	97.4%

Quinta das Vistas	Y1	Y2	Y1&2
Conjunto Original	357	346	703
Conjunto Pré-Processado	345	339	684
Volume Preservado	96.6%	99%	97.2%

Quinta S.João	Y1	Y2	Y1&2
Conjunto Original	360	362	722
Conjunto Pré-Processado	346	352	698
Volume Preservado	96.1%	97.2%	96.6%

4.3 Granularidade Horária

Para a análise de estimativa do consumo de serviços por hóspede em cada unidade hoteleira foram utilizados os conjuntos de dados com uma granularidade horária.

4.3.1 Definição do conjunto de dados

A construção dos conjuntos de dados perante uma granularidade horária implicou o produto cartesiano entre o conjunto de dados relativo à ocupação diária na unidade hoteleira com os conjuntos de dados sob uma granularidade horária, relativo aos sensores de consumo de serviços e às condições meteorológicas.

Tabela 4.4: Datasets Features

Tipo	**Designação**	**Unidade**	**Variável**
Numérico	Hóspedes	Hóspede	H
Discreto	Quartos Alugados	Quarto	Q
	Eletricidade	Watts p/Hora p/Hóspede	E
	Gás	Litros p/Hora p/Hóspede	G
	Água	Litros p/Hora p/Hóspede	A
	Temperature	Graus Celsius p/Hora	T

De forma a manter coerência com a análise efetuada sobre uma granularidade diária, selecionaram-se os conjuntos de dados pelo mesmo domínio temporal,i.e., o mesmo período de dois anos definido na Secção 4.2.1, compreendido entre Outubro 2010 e Outubro 2012.

4.3.2 Pre-Processamento de dados

Aquando da falha de um dos registos, quer dos sensores de consumo de serviços ou da estação meteorológica, foi adotada a mesma metodologia que a executada durante o processamento dos conjuntos de dados com uma granularidade diária, ou seja, a instância em causa foi excluída do conjunto de dados. Aquando da falha de um registo de ocupação foram excluídas as 24 instâncias respetivas.

No entanto, durante a análise dos resultados obtidos com as primeiras iterações com algoritmos de regressão, verificou-se que ao invés de estimar a performance do modelo com base no erro quadrado médio horário, a performance poderia ser estimada com base no erro médio diário, semanal ou mensal. A opção de observar o erro perante diferentes janelas temporais foi considerada devido à variabilidade do sinal dos sensores de consumo. Com esta abordagem foi possível atenuar o erro médio dada a compensação que ocorria entre erros de valor negativo e valor positivo dentro da janela temporal observada. Como por exemplo, a média do erro quadrado obtido ao estimar 24 horas consecutivas era superior ao somatório do erro das 24 horas consecutivas ao quadrado. Em suma, a atenuação do erro obtido com os modelos gerados seguiu a Equação 4.12 onde m define o número de instâncias em cada janela de observação.

$$|e_i^2| \geq |\sum_{i=1}^{m} e_i|^2, i \in [1, m] \tag{4.12}$$

De maneira a poder prosseguir com esta abordagem foi necessário atenuar o número de falhas nas leituras dos sensores de consumo que se localizavam espalhados em distintos períodos dos conjunto de dados. A necessidade surgiu visto que, ao adotar a observação do erro perante janelas temporais de maior dimensão, como por exemplo de 1 semana, uma instância em falha dentro de essa mesma janela implicaria a exclusão de todo esse período para análise, i.e., uma leitura horária em falha implicaria a exclusão da janela temporal de dimensão diária ou superior. Dado o volume limitado do conjunto de dados e para evitar a exclusão de janelas temporais optou-se por simular os valores em falta através de interpolação linear entre dois pontos. Aquando da presença de uma falha horária no sinal e na presença de valores anteriores e posteriores, foi simulado o valor em falta através da interpolação linear entre ambos os pontos circundantes. Na presença de falhas contínuas sucessivas manteve-se a opção de remover as instâncias em causa.

4.3.2.1 Escalonamento

Para efeitos de pré-processamento dos dados foi utilizado o escalonamento por *Z-Score*, Equação 4.10, de forma a obter um valor médio e variância comum ente variáveis.

De novo, sem a capacidade de explicar as observações com valores extremos nos sensores de consumo e, com base na sua baixa frequência e elevada influência na modelação dos algoritmos de regressão, optou-se por retirar as instâncias cujos valores em qualquer uma das variáveis em questão fosse superior ou inferior a $3 * IQR$ (intervalo entre quartis),

Tabela 4.5: Volume dos conjuntos de dados originais e após pré-processamento

Alpino Atlântico	Eletricidade	Água	Gás
Conjunto Original	17388	17423	17356
Conjunto Pré-Processado	17405	17440	17373
Percentagem Expectável	99.2%	99.4%	99.0%

Quinta das Vistas	Eletricidade	Água	Gás
Conjunto Original	16967	16960	16967
Conjunto Pré-Processado	16984	16984	16984
Percentagem Expectável	96.8%	96.8%	96.8%

Quinta S.João	Eletricidade	Água	Gás
Conjunto Original	17493	17495	17495
Conjunto Pré-Processado	17508	17512	17512
Percentagem Expectável	99.7%	99.8%	99.8%

valor este usualmente considerado como a fronteira para os valores caracterizados como extremos.

4.4 Conjuntos de Treino, Validação e Teste

Para a análise de perfis de consumo foi utilizado todo o conjunto de dados para a geração dos respetivos modelos de agrupamentos. Para a validação dos modelos gerados foram utilizados índices estatísticos aplicados a todo conjunto de dados e com agrupamentos respetivos.

Para a análise de estimativa de consumo de serviços foram definidos três sub-conjuntos, nomeadamente os conjuntos de treino, validação e teste. O conjunto de treino é utilizado para a geração do modelo, o de validação para estimar o erro dos modelos gerados com o propósito de escolher o melhor dos modelos e o de teste para a determinação do erro de generalização do modelo escolhido. O erro de generalização é o erro obtido por um modelo perante um conjunto de dados independente à sua geração ou à sua seleção entre outros modelos gerados. A constituição do número de observações em cada um dos conjuntos é dependente do volume de dados e no *signal-to-noise ratio*, i.e., o ratio estimado entre o sinal e ruído no conjunto de dados, Hastie et al. (2009), sendo que é usualmente definido um ratio de 50%, 25% e 25% para os conjuntos de treino, validação e teste respetivamente.

No entanto, a utilização da metodologia de treino, validação e teste através de subconjuntos independentes do total dos dados disponíveis implica um cenário onde o volume de dados é elevado. No projeto Soltgest, e após todo o processo de limpeza e validação, foi possível reunir um período total de 2 anos de informação para os 3 Hotéis em análise

onde estão presentes padrões de sazonalidades anuais. Dado o cenário da informação disponível, foi tomada a decisão de utilizar a primeira janela temporal anual, de Outubro 2010 a Outubro 2011, como conjunto de dados para treino e validação e a segunda janela temporal anual, de Outubro 2011 a Outubro 2012, como conjunto de dados para teste. De forma a utilizar todo o conjunto de dados destinado para treino e validação foi utilizada a técnica de validação cruzada, *cross-validation* (*CV*).

A validação cruzada é uma técnica usual para a seleção de modelos que permite utilizar todo um conjunto de dados para treino e teste. A técnica consiste na divisão do conjunto de dados em V partições, com $1 \leq V \leq N$ onde N é o número de instâncias do conjunto de dados, para estimar o risco (por.ex. o erro quadrado médio) de cada algoritmo. De forma iterativa, é selecionada cada uma das partições definidas para formar o novo subconjunto de dados para validação e as restantes partições não selecionadas formam o novo subconjunto de dados para treino. Um dos contributos da técnica de *CV* é permitir identificar o melhor algoritmo através do valor médio da estimativa de risco definida. As diversas estratégias de *CV* variam consoante a estratégia utilizada para a partição do conjunto de dados, Ounpraseuth et al. (2012).

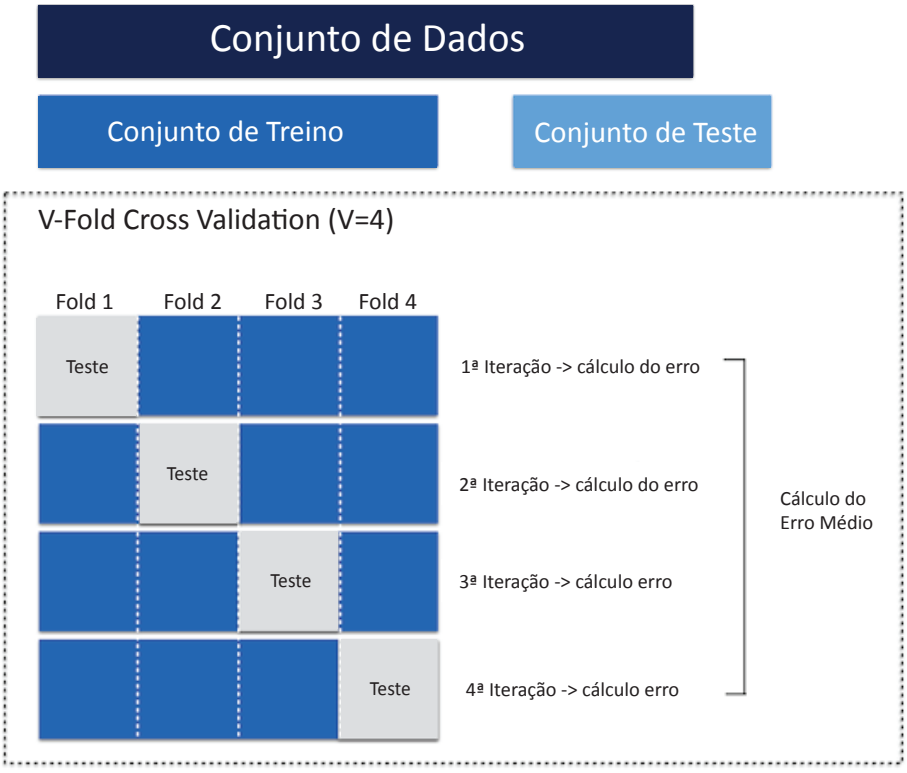


Figura 4.6: Exemplo da técnica *VFCV* para $V = 4$

A família de técnicas da metodologia *CV* é extensa, sendo as técnicas *VFCV* (*V-fold cross-validation*), *LOO* (*Leave-One-Out*) e *RLT* (*Repeated learning-testing*) provavelmente as mais conhecidas. A técnica *VFCV* reparte o conjunto de dados em V partições onde, em cada iterações, uma das partições é definida como conjunto de validação/teste e as

restantes *V-1* são agregadas e definidas como conjunto de treino. Um exemplo da técnica *VFCV* é apresentado na Figura 4.6. A técnica *LOO* é bastante semelhante, mas reparte o conjunto de dados em $n - 1$ partições sendo n correspondente ao volume do conjunto de dados. Iterativamente são agregadas as $n - 1$ partições e definidas como conjunto de treino e a instância restante avaliada como conjunto de validação/teste. A técnica *RLT* é semelhante à técnica *VFCV* no sentido em que a técnica força a repetição de x vezes a técnica *VFCV* sendo que em cada iteração re-ordena o conjunto de dados de forma a que as partições formadas por *VFCV* sejam formadas o mais aleatoriamente possível. Dado o volume de dados do projeto Soltgest utilizou-se a técnica de validação *VFCV*, *V-fold cross-validation*, visto que técnicas como *LOO* ou *RLT* são computacionalmente intensivas e não permitiam executar as experiências necessárias e/ou desejadas em tempo útil. A parametrização de V na técnica *VFCV* é ainda hoje objeto de estudo e apesar de ser recomendado $V = 2$ em Dietterich (1998) e Alpaydm (1999) para a comparação e identificação do modelo candidato entre algoritmos de aprendizagem (com a melhor performance), a sazonalidade anual presente na globalidade do conjunto de dados não permitia um conjunto de validação com uma dimensão tão extensa e, como tal, foi utilizado o parâmetro $V = 10$ sendo também este um dos valores mais recomendados para *VFCV*.

CAPÍTULO 5

IDENTIFICAÇÃO DE PERFIS DE CONSUMO

Neste capítulo é apresentado o estudo dedicado à identificação de perfis de consumo de serviços por hóspede nas unidades hoteleiras Alpino Atlântico, Quinta da Vistas e Quinta de S.João. Seguindo as etapas *Modelling* e *Evaluation* da metodologia CRISP-DM, foram modelados, validados e avaliados algoritmos não supervisionados de agrupamentos para a análise de perfis.

O capítulo está dividido em três secções. A Secção 5.1, referente à secção teórica relativa à identificação de perfis de consumo de serviços, apresenta os algoritmos de agrupamentos estudados e as respetivas técnicas de validação. A Secção 5.2 demonstra a aplicação dos algoritmos de agrupamentos e na Secção 5.3 são discutidos e interpretados os resultados obtidos com as técnicas de agrupamento.

5.1 Algoritmos de Agrupamento

As técnicas não-supervisionadas de agrupamento analisam os conjuntos de dados sem qualquer conhecimento prévio de distinção entre objetos. Os objetos são agrupados com base nos princípios de maximização de similaridade entre objetos da mesma classe e de maximização de dissimilariedade entre classes. Algumas técnicas de agrupamento possibilitam também a organização de objetos perante uma hierarquia de classes. Existem diversos algoritmos de agrupamento e são usualmente distinguidos entre algoritmos por partição e algoritmos hierárquicos. Os algoritmos por partição definem uma decomposição singular do conjunto de dados em k número de agrupamentos enquanto que os algoritmos hierárquicos, de aglomeração e divisórios, definem uma representação hierárquica da decomposição do conjunto de dados em agrupamentos, Everitt (2011); Jiawei Han e Pei (June 2011). Os algoritmos escolhidos para a elaboração de perfis de consumo englobam ambos os géneros, por partição e hierárquicos, e foram previamente mencionados na Tabela C.2. Os algoritmos por partição escolhidos foram os algoritmos *K-Means* e *PAM*, os

algoritmos hierárquicos de aglomeração escolhidos foram com base nas técnicas *Single*, *Complete*, *Average*, *Centroid* e *Ward.D2*, e o algoritmo hierárquico divisório escolhido foi com base na técnica *Diana*.

5.1.1 Algoritmos por Partição

Os algoritmos de agrupamento por partição organizam as instâncias dos conjuntos de dados em conjuntos exclusivos. O número de agrupamentos a formar necessita ser parametrizado à partida e os agrupamentos são formados com base na otimização de um critério, como por ex. a minimização da dispersão das instâncias com base na distância Euclidiana. Assumindo uma parametrização inicial de k agrupamentos, os algoritmos por partição selecionados, *K-Means* e *PAM*, começam por definir k instâncias como as posições iniciais dos centros de massa dos seus agrupamentos, designados por *centróides* e *medóides* respetivamente. De seguida, iterativamente e com base na função de similaridade definida, os algoritmos processam duas etapas, a de atribuição e a de re-ajustamento. No processo de atribuição, os algoritmos atribuem as instâncias ao agrupamento cujo *centróide* ou *medóide* é mais similar. No processo de re-ajustamento os centros de massa são recalculados com base no novo conjunto de instâncias que definem cada agrupamento. O algoritmo *K-Means* re-ajusta os seus centróides para a posição dos recém calculados centros de massa enquanto que o algoritmo *PAM* re-ajusta os seus *medóides* para a posição das instâncias mais similares aos recém calculados centros de massa. Os algoritmos repetem este processo até que não ocorra qualquer re-atribuição de instâncias entre agrupamentos ou até que valide um parâmetro de término pré-definido.

5.1.2 Algoritmos Hierárquicos

Os algoritmos de agrupamento hierárquicos de aglomeração seguem uma estratégia *bottom-up*. Durante o processo de inicialização é atribuído um agrupamento a cada instância e perante um processo iterativo são sucessivamente agregados pares de agrupamentos com base num critério de agrupamento definido, *Single*, *Complete*, *Average*, *Centroid* e *Ward.D2* Murtagh (2014), até um determinado critério de paragem. Os critérios de agrupamento são caracterizados por: o critério *Single* procura agrupar o par de agrupamentos cuja distância entre o par de instâncias mais próximas é mínima; o critério *Complete* procura agrupar o par de agrupamentos cuja distância entre o par de instâncias mais distantes é máxima; o critério *Average* procura agrupar os agrupamentos cuja média da distância entre todos os pontos de cada agrupamento é mínima; o critério *Centroid* procura agrupar os agrupamentos cujos os centros de massa são mais similares; o critério *Ward.D2* procura agrupar os agrupamentos cujo par de agrupamentos fundido aumenta o mínimo do total da dispersão em todos os agrupamentos, Equação 4.4. Cada etapa do processo iterativo de agrupamento é identificada como um nível hierárquico do modelo onde todos os agrupamentos definidos até essa etapa são visíveis e o conjunto ordenado de níveis definem o modelo hierárquico.

Por sua vez, os algoritmos hierárquicos divisórios seguem uma estratégia *top-down*. Durante o processo de inicialização é definido um único agrupamento englobando todas as instâncias do conjunto de dados. Iterativamente e consoante a métrica de desagrupamento definida, é sucessivamente selecionado e dividido um agrupamento em dois sub-conjuntos até um determinado critério de paragem. O critério de divisão *Diana* seleciona sucessivamente o agrupamento com o maior diâmetro para o processo de divisão sendo que o diâmetro de uma agrupamento é definido pelo maior valor de dissimilariedade obtido entre cada uma das suas instâncias e as restantes. A instância com o valor mais elevado de dissimilariedade é separada do agrupamento, usualmente denominado de *old party*, e instância um novo conjunto, usualmente denominado de *splinter group*. Repetidamente, as instâncias são atribuídas a cada um dos conjuntos com base na dissimilariedade com as instâncias de cada conjunto. Aquando da convergência do processo de atribuição é definido um novo nível no modelo onde o agrupamento que sofreu o processo de divisão é repartido em dois novos agrupamentos constituídos pelos conjuntos de instâncias *old party* e *splinter group* respetivamente. Por fim, o conjunto ordenado de níveis formado pelo processo de desagrupamento define o modelo hierárquico.

5.1.3 Processo de Inicialização dos Algoritmos

Os algoritmos de agrupamento por partição são algoritmos especialmente sensíveis à sua inicialização. A definição das instâncias que identificam a inicialização dos *centróides* ou *medóides*, no processo de inicialização dos algoritmos, definem os seus consequentes comportamentos sendo os algoritmos determinísticos após este passo. Uma das abordagens mais usuais para contornar a sensibilidade dos algoritmos por partição ao processo de inicialização é a realização de diversas experiências onde são definidas, aleatoriamente e com base numa distribuição normal, k instâncias do conjunto de dados. Dos modelos gerados é selecionado o que melhor otimiza a função objetivo.

Um método alternativo ao processo de inicialização aleatório do algoritmo *K-Means* é proposto em Arthur e Vassilvitskii (2007) e denominado de *K-Means++*. Este método define os *centróides* através um processo probabilístico e iterativo. A primeira instância é aleatoriamente selecionada como *centróide* inicial e os restantes são iterativamente selecionados de forma aleatória mas com base numa função de probabilidade definida pela dissimilariedade entre as observações e os centroídes previamente escolhidos. Quanto maior for a dissimilariedade entre uma instância e os *centróides* previamente selecionados maior a probabilidade de a instância ser escolhida como próximo *centróide*. Apesar de não ser o foco do estudo, o mesmo princípio foi aplicado ao algoritmo *PAM*, também pertencente à família dos algoritmos de agrupamento por partição.

Para efeitos de distinção, os algoritmos inicializados com o processo aleatório com base numa distribuição normal foram designados por *K-Means* e *PAM* e os inicializados com a técnica referida por *K-Means++* designados por *K-Means++* e *PAM++*. Dado que ambos os processos assumem técnicas de escolha aleatória foram realizadas 100 experiências para

cada técnica e escolhido o modelo que entre as experiências gerou o melhor valor relativo à função objetivo. Em todas as experiências foi parametrizada a permissão dos algoritmos executarem as suas iterações até à conversão da função objetivo.

Os algoritmos hierárquicos escolhidos são por sua vez determinísticos e usualmente não parametrizáveis.

5.1.4 Índices de identificação do Número de Agrupamentos

O processo de validação dos algoritmos de agrupamento tem por base a identificação do número de agrupamentos em que o conjuntos de dados pode ser particionado e que aparenta ser o mais adequado para a representação de distintos padrões.

Apesar da utilização do índice *Calinski and Harabasz* para a comparação entre técnicas de escalonamento, Equação 4.2, o índice de validação *TraceW* é provavelmente o índice mais comum entre todos os índices paramétricos, Equação 4.6.

O índice *TraceW* indica que quanto menor for o seu valor, maior é a coesão entre agrupamentos, no entanto, o aumento do número de agrupamentos formados por qualquer algoritmo usualmente decresce o valor de *TraceW*. No momento de escolha do número mais adequado de agrupamentos é necessário ter em conta o erro de generalização. Quando um modelo é excessivamente ajustado a um conjunto de dados específico o mesmo revela dificuldades em avaliar um novo conjunto de dados dado que não possui a capacidade de generalizar perante dados nunca antes vistos. De forma a evitar o ajustamento excessivo a um conjunto de dados, têm sido propostas algumas técnicas de validação dos índices paramétricos. Nenhuma das técnicas foi no entanto considerada, até hoje, a que identifica corretamente o número de agrupamentos a adotar perante qualquer cenário, índice paramétrico, conjunto de dados ou algoritmo de agrupamentos. As técnicas de validação mais comuns procuram identificar o valor mais adequado com base na curvatura resultante entre os diversos valores do índice, usualmente denominado de *Elbow point*. O ponto *Elbow point* é usualmente definido como o ponto de máxima curvatura da função e é considerado ser o ponto que possibilita o ajuste mais adequado ao conjunto de dados sem que o mesmo seja excessivo. Uma caracterização alternativa ao ponto *Elbow point* é a do ponto k para qual o ponto $k + 1$ não acresce um ganho de informação significativo. Uma das propostas para determinar o ponto *Elbow point* baseia-se na determinação visual através da projeção gráfica dos valores do índice para cada valor de k. Visualmente, procura-se determinar o ponto p para qual o ponto p+1 não apresenta um acréscimo significativo de performance ou ganho de informação. Um outro método proposto define o ponto *Elbow point* através do valor máximo absoluto (ou valor acima de um determinado *threshold*) da segunda derivada entre pontos. Durante o período de pesquisa realizado sobre a determinação do ponto *Elbow point* foi recorrente a recomendação da utilização do método visual em conjunção com o valor máximo absoluto da segunda derivada da função, sendo que a última não possibilita a identificação de uma estrutura com 2 ou menos agrupamentos. Durante o estudo de análise de perfis foram utilizadas as técnicas de visualização e da

segunda derivada da função para a determinação do ponto *Elbow point*.

Nem todos os índices paramétricos baseiam-se na identificação do ponto *Elbow point* para indicar qual o número de agrupamentos mais adequado. O índice *Calinski and Harabasz* sugere o valor obtido mais elevado como o número mais adequado de agrupamentos visto que, quanto maior é o afastamento entre os agrupamentos maior é o valor de *trace(B)* (numerador) e quanto mais compactos forem os agrupamentos menor é o valor de *trace(W)* (denominador).

Por sua vez, o índice *Silhouette* permite identificar o número mais adequado de agrupamentos como também avaliar a coesão de cada instância a cada agrupamento associado, indicando a *força* da similaridade de cada instância ao agrupamento atribuído. O índice é definido pela Equação 5.1, onde $a(i)$ é o valor médio de similaridade entre a instância i e todas as instâncias do mesmo agrupamento, C, e $b(i)$ é o valor médio de similaridade entre a instância i e todas as instâncias do agrupamento mais próximo de i excluindo o agrupamento C. A identificação do valor adequado de k rege-se pela procura do valor máximo do índice de *Silhouette*, $s(i)$.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{5.1}$$

Os índices *TraceW*, *Calinski and Harabasz* e *Silhouette* avaliam os agrupamentos definidos com base no conjunto de dados com que os algoritmos foram modelados. Alternativamente, o índice *Gap Statistic* propõe a análise comparativa entre a coesão dos agrupamentos definidos com o conjunto de dados inicial e a coesão dos agrupamentos obtidos com novos conjuntos de dados simulados. O método usualmente mais utilizado para a simulação dos novos conjuntos de dados usufrui da técnica *PCA/SVD*, *Principal Component Analysis* ou *Single Value Decomposition*, de forma a procurar manter a estrutura do conjunto de dados inicial. A geração dos dados decorre então da simulação de valores aleatórios, com distribuições uniformes, mas contidos no domínio das variáveis do conjunto de dados. O processo completo é descrito por:

Dado um conjunto de dados inicial X, assumindo que todas as variáveis possuem um valor médio de 0, é calculada a decomposição em valores singulares $X = UDV^T$. A matriz $X' = XV$ é calculada e consoante o domínio das colunas da matriz X' são geradas variáveis com distribuições uniformes Z'. Os dados gerados são posteriormente re-transformados para $Z = Z'V^T$.

Dada a aleatoriedade das simulações, o índice propõe a geração de b simulações para cada número de agrupamentos a avaliar. A cada novo conjunto de dados simulado é aplicado o algoritmo de agrupamentos em análise e calculado o valor de *trace(W)*. Das b simulações é calculado o valor médio de *trace(W)*, sob a escala logarítmica, $E[log(TraceW_{bk})]$, ao qual é subtraído o valor de *trace(W)* obtido com o conjunto de dados inicial, também à escala logarítmica $logTraceW_k$. Este processo é repetido para cada valor de k ,i.e., para

cada número de agrupamentos que se pretenda avaliar, e a curva resultante denominada de *Gap Statistic*, definida pela Equação 5.2.

$$Gap = E[log(TraceW_{bk})] - log(TraceW_k) \tag{5.2}$$

Para a validação do ponto mais adequado perante a curva de *Gap Statistic* foi adotado o método proposto em Tibshirani et al. (2001), sendo este considerado como um dos mais consistentes perante conjuntos de dados com diferentes características. O método identifica o valor de k mais adequado como o primeiro valor de Gap_k para o qual o valor de $Gap_{k+1} - s_{k+1}$ é inferior, Equação 5.3. O valor de s'_k corresponde ao desvio padrão dos valores obtidos em $log(TraceW_{bk})$ para cada k durante as b simulações, Equação 5.4.

$$k^* = argmin_k\{k|Gap(k) \geq Gap(k+1) - s'_{k+1}\} \tag{5.3}$$

$$s'_K k = s_k\sqrt{\frac{1+1}{b}} \tag{5.4}$$

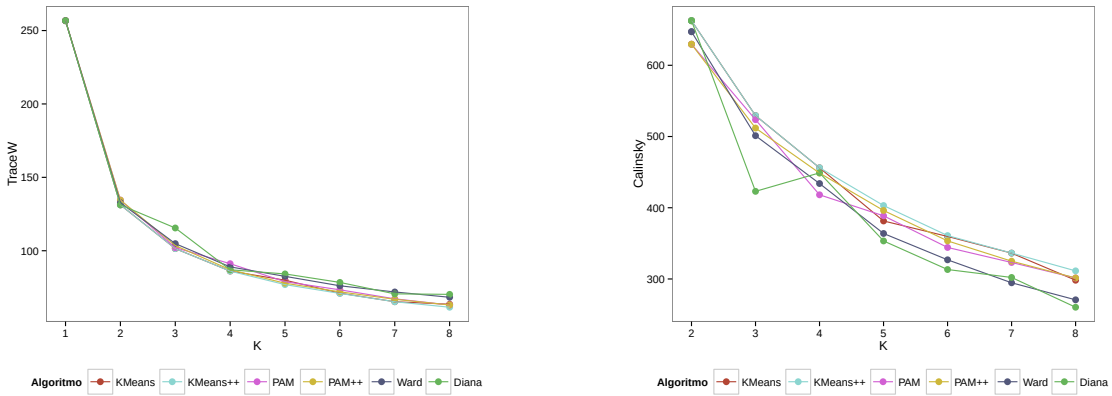
Os índices *TraceW*, *Calinski and Harabasz*, *Silhouette* e *Gap Statistic* foram os índices selecionados como principais pontos de referência para a validação do número mais adequado de agrupamentos a escolher para cada modelo gerado.

5.2 Modelação e Validação de Algoritmos de Agrupamentos

Nesta secção é apresentado o processo de modelação e validação dos algoritmos de agrupamentos e, identificado para cada conjunto de dados, o modelo candidato, i.e., o que aparenta ser o modelo mais adequado. Para a modelação dos algoritmos de agrupamento foi utilizado todo o conjunto de dados para a modelação dos algoritmos.

Na Figura 5.1 são ilustrados os valores dos índices *TraceW* e *Calinski and Harabasz* obtidos com os modelos gerados pelos algoritmos de agrupamentos escolhidos, Secção 5.1, e para diferentes valor de k, i.e. números de agrupamentos.

Os valores do índice *TraceW*, Figura 5.1(a), apresentam-se bastante similares entre todos os modelos. Existe no entanto uma pequena distinção entre os modelos por partição e os modelos hierárquicos, onde os últimos apresentam uma performance pior que os primeiros aquando do acréscimo do número de agrupamentos. A utilização da técnica de inicialização *K-Means++* com os algoritmos *K-Means* e *PAM* aumentou, apesar de minimamente, a performance dos algoritmos tendo sido o contributo mais acentuado com algoritmo *PAM*. O modelo *K-Means++* foi o modelo que obteve a melhor performance entre todos os modelos de agrupamentos e para todos valores de k. A determinação visual do ponto *Elbow point* no índice *TraceW* para o modelo *K-Means++* não é de fácil resolução, sendo que a proposta poderia possivelmente variar entre os valores de *k=2,3 e 4*. Recorrendo ao método da segunda derivada, Tabela 5.1, o valor máximo absoluto obtido foi para *k=3*. Como tal, e por consenso, é possível afirmar que um possível número



(a) *TraceW* (b) *Calinski and Harabasz*

Figura 5.1: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Alpino *Year Both*

adequado de agrupamentos indicado pelo índice *TraceW* para o modelo *K-Means++* seria o de *k=3*.

Tabela 5.1: Segunda Derivada do índice *TraceW* com o conjunto de dados Alpino *Year Both*

K	KMeans++
3	14.43
4	6.13
5	3.21
6	0.32
7	1.94
8	0.07
Recomendado	3

Os valores do índice *Calinski and Harabasz* atingem o seu valor máximo para *k=2* em todos os modelos e decrescem sucessivamente até *k=8*. O modelo *Diana* difere ligeiramente ao apresentar uma queda acentuada para *k=3*. No entanto, analisando a Tabela C.2 verifica-se que os modelos *K-Means++* e *Diana* foram os que obtiveram melhores resultados de *Calinski and Harabasz* para *k=2* e que para valores superiores de *k*, o algoritmo *K-Means++* foi o que apresentou, na sua maioria, os melhores resultados.

Com base nos índices *TraceW* e *Calinski and Harabasz*, o modelo *K-Means++* aparenta ser o que promove a maior coesão dos agrupamentos e uma maior dispersão entre os mesmos. No entanto, a identificação do número mais adequado de agrupamentos entre os dois índices não é consensual e como tal foram aplicados os índices *Silhoutte* e *Gap Statistic*.

A Figura 5.2 demonstra os valores obtido para o índice *Gap Statistic* com o algoritmo *K-Means++* aplicado ao conjunto de dados Alpino *Year Both*.

Analisando a Figura 5.2, pela estratégia de seleção do número mais adequado de

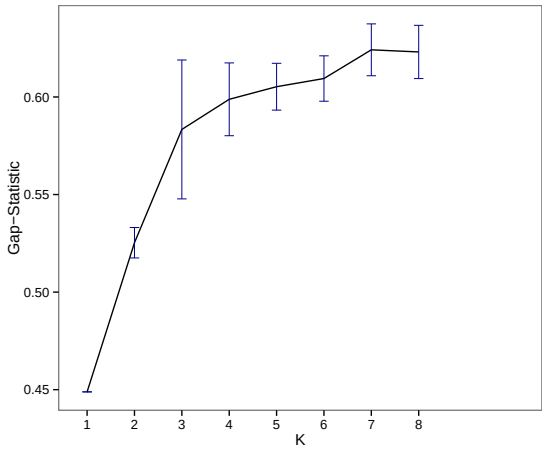
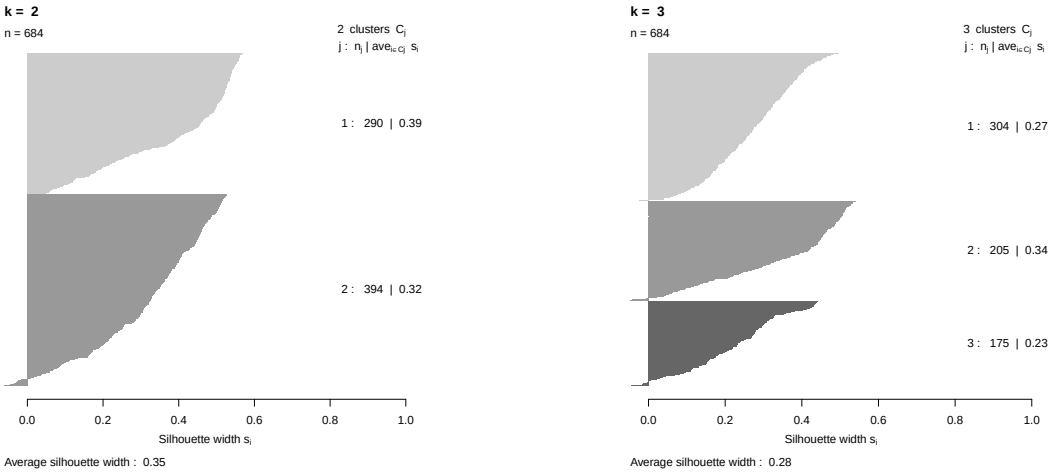


Figura 5.2: Gap Statistic com o conjunto de dados Alpino Atlântico *Y1&2*

agrupamentos para o índice *Gap Statistic* definida por Tibshirani et al. (2001), o ponto *k=4* foi o proposto como o mais adequado para o modelo *K-Means*. O ponto *k=3* não foi considerado como o mais indicado por uma margem mínima no entanto o ponto *k=2* foi considerado bastante desadequado.



(a) *k=2* (b) *k=3*

Figura 5.3: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Alpino Atlântico *Y1&2*

A Figura 5.3 apresenta o resultado obtido com o índice *Silhoutte* para *k=2* e *k=3*. Por tons de cinzento diferentes encontram-se distinguidos os agrupamentos formados para cada valor de *k*. O índice *Silhoutte* pode variar entre -1 e 1 que indica, respetivamente, o quanto a instância está desajustada ou ajustada ao agrupamento atribuído. As Figuras 5.3(a) e 5.3(b) ilustram, através de cada linha que compõe a figura de cada agrupamento, o valor do índice *Silhoutte* obtido para a respetiva instância. O volume e valor médio do índice de *Silhoutte* é também apresentado ao lado de cada agrupamento e em baixo na sua

totalidade. O índice *Silhoutte* sugere *k=2* como o número de agrupamentos mais adequado dado o valor médio de *Silhoutte* ser mais elevado, no entanto, é também possível verificar que para *k=3* o número de instâncias com valor negativo de *Silhoutte*, i.e., as instâncias que são consideradas como mal atribuídas, é reduzido.

A identificação do número de agrupamentos considerado como o mais adequado com o modelo *K-Means++* para o conjunto de dados Alpino *Y1&2* não é totalmente consensual entre os índices *TraceW*, *Calinski and Harabasz*, *Gap Statistic* e *Silhoutte*.

Para colmatar a falta consenso optou-se por incluir e analisar mais índices de validação tais como os índices *Davies Bouldin*, *Dunn*, *C-index*, *PBM*, entre outros Vendramin et al. (2010). Para o efeito foi utilizada uma biblioteca *NbClust*, Charrad et al. (2014), de forma a analisar 25 índices paramétricos para identificação do número mais adequado de agrupamentos, Apêndice D, entre os quais os quatro previamente analisados estão incluídos, *TraceW*, *Calinski and Harabasz*, *Gap Statistic* e *Silhoutte*. Para cada índice foi obtido o número mais adequado de agrupamentos com base no seu próprio critério. A Figura 5.4 apresenta o número de votos para cada valor de *k*..

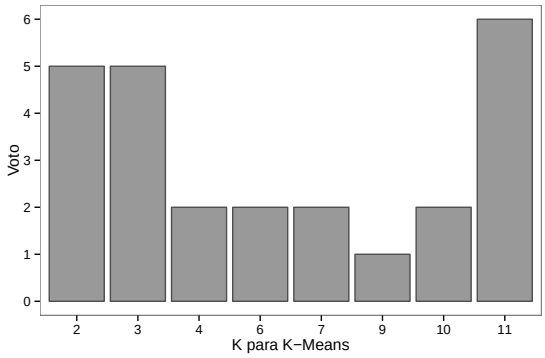


Figura 5.4: Identificação do número adequado de agrupamento por voto com o modelo *K-Means++* e o conjunto de dados Alpino Atlântico *Y1&2*

O voto maioritário para o número mais adequado de agrupamentos propõe *k=11*, no entanto, por repetidas experiências realizadas, o número recomendado pelos índices que recomendaram *k=11*, *Banfeld Raftery*, *C index*, *Davies Bouldin*, *SD*, *McClain Rao* e *SD* sempre foi o valor mais elevado para o qual *k* foi analisado. Dado o mesmo número de votos para *k=2* e *k=3*, foi tomada a decisão de optar por um número maior de agrupamentos visto que a opção, além de correta, traduzir-se-ia num acréscimo de informação aquando da avaliação da semântica de cada um dos agrupamentos.

Por fim, o modelo *K-Means++* foi considerado o modelo candidato para o conjunto de dados Alpino Atlântico *Y1&2* e *k=3* como o número mais adequado de agrupamentos.

5.2.0.1 Síntese da etapa de Modelação

O mesmo processo de modelação e validação foi aplicado a todos os conjuntos de dados definidos na Secção 4.2. Em todos os conjuntos de dados, o modelo gerado pelo algoritmo

K-Means++ foi o que apresentou a melhor performance perante os índices de validação apresentados. Os índices *Calinski and Harabasz* e *Silhoutte* indicaram *k=2* como o número mais adequado em todas as experiências realizadas com os diferentes conjuntos de dados enquanto que os índices *TraceW* e *Gap Statistic* indicaram, na sua grande maioria, *k=3* como o valor mais adequado. Perante o mesmo cenário de ambiguidade aquando da identificação do número mais adequado de agrupamentos que no conjunto de dados Alpino Atlântico *Y1&2*, recorreu-se ao sistema de votos para desambiguar os quatro índices paramétricos referidos. O sistema de votos indicou para todos os conjuntos de dados, muitas vezes por maioria, *k=3* como o número mais adequado de agrupamentos. Em Apêndice D encontram-se, sucintamente, as análises realizadas com os restantes conjuntos de dados.

5.3 Interpretação de Resultados

O objetivo de esta secção é a análise semântica dos agrupamentos obtidos com os modelos candidatos gerados na Secção 5.2. Para analisar cada um dos agrupamentos foram estabelecidas categorias de consumo de serviços, de hóspedes, de quartos ocupados e de temperatura ambiente de forma a enquadrar as instâncias respetivas a cada agrupamento. A análise da informação auferida permitiu definir perfis por cada unidade hoteleira.

5.3.1 Estratégia e Análise Semântica dos Agrupamentos

De forma a avaliar os agrupamentos resultantes dos modelos candidatos foi proposta uma estratégia de categorização das instâncias de cada agrupamento. As categorias foram definidas pela divisão do domínio de cada variável em cinco intervalos de valor com igual número de instâncias. Como tal, para a definição das fronteiras entre intervalos foram utilizados os percentis 20, 40, 60 e 80. Os intervalos definidos foram categoricamente definidos como *Low*, *Medium Low*, *Medium*, *Medium High* e *High*, indicados respetivamente por ordem de valor crescente. Esta metodologia permitiu analisar qual a frequência relativa do número de observações de cada um dos agrupamentos em cada um dos intervalos definidos. A distribuição da frequência relativa entre agrupamentos em cada intervalo permitiu classificar quais os agrupamentos mais relevantes. O critério de relevância baseou-se na frequência relativa de instâncias de um agrupamento c num determinado intervalo i, $\frac{N_{ci}}{N_i}$, perante um valor de *threshodl* adotado, τ, definido pela Equação 5.5. O valor de *threshodl* tem por base o número total de agrupamentos definidos pelo modelo, k, atenuado por um valor ε também este baseado no número total de agrupamentos k. O ajuste da condição através do valor ε foi considerado necessário dado que, ao longo da análise efetuada, observou-se que alguns dos agrupamentos obtiveram um valor de *IsRelevant* muito próximo mas abaixo de $\frac{1}{k}$. De forma a não ignorar estes agrupamentos pois a sua frequência relativa no intervalo ainda era considerável e tendo em conta que o índice *IsRelevant* baseia-se na frequência relativa , ou seja, quando um dos agrupamentos obtém

um valor de *IsRelevant* superior a $\frac{1}{k}$ implica que pelo menos um dos restantes não atinja o valor de $\frac{1}{k}$, foi adotado um ajuste à condição, determinado pelo valor ε. Desta forma, se um dos agrupamentos obtivesse um valor ligeiramente superior a $\frac{1}{k}$, com o ajuste da condição por ε, e se por sua vez um dos agrupamentos obtivesse um valor ligeiramente inferior a $\frac{1}{k}$ este não era desconsiderado como relevante.

$$IsRelevant_i = \frac{N_{ci}}{N_i} \geq \tau, \quad \tau = \frac{1}{k} - \varepsilon, \quad \varepsilon = \frac{0.1}{k} \tag{5.5}$$

Com o conhecimento de que $c = 3$ para todos os conjuntos de dados em análise, Secção 5.2.0.1, a Tabela 5.2 apresenta o valor de $\frac{1}{k}$, ou seja a frequência relativa de cada agrupamento do modelo candidato por categoria no conjunto de dados Alpino Atlântico *Y1&2*. A *negrito* são ilustrados os valores para qual a Equação 5.5 é verdadeira. É possível verificar que o agrupamento *C3* é o agrupamento com maior frequência relativa no nível *Medium Low*, seguido dos agrupamentos *C1* e *C2*. Pela Equação 5.5, com $k = 3$, $\tau = 0.3$, o agrupamento *C1* não é considerado relevante para o nível *Medium Low*. Seguindo a estratégia de categorização o agrupamento *C1* é considerado relevante nos níveis *Medium Low*, *Medium* e *Medium High*, o agrupamento *C2* nos níveis *Medium*, *Medium High* e *High* e o agrupamento *C3* nos níveis *Low* e *Medium Low* da variável *H*, número de hóspedes, no conjunto de dados *Y1&2* da unidade hoteleira Alpino Atlântico.

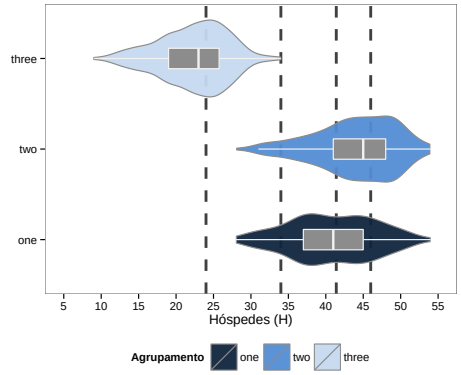
Tabela 5.2: Frequência Relativa do número de Hóspedes (H) distribuído por agrupamentos e níveis no conjunto de dados Alpino Atlântico *Y1&2*

Levels	C1	C2	C3
High	0.24	**0.76**	0.00
Medium High	**0.32**	**0.68**	0.00
Medium	**0.54**	**0.46**	0.00
Medium Low	**0.33**	0.13	**0.54**
Low	0.00	0.00	**1.00**

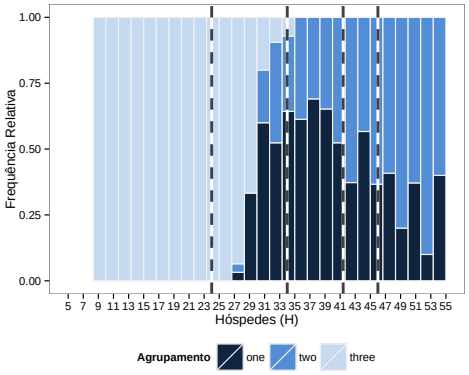
Além da identificação dos agrupamentos relevantes em cada intervalo através de tabelas como a Tabela 5.2, foram realizadas projeções visuais dos dados já categorizados de forma a percecionar melhor a distribuição dos mesmos. Ilustrado na Figura 5.5 os três agrupamentos de observações definidos para o conjunto de dados Alpino Atlântico *Y1&2* foram identificados como *C1*,*C2* e *C3*, unicamente como título de distinção e sem qualquer noção de categorização ordinal. Ambas as projeções ilustram as fronteiras de cada intervalo através de um tracejado vertical a cinzento.

As projeções na Figura 5.5(a) apresentam a distribuição de valores da variável *H* pelos três agrupamentos. Os agrupamentos foram caracterizados por tons de cor diferentes e as fronteiras dos intervalos de domínio, i.e., os percentis 20, 40, 60 e 80 da variável *H* foram assinaladas com linhas verticais a tracejado e por ordem respetiva.

Através de projeções como a ilustrada na Figura 5.5(a) procurou-se visualizar o domínio e a dispersão de valores de cada agrupamento. Esta projeção permite essencialmente



(a) Boxplots do número de Hóspedes (H) por agrupamento



(b) Frequência relativa do número de Hóspedes (H) por agrupamento

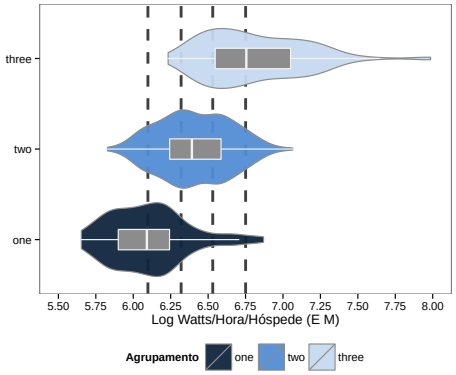
Figura 5.5: Número de Hóspedes (H) distribuído por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*

verificar a dispersão dos valores por todo o domínio da variável, comparar visualmente o volume de cada agrupamento através da área de cada *violin* e inferir qual a tendência do aglomerado de valores de cada agrupamento. Para ajudar na visualização foi adicionado um *boxplot* a cada agrupamento de forma a ilustrar o domínio de 50% do volume de cada agrupamento ao redor da sua mediana. É possível verificar que o agrupamento *C1* é o agrupamento que possui a dispersão mais equiparada entre níveis, encontrando-se na sua maioria disperso pelos níveis *Medium High* e *Medium*, mas também disperso pelos níveis *High* e *Medium*. O agrupamento *C2* possui pelo menos 75% das suas observações compreendidas entre os níveis *Medium High* e *High*. O agrupamento *C3* enquadra-se na sua totalidade nos níveis *Low* e *Medium Low* e praticamente 50% das suas observações encontram-se no nível *Low* no entanto apresenta uma grande concentração de instâncias próximo da fronteira entre os intervalos *Low* e *Medium Low*.

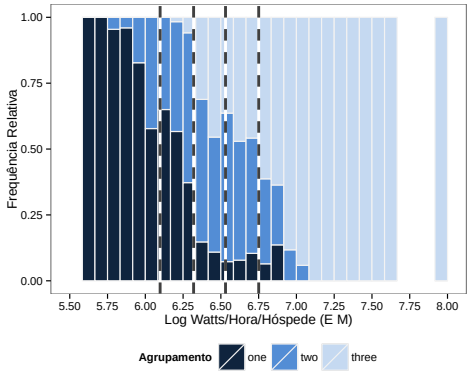
No entanto, com esta projeção visual a distribuição das instâncias em cada intervalo não era facilmente visualizada. Para auxílio, recorreu-se a projeções como a ilustrada na Figura 5.5(b). Na Figura é demonstrada a frequência relativa de observações da variável *H* por agrupamento, distinguidos pelas mesmas cores que na Figura 5.5(a). As observações foram agrupadas por cada par de hóspedes e as fronteiras dos intervalos de valor foram igualmente assinalas a tracejado.

À procura de uma maior interpretabilidade da semântica de cada agrupamento foram ilustradas as variáveis correspondentes ao valor médio diário, i.e., às variáveis *E M*, *A M* e *G M*, Figuras 6.1, 5.7 e 5.8 respetivamente, que correspondem aos valores médios dos sensores de consumo por hóspede e por dia.

O consumo médio diário de eletricidade, Figura 6.1 apresenta uma categorização bem diferenciada entre agrupamentos, tal qual como o consumo médio diário de gás, Figura 5.7. Os agrupamentos em ambos os consumos médios diários referidos variam de forma inversa aos agrupamentos da variável relativa ao número de Hóspedes (H), i.e., os

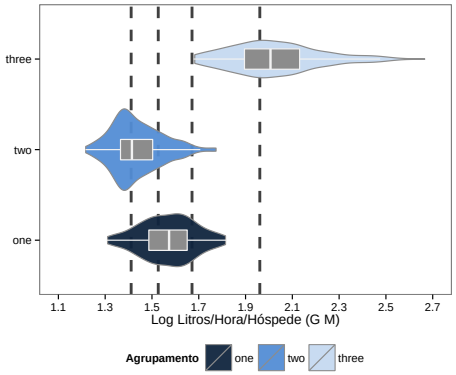


(a) Boxplots do consumo de Eletricidade (E M) por agrupamento

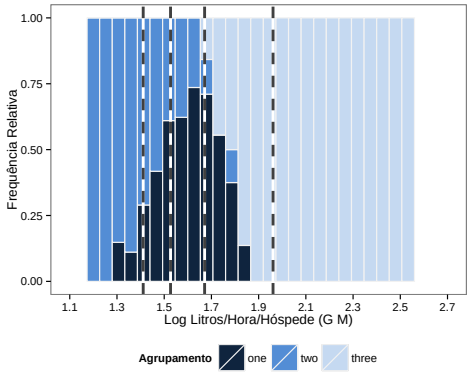


(b) Frequência relativa do consumo de Eletricidade (E M) por agrupamento

Figura 5.6: Consumo de Eletricidade (E M) distribuído por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*



(a) Boxplots do consumo de Gás (G M) por agrupamento

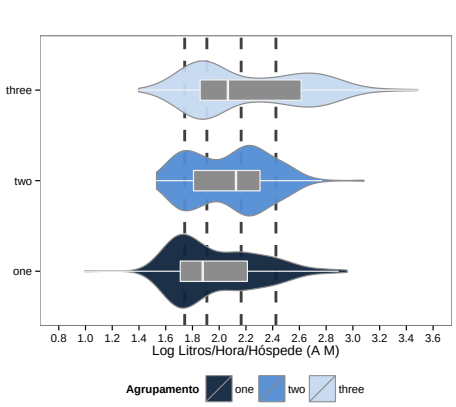


(b) Frequência relativa do consumo de Gás (G M) por agrupamento

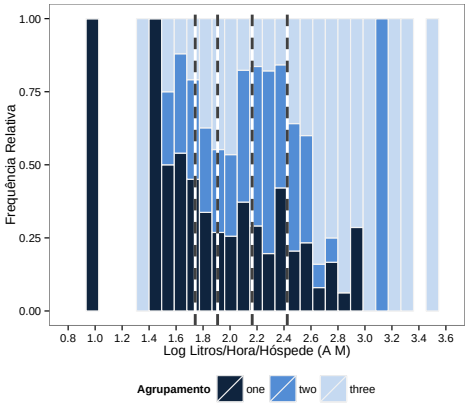
Figura 5.7: Consumo de Gás (G M) distribuído por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*

agrupamentos representativos de um valor mais elevado de ocupação são representativos de um menor consumo de eletricidade e de gás e vice-versa. No entanto, os agrupamentos *C1* e *C2* posicionam-se em categorias diferentes no que diz respeito ao consumo médio diário de eletricidade e gás. A diferenciação entre agrupamentos do consumo médio diário de água, Figura 5.8 é mais ambígua que nos outros dois serviços. No entanto, apresenta no entanto uma semelhança remota com as características dos agrupamentos do consumo médio de eletricidade onde o agrupamento *C1* caracteriza os valores de mais baixo consumo, o agrupamento *C2* de um consumo moderado e o agrupamento *C3* tendencialmente para consumos mais elevados.

À análise ilustrativa adicionou-se as variáveis $T\,M$ e Q, Figuras 5.9 e 5.10 respetivamente, que correspondem ao valor médio diário de temperatura ambiente e ao número



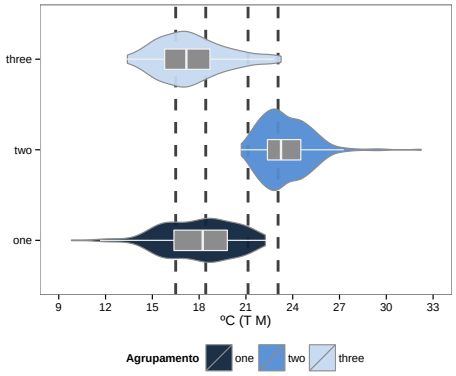
(a) Boxplots do consumo de Água (A M) por agrupamento



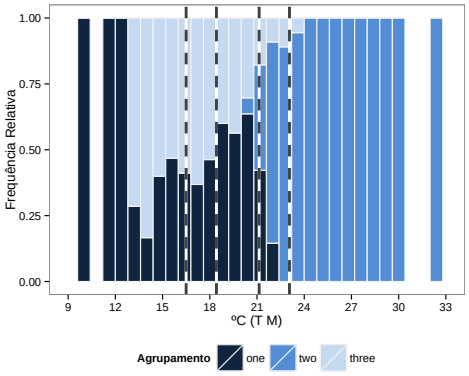
(b) Frequência relativa do consumo de Água (A M) por agrupamento

Figura 5.8: Consumo de Água (A M) distribuído por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*

diário de quartos ocupados na unidade hoteleira.



(a) Boxplots da Temperatura Ambiente (T M) por agrupamento

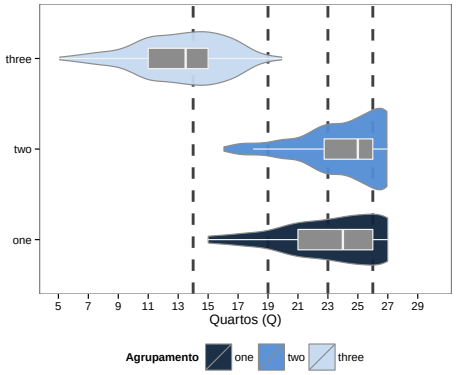


(b) Frequência relativa da Temperatura Ambiente (T M) por agrupamento

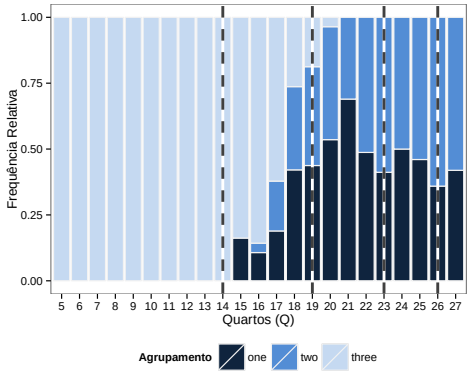
Figura 5.9: Temperatura Ambiente (T M) distribuída por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*

Os agrupamentos *C1* e *C3* englobaram os valores de temperatura média ambiente (T M) mais baixos e moderados, 5.9. Por sua vez, ao agrupamento *C2* foram atribuídas as instâncias com valores de temperatura média ambiente mais elevadas. Dada a noção temporal e período sazonal associado aos momentos em que a temperatura média ambiente é mais elevada na RAM, o agrupamento *C2* encontra-se aparentemente bem definido ao longo da estação de Verão, finais de Primavera e inícios de Outono.

Os agrupamentos definidos perante a variável Q, relativa ao número de quartos ocupados, Figura 5.10, é bastante semelhante à do número de hóspedes (H), no entanto, a dispersão de valores do agrupamento *C1* é mais concentrada no intervalo *High* que na



(a) Boxplots do número de Quartos ocupados (Q) por agrupamento



(b) Frequência relativa do número de Quartos ocupados (Q) por agrupamento

Figura 5.10: Número de Quartos ocupados (Q) distribuído por agrupamento no conjunto de dados Alpino Atlântico *Y1&2*

variável H.

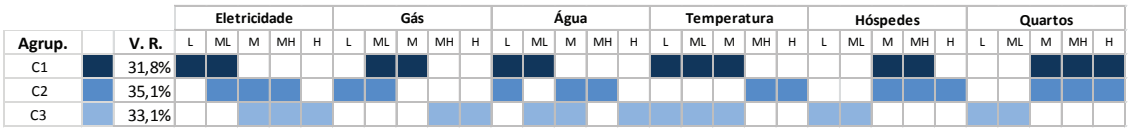
5.3.2 Caracterização dos Agrupamentos

Com uma melhor perceção da distribuição de valores de cada variável por agrupamento e através da condição *IsRelevant*, foi sumariado a categorização das instâncias em cada variável, i.e, as associações de cada instância por agrupamento a cada intervalo de valor.

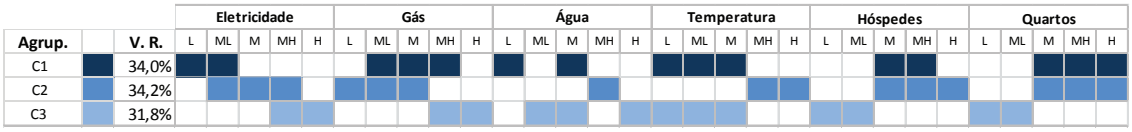
A Figura 5.11 apresenta sucintamente a categorização dos agrupamentos por cada conjunto de dados da unidade hoteleira Alpino Atlântico e respetivas variáveis. Entre os conjuntos de dados *Y1&2* e os sub-conjuntos *Y1* e *Y2* foram atribuídos aos agrupamentos obtidos os nomes *C1*, *C2* e *C3*, consoante a sua parecença semântica de forma a facilitar a comparação entre os diferentes conjuntos de dados. Na segunda coluna das figuras, *V.R.*, é apresentado o volume relativo de cada agrupamento em relação ao volume total do conjunto de dados respetivo. Nas seguintes colunas são apresentadas cada uma das variáveis e os respetivos intervalos de valor por sigla. Aos agrupamentos considerados relevantes num determinado nível de valor para uma determinada variável, com base no critério *isRelevant* Equação 5.5, foi atribuída a cor do respetivo agrupamento.

A dimensão relativa dos respetivos agrupamentos entre conjunto de dados foram semelhantes, no entanto, o agrupamento *C2* foi tendencialmente superior aos restantes, seguido dos agrupamentos *C1* e *C3* por ordem respetiva. A caracterização dos agrupamentos dos conjuntos de dados *Y1&2* e *Y1* é muito semelhante porém entre os conjunto *Y1&2* e *Y2* difere ligeiramente.

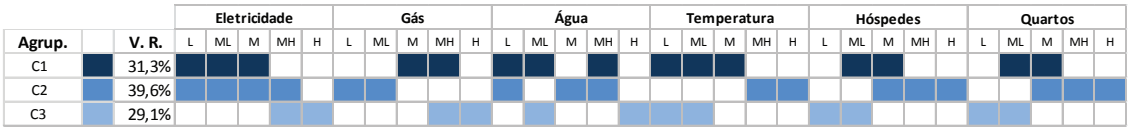
A Tabela 5.3 apresenta a concordância da atribuição das instâncias aos agrupamentos com semânticas semelhantes entre os conjuntos de dados da unidade hoteleira Alpino Atlântico. Os grupos *C1*, *C2*, e *C3* que se obtém para os conjuntos de dados em anos consecutivos, aparentam uma estabilidade tanto na quantidade de elementos como na



(a) Categorização dos agrupamentos no conjunto de dados Alpino Atlântico *Y1&2*



(b) Categorização dos agrupamentos no conjunto de dados Alpino Atlântico *Y1*



(c) Categorização dos agrupamentos no conjunto de dados Alpino Atlântico *Y2*

Figura 5.11: Categorização dos agrupamentos dos conjuntos de dados Alpino Atlântico

interpretação relativa a cada uma das variáveis considerada. Com base nas interpretações para cada uma das variáveis, foram então interpretados os grupos usando mais do que uma variável.

Tabela 5.3: Atribuição de instâncias a agrupamentos com semânticas similares nos conjuntos de dados da unidade Alpino Atlântico

		Y1			*Y2*		
	Agrupamento	C1	C2	C3	C1	C2	C3
Y1&2	C1	97	0	4	77	5	0
	C2	18	115	6	7	136	0
	C3	0	0	98	28	0	103

O agrupamento *C1* e *C2* repartem, na sua maioria, os valores de maior ocupação de quartos na unidade hoteleira, sendo que, o agrupamento *C2* enquadra os momentos em que existe uma maior taxa de ocupação dada a repartição da variável do número de Hóspedes. Por sua vez, o Agrupamento *C3* representa os momentos em que existe uma menor afluência de hóspedes na unidade hoteleira. A divisão entre os valores de temperatura ambiente é bastante ilustrativa ao separar os agrupamentos *C2* e *C1* entre temperaturas mais elevadas e temperaturas baixas a moderadas respetivamente. Os agrupamentos *C1* e *C3*, que repartem os valores de temperatura ambiente baixa a moderada, já haviam sido bem identificados relativamente à afluência de hóspedes na unidade hoteleira.

Em relação ao consumo de serviços, é praticamente consensual entre os serviços de eletricidade, gás e água, que quanto menor a presença de hóspedes na unidade hoteleira maior o consumo por hóspede. O agrupamento *C3* caracteriza os períodos de menor afluência na unidade hoteleira e enquadra os valores de maior consumo de serviços. Aquando de uma presença de hóspedes mais elevada, agrupamentos *C1* e *C2*, o consumo de eletricidade e de gás aparentam variar de forma inversa respetivamente aos valores de temperatura ambiente. O consumo de eletricidade aparenta ser menor quando os valores de temperatura são mais baixos, agrupamento *C1*, e o consumo de gás aparenta ser menor quando os valores de temperatura são mais altos, agrupamento *C2*. A categorização do consumo de água é no entanto algo ambígua sendo difícil caracterizar mais do que a relação inicial, ou seja, a relação entre o consumo de água de forma inversa ao número de hóspedes presente no hotel.

Como os valores de temperatura ambiente estão diretamente relacionados com períodos temporais, é possível afirmar que os agrupamentos *C1* e os agrupamentos *C2* e *C3* encontram-se também temporalmente caracterizados.

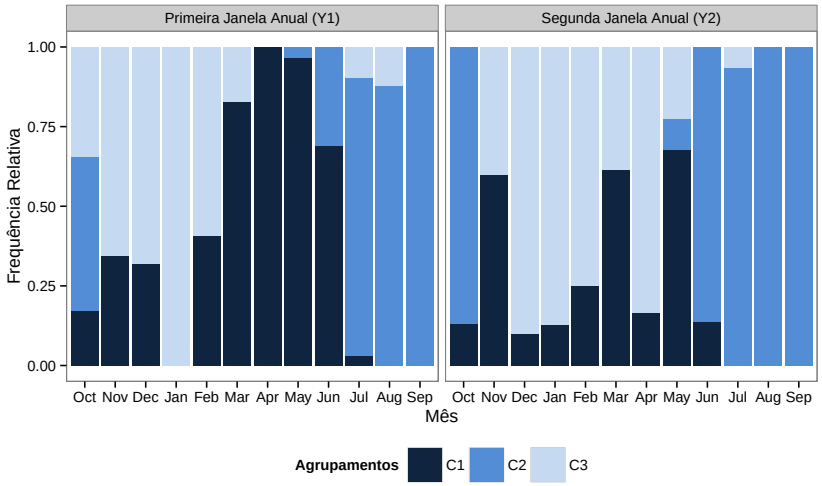


Figura 5.12: Caracterização temporal dos agrupamentos definidos com o conjunto de dados Alpino Atlântico *Y1&2*

A Figura 5.12 caracteriza mensalmente os agrupamentos do conjunto de dados Alpino

Atlântico *Y1&2* e perante as duas janelas anuais temporais, correspondentes também aos períodos dos conjuntos de dados *Y1* e *Y2*. As instâncias ocorrentes nos meses de Verão foram quase na sua totalidade atribuídas ao agrupamento *C2*, caracterizado pelos períodos de temperatura ambiente e de taxas de ocupação elevadas. Os agrupamentos *C1* e *C3* apresentaram-se de forma distinta entre os dois períodos anuais. Na primeira janela temporal o agrupamento *C1* englobou na sua maioria a estação de Primavera e repartiu, minoritariamente, as estações de Outono e Inverno com o agrupamento *C3*. Na segunda janela temporal ambos os agrupamentos repartiram as estações de Outono, Inverno e Primavera, onde o agrupamento *C3* teve relativamente mais volume que o agrupamento *C1*. Dado que o agrupamento *C1* e *C3* distinguem-se pelo número de turistas na unidade hoteleira, para efeitos de validação e compreensão dos motivos da diferença de volume relativo dos dois agrupamentos entre as duas janelas temporais, comparou-se o número total de hóspedes por mês entre as janelas temporais respetivas, Figura 5.13.

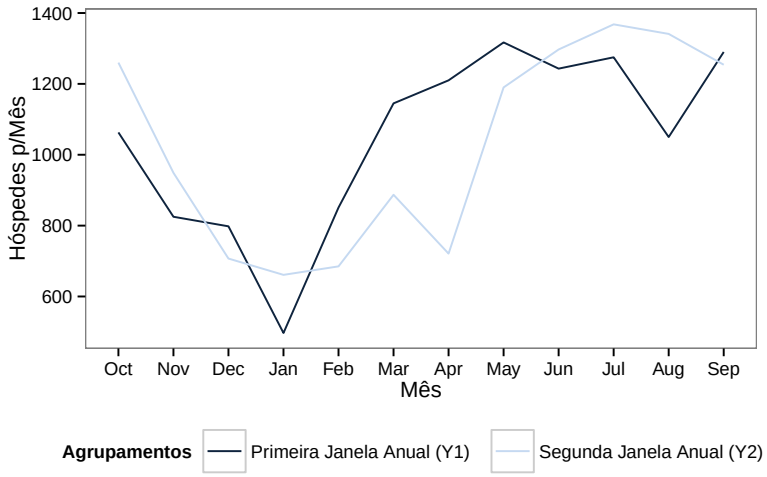


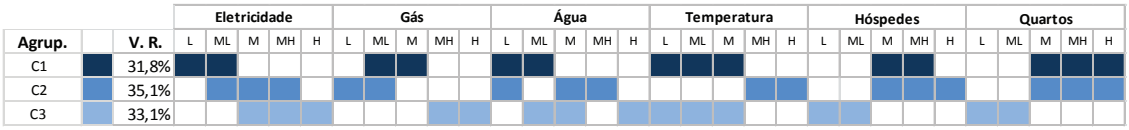
Figura 5.13: Caracterização da variável Hóspedes (H) por agrupamento no conjunto de dados Alpino *Year Both*

A diferença entre número total de hóspedes por mês em cada janela temporal valida a caracterização dos agrupamentos. Na segunda janela temporal houve mais hóspedes nos meses de Outubro e Novembro e no mês de Junho, o que valida a maior presença do agrupamento *C2* nesses mesmos meses na segunda janela temporal, Figura 5.12. Da mesma forma, houve muito menos hóspedes de Fevereiro a Junho na segunda janela temporal o que valida a maior presença do agrupamento *C3* em relação ao agrupamento *C1* durante esse período. O inverso sucedeu na primeira janela temporal, onde o agrupamento *C1* teve mais presença. Por fim, o número mais baixo de hóspedes em todo o período, ocorrido em Janeiro da primeira janela temporal, foi totalmente atribuído ao agrupamento *C3*.

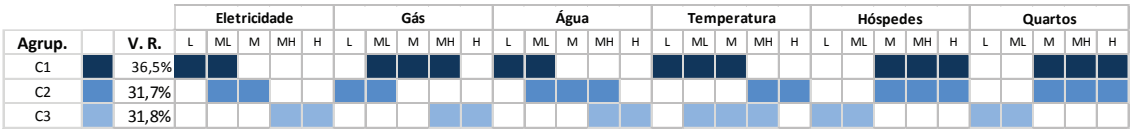
A caracterização dos agrupamentos por dias possivelmente relevantes, eventos socio-culturais e/ou feriados nacionais e regionais, não demonstrou uma diferença significativa que pudesse caracterizar os agrupamentos por estes períodos turísticos especiais.

5.3.3 Síntese de Resultados

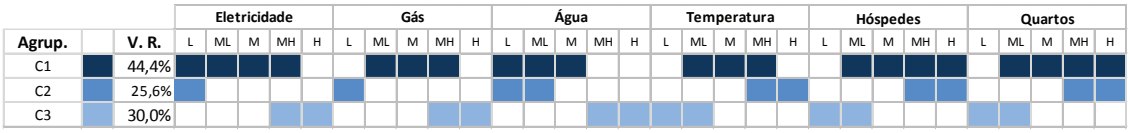
As três unidades hoteleiras mostraram coerência e estabilidade, quer na interpretação quer na quantidade de instâncias em cada agrupamento, entre os conjuntos de dados *Y1&2*, *Y1* e *Y2*. A síntese da caracterização dos agrupamentos das unidades hoteleiras Quinta das Vistas e Quinta de S.João encontram-se em Apêndice D. A Figra 5.14 sintetiza a análise a todas as unidades hoteleiras de forma que seja possível comparar os perfis definidos entre os três estabelecimentos. Para o efeito, os agrupamentos foram ordenados de forma a maximizar a similaridade da semântica entre agrupamentos de diferentes unidades hoteleiras. Por sua vez, os índices dos agrupamentos foram renomeados e as respetivas cores alteradas para o obter o mesmo efeito de comparação.



(a) Alpino Atlântico *Y1&2*



(b) Quinta de S.João *Y1&2*



(c) Quinta das Vistas *Y1&2*

Figura 5.14: Categorização dos agrupamentos

Os perfis entre as três unidades hoteleiras são bastantes semelhantes, em termos quantitativos e em termos de interpretação dos agrupamentos, especialmente entre as unidades hoteleiras Alpino Atlântico e Quinta de S.João. Porém, a unidade hoteleira Quinta das Vistas apresentou um volume menor no agrupamento *C2* em prol do agrupamento *C1* e como tal, uma pequena diferenciação no que diz respeito à interpretação de estes dois agrupamentos. Os agrupamentos definidos mantiveram a característica que relaciona um maior consumo de serviços por hóspede aquando de uma menor taxa de ocupação. Por sua vez, aquando de uma maior taxa de ocupação foi possível identificar dois grupos, um relativo aos períodos onde a temperatura ambiente é baixa a moderada e outro relativo aos períodos onde a temperatura ambiente é moderada a elevada. Como tal, os agrupamentos mantiveram também as mesmas características a nível temporal. Na unidade hoteleira

Quinta das Vistas, a separação entre o agrupamento *C2*, relativo aos períodos de maior ocupação e temperaturas mais elevadas, e o agrupamento *C1*, relativo aos períodos de maior ocupação mas temperaturas mais moderadas, foi mais tendencioso para o agrupamento *C1*. Denotar também que, a unidade hoteleira Quinta das Vistas foi o estabelecimento que obteve a menor taxa de ocupação ao longo de todo o período, Figura 3.6(b), o que pode justificar o menor volume do agrupamento *C2* dada as suas características.

5.3.4 Conclusão

Neste capítulo foram definidos perfis de consumo de serviços por dia e por hóspede em cada unidade hoteleira. Os perfis identificados foram caracterizados com base no consumo de eletricidade, água e gás, pela ocupação de hóspedes presente nas unidades hoteleiras e pelas condições atmosféricas no local, mais precisamente, pela temperatura ambiente no local. Observou-se os perfis de consumo identificados podem ser igualmente caracterizados temporalmente. Em suma, a comparação entre os perfis obtidos para cada unidade hoteleira revelou uma enorme semelhança. Mesmo tendo em conta as diferentes realidades de cada hotel, em termos de dimensão, taxa de ocupação, categoria da unidade hoteleira e magnitude de consumos de serviços, foi possível com esta análise caracterizar, de uma forma abrangente, um perfil de consumo de serviços por dia e por hóspede comum para as três unidades hoteleiras situadas na RAM.

CAPÍTULO 6

ESTIMATIVA DE CONSUMO DE SERVIÇOS

Neste capítulo é apresentado o estudo realizado para a estimativa de consumo de serviços nas unidades hoteleiras Alpino Atlântico, Quinta da Vistas e Quinta de S.João. Seguindo as etapas *Modelling* e *Evaluation* da metodologia CRISP-DM foram modelados, validados e avaliados algoritmos supervisionados de regressão para a análise preditiva.

A Secção 6.1 apresenta a secção teórica referente à estimativa de consumo de serviços, onde são apresentados os algoritmos de regressão estudados e as respetivas técnicas de validação e avaliação de esses mesmos algoritmos. A Secção 6.2 e a Secção 6.3 apresentam o estudo à estimativa do consumo de serviços sob uma granularidade diária e sob uma granularidade horária, respetivamente. Em cada um das secções é apresentada uma conclusão à respetiva análise.

6.1 Algoritmos de Regressão

As técnicas de aprendizagem supervisionada são usualmente denominadas como técnicas de classificação ou de regressão. As técnicas de classificação diferem das de regressão no sentido em que as primeiras procuram distinguir classes de dados ou conceitos enquanto as últimas procuram prever valores de dados perdidos ou indisponíveis. Como tal, para a estimativa de consumo de serviços nas unidades hoteleiras foram analisadas técnicas supervisionados de regressão. Os algoritmos de regressão utilizados para análise foram os seguintes: Regressão Linear, (RL); *K-Nearest Neighbors*, (KNN); *Support Vector Machine*, (SVM). Os algoritmos RL e SVM definem uma função global com base em todo o conjunto de dados para explicar a variável a estimar sendo que, o primeiro define uma função linear e o segundo possui a capacidade de definir uma função não-linear. Por sua vez, o algoritmo KNN define a sua estimativa com base numa função *local*, i.e., com base nas x observações mais próximas ao que pretende estimar.

O algoritmo Regressão Linear define uma relação linear entre uma variável escalar dependente $E(Y|X)$ com base em uma ou em um conjunto de variáveis explicativas, $X_1, ..., X_p$. A definição de linear caracteriza uma função linear objetiva definida pelas variáveis explicativas e os seus coeficientes respetivos, Equação 6.1.

$$\hat{Y} = f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{6.1}$$

A função linear é usualmente modelada através da minimização do erro quadrado da estimativa $f(x)$ ao verdadeiro valor y, para todas as N observações, Equação 6.2.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{6.2}$$

O algoritmo de Regressão Linear é bastante simples o que faz com que seja mais facilmente interpretado. No entanto, o algoritmo pode obter melhores resultados que algoritmos de regressão não-lineares aquando de cenários onde a razão sinal-ruído da variável a estimar é baixo, Hastie et al. (2009).

O algoritmo *K-Nearest Neighbors*, (KNN), define a sua estimativa utilizando as observações mais próximas às características da observação pretende estimar. Tipicamente o algoritmo é definido pela Equação 6.3 onde N_k corresponde às k observações mais próximas de x, usualmente denominadas de observações vizinhas. Por sua vez, a métrica de proximidade, ou similaridade, mais comum na utilização do algoritmo KNN é a distância Euclidiana.

$$\hat{Y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{6.3}$$

A estimativa do algoritmo KNN pode ser simplesmente definida como a média da variável de resposta y das k observações mais próximas de x. Dada a definição do algoritmo KNN e noção espacial de proximidade, o número de variáveis a utilizar para estimar uma variável de resposta está diretamente relacionado com o volume do conjunto de dados, Hastie et al. (2009).

O algoritmo *Support Vector Machine*, (SVM), foi inicialmente concebido como uma técnica de classificação. Em 1996 foi proposto em Drucker et al. (1997) uma adaptação para a sua utilização como técnica de regressão, também denominado como *Support Vector Regression*, (SVR). O algoritmo SVM é um algoritmo complexo que permite definir uma função não-linear através do uso de *kernels*, algoritmos capazes de definir relações entre conjuntos de variáveis em dimensões mais elevadas. Neste estudo foi utilizado o *Gaussian/(RBF) kernel* para a transformação do conjunto de variáveis em dimensões mais altas, baseado na distribuição Gaussiana.

6.1.1 Índices de Validação

O índice estatístico mais comum aquando da validação dos modelos de regressão é o índice R^2 que indica o quanto o modelo se encontra ajustado ao conjunto de dados, Equação 6.4, variando entre 0.0 e 1.0. O índice caracteriza-se por representar a percentagem da variância da variável dependente explicada pelo modelo. A variância explicada pelo modelo é calculada através da variância não explicada pelo modelo (ou variância do erro residual), Equação 6.6, sobre a variância total da variável a estimar, Equação 6.5.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6.4}$$

$$SS_{tot} = \sum_{i}^{n} (y_i - \bar{y})^2 \tag{6.5}$$

$$SS_{res} = \sum_{i}^{n} (y_i - f_i)^2 \tag{6.6}$$

No entanto, o índice R^2 acresce sempre o seu valor aquando da inclusão de novas variáveis explicativas aos modelos. De forma a que os modelos pudessem ser comparados entre si aquando de diferentes cardinalidades de variáveis explicativas foi utilizado o índice *Adjusted* R^2, Equação 6.7, que simplesmente ajusta o índice R^2 ao número p de termos independentes do modelo penalizando modelos mais complexos.

$$Ajusted\, R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{6.7}$$

Aquando da modelação dos algoritmos de Regressão Linear é usual a aplicação de técnicas que procurem identificar o menor número de variáveis que caracterizem o verdadeiro sinal a estimar e por sua vez minimizem o ruído. A família de técnicas *Selective selection* procura, sob processos iterativos, definir qual o menor conjunto de variáveis pertences a um conjunto que seja capaz de exprimir o mesmo detalhe de informação que todo o conjunto. Uma das técnicas pertences à família *Selective selection* é a *Exhaustive selection*, Hastie et al. (2009). Esta técnica procura, de forma exaustiva, todas as possíveis combinações das variáveis disponíveis e analisa-as perante diferentes índices providenciados, Hastie et al. (2009). Através da análise dos índices definidos é posteriormente escolhido o conjunto com melhor performance, i.e., o conjunto de variáveis que exprima a mesma quantidade/qualidade de informação da variável de resposta quanto todo o conjunto de variáveis. As métricas usualmente utilizadas para identificar qual o melhor conjunto de variáveis são os índices SS_{res}, *Mallow's Cp* e BIC (Bayesian Information Criterion), Hastie et al. (2009). O índice SS_{res}, Equação 6.6, define a variância do erro residual e como tal, o menor valor de SS_{res} é o mais indicado.

$$C_p = \frac{SS_{res_p}}{s^2} - n + 2p, \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 \tag{6.8}$$

O índice *Mallow's Cp*, Equação 6.8, procura relativizar a variância do erro de cada modelo de cardinalidade p com o erro quadrado médio, s^2, de um modelo gerado com todas as variáveis do conjunto. Como tal, o menor valor de *Mallow's Cp* é indicativo de um conjunto de variáveis mais adequado.

$$BIC = n \cdot ln(\frac{SS_{res}}{N}) + k \cdot ln(n) \tag{6.9}$$

O índice BIC identifica qual o modelo candidato mais adequado através da Equação 6.9, onde o aumento da variância não explicada da variável a estimar aumenta o valor de BIC. Como tal, o menor valor de BIC melhor o ajuste do modelo.

6.1.2 Métricas de Performance

As métricas de performance baseiam-se, na sua maioria, na diferença entre o valor verdadeiro e a estimativa do modelo, denominado por erro, e. Para avaliar a performance dos algoritmos de regressão com os conjuntos de dados do projeto Soltgest foram utilizadas as seguintes métricas: raiz quadrada do erro quadrado médio, RMSE ou *Round Mean Squared Error*, Equação 6.10; erro absoluto médio, MAE ou *Mean Absolute Error*, Equação 6.11; percentagem do erro absoluto médio, MAPE ou *Mean Absolute Percentage Error*, Equação 6.12. O RMSE indica a raiz quadrada do erro quadrado entre a estimativa e a verdadeira instância e é bastante utilizado para a modelação dos algoritmos de regressão pois a sua definição agrava o erro quanto maior for a disparidade entre a previsão e o valor verdadeiro. Por sua vez, o MAE indica o verdadeiro erro médio ocorrente entre a estimativa e o valor verdadeiro ao utilizar o valor absoluto da diferença. De forma a relativizar o erro e possibilitar a comparação entre serviços e unidades hoteleiras distintas foi introduzida a métrica MAPE que relativiza percentualmente a diferença ao valor verdadeiro, Hastie et al. (2009).

$$RMSE = \sqrt{mean(e_i^2)} \tag{6.10}$$

$$MAE = mean(|e_i|) \tag{6.11}$$

$$MAPE = mean(|p_i|), p_i = 100 * \frac{e_i}{y_i} \tag{6.12}$$

Com o decorrer do estudo verificou-se que o sinal do erro obtido oscilava bastante entre valores negativos e positivos. Como tal, foram adotadas duas novas métricas de performance onde o erro passou a ser analisado sob granularidades mais elevadas. Por exemplo, aquando do estudo da estimativa sob uma granularidade diária, verificou-se que o erro total ao fim de, por exemplo, uma semana era amortizado pela oscilação do sinal. O erro acumulado passou assim a ser definido por e_t onde t corresponde à dimensão da janela temporal, Equação 6.13. De forma a possibilitar a comparação entre estudos foi definido o erro médio absoluto por janela temporal, MAE_t, Equação 6.14. Em MAE_t foi

obtido o erro absoluto resultante de cada janela temporal e dividido pelo tamanho da janela forma a obter o erro absoluto sob a granularidade inicial. O erro médio absoluto percentual para cada período t também foi definido, $MAPE_t$ *Mean Absolute Percentage Error*, Equação 6.15, sendo este relativo ao erro ao final do período t.

$$e_t = \sum_{i=1}^{t} e_i \ , \quad t \geq 1 \tag{6.13}$$

$$MAE_t = mean(|\frac{e_t}{t}|) \tag{6.14}$$

$$MAPE_t = mean(p_t), p_t = \frac{|\sum_{i=1}^{t} e_i|}{\sum_{i=1}^{t} y_i} * 100 \tag{6.15}$$

Com esta abordagem foi possível observar o erro perante uma perspetiva de granularidade superior à do sinal.

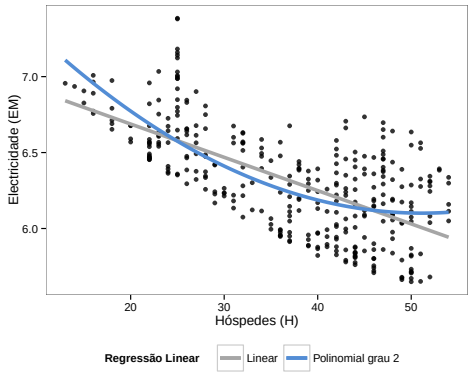
6.2 Estimativa perante Granularidade Diária

Recordando o processo de preparação dos dados na Secção 4.2 foram atribuídos dois conjuntos de dados para o estudo de estimativa de consumo de serviços, um para a fase de modelação dos algoritmos, para efeitos de treino e validação, e outro para a fase de teste. Os algoritmos de regressão modelados foram os apresentados na Secção 6.1, respetivamente os algoritmos Regressão Linear, KNN e SVM e a sua modelação teve como objetivo a diminuição da métrica RMSE, i.e., do erro quadrado entre o verdadeiro valor e a estimativa do modelo. A primeira abordagem ao estudo regressivo implicou presumir a independência entre instâncias e como tal o estudo da estimativa de um sensor de consumo não teve em consideração consumos passados. Os consumos de eletricidade, água e gás ocorrentes no mesmo dia estão muito possivelmente interligados e como tal, a análise a cada serviço teve por base unicamente os valores de ocupação e temperatura ambiente. A secção está dividia em seis sub-secções. As quatro primeiras sub-secções discutem as técnicas de análises preditivas utilizadas. A título ilustrativo foi apresentado o processo à análise de estimativa do consumo de eletricidade na unidade Alpino Atlântico sendo que as restantes análises aos restantes serviços e respetivas unidades hoteleiras encontram-se em Apêndice E. Na primeira sub-secção, a análise preditiva teve, exclusivamente, por base o número de hóspedes presentes na unidade hoteleira. Dada a semelhança da semântica das variáveis foi igualmente analisado o número de quartos ocupados como variável explicativa. Na segunda sub-secção foram adicionadas às variáveis de ocupação as variáveis relativas à temperatura ambiente, valor médio e respetivos percentis. Com base nos resultados da segunda sub-secção foi elaborada a terceira sub-secção onde foi incluído o estudo de perfis de consumo de serviços realizado no Capítulo 5. Na quarta sub-secção foi analisada a possibilidade e o benefício de incluir valores históricos de consumo de serviços.

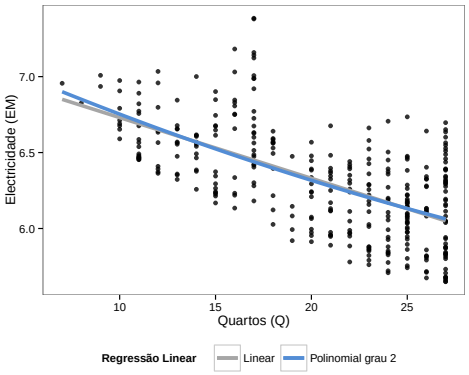
A quinta sub-secção apresenta uma análise ao erro obtido pelos modelos gerados sob diferentes perspetivas que, no cenário em causa, promove a performance dos modelos. A sexta e última sub-secção apresenta sucintamente os melhores resultados obtidos e modelos respetivos na estimativa do consumo de serviços em cada unidade hoteleira.

6.2.1 Estimativa com base na Ocupação

Esta secção apresenta a análise de estima de consumo de eletricidade com base no número de hóspedes ou com base no número de quartos ocupados. O objetivo principal foi verificar a relação entre ambas as variáveis. O algoritmo Regressão Linear foi modelado com todo o conjunto de dados de treino visto não implicar nenhum processo de parametrização/afinação do modelo. Aquando de uma primeira análise sob simples regressões lineares e com apenas uma variável independente observou-se que a tendência da dispersão entre do número de hóspedes (H) e o consumo de eletricidade era tendencialmente mais quadrática que retilínea, Figura 6.1(a). Por sua vez, a variável Q, número de quartos ocupados, apresentava a mesma tendência mas de uma forma muito mais discreta que a variável H, Figura 6.1(b).



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial de grau 2 com quartos ocupados (Q) como variável independente

Figura 6.1: Regressão Linear e Polinomial perante o consumo médio de eletricidade diário (EM) à escala logarítmica com o conjunto de dados de treino da unidade Alpino Atlântico

Na Figura 6.1 são apresentas ambas as variáveis em questão, H e Q do conjunto de dados de treino, e as respetivas funções dos modelos gerados. Para comparação foi aplicado a cada variável uma regressão linear de primeiro grau, ilustrada a cinzento, e uma regressão linear polinomial de segundo grau, ilustrada a azul. A Tabela 6.1 apresenta os modelos referidos e os resultados dos índices de validação e das métricas de performance. Os índices de validação foram calculados perante o comportamento do modelo com o conjunto de dados de treino e as métricas de performance com o conjunto de dados de teste.

Tabela 6.1: Sumário de Regressões Lineares para a estimativa de consumo de Eletricidade (EM) com o conjunto de dados Alpino Atlântico

Regressão Linear log (EM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.427	0.264	331.1	217.5	23.4
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.463	0.255	291.9	198.6	21.9
$\beta_0 + \beta_1 Q$	0.389	0.272	350.4	231.9	24.6
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.389	0.272	344.8	229.0	24.4

A utilização da função polinomial resultou num aumento do *Adjusted* R^2 e uma diminuição da variância do erro residual com a variável H, o que representa um melhor ajustamento do modelo polinomial aos conjuntos de dados de treino. A variável Q não apresentou melhorias no conjunto de dados de treino com a adição de complexidade de um polinomial de grau 2.

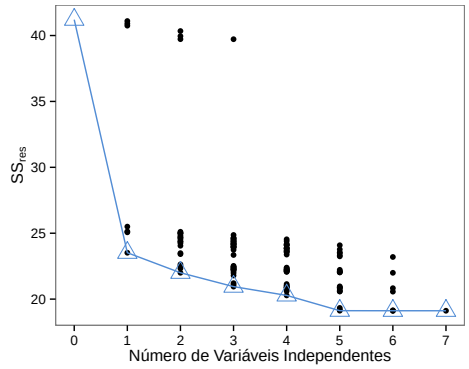
As métricas de performance apresentadas, RMSE, MA e MAPE, foram calculadas após a transformação da estimativa à escala original e não sobre a escala logarítmica. O erro residual não pode ser diretamente traduzido dos modelo gerados dado uso da escala logarítmica nas variáveis dependentes (sensores de consumo). Por sua vez, o uso da escala logarítmica altera também a interpretação dos coeficientes das variáveis independentes sendo que o acréscimo de uma unidade nas variáveis independentes caracteriza a diferença percentual entre o valor da observação e a estimativa. No entanto, a atribuição de um valor nulo aos coeficientes das variáveis independentes de um modelo, $\forall i \geq 1, \beta_i = 0$, possibilita verificar o valor estimado de consumo de eletricidade diário para base de operações. O modelo candidato a verificar a estimativa dos custos base de operação seria o modelo polinomial de segundo grau com base na variável H pois foi o que obteve o menor valor de MAE, MAPE e RMSE. Aquando da atribuição do valor nulo aos coeficientes das variáveis, i.e, a não existência de hóspedes, o modelo respetivo indica um consumo de 2,730k Watts/dia que pode ser interpretado como o consumo de eletricidade que decorre na unidade hoteleira como suporte de operação. Aquando de zero quartos ocupados o modelo, polinomial com base na variável Q, indica um consumo 1,443k Watts/dia sendo a interpretação idêntica à do modelo anterior dado a semântica das variáveis. A diferença da estimativa entre ambos os modelos é significativa o que pode implicar que uma das variáveis independentes não é capaz de exprimir totalmente a variável dependente ou, ambas as variáveis não o conseguem sendo a última alternativa a mais provável dada a variabilidade visível nas Figuras 6.1(a) e 6.1(b).

O modelo candidato de esta análise, Regressão Linear polinomial de segundo grau com base exclusiva no número de hóspedes (H), apresentou um erro percentual de consumo de eletricidade diário de 21.9% e com um erro absoluto de 199 Watts/Dia, um valor elevado dado os 780 Watts/Dia consumidos em média por cada hóspede na unidade hoteleira.

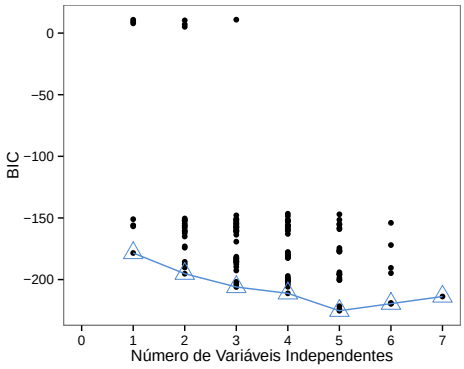
6.2.2 Estimativa com base na Ocupação e Temperatura Ambiente

Com o objetivo de melhorar a performance do modelo candidato até ao momento, foram incluídas mais variáveis à análise tais como os valores de temperatura ambiente e a união entre o número de hóspedes e o número de quartos ocupados no conjunto de dados.

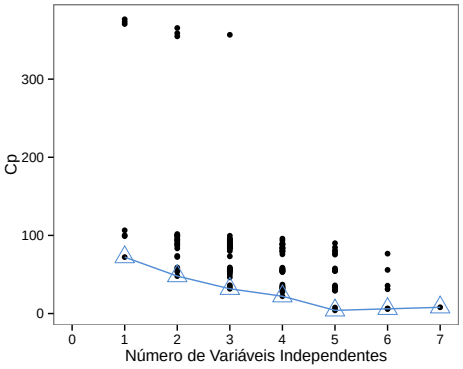
O primeiro passo para a modelação do algoritmo de regressão linear foi a escolha de uma estratégia que identificasse qual o melhor conjunto de variáveis com que modelar o algoritmo. Dado o volume dos conjuntos de dados e a cardinalidade de variáveis foi possível utilizar o processo *Exhaustive selection* da técnica *Stepwise selection*. Para a identificação da cardinalidade e do melhor conjunto de variáveis a adotar utilizaram-se os índices SS_{res}, Figura, BIC e *Mallow's Cp*, Figura 6.2.



(a) Índice SS_{res}



(b) Índice BIC



(c) Índice Mallow's Cp

Figura 6.2: Validação da melhor combinação de variáveis na modelação de algoritmos de Regressão Linear para a estimativa de Eletricidade (EM) com o conjunto de dados Alpino Atlântico

Na Figura 6.2 são apresentados os diversos resultados obtidos por cada índice, com diferentes combinações do conjunto de variáveis (representadas a preto em cada figura) e por cardinalidade das combinações. O valor mínimo no índice BIC está bem identificado na Figura 6.2(b) e indica um conjunto composto por cinco variáveis, o mesmo valor que o indicado pelo índice *Mallow's Cp* na Figura 6.2(c) mas de uma forma mais discreta. Por

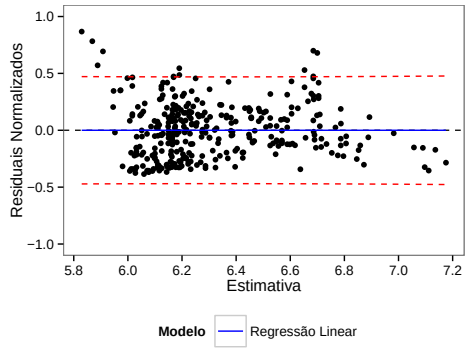
sua vez, os índices SS_{res} não apresentam ganhos para uma cardinalidade superior a 5 e, por consenso, foi adotado o conjunto indicado por todos os índices de cardinalidade igual a 5. O conjunto escolhido é composto pelas variáveis indicadas na Equação 6.16 e sua performance caracterizada na Tabela 6.2.

$$RL = \beta_0 + \beta_1 H + +\beta_2 H^2 + \beta_3 Q + \beta_4 Q^2 + \beta_5 TP25 \tag{6.16}$$

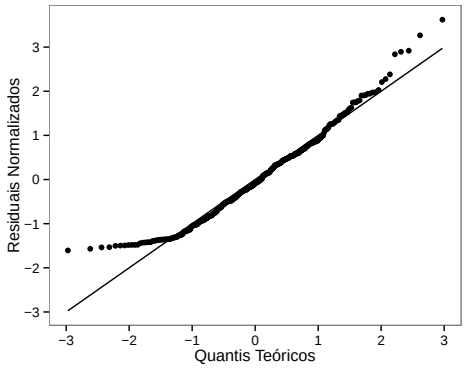
Tabela 6.2: Performance do modelo *RL* para a estimativa de eletricidade (EM) no conjunto de dados Alpino Atlântico

Regressão Linear log (EM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.529	0.239	293.9	191.5	21.2

O modelo *RL* foi definido com base nas variáveis de ocupação e respetivos polinomiais e com base no percentil 25 de temperatura ambiente. O modelo apresentou melhorias em todos os aspetos, no ajustamento ao conjunto de dados de treino com melhorias nos índices *Adjusted* R^2 e SS_{res}, e no erro de generalização indicado pelas métricas de performance MAE, MAPE e RMSE. No entanto, a percentagem do erro médio absoluto manteve-se acima dos 20%. De forma a avaliar a qualidade do ajuste do modelo foi observada a distribuição do erro residual Figuras 6.3(a) e 6.3.



(a) Distribuição do Erro residual normalizado perante a estimativa do modelo



(b) Distribuição do Erro residual normalizado perante uma hipotética distribuição Gaussiana

Figura 6.3: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Alpino Atlântico

A distribuição do erro ao longo das estimativas do modelo, Figura 6.3(a), indica que o erro residual aparenta ser não-correlacionado e a sua variância homogénea. A azul foi definida uma regressão linear, com o erro residual como variável dependente e a estimativa como variável independente, e a vermelho o intervalo de confiança da estimativa que delimita a área onde se espera que 95% das observações estejam contidas. O modelo da figura em questão, com coeficiente de correlação próximo de zero, indica que não existe

qualquer tendência no erro residual. Por sua vez, a magnitude entre ambas as fronteiras do intervalo de confiança e o modelo é bastante similar. A Figura 6.3(b) apresenta a distribuição dos valores do erro residual perante o que se define como uma hipotética distribuição Gaussiana (em recta). É visível que os valores do erro residual desenquadram-se de uma distribuição verdadeiramente gaussiana aquando de valores mais distantes ao seu valor médio, no entanto, a grande maioria dos valores seguem a hipotética distribuição gaussiana. Com base na leitura do erro residual, o modelo *RL* aparenta estar adequado à interpretação da variável dependente com as variáveis independentes disponíveis, Faraway (2004), apesar do elevado valor de MAPE.

Com o intuito de obter um modelo com melhor performance foram estudados outros algoritmos de regressão possíveis, nomeadamente o algoritmo KNN (*K-Nearest Neighbours*) e o algoritmo SVM (*Support Vector Machine*). Os algoritmos referidos requerem um processo de parametrização para um melhor ajustamento do modelo ao conjunto de dados e como tal, foi utilizada a técnica de *10-fold cross validation* com seleção aleatória das instâncias para a fase de modelação.

Tabela 6.3: Performance dos modelos de regressão para a estimativa de eletricidade (EM) no conjunto de dados Alpino Atlântico

Regressão Linear log (EM)	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	293.9	191.5	21.2
KNN	318.2	192.2	20.2
SVM	263.8	160.1	17.5

A Tabela 6.3 apresenta a comparação de performance entre os modelos *RL*, *KNN* e *SVM*. Ambos os modelos adicionados ao estudo obtiveram uma performance superior, em especial o modelo SVM que obteve um valor de MAPE de 17.5% e um erro de estimativa real de consumo de 160 Watts/Dia por hóspede.

6.2.3 Estimativa utilizando Perfis de Consumo

Observando a tendência do erro residual do modelo SVM ao longo do período de treino, Figura 6.4, e assinalando as instâncias pelos agrupamentos definidos no Capítulo 5, é possível observar alguma relação entre a magnitude do erro e os respetivos agrupamentos. Como tal, foi analisado nesta secção qual o contributo que a classificação e separação prévia das observações, com base nos agrupamentos definidos, poderia trazer. A estratégia adotada definiu assim x modelos de acordo com o número de agrupamentos. Como o Capítulo 5 demonstrou que, para as unidades hoteleiras em causa foram identificados 3 agrupamentos, foram definidos 3 modelos para cada sub-conjunto de dados correspondente a cada agrupamento.

O conjunto de dados de teste é "invisível"ao processo de modelação de algoritmos e como tal não foi possível utilizar a definição dos agrupamentos para todo o conjunto de

dados. A solução passou por recorrer a uma técnica de classificação.

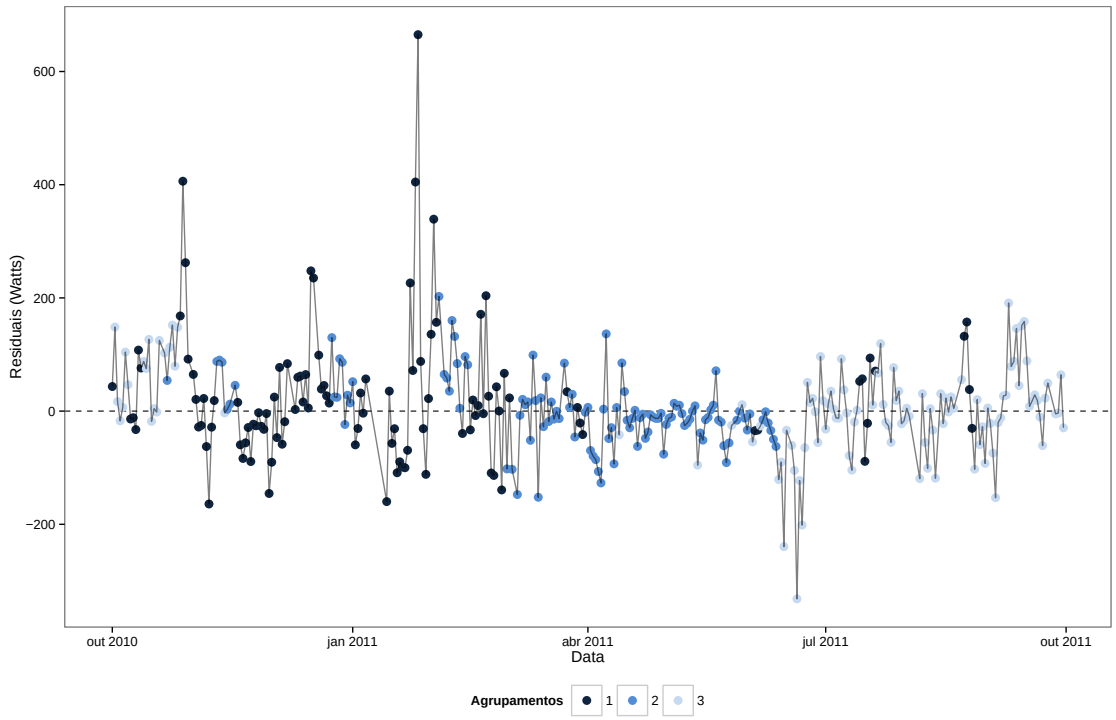


Figura 6.4: Distribuição do Erro residual da estimativa do consumo de Eletricidade (EM) perante o modelo SVM com o conjunto de dados Alpino Atlântico

Para poder classificar previamente as observações de forma a separar as instâncias em sub-conjuntos diferentes, recorreu-se a uma técnica de aprendizagem supervisionada de classificação. Ao conjunto de dados de treino foi adicionada uma variável *Agrupamento* (A) equivalente ao agrupamento atribuído na elaboração de perfis de consumo para o conjunto de dados *Y1* (idêntico ao conjunto de dados de treino). Com este novo conjunto de dados foi modelado o algoritmo *Decision Tree* (CART), sob a técnica *10-fold cross-validation*, para classificar o agrupamento de cada observação com base nos valores de ocupação e temperatura ambiente. O algoritmo *Decision Tree* (CART), gera um modelo de decisão com uma estrutura em forma de árvore, onde, na sua forma mais simples, avalia por etapas uma variável do conjunto de dados e gera uma regra que define qual o ação a tomar ou resultado a retornar.

A Figura 6.5 apresenta as regras definidas pelo modelo CART. Em cada nó é apresentado o agrupamento dominante, i.e., o agrupamento com maior volume no nó respetivo e as respetivas frequências relativas de cada agrupamento. Por exemplo, no nó inicial indica o agrupamento 2 (*C2*) visto ser o mais volumoso entre todos. Os nós terminais estão identificados com as mesmas cores adotadas para os agrupamentos *C1*, *C2* e *C3* na Secção 5.3. Os nós encontram-se também referenciados por índice no topo do nó.

A primeira regra após o nó inicial separa o agrupamento *C3* dos restantes com base no número de Hóspedes e encaminha-os para um nó (à esquerda) onde 92% de todas

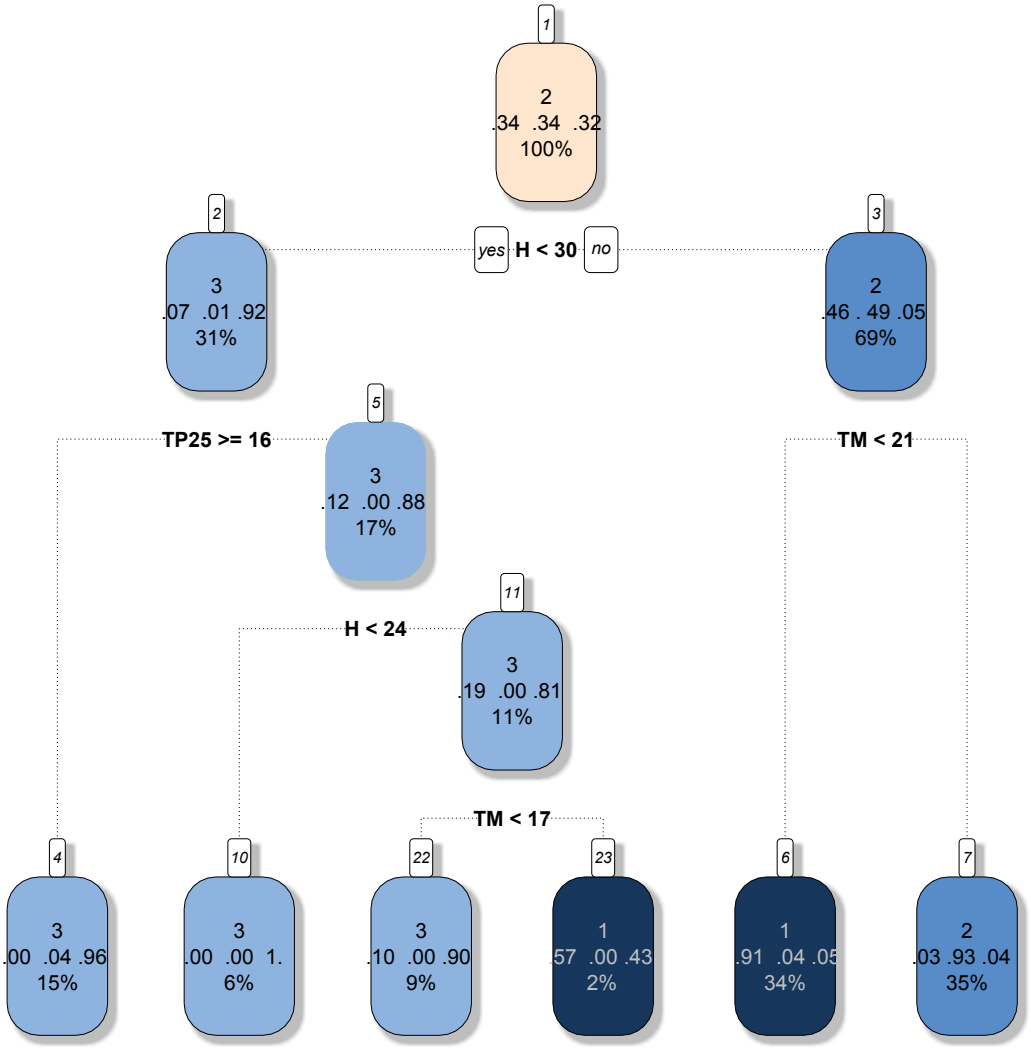


Figura 6.5: Representação do modelo *Decision Tree, CART* treinado com o conjunto de dados de treino da unidade Alpino Atlântico

as instâncias pertencem ao agrupamento em questão. No sentido inverso à regra está definido o terceiro nó onde estão presentes, quase exclusivamente, os agrupamentos *C1* e *C2*. Entre os dois últimos agrupamentos, a regra que melhor os separa baseia-se no valor de temperatura média ambiente donde resultam dois nós com frequências relativas de cada agrupamento superiores a 90%. Como observado também durante a análise de perfis de consumo, Secção 5.3, a distinção entre o agrupamento *C1* e *C3* é mais ambígua, refletido também no modelo de classificação que elaborou mais duas regras para encontrar a melhor classificação possível.

Após o processo de modelação do algoritmo de classificação, e ainda no processo de treino, o conjunto de dados de treino foi separado em sub-conjuntos com base na

classificação proveniente do estudo de perfis. Para cada sub-conjunto foram então modelados algoritmos de regressão. O processo de avaliação ou estimativa do conjunto de dados de teste foi definido por duas etapas. Na primeira etapa a instância do conjunto era classificado pelo modelo CART e, consoante a classificação atribuída, a instância era posteriormente avaliada pelo modelo de regressão respetivo.

Aquando da geração do algoritmo Regressão Linear com os diversos sub-conjuntos foi utilizada a técnica *Stepwise selection* que resultou nas funções apresentadas nas Equações 6.17, 6.18 e 6.19.

$$RL_{g1} = \beta_0 + \beta_1 H + \beta_2 Q^2 + \beta_3 TM \tag{6.17}$$

$$RL_{g2} = \beta_0 + \beta_1 H + +\beta_2 TP25 \tag{6.18}$$

$$RL_{g3} = \beta_0 + \beta_1 H + \beta_2 Q^2 \tag{6.19}$$

A Tabela 6.4 apresenta as métricas de performance para cada algoritmo e por cada sub-conjunto. As métricas são apresentas por sub-conjunto p e por todo o conjunto, por exemplo, o modelo composto RL, $RL_{p1,p2,p3}$, representa a avaliação de todo o conjunto pelo modelo de regressão linear para os sub-conjuntos 1,2 e 3.

A estratégia não foi benéfica para com todos os algoritmos e os modelos compostos SVM e KNN obtiveram piores resultados. O menor volume dos novos sub-conjuntos de dados de treino, mais especificamente, um terço do volume anterior, pode ter sido um dos motivos. Para o modelo composto KNN, um outro motivo pode ter sido a perda de algumas referências locais que agora se encontrarem em sub-conjuntos diferentes. Por sua vez, o modelo composto RL consegui ajustar-se melhor às possíveis tendências de cada um dos sub-conjuntos e obteve a melhor performance entre todos os modelos. No entanto, é de denotar a diferença do erro aquando da análise do sub-conjunto 3 dos restantes dois sub-conjuntos. A análise da dispersão do erro residual do modelo composto RL identificou a não-correlação entre o erro residual, homogeneidade de variância e uma distribuição quase gaussiana nos três modelos gerados. Porém os modelos RL_{g1} e RL_{g3} revelaram alguns valores extremos desviantes de uma hipotética distribuição gaussiana, especialmente o erro residual do modelo RL_{g1}, o que já seria esperado dado elevado valor de RMSE e a sua diferença para o valor de MAE.

Como referido, em todos os sub-conjuntos, o modelo com melhor performance foi o modelo RL e como tal o melhor conjunto foi definido como a combinação dos modelos RL, mais propriamente $RL_{p1,p2,p3}$. Visto que o modelo obteve a melhor performance até ao momento foi igualmente considerado como modelo candidato.

6.2.4 Estimativa utilizando Janelas Temporais

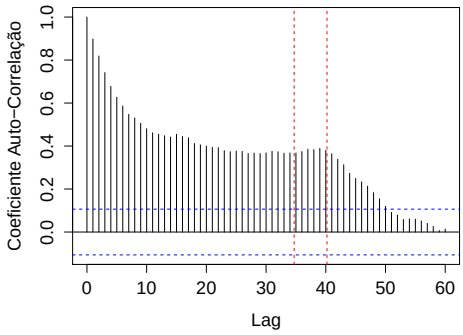
Os modelos gerados até esta fase foram baseados exclusivamente nos valores de ocupação na unidade hoteleira e nos valores de temperatura ambiente no local. Com o objetivo de

Tabela 6.4: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos

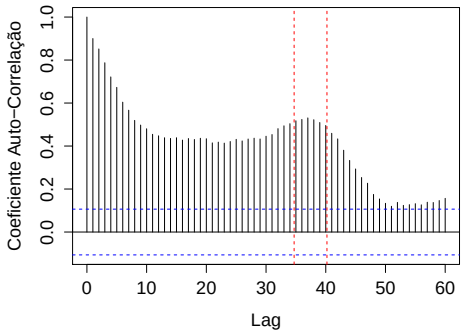
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	146.7	105.9	16.2
	2	125.6	96.7	15.4
	3	412.0	270.5	20.6
$RL_{p1,p2,p3}$	-	261.9	156.9	17.3
KNN	1	182.4	133.2	20.6
	2	143.7	113.2	16.8
	3	485.8	330.4	24.5
$KNN_{p1,p2,p3}$	-	309.1	190.5	20.3
SVM	1	208.9	142.9	22.1
	2	143.7	113.2	16.8
	3	512.4	372.3	29.1
$SVM_{p1,p2,p3}$	-	331.6	206.7	22.1
Melhor Conjunto				
$RL_{p1,p2,p3}$	-	261.9	156.9	17.3

diminuir o valor de MAPE do modelo candidato foi estudada a hipótese de incluir valores de consumo prévios como variáveis explicativas. Para o efeito, considerou-se analisar a possível dependência entre instâncias, i.e., a possível correlação entre o consumo de um serviço a estimar e o consumo do mesmo serviço z dias antes. A Figura 6.6 apresenta, por serviço, a média de auto-correlação das observações por cada z dias anteriores, ao longo de 60 dias. Para esta ilustração foi utilizado todo o conjunto de dados da unidade Alpino Atlântico pois o intuito era verificar se de facto existe alguma relação entre consumos prévios e o corrente. Na análise de auto-correlação de uma variável, o período t pelo qual se analisa a possível auto-correlação, neste caso o número z de dias, é usualmente denominado de *Lag*. Na Figura 6.6 o tracejado azul define o intervalo de confiança para o qual o coeficiente de correlação obtido poderá ter ocorrido simplesmente por um cenário aleatório.

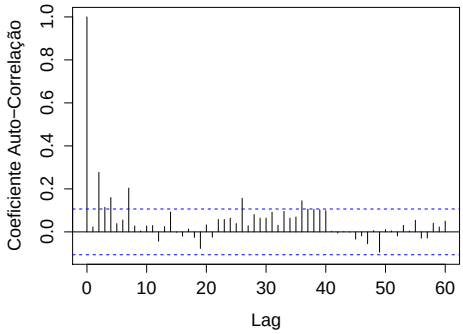
Os consumos de eletricidade e água, Figuras 6.6(a) e 6.6(b), juntamente com as variáveis de ocupação, hóspedes e quartos ocupados Figuras 6.6(d) e 6.6(e), apresentam uma tendência decrescente no coeficiente de auto-correlação. No entanto, a auto-correlação dos consumos de eletricidade e água tende a estabilizar até muito perto dos 40 dias de *Lag* onde decresce abruptamente. A auto-correlação das variáveis de ocupação decresce rapidamente até perto do 15º dia onde volta a crescer até muito perto do 40º dia onde, de igual forma, decresce abruptamente. Por sua vez, o consumo de água, Figura 6.6(c) não apresenta praticamente qualquer indício de auto-correlação ao longo dos 60 dias de *Lag* dado que o coeficiente de correlação encontra-se na maioria das vezes dentro do intervalo de confiança. Nas figuras relevantes foram assinaladas duas fronteiras, a



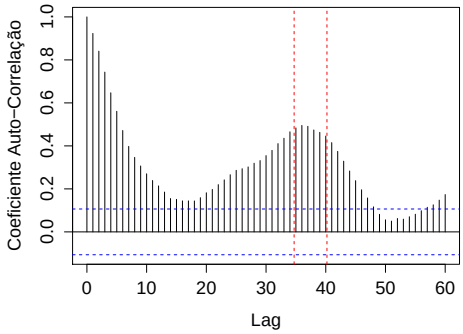
(a) Electricidade (EM)



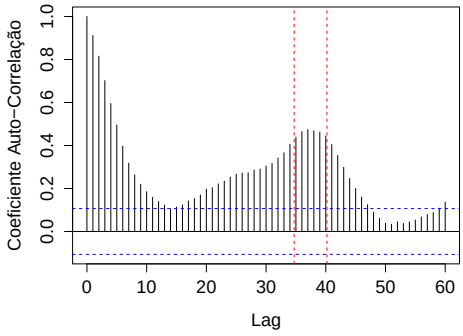
(b) Gás (GM)



(c) Água (AM)



(d) Hóspedes (H)



(e) Quartos (Q)

Figura 6.6:

tracejado vermelho, onde foi identificado o período mais longínquo em que o índice de auto-correlação indicou uma pequena subida e ainda se manteve razoavelmente estável. O objetivo proposto foi avaliar se a inclusão do valor médio das variáveis dentro do período assinalado era benéfico à modelação dos algoritmos de regressão. Foi traçado um período de 5 dias, do 35º ao 40º dia ao invés de um único dia de forma a obter um valor

médio esperado de correlação e não estar dependente de flutuações diárias. O período foi definido como o mais longínquo para maximizar o período entre o valor consumido e a estimativa, .i.e., os modelos gerados por esta técnica permitem, com base no consumo médio de serviços e de ocupação nos últimos 5 dias e com os valores esperados de ocupação e de temperatura a 35 dias, obter estimativas de consumo de serviços para esse período. Por sua vez, o comportamento do índice de auto-correlação das variáveis de ocupação também contribuíram para a definição do período em causa.

Tabela 6.5: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos

Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	149.8	116.3	18.9
	2	114.4	94.1	15.4
	3	401.2	274.2	20.5
$RL_{g1,g2,g3}$	-	254.3	159.5	17.9
KNN	1	190.4	136.3	22.0
	2	173.9	142.0	22.1
	3	541.6	384.7	29.6
$KNN_{g1,g2,g3}$	-	345.8	221.8	24.6
$KNN_{completo}$	-	327.58	214.8	24.5
SVM	1	172.3	123.9	20.2
	2	140.0	113.9	18.1
	3	360.4	247.8	19.0
$SVM_{g1,g2,g3}$	-	242.5	161.0	18.9
$SVM_{completo}$	-	245.2	157.4	17.6
Melhor Conjunto				
$RL_{g1,g2}$ SVM_{g3}	-	233.0	150.7	17.4

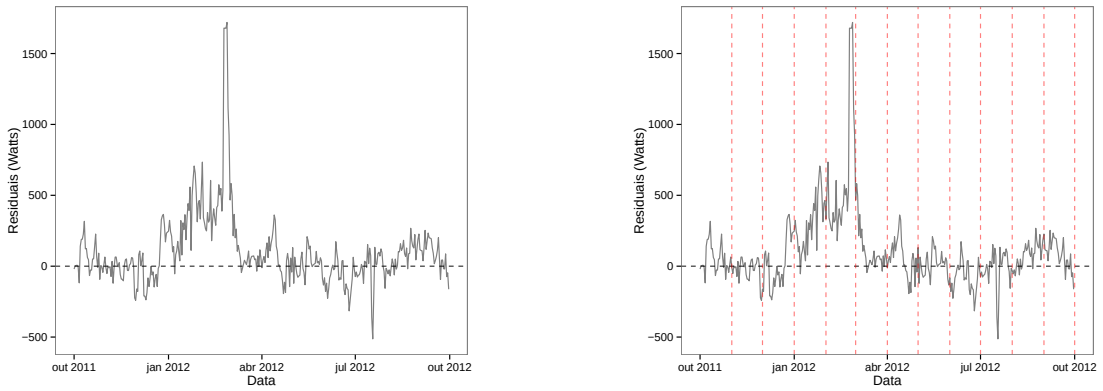
A cada instância do conjunto de dados foi adicionado o valor médio, de todas as variáveis, do período entre o 35º e o 40º dia anterior. Isso significou a exclusão das primeiras 40 instâncias do conjunto de dados, cerca de 10% do volume total, que por sua vez eram maioritariamente do sub-conjunto 3. A Tabela 6.5 apresenta os resultados obtidos com esta estratégia. Os algoritmos que foram modelados com todo o conjunto de dados de treino foram referenciados por *completo*, e os modelados com a estratégia de perfis de consumo definida na secção anterior foram agora referenciados por $g1, g2, g3$. O acréscimo das variáveis de consumos históricos não se mostrou benéfico para a modelação dos algoritmos RL e KNN. De novo, a cardinalidade das variáveis dos conjuntos em relação ao seu volume não é benéfico para o algoritmo KNN que apresentou um decréscimo de performance em todos os sub-conjuntos e na sua globalidade. Por sua vez, e mesmo aplicando *Stepwise Selection*, o algoritmo RL apresentou um declínio de performance no primeiro e terceiro sub-conjunto, precisamente os sub-conjuntos aos quais foram retiradas

algumas instâncias. No entanto, o algoritmo SVM beneficiou bastante com a inclusão das novas variáveis atingindo a melhor performance entre todos para a avaliação do agrupamento 3.

A definição do melhor conjunto passou pela escolha dos modelos que melhor se comportaram em cada sub-conjunto, tendo sido assim definido pelos modelos $RL_{g1,g2}$ SVM_{g3}. Apesar de o modelo $RL_{g1,g2}$ SVM_{g3} ter obtido um valor de MAPE um ponto percentual acima do modelo $RL_{p1,p2,p3}$ da secção anterior, os valores de RMSE e MAE foram inferiores e como tal o modelo $RL_{g1,g2}$ SVM_{g3} foi considerado como o novo modelo candidato.

6.2.5 Análise do Erro de Estimativa sob Diferentes Perspetivas

A Figura 6.7(a) apresenta o erro residual do modelo candidato $RL_{g1,g2}$ SVM_{g3} ao longo do período do conjunto de dados de teste. O erro residual apresenta uma oscilação entre valores positivos e negativos bastante elevada. Como tal, o somatório do erro ao final de um período t pode ser amortizado por esta oscilação. A Figura 6.7(b) apresenta, em linhas verticais vermelhas e a tracejado, períodos mensais aos quais se poderia calcular os valores de MAE_t e MPE.

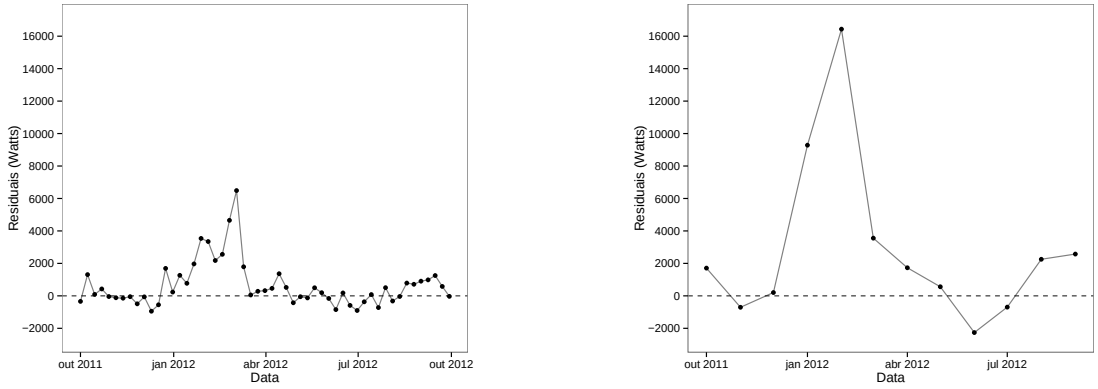


(a) (b)

Figura 6.7: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo candidato e o conjunto de dados de teste da unidade Alpino Atlântico

O erro acumulado, e_t para os períodos semanal e mensal são apresentados, respetivamente, nas Figuras 6.8(a) e 6.8(b).

O erro acumulado semanal, Figura 6.8(a), é um sinal que ainda oscila bastante no entanto a oscilação do sinal no erro acumulado mensal, Figura 6.8(b), é mais subtil. Em ambas as figuras verifica-se que o modelo sub-estimou o sinal no período compreendido entre Janeiro de 2012 e Abril de 2012, visível no pico de erro acumulado apresentado em ambas as Figuras 6.8(a) e 6.8(b). Como a métrica MAE_t é interpretada sobre a mesma granularidade que a inicial, i.e., sob uma granularidade diária, a sua comparação com os resultados obtidos anteriormente é direta. O seu significado porém indica que é o valor



(a) (b)

Figura 6.8: Análise do erro acumulado da estimativa de Eletricidade (EM) com o modelo de candidato e o conjunto de dados de treino da unidade Alpino Atlântico

médio de erro esperado por dia se for realizada uma estimativa para o total do consumo semanal ou mensal. O $MAPE_t$ é por sua vez a erro percentual obtido no final do período t, isto é, a diferença percentual entre o verdadeiro consumo total e o total estimado relativo ao verdadeiro consumo total.

Tabela 6.6: Alpino Atlântico

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Semana)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Mês)
Eletricidade	$RL_{g1,g2}\ SVM_{g3}$	150.7	17.4	132.4	14.6	107.1	11.1

A Tabela 6.6 apresenta o erro obtido sobre as diferentes perspetivas temporais. Perante uma perspetiva semanal do erro do modelo candidato, o erro médio absoluto diário reduziu perto de 12,1% em comparação com a perspetiva diária. Como tal, o menor erro médio absoluto diário do modelo candidato estabeleceu-se em 107,1 Watts/Dia perante o consumo médio de 780 Watts/Dia por hóspede.

6.2.6 Síntese de Resultados

Nesta secção são apresentados os modelos candidatos, para cada serviço e unidade hoteleira, e as respetivas métricas de performance sob as diferentes perspetivas temporais do erro dos modelos. O resumo do processo de análise por cada serviço e unidade hoteleira encontra-se em Apêndice E.

Na unidade Alpino Atlântico, Tabela 6.7, apenas o modelo para a estimativa do consumo de eletricidade beneficiou quer da estratégia de classificação do conjunto de dados quer da estratégia de inclusão de consumos históricos. Os modelos para a estimativa do consumo de água e gás que obtiveram melhores resultados foram os de RL, com base na ocupação e com base na ocupação e temperatura ambiente respetivamente. Os sinais de

Tabela 6.7: Modelos Candidatos em Alpino Atlântico

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Semana)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Mês)
Eletricidade	$RL_{g1,g2}\ SVM_{g3}$	150.7	17.4	132.4	14.6	107.1	11.1
Água	$RL(Q)$	2.898	29.8	1.56	15.0	1.32	13.3
Gás	RL	0.337	6.3	0.268	5.25	0.214	4.3

erro residual dos modelos para a estimativa de eletricidade e água eram os mais oscilantes e como tal obtiveram substancialmente melhores resultados sob sob a perspetiva semanal e mensal.

Tabela 6.8: Modelos Candidatos em Quinta de S.João

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Semana)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Mês)
Eletricidade	$RL_{p1,p2,p3}$	68.7	7.1	58.8	6.2	56.3	6.1
Água	$RL(H+H^2)$	1.795	12.1	1.04	7.2	0.752	5.3
Gás	RL	0.729	8.3	4.74	34.1	4.70	33.9

Na unidade Quinta de S.João o algoritmo preponderante foi o de Regressão Linear. Repetindo as características da unidade hoteleira Alpino Atlântico, apenas o modelo para a estimativa de eletricidade obteve benefícios em utilizar a estratégia de classificação prévia do conjunto de dados. O sinal do erro residual do modelo para a estimativa de consumo de gás sub-estimou o consumo real durante todo o período. Por consequência e perante este cenário a perceção do erro absoluto diário perante outras perspetivas é prejudicial, tendo o valor de MAE_t subido de 0,7l na perspetiva diária para 4,74l e 4,70l nas perspetivas semanal e mensal respetivamente. O modelo para a estimativa de água, tal como no Hotel Alpino Atlântico, foi o que beneficiou mais da leitura do erro perante outras perspetivas.

Tabela 6.9: Modelos Candidatos em Quinta das Vistas

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Semana)	MAE_t (Watts/Dia)	$MAPE_t$ (% /Mês)
Eletricidade	$RL_{p1,p2,p3}$	114.2	6.4	94.0	5.4	87.4	4.9
Água	$RL(H)$	6.807	28.3	5.29	22.0	4.28	18.2
Gás	SVM_{p1} RL_{p2} SVM_{p3}	6.27	131.1	4.87	82.8	3.20	38.5

Na unidade hoteleira Quinta das Vistas verificou-se o cenário contrário à unidade hoteleira Quinta de S.João com o modelo para a estimativa do consumo de gás. Neste

estabelecimento o modelo sob-estimou o verdadeiro consumo, especialmente durante os meses do Verão, e como tal agravou o MAE_t perante analise do erro sobre outras perspetivas. O modelo para a estimativa do consumo de eletricidade voltou a beneficiar da estratégia de classificação prévia e o modelo para a estimativa do consumo de água foi definido com uma RL com base no número de hóspedes presentes na unidade hoteleira. Ambos os modelos para a estimativa de eletricidade e água beneficiaram da análise do erro perante outras perspetivas.

6.2.7 Conclusão

Após o processo de modelação e identificação dos modelos que melhor estimam o consumo de cada serviço em cada unidade hoteleira surge a questão sobre a sua verdadeira utilidade. Dada a inexistência de outros estimadores para comparação foi adotado um estimador que define a sua estimativa como a média do consumo ocorrido no passado, i.e., utilizará a média do consumo de cada serviço do conjunto de dados de treino como estimativa para o conjunto de dados de teste.

Tabela 6.10: Comparação entre o modelo candidato para cada serviço e unidade hoteleira com um estimador com base no valor médio do serviço

Hotel	Serviço	Consumo Médio (Dia/Hóspede)	Modelo Candidato ($MAPE_t$)	Modelo V. Médio (MAPE)	Benefício (%)
Alpino Atlântico	Eletricidade	780.0 Watts	11.1	29.8	62.7
	Água	9.08 Litros	13.3	29.8	55.3
	Gás	5.73 Litros	4.3	26.0	83.4
Quinta das Vistas	Eletricidade	1742.1 Watts	4.9	35.9	86.3
	Água	23.9 Litros	18.2	38.3	52.4
	Gás	10.3 Litros	38.5	196.0	80.3
Quinta de S.João	Eletricidade	878.8 Watts	6.1	21.3	71.3
	Água	13.9 Litros	5.3	15.8	66.4
	Gás	9.1 Litros	33.9	27.7	-22.0

A Tabela 6.10 apresenta a comparação entre o modelo candidato de cada serviço em cada unidade hoteleira e o modelo de valor médio. É igualmente apresentado o consumo médio por serviço, dia e hóspede em cada estabelecimento. O benefício da utilização do modelo de regressão excedeu os 52% em todas as unidades e serviços à exceção da estimativa do consumo de gás na unidade hoteleira Quinta de S.João. Não obstante, a referida unidade hoteleira apresentava valores atípicos no consumo de gás que influenciou a modelação dos modelos analíticos ao longo de todo processo.

Em suma, a utilização de modelos analíticos de regressão para a estimativa do consumo de serviços revelou-se benéfico em comparação com estimadores mais simples como valores médios de dados históricos.

6.3 Estimativa perante Granularidade Horária

Da informação disponível nos conjuntos de dados apenas as variáveis relativas ao consumo de serviços e a variável relativa à temperatura ambiente se encontravam sob uma granularidade horária enquanto que a informação relativa à ocupação se encontrava sob uma granularidade diária. Como tal, foi tomada a opção de, perante uma granularidade horária, procurar estimar o consumo do serviços com base no próprio sinal.

A secção está estruturada em três sub-secções. A primeira sub-secção apresenta o estudo Mobis (2012) que procura estimar a curto prazo um sinal perante um cenário semelhante. A segunda sub-secção demonstra o sinal das variáveis de consumo de serviços e a terceira sub-secção apresenta o estudo à estimativa dos consumos de serviços na unidade hoteleira Alpino Atlântico. A quarta sub-secção apresenta uma síntese dos resultados da análise aplicada a todos os serviços e unidades hoteleiras.

Os conjuntos de dados de cada unidade hoteleira foram repartido em dois conjuntos, um conjunto de dados para treino (e validação) correspondente à primeira janela anual e outro para a fase de teste correspondente à segunda janela anual, Secção 4.3.

6.3.1 Técnica de estimativa com base no próprio sinal

A análise à estimativa do consumo de serviços apoiou-se no estudo realizado em Mobis (2012). O estudo desenvolvido procurou estimar o volume de tráfico com base em estimadores de informação corrente e estimadores de valor médio histórico. A razão pela qual ambos os estimadores foram adotados foi porque os estimadores de informação corrente tinham provado ser mais úteis aquando de previsões a curto prazo e os estimadores de valor médio histórico mais interessantes aquando de previsões a longo termo. A utilização de ambos os estimadores provou ser útil visto que, no estudo em causa, não era expectável que houvesse alterações abruptas ao valor corrente num curto período de tempo e que, a longo termo, a falta de influência do valor corrente fosse colmatada pelo estimador de médias históricas. Para conjugar ambos os estimadores foi utilizado o algoritmo de regressão linear. Para cada período a estimar foi gerado um modelo de regressão linear cujas variáveis independentes eram definidas pelas previsões dos estimadores de valor corrente e de médias históricas às quais o modelo de regressão linear atribuía coeficientes consoante o seu peso na previsão do período a estimar. Uma última estratégia do estudo passou por adicionar mais variáveis ao modelo de regressão linear, tais como a velocidade média no local, de forma a auxiliar o modelo a explicar a variação do tráfico.

A técnica do valor corrente, VC define a sua estimativa $f(x)$ com base no último valor conhecido X_t, Equação 6.21, e é uma técnica que apesar de simples é bastante utilizada para a estimativa do valor seguinte dada a sua precisão neste cenário específico, Hyndman (2014).

Por sua vez, a técnica de médias históricas, MH, também conhecida por médias sazonais, procura definir valores médios para cada momento do período cíclico do sinal. Por

exemplo, na presença de um conjunto de dados com granularidade horária e supondo um período cíclico de 24 horas, a técnica de médias históricas define uma estimativa para cada hora através do valor médio dos dados histórico para essa mesma hora. No estudo Mobis (2012), a média histórica foi aplicada sobre o período de 24 horas,ou seja, foi obtido o valor médio para cada hora de todo o conjunto de dados de treino. Adicionalmente, foi considerado que o ciclo do sinal poderia ser diferenciado por ser ou não fim-de-semana e como tal foram definidas médias históricas por hora e pela diferenciação de ser ou não fim-de-semana.

A conjugação do algoritmo de Regressão Linear, Secção 6.1, com os estimadores VC e MH caracteriza-se pela utilização dos estimadores simples como variáveis independentes na modelação do algoritmo Regressão Linear. Por exemplo, considerando um cenário com um conjunto de dados com granularidade horária e um período de 24 horas, se às 0 horas pretender-se estimar o sinal ao fim de 12 horas, o estimador VC irá indicar o último valor conhecido até ao momento X_t ou seja o valor da última instância que corresponde às 0 horas. Por sua vez, o estimador MH irá indicar o valor médio do conjunto de dados de treino para as 12 horas. Ambos os resultados são então utilizados pelo modelo de regressão linear para obter a estimativa final. A função do modelo de RL é definida pela Equação 6.20.

$$RL(x_p) = \beta 0 + \beta 1 VC(x_p) + \beta 2 MH(x_p) \tag{6.20}$$

A técnica estuda em Mobis (2012) apresentou-se como interessante para a estimativa de consumo de serviços visto que era expectável que o consumo de cada serviço em cada unidade hoteleira, dada a natureza do consumo, estivesse relacionado com o consumo ocorrido no momento anterior. Era igualmente expectável que o consumo de cada serviço tivesse um comportamento diário característico e que por sua vez, o número de hóspedes presentes no estabelecimento e a temperatura ambiente no local, pudessem auxiliar na caracterização do sinal.

6.3.2 Análise dos sinais de consumo de serviços

Previamente à análise de estimativa dos consumos de serviços na unidade hoteleira Alpino Atlântico analisou-se o comportamento do sinal de consumo de cada serviço a cada 24 horas, Figuras 6.9, 6.10 e 6.11. O sinal de consumo foi igualmente separado por estações do ano dado que durante a análise exploratória observou-se que existiam diferenças de valor entre estações do ano em alguns serviços. Foram também assinalados nas figuras os valores médios de cada sinal por estação do ano com a cor respetiva a cada estação do ano. Com esta análise procurou-se observar se o sinal de consumo de cada serviço a 24 horas apresentava algum padrão e qual a variação do mesmo.

Os sinais do consumo de eletricidade e de gás, Figuras 6.9 e 6.11, aparentam ter ambos uma menor variância durante a estação de Verão e uma maior variância durante a estação de Inverno. Por sua vez, o valor de consumo médio de eletricidade difere entre

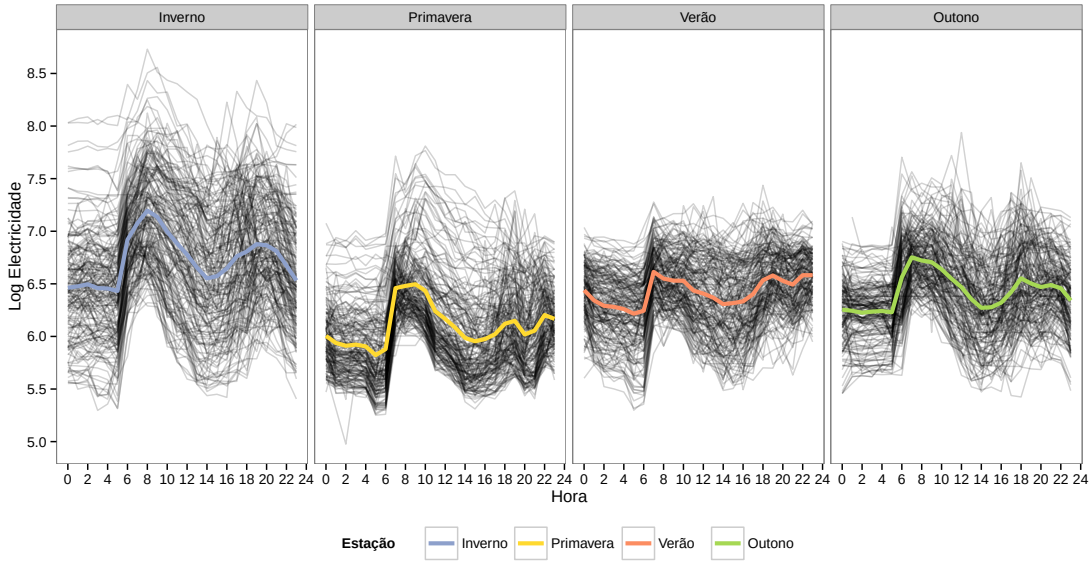


Figura 6.9: Comportamento do sinal de consumo de Eletricidade a 24 horas e por estação do ano na unidade hoteleira Alpino Atlântico

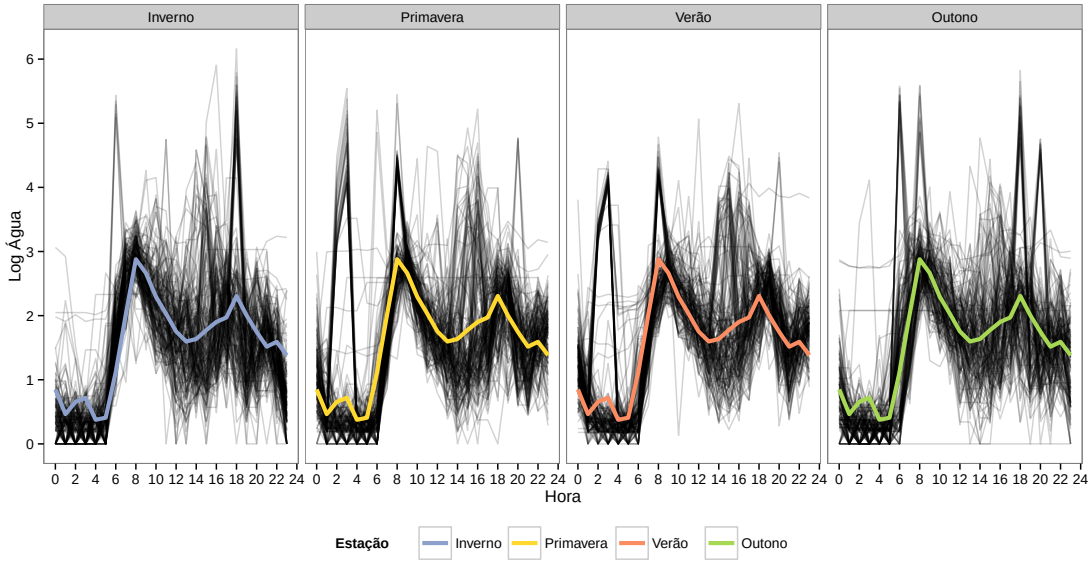


Figura 6.10: Comportamento do sinal de consumo de Água a 24 horas e por estação do ano na unidade hoteleira Alpino Atlântico

estações do ano enquanto que o valor de consumo médio de gás aparenta variar muito pouco entre diferentes estações do ano. O sinal de consumo água, Figura 6.10, apresenta alguns comportamentos distintos entre as estações de Outono e Inverno e as estações de Primavera e Verão. No entanto, o valor médio de consumo não aparenta variar entre as diferentes estações do ano.

Todos os sinais de consumo aparentam apresentar um comportamento cíclico de 24

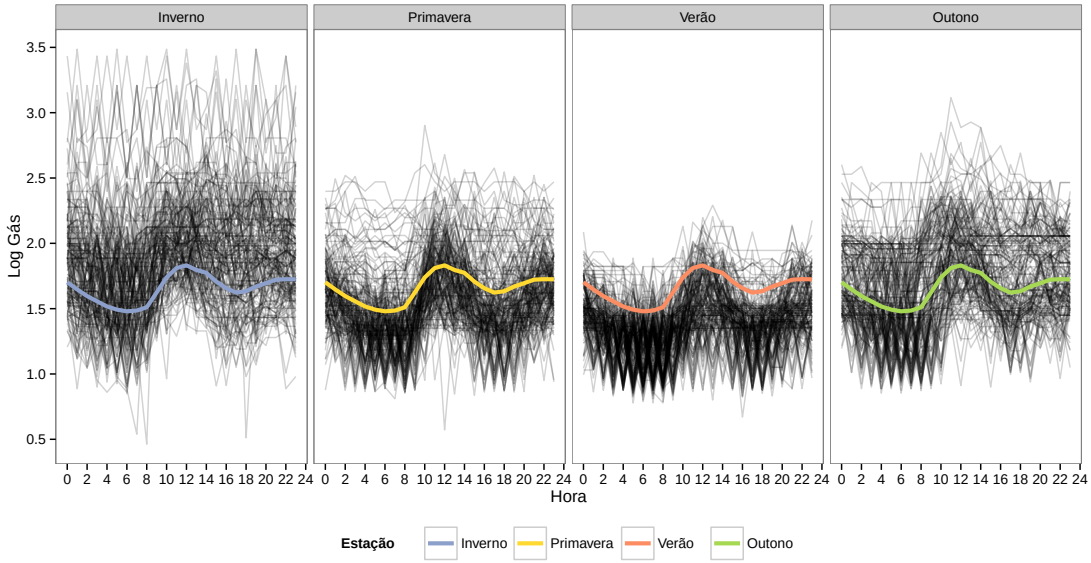


Figura 6.11: Comportamento do sinal de consumo de Gás a 24 horas e por estação do ano na unidade hoteleira Alpino Atlântico

horas, com um período de baixo consumo durante a madrugada, um pico de consumo durante as horas da manhã, um decréscimo de consumo durante as primeiras horas da tarde e uma ligeira subida nas últimas horas da tarde até ao fim da noite.

6.3.3 Análise da estimativa do sinal

.

Tendo em conta a granularidade horária dos conjunto de dados, o período considerado a estimar variou entre uma 1 hora de desfasamento do valor corrente a 47 horas do valor corrente. Seguindo a estratégia utilizada no estudo Mobis (2012), foram gerados, para cada período (hora), os seguintes modelos para a estimativa de consumo de cada serviço: valor corrente, cuja estimativa tem por base na última instância conhecida x_t, Equação 6.21; regressão linear com base no valor corrente, Equação 6.22; regressão linear com base no valor corrente e no valor médio histórico da hora a estimar consoante a estação do ano, Equação 6.23; regressão linear com base no valor corrente, no valor médio histórico da hora a estimar consoante a estação do ano, na temperatura ambiente, no número de hóspedes presentes e no número de quartos ocupados, Equação 6.24.

$$VC(x) = x_t \tag{6.21}$$

$$RL_{cv}(x) = \beta_0 + \beta_1 VC(x) \tag{6.22}$$

$$RL_{CV+MH}(x) = \beta_0 + \beta_1 VC(x) + \beta_2 MH(x) \tag{6.23}$$

$$RL_{CV+MH+TM+H+Q}(x) = \beta_0 + \beta_1 VC(x) + \beta_2 MH(x) + \beta_3 TM + \beta_4 H + \beta_5 Q \quad (6.24)$$

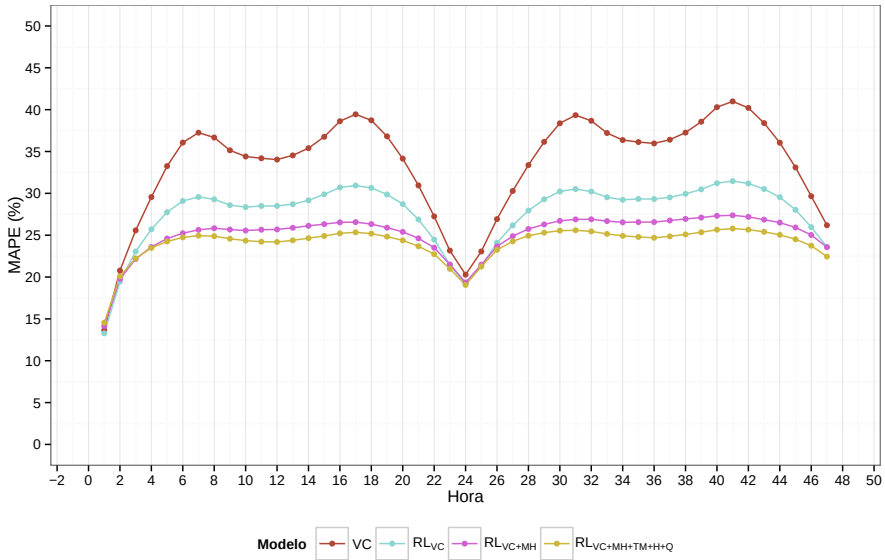


Figura 6.12: Índice MAPE para a estimativa de consumo de Eletricidade a 48 horas no conjunto de dados Alpino Atlântico

A Figura F.10 apresenta o comportamento dos modelos, por cada hora de desfasamento, relativamente ao consumo de eletricidade na unidade hoteleira Alpino Atlântico. Dentro do período a estimar selecionado, o modelo de regressão linear com base no valor corrente é o modelo com melhor performance para previsões até 2 horas de desfasamento da hora do valor corrente. Após um desfasamento igual ou superior a 3 horas, o modelo de regressão linear que engloba as variáveis de ocupação e temperatura ambiente além dos estimadores base (valor corrente e médias históricas) foi o que obteve melhor performance não excedendo os 25% de MAPE dentro de esse período. O comportamento dos modelos refletem-se a cada ciclo horário de 24 horas.

A Figura F.10 apresenta o comportamento dos modelos, por cada hora de desfasamento, relativamente ao consumo de água na unidade hoteleira Alpino Atlântico. Os modelos que incluem o estimador de médias histórias apresentam a melhor performance e um valor de MAPE inferior e praticamente constante a 50%, no entanto, o estimador de valor corrente apresenta valores de MAPE bastante atípicos. De forma a tentar perceber o motivo pelo qual os valores de MAPE eram tão elevados para o modelo VC, observou-se o valor de MAPE, por hora do dia, para o modelo VC com um desfasamento de uma hora (modelo que obteve perto de 150% de MAPE), Figura 6.14. Pela figura foi possível verificar que o grande impacto no valor de MAPE ocorria entre as 23 e as 6 horas. Durante a análise exploratória, Figura 3.18(a), observou-se que durante este período, o consumo de água na unidade hoteleira Alpino Atlântico registou por diversas vezes um consumo de zero.

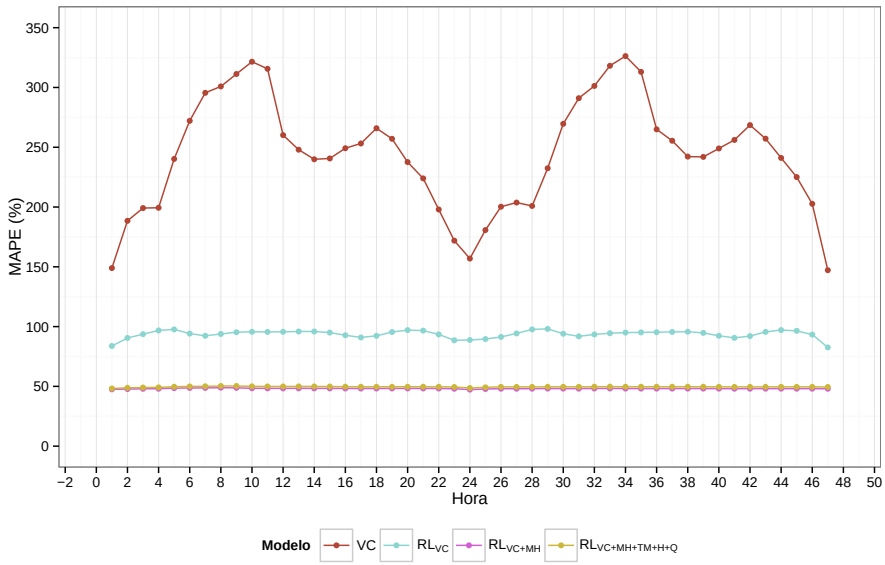


Figura 6.13: Índice MAPE para a estimativa de consumo de Água a 48 horas no conjunto de dados Alpino Atlântico

A utilização da escala logarítmica ($log(x+1)$) manteve estes registos como zero após a transformação e como tal, aquando do cálculo de MAPE, foi necessário utilizar uma aproximação do valor zero à primeira casa decimal, i.e., $0,1$. A utilização de um valor tão baixo provoca um elevado valor de MAPE aquando de uma estimativa mal adequada. O valor corrente estima com base no valor precedente e como tal a estimativa é muitas vezes desadequada o que por sua vez acabou por contribuir para um elevado valor de MAPE.

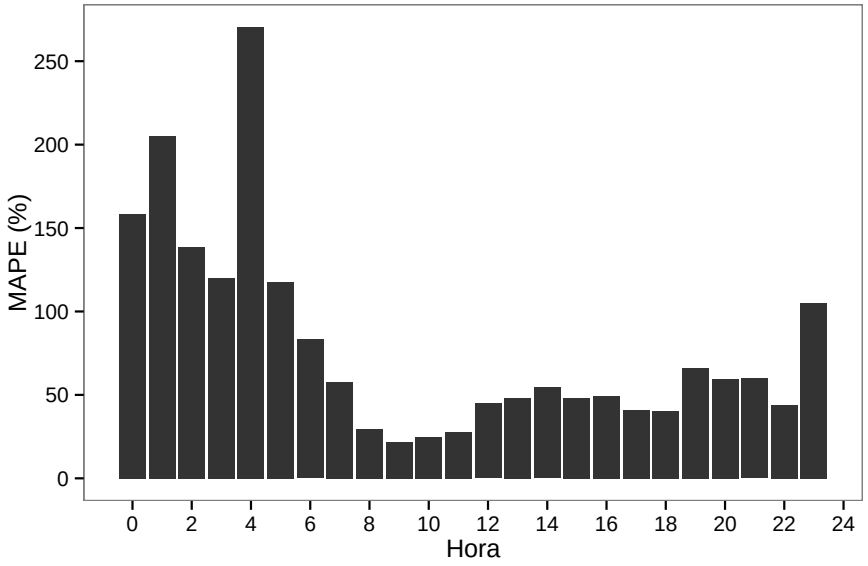


Figura 6.14: Índice MAPE para a estimativa de consumo de Água a 48 horas no conjunto de dados Alpino Atlântico

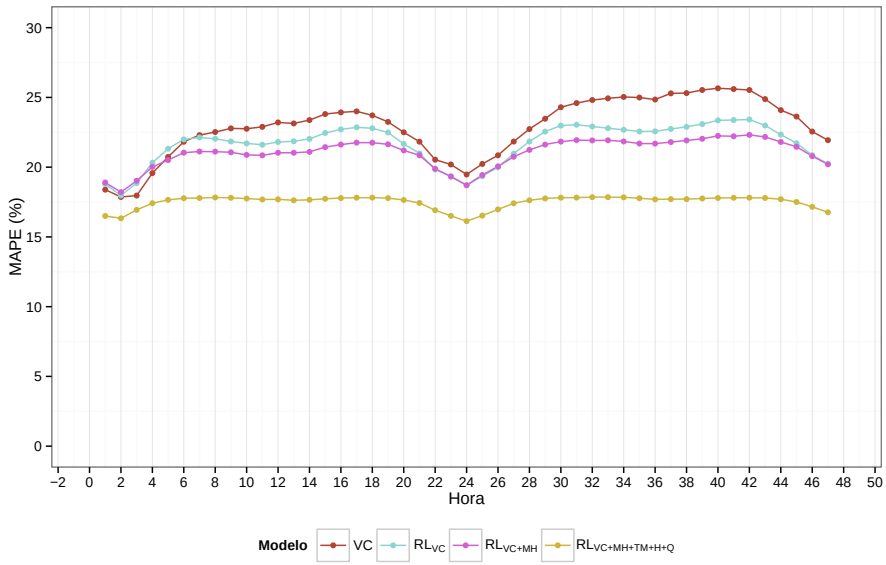


Figura 6.15: Índice MAPE para a estimativa de consumo de Gás a 48 horas no conjunto de dados Alpino Atlântico

No caso do consumo de gás na unidade hoteleira Alpino Atlântico, o modelo de regressão linear que engloba as variáveis de ocupação e temperatura ambiente além dos estimadores base (valor corrente e médias históricas) foi o que obteve a melhor performance independentemente do período a estimar. O modelo apresentou também uma variação muito ligeira ao longo do período em comparação com os restantes modelos.

6.3.4 Síntese de Resultados

As análises de estimativa dos sinais de consumo de serviços das unidades hoteleiras Quinta das Vistas e Quinta S.João são apresentadas em Apêndice F.

De um modo geral, para todos os serviços e unidades hoteleiras, os modelos de regressão linear com base nas variáveis de ocupação e temperatura ambiente além dos estimadores base (valor corrente e médias históricas), apresentaram a melhor performance ao longo de todo o período de 48 horas selecionado. A performance dos modelos regressivos para a estimativa do consumo de eletricidade foi bastante semelhantes à dos modelos regressivos para a estimativa do consumo de gás.

Os modelos, baseados unicamente no valor corrente, destinados à estimativa do consumo de água revelaram valores atípicos de MAPE em todas as unidades hoteleiras. Como referido na Secção 6.3.3, uma possível explicação relaciona o valor corrente com o número de observações no conjunto de dados com um consumo de água idêntico a zero.

6.3.5 Conclusão

A utilização da técnica apresentada no estudo Mobis (2012) demonstrou ser também adequada à estimativa do consumo de serviços por hóspedes em unidades hoteleiras. Em

comparação com os estimadores base de valor corrente, os modelos de regressão linear revelaram uma melhor performance para todo o período de 48 horas. A utilização de variáveis auxiliares como a temperatura ambiente, o número de hóspedes e o número de quartos alugados auxiliaram os modelos ao não penalizar muito a sua performance aquando do maior desfasamento do período da estimativa ao período do valor corrente.

CAPÍTULO 7

CONCLUSÃO E TRABALHO FUTURO

Dada a crescente relevância da atividade turística na economia global e os inerentes custos de operação relacionados com o consumo de serviços nos estabelecimentos hoteleiros, os resultados obtidos nesta dissertação tornam-se uma mais valia para futuros estudos que visam incorporar técnicas de data mining em soluções de gestão eficiente de consumo de serviços. Os objetivos inicialmente propostos por esta dissertação para o projeto Soltgest foram atingidos e com resultados interessantes.

A primeira análise do estudo concluiu que é possível identificar um perfil de consumo de serviços comum entre os estabelecimentos hoteleiros da Região Autónoma da Madeira pertencentes ao projeto Soltgest. O perfil identificado caracteriza, quantitativamente e temporalmente, o consumo dos serviços de eletricidade, água e gás por hóspede presente em cada unidade hoteleira e consoante as condições meteorológicas no local.

A segunda análise permitiu identificar modelos preditivos com a capacidade de estimar o consumo de serviços como a eletricidade, a água e o gás, com uma maior precisão que hipotéticos estimadores base de consumo de serviços. A estimativa de consumo de serviços foi estudada sob uma perspetiva de consumo diário por hóspede presente em cada unidade hoteleira e sob uma perspetiva de consumo horário a curto prazo.

Durante o estudo realizado foi incluído ao projeto informação relevante relativa às condições meteorológicas na RAM, proveniente do grupo ANA, Aeroportos de Portugal. A informação recolhida foi essencial para a determinação, mais rigorosa possível, dos valores de temperatura ambiente na região. Foi igualmente incluída informação relativa aos períodos de ocorrência de eventos sociais e culturais influentes na atividade turística, proveniente da Direção Regional do Turismo do Governo Regional da Madeira, mas a sua identificação não contribuiu para a definição de perfis de consumo por unidade hoteleira nem para a análise de estimativa de consumos de serviços.

Por fim, foram elaborados ao longo de este estudo relatórios de qualidade dos dados do projeto Solgest e foi apresentada sucintamente nesta a dissertação os factos considerados mais relevantes da análise exploratória.

O estudo realizado identificou perfis de consumo de serviços em cada unidade hoteleira e gerou modelos com capacidades preditivas de consumo de serviços, no entanto, estes baseiam-se no consumo geral de toda a infra-estrutura hoteleira e no número de hóspedes presentes diariamente no estabelecimento. Uma opinião resultante do estudo realizado é a de que uma maior sensibilidade da informação relativa à ocupação, turística e funcional, seria um requisito principal para um estudo futuro. Como por exemplo, a informação relativa ao número de hóspedes presentes em cada quarto de hóspedes e o respetivo período do dia em que o ocupavam, a faixa etária e nacionalidade de cada hóspede e o número de funcionários presentes na unidade hoteleira são apenas alguns dos exemplos que poderiam contribuir para definição de perfis de consumo de serviços e para a análise de estimativa de consumo de serviços. Além do acréscimo de informação relativa à ocupação na unidade hoteleira, seria igualmente importante o conhecimento de serviços de rotina ou pré-agendados realizados pelos funcionários da unidade hoteleira que afetassem o consumo de serviços.

BIBLIOGRAFIA

Alpaydm, E. (1999). "Combined 5× 2 cv F test for comparing supervised classification learning algorithms". Em: *Neural computation*, pp. 1885–1892.

Arthur, D. e S. Vassilvitskii (2007). "K-means++: The Advantages of Careful Seeding". Em: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. Society for Industrial e Applied Mathematics: New Orleans, Louisiana, pp. 1027–1035. ISBN: 978-0-898716-24-5.

Azevedo, A. I. R. L. (2008). "KDD, SEMMA and CRISP-DM: a parallel overview". Em:

Box, G. e N. Draper (1987). *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. ISBN: 9780471810339.

Caliński, T. e J. Harabasz (1974). "A dendrite method for cluster analysis". Em: *Communications in Statistics-theory and Methods* 3, pp. 1–27.

Casca, S. (2013). *Solgest Project Context Study .pdf*. Rel. téc. URL: `http://centria.di.fct.unl.pt/~jmp/mscs/sc/prep-d/SolgestProjectContextStudy.pdf`.

Charrad, M., N. Ghazzali, V. Boiteau e A. Niknafs (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set". Em: *Journal of Statistical Software* 61(6), pp. 1–36.

Dietterich, T. G. (1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". Em: *Neural Computation* 10(7), pp. 1895–1923. ISSN: 0899-7667.

Drucker, H., C. J. Burges, L. Kaufman, A. Smola, V. Vapnik et al. (1997). "Support vector regression machines". Em: *Advances in neural information processing systems* 9, pp. 155–161.

EnergyStar (2007a). *Facility Type:Hotels and Motels*.

EnergyStar (2007b). *Hotels : An Overview of Energy Use and Energy Efficiency Opportunities*.

Everitt Brian S., e. a. (2011). *Cluster Analysis*. ISBN: 9780470749913.

Faraway, J. J. (2004). *Linear Models with R*. Chapman snd Hall.

Fonseca, A. M. A. G. da (2006). *Oferta Turística e Relação Turismo-Ambiente na Região Autónoma da Madeira*. Rel. téc.

Governo da Madeira, D. a. R. d. E. (2012a). *Anuário Estatístico da Região Autónoma da Madeira*. Rel. téc.

Governo da Madeira, D. a. R. d. E. (2012b). *Estatísticas demográficas da Região Autónoma da Madeira*. Rel. téc.

Governo da Madeira, D. a. R. d. E. (2012c). *Madeira em Numeros 2011*. Rel. téc. Madeira.

Governo da Madeira, D. a. R. d. T. (2011). *Calendário de Animação Turística*. Rel. téc., pp. 17–18.

Governo da Madeira, D. a. R. d. T. (2012d). *Calendário de Animação Turística*. Rel. téc.

Hastie, T. J., R. J. Tibshirani e J. H. Friedman (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer: New York. ISBN: 978-0-387-84857-0.

Hyndman, L. R. J. (2014). *Forecasting : Principles & Practice*.

IBM, C.-D. (2011). *IBM SPSS Modeler CRISP-DM Guide*.

IBM, C.-D. (2012). *IBM SPSS Modeler Professional*.

Jiawei Han, M. K. e J. Pei (June 2011). *Data Mining, Concepts and Techniques*. Third. Morgan Kaufmann. ISBN: 978-0-12-381479-1.

Kurgan, L. A. e P. Musilek (2006). "A survey of Knowledge Discovery and Data Mining process models". Em: *The Knowledge Engineering Review* 21(01), pp. 1–24.

Leitao, N. C. (2011). "Tourism and economic growth: a panel data approach". Em:

Liu, M., S. L. Lai e K. C. Kuo (2012). "Economic Growth, Energy Consumption and Tourism Development in Taiwan: A Granger Causality Approach". Em: *Advanced Materials Research* 524-527, pp. 3376–3379. ISSN: 10226680.

Milligan, G. W. e M. C. Cooper (1985). "An examination of procedures for determining the number of clusters in a data set". Em: *Psychometrika* 50(2), pp. 159–179.

Milligan, G. W. e M. C. Cooper (1987). *A study of variable standardization*. College of Business, Ohio State University.

Mobis, C. (2012). *Mobis Project website*.

Murtagh, F. (2014). "Ward ' s Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward ' s Criterion ?" Em: 295(October), pp. 274–295.

Nissan, E., M.-A. Galindo e M. T. Méndez (2011). "Relationship between tourism and economic growth". Em: *The Service Industries Journal* 31(10), pp. 1567–1572.

Ounpraseuth, S., S. Y. Lensing, H. J. Spencer e R. L. Kodell (2012). "Estimating misclassification error: a closer look at cross-validation based methods." Em: *BMC research notes* 5(1), p. 656.

Rubel, F. e M. Kottek. "Observed and projected climate shifts 19012100 depicted by world maps of the Koppen-Geiger climate classification". Em: *Meteorologische Zeitschrift*.

Steinley, D. (2004). "Standardizing Variables in K-means Clustering". English. Em: *Classification, Clustering, and Data Mining Applications*. Ed. por D. Banks, F. McMorris, P. Arabie e W. Gaul. Studies in Classification, Data Analysis, and Knowledge Organisation. Springer Berlin Heidelberg, pp. 53–60. ISBN: 978-3-540-22014-5.

Tibshirani, R, G Walther e T Hastie (2001). *Estimating the number of clusters in a data set via the gap statistic*.

Turismo, O. M. do (2014). *Panorama OMT del turismo internacional*. Rel. téc.

United Nations World Tourism Organization (UNWTO) (2011). *Tourism towards 2030. Global overview. UNWTO General Assembly, 19th Session. Gyeongju, Republic of Korea, 10 October 2011*. Rel. téc.

Vendramin, L., R. J. Campello e E. R. Hruschka (2010). "Relative clustering validity criteria: A comparative overview". Em: *Statistical Analysis and Data Mining: The ASA Data Science Journal*, pp. 209–235.

World Meteorological Organization, W. (2015). *World Weather Information Service*. Acedido: 2015-03-01.

World Travel & Tourism Council (2013). *Travel & Tourism Economic Impact 2013 Portugal*. Rel. téc.

APÊNDICE

INTERPRETAÇÃO DO PROJETO

Evento	Primeiro Dia	Último Dia
Festival Passeios a Pé	terça-feira, 12 de Janeiro de 2010	sábado, 16 de Janeiro de 2010
Festival de Orientação da Madeira	terça-feira, 26 de Janeiro de 2010	quinta-feira, 28 de Janeiro de 2010
Festas de Carnaval	sábado, 13 de Fevereiro de 2010	terça-feira, 16 de Fevereiro de 2010
Madeira Island Open	quinta-feira, 8 de Abril de 2010	domingo, 11 de Abril de 2010
Festa da Flor da Madeira	sábado, 17 de Abril de 2010	domingo, 18 de Abril de 2010
Festival do Atlântico	sábado, 5 de Junho de 2010	sábado, 5 de Junho de 2010
Festival do Atlântico	sábado, 12 de Junho de 2010	sábado, 12 de Junho de 2010
Festival do Atlântico	sábado, 19 de Junho de 2010	sábado, 19 de Junho de 2010
Festival do Atlântico	sábado, 26 de Junho de 2010	sábado, 26 de Junho de 2010
Classic Rally	quarta-feira, 23 de Junho de 2010	sexta-feira, 25 de Junho de 2010
Rali Vinho da Madeira	quinta-feira, 5 de Agosto de 2010	sábado, 7 de Agosto de 2010
Festa do Vinho da Madeira	segunda-feira, 2 de Agosto de 2010	quinta-feira, 5 de Agosto de 2010
Arraial do Monte	domingo, 15 de Agosto de 2010	domingo, 15 de Agosto de 2010
Festival de Colombo	segunda-feira, 16 de Agosto de 2010	quarta-feira, 18 de Agosto de 2010
Festas do Fim-do-Ano	domingo, 26 de Dezembro de 2010	quinta-feira, 6 de Janeiro de 2011

Figura A.1: Eventos Socio-Culturais 2010

Evento	Primeiro Dia	Último Dia
Festas de Carnaval	sexta-feira, 4 de Março de 2011	terça-feira, 8 de Março de 2011
Madeira Island Open (Main Tour da PGA European Tour)	quinta-feira, 19 de Maio de 2011	domingo, 22 de Maio de 2011
Festa da Flor da Madeira	quinta-feira, 5 de Maio de 2011	domingo, 8 de Maio de 2011
Festival do Atlântico	sábado, 4 de Junho de 2011	sábado, 4 de Junho de 2011
Festival do Atlântico	sábado, 11 de Junho de 2011	sábado, 11 de Junho de 2011
Festival do Atlântico	sábado, 18 de Junho de 2011	sábado, 18 de Junho de 2011
Festival do Atlântico	sábado, 25 de Junho de 2011	sábado, 25 de Junho de 2011
Classic Rally	quinta-feira, 23 de Junho de 2011	sábado, 25 de Junho de 2011
Rali Vinho da Madeira	quinta-feira, 4 de Agosto de 2011	sábado, 6 de Agosto de 2011
Arraial do Monte	segunda-feira, 15 de Agosto de 2011	segunda-feira, 15 de Agosto de 2011
Festa do Vinho da Madeira	quinta-feira, 1 de Setembro de 2011	domingo, 4 de Setembro de 2011
Festival de Colombo	quinta-feira, 15 de Setembro de 2011	sábado, 17 de Setembro de 2011
Festival da Natureza	quarta-feira, 12 de Outubro de 2011	sábado, 15 de Outubro de 2011
Festas do Fim-do-Ano	sexta-feira, 25 de Novembro de 2011	quinta-feira, 6 de Dezembro de 2012

Figura A.2: Eventos Socio-Culturais 2011

Evento	Primeiro Dia	Último Dia
Festival Passeios a Pé	terça-feira, 12 de Janeiro de 2012	sábado, 16 de Janeiro de 2012
MOF - Festival de Orientação da Madeira	terça-feira, 26 de Janeiro de 2012	quinta-feira, 28 de Janeiro de 2012
Festas de Carnaval	sexta-feira, 17 de Fevereiro de 2012	terça-feira, 21 de Fevereiro de 2012
Madeira Island Open (Main Tour da PGA European Tour)	quinta-feira, 10 de Maio de 2012	domingo, 13 de Maio de 2012
Festa da Flor da Madeira	quinta-feira, 19 de Abril de 2012	domingo, 22 de Abril de 2012
Festival do Atlântico	sábado, 9 de Junho de 2012	sábado, 9 de Junho de 2012
Festival do Atlântico	sábado, 16 de Junho de 2012	sábado, 16 de Junho de 2012
Festival do Atlântico	sábado, 23 de Junho de 2012	sábado, 23 de Junho de 2012
Festival do Atlântico	sábado, 30 de Junho de 2012	sábado, 30 de Junho de 2012
Classic Rally	sexta-feira, 22 de Junho de 2012	terça-feira, 26 de Junho de 2012
Rali Vinho da Madeira	sexta-feira, 3 de Agosto de 2012	domingo, 5 de Agosto de 2012
Festa do Vinho da Madeira	quinta-feira, 30 de Agosto de 2012	quinta-feira, 2 de Agosto de 2012
Arraial do Monte	quarta-feira, 15 de Agosto de 2012	quarta-feira, 15 de Agosto de 2012
Festival da Natureza	sábado, 1 de Outubro de 2012	sexta-feira, 7 de Outubro de 2012
Festival de Colombo	quinta-feira, 13 de Setembro de 2012	sábado, 15 de Setembro de 2012
Festas do Fim-do-Ano	sexta-feira, 30 de Novembro de 2012	domingo, 6 de Janeiro de 2013

Figura A.3: Eventos Socio-Culturais 2012

APÊNDICE B

INTERPRETAÇÃO DOS DADOS

B.1 Descrição do Conjunto de Dados

Cardinal Direction	Degree Direction
N	348.75 - 11.25
NNE	11.25 - 33.75
NE	33.75 - 56.25
ENE	56.25 - 78.75
E	78.75 - 101.25
ESE	101.25 - 123.75
SE	123.75 - 146.25
SSE	146.25 - 168.75
S	168.75 - 191.25
SSW	191.25 - 213.75
SW	213.75 - 236.25
WSW	236.25 - 258.75
W	258.75 - 281.25
WNW	281.25 - 303.75
NW	303.75 - 326.25
NNW	326.25 - 348.75

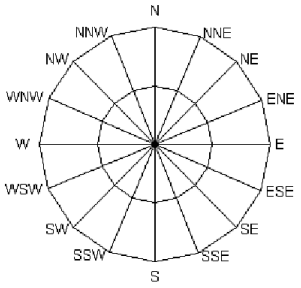


Figura B.1: Direção Cardinal e ângulo respectivo

B.2 Volume e Periodicidade dos Conjuntos de Dados

B.3 Análise Exploratória

Tabela B.1: Sumário dos registos dos sensores de consumo de serviços

Hotel	Sensor	n	#NA	Min	Max
Porto Santo Hotel	Agua	23675	732	2010-09-21 01:00:00	2013-06-03 11:00:00
	Eletricidade	676	23731	2010-09-21 01:00:00	2010-10-19 04:00:00
Meliã Hotel	Agua	15573	8834	2010-10-22 15:00:00	2012-08-03 16:00:00
	Eletricidade	11177	13230	2011-04-25 22:00:00	2012-08-03 16:00:00
	Gas	15573	8834	2010-10-22 15:00:00	2012-08-03 16:00:00
CS Madeira Hotel	Agua	15622	8785	2010-09-21 01:00:00	2012-07-02 22:00:00
	Eletricidade	15622	8785	2010-09-21 01:00:00	2012-07-02 22:00:00
	Gas	15622	8785	2010-09-21 01:00:00	2012-07-02 22:00:00
Alpino Atlântico Hotel	Agua	18855	5552	2010-09-21 01:00:00	2012-11-14 15:00:00
	Eletricidade	18855	5552	2010-09-21 01:00:00	2012-11-14 15:00:00
	Gas	18855	5552	2010-09-21 01:00:00	2012-11-14 15:00:00
Quinta das Vistas Hotel	Agua	17995	6412	2010-09-21 01:00:00	2012-10-09 20:00:00
	Eletricidade	17995	6412	2010-09-21 01:00:00	2012-10-09 20:00:00
	Gas	17995	6412	2010-09-21 01:00:00	2012-10-09 20:00:00
Reid's Hotel	Agua	23031	1376	2010-09-21 01:00:00	2013-05-07 16:00:00
	Eletricidade	23031	1376	2010-09-21 01:00:00	2013-05-07 16:00:00
	Gas	19311	5096	2011-02-22 23:00:00	2013-05-07 16:00:00
Baía Azul Hotel	Agua	24406	1	2010-09-21 01:00:00	2013-07-03 23:00:00
	Eletricidade	24406	1	2010-09-21 01:00:00	2013-07-03 23:00:00
	Gas	24406	1	2010-09-21 01:00:00	2013-07-03 23:00:00
Quintinha S. João Hotel	Agua	24407	0	2010-09-21 01:00:00	2013-07-03 23:00:00
	Eletricidade	24407	0	2010-09-21 01:00:00	2013-07-03 23:00:00
	Gas	24407	0	2010-09-21 01:00:00	2013-07-03 23:00:00

Tabela B.3: Sumário dos registos das estações meteorológicas no conjunto de dados

Hotel	Estação	n	#NA	Min	Max
Alpino Atlântico Hotel	Exterior	17525	6883	2010-09-22 01:00:00	2012-11-05 16:00:00
Alpino Atlântico Hotel	Interior	17525	6883	2010-09-22 01:00:00	2012-11-05 16:00:00
Baía Azul Hotel	Exterior	18049	6359	2010-09-21 01:00:00	2013-02-03 19:00:00
Baía Azul Hotel	Interior	18049	6359	2010-09-21 01:00:00	2013-02-03 19:00:00
Quinta das Vistas Hotel	Exterior	17820	6588	2010-09-21 01:00:00	2012-12-17 15:00:00
Quinta das Vistas Hotel	Interior	17820	6588	2010-09-21 01:00:00	2012-12-17 15:00:00

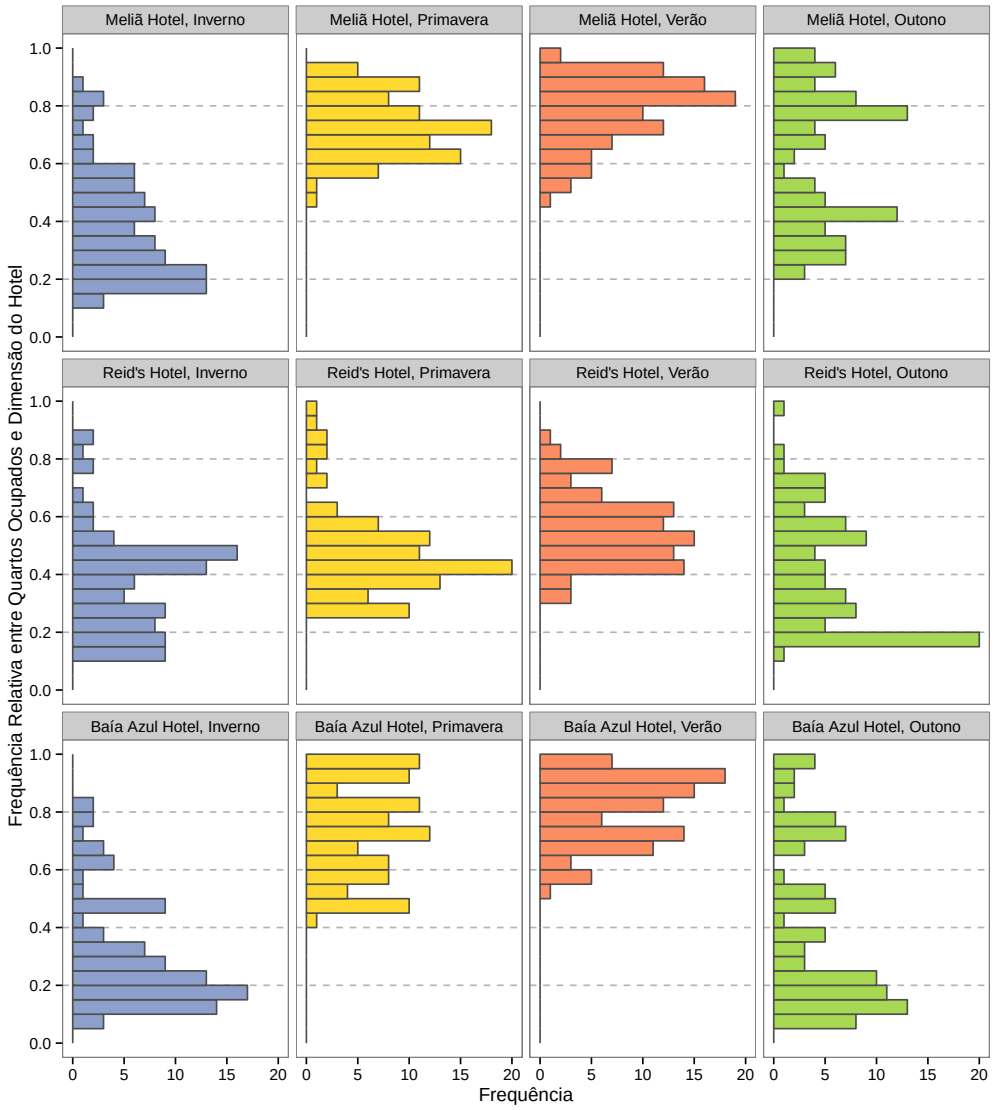


Figura B.2: Distruição da Frequência Relativa entre o número de quartos ocupados e a dimensão do hotel por cada estabelecimento

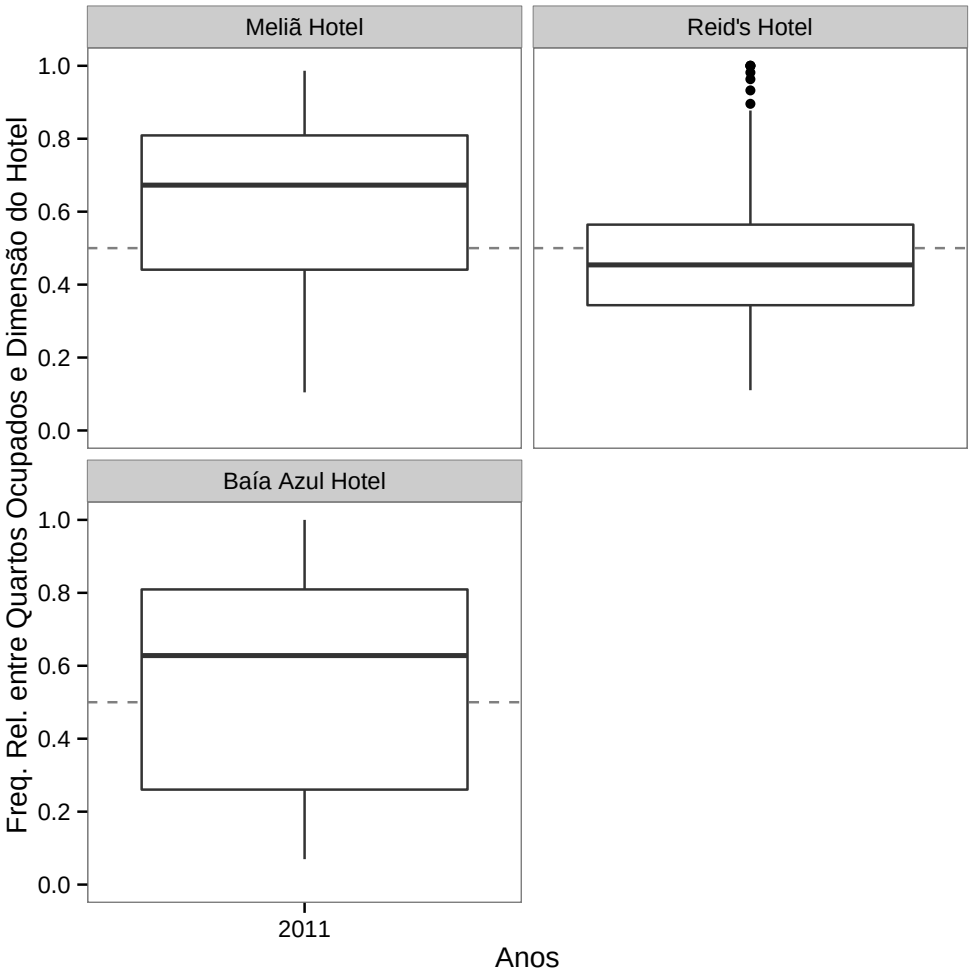


Figura B.3: Distribuição da Frequência Relativa entre o número de quartos ocupados e a dimensão do hotel por cada estabelecimento e ano

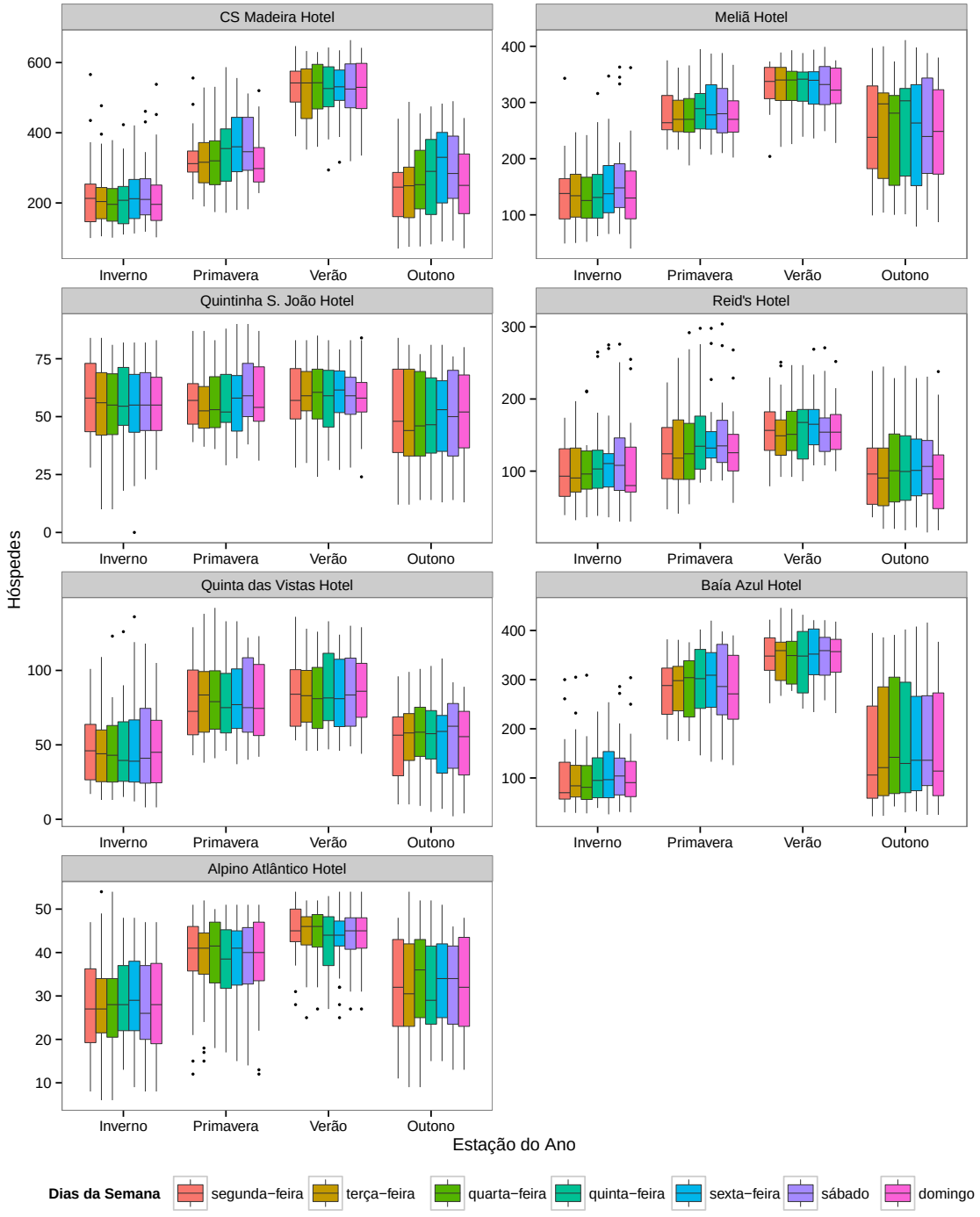


Figura B.4: Distribuição e variância do número de quartos alugados por ano e dia da semana em cada estabelecimento

B.3.0.1 Quinta de S.João

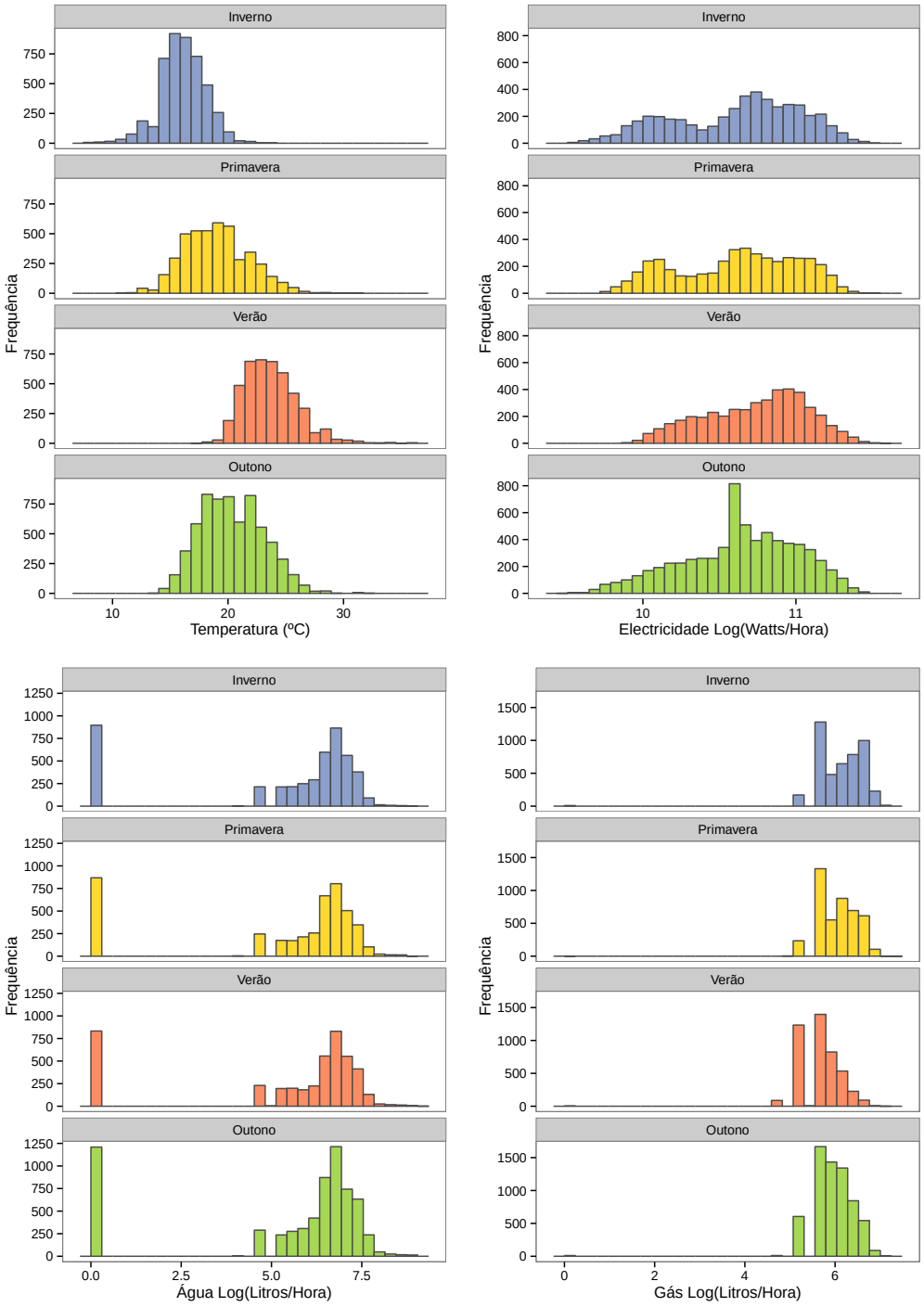


Figura B.5: Distribuição de valores à escala logarítmica das leituras dos sensores de consumo do Hotel Quinta de S.João e as leituras de temperatura ambiente por estação do ano

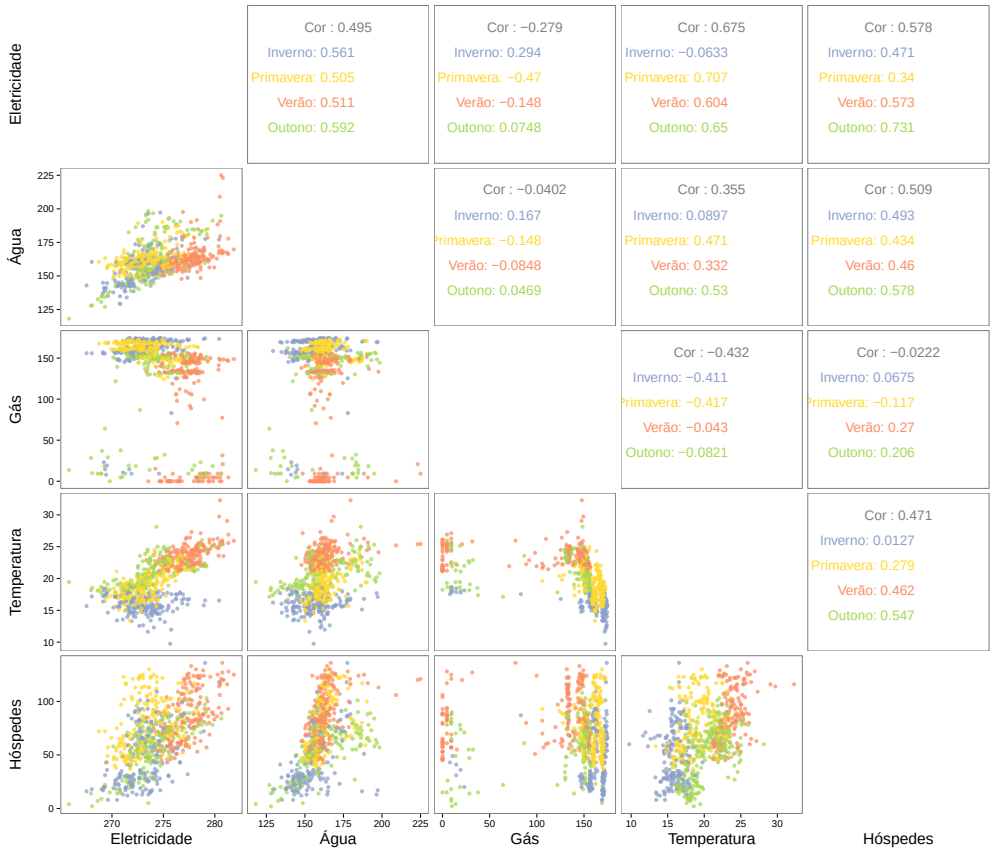


Figura B.6: Distribuição de valores à escala logarítmica das leituras dos sensores no Hotel Quinta de S.João por estação do ano

B.3.0.2 Quinta das Vistas

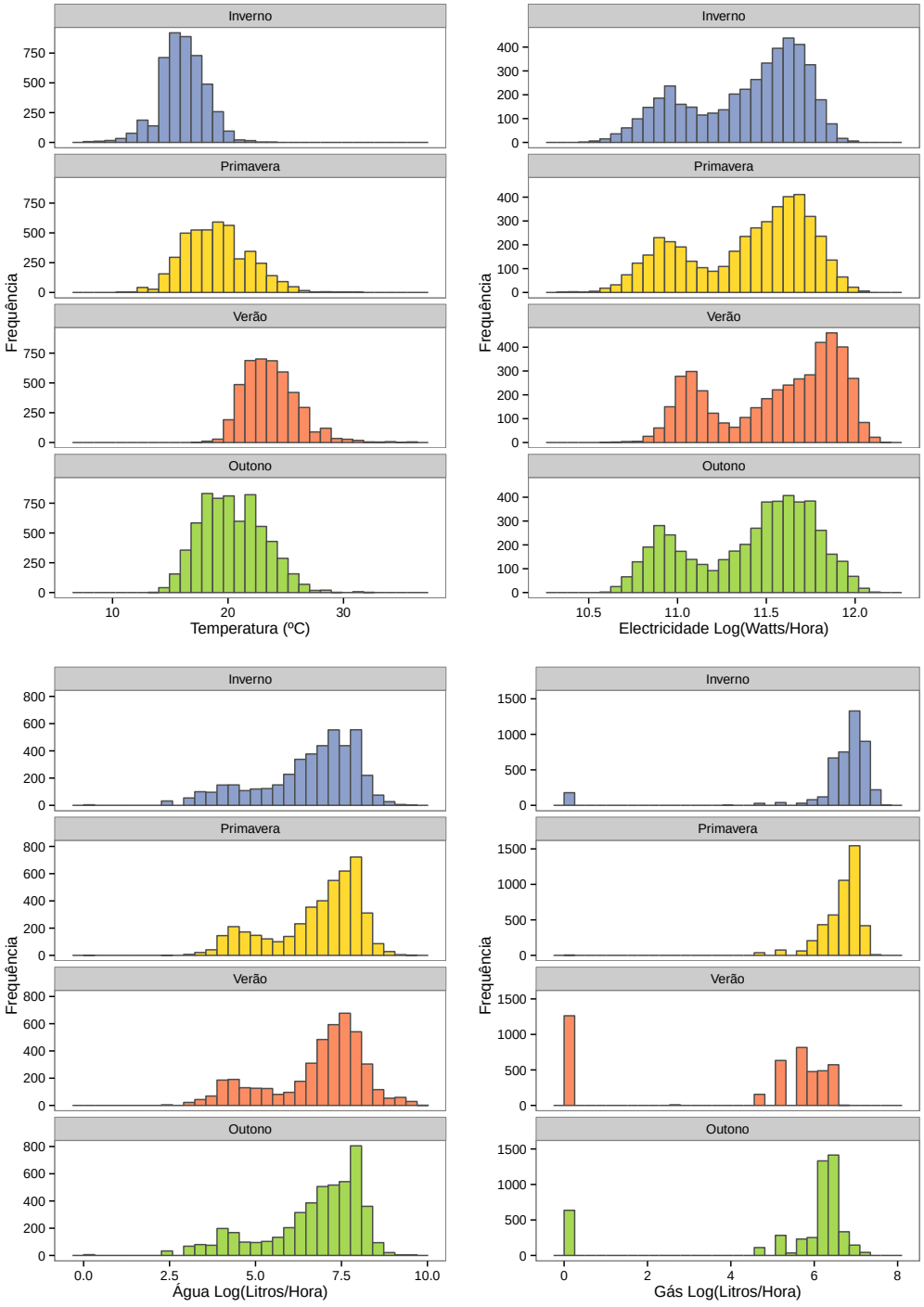


Figura B.7: Distribuição de valores à escala logarítmica das leituras dos sensores de consumo do Hotel Quinta das Vistas e as leituras de temperatura ambiente por estação do ano

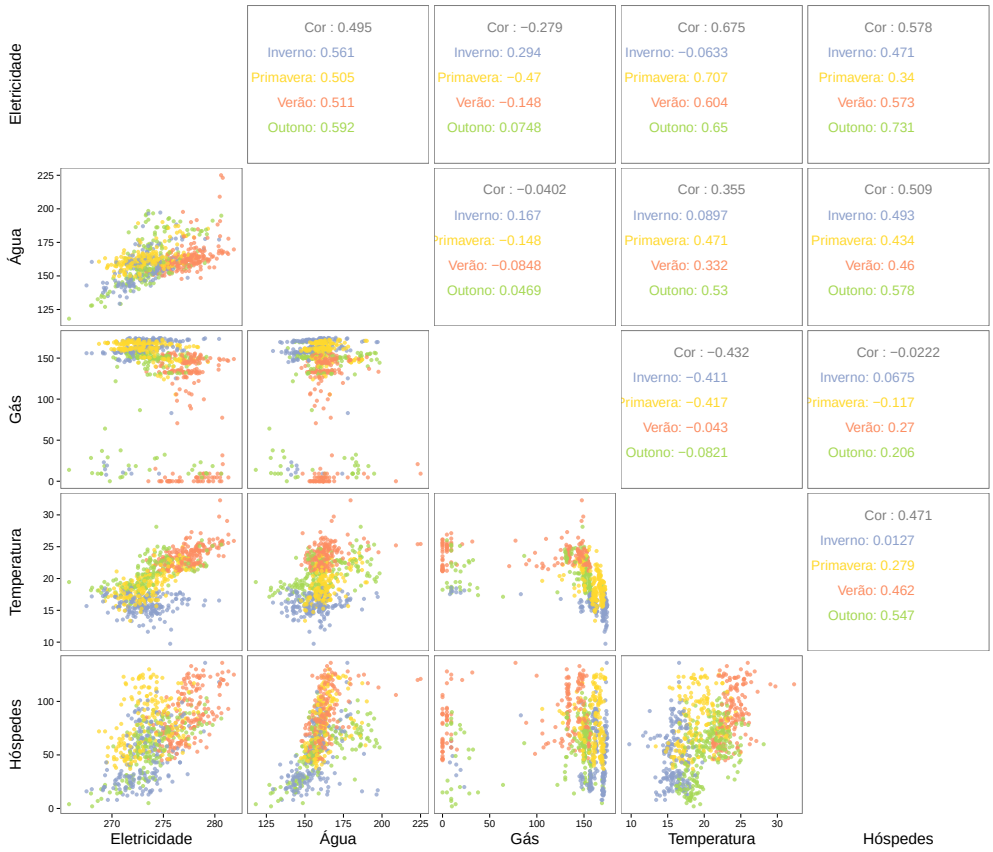


Figura B.8: Distribuição de valores à escala logarítmica das leituras dos sensores no Hotel Quinta das Vistas por estação do ano

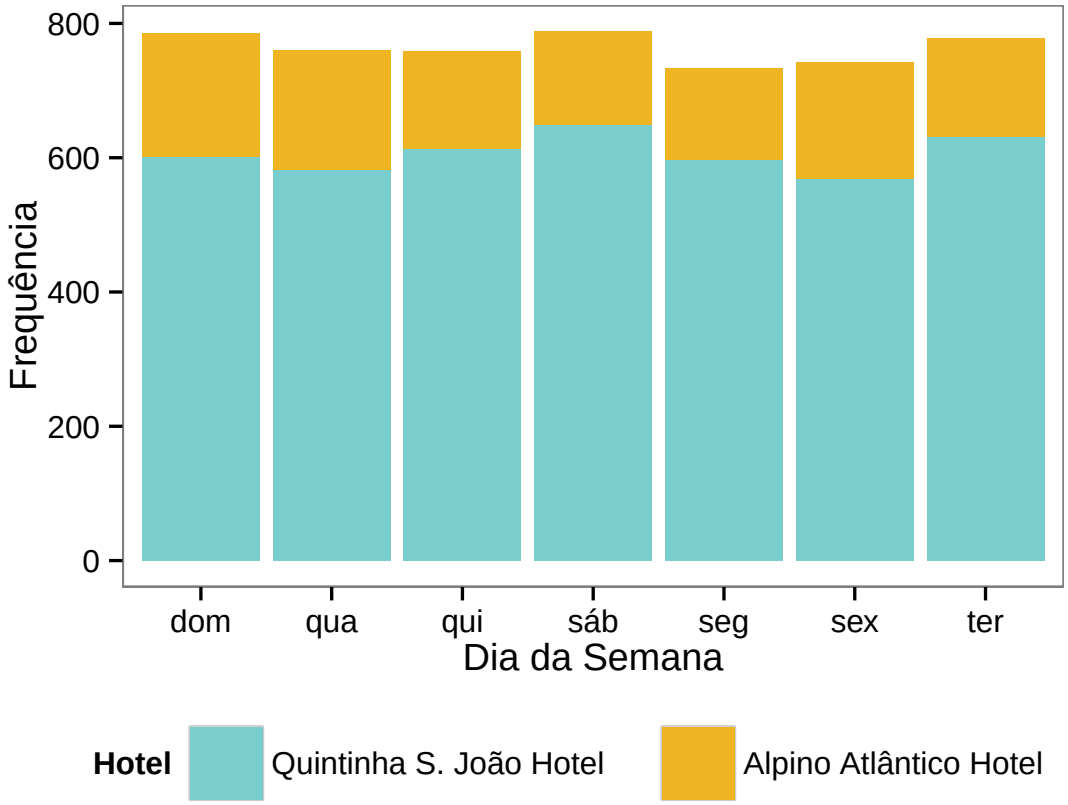


Figura B.9: Frequência de valores do sensor de Água aquando de valor zero nos Hotéis Alpino Atlântico e Quinta de S.João por dia de semana

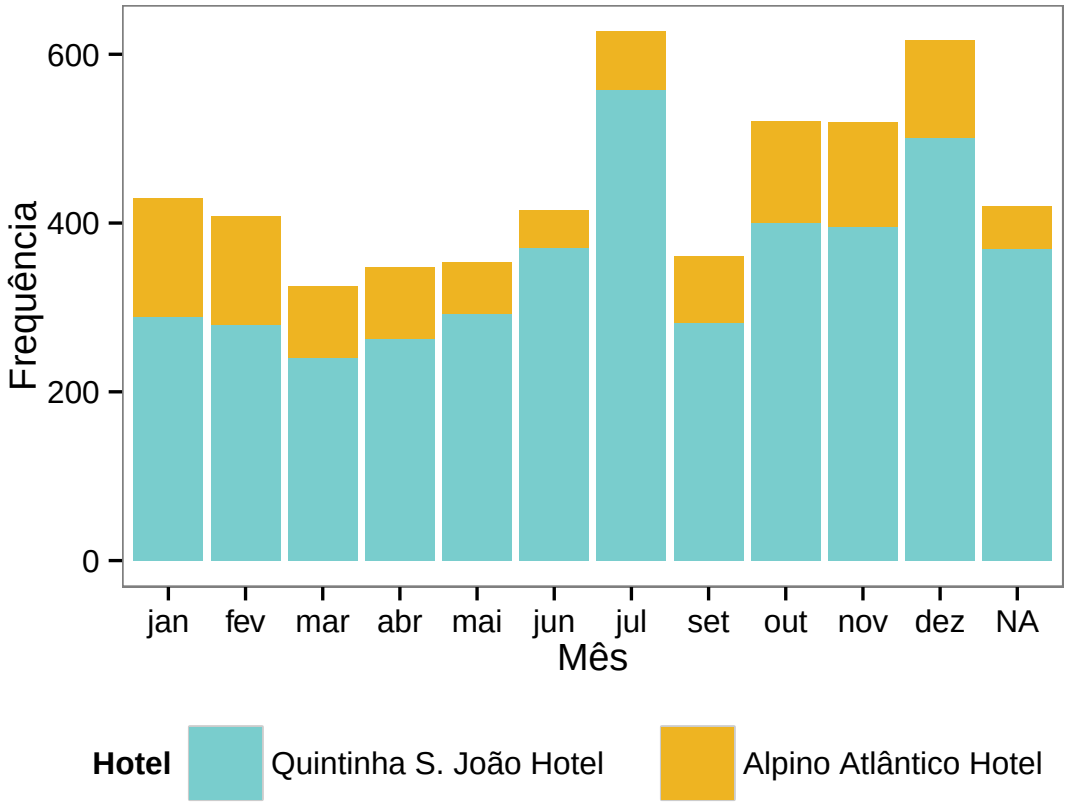


Figura B.10: Frequência de valores do sensor de Água aquando de valor zero nos Hotéis Alpino Atlântico e Quinta de S.João por dia de semana

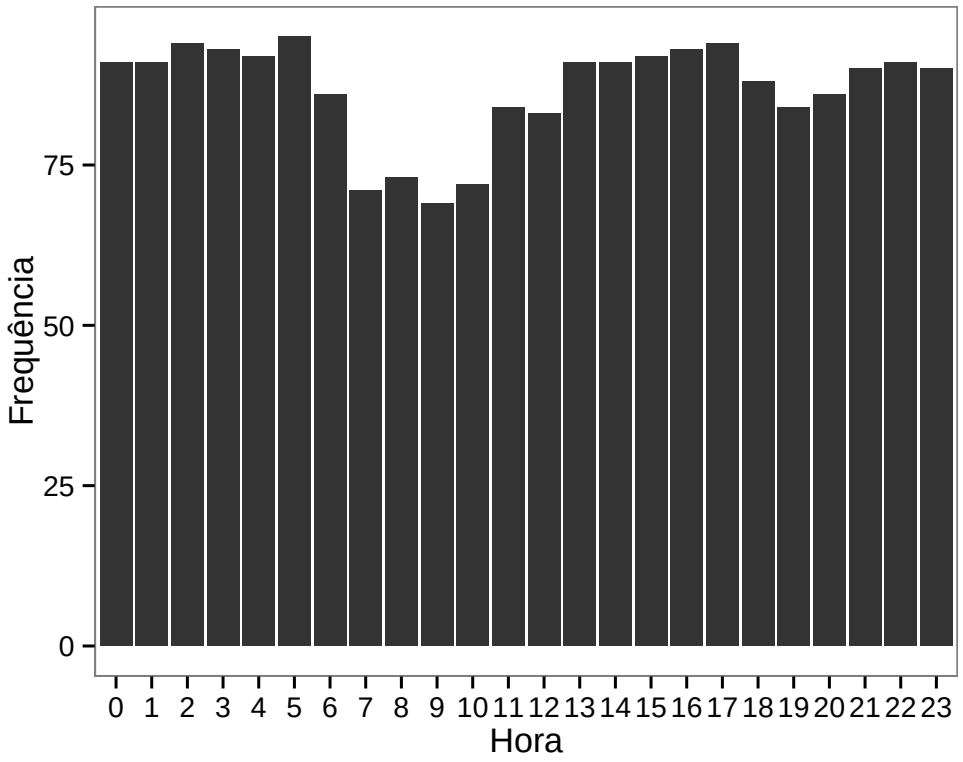


Figura B.11: Frequência de valores por Hora do sensor de Gás aquando de valor zero no Hotel Quinta das Vistas

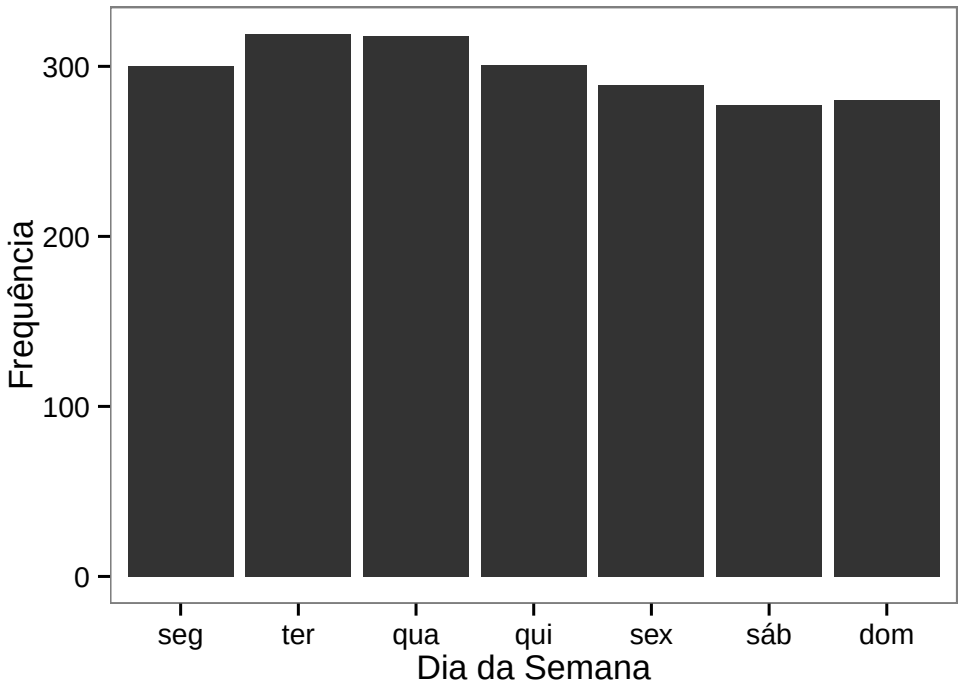


Figura B.12: Frequência de valores por dia de semana do sensor de Gás aquando de valor zero no Hotel Quinta das Vistas

APÊNDICE C

PREPARAÇÃO DOS DADOS

Tabela C.1: Validação de técnicas de escalonamento por *Calinski and Harabasz* para o conjunto de dados Quinta das Vistas *Y1&2*

Agrup.		Hierarchical Clustering						Partitional Clustering	
	K	Single	Complete	Average	Centroid	Ward.D2	Diana	**K-Means**	PAM
Range	2	15.8	297.5	580.7	48.1	647.2	662.6	662.6	629.6
	3	11.4	218.5	294.2	25.7	501.2	423.0	529.4	523.7
	4	8.0	346.4	211.0	18.1	434.0	449.0	456.1	418.2
	5	6.4	342.5	239.3	180.9	363.9	353.4	381.4	388.7
	6	5.5	286.8	212.7	145.2	326.9	313.3	359.7	344.3
Min-Max	2	6.4	277.2	23.6	6.4	428.6	480.9	486.4	466.0
	3	3.9	315.7	249.1	3.9	351.4	267.9	395.1	368.4
	4	3.3	275.9	264.9	11.1	329.5	273.4	353.9	350.2
	5	2.8	282.5	205.0	16.5	316.3	299.8	338.6	357.6
	6	2.5	332.4	166.2	13.5	307.2	261.2	374.1	363.5
Z-Score & Median/IQR	2	15.3	410.4	30.2	3.5	427.1	453.8	454.0	438.6
	3	8.8	263.4	16.3	3.1	348.8	287.2	396.0	386.2
	4	6.3	187.8	154.8	2.8	286.1	317.1	325.2	319.5
	5	5.6	214.7	118.3	2.7	248.1	258.3	275.7	279.5
	6	4.6	179.0	96.0	2.5	219.7	232.9	253.0	243.6

Tabela C.2: Validação de técnicas de escalonamento por *Calinski and Harabasz* para o conjunto de dados Quinta de S. João *Y1&2*

Agrup.		Hierarchical Clustering						Partitional Clustering	
	K	Single	Complete	Average	Centroid	Ward.D2	Diana	**K-Means**	PAM
Range	2	15.8	297.5	580.7	48.1	647.2	662.6	662.6	629.6
	3	11.4	218.5	294.2	25.7	501.2	423.0	529.4	523.7
	4	8.0	346.4	211.0	18.1	434.0	449.0	456.1	418.2
	5	6.4	342.5	239.3	180.9	363.9	353.4	381.4	388.7
	6	5.5	286.8	212.7	145.2	326.9	313.3	359.7	344.3
Min-Max	2	2.9	373.5	57.8	19.7	369.3	423.2	423.7	412.8
	3	5.7	354.0	32.4	14.9	346.9	398.1	415.9	403.8
	4	5.2	265.4	24.2	10.7	319.6	347.0	369.8	363.5
	5	4.5	241.1	121.2	8.2	286.7	281.3	336.8	331.3
	6	4.0	198.7	175.2	6.8	259.4	258.7	297.3	296.5
Z-Score & Median/IQR	2	15.3	410.4	30.2	3.5	427.1	453.8	454.0	438.6
	3	8.8	263.4	16.3	3.1	348.8	287.2	396.0	386.2
	4	6.3	187.8	154.8	2.8	286.1	317.1	325.2	319.5
	5	5.6	214.7	118.3	2.7	248.1	258.3	275.7	279.5
	6	4.6	179.0	96.0	2.5	219.7	232.9	253.0	243.6

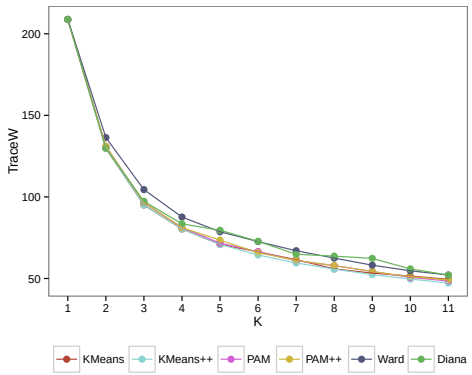
APÊNDICE D

PERFIS DE CONSUMO

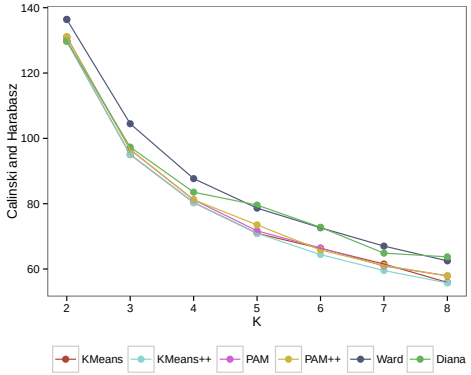
D.1 Modelação

D.1.1 Quinta de S.João

D.1.1.1 Y1&2



(a) *TraceW*

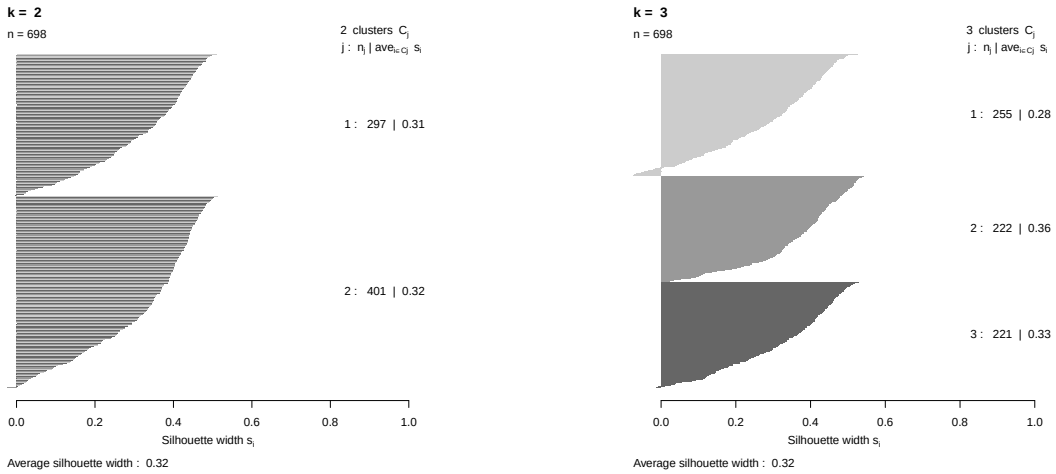


(b) *Calinski and Harabasz*

Figura D.1: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta de S.João *Y1&2*

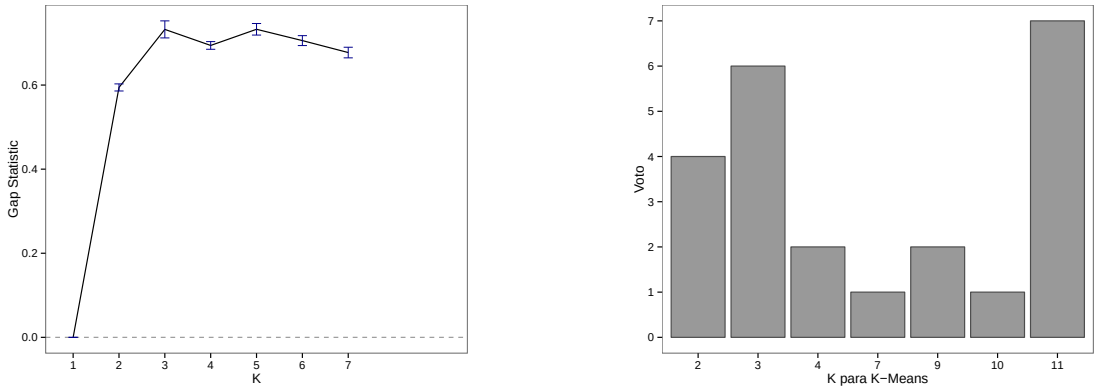
Tabela D.1: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta de S.João *Y1&2*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	44.26	44.26	43.24	43.24	40.46	46.42
4.00	20.04	20.03	19.12	19.12	15.11	18.73
5.00	5.28	5.31	5.86	7.70	7.78	9.85
6.00	4.82	2.89	4.27	0.02	3.01	2.87
7.00	0.19	1.63	0.30	2.97	0.40	1.10
8.00	0.84	1.09	2.54	1.53	1.09	6.77
9.00	3.01	0.48	0.60	0.42	0.19	0.25
10.00	0.85	0.51	0.07	0.44	0.88	5.06
11.00	0.24	0.47	1.55	1.17	0.89	2.70
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

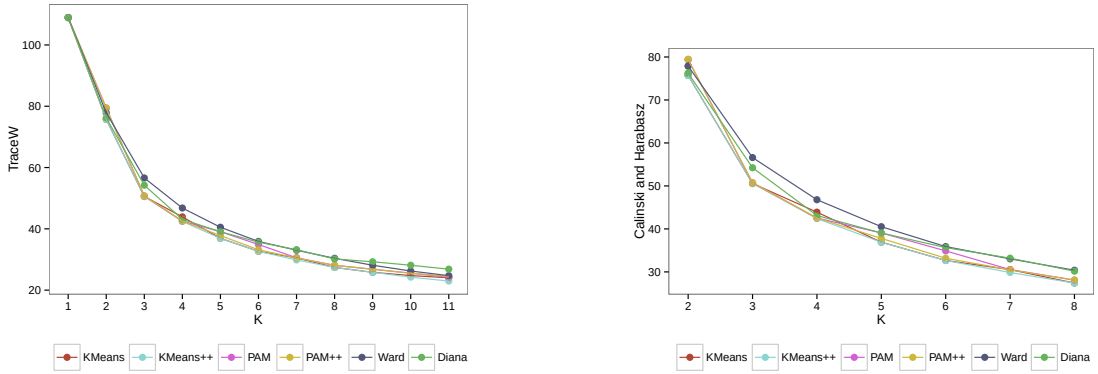
Figura D.2: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta de S.João *Y1&2*



(a) *K-Means* (b) *Ward*

Figura D.3: Gap Statistic com o conjunto de dados Quinta de S.João *Y1&2* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.1.2 Y1

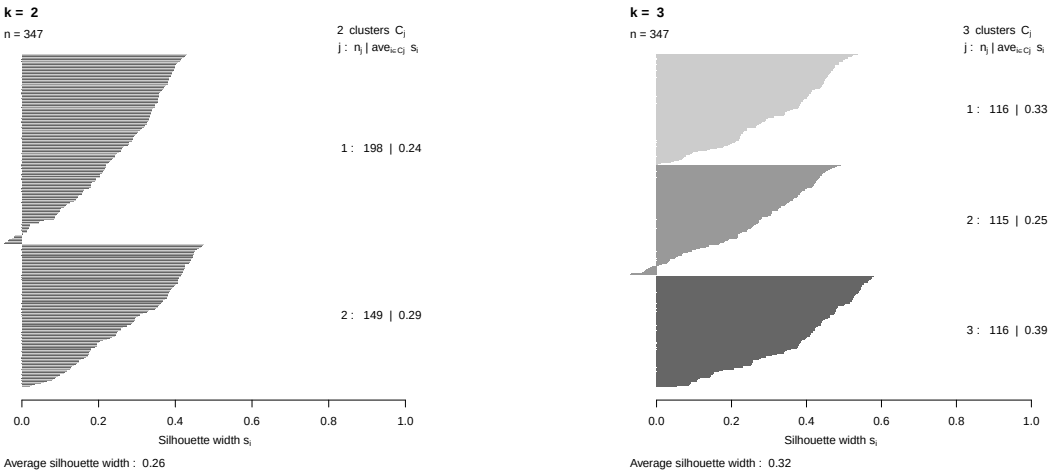


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.4: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta de S.João *Y1&2*

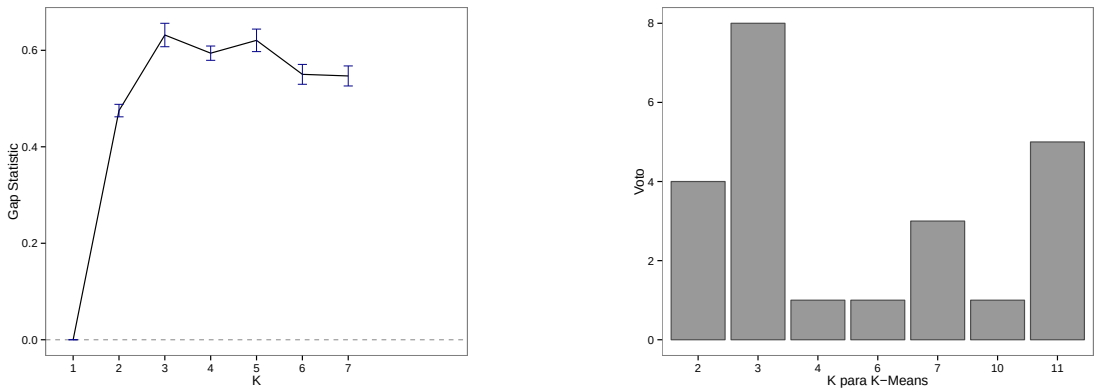
Tabela D.2: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta de S.João *Y1*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	8.07	8.07	0.82	0.82	9.76	10.72
4.00	18.49	17.00	20.50	20.50	11.48	10.87
5.00	0.33	2.63	4.72	3.45	3.54	7.06
6.00	2.73	1.27	0.70	0.12	1.63	0.66
7.00	2.13	1.44	0.17	1.92	1.78	0.94
8.00	1.00	0.35	1.89	0.32	0.21	0.57
9.00	1.54	0.86	1.17	1.10	0.37	2.12
10.00	0.53	0.01	0.01	0.01	0.39	0.23
11.00	0.33	0.30	0.09	0.27	0.36	0.13
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

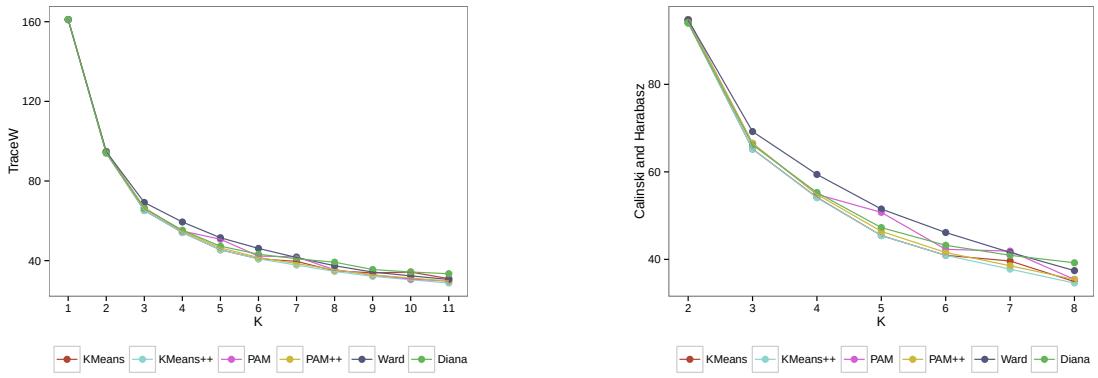
Figura D.5: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta de S.João *Y1*



(a) *K-Means* (b) *Ward*

Figura D.6: Gap Statistic com o conjunto de dados Quinta de S.João *Y1* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.1.3 Y2

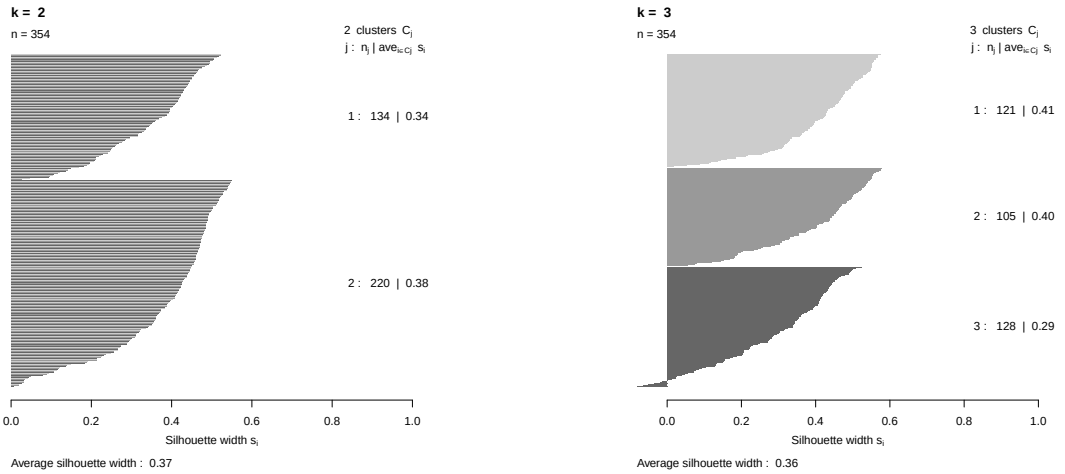


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.7: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta de S.João *Y2*

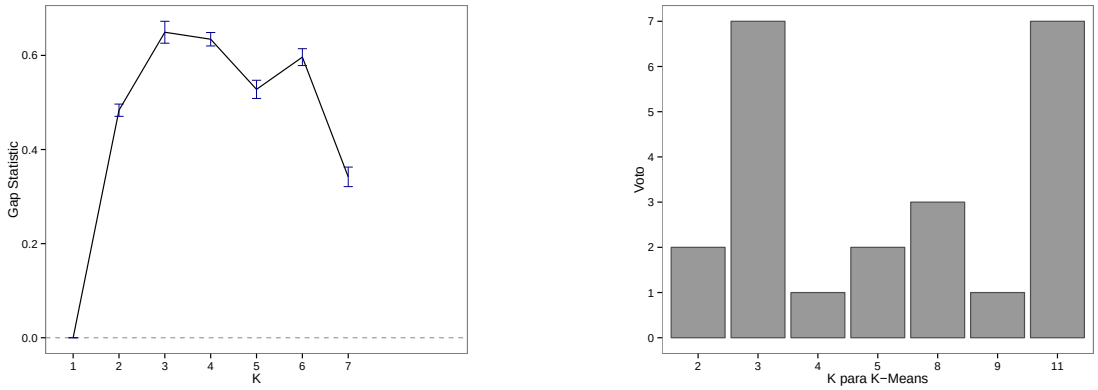
Tabela D.3: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta de S.João *Y2*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	38.15	38.15	37.99	37.99	40.88	39.09
4.00	17.81	17.72	16.64	16.64	15.69	17.03
5.00	2.33	2.49	7.54	3.20	1.90	2.87
6.00	4.21	4.14	4.34	3.62	2.54	3.98
7.00	3.22	1.35	8.05	1.81	0.82	1.75
8.00	3.43	0.02	6.06	0.10	0.40	0.57
9.00	3.65	0.69	3.86	0.52	0.99	1.99
10.00	1.48	0.80	0.57	1.08	1.26	2.46
11.00	3.68	0.03	0.97	0.06	0.07	0.38
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

Figura D.8: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta de S.João *Y2*

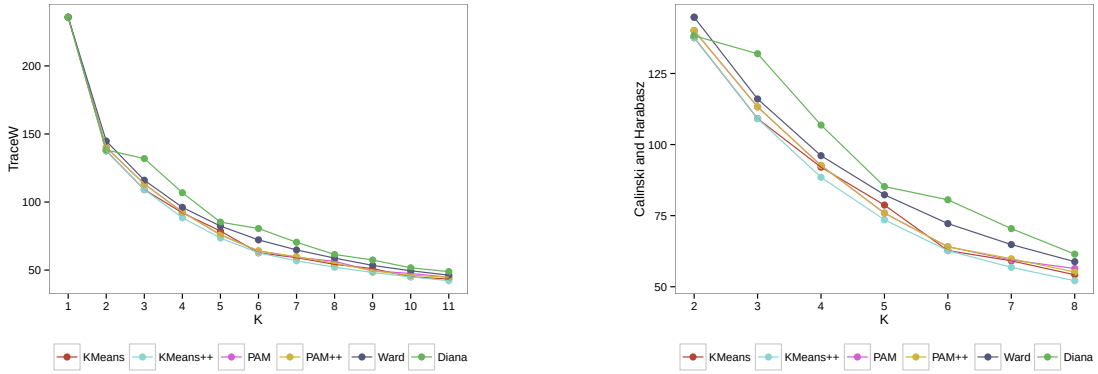


(a) *K-Means* (b) *Ward*

Figura D.9: Gap Statistic com o conjunto de dados Quinta de S.João *Y2* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.2 Quinta das Vistas

D.1.2.1 Y1&2

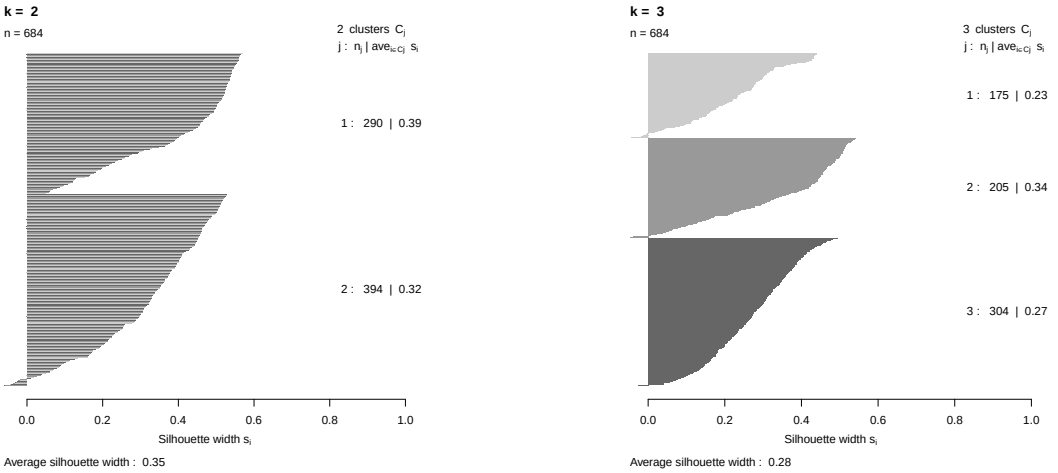


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.10: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta das Vistas *Y1&2*

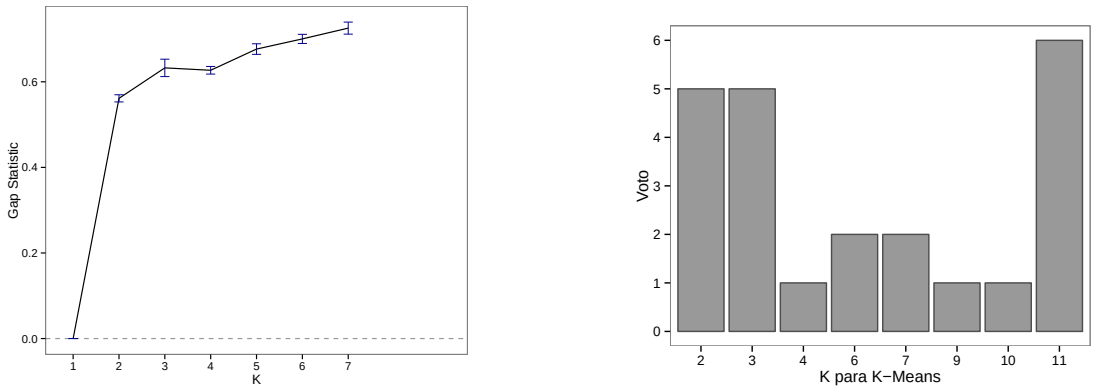
Tabela D.4: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta das Vistas *Y1&2*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	69.68	69.68	68.88	68.88	62.23	91.18
4.00	11.39	7.82	6.23	6.23	8.83	18.76
5.00	3.77	5.71	3.84	3.84	6.17	3.46
6.00	2.68	4.12	4.93	4.93	3.63	16.97
7.00	12.40	4.91	7.19	7.62	2.76	5.48
8.00	1.30	1.24	1.62	0.48	1.36	1.18
9.00	1.81	0.80	3.90	0.98	0.64	4.92
10.00	2.89	0.64	4.94	2.42	1.45	1.72
11.00	4.23	0.42	0.77	1.84	0.70	2.98
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

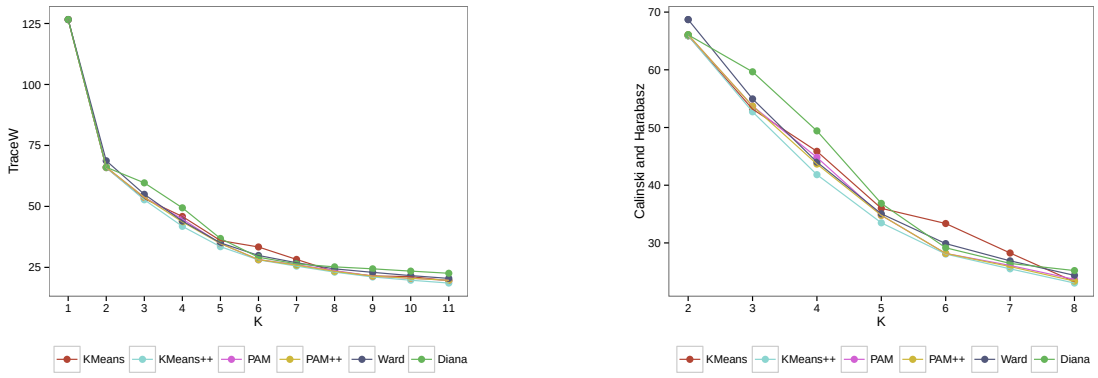
Figura D.11: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta das Vistas *Y1&2*



(a) *K-Means* (b) *Ward*

Figura D.12: Gap Statistic com o conjunto de dados Quinta das Vistas *Y1&2* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.2.2 Y1

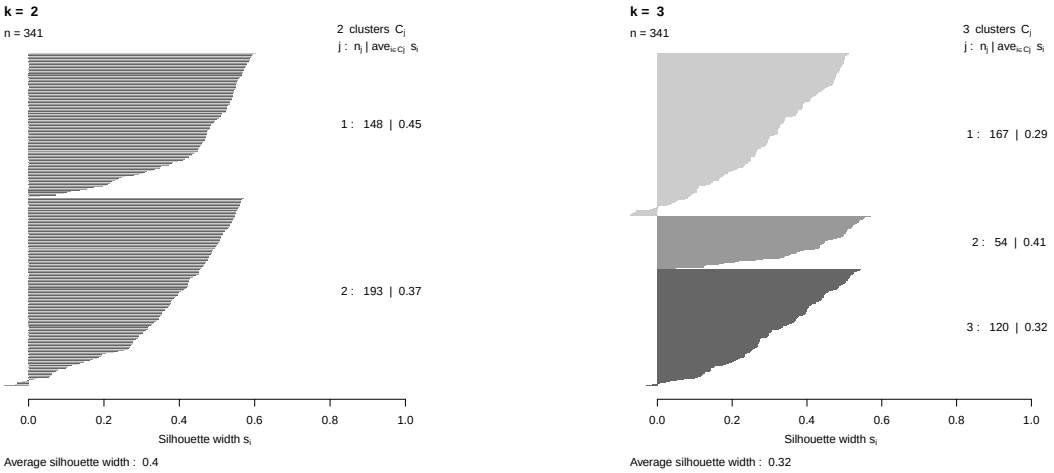


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.13: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta de S.João *Y1*

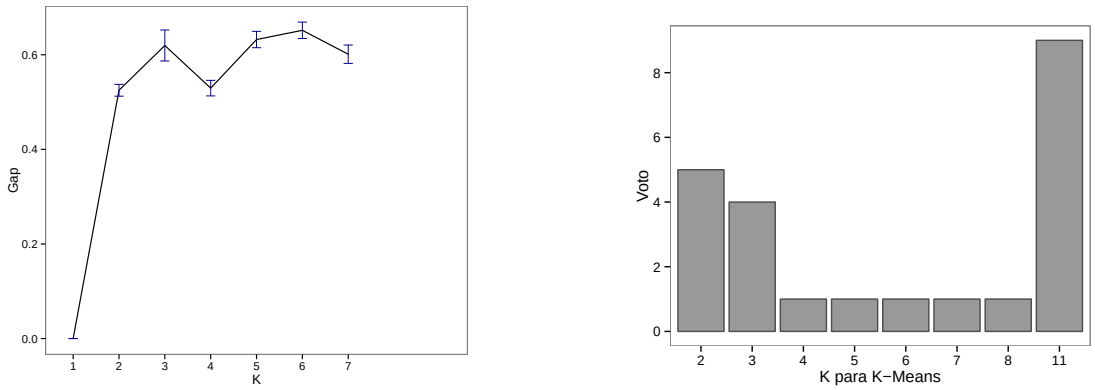
Tabela D.5: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta de S.João *Y1*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	47.99	47.53	48.22	48.22	44.16	54.15
4.00	5.40	2.29	3.44	2.29	2.79	3.84
5.00	2.57	2.55	1.14	1.17	2.00	2.33
6.00	7.25	2.93	3.43	2.27	3.78	4.88
7.00	2.48	2.85	4.53	4.33	2.19	4.98
8.00	0.13	0.10	0.33	0.24	0.48	1.47
9.00	3.17	0.47	0.15	0.55	1.09	0.43
10.00	1.54	0.67	1.37	1.08	0.09	0.13
11.00	1.41	0.15	0.04	0.04	0.19	0.05
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

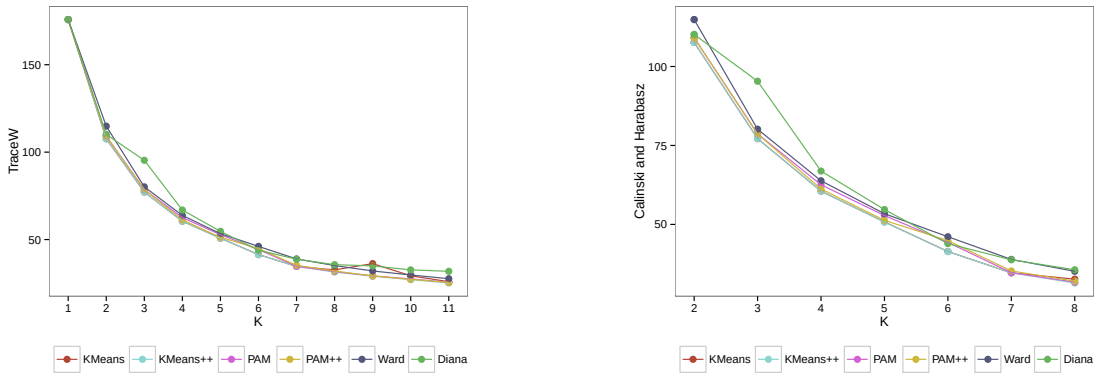
Figura D.14: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta das Vistas *Y1*



(a) *K-Means* (b) *Ward*

Figura D.15: Gap Statistic com o conjunto de dados Quinta das Vistas *Y1* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.2.3 Y2

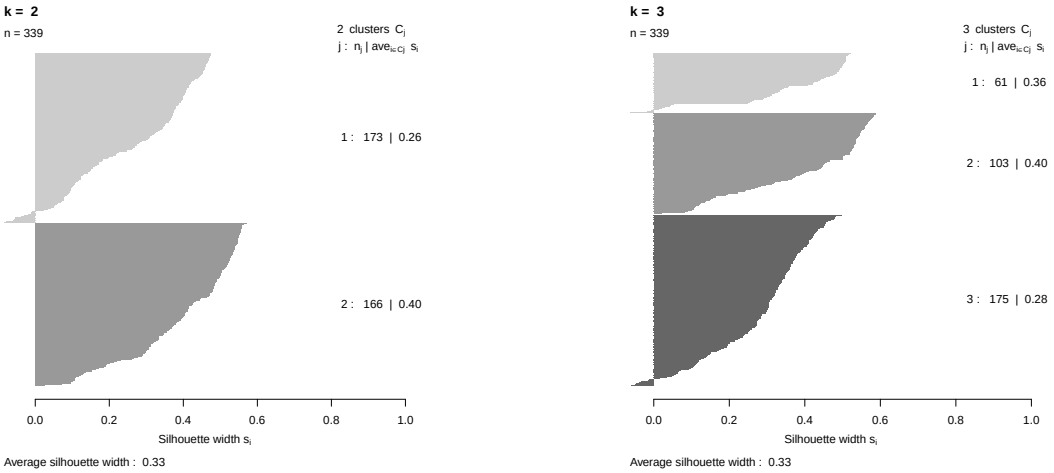


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.16: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta de S.João *Y2*

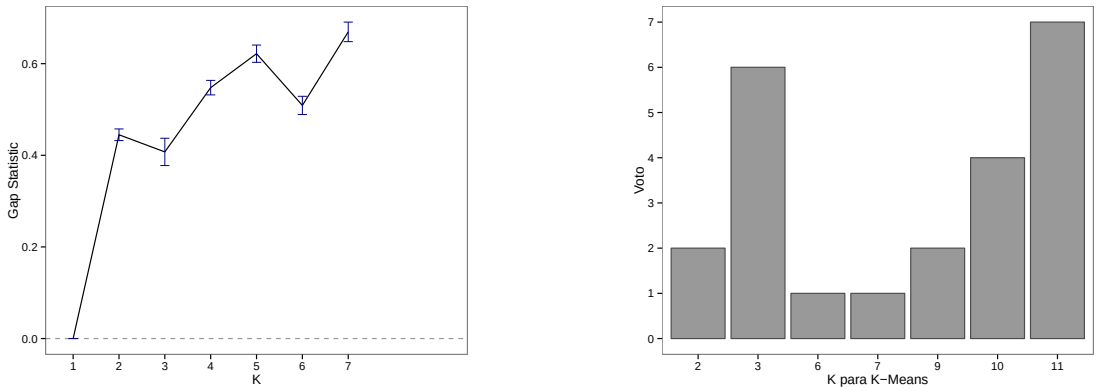
Tabela D.6: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta de S.João *Y2*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	37.73	37.73	36.26	36.26	26.24	50.86
4.00	13.81	13.81	14.36	13.01	18.37	13.64
5.00	6.91	6.85	6.35	7.51	5.91	16.26
6.00	0.37	0.45	1.24	3.53	3.13	1.40
7.00	2.62	2.61	1.08	3.20	0.11	5.68
8.00	4.76	3.50	6.58	6.35	3.41	1.97
9.00	5.59	0.82	0.52	0.44	0.68	2.46
10.00	10.55	0.47	0.81	0.86	0.79	1.61
11.00	3.58	0.01	0.26	0.17	0.17	1.49
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

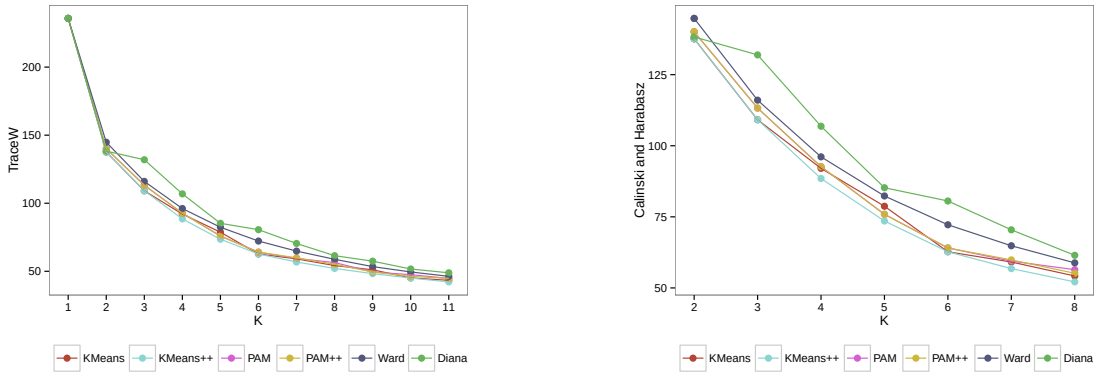
Figura D.17: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta das Vistas *Y1*



(a) *K-Means* (b) *Ward*

Figura D.18: Gap Statistic com o conjunto de dados Quinta das Vistas *Y1* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.3 Alpino Atlântico

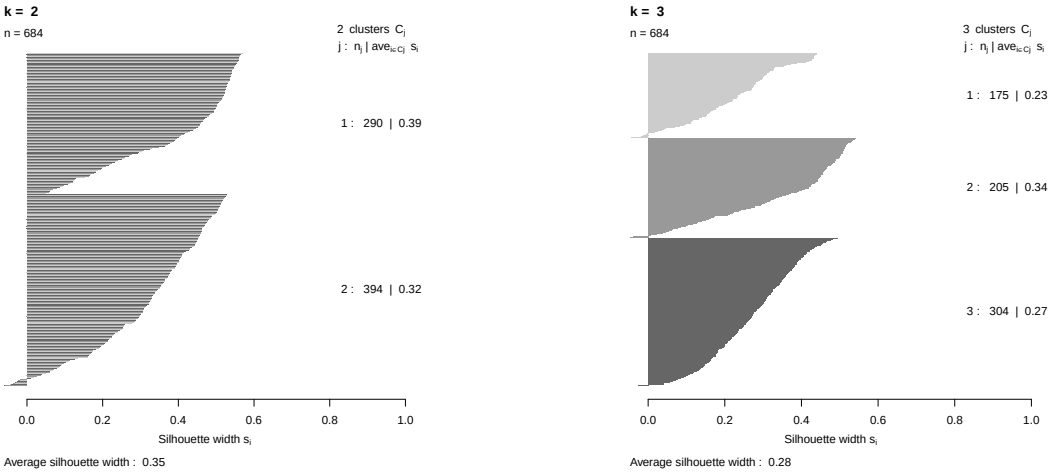


(a) *TraceW* (b) *Calinski and Harabasz*

Figura D.19: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Quinta das Vistas *Y1&2*

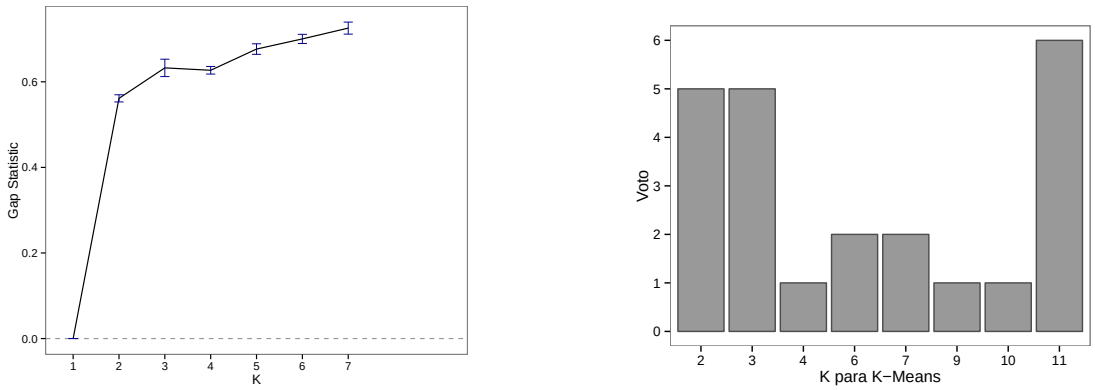
Tabela D.7: Segunda Derivada do índice *TraceW* com o conjunto de dados Quinta das Vistas *Y1&2*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	69.68	69.68	68.88	68.88	62.23	91.18
4.00	11.39	7.82	6.23	6.23	8.83	18.76
5.00	3.77	5.71	3.84	3.84	6.17	3.46
6.00	2.68	4.12	4.93	4.93	3.63	16.97
7.00	12.40	4.91	7.19	7.62	2.76	5.48
8.00	1.30	1.24	1.62	0.48	1.36	1.18
9.00	1.81	0.80	3.90	0.98	0.64	4.92
10.00	2.89	0.64	4.94	2.42	1.45	1.72
11.00	4.23	0.42	0.77	1.84	0.70	2.98
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

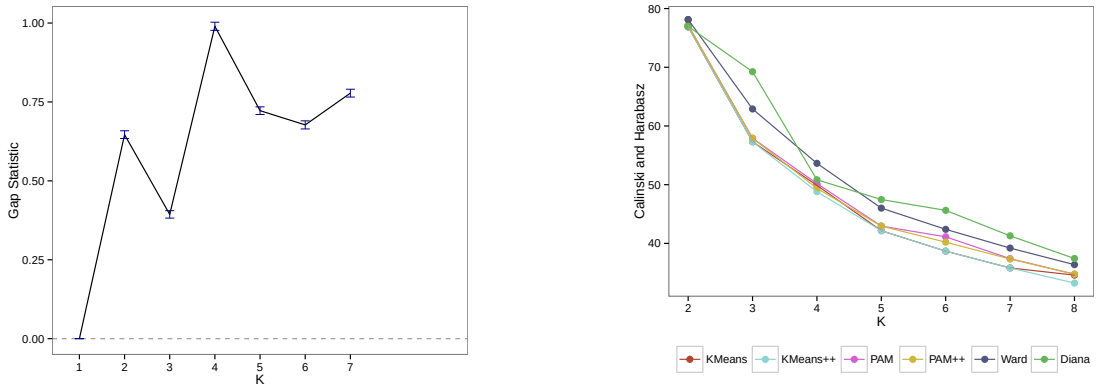
Figura D.20: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Quinta das Vistas *Y1&2*



(a) *K-Means* (b) *Ward*

Figura D.21: Gap Statistic com o conjunto de dados Quinta das Vistas *Y1&2* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.3.1 Y1



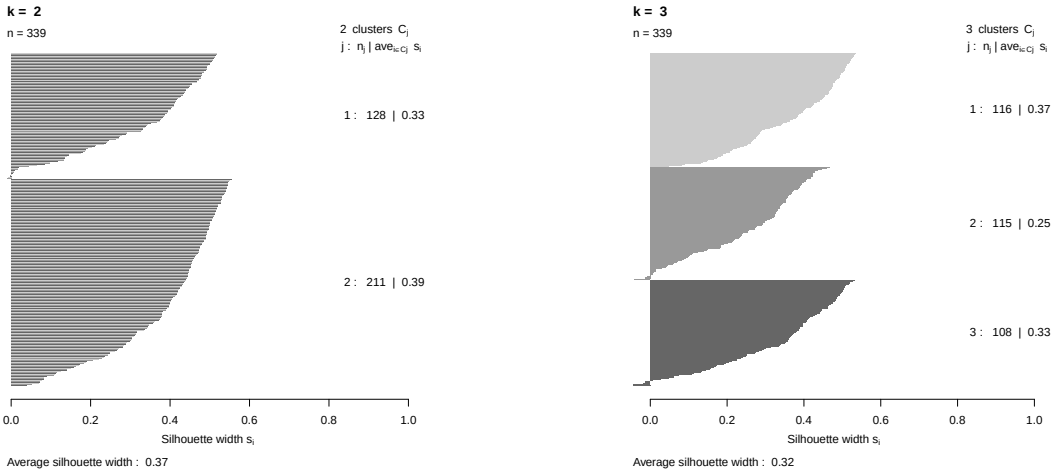
(a) *TraceW*

(b) *Calinski and Harabasz*

Figura D.22: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Alpino Atlântico *Y1*

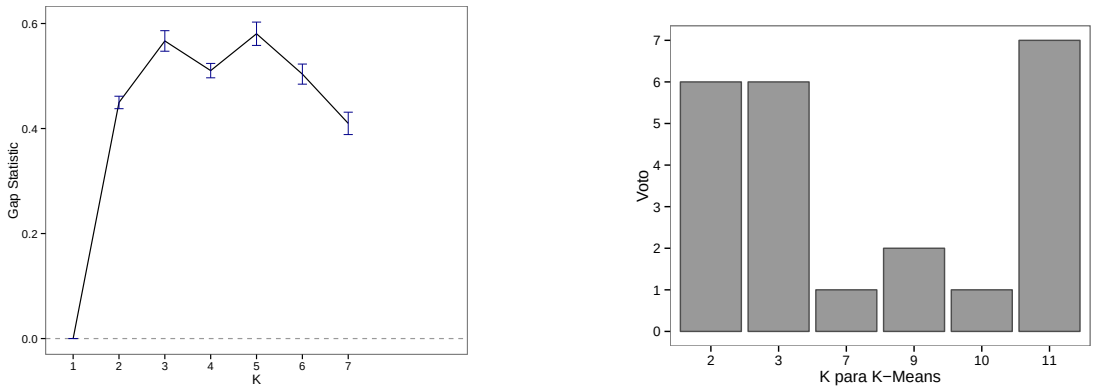
Tabela D.8: Segunda Derivada do índice *TraceW* com o conjunto de dados Alpino Atlântico *Y1*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	33.40	33.40	33.49	33.49	36.54	45.21
4.00	12.17	11.09	11.52	10.73	5.97	10.69
5.00	0.30	1.85	0.49	2.08	1.63	15.04
6.00	4.24	3.17	5.40	3.70	4.02	1.52
7.00	0.62	0.62	1.88	0.13	0.42	2.49
8.00	1.64	0.29	1.07	0.30	0.37	0.48
9.00	1.89	0.30	0.12	0.61	0.12	0.01
10.00	3.25	0.38	1.74	1.30	0.62	3.42
11.00	0.96	0.32	0.83	0.76	0.44	1.02
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

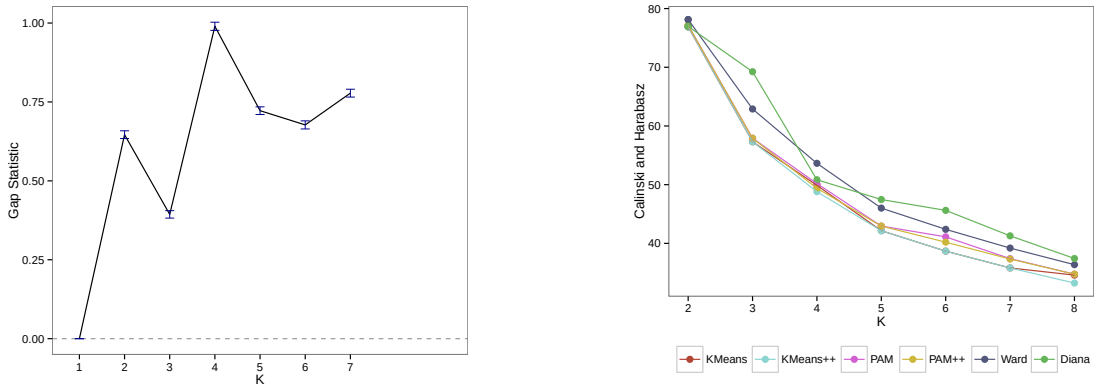
Figura D.23: Índice *Silhouette* com o modelo *K-Means*++ com o conjunto de dados Alpino Atlântico *Y1*



(a) *K-Means* (b) *Ward*

Figura D.24: Gap Statistic com o conjunto de dados Alpino Atlântico *Y1* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means*++

D.1.3.2 Y2



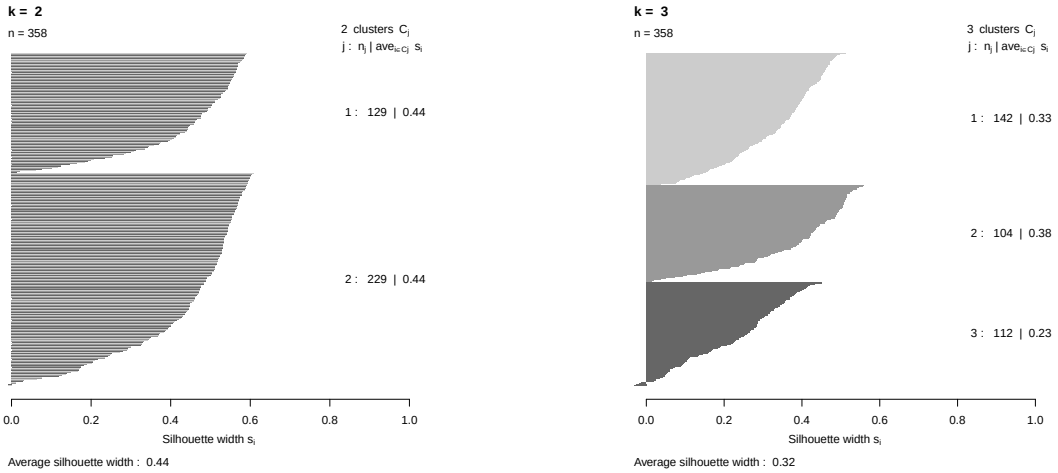
(a) *TraceW*

(b) *Calinski and Harabasz*

Figura D.25: Índices *TraceW* e *Calinski and Harabasz* nos modelos gerados com o conjunto de dados Alpino Atlântico *Y1*

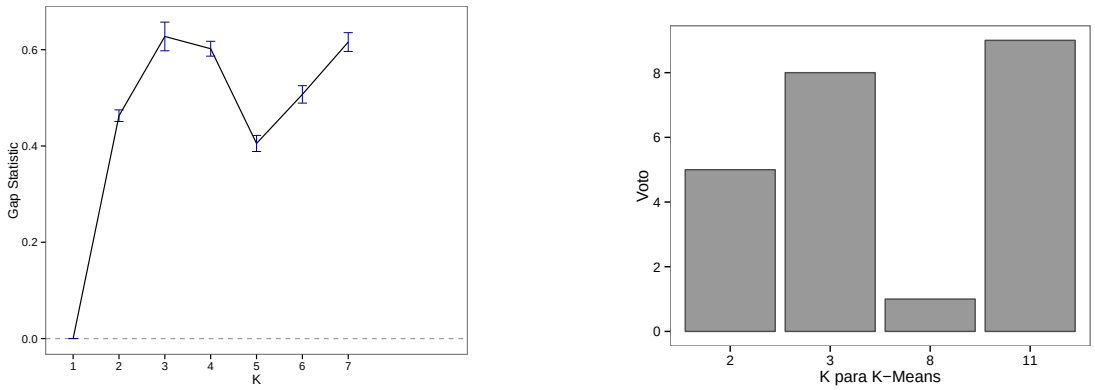
Tabela D.9: Segunda Derivada do índice *TraceW* com o conjunto de dados Alpino Atlântico *Y1*

K	KMeans	KMeans++	PAM	PAM++	Ward	Diana
3.00	86.08	86.08	88.53	88.53	82.45	97.15
4.00	13.72	13.69	8.18	7.90	8.68	7.90
5.00	5.80	3.32	4.94	5.49	8.86	17.33
6.00	2.29	2.11	3.63	2.70	0.44	2.79
7.00	1.82	0.29	0.61	0.98	0.20	2.27
8.00	1.84	0.69	0.09	1.21	0.40	0.32
9.00	0.49	0.68	0.61	0.49	0.39	1.20
10.00	3.68	0.61	0.12	0.20	0.61	0.20
11.00	3.58	0.31	0.58	0.65	0.78	0.41
Recomendado	3	3	3	3	3	3



(a) *k=2* (b) *k=3*

Figura D.26: Índice *Silhouette* com o modelo *K-Means++* com o conjunto de dados Alpino Atlântico *Y2*



(a) *K-Means* (b) *Ward*

Figura D.27: Gap Statistic com o conjunto de dados Alpino Atlântico *Y2* e Identificação do número adequado de agrupamento por consenso com o modelo *K-Means++*

D.1.4 Índices de Validação

Tabela D.10: Índices paramétricos utilizados para identificação do número mais adequado de agrupamentos

Índice	Técnica Validação
Ball Hall	Segundo Derivada
Banfeld Raftery	Minimo
C index	Mínimo
Calinski Harabasz	Máximo
Davies Bouldin	Minimo
Det. Ratio	Segunda Derivada
Dunn	Maximo
GDI	Maximo
Gamma	Maximo
Gap Statistic	Maximo
Log Det Ratio	Segunda Derivada
Log SS Ratio	Segunda Derivada
McClain Rao	Minimo
PBM	Maximo
Point Biserial	Maximo
Ratkowsky Lance	Maximo
Ray Turi	Minimo
Scott Symons	Minimo
SD	Minimo
Silhouette	Maximo
Tau	Maximo
Trace(W)	Segunda Derivada
Trave(WiB)	Segunda Derivada
Wemmert Gancarski	Maximo
Xie Beni	Minimo

D.1.5 Interpretação

D.1.5.1 Quinta de S.João

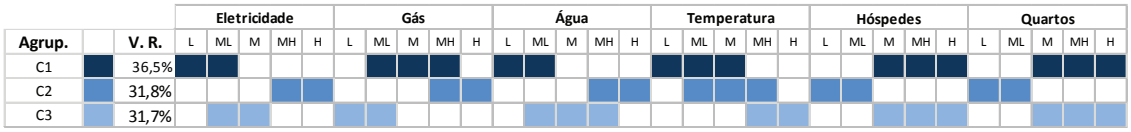


Figura D.28: Categorização dos agrupamentos no conjunto de dados Quinta de S.João *Y1&2*

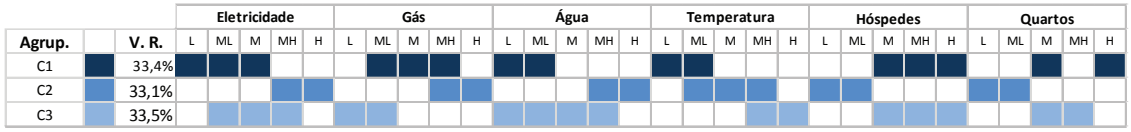


Figura D.29: Categorização dos agrupamentos no conjunto de dados Quinta de S.João *Y1*

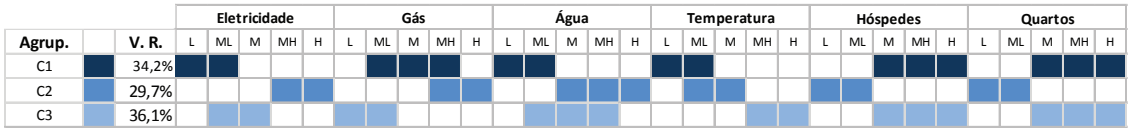


Figura D.30: Categorização dos agrupamentos no conjunto de dados Quinta de S.João *Y2*

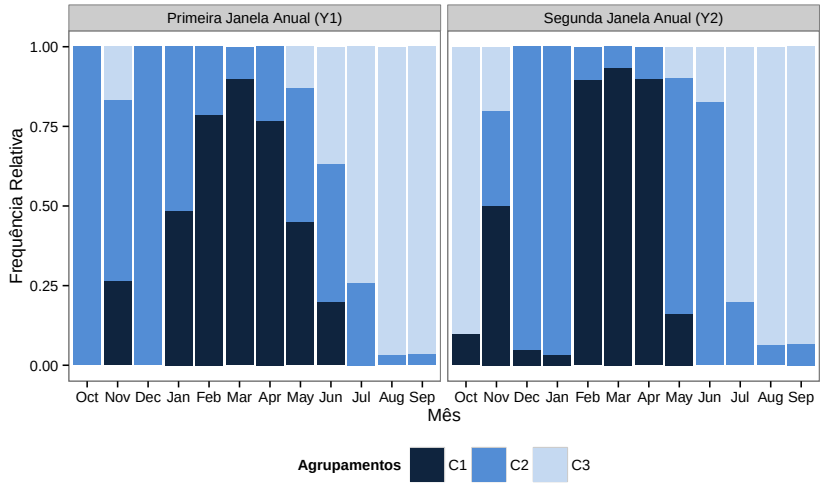


Figura D.31: Caracterização temporal dos agrupamentos definidos com o conjunto de dados Quinta de S.João *Y1&2*

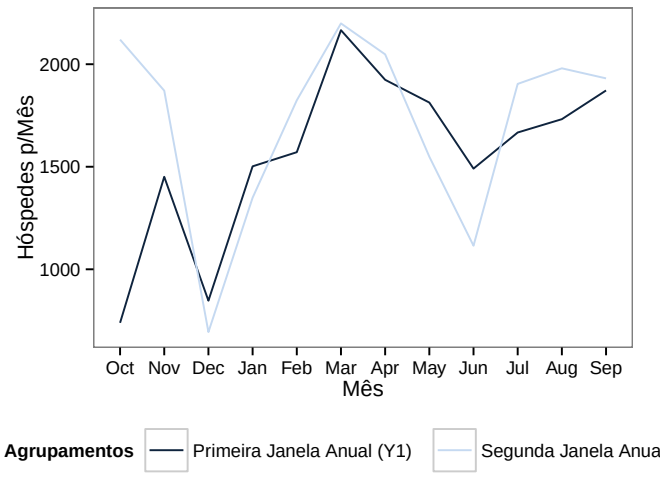
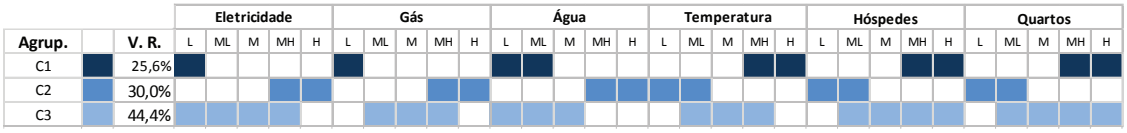


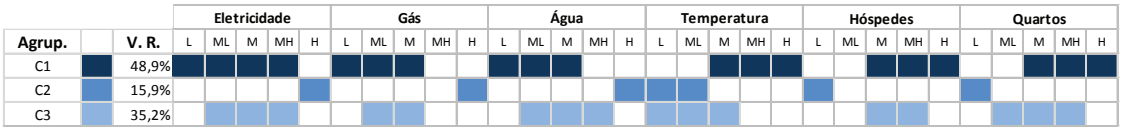
Figura D.32: Caracterização da variável Hóspedes (H) por agrupamento no conjunto de dados Quinta de S.João *Y1&2*

D.1.5.2 Quinta das Vistas



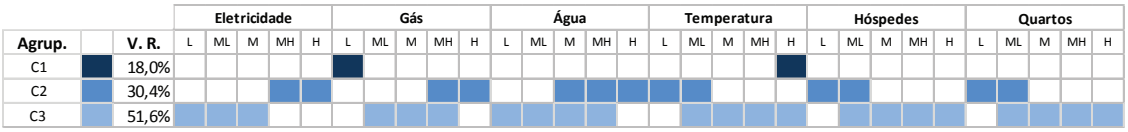
		Eletricidade					Gás					Água					Temperatura					Hóspedes					Quartos				
Agrup.	V. R.	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H
C1	25,6%																													
C2	30,0%																													
C3	44,4%																													

Figura D.33: Categorização dos agrupamentos no conjunto de dados Quinta das Vistas *Y1&2*



		Eletricidade					Gás					Água					Temperatura					Hóspedes					Quartos				
Agrup.	V. R.	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H
C1	48,9%																													
C2	15,9%																													
C3	35,2%																													

Figura D.34: Categorização dos agrupamentos no conjunto de dados Quinta das Vistas *Y1*



		Eletricidade					Gás					Água					Temperatura					Hóspedes					Quartos				
Agrup.	V. R.	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H	L	ML	M	MH	H
C1	18,0%																													
C2	30,4%																													
C3	51,6%																													

Figura D.35: Categorização dos agrupamentos no conjunto de dados Quinta das Vistas *Y2*

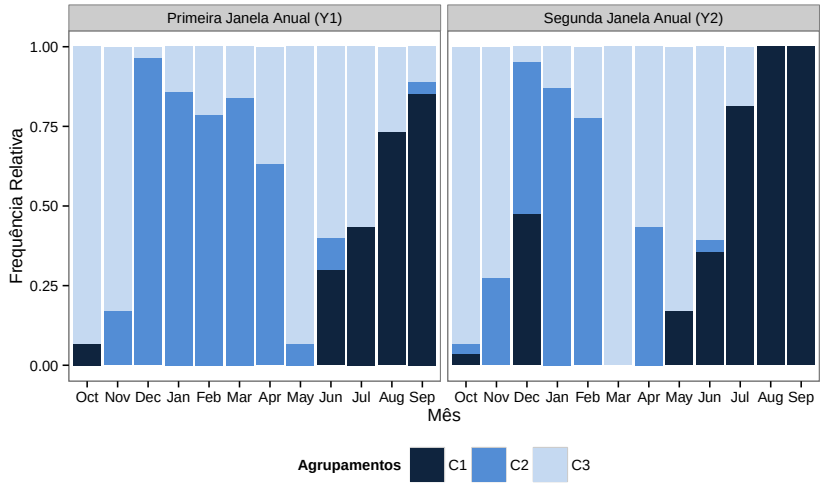


Figura D.36: Caracterização temporal dos agrupamentos definidos com o conjunto de dados Quinta das Vistas *Y1&2*

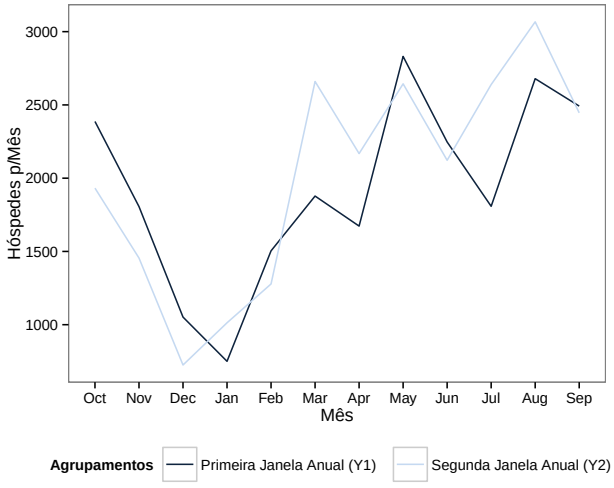


Figura D.37: Caracterização da variável Hóspedes (H) por agrupamento no conjunto de dados Quinta das Vistas *Y1&2*

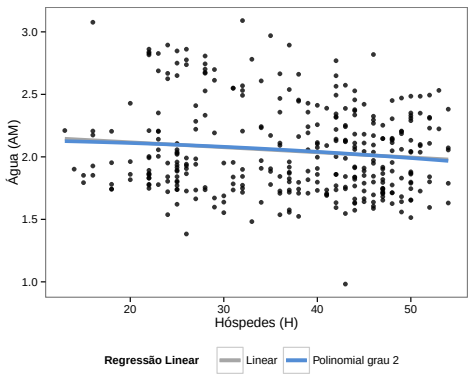
APÊNDICE E

CONSUMO DE SERVIÇOS GRANULARIDADE DIÁRIA

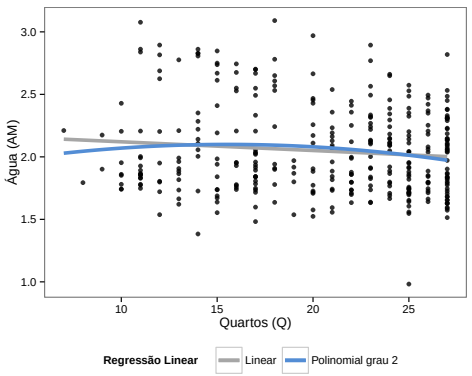
E.1 Alpino Atlântico

E.1.1 Consumo de Água

E.1.1.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.1: Regressão Linear e Polinomial perante consumo médio de Água (AM) à escala logarítmica com o conjunto de dados Alpino Atlântico

Tabela E.1: Sumário de Regressões Lineares para a estimativa de consumo de Água (AM) com o conjunto de dados Alpino Atlântico

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.01	0.353	3.986	2.897	30.0
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.008	0.353	4.004	2.900	30.0
$\beta_0 + \beta_1 Q$	0.008	0.353	4.011	2.898	29.8
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.010	0.353	4.088	2.914	29.8

E.1.1.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H^2 + +\beta_2 TP75 \tag{E.1}$$

Tabela E.2: Performance dos modelos de regressão para a estimativa de eletricidade (AM) no conjunto de dados Alpino Atlântico

Regressão Linear log (AM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.121	0.332	4.13	2.96	32.0
KNN	-	-	4.22	2.97	31.3
SVM	-	-	4.34	2.98	30.9

E.1.1.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 TP25 \tag{E.2}$$

$$RL_2 = \beta_0 + \beta_1 TP75 + \beta_2 H^2 \tag{E.3}$$

$$RL_3 = \beta_0 + \beta_1 TP25 \tag{E.4}$$

Tabela E.3: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Alpino Atlântico

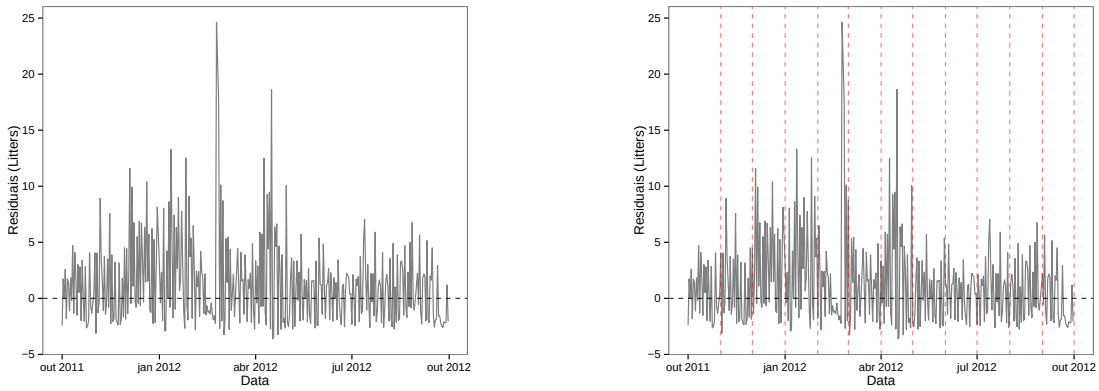
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	3.054	2.538	29.0
	2	2.392	2.054	28.3
	3	6.076	4.257	34.0
$RL_{p1,p2,p3}$	-	4.122	2.903	30.4
KNN	1	3.148	2.569	29.0
	2	2.445	2.126	29.3
	3	6.205	4.320	34.9
$KNN_{p1,p2,p3}$	-	4.215	2.962	31.1
SVM	1	3.196	2.528	27.4
	2	2.445	2.126	29.3
	3	.737	4.223	34.9
$SVM_{p1,p2,p3}$	-	3.99	2.92	30.7
Melhor Conjunto				
$SVM_{p1}\ RL_{p2,p3}$	-	4.14	2.90	30.0

E.1.1.4 Estimativa utilizando Janelas Temporais

Tabela E.4: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Alpino Atlântico

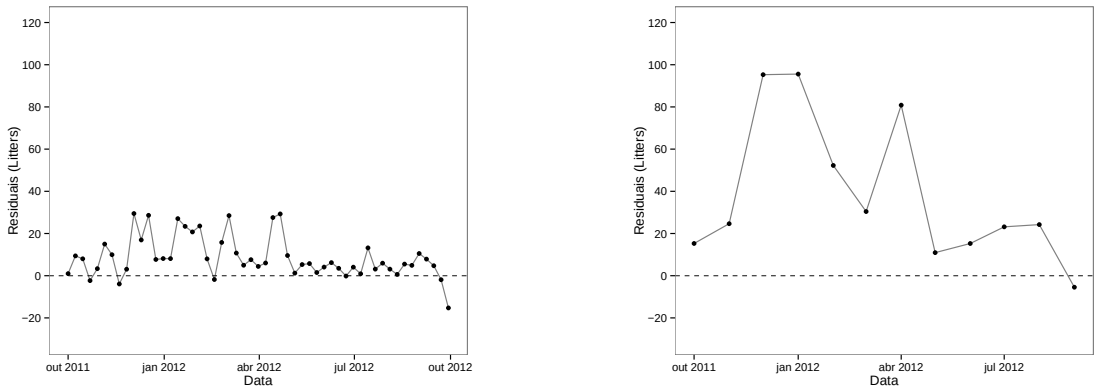
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	3.338	2.614	28.3
	2	2.889	2.405	33.2
	3	5.954	4.353	35.9
$RL_{g1,g2,g3}$	-	4.249	3.105	33.0
KNN	1	3.161	2.466	26.3
	2	2.523	2.157	29.3
	3	6.366	4.512	35.4
$KNN_{g1,g2,g3}$	-	4.316	3.016	30.6
$KNN_{completo}$	-	4.288	2.994	30.6
SVM	1	3.268	2.534	27.4
	2	3.025	2.406	30.3
	3	6.301	4.463	33.9
$SVM_{g1,g2,g3}$	-	4.441	3.123	30.8
$SVM_{completo}$	-	4.282	3.039	30.6
Melhor Conjunto				
$KNN_{g1,g2}$ SVM_{g3}	-	4.284	3.000	30.1

E.1.1.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.2: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Alpino Atlântico



(a) (b)

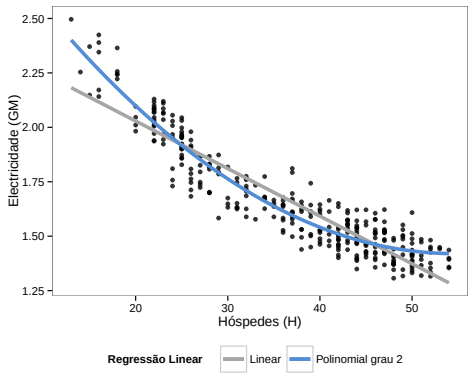
Figura E.3: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Alpino Atlântico

Tabela E.5: Alpino Atlântico

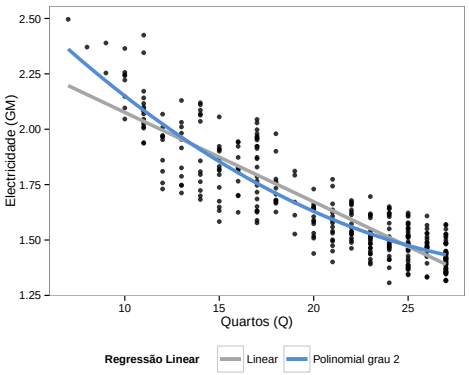
		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Água	$RL(Q)$	2.898	29.8	1.56	15.0	1.32	13.3

E.1.2 Consumo de Gás

E.1.2.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.4: Regressão Linear e Polinomial perante consumo médio de Gás (GM) à escala logarítmica com o conjunto de dados Alpino Atlântico

Tabela E.6: Sumário de Regressões Lineares para a estimativa de consumo de Gás (GM) com o conjunto de dados Alpino Atlântico

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE*** (Watts/Dia)	**MAE*** (Watts/Dia)	**MAPE*** (%)
$\beta_0 + \beta_1 H$	0.854	0.094	0.783	0.533	9.0
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.903	0.076	0.474	0.388	7.5
$\beta_0 + \beta_1 Q$	0.796	0.111	0.949	0.682	11.0
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.817	0.105	0.810	0.623	10.7

E.1.2.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + +\beta_2 H^2 + \beta_3 TM \tag{E.5}$$

Tabela E.7: Performance dos modelos de regressão para a estimativa de eletricidade (GM) no conjunto de dados Alpino Atlântico

Regressão Linear log (GM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.911	0.073	0.422	0.337	6.3
KNN	-	-	0.679	0.492	8.3
SVM	-	-	0.491	0.356	6.4

E.1.2.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 TP25 \tag{E.6}$$

$$RL_2 = \beta_0 + \beta_1 TP75 + \beta_2 H^2 \tag{E.7}$$

$$RL_3 = \beta_0 + \beta_1 TP25 \tag{E.8}$$

Tabela E.8: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Alpino Atlântico

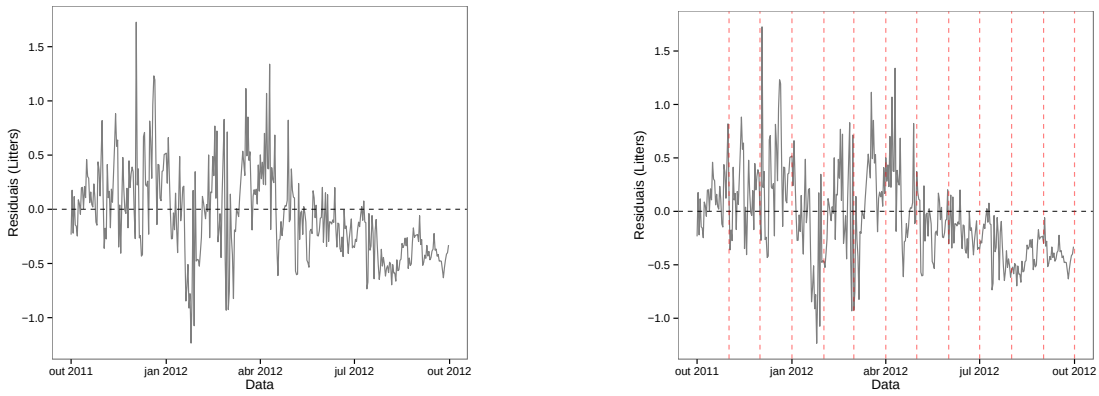
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	0.451	0.338	6.19
	2	0.422	0.361	9.0
	3	0.583	0.444	5.38
$RL_{p1,p2,p3}$	-	0.488	0.383	7.1
KNN	1	0.631	0.489	9.0
	2	0.361	0.310	7.7
	3	0.961	0.749	8.8
$KNN_{p1,p2,p3}$	-	0.677	0.499	8.4
SVM	1	0.564	0.450	8.2
	2	0.361	0.310	7.7
	3	1.147	0.879	9.9
$SVM_{p1,p2,p3}$	-	0.755	0.533	8.6
Melhor Conjunto $RL_{p1}\ KNN_{p2}\ RL_{p3}$	-	0.466	0.361	6.5

E.1.2.4 Estimativa utilizando Janelas Temporais

Tabela E.9: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Alpino Atlântico

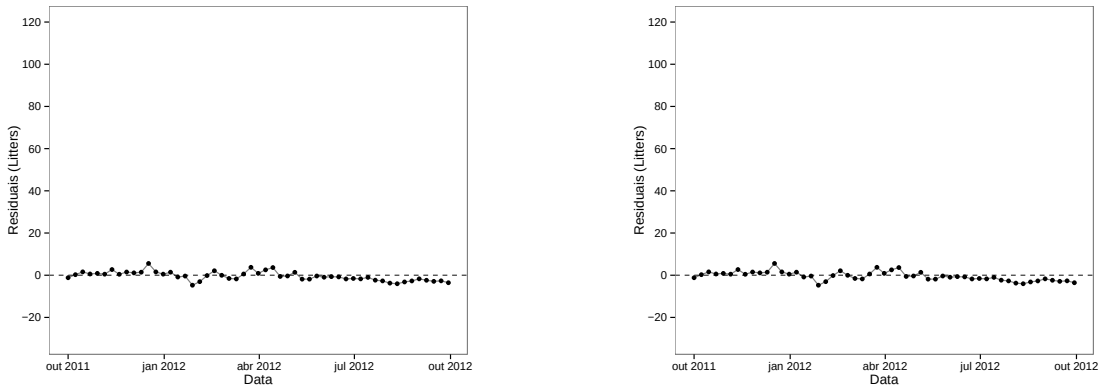
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	0.529	0.425	8.1
	2	0.662	0.551	13.4
	3	1.071	0.776	9.0
$RL_{g1,g2,g3}$	-	0.799	0.597	10.7
KNN	1	0.623	0.513	9.6
	2	0.480	0.416	10.3
	3	1.855	1.362	15.0
$KNN_{g1,g2,g3}$	-	1.158	0.755	11.7
$KNN_{completo}$	-	1.274	0.859	13.5
SVM	1	0.586	0.483	9.2
	2	0.396	0.348	8.7
	3	0.953	0.700	7.9
$SVM_{g1,g2,g3}$	-	0.672	0.497	8.6
$SVM_{completo}$	-	0.673	0.507	8.9
Melhor Conjunto				
RL_{g1} $SVM_{g2,g3}$	-	0.749	0.556	8.4

E.1.2.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.5: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados de treino da unidade Alpino Atlântico



(a) (b)

Figura E.6: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados de treino da unidade Alpino Atlântico

Tabela E.10: Alpino Atlântico

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Gás	*RL*	0.337	6.3	0.268	5.25	0.214	4.3

E.1.3 Árvore de Decisão de apoio à estimativa

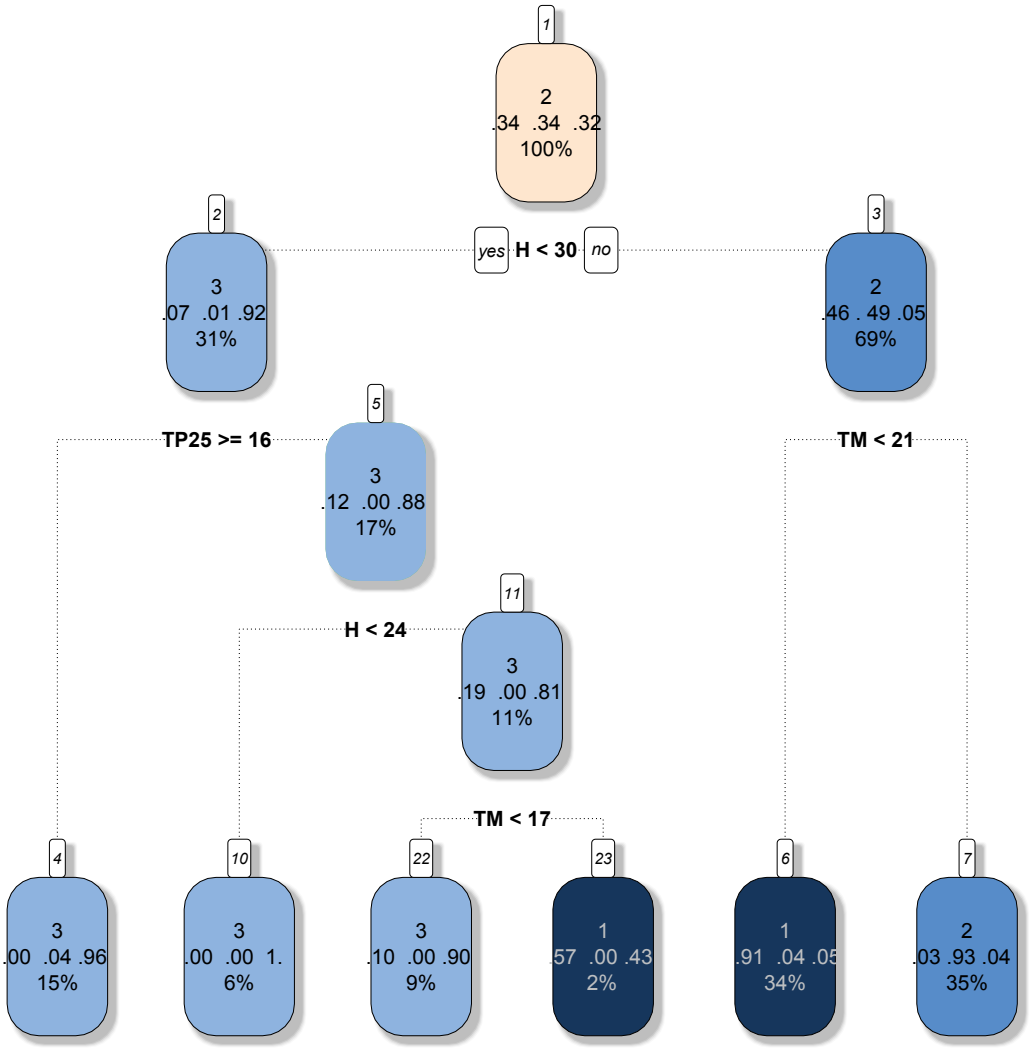
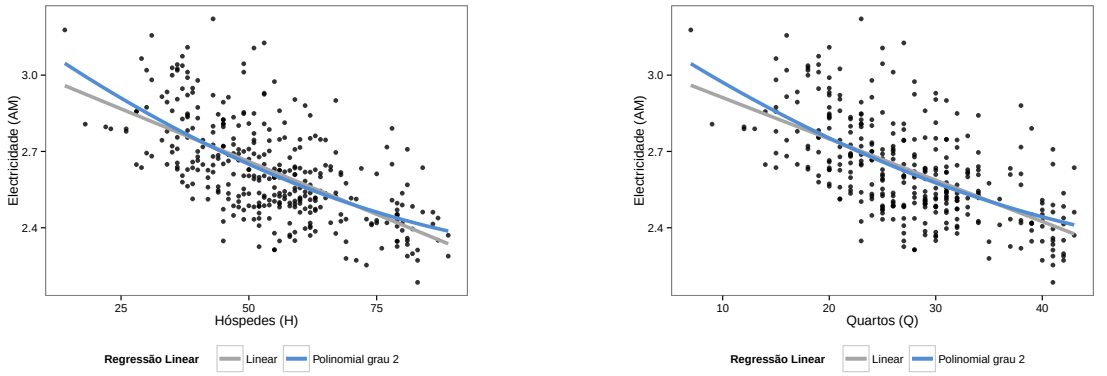


Figura E.7: Árvore de Decisão modelada e de apoio à estimativa com o conjunto de dados Alpino Atlântico

E.2 Quinta de S.João

E.2.1 Consumo de Água

E.2.1.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente

(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.8: Regressão Linear e Polinomial perante consumo médio de Água (AM) à escala logarítmica com o conjunto de dados Quinta de S.João

Tabela E.11: Sumário de Regressões Lineares para a estimativa de consumo de Água (AM) com o conjunto de dados Quinta de S.João

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.392	0.150	2.550	1.820	12.2
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.397	0.149	2.504	1.795	12.1
$\beta_0 + \beta_1 Q$	0.367	0.153	2.567	1.841	12.5
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.370	0.153	2.522	1.813	12.3

E.2.1.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H^2 + +\beta_2 TP75 \tag{E.9}$$

Tabela E.12: Performance dos modelos de regressão para a estimativa de eletricidade (AM) com o conjunto de dados Quinta de S.João

Regressão Linear log (AM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.400	0.149	2.504	1.780	12.0
KNN	-	-	2.610	1.860	12.6
SVM	-	-	2.610	1.824	12.1

E.2.1.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 H + \beta_2 Q2 \tag{E.10}$$

$$RL_2 = \beta_0 + \beta_1 TP75 + \beta_2 H + \beta_3 TM \tag{E.11}$$

$$RL_3 = \beta_0 + \beta_1 TP25 + \beta_2 TP75 + \beta_3 H \tag{E.12}$$

Tabela E.13: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

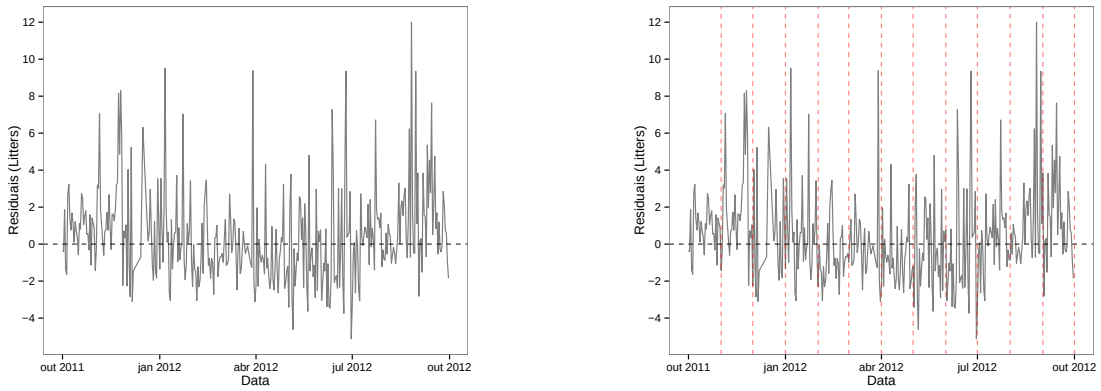
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	2.492	1.741	11.6
	2	1.908	1.443	11.2
	3	3.357	2.582	15.7
$RL_{p1,p2,p3}$	-	2.558	1.839	12.4
KNN	1	2.664	1.730	11.2
	2	1.946	1.449	11.3
	3	3.516	2.808	17.5
$KNN_{p1,p2,p3}$	-	2.690	1.888	12.7
SVM	1	2.495	1.795	12.2
	2	1.946	1.449	11.3
	3	3.612	2.851	17.7
$SVM_{p1,p2,p3}$	-	2.648	1.927	13.2
Melhor Conjunto				
$KNN_{p1}\ RL_{p2,p3}$	-	2.633	1.834	12.3

E.2.1.4 Estimativa utilizando Janelas Temporais

Tabela E.14: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

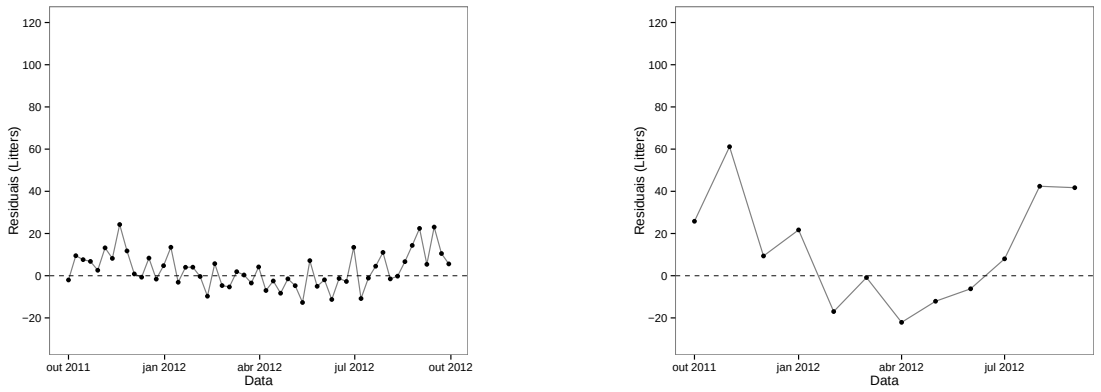
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	2.592	1.792	12.2
	2	1.967	1.553	12.5
	3	5.092	3.881	24.3
$RL_{g1,g2,g3}$	-	3.202	2.200	15.1
KNN	1	2.751	1.806	11.7
	2	1.931	1.495	11.9
	3	3.591	2.799	16.8
$KNN_{g1,g2,g3}$	-	2.747	1.935	13.0
$KNN_{completo}$	-	2.710	1.855	12.3
SVM	1	2.707	1.792	11.7
	2	2.098	1.648	13.0
	3	3.538	2.718	16.4
$SVM_{g1,g2,g3}$	-	2.751	1.960	13.2
$SVM_{completo}$	-	2.752	1.912	12.6
Melhor Conjunto				
SVM_{g1} KNN_{g2} SVM_{g3}	-	2.711	1.910	12.8

E.2.1.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.9: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta de S.João



(a) (b)

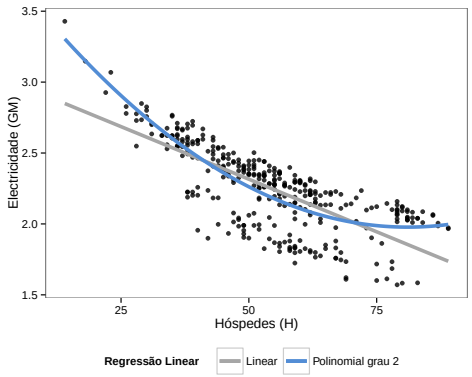
Figura E.10: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta de S.João

Tabela E.15: qsjoao Atlântico

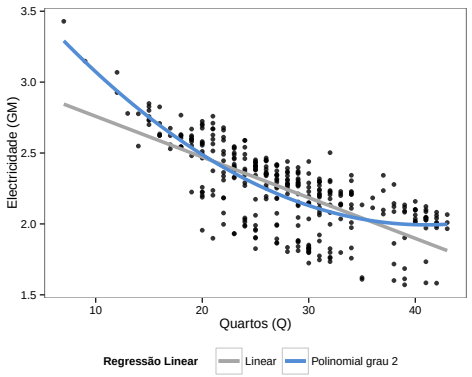
		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Água	$RL(H + H^2)$	1.795	12.1	1.04	7.2	0.752	5.3

E.2.2 Consumo de Gás

E.2.2.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.11: Regressão Linear e Polinomial perante consumo médio de Gás (GM) à escala logarítmica com o conjunto de dados Quinta de S.João

Tabela E.16: Sumário de Regressões Lineares para a estimativa de consumo de Gás (GM) com o conjunto de dados qsjoao Atlântico

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE*** (Watts/Dia)	**MAE*** (Watts/Dia)	**MAPE*** (%)
$\beta_0 + \beta_1 H$	0.552	0.194	1.644	1.283	14.80
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.628	0.177	1.652	1.256	14.83
$\beta_0 + \beta_1 Q$	0.504	0.205	1.675	1.298	15.0
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.561	0.191	1.69	1.27	15.16

E.2.2.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + + \beta_2 H^2 + \beta_3 TM \tag{E.13}$$

Tabela E.17: Performance dos modelos de regressão para a estimativa de eletricidade (GM) com o conjunto de dados Quinta de S.João

Regressão Linear log (GM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.830	0.119	0.979	0.729	8.3
KNN	-	-	1.238	0.928	10.4
SVM	-	-	1.108	0.868	9.8

E.2.2.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 TM + \beta_2 H + + \beta_3 H2 \tag{E.14}$$

$$RL_2 = \beta_0 + \beta_1 TP25 + \beta_2 Q + + \beta_3 H^2 \tag{E.15}$$

$$RL_3 = \beta_0 + \beta_1 TP25 + + \beta_1 H + + \beta_2 H^2 \tag{E.16}$$

Tabela E.18: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

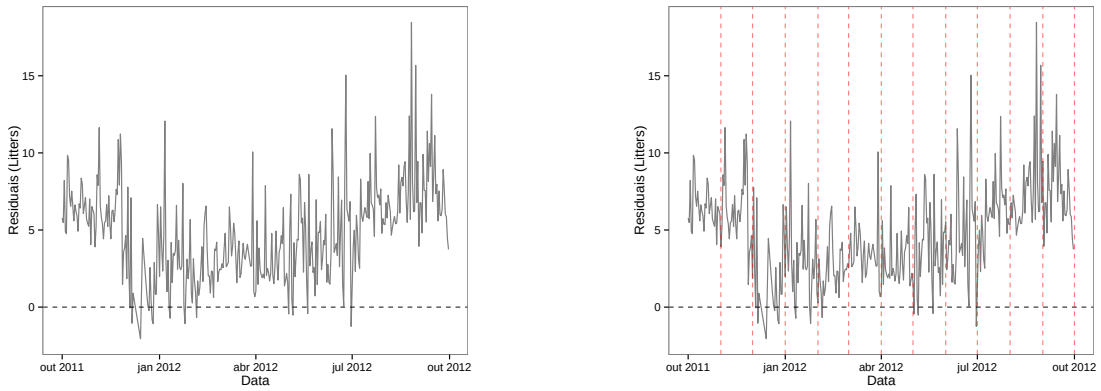
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	1.559	1.149	14.4
	2	0.834	0.656	7.7
	3	1.202	0.858	7.0
$RL_{p1,p2,p3}$	-	1.277	0.9195	10.4
KNN	1	1.479	1.033	12.4
	2	1.031	0.8106	9.3
	3	1.527	1.185	9.6
$KNN_{p1,p2,p3}$	-	1.360	0.995	10.7
SVM	1	1.232	0.9448	12.3
	2	1.031	0.8106	9.3
	3	1.571	1.075	8.6
$SVM_{p1,p2,p3}$	-	1.260	0.931	10.5
Melhor Conjunto				
$SVM_{p1}\ RL_{p2,p3}$	-	1.109	0.829	9.5

E.2.2.4 Estimativa utilizando Janelas Temporais

Tabela E.19: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

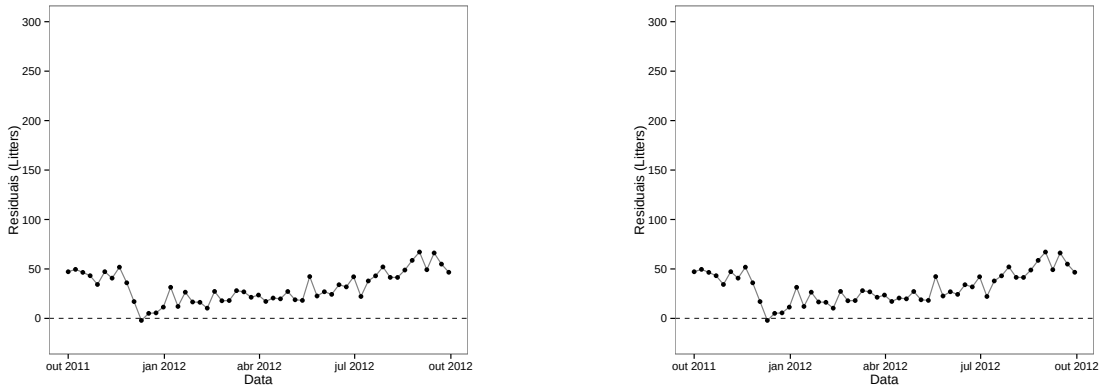
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	2.068	1.836	27.7
	2	0.918	0.7143	8.2
	3	1.678	1.309	11.0
$RL_{g1,g2,g3}$	-	1.675	1.344	17.4
KNN	1	1.705	1.287	16.0
	2	1.253	1.017	11.9
	3	1.761	1.335	10.5
$KNN_{g1,g2,g3}$	-	1.585	1.210	13.4
$KNN_{completo}$	-	1.759	1.270	13.9
SVM	1	1.487	1.124	15.5
	2	0.983	0.788	9.2
	3	2.267	1.877	14.97
$SVM_{g1,g2,g3}$	-	1.576	1.188	13.3
$SVM_{completo}$	-	2.329	1.680	19.3
Melhor Conjunto				
$SVM_{g1}\ RL_{g2}\ KNN_{g3}$	-	1.402	1.038	11.9

E.2.2.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.12: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados Quinta de S.João



(a) (b)

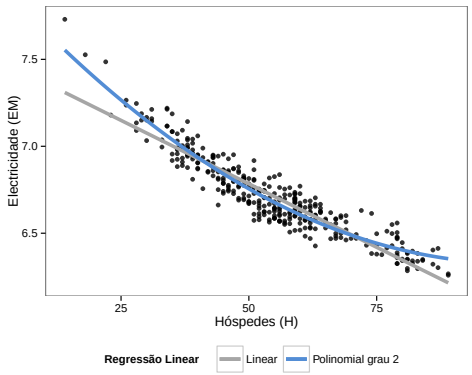
Figura E.13: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados Quinta de S.João

Tabela E.20: qsjoao Atlântico

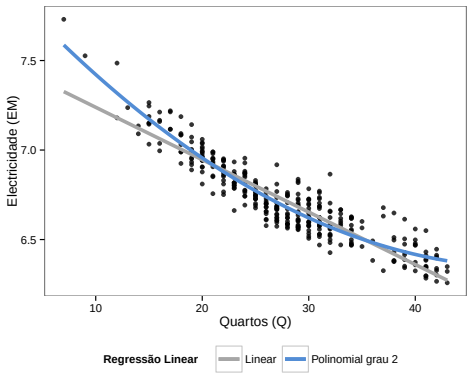
		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Gás	*RL*	0.729	8.3	4.74	34.1	4.70	33.9

E.2.3 Consumo de Eletricidade

E.2.3.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.14: Regressão Linear e Polinomial perante consumo médio de Água (EM) à escala logarítmica com o conjunto de dados Quinta de S.João

Tabela E.21: Sumário de Regressões Lineares para a estimativa de consumo de Água (EM) com o conjunto de dados Quinta de S.João

Regressão Linear log (EM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.862	0.085	146.2	93.40	9.3
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.898	0.073	118.1	84.28	8.8
$\beta_0 + \beta_1 Q$	0.846	0.089	143.0	94.10	9.6
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.882	0.078	113.1	83.72	8.9

E.2.3.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + \beta_2 Q + \beta_3 Q^2 + \beta_4 TP75 \tag{E.17}$$

Tabela E.22: Performance dos modelos de regressão para a estimativa de eletricidade (EM) com o conjunto de dados Quinta de S.João

Regressão Linear log (EM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.910	0.068	105.8	75.4	7.8
KNN	-	-	134.9	84.4	8.3
SVM	-	-	106.5	74.3	7.6

E.2.3.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 H + \beta_2 Q + \beta_2 H^2 \tag{E.18}$$

$$RL_2 = \beta_0 + \beta_1 H + \beta_2 H^2 + \beta_3 TP25 \tag{E.19}$$

$$RL_3 = \beta_0 + \beta_1 TP25 + \beta_2 H \tag{E.20}$$

Tabela E.23: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

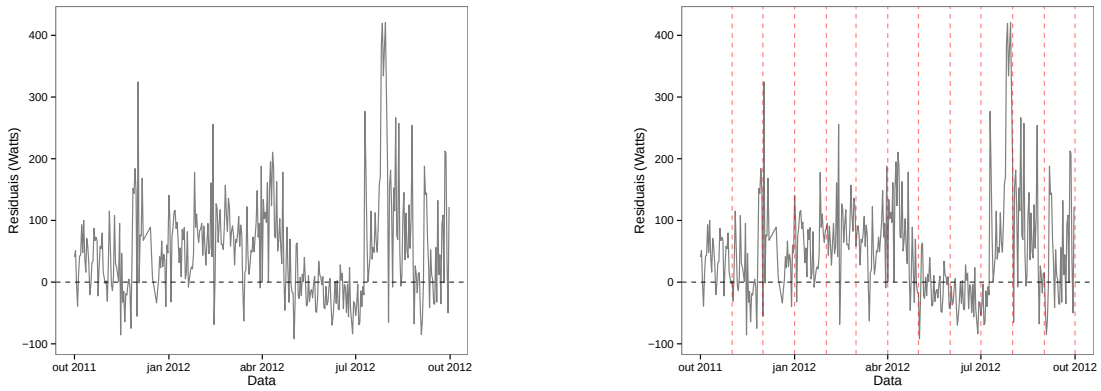
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	80.0	63.3	7.2
	2	68.5	49.1	6.4
	3	146.9	106.5	8.1
$RL_{p1,p2,p3}$	-	96.8	68.7	7.1
KNN	1	161.7	103.0	10.4
	2	77.9	55.2	7.1
	3	202.7	137.8	10.2
$KNN_{p1,p2,p3}$	-	151.7	95.3	9.3
SVM	1	187.2	118.6	11.8
	2	77.9	55.2	7.1
	3	218.1	131.3	9.3
$SVM_{p1,p2,p3}$	-	168.6	100.6	9.7
Melhor Conjunto				
$RL_{p1,p2,p3}$	-	96.8	68.7	7.1

E.2.3.4 Estimativa utilizando Janelas Temporais

Tabela E.24: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta de S.João

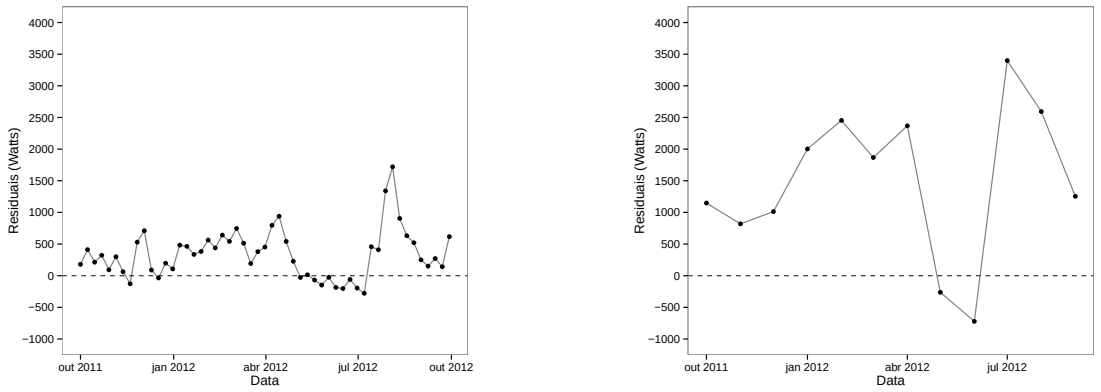
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	95.8	80.3	9.4
	2	76.4	56.1	7.3
	3	146.5	104.9	7.9
$RL_{g1,g2,g3}$	-	104.6	78.1	8.4
KNN	1	164.8	110.1	11.3
	2	96.1	75.5	10.1
	3	245.1	161.4	11.9
$KNN_{g1,g2,g3}$	-	170.1	110.6	11.1
$KNN_{completo}$	-	191.6	121.2	11.8
SVM	1	112.2	84.4	9.5
	2	85.6	65.7	9.0
	3	143.6	108.4	8.5
$SVM_{g1,g2,g3}$	-	112.8	83.9	9.1
$SVM_{completo}$	-	103.9	79.2	9.2
Melhor Conjunto				
$RL_{g1,g2,g3}$	-	146.5	104.9	7.9

E.2.3.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.15: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta de S.João



(a) (b)

Figura E.16: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta de S.João

Tabela E.25: qsjoao Atlântico

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Eletricidade	$RL_{p1,p2,p3}$	68.7	7.1	58.8	6.2	56.3	6.1

E.2.4 Árvore de Decisão de apoio à estimativa

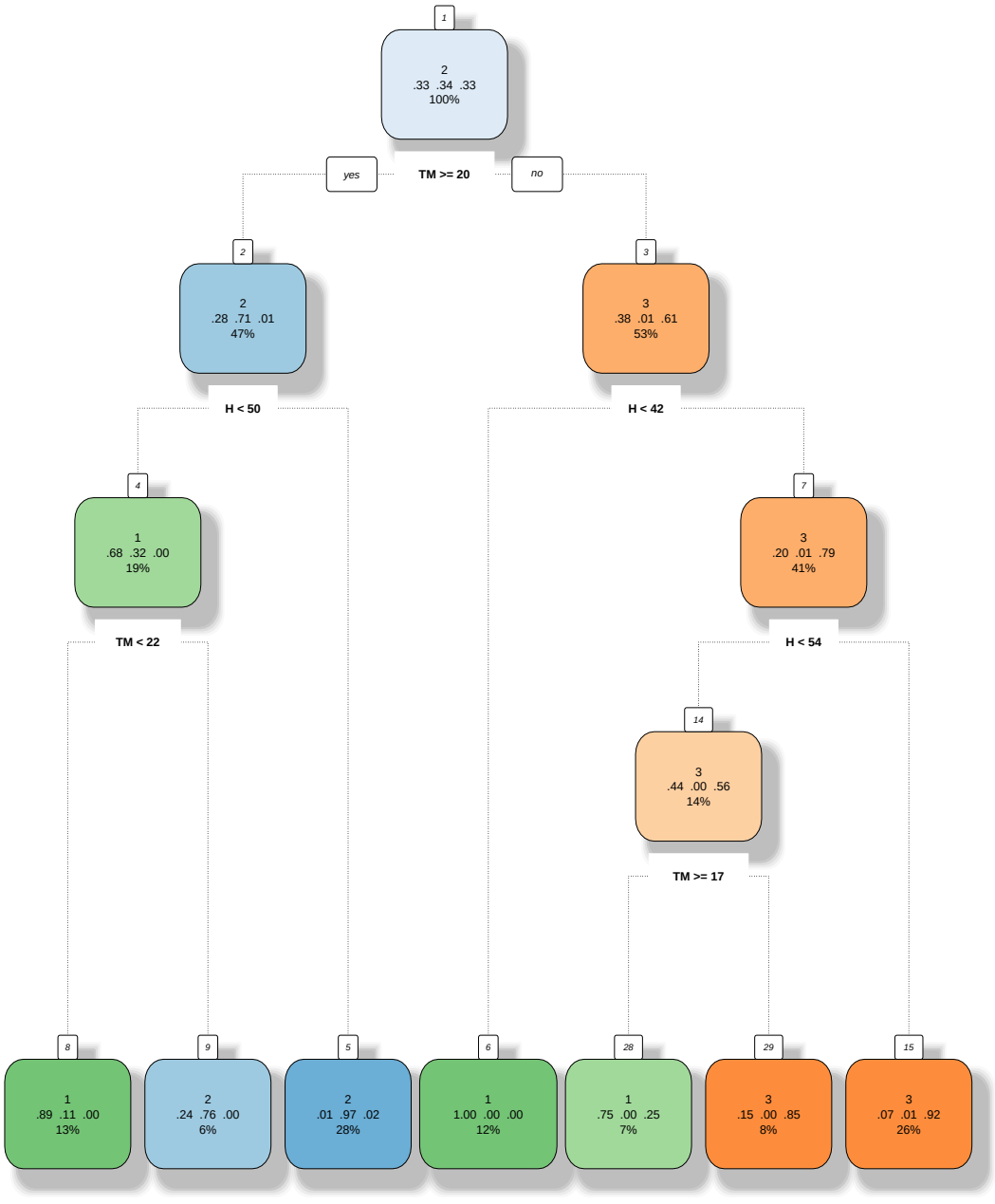
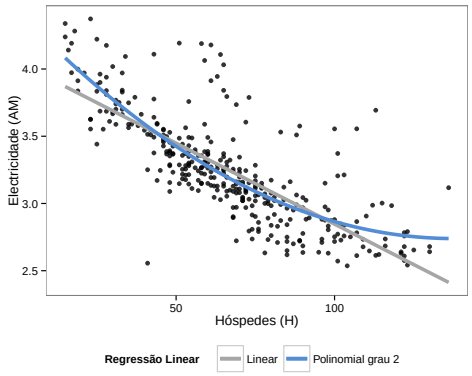


Figura E.17: Árvore de Decisão modelada e de apoio à estimativa com o conjunto de dados Quinta de S.João

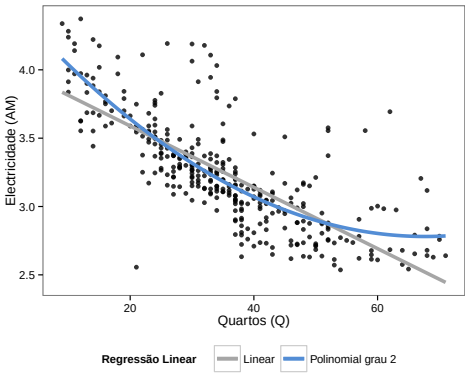
E.3 Quinta das Vistas

E.3.1 Consumo de Água

E.3.1.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.18: Regressão Linear e Polinomial perante consumo médio de Água (AM) à escala logarítmica com o conjunto de dados Quinta das Vistas

Tabela E.26: Sumário de Regressões Lineares para a estimativa de consumo de Água (AM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.604	0.250	9.129	6.807	28.3
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.639	0.239	9.855	7.283	29.4
$\beta_0 + \beta_1 Q$	0.581	0.257	9.10	6.860	28.8
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.632	0.241	10.26	7.574	30.6

E.3.1.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + +\beta_2 H^2 \tag{E.21}$$

Tabela E.27: Performance dos modelos de regressão para a estimativa de eletricidade (AM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (AM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.639	0.239	9.85	7.28	29.4
KNN	-	-	10.4	7.81	32.0
SVM	-	-	10.7	8.05	32.1

E.3.1.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 Q + \beta_2 Q^2 \tag{E.22}$$

$$RL_2 = \beta_0 + \beta_1 H + \beta_2 H^2 \tag{E.23}$$

$$RL_3 = \beta_0 + \beta_1 TP25 \tag{E.24}$$

Tabela E.28: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

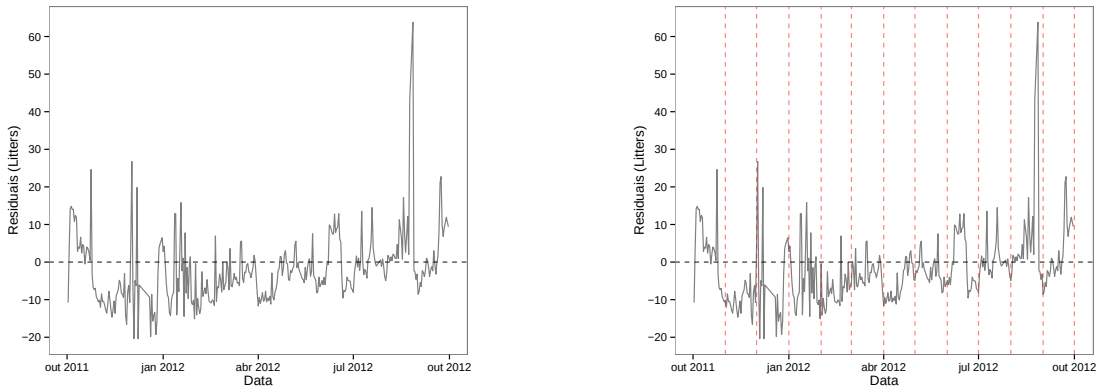
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	15.00	13.60	41.09
	2	9.613	5.933	21.35
	3	8.643	7.173	36.7
$RL_{p1,p2,p3}$	-	10.17	7.553	30.7
KNN	1	14.67	13.03	37.9
	2	10.52	6.733	24.4
	3	9.334	7.675	40.24
$KNN_{p1,p2,p3}$	-	10.74	8.031	31.1
SVM	1	16.35	13.80	39.56
	2	10.52	6.733	24.4
	3	9.017	7.903	43.8
$SVM_{p1,p2,p3}$	-	10.96	8.23	34.9
Melhor Conjunto				
$KNN_{p1}\ RL_{p2,p3}$	-	10.10	7.473	30.3

E.3.1.4 Estimativa utilizando Janelas Temporais

Tabela E.29: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

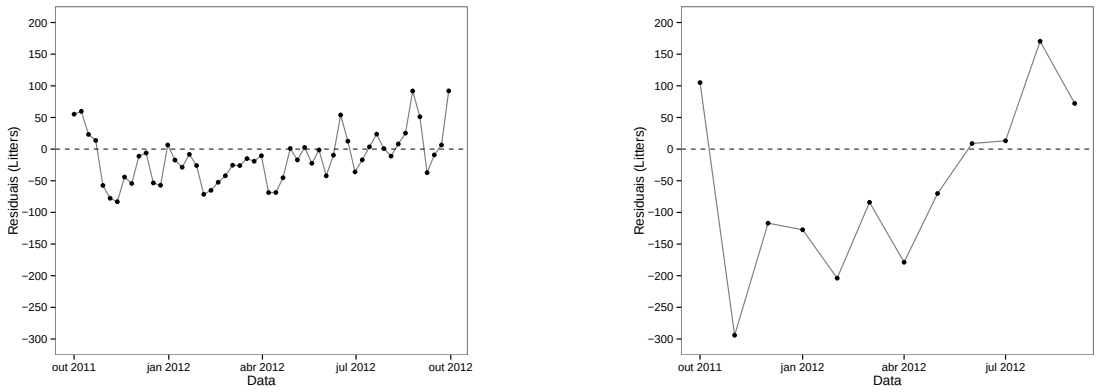
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	22.04	18.76	53.1
	2	9.483	6.252	24.04
	3	23.86	11.26	54.4
$RL_{g1,g2,g3}$	-	18.77	10.18	41.2
KNN	1	12.95	10.90	29.8
	2	9.925	6.435	26.65
	3	9.970	8.258	44.66
$KNN_{g1,g2,g3}$	-	10.42	7.8534	34.85
$KNN_{completo}$	-	10.77	7.862	33.5
SVM	1	12.74	10.015	28.2
	2	9.613	5.865	20.6
	3	8.0811	6.856	36.0
$SVM_{g1,g2,g3}$	-	9.521	6.879	28.3
$SVM_{completo}$	-	10.28	7.157	28.5
Melhor Conjunto				
$SVM_{g1,g2,g3}$	-	9.521	6.879	28.3

E.3.1.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.19: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta das Vistas



(a) (b)

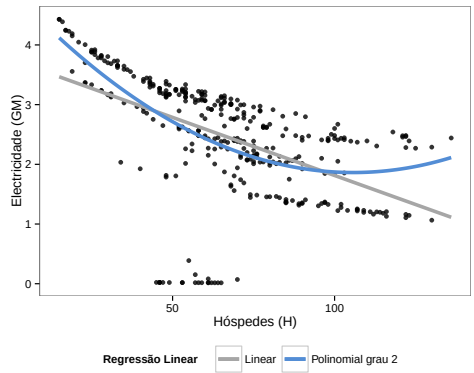
Figura E.20: Análise do Erro residual do consumo de Eletricidade (AM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta das Vistas

Tabela E.30: qvistas Atlântico

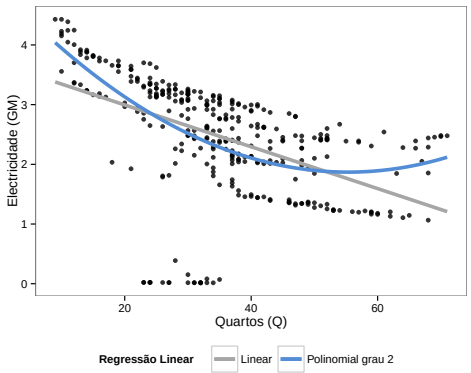
		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Água	$RL(H)$	6.807	28.3	5.29	22.0	4.28	18.2

E.3.2 Consumo de Gás

E.3.2.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.21: Regressão Linear e Polinomial perante consumo médio de Gás (GM) à escala logarítmica com o conjunto de dados Quinta das Vistas

Tabela E.31: Sumário de Regressões Lineares para a estimativa de consumo de Gás (GM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (AM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE*** (Watts/Dia)	**MAE*** (Watts/Dia)	**MAPE*** (%)
$\beta_0 + \beta_1 H$	0.552	0.194	6.936	4.644	175.1
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.628	0.177	11.00	6.062	205.9
$\beta_0 + \beta_1 Q$	0.504	0.205	6.919	4.681	177.1
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.561	0.191	11.55	6.127	211.9

E.3.2.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + + \beta_2 H^2 + \beta_3 TM \tag{E.25}$$

Tabela E.32: Performance dos modelos de regressão para a estimativa de eletricidade (GM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (GM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.613	0.598	8.617	5.546	134.7
KNN	-	-	9.86	6.51	171.3
SVM	-	-	50.21	11.45	274.6

E.3.2.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 TM + \beta_2 H + + \beta_3 H2 \tag{E.26}$$

$$RL_2 = \beta_0 + \beta_1 TM + \beta_2 H + + \beta_3 H2 \tag{E.27}$$

$$RL_3 = \beta_0 + \beta_1 TP25 + + \beta_2 TP75 + \tag{E.28}$$

Tabela E.33: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

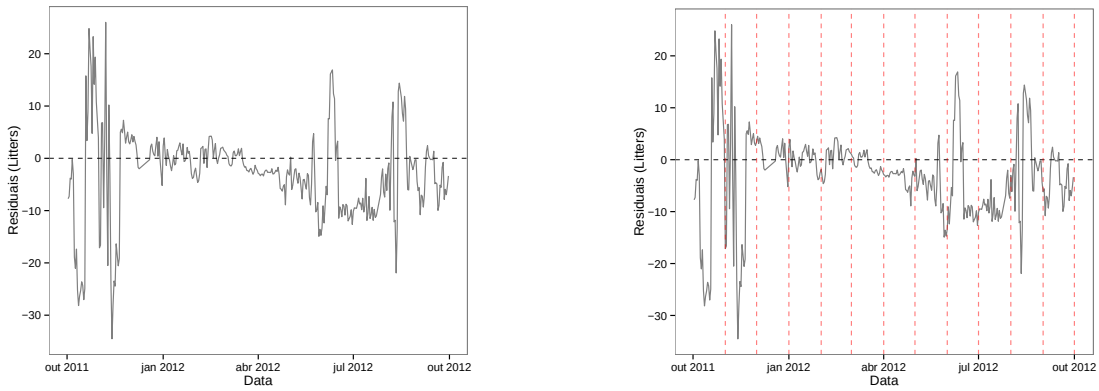
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	25.04	22.09	523.5
	2	2.713	2.285	96.38
	3	7.716	6.171	99.61
$RL_{p1,p2,p3}$	-	10.84	6.764	158.2
KNN	1	23.23	19.08	480.4
	2	3.117	2.549	131.7
	3	9.118	7.441	133.5
$KNN_{p1,p2,p3}$	-	10.78	6.998	181.8
SVM	1	18.38	16.01	335.2
	2	3.117	2.549	131.7
	3	8.207	7.036	98.6
$SVM_{p1,p2,p3}$	-	9.00	6.388	146.2
Melhor Conjunto				
$SVM_{p1}\ RL_{p2}\ SVM_{p3}$	-	8.94	6.27	131.1

E.3.2.4 Estimativa utilizando Janelas Temporais

Tabela E.34: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

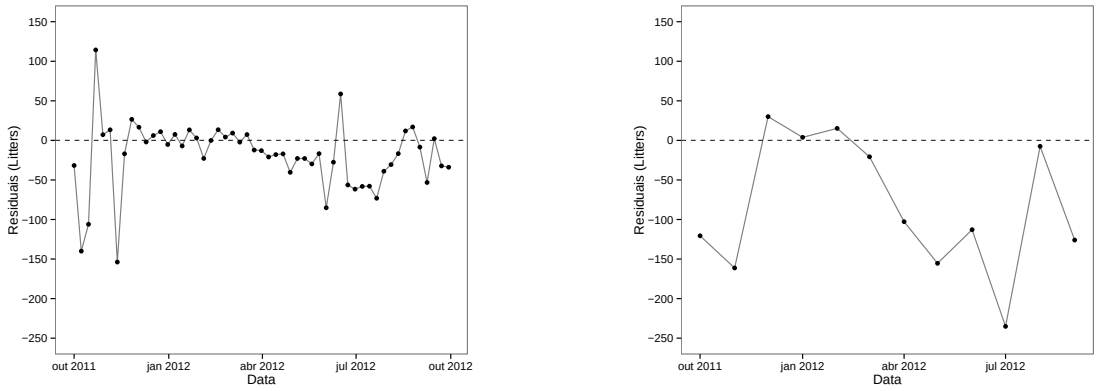
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	25.18	21.95	522.6
	2	5.802	4.598	161.6
	3	2045.4	276.7	1613.2
$RL_{g1,g2,g3}$	-	1342.3	124.2	837.9
KNN	1	19.73	15.62	256.8
	2	5.267	4.249	201.4
	3	10.36	8.585	149.5
$KNN_{g1,g2,g3}$	-	10.64	7.727	186.9
$KNN_{completo}$	-	12.67	8.38	183.0
SVM	1	20.33	17.97	374.5
	2	678.4	184.93	1556.2
	3	17.24	12.94	179.0
$SVM_{g1,g2,g3}$	-	443.9	87.22	678.2
$SVM_{completo}$	-	2676.2	355.7	2996.4
Melhor Conjunto				
KNN_{g1} RL_{g2} SVM_{g3}	-	14.05	9.753	182.5

E.3.2.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.22: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados de treino da unidade Quinta das Vistas



(a) (b)

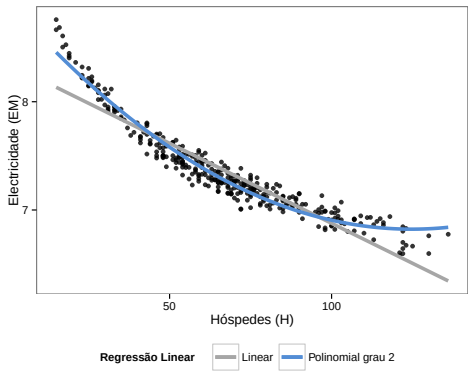
Figura E.23: Análise do Erro residual do consumo de Gás (GM) com o conjunto de dados de treino da unidade Quinta das Vistas

Tabela E.35: qvistas Atlântico

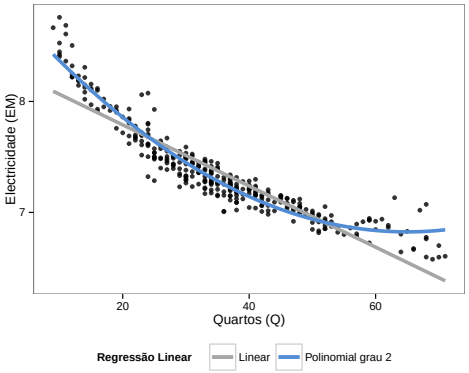
		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Gás	$SVM_{p1}\ RL_{p2}\ SVM_{p3}$	6.27	131.1	4.87	82.8	3.20	38.5

E.3.3 Consumo de Eletricidade

E.3.3.1 Estimativa com base na Ocupação ou Quartos Ocupados



(a) Regressão Linear e Polinomial de grau 2 com número de hóspedes (H) como variável independente



(b) Regressão Linear e Polinomial com quartos ocupados (Q) como variável independente

Figura E.24: Regressão Linear e Polinomial perante consumo médio de Água (EM) à escala logarítmica com o conjunto de dados Quinta das Vistas

Tabela E.36: Sumário de Regressões Lineares para a estimativa de consumo de Água (EM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (EM)	***Adjusted*** R^2	**Variância do Erro Residual**	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
$\beta_0 + \beta_1 H$	0.872	0.145	407.3	243.5	12.6
$\beta_0 + \beta_1 H + \beta_2 H^2$	0.950	0.090	195.5	146.8	9.1
$\beta_0 + \beta_1 Q$	0.838	0.164	425.5	262.1	13.5
$\beta_0 + \beta_1 Q + \beta_2 Q^2$	0.929	0.108	221.2	171.3	10.4

E.3.3.2 Estimativa utilizando Stepwise Selection

$$RL = \beta_0 + \beta_1 H + \beta_2 Q + \beta_3 Q^2 + \beta_4 TP75 \tag{E.29}$$

Tabela E.37: Performance dos modelos de regressão para a estimativa de eletricidade (EM) com o conjunto de dados Quinta das Vistas

Regressão Linear log (EM)	***Adjusted*** $\boldsymbol{R^2}$	$\boldsymbol{SS_{res}}$	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
RL	0.957	0.083	168.6	125.3	7.4
KNN	-	-	266.1	159.5	8.6
SVM	-	-	174.5	123.2	6.8

E.3.3.3 Estimativa utilizando Perfis de Consumo

$$RL_1 = \beta_0 + \beta_1 H + \beta_2 H^2 \tag{E.30}$$

$$RL_2 = \beta_0 + \beta_1 H + \beta_2 H^2 \tag{E.31}$$

$$RL_3 = \beta_0 + \beta_1 TM + \beta_2 H + \beta_3 Q + \beta_2 Q^2 \tag{E.32}$$

Tabela E.38: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

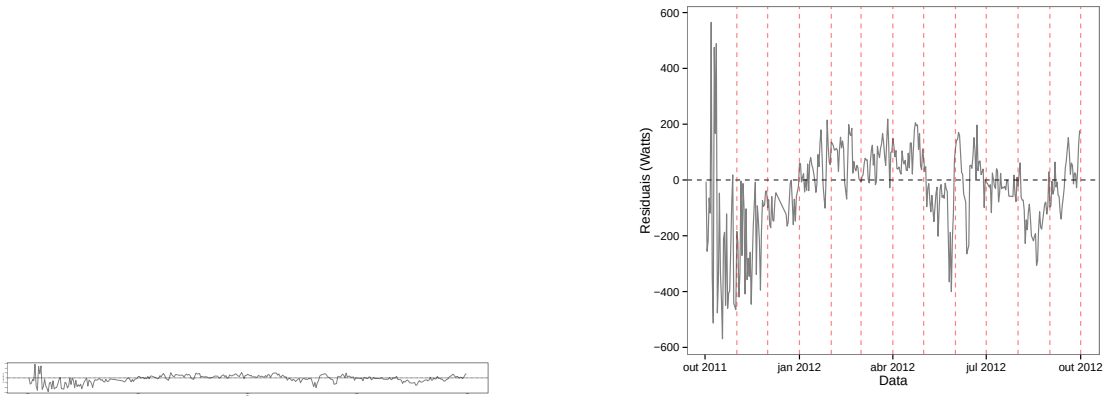
Algoritmo	**Agrupamento** p	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	324.9	280.8	7.8
	2	106.1	87.1	6.4
	3	114.1	86.3	6.0
$RL_{p1,p2,p3}$	-	159.3	114.2	6.4
KNN	1	784.7	568.2	15.2
	2	132.2	108.5	7.9
	3	168.8	118.8	8.2
$KNN_{p1,p2,p3}$	-	327.0	178.0	9.1
SVM	1	438.9	358.6	9.3
	2	132.2	108.5	7.9
	3	196.2	141.2	9.6
$SVM_{p1,p2,p3}$	-	226.6	158.0	8.8
Melhor Conjunto				
$RL_{p1,p2,p3}$	-	159.3	114.2	6.4

E.3.3.4 Estimativa utilizando Janelas Temporais

Tabela E.39: Melhor conjunto de variáveis com os vários modelos por sub-conjuntos com o conjunto de dados Quinta das Vistas

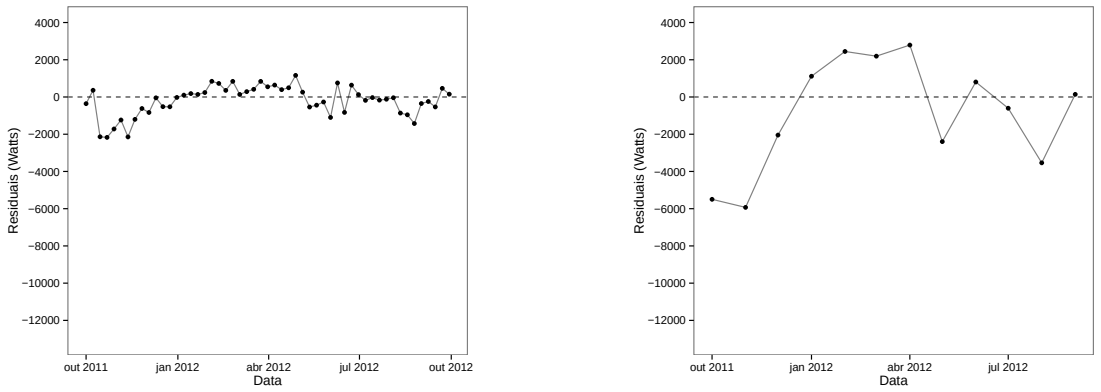
Algoritmo	**Agrupamento** g	**RMSE** (Watts/Dia)	**MAE** (Watts/Dia)	**MAPE** (%)
Regressão Linear	1	594.9	519.5	14.1
	2	103.1	77.5	5.7
	3	177.0	136.4	9.2
$RL_{g1,g2,g3}$	-	261.0	165.4	8.4
KNN	1	1370.2	1068.1	26.1
	2	230.1	175.1	13.1
	3	374.5	288.9	20.3
$KNN_{g1,g2,g3}$	-	590.6	350.5	18.0
$KNN_{completo}$	-	669.9	353.7	17.2
SVM	1	934.8	733.2	18.3
	2	120.5	98.0	7.1
	3	287.6	218.3	15.8
$SVM_{g1,g2,g3}$	-	406.9	239.7	12.4
$SVM_{completo}$	-	402.5	247.4	13.4
Melhor Conjunto				
$RL_{g1,g2,g3}$	-	261.0	165.4	8.4

E.3.3.5 Análise do Erro de Estimativa sob Diferentes Perspetivas



(a) (b)

Figura E.25: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta das Vistas



(a) (b)

Figura E.26: Análise do Erro residual do consumo de Eletricidade (EM) com o modelo de Regressão Linear *RL* e o conjunto de dados de treino da unidade Quinta das Vistas

Tabela E.40: qvistas Atlântico

		Diário		Semanal		Mensal	
Sensor	Modelo	MAE (Watts/Dia)	MAPE (% /Dia)	ME (Watts/Dia)	MPE (% /Semana)	ME (Watts/Dia)	MPE (% /Mês)
Eletricidade	$RL_{p1,p2,p3}$	114.2	6.4	94.0	5.4	87.4	4.9

APÊNDICE F

CONSUMO DE SERVIÇOS GRANULARIDADE HORÁRIA

F.1 Quinta de S.João

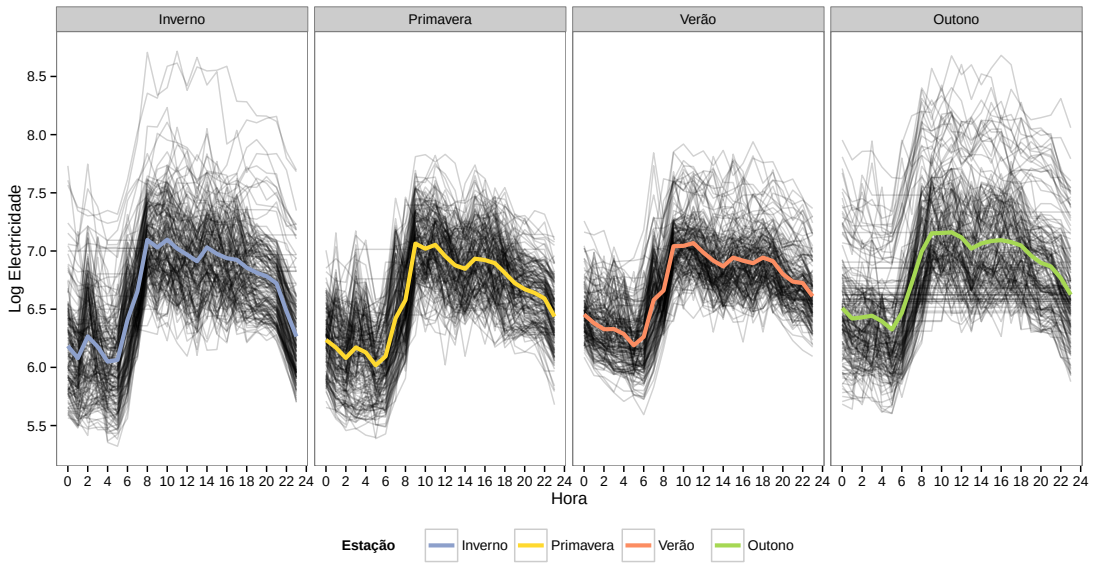


Figura F.1: Comportamento do sinal de consumo de Eletricidade a 24 horas e por estação do ano na unidade hoteleira Quinta de S.João

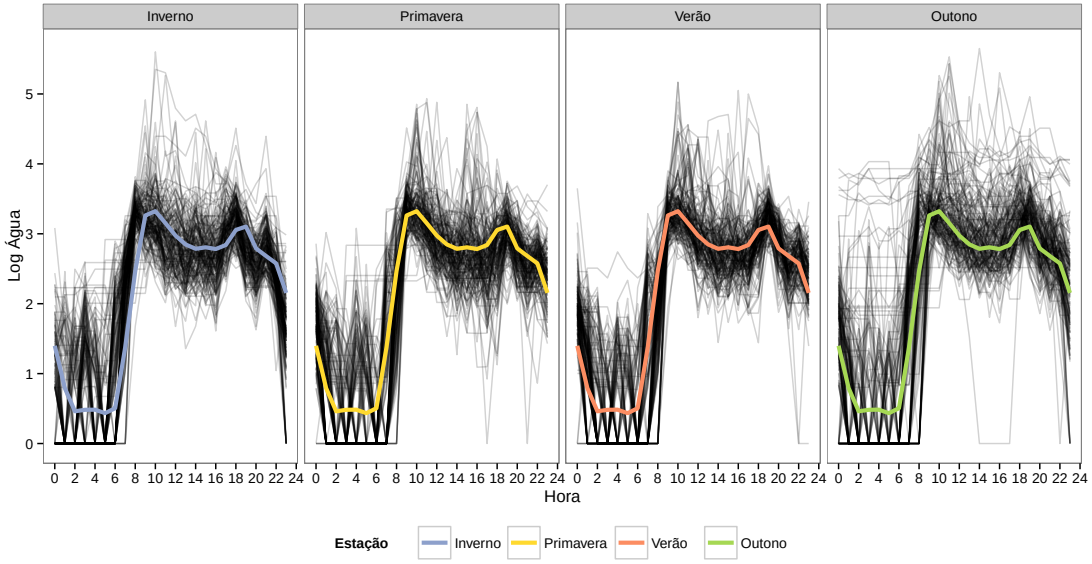


Figura F.2: Comportamento do sinal de consumo de Água a 24 horas e por estação do ano na unidade hoteleira Quinta de S.João

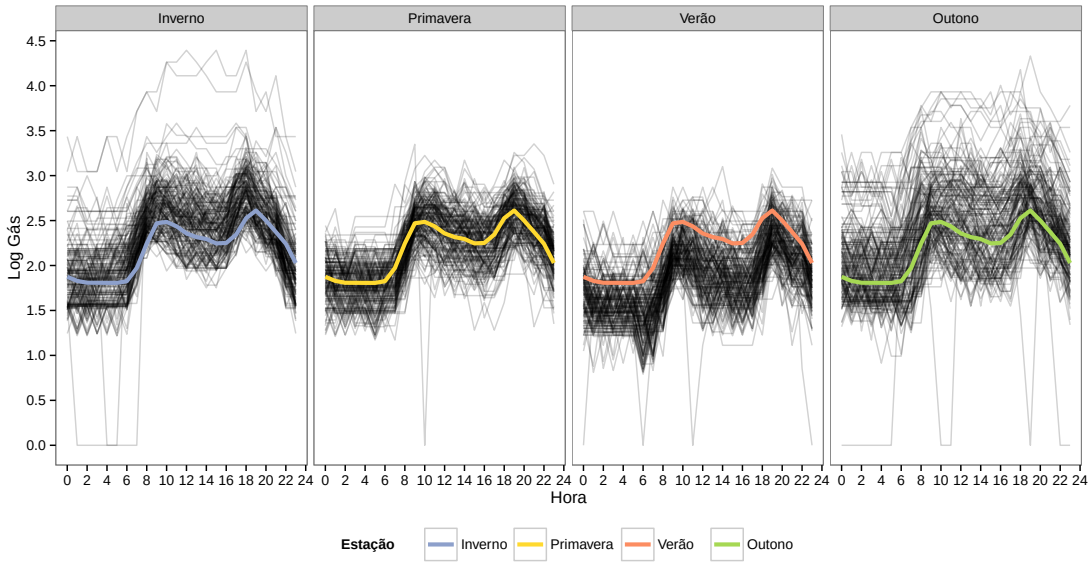


Figura F.3: Comportamento do sinal de consumo de Gás a 24 horas e por estação do ano na unidade hoteleira Quinta de S.João

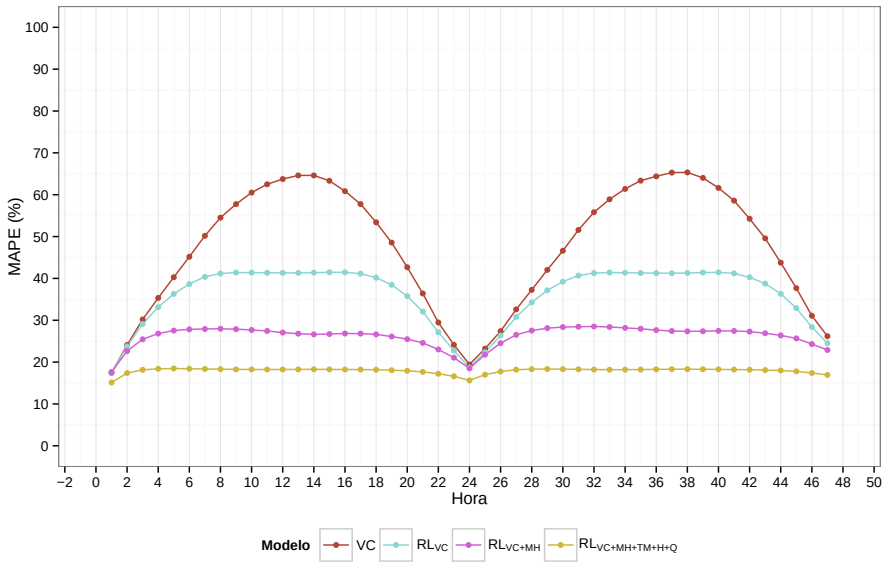


Figura F.4: Índice MAPE para a estimativa de consumo de Eletricidade a 48 horas no conjunto de dados Quinta de S.João

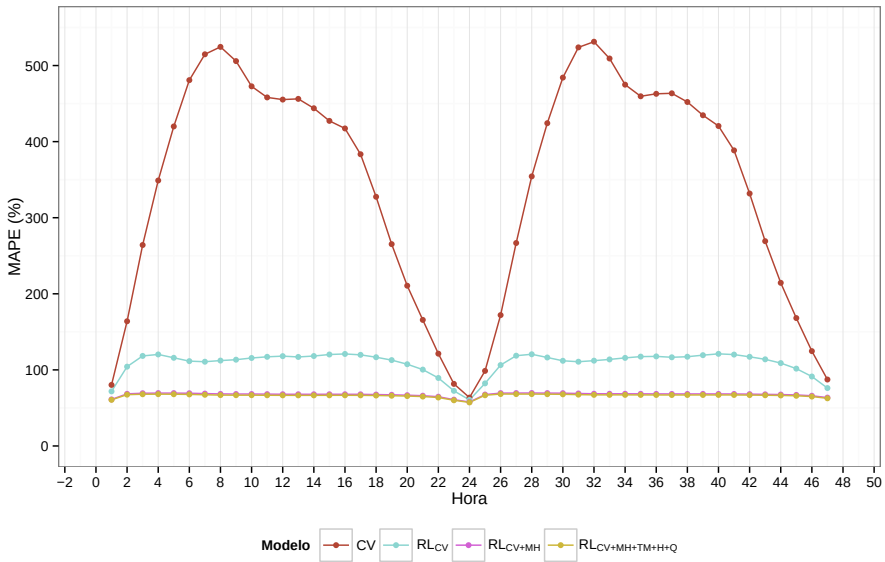


Figura F.5: Índice MAPE para a estimativa de consumo de Água a 48 horas no conjunto de dados Quinta de S.João

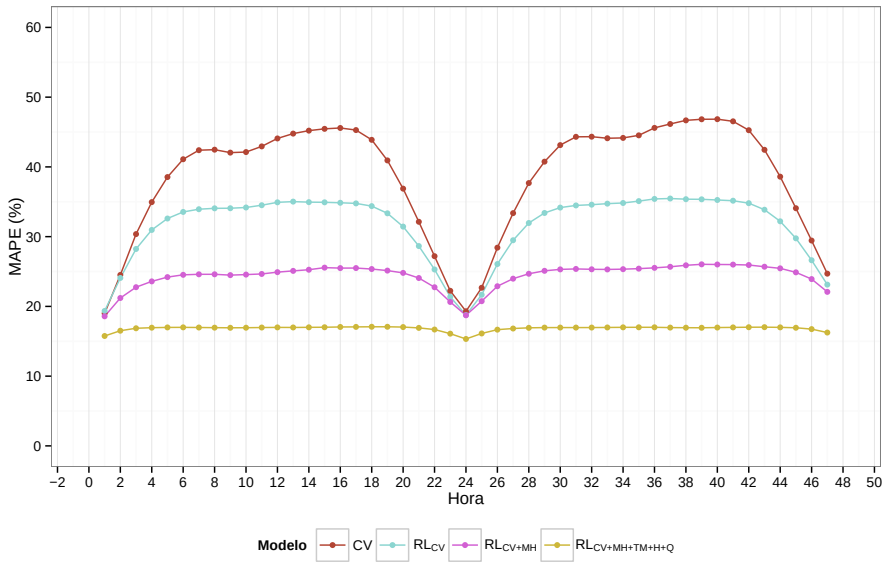


Figura F.6: Índice MAPE para a estimativa de consumo de Gás a 48 horas no conjunto de dados Quinta de S.João

F.2 Quinta das Vistas

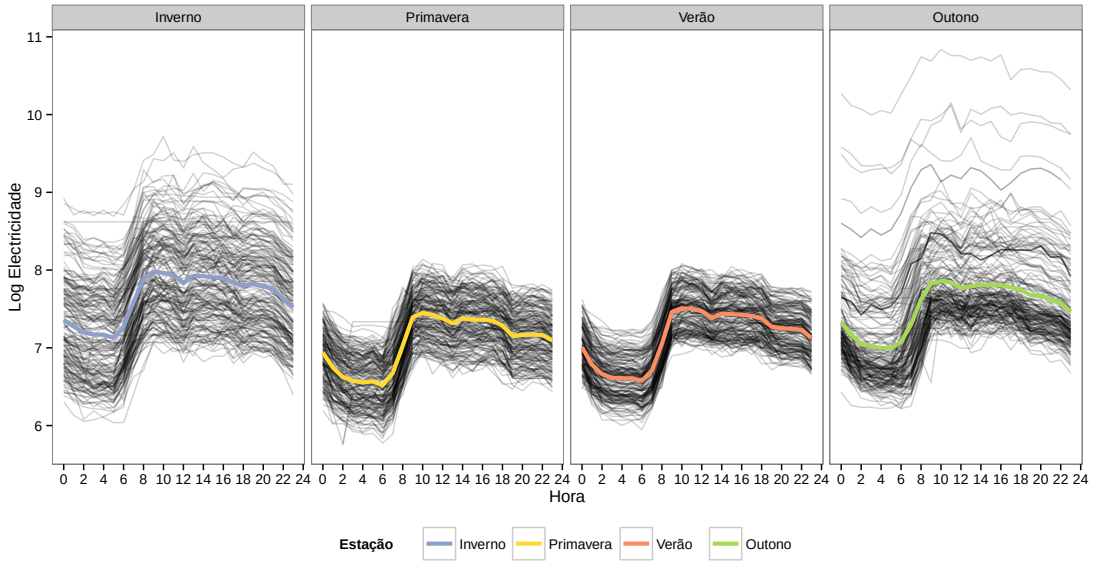


Figura F.7: Comportamento do sinal de consumo de Eletricidade a 24 horas e por estação do ano na unidade hoteleira Quinta das Vistas

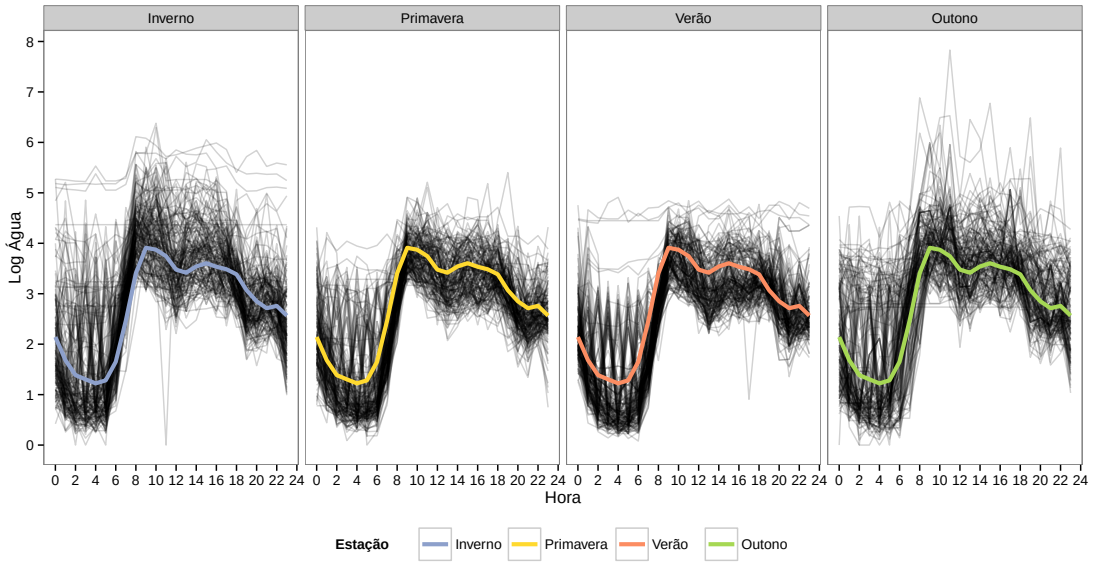


Figura F.8: Comportamento do sinal de consumo de Água a 24 horas e por estação do ano na unidade hoteleira Quinta das Vistas

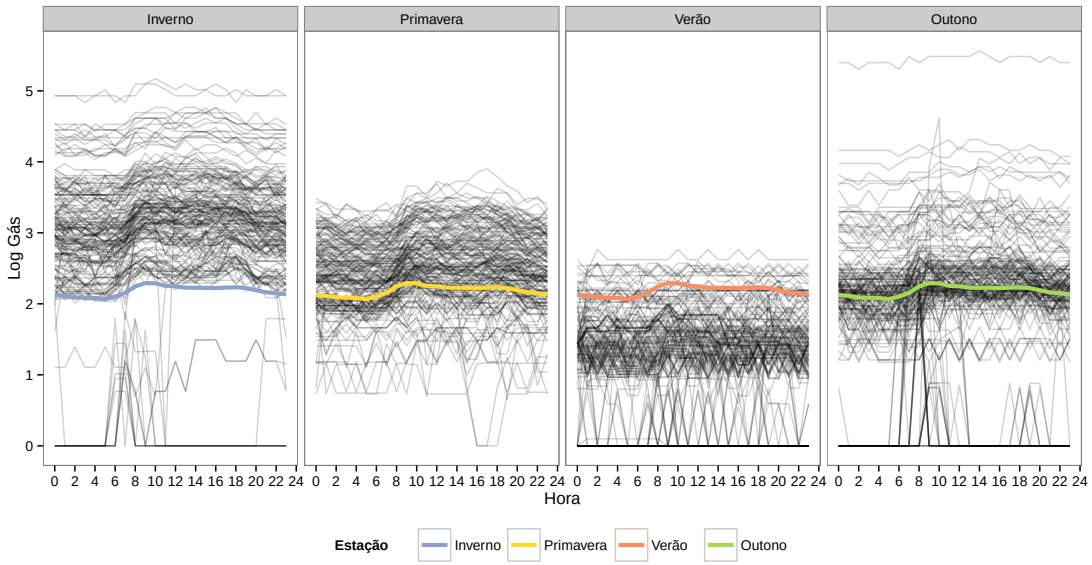


Figura F.9: Comportamento do sinal de consumo de Gás a 24 horas e por estação do ano na unidade hoteleira Quinta das Vistas

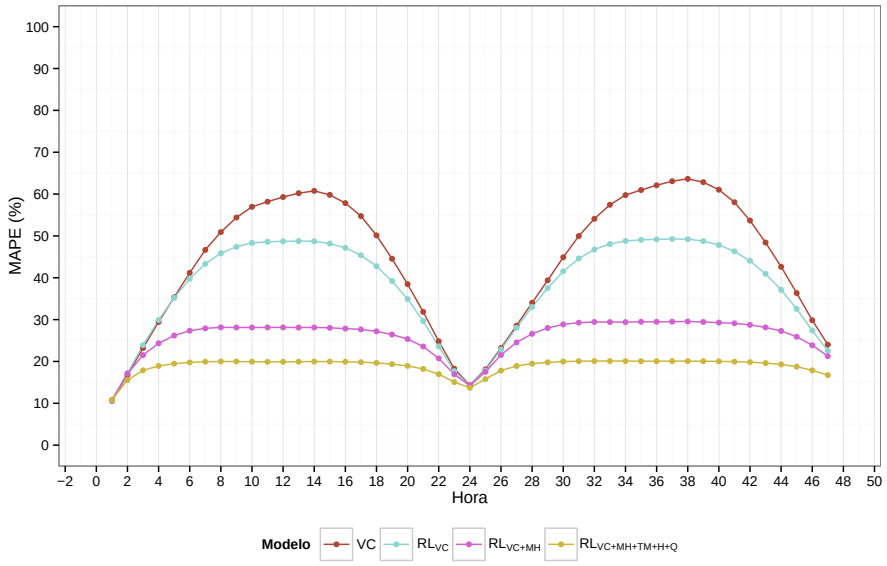


Figura F.10: Índice MAPE para a estimativa de consumo de Eletricidade a 48 horas no conjunto de dados Quinta das Vistas

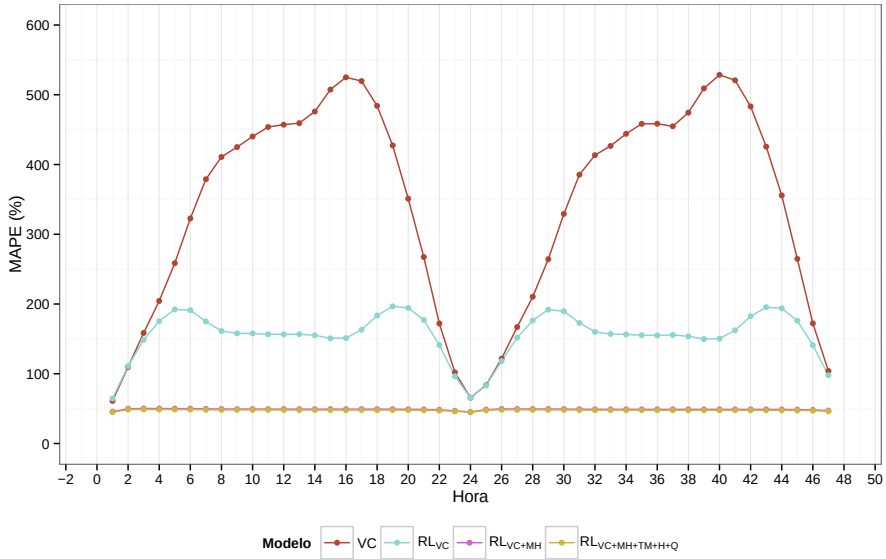


Figura F.11: Índice MAPE para a estimativa de consumo de Água a 48 horas no conjunto de dados Quinta das Vistas

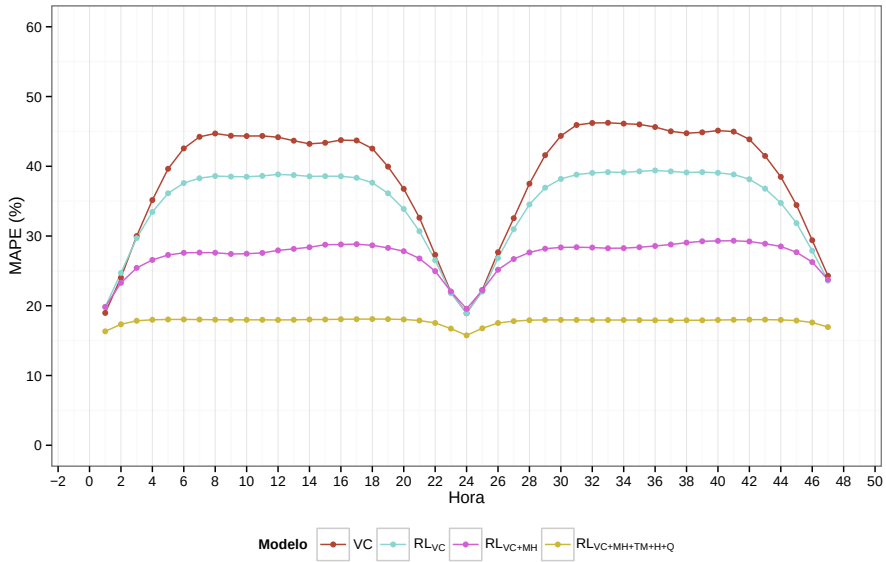


Figura F.12: Índice MAPE para a estimativa de consumo de Gás a 48 horas no conjunto de dados Quinta das Vistas