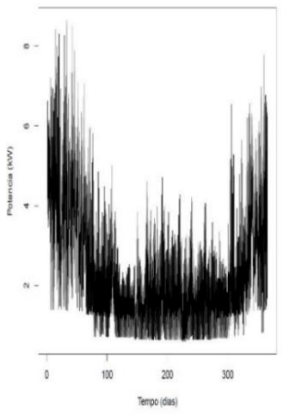
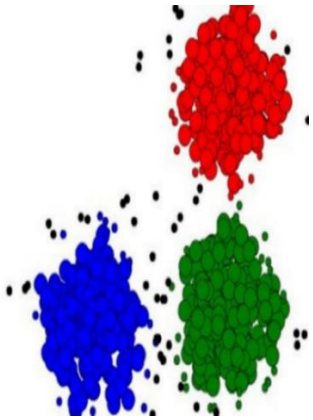
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Eletrotécnica de Energia e Automação





Segmentação de perfis de consumo de energia elétrica e gás

NUNO ANDRÉ SOBRAL DE SOUSA

(Licenciado em Engenharia Eletrotécnica)

Trabalho Final de Mestrado para obtenção do grau de Mestre
em Engenharia Eletrotécnica – Ramo Energia

Orientador (es):

Prof. Dr. João Hermínio Ninitas Lagarto

Profª. Drª. Ana Alexandra Antunes Figueiredo Martins

Júri:

Presidente: Prof. Dr. Filipe André de Sousa Figueira Barata

Vogais:

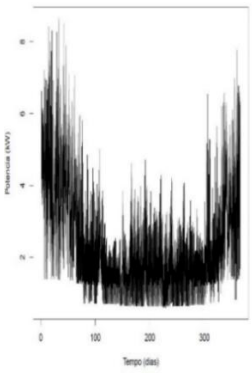
Profª. Drª. Alda Cristina Jesus Valentim Nunes de Carvalho

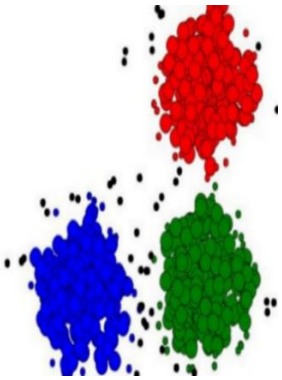
Prof. Dr. João Hermínio Ninitas Lagarto

Dezembro 2022

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Eletrotécnica de Energia e Automação





Segmentação de perfis de consumo de energia elétrica e gás

NUNO ANDRÉ SOBRAL DE SOUSA

(Licenciado em Engenharia Eletrotécnica)

Trabalho Final de Mestrado para obtenção do grau de Mestre
em Engenharia Eletrotécnica – Ramo Energia

Orientador (es):

Prof. Dr. João Hermínio Ninitas Lagarto

Profª. Drª. Ana Alexandra Antunes Figueiredo Martins

Júri:

Presidente: Prof. Dr. Filipe André de Sousa Figueira Barata

Vogais:

Profª. Drª. Alda Cristina Jesus Valentim Nunes de Carvalho

Prof. Dr. João Hermínio Ninitas Lagarto

Dezembro 2022

Agradecimentos

À minha mulher, Liliana, pelo apoio nesta aventura e paciência pela minha ausência no tempo despendido no mestrado.

Aos meus pais por me terem dado a oportunidade de estudar no Ensino Superior e assim poder continuar os meus estudos na área que gosto.

A todos os colegas me ajudaram a realizar trabalhos de grupo e apresentações durante o mestrado.

Aos meus orientadores por terem aceitado o meu convite para orientação, proporem o tema para este trabalho final e também a sua paciência durante a elaboração desta dissertação.

Ao ISEL por me ter dado a oportunidade de obter o grau de Mestre em Engenharia Eletrotécnica.

Muito obrigado a todos!

Resumo

No sector de energia elétrica é fundamental efetuar uma gestão eficiente e integrada da procura, de modo a melhorar a planificação de despacho da energia elétrica, identificar tendências de consumo de energia, falhas de despacho e consumos anormais. Atualmente com o avanço da tecnologia, é possível obter dados de consumidores de energia elétrica existentes para estudo, de uma maneira mais rápida e fidedigna através de contadores inteligentes (*smart meters)*.

O objetivo desta dissertação é identificar padrões de consumo residencial de energia elétrica e gás em diversas cidades dos EUA utilizando o *clustering*, também designado por análise de agrupamento, análise classificatória ou análise de *clusters*.

O *clustering* foi aplicado a dados de consumo residencial de energia elétrica e gás de 936 cidades dos EUA, utilizando o algoritmo *k-medoids*. Este algoritmo consiste na determinação de um elemento central para cada grupo distinto de consumidores, o medoide, em que a cada medoide é atribuído o(s) elemento(s) (consumidor(es)) mais próximos, consumidores com perfil aproximados. A proximidade dos consumidores é determinada pela média da combinação de quatro tipos de medidas de proximidade: distância euclidiana, distância baseada no coeficiente de correlação de Pearson, distância euclidiana entre funções de autocorrelação (ACF) e distância euclidiana entre periodogramas. O número de grupos de perfis de consumo residencial distintos indicado é determinado através do critério do índice *Silhouette*.

Os resultados obtidos da aplicação desta técnica, demonstraram que esta foi eficaz na tipificação dos consumos residenciais das cidades analisadas dos EUA. De um modo geral, o agrupamento definiu dois grupos distintos: um grupo que abrangia cidades dos estados do norte e centro, e outro grupo em que pertenciam cidades dos estados do sul dos EUA.

Palavras Chave: *Séries Temporais, Clustering, Consumo de energia elétrica e gás , k-medoids*

Abstract

In the electricity sector, it is essential to carry out an efficient and integrated management of demand, to improve the planning of electricity dispatch, identify trends in energy consumption, dispatch failures and abnormal consumption. Nowadays with the advancement of technology, it is possible to obtain data from existing electricity consumers for study, in a faster and more reliable way through smart meters.

The objective of this dissertation is to identify patterns of residential electricity and gas consumption in several US cities using clustering, also known as cluster analysis.

Clustering was applied to electricity and gas consumers data from 936 US cities, using the k-medoids clustering algorithm. This algorithm consists of determining a central element for each distinct group of consumers, the medoid, and for each medoid is assigned the closest element(s) (consumer(s)), consumers with similar consumption profile. Consumer proximity is determined by the average of the combination of four types of proximity measures: Euclidean distance, distance based on Pearson's correlation coefficient, Euclidean distance between autocorrelation functions (ACF) and Euclidean distance between periodograms. The number of groups of consumption profiles is obtained using the *Silhouette* index criteria.

The results obtained from the application of this technique showed that it was effective in typifying the residential consumption in the analyzed US cities. In general, the typification of different residential consumption patterns defined two distinct groups: a group that included cities in the northern and central states, and another group in which cities in the southern states of the USA belonged.

Keywords: *Time Series, Clustering, Electricity and gas consuption,k-medoids*

Índice

Índice de Figuras

Índice de Tabelas

Listagem de siglas

ACF	AutoCorrelation Function (Função de Autocorrelação)
BEopt	Building Energy Optimization Tool (Ferramenta de Otimização de Energia em Edifícios)
CCM	Consumer Mixture Model (Modelo de mistura de consumidor)
CNY	Yuan Renminbi (Moeda chinesa)
DBI	Davies-Bouldin Index (Índice Davies-Bouldin)
DOE	Department of Energy (Departamento de Energia dos EUA)
DTW	Dynamic Time Warping
EBF	Bllipsoidal-Basis-Function (Função Base Elipsoidal)
EUA	Estados Unidos da América
FCM	Fuzzy C-Means
FV	Fotovoltaico
GGM	Modelos Probabilisticos Gausianos
GIS	Geographical Information System (Sistema de informação Geográfica)
IBT	Increasing-Block Tariff (Tarifa de Bloco Crescente)
IECC	International Energy Conservation Code (Código Internacional de Conservação de Energia)
kW	Quilowatt
kWh	Quilowatt-hora
MSE	Mean Squared Error (Erro Quadrático Médio)
NREL	National Renewable Energy Laboratory (Laboratório Americano de Energia Renovável)
SIL	Índice *Silhouette*
SOFNN	Self-Organizing Fuzzy Neural Networks
SOM	Self-Organizing Map
WCBCR	Razão da soma dos quadrados dentro de um cluster

Capítulo 1 - Introdução

No presente capítulo são apresentados o enquadramento, o objetivo, a metodologia aplicada e a estrutura deste trabalho.

1 Introdução

1.1 Enquadramento

Atualmente e com grande tendência no futuro, no mercado de energia elétrica é fundamental uma gestão eficiente e integrada da procura, de modo a melhorar a planificação de despacho da energia elétrica, identificar tendências de consumo de energia, falhas de despacho e consumos anormais.

Com o avanço da tecnologia, é possível obter dados de consumidores de energia elétrica existentes para estudo, de uma maneira mais rápida e fidedigna através de contadores inteligentes (*smart meters*).

Além da caracterização dos perfis de consumo agregado, é também importante identificar o comportamento de consumo dos consumidores finais durante o ano, quais as alturas do ano com maior/menor consumo e a sua relação com a região geográfica/climática que se inserem esses consumidores [1].

1.2 Objetivo

O objetivo desta dissertação é identificar padrões de consumo residencial de energia elétrica e gás em diversas cidades dos EUA utilizando o *clustering*, também designado por análise de agrupamento, análise classificatória ou análise de *clusters*.

1.3 Metodologia

A metodologia utilizada para este trabalho é a aplicação de técnicas de clustering a um conjunto de dados de consumo residencial de energia elétrica e gás através da aplicação do algoritmo de agrupamento, o *k-medoids*. Este algoritmo consiste na determinação de um elemento central para cada grupo de consumo padrão distinto, o medoide, em que a cada medoide é atribuído o(s) elemento(s) (consumidor(es)) mais próximos, consumidores com perfis

de consumo similar. A proximidade dos consumidores é determinada pela média da combinação de quatro tipos de medidas de proximidade: distância euclidiana, distância baseada no coeficiente de correlação de *Pearson*, distância euclidiana entre funções de autocorrelação (ACF) e distância euclidiana entre periodogramas.

Para a determinação do número de grupos (*clusters*) de perfis de consumo residencial padrão distintos é utilizado o índice *Silhouette* como critério.

Os dados para este trabalho são referentes a consumos anuais de consumidores residenciais de 936 cidades de vários estados dos EUA que são analisados e aos quais é efetuado o *clustering.* A análise de dados e obtenção de resultados são realizados com recurso a uma linguagem de programação designada de R.

1.4 Estrutura

A presente dissertação divide-se em cinco capítulos (introdução, estado da arte, metodologia, segmentação de perfis de consumo e, conclusões e trabalho futuro), bibliografia e anexos.

O capítulo 1, refere-se à introdução do tema desta dissertação, qual a motivação para realização desta, objetivos pretendidos e metodologia adotada.

No capítulo 2, descreve-se alguns estudos efetuados anteriormente utilizando o *clustering* aplicado à engenharia eletrotécnica, os métodos utilizados por estes e respetivas conclusões.

No capítulo 3, começa-se por explicar a importância da análise de dados através de medidas descritivas, *boxplots* e *outliers*, histogramas, ACF, periodogramas e cronogramas. Além disto, também são apresentadas quais foram as medidas de proximidade utilizadas, que método de agrupamento foi aplicado, bem como qual critério foi utilizado para obtenção do número indicado de grupos para efetuar o *clustering*.

No capítulo 4, apresentam-se os resultados relativos ao *clustering* efetuado aos consumidores residenciais de 50 cidades, uma cidade por cada estado dos EUA e a consumidores residenciais de 936 cidades, a totalidade dos dados obtidos. Também são apresentados resultados das medidas descritivas de duas cidades que representam os diferentes grupos de consumidores residenciais padrão existentes (medoides) bem como os respetivos resultados através da análise de histogramas, *boxplots* e *outliers*, ACF e cronogramas. Além disto, são apresentados os resultados geográficos do *clustering*, a relação existente entre o *clustering* efetuado e as diferentes regiões climáticas existentes nos EUA e ainda os resultados do agrupamento de cidades dentro de cada estado.

No capítulo 5, apresentam-se as principais conclusões tendo em conta os resultados obtidos no capítulo 4 com aplicação do método de agrupamento escolhido nesta dissertação, *o k-medoids*, bem como algumas sugestões para futuros trabalhos.

Por fim, nos anexos, apresentam-se todos resultados detalhados da análise efetuada nesta dissertação bem como o código em R para aplicação da metodologia proposta.

Capítulo 2 – Estado da Arte

O presente capítulo descreve a importância do *clustering* de dados e são citados alguns artigos em que utilizaram esta técnica aplicada à engenharia eletrotécnica.

2 Estado da Arte

2.1 Análise de agrupamento

Atualmente com a introdução de contadores inteligentes nas redes elétricas por todo o mundo tornou-se necessário interpretar os dados provenientes destes e em que medida estes podem ajudar, por exemplo, no ajuste de tarifas de eletricidade, na identificação de perfis de consumo, ajuste de uma rede elétrica e até na definição de locais prováveis para implementação de uma central elétrica.

O *clustering* é um processo que se efetua de forma intuitiva no nosso dia a dia sem dar conta. Exemplo disso, é a forma como se classificam pessoas ou objetos consoante as suas características tais como: cor, volumetria, cheiro, altura, comportamento entre outras. Desta forma, agrupa-se objetos, pessoas ou dados pela maior semelhança possível de uma ou mais características de modo que esses grupos sejam os mais distintos possíveis.

2.2 Técnicas de agrupamento aplicadas à engenharia eletrotécnica

Em seguida são descritos alguns exemplos da aplicação de *clustering* na engenharia eletrotécnica.

Um exemplo, é a sua utilização para determinar possíveis locais onde implementar uma rede de energia elétrica, descrito no artigo [2]. Neste estudo, foi proposta uma nova metodologia para selecionar locais candidatos para centrais de energia solar, incluindo o desenvolvimento de mapas, utilizando técnica de agrupamento por forma a determinar regiões de atributos de qualidade solar coerentes. Estas regiões são definidas por um atributo que considera tanto a claridade solar quanto a variabilidade solar. A metodologia proposta foi uma combinação de dois algoritmos de *clustering*: a propagação de afinidade **(***Affinity Propagation***)** e o *k-means* (valor médio entre elementos da amostra), a fim de produzir partições dos dados para diversos números de grupos (*clusters)*. A propagação de afinidade fornece

os medoides iniciais por um mecanismo que considera todos os pontos de dados como potenciais medoides de *clusters*.

Para testar a metodologia, foram utilizados dados provenientes do Laboratório Americano de Energia Renovável (*NREL*) sob a forma de mapas e dados do Sistema de Informação Geográfica (*GIS*). O local de teste foi uma ilha americana no Pacífico.

Ao utilizar índices de validação interna, atribui-se o grau de "bom" agrupamento, que ocorre para cada partição numa variedade de k número de *clusters* em termos de maximização, ocorrendo a atribuição dos elementos dos *clusters* e a separação entre *clusters*. O método L é implementado para determinar o ponto de deflexão onde o número de *clusters* começa a revelar agrupamentos confiáveis de índices de validade qualitativos e inalterados.

De acordo com esta metodologia de aprendizagem não supervisionada, o intervalo restrito proposto do número de *clusters* apropriado, conjuntamente com a sua segmentação espacial pode fornecer um esquema mínimo razoável de agrupamento de modo que este apresente uma estrutura coerente de atributos solares.

A validade dos mapas de agrupamento foi avaliada de várias maneiras:
primeiro, uma análise de correlação dentro e entre os *clusters* foi usada para identificar o grau de distinção entre diferentes agrupamentos e indicar aquele que satisfaz um determinado critério de máxima distinção.
Em segundo lugar, a reprodutibilidade de um mapa de agrupamento também é confirmada em relação à semelhança entre a partição existente e a nova partição gerada.

Finalmente, destaca-se a contribuição de um mapa de agrupamento para a instalação eficiente de parques solares provando que a seleção de locais entre diferentes *clusters* pode apenas ter em conta as centrais solares candidatas. Concluiu-se também neste estudo que é possível atribuir informações relacionadas com a variabilidade solar e taxas de aumento de irradiância para cada cluster e para cada combinação de dois *clusters* de modo que se possa produzir mapas que complementem os mapas tradicionais de disponibilidade solar.

Os mapas de agrupamento resultantes demonstraram ser importantes para decidir onde implantar a telemetria solar por forma a melhorar a precisão de previsão da geração solar. Além disto, estes também são importantes para decidir onde implantar centrais de energia solar de modo que o agregado de potência minimize a ocorrência de grandes quebras de fornecimento, para decidir onde implantar sensores solares a fim de aumentar as variáveis de entrada para melhorar a previsão e assim reduzir a incerteza de recurso solar.

Ainda sobre o tema de *clustering* aplicado à energia solar, em outro artigo [3] , foi abordado o estudo de séries temporais e as implicações físicas em sistemas fotovoltaicos sob condições desconhecidas de funcionamento.

O estudo teve como objetivo desenvolver uma ferramenta que permita de uma forma dinâmica detetar as falhas num sistema solar fotovoltaico ao processar séries temporais, que contenham dados de tensão e corrente do sistema, temperatura e irradiância só em relação à variação de irradiação. Com este propósito, foi usado o método *k-means* para efetuar o *clustering* em séries temporais, por forma a categorizar as falhas no sistema e distingui-las do seu funcionamento normal.

Para testar a metodologia foi utilizado um leque muito restrito de condições de funcionamento usuais num sistema fotovoltaico, aplicando individualmente, o sombreamento total, o sombreamento parcial, falhas em circuito aberto (falha em um dos módulos) ou o funcionamento normal sem qualquer falha.

Após analisar os dados e efetuar o teste desta metodologia, os resultados experimentais revelaram que diferentes condições de um sistema fotovoltaico estão fisicamente relacionadas com grupos distintos de séries temporais, relacionando-os com correntes e tensões. Os resultados também revelaram que o número ideal para o número de *clusters* é 4, exatamente o número de tipo de situações que se pretendia testar (condições normais, sombreamento total, parcial ou falha em um dos módulos).

Este método demonstrou ser capaz de distinguir com clareza as diferentes falhas no sistema, distinguindo a falha de origem elétrica (falha num dos módulos) das falhas ambientais (sombreamento total ou parcial).

O *clustering* também pode ser utilizado no teste de esquemas tarifários de eletricidade aplicando estes ao agrupamento de perfis de consumo, exemplo disto é o descrito no artigo [4]. Neste, foi sugerido um novo esquema tarifário, uma tarifa por blocos crescentes (*IBT- Increasing-Block Tariff*), que não afetará apenas a conta de eletricidade para os residentes, mas também a mudança nos comportamentos de consumo residencial de eletricidade. O IBT proposto foi considerado para a segmentação dos consumidores sendo formado por 3 blocos: o primeiro englobava consumidores com consumo inferior a 200 kWh/mês com um preço de 0,446 CNY/ kWh, o segundo englobava consumidores com consumo mensal entre os 201 e os 400 kWh com um preço de 0,5483 CNY/ kWh e o terceiro englobava consumidores em que o seu consumo estava acima dos 400 kWh/mês com um preço de 0,7983 CNY/ kWh, considerando que o yuan renminbi (CNY) é a moeda oficial da República Popular da China.

Este novo esquema tarifário foi aplicado a um conjunto de dados de consumo diário de 533 famílias de abril de 2014 a fevereiro de 2015 obtidos através de contadores inteligentes. O seu objetivo era analisar perfis de consumo através de uma previsão de carga dos consumidores a curto prazo, a fim de se aplicar uma tarifa justa para os consumidores.

Para se efetuar este estudo, utilizou-se um algoritmo de previsão de redes neuronais difusas auto-organizáveis (SOFNN, Self-Organizing Fuzzy Neural Networks). Este algoritmo tem a característica de auto-organizar os seus próprios neurónios no processo de aprendizagem, em que os parâmetros de aprendizagem e a estrutura da rede são automaticamente atualizados, ajudando a alcançar uma melhor precisão na previsão. Este é estruturado em cinco camadas: camada de entrada, camada de função base elipsoidal (EBF), camada normalizada, camada ponderada, e camada de saída.

Para a realização do agrupamento de perfis de consumo em vários segmentos, foi utilizado o método *fuzzy C-means* (FCM). Estes segmentos foram utilizados pelo SOFNN, em que em cada segmento se efetuou a respetiva previsão de carga. Os segmentos gerados eram distintos entre eles em termos de estrutura de rede neuronal bem como nos seus parâmetros. O método *K-means* / FCM foi utilizado para a determinação do número de grupos.

Os resultados experimentais desta metodologia, demonstraram que os consumidores domésticos analisados foram classificados em cinco grupos com padrões de consumo distintos tendo em conta o IBT descrito anteriormente: consumidores com baixo consumo de energia elétrica (abaixo dos 10 kWh/dia e de um modo estável) e insensibilidade a altas temperaturas *(Cluster 1)*, consumidores comuns (consomem mais energia elétrica na época de Verão) e sensibilidade a altas temperaturas *(Cluster 2)*, consumidores comuns e sensibilidade à IBT *(Cluster 3)*, consumidores com alto consumo de energia elétrica e sensibilidade as altas temperaturas *(Cluster 4)* e consumidores de luxo (consomem um nível mais alto de energia elétrica, sendo esta gradual ao logo do dia) *(Cluster 5)*.

Também ficou demonstrado que este esquema pode alcançar uma melhor precisão de previsão de carga e flexibilidade usando um modelo híbrido.

Do ponto de vista de aplicação prática, o modelo de previsão proposto pode ajudar as empresas de energia elétrica a fornecerem eletricidade de uma forma confiável, sem interrupções e de melhor qualidade bem como para estabelecer cronogramas apropriados de operação e manutenção dentro de uma determinada área.

Além disso, os comportamentos de consumo identificados podem ser analisados e usados para melhorar o design e promover a conscientização/ aceitação da IBT proposta.

Um exemplo de aplicação de *clustering* na previsão de preços de eletricidade, é a abordagem efetuada no estudo [5]. Este estudo teve como objetivo a aplicação de redes neuronais ou modelos híbridos (redes

neuronais e algoritmos de *clustering)* para previsão do preço de eletricidade para o dia seguinte, neste caso, para a Itália. Para isto, os dados foram agrupados em grupos homogéneos e a cada grupo foi aplicado um tipo de modelo de previsão.

Os modelos propostos são caracterizados por uma grande flexibilidade, com diferentes tipos de dados de entrada nas redes neuronais. Os modelos foram testados em conjuntos de dados que incluem padrões de preços atípicos e muitos *outliers*.

De acordo com outros estudos anteriores, o preço da eletricidade é tipicamente uma função não linear com várias variáveis a serem consideradas, como por exemplo, o histórico de preço, a previsão de carga, a capacidade de geração entre outras. Com base nisto, foram incluídos neste estudo dados como por exemplo, o histórico de carga a satisfazer bem como a sua previsão para o dia seguinte. Dados como a temperatura ou outros dados climáticos, não foram considerados.

Para este estudo foram criados seis modelos de previsão, constituídos por dados com várias características para servirem como dados de entrada de redes neuronais, obtendo os seus resultados separadamente e por fim, efetuar a sua comparação por forma a identificar qual ou quais os modelos mais fiáveis e robustos a utilizar para a previsão de preços para o dia seguinte. Abaixo descrevem-se os seis modelos propostos:

Modelo A:

Modelo de referência para comparação com os restantes modelos propostos. Tem como dados de entrada, dias da semana, fim de semana ou feriados, com os preços horários da última semana incluindo o dia anterior.

Modelo B:

Tem como dados de entrada, dados com as características iguais ao Modelo A e com os preços horários das últimas 168 horas.

Modelo C:

Considera variáveis externas como as cargas a satisfazer, dados de geração de energia renovável (eólica mais solar) e preços de gás natural com as seguintes características: iguais ao Modelo B mais o valor médio de carga do último dia, valor de carga horária da última semana, preço médio da energia do último dia, preço de gás do dia anterior, previsão horária da geração de energia renovável para o dia específico de previsão e preço do gás natural para o dia específico de previsão.

Modelo D:

Utiliza dados dados com características iguais ao Modelo C, adicionando os preços previstos para os seguintes países: Grécia, França, Alemanha e Suíça.

Modelo E:

Constituído por duas redes neuronais em cascata, em que os dados de entrada para a primeira rede neuronal serão semelhantes aos utilizados no Modelo D. Os resultados de preço horário desta primeira rede neuronal servirão de dados de entrada para a segunda rede neuronal mais os dados de entrada semelhantes ao Modelo D.

A segunda rede neuronal tem como objetivo normalizar os preços resultantes da análise da primeira rede neuronal, sendo os resultados desta, os finais deste modelo.

Modelo F:

Modelo híbrido composto por dois estágios: o primeiro estágio responsável por pré-processar dados semelhantes ao Modelo D e efetuar o agrupamento de dados utilizando o *K-means* e a distância euclidiana, com o objetivo de agrupamentos de preços tendo em conta a sua semelhança de perfil de curva de consumo e tendência. O segundo estágio será responsável por processar os dados vindos do primeiro estágio, por *cluster*, numa rede neuronal para obtenção da previsão dos preços.

Após testar os modelos descritos anteriormente e comparando-os, verificou-se que ao usar dados não pré-processados, o erro de previsão é aproximadamente de 20%. Com isto pode afirmar-se que este tipo de modelo requer mais desenvolvimento para ser aperfeiçoado já que tem uma percentagem de erro grande.

Ao comparar os vários modelos, os resultados demonstraram que o pré-processamento de dados através de modelos de dois estágios em cascata (Modelo F), será o melhor modelo para a previsão de preços para o dia seguinte. Também ficou demonstrado que os dados que serviram de dados de entrada para tratamento de redes neuronais deveriam incluir preços de outros mercados, preços de geração de energia renovável e capacidade de geração renovável de modo a obter resultados mais robustos e fiáveis, e não conterem apenas o histórico de preços de uma determinada área ou país.

Quanto aos modelos híbridos aplicados a modelos em estágios, estes não demonstraram ser uma mais-valia para o objetivo deste estudo. Apesar disto, demonstraram ter um baixo valor de erro de previsão comparando com outros modelos.

Com base neste estudo, os modelos com redes neuronais na previsão de preços para o dia seguinte evidenciaram ser ferramentas fiáveis e robustas para todos intervenientes no mercado energético, sendo que estes ainda requerem aperfeiçoamento ao longo dos tempos perante novos comportamentos de consumo de energia elétrica por parte dos consumidores.

Com a introdução de contadores inteligentes em redes elétricas, atualmente, temos acesso a informação que dantes não era possível. Devido ao grande volume de dados obtido, tornou-se necessário desenvolver técnicas que nos permitissem analisar estes dados e apresentar seus resultados a partes interessadas tais como operadores de rede, distribuidores e comercializadores de energia elétrica. Para a análise destes dados, o *clustering* tornou-se uma das técnicas mais conhecidas com o objetivo de descobrir quais os perfis de consumo existentes numa rede elétrica.

Dentro desta temática, existe o exemplo do artigo [6], em que se propôs estudar várias técnicas de *clustering*, por forma a obter características de perfis de consumidores (diagrama de carga), produção de energia eólica e preço de eletricidade.

Para obtenção dos perfis desejados foram utilizadas técnicas de análise através do método hierárquico e rede neuronal. Os dados utilizados no estudo foram provenientes do consumo doméstico diário de uma pequena cidade portuguesa, da produção de energia eólica e do preço da eletricidade do MIBEL durante 1 ano.

Para efetuar o *clustering* foram usados dois tipos de dados:

- Para o consumo de energia diário e geração de energia eólica foram usados dados divididos por quartos de hora durante um ano.
- Para o preço de energia foram usados dados divididos por hora durante um ano.

De modo a encontrar o número de *clusters*, dividiram-se os dados em grupos distintos segundo características comuns como por exemplo: o tipo de dias que os compõem, se são dias de semana, fim de semana ou férias; e o seu consumo. Foram usados dados de teste e os seus resultados foram comparados entre os métodos hierárquico e rede neuronal.

Comparando os dois métodos, verificou-se que usaram dois processos diferentes e que os elementos dos seus *clusters* também foram diferentes. No entanto, os dois métodos determinaram que 5 é o número ideal de *clusters*, determinado pela distância dos seus elementos, usando a distância euclidiana.

Cada *cluster* identificado representou um perfil de consumo diferente. Para cada método, também se efetuou o estudo do perfil de carga de cada cluster obtido.

Este estudo deu a conhecer dois métodos que permitem uma forma de projetar uma *smartgrid* em termos de gestão (preço de eletricidade) e armazenamento para despacho de energia analisando os dados de consumo

de determinado local, identificando os perfis de consumo dos respetivos consumidores.

Dentro do estudo de técnicas de *clustering* usadas para identificação de padrões de consumo numa rede elétrica, o artigo [7] teve como objetivo a comparação de várias técnicas conhecidas de *clustering:*

- *K-center Clustering*

  Técnica que inclui técnicas como *K-means, K-medians, K-medoids, FCM* que se baseiam em medição de dissemelhanças entre perfis de consumo baseado em distâncias entre perfis, usando a distância euclidiana, Manhattan e outras.

- *Clustering* hierárquico

  Habitualmente este método é mais flexível e simples que os métodos *k-center*, mas tem um custo de processamento computacional maior.

  Através deste método é criada uma árvore hierárquica sobre os dados a serem analisados.

- Self-organizing map (SOM)

  É uma rede neuronal não supervisionada em que resulta uma representação gráfica que permite uma análise de dados mais fácil e um agrupamento de dados através da sua observação.

- Modelos Probabilísticos (GGM)

  Este método usa a distribuição Gaussiana, onde várias componentes são várias distribuições Gaussianas com as suas médias e variâncias.

Para efetuar as comparações entre os métodos, teve-se em conta os seguintes indicadores:

- Erro médio quadrático (MSE)
- Índice *Silhouette* (SIL)
- Índice Davies-Bouldin (DBI)
- Adequação do índice médio (MIA)
- Índice de Dunn
- Razão da soma dos quadrados dentro do cluster para a variação entre *clusters* (WCBCR)

De modo a determinar:

✓ o número adequado de *clusters*

✓ o desempenho

✓ o efeito dos parâmetros do método nos resultados de agrupamento

✓ o desempenho do agrupamento quando alguns atributos são adicionados ou removidos

Após análise dos resultados das várias técnicas de *clustering* referidas anteriormente, este estudo revelou que o *clustering* é uma técnica importante para otimizar a rede de distribuição de energia, detetar falhas, tipificar perfis de consumo associados, perceber comportamentos e tendências dos consumidores, ajustar tarifas de modo a serem atrativas para os consumidores e impedir muitos picos de carga na rede.

Também demonstrou que técnicas que utilizam séries temporais normalmente usam a distância *Dynamic Time Warping* (DTW), que pode ser benéfico, visto que este tipo de distância não é sensível à escala temporal ou mudanças temporais, tal como acontece com a distância euclidiana ou a distância de *Minkowski*.

Demonstrou igualmente que técnicas baseadas em séries temporais têm um custo computacional grande comparativamente com outras técnicas.

Quanto às tendências para aplicações futuras dos métodos estudados, os autores deste estudo referem que:

- Técnicas de *clustering* como *K-means* ou FCM podem ser utilizadas para efetuar *clustering* em tempo real.
- Técnicas de *clustering* baseadas em redes neuronais mostraram ser fiáveis em tempo real com vantagem sobre outras técnicas: não foi necessário saber nada sobre quais os dados a processar nem foi preciso utilizar métricas de medição de distância como a euclidiana.
- Técnicas baseadas em modelos probabilísticos mostraram ser úteis para aplicação no fornecimento de energia numa rede elétrica.

- Técnicas baseadas em redes neuronais em conjunto com o *algoritmo K-shape* mostraram ser úteis para previsão de carga elétrica.
- Técnicas baseadas em duas camadas de *clustering* mostraram ser úteis para previsão diária de carga e sua variação ao longo do dia, e na identificação de padrões sociodemográficos através dos respetivos dados de consumo elétrico.

Embora a maior parte das técnicas de *clustering* utilizem dados off-line como seus dados de entrada, este estudo reforça como melhoria a importância de se apostar em técnicas que utilizem dados em tempo real de modo que se possa efetuar uma previsão de carga a curto prazo ou projetar uma rede elétrica com a capacidade de fornecimento de energia mais dinâmico.

Noutro artigo, [8], a aplicação do *clustering* teve como objetivo identificar e comparar numa *smartgrid*, perfis de consumo que usam energia fotovoltaica (FV) dos que não usam, já que os perfis que usam energia fotovoltaica têm como característica no perfil de consumo apresentarem pequenas distorções na sua curva.

Neste estudo, foram considerados dados de consumo sem contar com o fim de semana, já que aos fins de semana o comportamento de consumo dos utilizadores altera consideravelmente face aos dias de semana.

Para efetuar a análise de perfis de consumo numa rede de distribuição tendo consumidores com e sem recurso a energia solar, neste estudo foram propostas duas técnicas de agrupamento de perfis que usam o Modelo de Mistura de Consumidor (CMM): a primeira técnica usa ponderação de combinação linear para atribuir aos clientes-padrão valores maiores, enquanto a segunda técnica realiza a desagregação de dados de consumos de energia solar e efetua o agrupamento destes dados. O CMM modela os padrões de carga de clientes com FV (clientes FV) como uma combinação dos padrões de carga de clientes sem FV (clientes não-FV).

De modo a realizar este estudo, assumiu-se que:

1. O conjunto de dados contém uma mistura de clientes FV e não-FV.
2. Os clientes residem na mesma região.
3. Um pequeno número de clientes FV tem os seus sistemas solares de geração e consumo medidos separadamente como referência para a metodologia.
4. Os dados são referentes a clientes que têm apenas dispositivos de medição de consumo elétrico através da rede pública e de sistemas FV. Para simplificar, não foram consideradas outras formas de geração de energia.

A primeira etapa da metodologia foi realizar o agrupamento em duas camadas. A primeira camada de agrupamento define o número de *clusters*, que neste caso é dois, onde o objetivo é separar os clientes em clientes FV e clientes não-FV.

A segunda camada de agrupamento é efetuada no cluster de clientes não-FV, onde o objetivo é encontrar os principais padrões de consumo deste tipo de cliente. Para o agrupamento de perfis foi usado o *k-medoids* em conjunto com *o DTW* como medida de distância. O melhor número de *clusters* foi escolhido com base no SIL, enquanto o medoide de cada *cluster* foi obtido pelo algoritmo *DTW Barycenter Averaging*.

Os resultados deste estudo mostraram que o uso de medidas desagregadas para obter o perfil de consumo melhoraram a qualidade do *cluster* usando o SIL, enquanto algoritmos mais comuns como a distância euclidiana, separaria consumidores FV e não-FV com padrões semelhantes de consumo. Esta metodologia demonstrou ser capaz de agrupar consumidores FV e não-FV com perfis semelhantes nos mesmos *clusters*. Também ficou demostrado que ao efetuar o agrupamento usando a ponderação em relação ao consumidor, os resultados foram piores ao refazer o agrupamento em medições desagregadas. O CMM proposto não refina as definições de *cluster* após a atribuição de consumidores FV para *clusters* existentes, mostrando resultados insatisfatórios. O proposto como melhoria seria desenvolver um algoritmo CMM que iterativamente refine a definição de *cluster* para

desagregação de dados por forma a fornecer melhores resultados de agrupamento e desagregação.

Um outro exemplo de análise de perfis de consumo, é o descrito no artigo [9] em que se propôs segmentar perfis de consumo entre 2008 e 2009 através de uma amostra de consumidores residenciais espanhóis.

Neste estudo, foi aplicada uma técnica de *clustering* dinâmica numa amostra de dados horária referente a 759 consumidores residenciais, que teve como objetivo a análise de duas vertentes: por um lado, classificar os consumidores consoante o seu perfil de consumo tendo em conta também as leis do mercado de energia elétrica espanhol e por outro mostrar a utilidade de uma ferramenta de análise de perfis de consumo por forma a ajudar a segmentar perfis de consumo de clientes, avaliar falhas no consumo de energia, avaliar tendências de consumo e ações de gestão da procura (DSM).

Para o desenvolvimento do algoritmo de *clustering* foi utilizado o *K-means* devido à sua robustez e eficiência. Para a determinação de uma solução inicial os medoides de cada grupo podem ser obtidos utilizando métodos heurísticos, o conhecimento prévio dos dados ou utilizando o algoritmo *K-means++*.

Foi utilizada a distância euclidiana para determinar os grupos, em que os elementos de cada grupo são os que têm a menor distância entre eles.

Dependendo do objetivo, as opções para o nível de granularidade (detalhe) do *clustering* deste estudo tiveram em conta os seguintes critérios:

- *n* perfis de carga diários dos consumidores
- Perfis de carga acumulado ou a sua média mensal, para evitar flutuações de consumo semanais devido aos dias de trabalho, normalmente dias de semana.
- *n* níveis de carga diários ordenados por tipo de dia para evitar flutuações de consumo semanais devido aos dias de trabalho.

Os resultados mostraram que a metodologia apresentada é uma ferramenta eficiente para classificação dos clientes consoante o seu consumo de energia elétrica e tendências. Além disto, a deteção de fraude também pode ser outra possibilidade para utilização desta metodologia: um rápido agrupamento de consumidores em padrões de consumo permitirá a deteção de grupos específicos de clientes cujo nível e perfil de consumo de energia podem não se adequar em termos de sua potência contratada e tarifas aplicadas. Também demonstraram ser uteis para serem representativos do tipo de clientes residenciais em Espanha, se não houver outras variáveis além das que foram levadas em consideração. Se se tiver em conta variáveis como por exemplo, as diferentes regiões climáticas de Espanha e/ou a idoneidade dos clientes analisados, deve ser abordada adequadamente a metodologia.
Esta metodologia demonstrou também que a previsão de carga pode ser combinada com o agrupamento dinâmico de modo a fornecer estimativas úteis de padrões de consumo a médio prazo.

A partir da análise dos resultados desta metodologia, um especialista ou operador deve poder identificar e classificar possíveis clientes. Esta decisão poderá ser apoiada por sistemas de apoio à decisão e uma análise automatizada dos resultados de *clusters*, realizando avaliação de tendências, deteção de anomalias nos comportamentos de consumo e sugerir automaticamente a grupos de clientes, ações específicas como ofertas comerciais ou pedidos de redução de energia.

Capítulo 3 – Metodologia

O presente capítulo descreve a metodologia utilizada no trabalho apresentado, explicando quais as ferramentas estatísticas utilizadas para analisar os dados do estudo, quais as medidas de proximidade e método de *clustering* utilizados bem como a descrição da linguagem informática a que se recorreu para efetuar o estudo, o R.

3 Metodologia

Atualmente, devido ao grande volume de dados disponível sobre consumos de energia elétrica, é muito importante analisar os perfis dos consumidores e perceber se estes demonstram algum padrão de consumo ou tendência.

A interpretação de resultados de uma análise de dados, pode também permitir mais facilmente a tomada de decisão em termos de implementação de novas metodologias e/ou identificar/corrigir falhas existentes.

Em seguida são apresentadas algumas ferramentas utilizadas neste estudo para análise de dados.

3.1 Medidas descritivas

As medidas descritivas são utilizadas para estudar uma distribuição de frequências. Estas podem ser classificadas consoante [10]:

- **Localização:** localizam valores observados numa distribuição.
  Exemplos: média, mediana, quartis, extremos
- **Dispersão:** medem o grau de dispersão dos dados.

  Exemplos: amplitude interquartil, desvio padrão e variância, coeficiente de variação
- **Assimetria:** medem o grau de simetria da distribuição.

  Exemplos: coeficiente de assimetria
- **Achatamento:** Medem o grau de achatamento da distribuição

  Exemplos: coeficiente de *kurtosis*, compara o achatamento da distribuição face à curva de *Gauss*

3.2 Histograma

Um histograma é um gráfico representativo da distribuição de frequência, Figura 3.1, onde são representados o número de observações de cada um dos intervalos, no caso deste estudo, são representados o consumo nas abcissas em kWh e o número de ocorrências nas ordenadas.

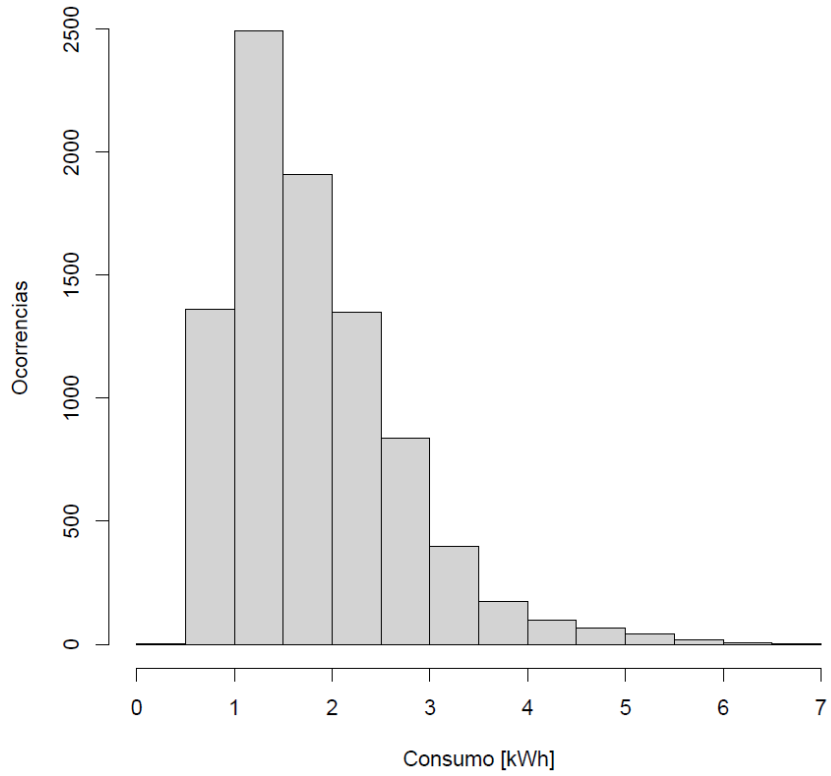


Figura 3.1 - Exemplo de histograma de consumo

3.3 *Boxplot* e identificação de *Outliers*

O *boxplot*, diagrama de caixa ou de caixa com bigodes é uma ferramenta gráfica que permite visualizar a distribuição de frequências e representar algumas medidas de estatística descritiva como o mínimo, máximo, primeiro quartil, segundo quartil ou mediana e o terceiro quartil.

Observando a Figura 3.2, verifica-se que o local onde o eixo horizontal começa junto à *boxplot* (da esquerda para a direita) indica o mínimo (excetuando algum possível *outlier*) e onde o eixo termina indica o máximo (também excetuando algum possível *outlier*). O IQ representa o intervalo interquartílico, o $Q_{1/4}$ representa o 1º quartil e o $Q_{3/4}$ representa o 3º quartil.

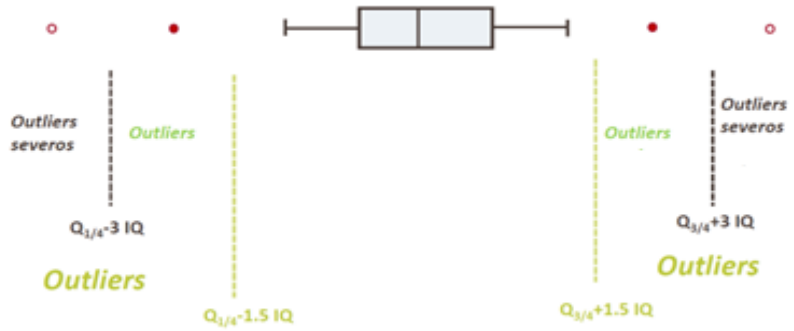


Figura 3.2 – Bloxplot e Outliers [11]

O *boxplot* também permite uma análise visual da simetria, caudas e *outliers* de um conjunto de dados:

- **Localização central** – A linha central da "caixa" representa a mediana.
- **Dispersão** – A dispersão dos dados pode ser representada pelo intervalo interquartílico (IQ) que é a diferença entre o terceiro quartil e o primeiro quartil (tamanho da caixa), ou ainda pela amplitude que é calculada pela diferença entre o valor máximo e o valor mínimo.

- **Simetria** – Um conjunto de dados que tem uma distribuição simétrica, terá a linha da mediana no centro do retângulo e os "bigodes" com uma dimensão igual. Quando a linha da mediana está próxima ao primeiro quartil, os dados são assimétricos positivos e quando a posição da linha da mediana é próxima ao terceiro quartil, os dados são assimétricos negativos. De ressalvar que a mediana é a medida de tendência central mais indicada quando os dados possuem distribuição assimétrica, uma vez que a média aritmética é influenciada pelos *outliers*.
- **Caudas ou bigodes** – As linhas que vão do retângulo até aos *outliers* podem fornecer o comprimento das caudas da distribuição.
- ***Outliers*** – indicam possíveis valores atípicos, no caso deste estudo, são valores atípicos de consumo de energia elétrica e gás. São identificados num *boxplot*, por se localizarem à esquerda ou à direita do limite de deteção de *outliers*, Figura 3.2.

O limite de deteção de *outliers* é construído utilizando o intervalo interquartílico, assim os limites inferior e superior de deteção de *outlier* são definidos por:

$$Limite\ \text{à}\ esquerda = Primeiro\ Quartil - 1{,}5 * (Terceiro\ Quartil - Primeiro\ Quartil) \quad (3.1)$$

$$Limite\ \text{à}\ direita = Terceiro\ Quartil + 1{,}5 * (Terceiro\ Quartil - Primeiro\ Quartil) \quad (3.2)$$

Para os *outliers* severos, considera-se o valor de 3 nas equações (3.1) e (3.2) em vez de 1,5, conforme indicado na Figura 3.2.

3.4 Séries Temporais

Seja um conjunto de observações (dados) definidas por x_1, x_2,…, x_T feitas nos períodos $1,2,\ldots,T$ contados a partir de uma determinada origem. Considerando as observações espaçadas no tempo, uma série temporal pode ser definida por:

$$x_t,\ t = 1,2,\ldots,T \tag{3.3}$$

Uma série temporal pode ser representada de uma forma gráfica, o cronograma, em que o eixo das abcissas representa o tempo e o eixo das ordenadas representa os valores da série.

As séries temporais existem em várias áreas tais como: finanças, marketing, economia, seguros, demografia, ciências sociais, meteorologia, energia, epidemiologia, entre outras [12].

O valor médio, μ, e a variância, σ^2, da série podem ser estimados a partir das observações, $x_1, x_2, \ldots, x_T$:

$$\hat{\mu} = \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{3.4}$$

$$\hat{\sigma}^2 = s^2 = \widehat{\gamma_0} = \frac{1}{T-1}\sum_{t=1}^{T} (x_t - \bar{x})^2 \tag{3.5}$$

3.4.1 Função de autocorrelação (ACF)

É frequentemente utilizada para interpretar como valores presentes se relacionam com valores passados. O valor da função de autocorrelação, ρ_k, é uma medida da correlação entre as observações de uma série temporal X, que são separadas por k unidades de tempo (x_t e x_{t-k}).

Estimando a autocovariância e a autocorrelação ,γ_k e ρ_k respetivamente, a partir das observações, $x_1, x_2, \dots, x_T$:

$$\hat{\gamma}_k = \frac{1}{T-1}\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x}) \tag{3.6}$$

$$\hat{\rho}_k = \hat{\rho}(x_t, x_{t-k}) = r_k = \frac{\widehat{\gamma_k}}{\hat{\gamma}_0} = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{3.7}$$

Onde,

$$-1 \leq \widehat{\rho_k} \leq 1\,,\, \widehat{\rho_0} = 1 \; e \; \widehat{\rho_k} = \widehat{\rho_{-k}} \tag{3.8}$$

Em geral, à medida que k aumenta, a capacidade de memória do processo diminui e consequentemente a autocorrelação vai igualmente diminuindo. Assim, a memória do processo caracteriza-se pela forma como ρ_k decai.

O gráfico das estimativas da ACF para diferentes valores de k é designado de correlograma, apresentado na Figura 3.3.

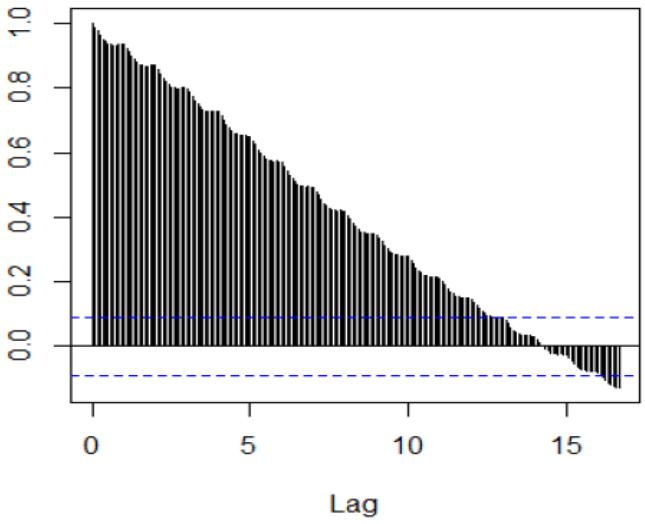


Figura 3.3 - Exemplo de um correlograma [13]

3.4.2 Periodograma - Identificação de periocidades em séries temporais

O periodograma é uma função que estabelece a forma como a variabilidade total da série é particionada ao longo das várias componentes relativas a cada uma das frequências de Fourier [14].

Considerando a série temporal x_t com T observações, em que $x_1, x_2, \dots, x_T$ seguem o modelo:

$$x_t = \sum_{k=1}^{[T/2]} \left(a_k cos\left(\omega_k t\right) + b_k sen\left(\omega_k t\right)\right) + \varepsilon_t \qquad (3.9)$$

Sendo $t = 1,2, \dots, T$ e ω_k as frequências de Fourier ($\omega_k = \frac{2\pi k}{T}$ e $k = 1, \dots, \left[\frac{T}{2}\right]$, onde $\left[\frac{T}{2}\right]$ é o maior inteiro menor ou igual a $\frac{T}{2}$ [19],[20] e admitindo que ε_t *são* variáveis aleatórias não correlacionadas e tais que $\mathrm{E}(\varepsilon_t) = 0, \mathrm{Var}\,(\varepsilon_t) = \sigma^2$.

As estimativas dos valores a_k e b_k podem ser obtidos através da minimização de:

$$SQ = \sum_{t=1}^{T} \left(x_t - \sum_{k=1}^{[T/2]} \left(a_k cos\left(\omega_k t\right) + b_k sen\left(\omega_k t\right)\right) \right)^2 \qquad (3.10)$$

Assim, a função de um periodograma é a indicada na equação (3.11) [15].

$$I(\omega_k) = \begin{cases} TA_0^2 & k = 0 \\ \frac{T}{2}(A_k^2 + B_k^2) & k = 1, \dots, [\frac{T-1}{2}], \\ TA_{\frac{N}{2}}^2 & k = \frac{T}{2}(T\ par), \end{cases} \qquad (3.11)$$

Em cada frequência ω_k, existe uma ordenada $I(\omega_k)$ que procura estimar a contribuição dessa frequência para a série, em que A_k e B_k são os estimadores dos mínimos quadrados.

Para grandes valores das ordenadas do periodograma correspondem frequências que estão na série, enquanto para valores pequenos correspondem frequências que não estão na série. Nesta situação, o que se procura são as frequências que originam os picos num periodograma [14], [15]. Um exemplo de um periodograma é apresentado na Figura 3.4.

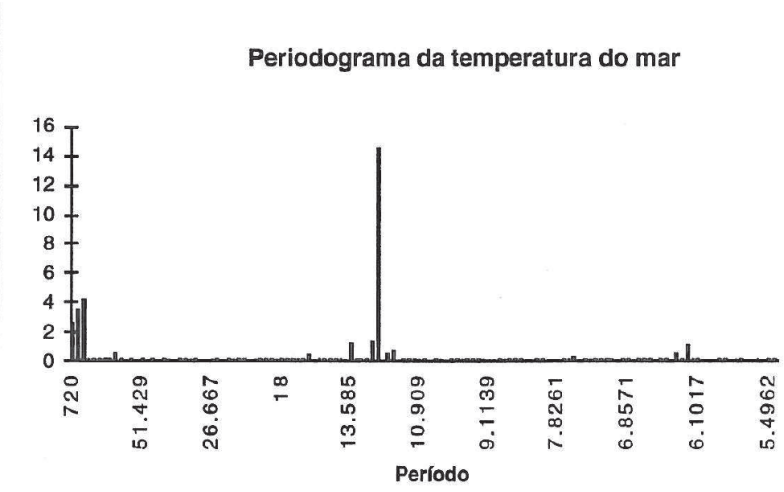


Figura 3.4 – Exemplo de um periodograma [15]

3.5 Clustering

Citando o artigo [16], uma forma para avaliar uma característica é efetuá-la por meio da observação, atribuindo um julgamento, tendo o "*cluster*" como valor. Muitos autores de artigos sobre o tema, apenas definem um *cluster* em termos de coesão interna (homogeneidade) e externa (isolamento – separação). Tendo em conta este conceito, o *clustering* é efetuado no nosso dia a dia, através de uma interpretação de indivíduos segundo certas características, em que se define uma distância entre eles, ou seja, indivíduos semelhantes têm uma distância menor e indivíduos diferentes têm uma distância maior. Assim surge o *clustering*, quando se tenta classificar ou organizar indivíduos em grupos coerentes, que através de uma função de distância, os indivíduos são divididos em grupos para que, intuitivamente, os indivíduos dentro do mesmo grupo estejam "próximos" e os que pertencem a diferentes grupos estejam "distantes".

Exemplo de um *clustering* efetuado de forma clara é a Figura 3.5.

Figura 3.5 – Clusters com coesão interna e / ou solução externa [16]

Mas nem sempre é fácil efetuar de uma forma clara ou nem existe maneira de efetuar o *clustering*, tendo como exemplo a Figura 3.6.

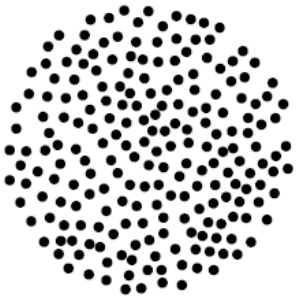


Figura 3.6 – Dados que não contêm *clusters* "naturais" [16]

O *clustering* de indivíduos é aplicado atualmente em diversas áreas, tais como, biologia, botânica, medicina, psicologia, geografia, marketing, processamento de imagem, psiquiatria, arqueologia, energia, entre outras [16].

Para a realização do *clustering* é necessário:

a. Definir a medida de proximidade (dissemelhança ou distância) entre indivíduos a agrupar.
b. Definir o método de agrupamento.

3.6 Medidas de proximidade

As medidas de proximidade são usadas para efetuar o agrupamento de dados de modo a formar grupos de dados e seus respetivos elementos [16]. Nesta dissertação, os elementos representam os perfis dos consumos anuais de eletricidade e gás nas cidades analisadas e a distância entre os consumos das cidades vai ser calculada através da média das seguintes distâncias utilizadas: distância euclidiana, distância baseada no coeficiente de correlação de *Pearson*, distância entre ACF e distância euclidiana entre periodogramas.

3.6.1 Distância Euclidiana

A distância euclidiana é a medida mais utilizada em processos de agrupamento, sendo definida pela expressão (3.12), considerando duas séries temporais x_t e y_t de dimensão *T*.

$$d_{euc}(x_t, y_t) = \sqrt{\sum_{i=1}^{T} (x_i - y_i)^2} \qquad (3.12)$$

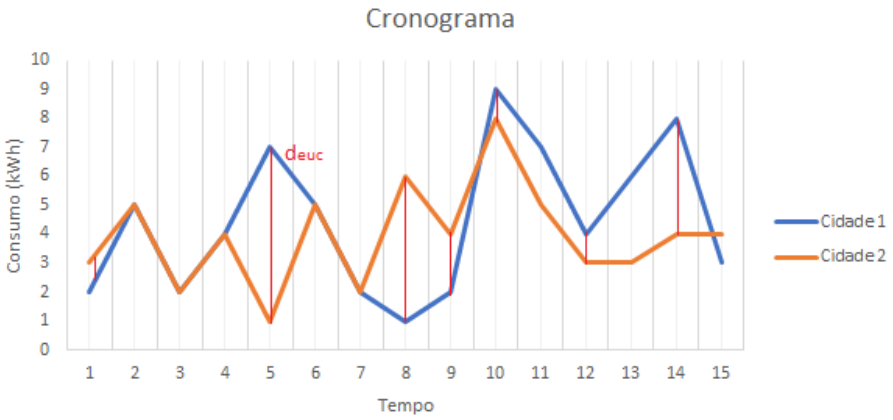


Figura 3.7 – Exemplo da distância euclidiana entre duas séries temporais

A distância euclidiana quantifica as diferenças de valor entre as duas series para cada instante.

3.6.2 Distância de Correlação de *Pearson*

A distância de correlação de *Pearson* é utilizada para medir o grau de relação linear entre duas séries temporais x_t e y_t (t=1,…,T) [17], segundo as expressões (3.13) e (3.14):

$$d_{cor}(x_t, y_t) = \sqrt{1 - r} \qquad (3.13)$$

onde r representa o coeficiente de correlação e $\bar{x}$, $\bar{y}$ representam o valor médio de cada perfil de consumo.

$$r = \frac{\sum_{i=1}^{T} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{T} (x_i - \bar{x})^2 \sum_{i=1}^{T} (y_i - \bar{y})^2}} \qquad (3.14)$$

3.6.3 Distância entre funções de autocorrelação (ACF)

No caso deste estudo, a autocorrelação de um dado perfil de consumo mostra como o consumo atual de energia elétrica é influenciado pelo consumo das horas anteriores.

Esta distância entre duas séries x_t e y_t (t=1,…,T), expressão (3.15), neste caso entre as autocorrelações das séries de consumo, utiliza a distância euclidiana entre as duas funções de autocorrelação, sendo r_l os respetivos coeficientes de autocorrelação para o desfasamento l [18] , [19]:

$$d_{Autocorr\,(x_t, y_t)} = \sqrt{\sum_{l=1}^{L} \left(r_l(x_t) - r_l(y_t)\right)^2} \qquad (3.15)$$

3.6.4 Distância euclidiana entre periodogramas

Neste estudo, será ainda considerada a distância entre duas series temporais definida como distância euclidiana entre os periodogramas associados às duas series temporais, x_t e $y_t (t = 1, \dots, T)$.

Sendo $I_x(w_k)$ e $I_y(w_k)$, os periodogramas das séries temporais x_t e y_t respetivamente, nas frequências $w_k = \frac{2\pi k}{n}$, $k = 1, \dots, \left[\frac{T}{2}\right]$, onde $\left[\frac{T}{2}\right]$ é o maior inteiro menor ou igual a $\frac{T}{2}$ [19],[20] .

A distância entre as séries x_t e y_t (t=1,…,T), é definida por:

$$d_P(x_t, y_t) = \sqrt{\sum_{k=1}^{[T/2]} \left[I_x(w_k) - I_y(w_k) \right]^2} \qquad (3.16)$$

3.6.5 Normalização

De modo a combinar as quatro distâncias utilizadas, recorreu-se a uma normalização através do método min-max [18]:

$$norm(d(x_t, y_t)) = \frac{d(x_t, y_t) - min\{d(x_t, y_t)\}}{max\{d(x_t, y_t)\} - min\{d(x_t, y_t)\}} \qquad (3.17)$$

3.7 Método de agrupamento

O *clustering* tem como objetivo o estudo de semelhanças entre dados, dos quais se formam grupos. Os grupos são formados de forma a maximizar as semelhanças entre elementos de um grupo (intragrupo) e minimizar as semelhanças entre elementos de vários grupos (entre grupos).

Os algoritmos que executam tarefas de análise de dados, muitas vezes, usam alguma(s) medida(s) de semelhança entre observações (dados) no seu processo de execução. Essas medidas servem para guiar o processo de construção da superfície de decisão que determina qual a região de abrangência de cada grupo de dados.

Exemplos de técnicas de agrupamento são: *k-medoids*, *k-means*, *k-shape*, *k-medians* entre outros [17].

3.7.1 Método de agrupamento *k-medoids*

Neste método, os *clusters* são representados por um dos pontos de dados no *cluster*, que são designados por medoides. Cada medoide sintetiza as informações do *cluster* e representa as características tipificadas dos *clusters*

e, em seguida, sintetiza as características dos elementos pertencentes a cada *cluster*. Aplicando o método de agrupamento *k-medoids*, minimiza-se a função objetivo representada pela soma (ou matematicamente equivalente, a média) da diferença de elementos para os seus elementos representativos mais próximos. O método de agrupamento *k-medoids* primeiro calcula um conjunto de elementos representativos, os medoides. Após encontrar o conjunto de medoides, a cada um dos medoides é atribuído os elementos mais próximos, formando um grupo.

O algoritmo é constituído por duas fases:

1º Fase: Esta é a fase de construção, sendo selecionados sequencialmente os dados "centrais" para serem usados como medoides iniciais.

2º Fase: Esta é a fase de troca de medoides efetuada durante o método até se encontrar os medoides finais. O seu objetivo é efetuar a troca de medoides com dados não medoides de forma continua até que não possa ser mais efetuada. Este método é caracterizado por uma função de minimização conforme a equação (3.18).

$$min: \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}\, d_{ic} \qquad (3.18)$$

$$\sum_{c=1}^{C} u_{ic} = 1, u_{ic} \geq 0, u_{ic} = \{0,1\} \qquad (3.19)$$

onde u_{ic} indica o grau de pertença da i-ésima unidade ao c-ésimo cluster; $u_{ic} = \{0,1\}$, ou seja, $u_{ic} = 1$ quando a i-ésima unidade pertence ao c-ésimo *cluster*.

Caso contrário, d_{ic} indica a distância entre o i-ésimo objeto e o medoide do c-ésimo *cluster* [21].

3.7.2 Avaliação de qualidade do agrupamento: índice *Silhouette* (SIL)

Considerando uma unidade i $(i = 1, \dots, I)$ pertencente ao *cluster* c, $(c = 1, \dots, C)$, pelo algoritmo de agrupamento, isto significa que a i-ésima unidade está mais próxima do centroide do c-ésimo *cluster* do que de qualquer outro medoide. Considere-se a_{ip} como a distância da i-ésima unidade para todas as outras unidades pertencentes ao *cluster* c $(i, p = 1, \dots, I)$ com i≠ p e a distância média dessa unidade para todas as unidades pertencentes a outro *cluster* $q, com\ q \neq c$, ser designada por d_{iq}. Assim, seja b_{ip} o mínimo de d_{iq}, calculado por $q = 1, \dots, C,\ q \neq c$, que representa a dissemelhança da i-ésima unidade pelo seu *cluster* vizinho mais próximo. Então, o índice *Silhouette* do i-ésimo objeto é definido do seguinte modo:

$$S_i = \frac{b_{ip} - a_{ip}}{max\{a_{ip}, b_{ip}\}} \qquad (3.20)$$

onde o denominador é um termo de normalização. Evidentemente, quanto maior o valor de S_i , melhor será a atribuição da i-ésima unidade ao c-ésimo *cluster*.

Assim, a avaliação de qualidade do agrupamento pode ser obtida através do índice *Silhouette* que é a média de S_i sobre $i = 1, \dots, \mathrm{I}$:

$$SIL = \frac{1}{I}\sum_{i=1}^{I} S_i \qquad (3.21)$$

A constituição dos *clusters* (medoides e respetivos elementos) é alcançada quando o índice *Silhouette* é o mais próximo possível de 1, o que implica a minimização da distância intracluster (a_{ip}) enquanto maximiza a distância entre *clusters* (b_{ip}) [21]. Neste estudo o índice *Silhouette* vai definir o número ideal de grupos tendo em conta que quanto mais próximo o valor estiver de 1, mais esse número de grupos é apropriado.

3.8 Programa R

O R é uma linguagem de programação orientada a objetos, dinâmica, tipificada para análise e visualização de dados [17].

Esta linguagem é aplicada em diversas áreas, como por exemplo, *data science, machine learning*, estatística computacional entre outras.

Para obtenção dos resultados desta dissertação utilizou-se o R como linguagem de programação para análise de dados por forma a realizar o *clustering* e obter os respetivos resultados. Para sua compilação utilizou-se o IDE RStudio, que é um ambiente de desenvolvimento integrado, dedicado à computação estatística e à geração de gráficos. Inclui editor de sintaxe que oferece suporte à execução de código, bem como ferramentas para efetuar a depuração e gestão do espaço de trabalho.

Capítulo 4 – Segmentação de perfis de consumo

O presente capítulo descreve os dados de entrada utilizados neste estudo, a análise exploratória dos dados de consumo de energia elétrica utilizados e os respetivos resultados dos agrupamentos de consumos efetuados.

4 Segmentação de perfis de consumo

4.1 Descrição dos dados de entrada

Os dados utilizados para aplicação e estudo da metodologia proposta neste trabalho são de 2010 e foram obtidos do Departamento de Energia dos Estados Unidos (DOE) [22]. Este departamento é responsável pela política de energia e segurança nuclear, incluindo o programa de armas nucleares, produção de reatores para a Marinha, conservação de energia, pesquisa no campo energético, administração de resíduos nucleares e produção de energia doméstica.

Estes dados são referentes ao programa do DOE designado "*Building America*", que aplica técnicas de engenharia para acelerar o desenvolvimento e adoção de tecnologias avançadas de energia em edifícios existentes e na construção de edifícios residenciais. Este programa apoia várias equipas na construção de edifícios residenciais avançados em escala comunitária. É aplicado um processo para realizar avaliações de custo e desempenho em relação a cada construtor ou contratante de reconstrução. O objetivo geral é reduzir significativamente o uso de energia com apenas o aumento nominal dos custos iniciais de construção. Os conceitos de eficiência energética incorporados nestas casas são avaliados por meio da realização de iterações sucessivas de projeto, teste, redesenho, incluindo compensações de custo e desempenho. Como resultado, este programa permitirá o desenvolvimento de inovações que podem ser usadas de maneira económica em habitações em escala de produção.

Os dados analisados referentes a consumos de energia elétrica e gás, frequentemente utilizadas para avaliação de consumo energético em edifícios. Estes são estimativas de hora em hora para avaliar os impactos de energia dependentes do tempo de sistemas avançados usados em casas. Efeitos dependentes do tempo como massa térmica, ganho de calor solar e infiltração de ar induzida pelo vento são alguns exemplos que podem ser modelados com precisão apenas usando um modelo que calcule a transferência de calor e a temperatura em intervalos de tempo curtos.

Estes também são necessários para estimar com precisão os picos de carga. Por ter sido especificamente desenvolvido e adaptado para atender às necessidades do programa do DOE, o BEopt (usando DOE-2 ou *EnergyPlus* como mecanismo de simulação) foi utilizada como ferramenta de simulação de consumo energético horário utilizada para novas construções.

No âmbito de avaliação de consumo para as novas construções, teve-se em conta o índice B10 *Benchmark*. Este representa uma casa construída de acordo com o *International Energy Conservation Code* (IECC) de 2009, bem como os padrões federais de eletrodomésticos dos EUA em vigor a partir de 1 de janeiro de 2010, as características de iluminação e diversas cargas elétricas mais comuns em 2010. O B10 *Benchmark* é usado como o ponto de referência para acompanhar o progresso para cumprir as metas plurianuais de economia de energia estabelecidas pelo programa *Building America*. De modo a cumprir estas metas, no âmbito da construção de uma nova habitação, o B10 *Benchmark* tem em conta a avaliação de caraterísticas de tetos, isolamentos térmicos, constituição de paredes, existência ou não de caves, dimensões das divisões, número de piso, exposição solar e dimensões de janelas, tipo de edifícios (unifamiliar ou multifamiliar), condutância térmica.

Em termos de equipamentos da habitação, tem em conta a sua existência ou não, tipo, as características, tipo de alimentação destes (a eletricidade ou gás) e tempo de utilização. Também tem em conta indicadores de referência por equipamento com funções de climatização, aquecimento de água, infiltração e ventilação de ar, iluminação e outros equipamentos como secadoras, frigoríficos, elevadores, televisões, micro-ondas entre outros.

No caso de habitações que foram remodeladas no âmbito do programa, além dos indicadores anteriormente referidos para as novas construções, adicionalmente é avaliado consoante o ano de construção e a presença de fontes renováveis no fornecimento de energia à habitação.

Os dados extraídos para o presente estudo são referentes a um ano de consumos de 936 cidades dos 50 estados dos EUA, no formato csv, de perfis de consumo residencial de energia elétrica e gás.

Como a discretização dos dados é horária e os dados são valores de potência em kW, considera-se que a potência é constante durante 1 hora pelo que os valores horários correspondem a consumos em kWh.

4.2 Análise exploratória dos dados

Como referido anteriormente, os dados analisados para aplicação desta metodologia são referentes a um ano de consumos residenciais horários de energia elétrica e gás, de 936 cidades dos EUA, Figura 4.1, em kWh. Os dados estão em formato csv tendo os respetivos ficheiros um nome que permite identificar a cidade e o respetivo estado [23].

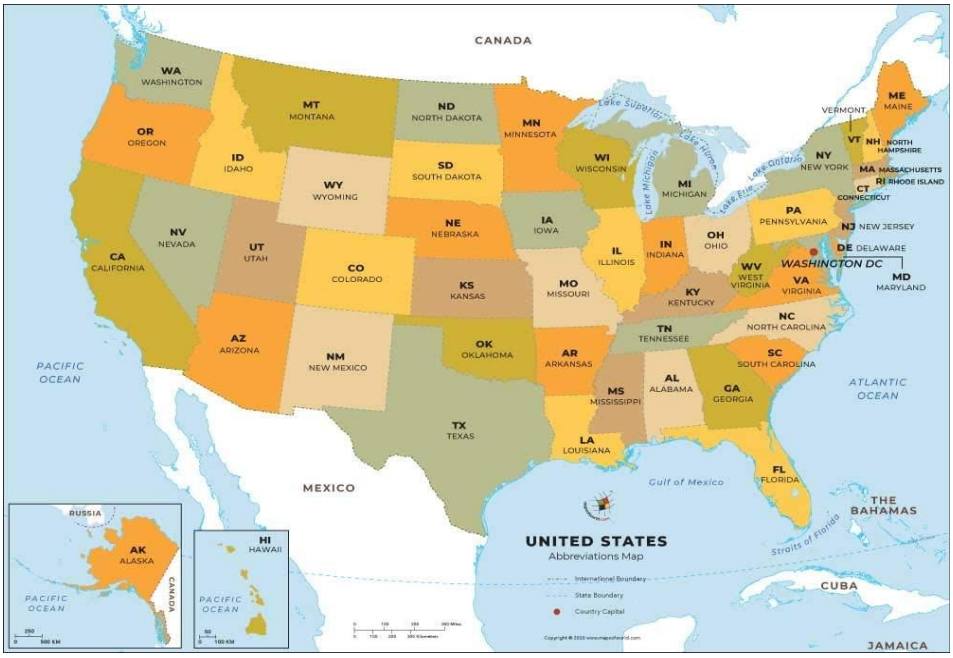


Figura 4.1 – EUA – Área geográfica (fonte: https://store.mapsofworld.com/digital-maps/us-maps-1-2/us-states-abbreviatios-map-1)

Na Tabela 4.1, estão identificadas quantas cidades por estado foram analisadas para este estudo.

Tabela 4.1 – Número de cidades analisadas por estado

Estados	Número total de Cidades	Estados	Número total de Cidades
Alabama (AL)	14	Michigan (MI)	32
Alasca (AK)	2	Minnesota (MN)	54
Arizona (AZ)	18	Mississippi (MS)	13
Arkansas (AR)	18	Missouri (MO)	17
Califórnia (CA)	73	Montana (MT)	16
Carolina do Norte (NC)	21	Nebrasca (NE)	26
Carolina do Sul (SC)	11	Nevada (NV)	10
Colorado (CO)	25	Nova Hampshire (NH)	8
Connecticut (CT)	7	Nova Iorque (NY)	24
Dakota do Norte (ND)	10	Nova Jérsia (NJ)	9
Dakota do Sul (SD)	11	Novo México (NM)	16
Delaware (DE)	2	Ohio (OH)	13
Flórida (FL)	42	Oklahoma (OK)	15
Geórgia (GA)	19	Oregon (OR)	19
Havai (HI)	10	Pensilvânia (PA)	21
Idaho (ID)	13	Rhode Island (RI)	3
Illinois (IL)	19	Tennessee (TN)	8
Indiana (IN)	10	Texas (TX)	61
Iowa (IA)	39	Utah (UT)	13
Kansas (KS)	23	Vermont (VT)	4
Kentucky (KY)	12	Virgínia (VA)	31
Louisiana (LA)	17	Virgínia Ocidental (WV)	11
Maine (ME)	15	Washington (WA)	29
Maryland (MD)	5	Wisconsin (WI)	19
Massachusetts (MA)	15	Wyoming (WY)	13

Conforme se vai validar mais à frente no subcapítulo 4.3.2, existem dois diferentes perfis típicos de consumo presentes nos dados analisados. Estes são representados pelas seguintes cidades, os medoides dos dois grupos: Walla Walla do estado de Washington (WA) e a Lake Charles do estado de Louisiana (LA).

Para estas duas cidades, caracterizou-se o comportamento de consumo dos seus consumidores através das suas medidas descritivas: a média, a

mediana, o valor mínimo e máximo, o primeiro e o terceiro quartis, a, variância, o desvio padrão e outros, apresentado na Tabela 4.2.

Tabela 4.2 – Medidas descritivas das cidades de Walla Walla e Lake Charles

	Walla Walla	Lake Charles
Número de dados (observações)	8.760	8.760
Número de dados erróneos	0	0
Valor Mínimo (kWh)	0,663	0,478
Valor Máximo (kWh)	11,902	6,998
1º Quartil (Q1) (kWh)	1,984	1,152
3º Quartil (Q3) (kWh)	7	2,269
Amplitude interquartil (kWh)	5,016	1,117
Média (kWh)	4,613	1,8
Mediana (kWh)	4,323	1,622
Soma (kWh)	40.414	15.773
Desvio Padrão (kWh)	2,791	0,899
Skewness (Assimetria da curva de distribuição)	0,326	1,261
Kurtosis (Achatamento da curva de distribuição)	-1,15	2,428

Em ambas as cidades, foram analisados 8760 dados que representam as leituras horárias de consumo residencial durante um ano para cada uma. Comparando os perfis de consumo das duas cidades, Tabela 4.2, verifica-se que os consumidores de Walla Walla têm valores superiores em relação ao valor máximo de consumo, consumo anual, amplitude interquartil, média, assim como o desvio padrão. O que significa que os consumidores desta cidade têm um maior consumo do que os de Lake Charles e que o seu consumo tem uma maior variação ao longo do ano.

Tendo em conta o índice *Skewness* da Tabela 4.2 dos consumidores das duas cidades, verifica-se que a distribuição dos consumidores de Walla Walla é mais simétrica do que os consumidores de Lake Charles que apresenta ter uma distribuição assimétrica positiva.

Analisando o achatamento da curva de distribuição dos dois medoides através da análise do índice *Kurtosis*, verifica-se que esta é mais achatada para os consumidores de Walla Walla que para os de Lake Charles, o que

mostra que os consumidores de Walla Walla têm um consumo mais disperso, o que também pode ser verificado pelo respetivo desvio padrão.

Em seguida, efetuou-se uma análise mais detalhada dos consumos dos consumidores dos medoides através de várias ferramentas. O respetivo código em R, é apresentado neste documento no Anexo E.

4.2.1 Histograma

Para a análise dos valores mais frequentes de consumo recorreu-se aos histogramas das cidades, conforme Figura 4.2 e Figura 4.3.

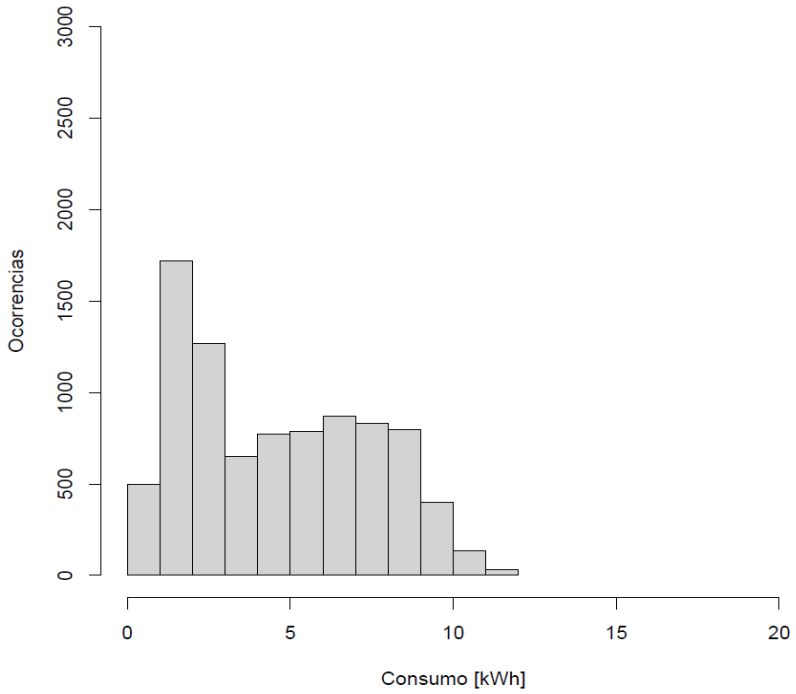


Figura 4.2 – Histograma dos consumos anuais da cidade de Walla Walla

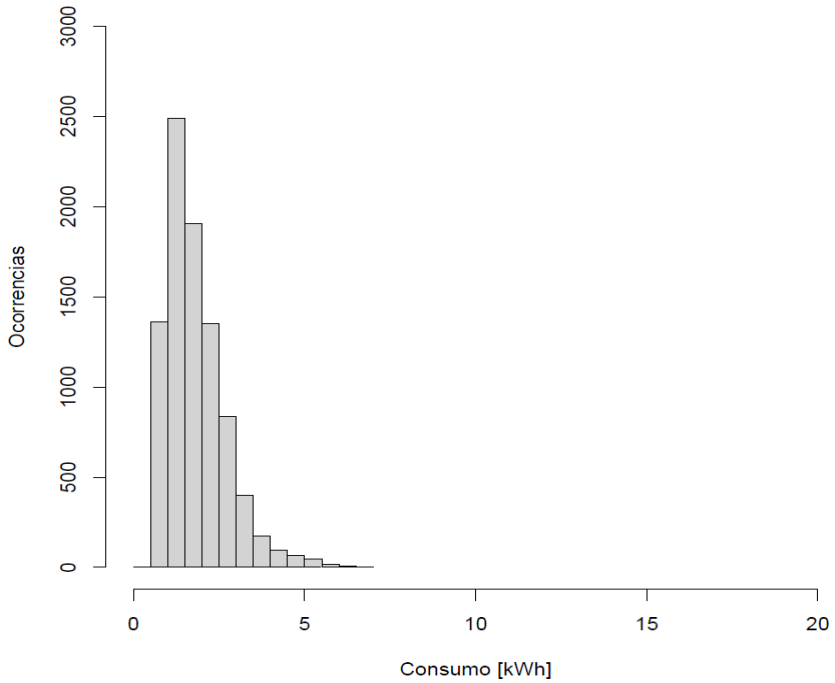


Figura 4.3 – Histograma dos consumos anuais da cidade de Lake Charles

Pela análise dos histogramas, verifica-se que os consumos dos consumidores de Walla Walla variam dentro do intervalo [0; 12] kWh ao longo do ano sendo mais frequente consumos entre os 1 kWh e 3 kWh. Para os consumidores Lake Charles, o seu consumo ao longo do ano varia entre os 0,5 kWh e os 7 kWh, sendo mais frequente os consumos entre os 1 kWh e 2 kWh. Visto isto, conclui-se que a amplitude de consumo dos consumidores de Walla Walla é maior que os de Lake Charles.

De acordo com o resultado do coeficiente de assimetria da Tabela 4.2, confirma-se também com o respetivo histograma que a distribuição de consumos dos consumidores de Lake Charles apresenta ser bastante assimétrica com enviesamento à direita.

4.2.2 *Boxplots, Outliers* e sua localização no cronograma de consumo

De modo a identificar o intervalo interquartilico e a presença de valores atípicos nos perfis de consumo dos consumidores das cidades recorreu-se aos *boxplots*, identificação de *outliers* e a sua localização no respetivo cronograma de consumo conforme as figuras, Figura 4.4 à Figura 4.9.

4.2.1.1 *Outliers*

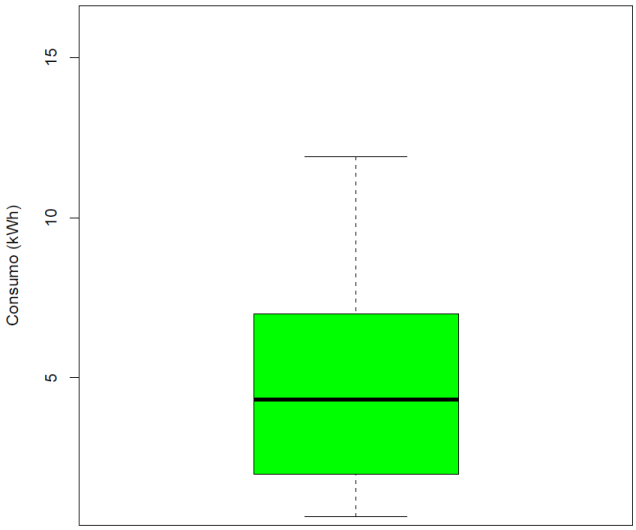


Figura 4.4 - Outliers da cidade Walla Walla

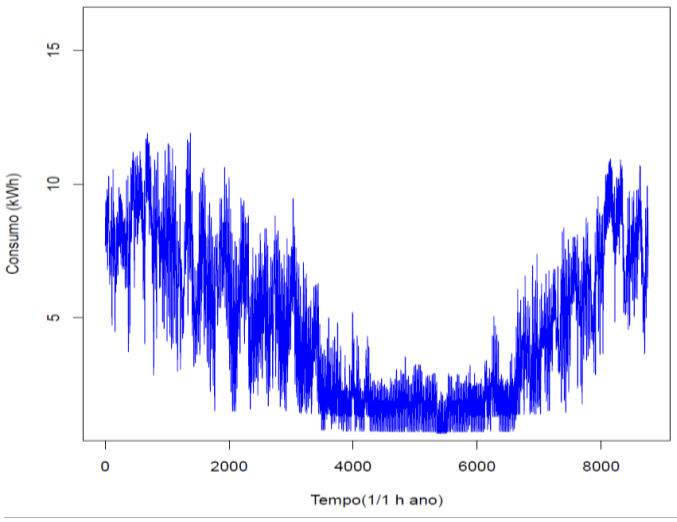


Figura 4.5 - Localização de outliers no cronograma da cidade Walla Walla

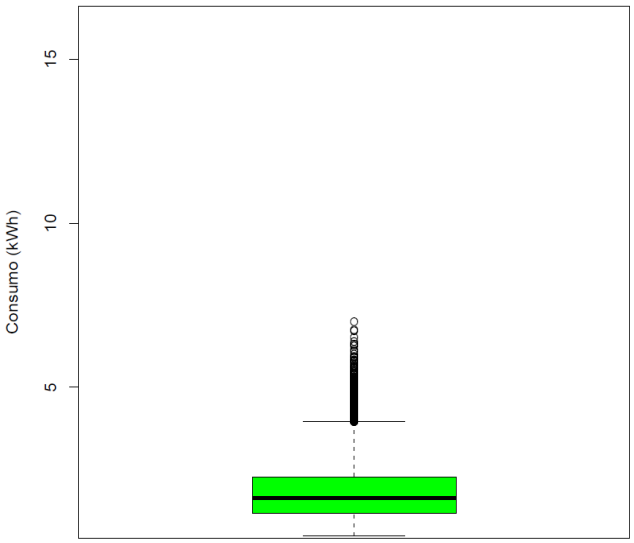


Figura 4.6 - Outliers da cidade Lake Charles

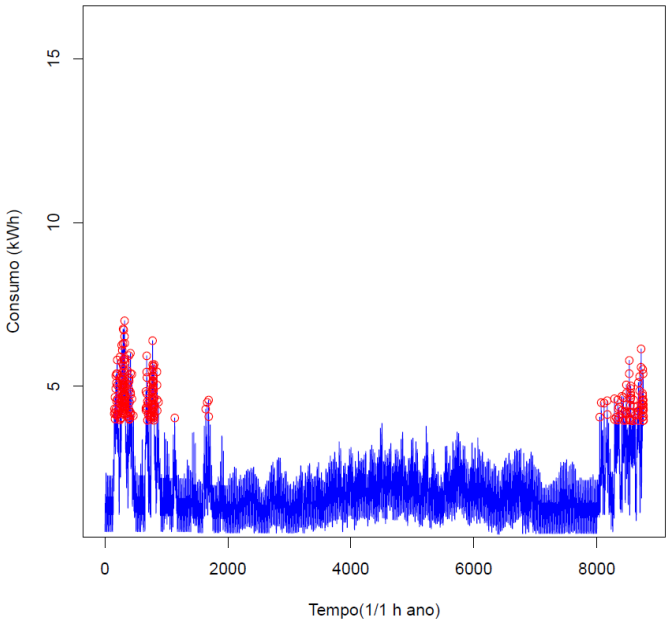


Figura 4.7 - Localização de outliers no cronograma da cidade Lake Charles

4.2.1.2 *Outliers* severos

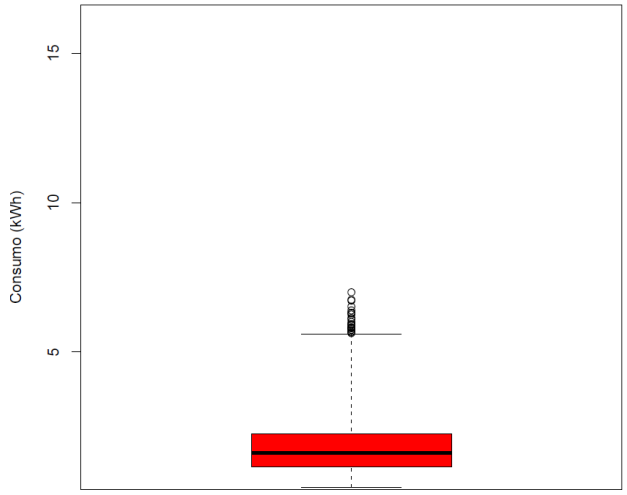


Figura 4.8 - Outliers severos da cidade Lake Charles

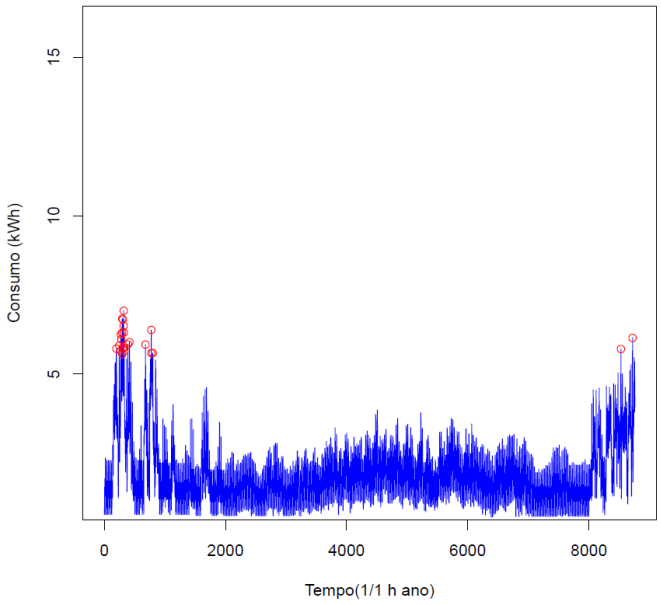


Figura 4.9 - Localização de outliers severos no cronograma da cidade Lake Charles

Ao analisar a presença de *outliers* nas respetivas cidades (da Figura 4.4 à Figura 4.9), verifica-se que a única cidade que os apresenta é Lake Charles. Os seus *outliers*, valores atípicos de consumo, localizam-se no início e fim

do ano, período de Inverno, e os seus valores se situam entre os 4 e os 7 kWh, não existindo *outliers* com valores negativos.

Também se pode verificar que o intervalo interquartilico dos consumidores de Walla Walla é maior que os de Lake Charles através da dimensão dos respetivos *boxplots*.

4.2.3 Cronograma

De modo a detetar a existência de sazonalidade e compreender os hábitos de consumo de energia elétrica e gás ao longo do ano dos consumidores das duas cidades, recorreu-se à caracterização das suas séries temporais de consumo através dos seus cronogramas, comparando-os, conforme a Figura 4.10.

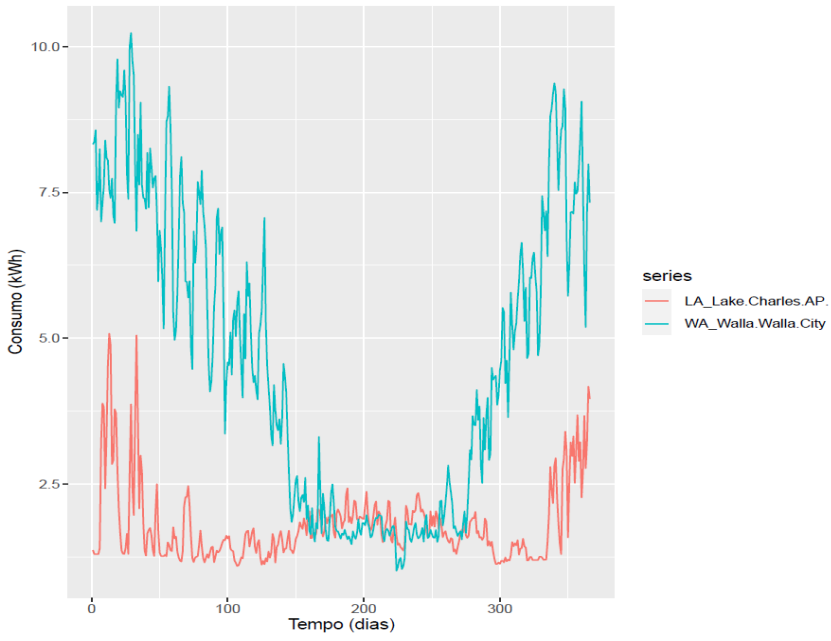


Figura 4.10 – Cronogramas das cidades de Walla Walla (a verde) e Lake Charles (a laranja)

Através da análise dos cronogramas, verificou-se o seguinte comportamento de consumo de Walla Walla, a verde na Figura 4.10, no início do ano existe um consumo alto que vai decaindo gradualmente até

meio do ano, altura do Verão e que sensivelmente a partir do final do Verão, o seu consumo vai aumentando gradualmente atingindo o seu pico no final do ano.

Pela Figura 4.10, comparando os consumos dos consumidores das duas cidades, verifica-se que os consumos de Walla Walla são mais altos ao longo do ano comparativamente com Lake Charles, com exceção da altura do Verão que apresenta consumos mais baixos.

Os resultados demonstram ainda que em ambas as cidades existem picos de consumo no início e final do ano, período de Inverno, embora os consumos de Walla Walla sejam bastante mais altos neste período do ano.

4.2.4 Função de autocorrelação (ACF)

De modo a compreender o comportamento de consumo horário dos consumidores das duas cidades recorreu-se à análise dos correlogramas com um desfasamento (lag) até 3 dias (últimas 72 horas), conforme a Figura 4.11 e Figura 4.12.

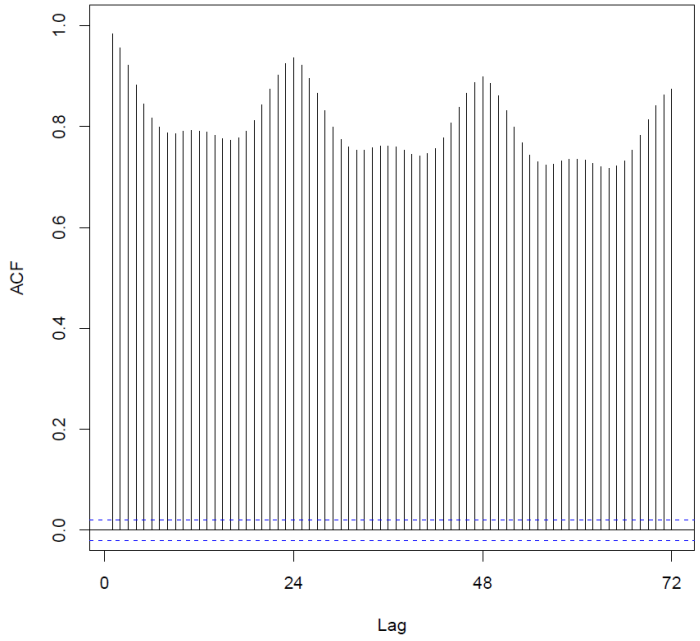


Figura 4.11 – ACF de Walla Walla

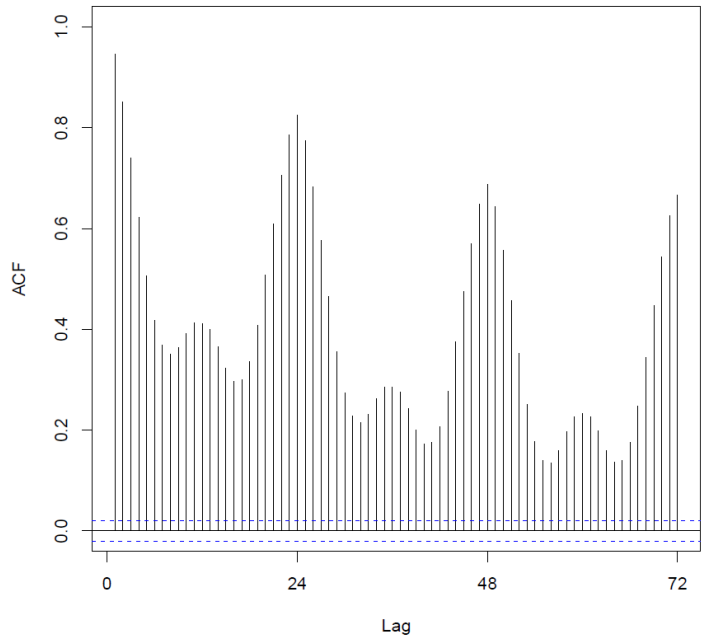


Figura 4.12 – ACF de Lake Charles

Através da análise da ACF, Figura 4.11, verifica-se que para a cidade de Walla Walla, que o consumo dos seus consumidores nas horas anteriores afeta o consumo presente, tendo um maior efeito a cada 24 horas. O decaimento lento indica que possivelmente a série tem uma tendência, sendo igualmente evidente alguma ciclicidade.

Em relação à análise da ACF, Figura 4.12, dos consumidores de Lake Charles verifica-se que o seu consumo varia consideravelmente da presente hora para hora anterior, mas que gradualmente vai afetar cada vez menos a hora presente, os consumos das 12 horas anteriores. Também se verifica que este comportamento é cíclico a cada 24 horas, mas de uma forma cada vez mais atenuada.

4.3 Análise de resultados

4.3.1 Análise preliminar

Devido ao grande volume de dados obtido da análise das 936 cidades, a totalidade dos dados, optou-se primeiro em aplicar a metodologia descrita no capítulo 3, a dados de consumo anual de eletricidade e gás de consumidores residências de 50 cidades, tendo sido escolhida uma cidade por cada estado. Estes representam 438.000 observações que foram consideradas como uma amostra de dados, que permitiram testar a metodologia e obter os resultados de uma forma mais breve.

Inicialmente determinou-se qual o número de grupos (*clusters)* aconselhado para efetuar o *clustering* através do índice *Silhouette*. Este número representa o número de grupos de consumos padrão distintos em que é possível agrupar o consumo dos consumidores destas 50 cidades, conforme demonstrado no gráfico da Figura 4.13.

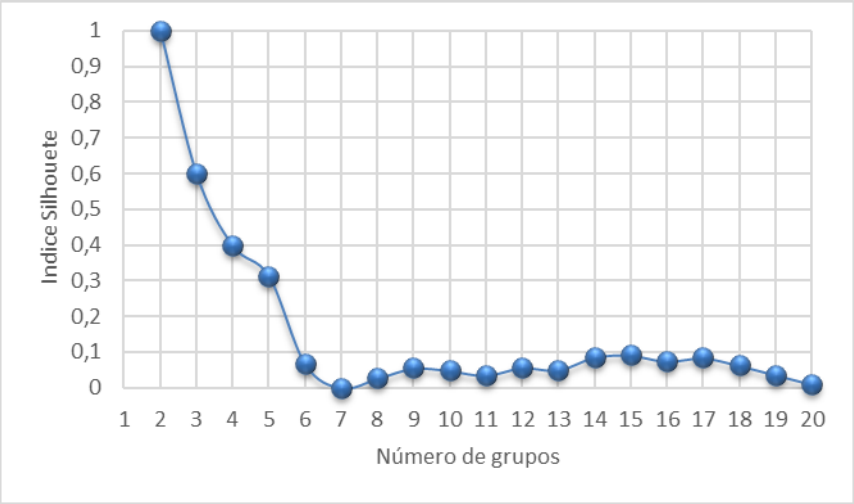


Figura 4.13 – Determinação do número de grupos para as 50 cidades analisadas

Verificou-se ao analisar a Figura 4.13, que o número aconselhado de grupos é de dois tendo como critério o índice *Silhouette*, visto que o seu valor está mais próximo de 1.

Uma vez determinado o número de grupos, efetuou-se o *clustering* tendo em conta o número de grupos indicado. Resumidamente, os consumos dos consumidores das 50 cidades ficaram distribuídos da seguinte forma por dois grupos distintos: 34 cidades no grupo 1 e 16 no grupo 2. Verificou-se que das 50 cidades da amostra escolhida, cerca de 68% das cidades os seus consumidores pertencem ao grupo 1 e 32% ao grupo 2.

Sendo o medoide o representante tipo de um grupo, verificou-se que o medoide do grupo 1 é o perfil de consumo dos consumidores da cidade de Providence do estado de Rhode Island e do grupo 2 é o perfil de consumo dos consumidores da cidade de Sacramento do estado da Califórnia.

Na Figura 4.14, estão representados os perfis de consumo dos consumidores das duas cidades através dos seus cronogramas.

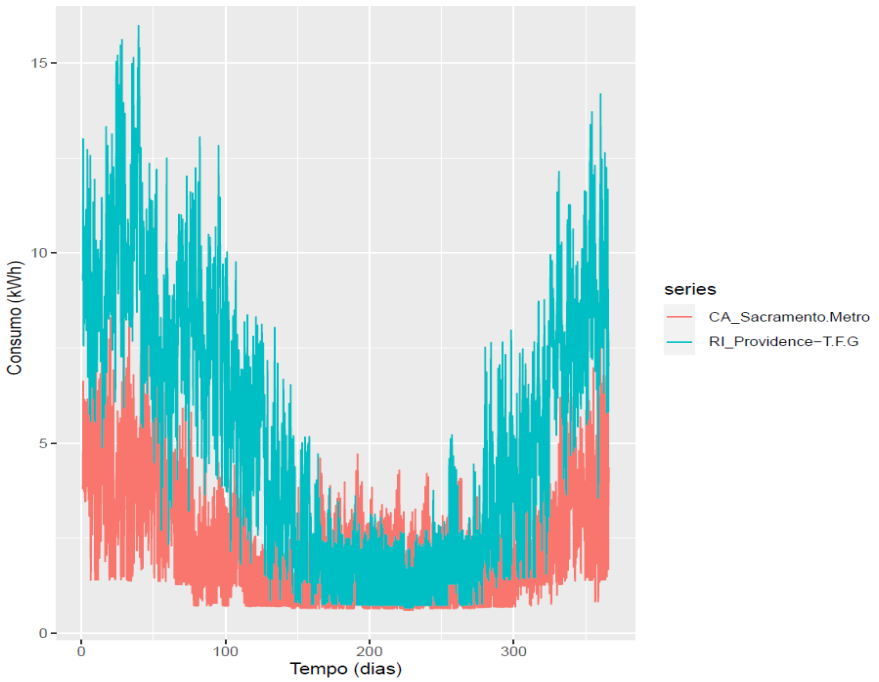


Figura 4.14 - Cronograma das cidades: Providence (a verde) e Sacramento (a laranja)

Por forma a analisar mais facilmente a característica de consumo dos consumidores das duas cidades, optou-se por alisar as curvas de consumo dos seus cronogramas usando a média móvel diária, conforme a Figura 4.15.

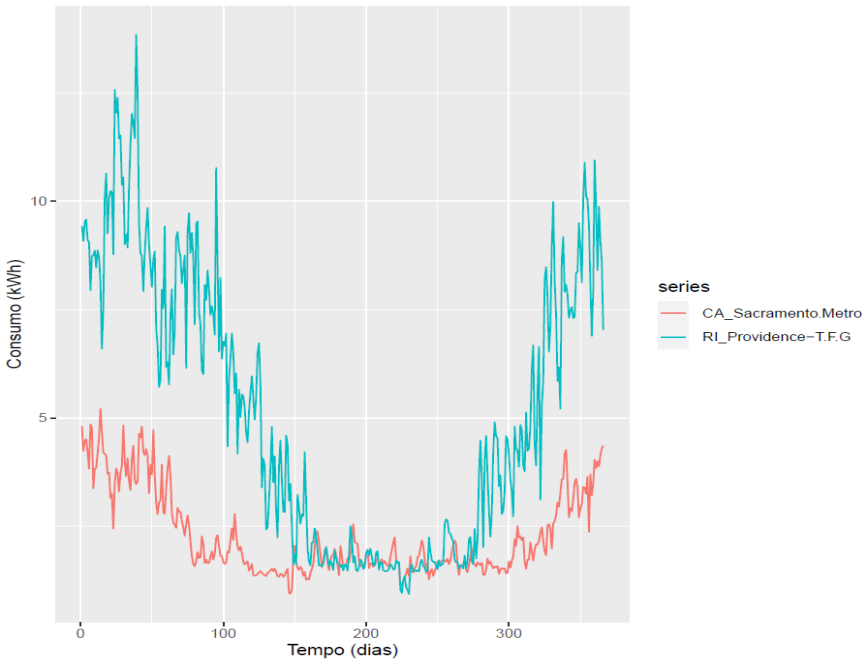


Figura 4.15 – Cronograma alisado das cidades de Providence (a verde) e Sacramento (a laranja)

Caracterizando os dois medoides, um de cada grupo, conclui-se que o perfil de consumo dos consumidores de uma cidade do grupo 1, representado pela cidade Providence (medoide do grupo 1), a verde na Figura 4.15, tem um perfil que consome grande quantidade de energia elétrica durante o Inverno (início e final do ano) mas que esta vai decaindo gradualmente durante o ano até à altura do Verão, sensivelmente, e que no final do Verão o seu consumo sobe gradualmente atingindo o seu pico no Inverno.

Para o perfil de consumo dos consumidores do grupo 2, representada pela cidade de Sacramento (medoide do grupo 2), pode ser caracterizado por apresentar também picos de consumo na altura do Inverno embora mais baixos comparativamente com os consumidores do grupo 1.

Comparando os perfis de consumo dos dois medoides, validou-se que os perfis de consumo dos consumidores das cidades do grupo 1 são mais altos

do que os do grupo 2 durante um ano, mas que na altura do Verão estes são muito semelhantes tendo por vezes o grupo 2 consumos mais altos.

Na Figura 4.16 é mostrado geograficamente nos EUA, o resultado do *clustering* efetuado.

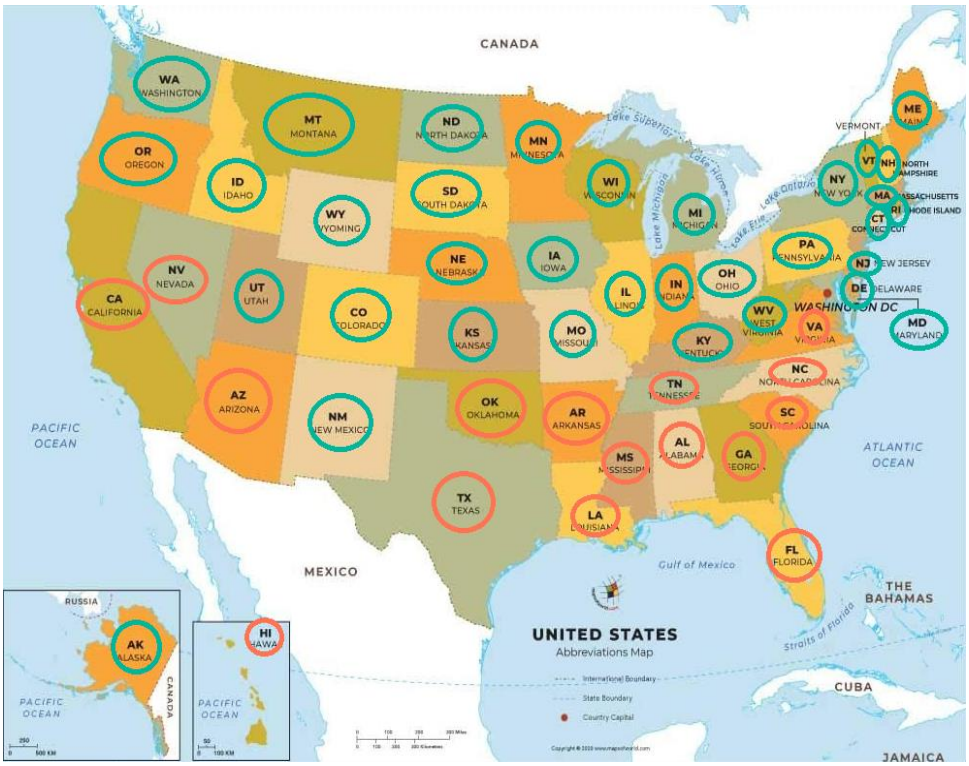


Figura 4.16 – Localização geográfica dos grupos para 50 cidades: grupo 1 a verde e grupo 2 a laranja

Pela análise da Figura 4.16, consegue-se ter mais a perceção de como ficou efetuado o *clustering* pela metodologia *k-medoids*. Claramente, verificou-se que os perfis de consumo de consumidores das cidades correspondentes ao grupo 1 são consumidores das cidades dos estados a norte (Alasca incluído) ou centrais dos EUA, ao passo que os perfis de consumo de consumidores das cidades do grupo 2 são cidades dos estados a sul dos EUA (incluído o Havai), tendo como exceção o estado do Novo México (NM) que pertence ao grupo 1.

Analisando o resultado anterior do *clustering* dos perfis de consumo de energia elétrica e gás com os dados das regiões climáticas dos EUA, segundo a Figura 4.17.

Figura 4.17 – Regiões climáticas dos EUA [24]

Obteve-se os seguintes resultados, conforme a Tabela 4.3.

Tabela 4.3 – Análise do agrupamento com as regiões climáticas para as 50 cidades

Grupo e região climática	Contagem de Cidades
Grupo 1	**34**
N/D	1
Muito fria (1)	4
Fria (2)	20
Fria (2) ou Húmida (4)	1
Amena (3)	2
Húmida (4)	6
Grupo 2	**16**
N/D	1
Húmida (4)	7
Húmida (4) ou Quente (5)	1
Quente (5)	4
Muito Seca (7)	3
Total Geral	**50**

De uma forma geral em relação às regiões climáticas, os resultados da Tabela 4.3, demonstraram que os perfis correspondentes a consumidores das cidades analisadas do grupo 1 predomina a região Fria (2) enquanto para o grupo 2 é a região Húmida (4). Também se verificou que para os perfis de consumo dos consumidores das cidades do grupo 1 as regiões climáticas presentes são Muito Fria (1), Fria (2), Amena (3), Húmida (4) sendo que as regiões climáticas Muito Fria (1), Fria (2), Amena (3) são exclusivas deste grupo.

Quanto para os perfis de consumo de consumidores pertencentes ao grupo 2, as regiões climáticas pertencentes são Húmida (4), Quente (5), Muito Seca (7), sendo exclusivas as regiões climáticas Quente (5), Muito Seca (7). Em relação à região climática Húmida (4), esta está presente nos dois grupos de perfis de consumo de consumidores ao que não se pode afirmar que esta é exclusiva de um grupo.

De notar que na Tabela 4.3 são apresentadas cidades classificadas nas regiões climáticas Fria (2) ou Húmida (4) e Húmida (4) ou Quente (5) pertencentes aos grupos de consumo 1 e 2 respetivamente, devido a que as cidades analisadas desses grupos de consumo, estarem na fronteira geográfica dessas regiões climáticas não sendo possível determinar com exatidão.

Ainda referir, que está presente o N/D como região climática tanto no grupo 1 como no grupo 2 devido a que no mapa apresentado na Figura 4.17 não está incluído o estado do Alasca nem o estado do Havai, respetivamente.

Os resultados da determinação do número indicado de grupos, *clustering* efetuado, grupos (*clusters*) e regiões climáticas correspondentes encontram-se no Anexo A. Os resultados dos histogramas, ACF, *outliers* e perfis de consumo dos medoides citados acima, encontram-se no Anexo B.

No final de validar com sucesso a aplicação da metodologia com os dados da amostra, escolha de 50 cidades, uma por estado, prosseguiu-se com a obtenção dos resultados para a totalidade dos dados.

4.3.2 Análise da totalidade dos dados

Conforme referido no final do subcapítulo 4.1, os resultados principais para este estudo foram obtidos através da análise de dados de consumo de energia elétrica e gás de consumidores residenciais de 936 cidades dos EUA. Como os dados de consumo anuais de consumidores de uma cidade representam 8760 observações, no total foram analisadas 8.199.360 observações.

Tal como aconteceu na análise com a amostra de dados, o número aconselhado de grupos (*clusters*) de consumos padrão distintos para efetuar o *clustering* é de dois, usando o índice *Silhouette* como critério de escolha, conforme apresentado no gráfico na Figura 4.18.

Figura 4.18 – Determinação do número de grupos para as 936 cidades analisadas

Tendo em conta o número de grupos determinado, o *clustering* ficou efetuado da seguinte forma: das 936 cidades, 635 cidades os seus consumidores pertencem ao grupo 1 e os consumidores das 301 cidades ao grupo 2. Tal como aconteceu com as 50 cidades analisadas, verificou-se que no *clustering* de consumo de energia elétrica e gás de consumidores das 936 cidades, a maior parte ficou agregado ao grupo 1 representando cerca de 68 % dos dados enquanto o grupo 2 ficou com cerca de 32% das cidades.

Os resultados detalhados deste *clustering* serão apresentados em anexo neste trabalho, Anexo C e D, devido ao grande volume de dados dos consumos das 936 cidades analisadas.

Quanto aos medoides de cada grupo, o medoide do grupo 1 é representado pelos perfis de consumo de consumidores da cidade de Walla Walla do estado de Washington e do grupo 2 é representado pelos perfis de consumo de consumidores da cidade de Lake Charles do estado do Louisiana.

As análises dos perfis de consumo de energia elétrica destes medoides foram descritas em detalhe anteriormente no subcapítulo 4.2 - "*Análise exploratória dos dados*" nas cidades exemplo.

De uma forma resumida, verificou-se que para o perfil de consumo dos consumidores de Lake Charles existe um pico de consumo no início e no final do ano, período de Inverno. Quanto ao resto do ano, o consumo é mais uniforme não apresentando grandes picos de consumo, conforme Figura 4.19. No caso da cidade de Walla Walla, o perfil de consumo dos seus consumidores no início do ano existe um consumo alto que vai decaindo gradualmente até meio do ano, altura do Verão e que a partir do final do Verão, o seu consumo vai aumentando gradualmente atingindo o seu pico de consumo no final do ano, conforme Figura 4.19.

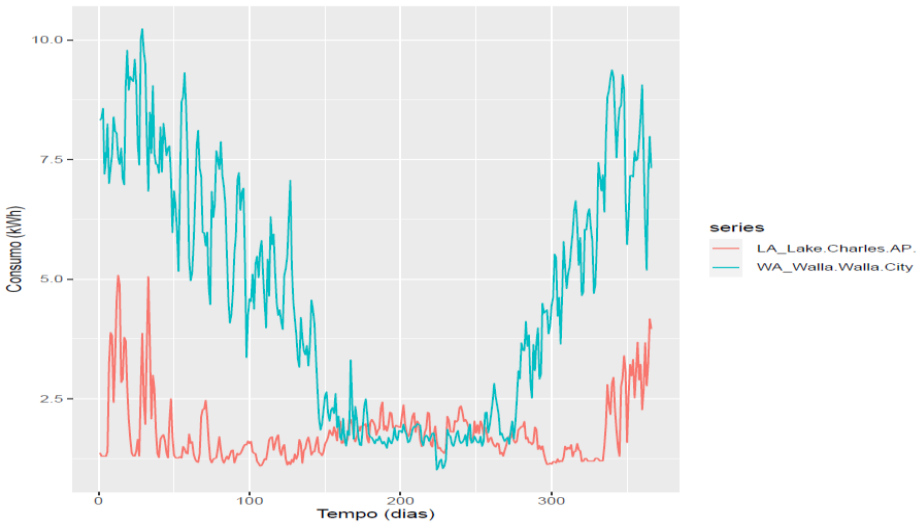


Figura 4.19 - Cronograma alisado para as cidades de Walla Walla (a verde) e Lake Charles (a laranja)

Comparando as duas cidades, observou-se que o perfil de consumo de consumidores de Walla Walla são mais alto ao longo do ano comparativamente com Lake Charles, com exceção da altura do Verão. Também se verificou que em ambas as cidades existem nos perfis de consumo dos seus consumidores, picos de consumo no início e final do ano, período de Inverno embora os consumos nessas alturas de consumidores de Walla Walla sejam bastante mais altos.

Na Figura 4.20 é apresentado o resultado do *clustering* geograficamente nos EUA.

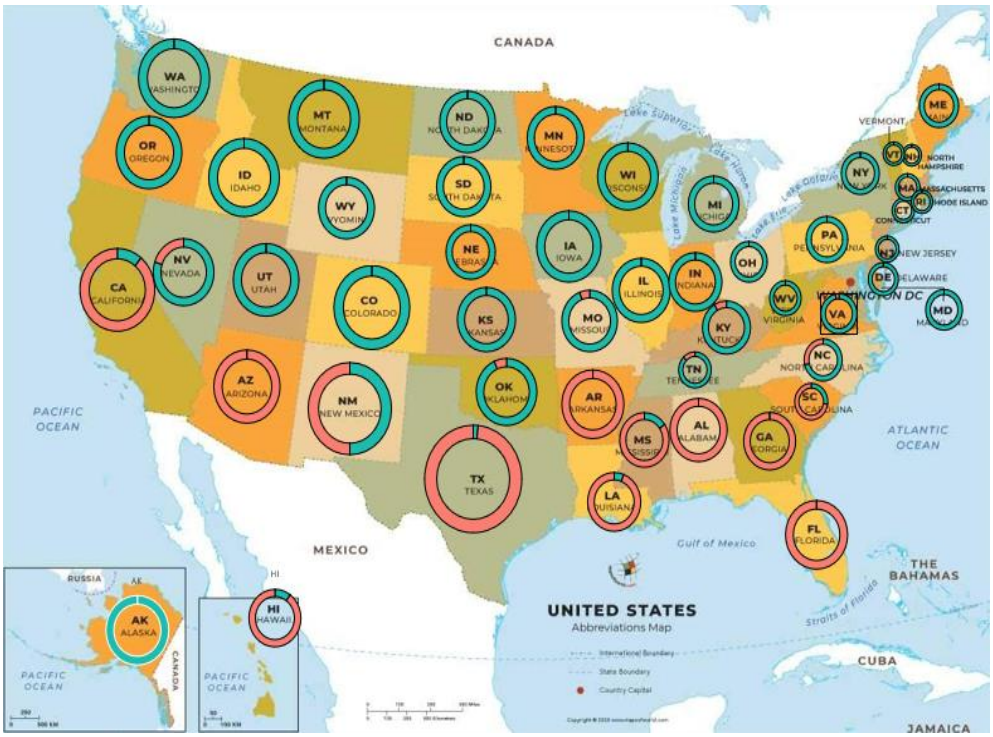


Figura 4.20 – Localização dos grupos de consumo a que pertencem as cidades por estado: grupo 1 a verde e grupo 2 a laranja

Na Figura 4.20 , é apresentado em cada estado, a proporção das cidades que têm consumidores com um consumo semelhante ao grupo 1, a verde, ou ao grupo 2, a laranja. Tendo como o exemplo o estado do Nevada (NV), verifica-se que os consumos dos consumidores das cidades analisadas deste estado maioritariamente pertencem ao grupo 1, no entanto, cerca de um quarto das cidades o seu consumo pertence ao grupo 2.

Ao analisar a Figura 4.20, tal como aconteceu nos resultados da amostra, verificou-se que os consumos dos consumidores das cidades correspondentes ao grupo 1 são as cidades dos estados a norte ou centrais dos EUA, ao passo que os das cidades do grupo 2 são cidades dos estados a sul dos EUA, tendo como exceção o estado do Novo México (NM), novamente. Ao contrário do que aconteceu nos resultados de *clustering* para as 50 cidades, em que os consumidores do Novo México pertenciam ao grupo 1, nos resultados da análise da totalidade dos dados, este estado mantém-se "neutro". A razão desta "neutralidade" deve-se a que dos consumos de consumidores das 16 cidades analisadas neste estado, metade pertence ao grupo 1 e a outra metade ao grupo 2, conforme demonstrado na Figura 4.20.

Ainda de referir que os consumidores das cidades do Alasca pertencem totalmente ao grupo 1 e as cidades do Havai maioritariamente pertencem ao grupo 2.

Nas seguintes tabelas, Tabela 4.4 e Tabela 4.5, são apresentados o número de cidades analisadas por estado e a identificação do respetivo grupo de consumo, em valor absoluto e em percentagem respetivamente.

Tabela 4.4 – Cidades analisadas por estado e a que grupo pertencem

	Grupo			Grupo	
Estado	1	2	**Estado**	1	2
Alabama (AL)		14	Michigan (MI)	32	
Alasca (AK)	2		Minnesota (MN)	54	
Arizona (AZ)		18	Mississippi (MS)	2	11
Arkansas (AR)		18	Missouri (MO)	16	1
Califórnia (CA)	8	65	Montana (MT)	16	
Carolina do Norte (NC)	15	6	Nebrasca (NE)	26	
Carolina do Sul (SC)	3	8	Nevada (NV)	8	2
Colorado (CO)	25		Nova Hampshire (NH)	8	
Connecticut (CT)	7		Nova Iorque (NY)	24	
Dakota do Norte (ND)	10		Nova Jérsia (NJ)	9	
Dakota do Sul (SD)	11		Novo México (NM)	8	8
Delaware (DE)	2		Ohio (OH)	13	
Flórida (FL)		42	Oklahoma (OK)	14	1
Geórgia (GA)		19	Oregon (OR)	19	
Havai (HI)	1	9	Pensilvânia (PA)	21	
Idaho (ID)	13		Rhode Island (RI)	3	
Illinois (IL)	19		Tennessee (TN)	7	1
Indiana (IN)	10		Texas (TX)	1	60
Iowa (IA)	39		Utah (UT)	13	
Kansas (KS)	23		Vermont (VT)	4	
Kentucky (KY)	11	1	Virgínia (VA)	30	1
Louisiana (LA)	1	16	Virgínia Ocidental (WV)	11	
Maine (ME)	15		Washington (WA)	29	
Maryland (MD)	5		Wisconsin (WI)	19	
Massachusetts (MA)	15		Wyoming (WY)	13	

Tabela 4.5 – Cidades analisadas por estado e a que grupo pertencem em percentagem

	Grupo			Grupo	
Estados	1	2	**Estados**	1	2
Alabama (AL)	0%	100%	Michigan (MI)	100%	0%
Alasca (AK)	100%	0%	Minnesota (MN)	100%	0%
Arizona (AZ)	0%	100%	Mississippi (MS)	15%	85%
Arkansas (AR)	0%	100%	Missouri (MO)	94%	6%
Califórnia (CA)	11%	89%	Montana (MT)	100%	0%
Carolina do Norte (NC)	71%	29%	Nebrasca (NE)	100%	0%
Carolina do Sul (SC)	27%	73%	Nevada (NV)	80%	20%
Colorado (CO)	100%	0%	Nova Hampshire (NH)	100%	0%
Connecticut (CT)	100%	0%	Nova Iorque (NY)	100%	0%
Dakota do Norte (ND)	100%	0%	Nova Jérsia (NJ)	100%	0%
Dakota do Sul (SD)	100%	0%	Novo México (NM)	50%	50%
Delaware (DE)	100%	0%	Ohio (OH)	100%	0%
Flórida (FL)	0%	100%	Oklahoma (OK)	93%	7%
Geórgia (GA)	0%	100%	Oregon (OR)	100%	0%
Havai (HI)	10%	90%	Pensilvânia (PA)	100%	0%
Idaho (ID)	100%	0%	Rhode Island (RI)	100%	0%
Illinois (IL)	100%	0%	Tennessee (TN)	88%	13%
Indiana (IN)	100%	0%	Texas (TX)	2%	98%
Iowa (IA)	100%	0%	Utah (UT)	100%	0%
Kansas (KS)	100%	0%	Vermont (VT)	100%	0%
Kentucky (KY)	92%	8%	Virgínia (VA)	97%	3%
Louisiana (LA)	6%	94%	Virgínia Ocidental (WV)	100%	0%
Maine (ME)	100%	0%	Washington (WA)	100%	0%
Maryland (MD)	100%	0%	Wisconsin (WI)	100%	0%
Massachusetts (MA)	100%	0%	Wyoming (WY)	100%	0%

Analisando a Tabela 4.4 e Tabela 4.5, tal como referido anteriormente, verificou-se que a maior parte das cidades analisadas têm um perfil de consumo semelhante ao grupo 1. Além disto, provou-se através da análise destas tabelas em termos de consumo, o que foi mostrado na Figura 4.16, que os perfis consumos de consumidores das cidades dos estados do norte dos EUA são totalmente pertencentes ao grupo 1. O contrário não acontece com os perfis consumos de consumidores das cidades dos estados do sul, ou seja, terem perfis de consumo exclusivamente semelhantes ao grupo 2, neste caso existem em alguns estados, cidades em que os perfis de consumo dos

seus consumidores são semelhantes ao grupo 1. Quanto aos perfis de consumo de consumidores das cidades dos estados do centro dos EUA, estes têm um perfil de consumo maioritariamente semelhante ao grupo 1.

Capítulo 5 – Conclusões e trabalho futuro

Neste capítulo são apresentadas as principais conclusões do trabalho bem como algumas sugestões para trabalhos futuros sobre o tema.

5 Conclusões e trabalho futuro

Com a implementação de redes de energia elétrica tornou-se importante perceber quais os grupos de consumidores que dependem desta e quais os seus hábitos de consumo. Além disto, com o aumento de consumidores de uma rede elétrica surgiu a preocupação de implementar uma rede suficientemente capaz no fornecimento sem interrupções e com o mínimo de falhas possível. Com este objetivo e devido ao avanço tecnológico, atualmente é possível ter acesso a dados de consumo de consumidores de energia elétrica através da instalação na rede de energia elétrica de contadores inteligentes (*smart meters*). Mas a obtenção dos dados só por si não basta e devido ao grande volume e variedade de dados obtido, torna-se necessário desenvolver técnicas eficazes na interpretação dos dados. Com base nesta ideia, surgiu a possibilidade de utilizar técnicas de *clustering* ou agrupamento de dados por forma a agrupar consumidores com perfis de consumo semelhantes e em que os grupos formados sejam os mais distintos possíveis.

Nesta dissertação, foi aplicada uma técnica de *clustering* de modo que permitisse interpretar dados provenientes de consumidores residenciais de energia elétrica e gás combinados, analisando 936 cidades dos 50 estados dos EUA. A técnica aplicada consistiu em aplicar o algoritmo *k-medoids* a dados de consumo residencial das 936 cidades, em que o número de grupos (*clusters*) de consumo padrão de energia elétrica e gás foi determinado pelo índice *Silhouette*.

Resumidamente, este *clustering* ficou efetuado em dois grupos distintos: um grupo que abrangia os consumidores das cidades dos estados do norte e centro, e outro grupo em que os seus consumidores eram de cidades dos estados do sul dos EUA, com exceção do Novo México (NM). O resultado do *clustering* como o número de grupos (*clusters*) distintos aconselhados

foram semelhantes tanto para a análise dos consumidores das cidades da amostra bem como para a totalidade das cidades.

Também foi possível caracterizar os grupos de consumidores consoante as regiões climáticas existentes nos EUA: consumidores das cidades do grupo 1 as regiões climáticas presentes são Muito Fria (1), Fria (2), Amena (3), Húmida (4) sendo que as regiões climáticas Muito Fria (1), Fria (2), Amena (3) são exclusivas deste grupo. Quanto consumidores das cidades do grupo 2, as regiões climáticas presentes são Húmida (4), Quente (5), Muito Seca (7) sendo exclusivas as regiões climáticas Quente (5), Muito Seca (7). Em relação à região climática Húmida (4), esta está presente nos dois grupos de consumidores ao que não se pode afirmar que esta é exclusiva de um grupo. Ainda de referir que nem o estado do Alasca nem o estado do Havai poderão ser classificados climaticamente devido a estes não estarem presentes no mapa climático dos EUA.

Quanto aos medoides de cada grupo, considerando como medoide os consumidores de uma cidade que representa o perfil de consumo padrão de cada grupo, estes demostraram ter comportamentos de consumo diferentes ao logo de um ano como esperado. Além disso, se compararmos os perfis de consumo dos consumidores dos medoides da análise da amostra com os da da totalidade dos dados, estes são muitos semelhantes para os respetivos grupos 1 e 2.

Também se concluiu, perante os resultados da identificação de *ouliers*, que estes existiam apenas para os consumidores da cidade de Lake Charles, perfil representante do grupo que englobava os estados no sul dos EUA, com exceção do Novo México. A presença de *outliers* na altura do Inverno poderá ser indicador de seja aconselhável reforçar a rede de fornecimento de energia elétrica e gás nesta altura do ano para estes estados.

Os resultados obtidos da aplicação desta técnica demonstraram que esta foi eficaz na tipificação dos consumos de consumidores das cidades analisadas dos EUA, tanto quando aplicada a uma amostra de dados (50 cidades) bem como na aplicação à totalidade dos dados (936 cidades). Além

da tipificação, também foi possível caracterizar o comportamento de consumo dos respetivos medoides de cada grupo, havendo distinção entre eles.

Como sugestão para trabalho futuro, de modo a refinar o estudo, seria interessante ter dados mais detalhados quanto às regiões climáticas existentes nos EUA, como por exemplo: amplitude térmica, valores máximos e mínimos de temperatura, por forma a poder caracterizar melhor os grupos existentes. Também seria interessante utilizar dados socioeconómicos como por exemplo o rendimento per capita, qual o horário laboral dos consumidores e/ou algum indicador de acessibilidade ao fornecimento de energia elétrica e gás para detetar falhas de abastecimento de modo a justificar o comportamento/tendência do seu consumo.

Bibliografia

Bibliografia

[1] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Trans Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015, doi: 10.1109/TSG.2014.2364233.

[2] A. Zagouras, H. T. C. Pedro, and C. F. M. Coimbra, "Clustering the solar resource for grid management in island mode," *Solar Energy*, vol. 110, pp. 507–518, Dec. 2014, doi: 10.1016/j.solener.2014.10.002.

[3] G. Liu, L. Zhu, X. Wu, and J. Wang, "Time series clustering and physical implication for photovoltaic array systems with unknown working conditions," *Solar Energy*, vol. 180, pp. 401–411, Mar. 2019, doi: 10.1016/j.solener.2019.01.041.

[4] X. Fu, X. J. Zeng, P. Feng, and X. Cai, "Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China," *Energy*, vol. 165, pp. 76–89, Dec. 2018, doi: 10.1016/j.energy.2018.09.156.

[5] I. P. Panapakidis and A. S. Dagoumas, "Day-ahead electricity price forecasting via the application of artificial neural network based models," *Appl Energy*, vol. 172, pp. 132–151, Jun. 2016, doi: 10.1016/j.apenergy.2016.03.089.

[6] P. Miguel, J. Gonçalves, L. Neves, and A. G. Martins, "Using clustering techniques to provide simulation scenarios for the smart grid," *Sustain Cities Soc*, vol. 26, pp. 447–455, Oct. 2016, doi: 10.1016/j.scs.2016.04.012.

[7] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renewable and Sustainable Energy Reviews*, vol. 120, Mar. 2020, doi: 10.1016/j.rser.2019.109628.

[8] C. M. Cheung, R. Kannan, and V. K. Prasanna, "Load Demand User Profiling in Smart Grids withDistributed Solar Generation," Los Angeles, 2020.

[9] I. Benítez, A. Quijano, J. L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *International Journal of Electrical Power and Energy Systems*, vol. 55, pp. 437–448, 2014, doi: 10.1016/j.ijepes.2013.09.022.

[10] A. Afonso and C. Nunes, *Versão revista e aumentada PROBABILIDADES E ESTATÍSTICA Aplicações e Soluções em SPSS*. Évora: Universidade de Évora, 2019.

[11] A. A. Martins, "EGER - Estatística descritiva," Lisboa, 2021.

[12] B. J. F. Murteira, D. A. Muller, and K. Feridun Tukman, *Análise de Sucessões Cronológicas*, McGraw-Hill. 1993.

[13] A. Alexandra Martins, "EGER - Séries Temporais e Previsão," Lisboa, 2021.

[14] J. T. de A. Fernandes, "Análise de Séries Temporais no Domínio da Frequência. Importância do Periodograma neste contexto.," 2012.

[15] T. Alpuim, "1995 - Alpuim - Detecção de periodicidades em series temporais o Periodograma," Lisboa, 19, 1995.

[16] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th Edition. West Sussex: John Wiley & Sons, Ltd, 2011.

[17] A. Kassambara, *Multivariate Analysis I Practical Guide To Cluster Analysis in R Unsupervised Machine Learning*, Edition 1. 2017. [Online]. Available: http://www.sthda.com

[18] M. G. M. S. Cardoso, A. Martins, and J. Lagarto, "Combining various dissimilarity measures for clustering electricity market prices," 2020.

[19] P. Montero and J. A. Vilar, "TSclust: An R Package for Time Series Clustering," 2014. [Online]. Available: http://www.jstatsoft.org/

[20] J. Caiado, N. Crato, and D. Peña, "A periodogram-based metric for time series classification," Jun. 2006. doi: 10.1016/j.csda.2005.04.012.

[21] E. D. Ann Maharaj Pierpaolo and U. Jorge Caiado, *Time Series Clustering and Classification*. London: CRC Press, 2019.

[22] R. Hendron and C. Engebrecht, "Building America House Simulation Protocols (Revised)," Oak Ridge, 2010. [Online]. Available: http://www.osti.gov/bridge

[23] R. Hendron and C. Engebrecht, "Building America House Simulation Protocols (Revised)," 2010. [Online]. Available: http://www.osti.gov/bridge

[24] P. Northwest National Laboratory, "Building America Top Innovations Hall of Fame Profile – Building Science-Based Climate Maps," 2004. [Online]. Available: http://resourcecenter.pnl.gov/cocoon/morf/

Anexos

Anexo A

5.1.1 A.1 Análise do número indicado de grupos para 50 cidades

Escolha do número indicado de grupos (a verde), usando como critério o índice *Silhouette*.

Tabela A.1 – Determinação do número ideal de grupos para clustering de 50 cidades

Ordem preferencial	k = número de grupos	Índice *Silhouette*
1	2	1
2	3	0,598848675
3	4	0,398583373
4	5	0,313538534
5	6	0,066935614
6	7	0
7	8	0,026128735
8	9	0,054768316
9	10	0,04775488
10	11	0,034578935
11	12	0,056117596
14	13	0,049442402
12	14	0,085489509
15	15	0,090927707
13	16	0,073748579
16	17	0,083504395
17	18	0,062265875
19	19	0,036316444
18	20	0,010159234

5.1.2 A.2 Agrupamento detalhado das 50 cidades

Agrupamento detalhado, a cinzento os medoids de cada grupo:

Tabela A.2 – Agrupamento para 50 cidades

	Grupo	
	1	2
Estado_Cidade	AK_Anchorage	AL_Montgomery
	CO_Denver	AR_Little Rock
	CT_Hartford	AZ_Phoenix
	DE_Dover	CA_Sacramento
	IA_Des Moines	FL_Jacksonville
	ID_Boise	GA_Atlanta
	IL_Springfield	HI_Honolulu
	IN_Indianapolis	LA_Baton Rouge
	KS_Topeka	MS_Jackson
	KY_Fort Knox	NC_Winston-Salem
	MA_Boston	NV_Las Vegas
	MD_Baltimore	OK_Oklahoma
	ME_Augusta	SC_Columbia
	MI_Lansing	TN_Nashville
	MN_St.Paul	TX_Austin
	MO_Columbia	VA_Richmond
	MT_Helena	
	ND_Bismarck	
	NE_Columbus	
	NH_Concord	
	NJ_Trenton	
	NM_Santa Fe	
	NY_Albany	
	OH_Columbus	
	OR_Salem	
	PA_Harrisburg	
	RI_Providence	
	SD_Pierre	
	UT_Salt Lake City	
	VT_Montpelier	
	WA_Olympia	
	WI_Madison	
	WV_Beckley	
	WY_Cheyenne	

5.1.3 A.3 Análise dos grupos e regiões climáticas das 50 cidades

Análise das regiões climáticas para as 50 cidades, a cinzento os medoids de cada grupo.

Tabela A.3 – Agrupamento para 50 cidades e respetivas regiões climáticas

Estado_Cidade	Grupo	Categoria de Clima	Estado_Cidade	Grupo	Categoria de Clima
AK_Anchorage	1	N/D	AL_Montgomery	2	5
CO_Denver	1	1	AR_Little Rock	2	4
CT_Hartford	1	2	AZ_Phoenix	2	7
DE_Dover	1	4	CA_Sacramento	2	7
IA_Des Moines	1	2	FL_Jacksonville	2	5
ID_Boise	1	2	GA_Atlanta	2	4
IL_Springfield	1	2	HI_Honolulu	2	N/D
IN_Indianapolis	1	2	LA_Baton Rouge	2	5
KS_Topeka	1	4	MS_Jackson	2	4 ou 5
KY_Fort Knox	1	4	NC_Winston-Salem	2	4
MA_Boston	1	2	NV_Las Vegas	2	7
MD_Baltimore	1	4	OK_Oklahoma	2	4
ME_Augusta	1	1	SC_Columbia	2	4
MI_Lansing	1	1	TN_Nashville	2	4
MN_St.Paul	1	2	TX_Austin	2	5
MO_Columbia	1	4	VA_Richmond	2	4
MT_Helena	1	2			
ND_Bismarck	1	1			
NE_Columbus	1	2			
NH_Concord	1	2			
NJ_Trenton	1	4			
NM_Santa Fe	1	2			
NY_Albany	1	2			
OH_Columbus	1	2			
OR_Salem	1	3			
PA_Harrisburg	1	2			
RI_Providence	1	2			
SD_Pierre	1	2			
UT_Salt Lake City	1	2			
VT_Montpelier	1	2			
WA_Olympia	1	3			
WI_Madison	1	2			
WV_Beckley	1	2 ou 4			
WY_Cheyenne	1	2			

Anexo B

5.1.4 B.1 Medoides da análise das 50 cidades

B.1.1 Medoide do grupo 1 (Cidade de Providence do estado de Rhode Island)

- **ACF**

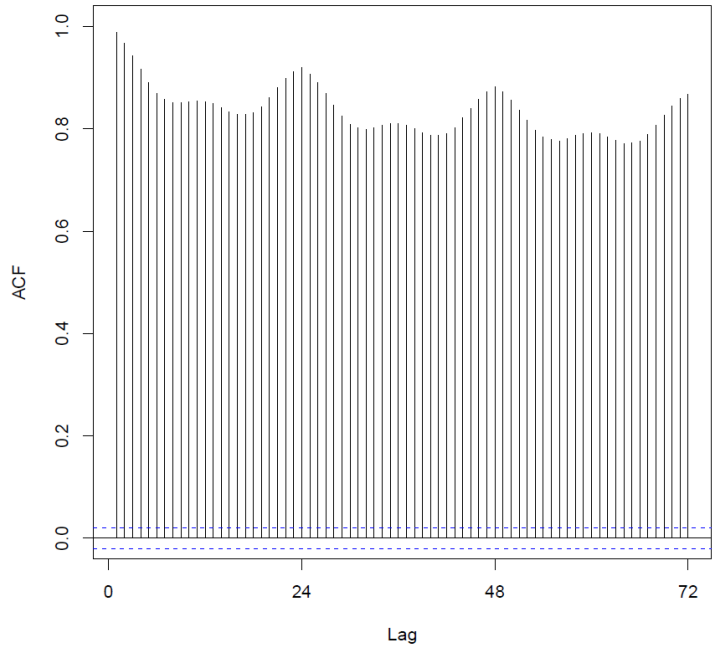


Figura B.1 – ACF de Providence

- **Histograma**

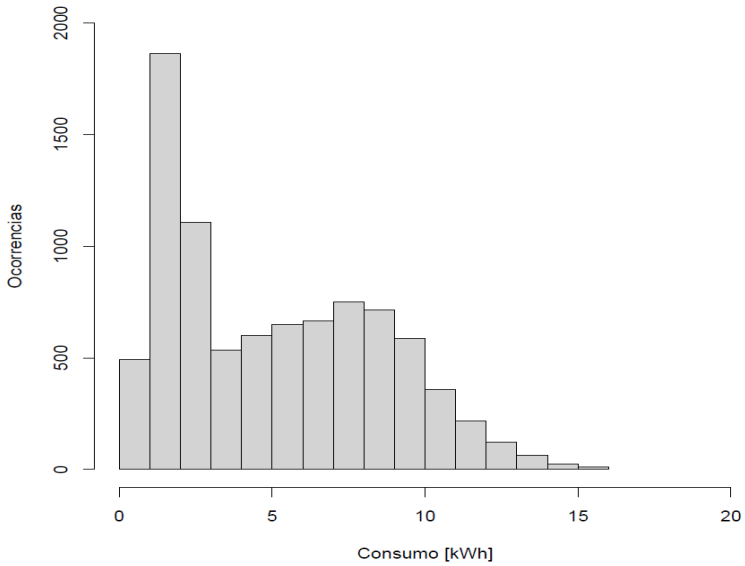


Figura B.2 – Histograma de Providence

- ***Outliers***

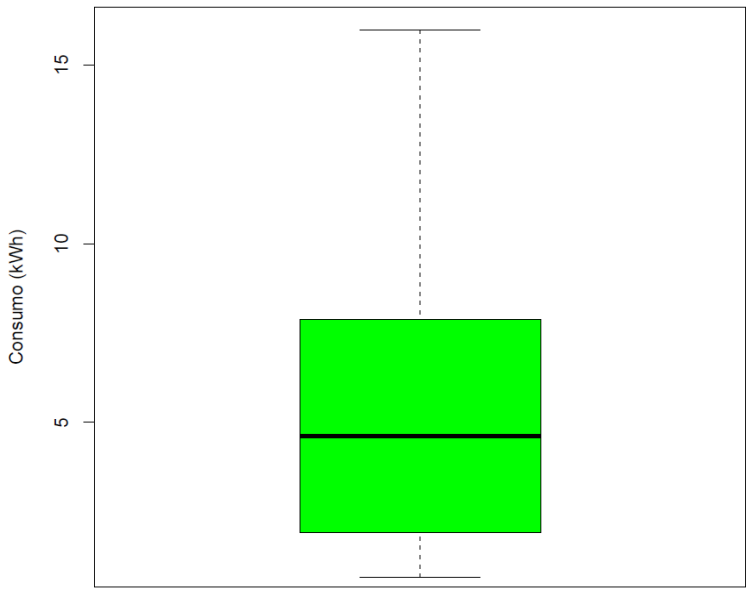


Figura B.3 – Outliers de Providence

- **Caracterização dos *outliers* no cronograma**

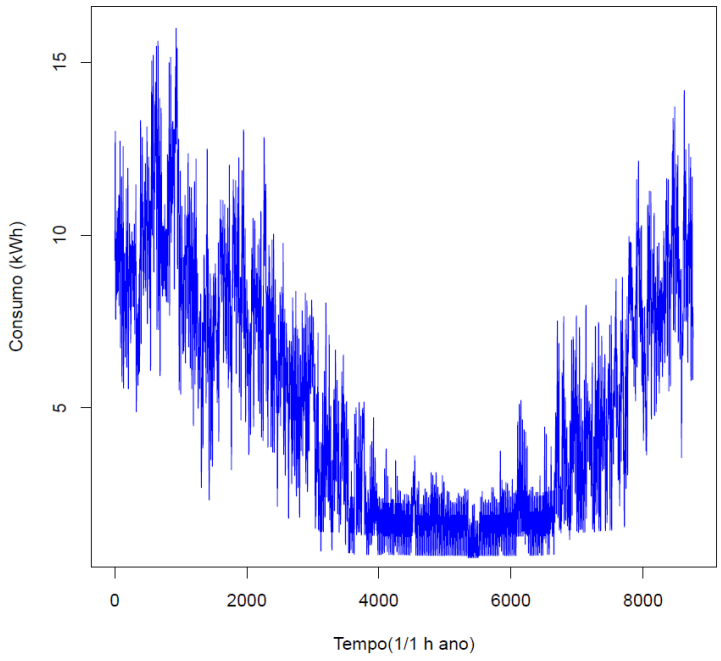


Figura B.4 – Caracterização e localização de outliers de Providence

- ***Outliers* severos**

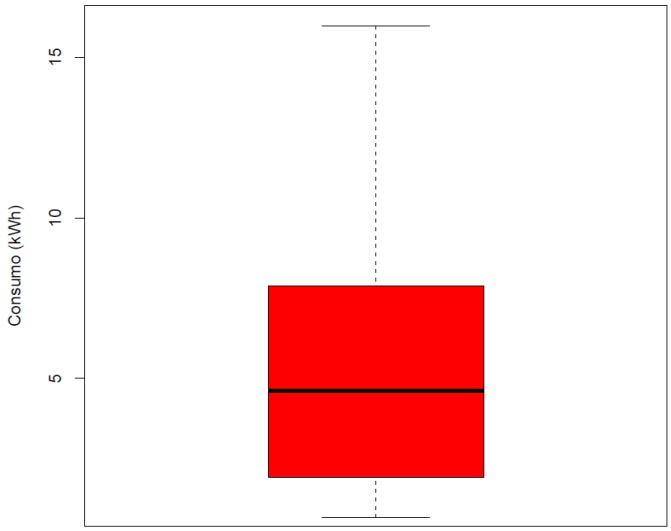


Figura B.5 – Outliers severos de Providence

- **Caracterização dos *outliers* no cronograma**

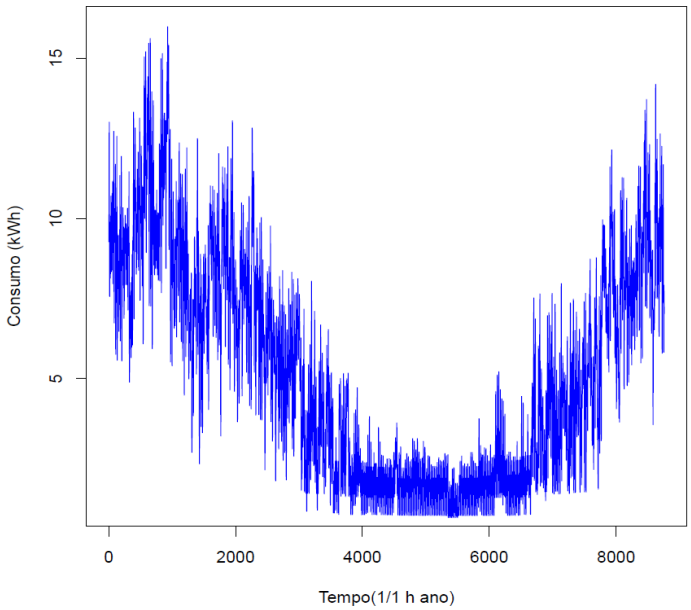


Figura B.6 – Caracterização e localização de outliers severos de Providence

- **Cronograma**

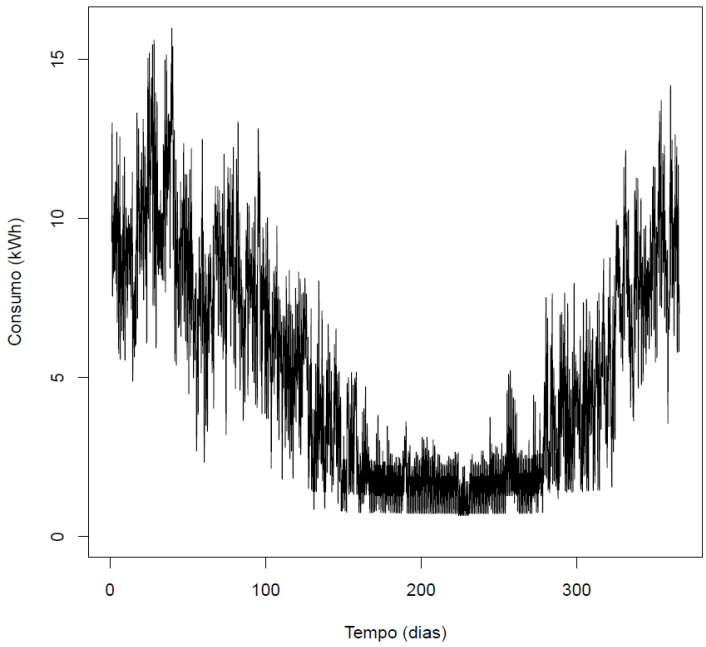


Figura B.7 – Cronograma de consumo de energia eléctrica e gás de Providence

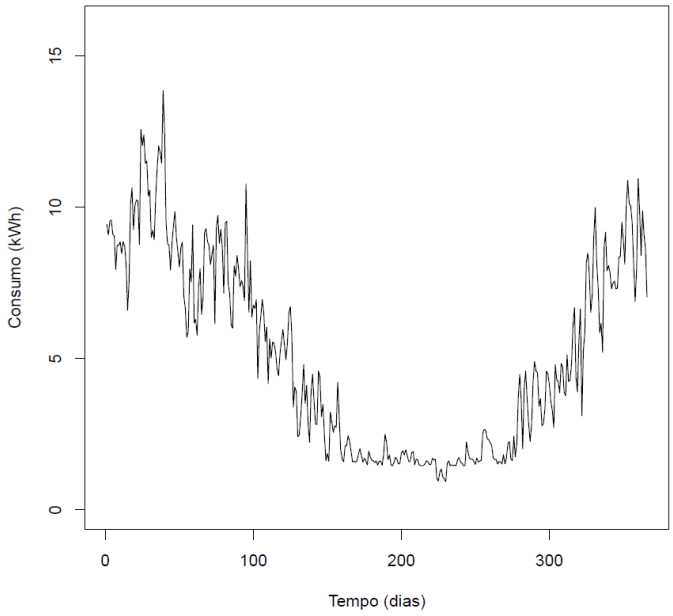


Figura B.8 – Cronograma alisado de consumo de energia elétrica e gás de Providence

B.1.2 Medoide do grupo 2 (Cidade de Sacramento do estado da Califórnia)

- **ACF**

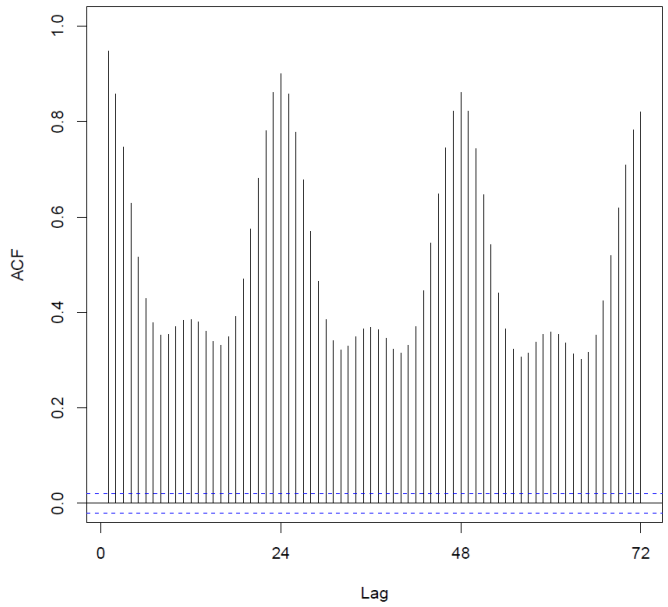


Figura B.9 – ACF de Sacramento

- **Histograma**

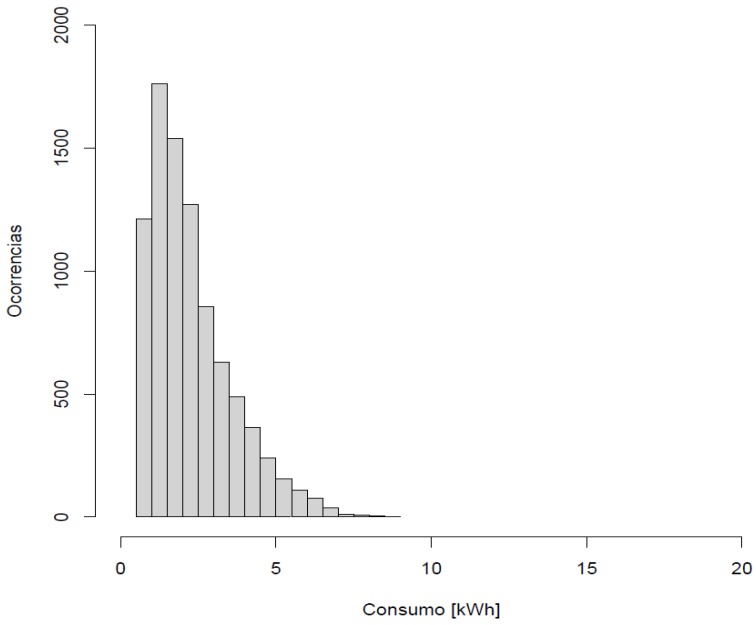


Figura B.10 – Histograma de Sacramento

- ***Outliers***

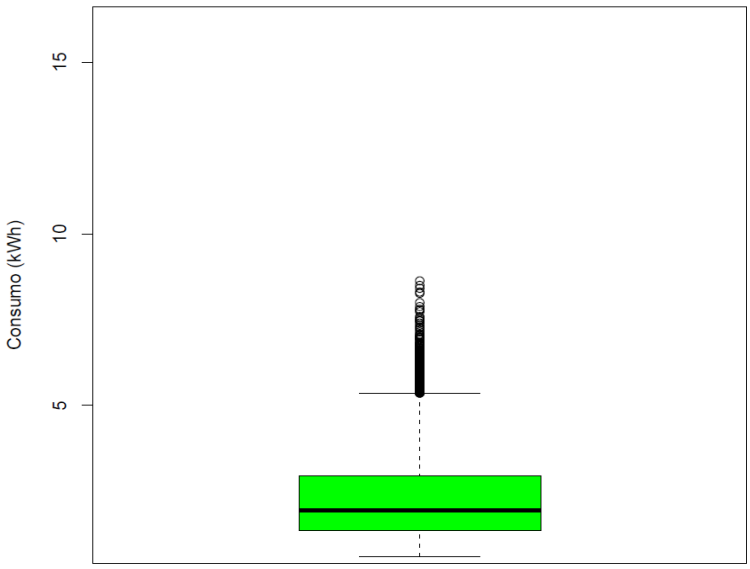


Figura B.11 – Outliers de Sacramento

- **Localização dos *outliers* no cronograma**

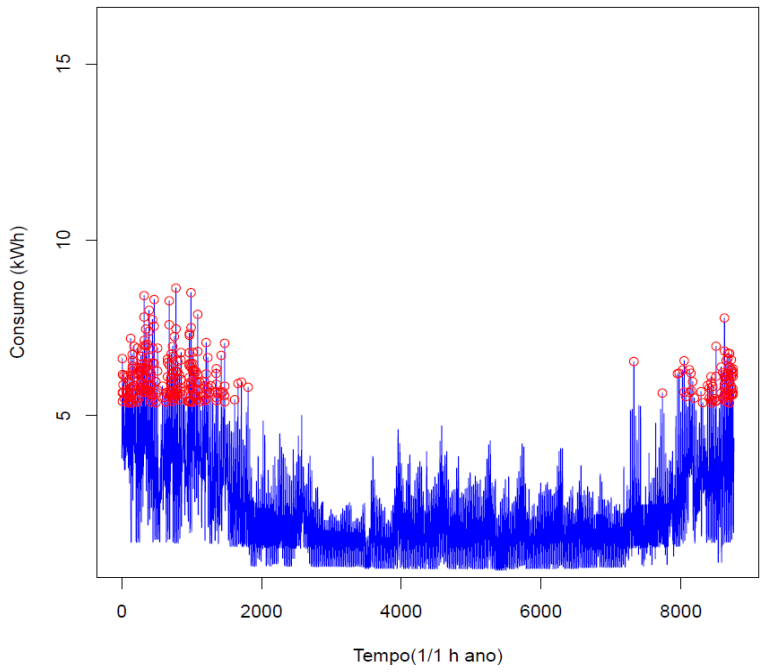


Figura B.12 – Caracterização e localização de outliers de Sacramento

- ***Outliers* severos**

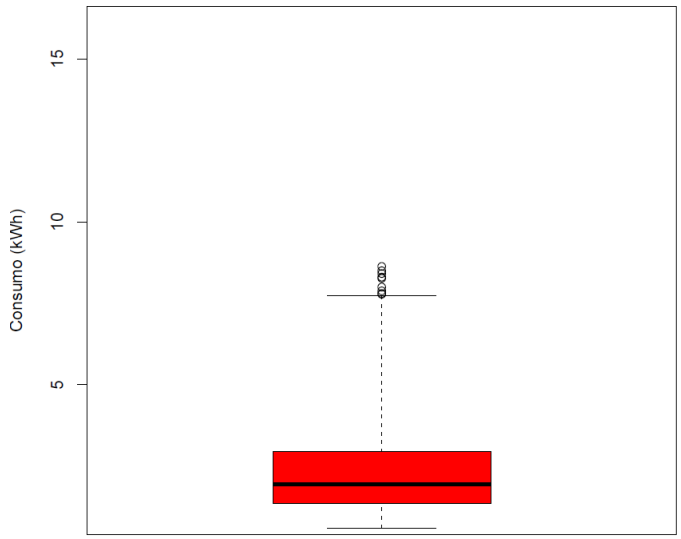


Figura B.13 – Outliers severos de Sacramento

- **Caracterização dos *outliers* serveros no cronograma**

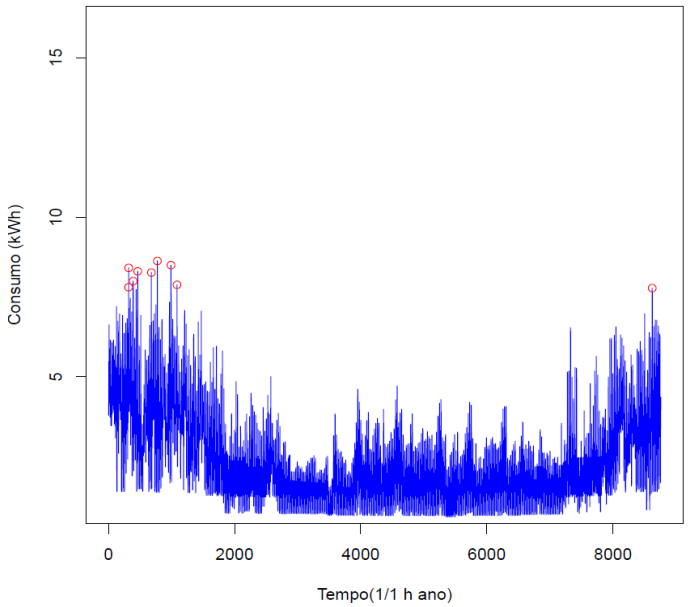


Figura B.14 – Caracterização e localização de outliers severos de Sacramento

- **Cronograma**

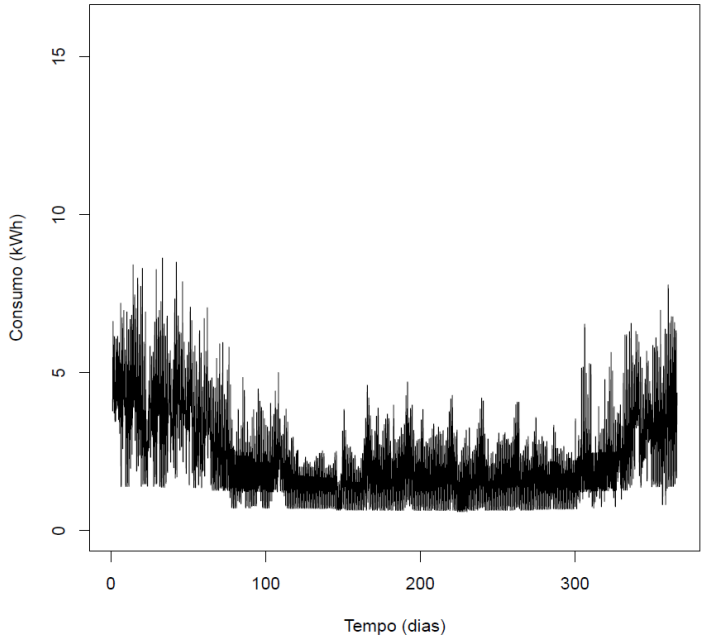


Figura B.15 – Cronograma de consumo de energia elétrica e gás de Sacramento

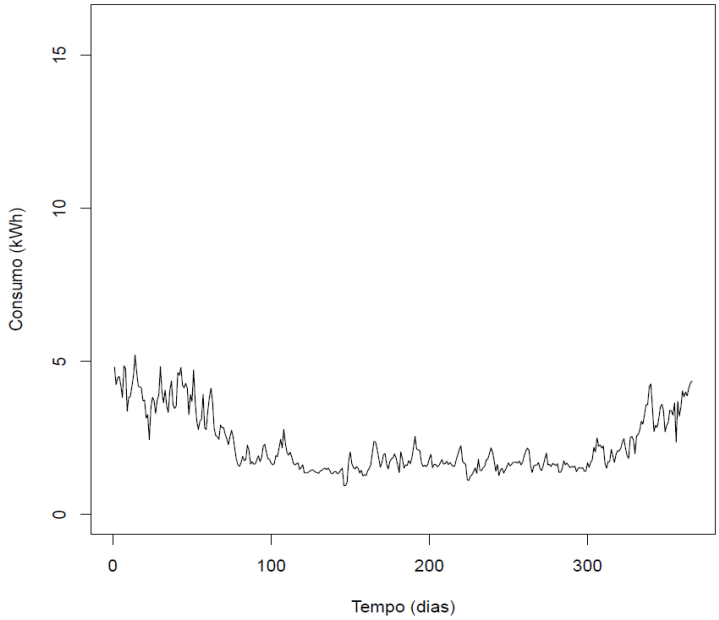


Figura B.16 – Cronograma de consumo de energia elétrica e gás alisado de Sacramento

Anexo C

5.1.5 C.1 Análise do número indicado de grupos para 936 cidades

Escolha do número indicado de grupos (a verde), usando como critério o índice *Silhouette*.

Tabela C.1 – Determinação do número ideal de grupos para clustering de 936 cidades

Ordem preferencial	k = número de grupos	Índice *Silhouette*
1	2	1
2	3	0,877448218
3	4	0,653199462
4	5	0,494816048
5	6	0,517485499
6	7	0,356225358
7	8	0,431339496
8	9	0,334208703
9	10	0,361796438
10	11	0,288796491
11	12	0,132445494
12	13	0,143914923
13	14	0,121808099
14	15	0,114349473
15	16	0,060765102
16	17	0,039986096
17	18	0,038123558
18	19	0,033046079
19	20	0

Anexo D

5.1.6 D.1 Análise dos medoides das 936 cidades

D1.1 Medoide grupo 1 (Cidade de Walla Walla do estado da Washington)

- **Cronograma**

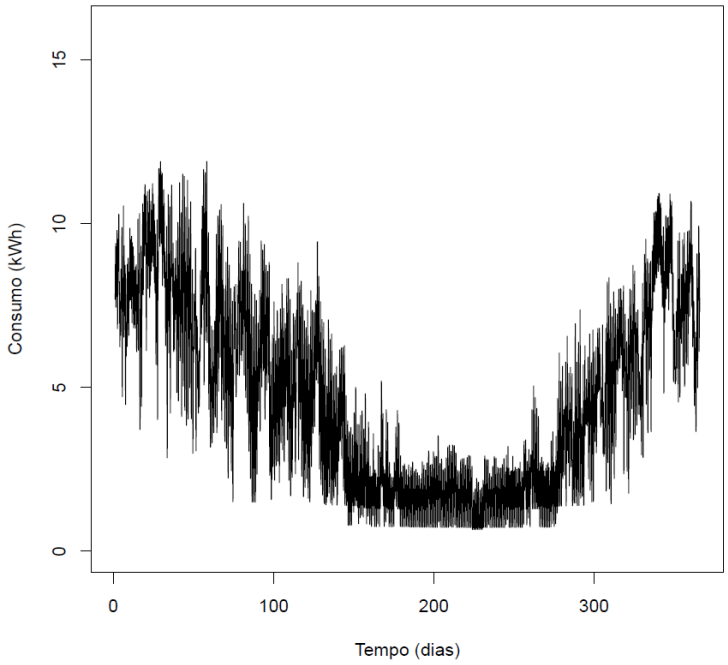


Figura D.1 – Cronograma de consumo de energia elétrica e gás de Walla Walla

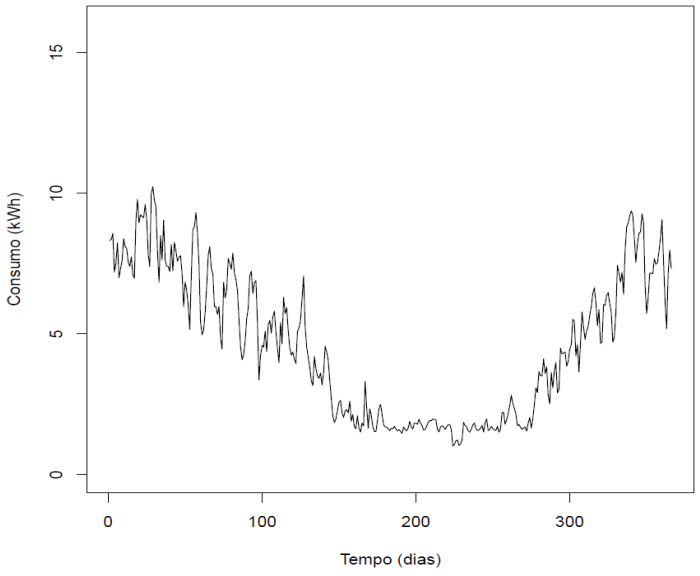


Figura D.2 – Cronograma alisado de consumo de Walla Walla

D1.2 Medoide grupo 2 (Cidade de Lake Charles do estado do Louisiana)

- **Cronograma**

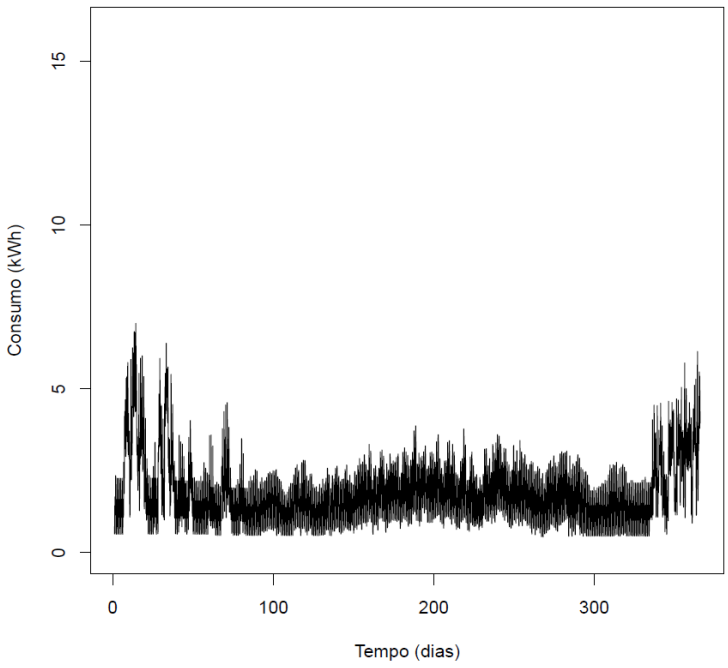


Figura D.3 – Cronograma de consumo de energia elétrica e gás de Lake Charles

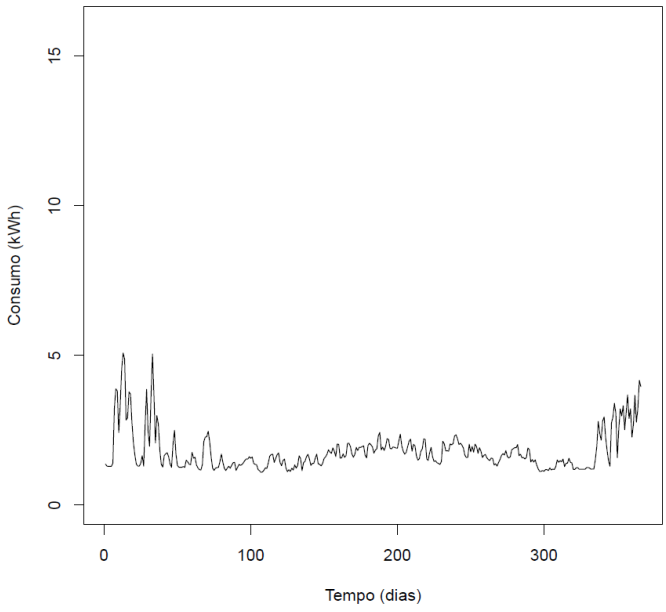


Figura D.4 – Cronograma alisado de consumo de energia elétrica e gás de Lake Charles

Anexos E

5.1.7 E.1 Código R

E1.1 Extração e tratamento de dados

Importar ficheiros csv de uma pasta para matriz newdata

```
path <- "C:/Users/Guest/Desktop/Tese/Proposta/BD/BASE/BASE/
"


files <- dir(path,pattern = "*.csv",all.files = F,
                    full.names = T, recursive = F, ignore.c
ase = F,
                    include.dirs = F, no.. = T)

newdata <- sapply(files,read_csv, simplify=TRUE) %>% bind_r
ows(.id = "id")
```

Renomear colunas id e Electricity:Facility

```
names(newdata)[1] <- "Area";

names(newdata)[3] <- "Electricity Consumption [kW]";
```

Converter coluna Date/Time de formato string para data (01/01 01:00:00)

```
newdata[2]<-as.POSIXct(newdata$`Date/Time`,format="%m/%d %H
:%M:%S",tz="GMT")
```

Retirar path da coluna id da coluna Area

```
newdata[1]<-str_replace_all(newdata$`Area`,"C:/Users/Guest/
Desktop/Tese/Proposta/BD/BASE/BASE/USA_"," ");

newdata[1]<-str_replace_all(newdata$`Area`,"_TMY3_BASE.csv"
," ");

newdata[1]<-sub("\\..*", "", newdata$`Area`) #####Nome da c
oluna Area  até ao "." - (estado e cidade)
```

Criar coluna com a soma de Eletricidade e Gás

```
sum <- newdata[3]+newdata[4]
```

Guardar matriz dados com dados da matriz newdata com a soma de eletricidade e gás

```
dados <- cbind(newdata[,1:2],sum)
```

Validar dados vazios

```
Checkflawdata<-is.na(dados);
```

Validar dados em falta

```
missingdata<-dados[!complete.cases(dados),];
```

Guardar na dataframe finaldata sem os dados omissos

```
finaldata <- na.omit(dados);
```

Número de ficheiros analisados

```
num_files <- length(list.files(path))
```

Obter nome das colunas de dados (nome das cidades e respetivo estado) sem a extensão csv

```
nomecolunas <- files
nomecolunas<-str_replace_all(nomecolunas,"C:/Users/Guest/Desktop/Tese/Proposta/BD/BASE/BASE/USA_"," ");

nomecolunas<-str_replace_all(nomecolunas,"_TMY3_BASE.csv","");

nomecolunas<-str_sub(nomecolunas,start = 1,end = 20);
```

Importar dados para a dataframe dadosnovos usando a matriz newdata e ordenar as colunas por cidades

```
dadosnovos<-data.frame(data=newdata$`Date/Time`[1:8760])

aux<-as.data.frame(matrix(finaldata$`Electricity Consumption [kW]`,nrow=8760))

dadosnovos<-cbind(dadosnovos,aux)

names(dadosnovos)<-c("data",nomecolunas)
```

E1.2 Obtenção resultados

- Histograma

```
for (i in 1:num_files) {

  hist(dadosnovos[,i+1],main=c(" Histograma",names(dadosnov
os)[i+1]),xlab = "Consumo [kWh]",ylab = "Ocorrências")
}
```

- *Outliers*, caracterização e localização

```
for (i in 1:num_files) {
boxplot(dadosnovos[,i+1],col="green",range=1.5)
outliers<-boxplot.stats(dadosnovos[,i+1],coef = 1.5)
print(outliers);
outliers$out
title(main = c("Outliers" ,"  ", nomecolunas[i]))

aux<-dadosnovos[,i+1] %in% outliers$out
x<-which(aux==TRUE)
y<-outliers$out
x1<-c(1:length(dadosnovos[,i+1]))
plot(x1,dadosnovos[,i+1],type="l",col="blue",ylab = "Consum
o (kWh)",xlab = "Tempo(1/1 h ano)")
points(x,dadosnovos[,i+1][x],col="red")

title(main = c(" Caracterização e localização - Outliers" ,
"  ", nomecolunas[i]))
}
```

- *Outliers* severos, caracterização e localização

```
for (i in 1:num_files) {

  boxplot(dadosnovos[,i+1],col="red",range=3)
  outliers_severos<-boxplot.stats(dadosnovos[,i+1],coef = 3
)
  print(outliers_severos);
  outliers_severos$out
  title(main = c("Outliers Severos" ,"  ", nomecolunas[i]))


  aux<-dadosnovos[,i+1] %in% outliers_severos$out
  x<-which(aux==TRUE)
  y<-outliers_severos$out
  x1<-c(1:length(dadosnovos[,i+1]))
  plot(x1,dadosnovos[,i+1],type="l",col="blue",ylab = "Cons
```

```
umo (kWh",xlab = "Tempo(1/1 h ano)")
  points(x,dadosnovos[,i+1][x],col="red")

  title(main = c(" Caracterização e localização - Outliers
Severos" ,"  ", nomecolunas[i]))

}
```

- Consumo de energia eléctrica e gás todos as cidades por coluna – Serie Temporal

```
consumo.ts<-ts(dadosnovos[,2:(num_files+1)],start=c(1,1),fre
quency = 24)
```

- Caracterização da serie temporal por cidade analisada

```
for (i in 1:num_files){

  plot(consumo.ts[,i],main=paste((" - Consumo eletricidade
e gás - "),nomecolunas[i]), xlab="Tempo (dias)" , ylab=" Co
nsumo (kWh)")
    #ggtitle(paste(" - Consumo eletricidade - "),nomecoluna
s[i])+ xlab("Tempo (dias)") + ylab("Potencia (kW)")

}
```

- Caracterização da serie temporal para os medoides

```
autoplot(consumo.ts) +
  ggtitle(" - Consumo eletricidade e gás - Medoides")+ xlab
("Tempo (dias)") + ylab("Potencia (kWh)")
```

- Função Autocorrelação (ACF)

```
for (i in 1:num_files){

autoc<-Acf(consumo.ts[,i])
title(main = c(" Autocorrelação" ," ", nomecolunas[i]))

}
```

E1.3 Algoritmo de *clustering*

- Função que calcula as distâncias para o *clustering*

```
distancias<-function(dadosparadiss,w){

  correlacoes<-cor(dadosparadiss)
  d<-as.dist(sqrt((1-correlacoes)/2))
  mx<-max(d)
  mn<-min(d)
  d_PEARSON2<-(d-mn)/(mx-mn)
  remove(d)
  remove(mx)
  remove(mn)

  d<-diss(dadosparadiss,"EUCL")
  mx<-max(d)
  mn<-min(d)
  d_EUCLID<-(d-mn)/(mx-mn)
  remove(d)
  remove(mx)
  remove(mn)

  d<-diss(dadosparadiss,"PER")
  mx<-max(d)
  mn<-min(d)
  d_PERIOD<-(d-mn)/(mx-mn)
  remove(d)
  remove(mx)
  remove(mn)

  d<-diss(dadosparadiss,"ACF",lag.max=70) # ACF
  mx<-max(d)
  mn<-min(d)
  d_ACF<-(d-mn)/(mx-mn)
  remove(d)
  remove(mx)
  remove(mn)

  d<-w[1]*d_PEARSON2+w[2]*d_EUCLID+w[3]*d_PERIOD+w[4]*d_ACF

  return(d)

} # final funcao distancias
```

- Função para *clustering*

```
clustering <-function(dados.clust,w){

  # inputs para esta função:
  # dados.clust - dados para agrupamento onde cada coluna c
orresponde a uma cidade
  # w - pesos para o calculo das distancias

  dados.clust.ts<-ts(dados.clust,start=1, frequency=1, clas
s="mts",names=names(dados.clust))
  # coloca os dados dados.clust num objeto ts onde cada col
una corresponde a uma cidade

  D<-distancias(dados.clust.ts,w)
  # calculo da matriz de distancias

  #selecao n de clusters
  n.max.grupos<-min(c(ncol(dados.clust)-1,20))

  result <-matrix(c(rep(0,4*n.max.grupos)),n.max.grupos,4)
  colnames(result)<-c("k","avgsilwidth","ch","dunn2")

  for (k in 2:n.max.grupos)
  {
    Cdat<-pam(D, k, diss=TRUE) #clustering
    Cstat<-cluster.stats(D, Cdat$clustering , Silhouette =
TRUE,wgap=TRUE,aggregateonly=TRUE )
    result[k,1]<-k
    result[k,2]<-Cstat$avg.silwidth # or Cdat$silinfo$avg.w
idth#
    result[k,3]<- Cstat$ch
    result[k,4]<-Cstat$dunn2
  }

  result<-result[-1,] # retira a primeira linha para k=1
  result.norm<-cbind(result[,1],data.Normalization(result[,
2:4],type="n4",normalization = "column"))
  colnames(result.norm)<-c("k","n_avgsilwidth","n_ch","n_du
nn2")
  # na primeira coluna: k; nas restantes colunas indices no
rmalizados min/range, ou seja, maior valor<->melhor
  total.resul<-apply(result.norm[,2:4],1,sum)
  final<-cbind(result.norm,total.resul)

 colnames(final)<-c("k","n_avgsilwidth","n_ch","n_dunn2","t
otal_norm")
  final<- final[order(final[,5],decreasing = TRUE),]
  # para cada k: valor da soma dos indices normalizados;ord
enado por ordem crecente;
  # menor valor da soma dos indices (primeiro valor) <-> me
lhor k
```

```
  clust<-pam(D, k=final[1,1], diss=TRUE) # faz o clustering
para o melhor k
  ##clust<-pam(D, k=4, diss=TRUE) # faz o clustering  para
o  k=4 (4 grupos)
  ##clust<-pam(D, k=5, diss=TRUE) # faz o clustering  para
o  k=5 (5 grupos)
  ##clust<-pam(D, k=2, diss=TRUE) # faz o clustering  para
o  k=2 (2 grupos)
  ##clust<-pam(D, k=14, diss=TRUE) # faz o clustering  para
o  k=14 (14 grupos)

  res<-list(clust,final,D)

  return(res)
}   # fim função clustering
```

E1.4 Obtenção dos resultados do *clustering*

```
dados.clust<-dadosnovos[,2:ncol(dadosnovos)]

## Error in eval(expr, envir, enclos): object 'dadosnovos' not found

w<-c(0.25,0.25,0.25,0.25)   #### ponderação de cada distanc
ia: Pearson,Euclidiada,ACF e Periodograma
tic()

res.clust<-clustering (dadosnovos[,2:ncol(dadosnovos)],w) #
######### Resultado do clustering

toc()
```

Localização dos consumos das cidades, divididos por grupos de consumo

```
res.clust[[1]]$medoids ## Localização dos k-medoids

xx<-res.clust[[1]]$clustering  ### Clustering  efectuado

ind_coesao_separacao<-res.clust[[2]] ### indices de coesão_
separação
```

E1.5 Forma alisada por dia dos cronogramas das cidades

```
dadosnovos <- dadosnovos %>%
  mutate(dia=format(dadosnovos$data,"%Y-%m-%d"))

medias.dia<-dadosnovos %>%
  group_by(dia) %>%
  summarise_at(vars(2:3),mean)

######## Consumo medio de energia elétrica e gás todos os f
icheiros por coluna #######

consumomedia.ts<-ts(medias.dia[,2:(num_files+1)],start=c(1,1
),frequency = 1)

######## Caracterizacao da Serie Temporal por cidade (media
) #########

pdf("Serie Temporal_Media.pdf")

for (i in 1:num_files){

 plot(consumomedia.ts[,i],main=paste((" - Consumo eletricid
ade e gás - "),nomecolunas[i]), xlab="Tempo (dias)" , ylab=
" Consumo (kWh)")
}

dev.off()

######## Caracterizacao da Serie Temporal - todos #########

pdf("Serie Temporal_Media - todos.pdf")

autoplot(consumomedia.ts) +
  ggtitle(" - Consumo eletricidade e gás - Medoides")+ xlab
("Tempo (dias)") + ylab("Consumo (kWh)")

dev.off()
```