

Mestrado em Engenharia Informática
Estágio
Relatório Final

Inteligência Computacional para a Gestão Eficiente de Energia e Conforto no Domínio das Casas Inteligentes

Marta Luísa Nunes Canelas Pais
mlpais@student.dei.uc.pt

Orientador (DEI):
Professor Doutor António Dourado Pereira Correia

Orientador (Whitesmith):
Miguel Gonçalves Tavares

Data: 1 de julho de 2016



**FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA**
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Mestrado em Engenharia Informática
Estágio
Relatório Final

Inteligência Computacional para a Gestão Eficiente de Energia e Conforto no Domínio das Casas Inteligentes

Marta Luísa Nunes Canelas Pais
mlpais@student.dei.uc.pt

Orientador (DEI):
Professor Doutor António Dourado Pereira Correia

Orientador (Whitesmith):
Miguel Gonçalves Tavares

Júri Arguente:
Professor Doutor Joel Perdiz Arrais

Júri Vogal:
Professor Doutor Luís Miguel M. L. Macedo

Data: 1 de julho de 2016



**FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA**
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Agradecimentos

Dedico este espaço a todos aqueles que me proporcionaram as condições ideais para a concretização deste trabalho, uma vez que sem o seu apoio tudo isto teria sido impossível. Não sendo exequível a nomeação de todos, há alguns a quem não posso deixar de expressar o meu reconhecimento e sinceros agradecimentos.

Ao Professor Doutor António Dourado que, ao partilhar todo o seu conhecimento científico e experiência, me guiou ao longo deste processo.

Ao meu orientador Miguel Tavares, cujas paciência e exigência fizeram com que explorasse ao máximo as minhas capacidades.

À empresa Whitesmith e em especial ao CEO Rafael Jegundo, por me terem acolhido e proposto a execução de um projeto cativante e enriquecedor.

Aos elementos do júri, Professor Doutor Joel Perdiz Arrais e Professor Doutor Luís Miguel M. L. Macedo, por todas as sugestões de aperfeiçoamento e pela avaliação crítica e construtiva para o desenvolvimento deste projeto.

A todos os meus professores do Departamento de Engenharia Informática, que me formaram e despertaram em mim uma constante curiosidade por todos estes assuntos, e aos colegas que me acompanharam ao longo deste percurso.

Agradeço, em especial, à minha família que me sempre me apoiou nos momentos mais difíceis e se juntou a mim na celebração dos momentos de maior alegria.

Muito obrigada.

Resumo

A previsão do consumo energético é uma necessidade para a gestão eficiente das *smart grids*, dos programas de *demand response* e da gestão da eficiência energética de uma habitação. Isto deve-se ao facto de, ao obter informação sobre o eventual estado futuro do sistema elétrico, ser possível tomar medidas preventivas que permitam ao produtor de energia uma melhor gestão da produção. Numa outra perspetiva, e sob o ponto de vista do consumidor, é possível identificar padrões de consumo e planear o consumo energético eficientemente, beneficiando de menores custos.

O presente trabalho de estágio, a decorrer na empresa Whitesmith, terá como objetivo contribuir para o desenvolvimento da plataforma Unplugg, especificamente para a definição e elaboração de um modelo de previsão, *clustering* de séries temporais e deteção de *change-points*. A literatura aponta no sentido de que os modelos ARMA, ARIMA, SARIMA e Holt-Winters apresentam bom desempenho na previsão do consumo a curto e médio prazo. Para a previsão do consumo energético será, portanto, utilizada uma abordagem baseada em modelos de séries temporais uni-variados e multi-variados que permitem a previsão do consumo energético residencial. Serão comparados modelos, janelas de histórico e previsão, aplicação de filtros, transformações, utilização de diferentes tipos de histórico e, ainda, a utilização de variáveis exógenas de fatores ambientais. Adicionalmente serão efetuados *clusterings* de séries temporais para várias granularidades do consumo de séries temporais normalizadas e não-normalizadas para o efeito de previsão em bloco e da tipificação de padrões de consumo. Para tal, ao longo do primeiro semestre foi efetuada uma investigação extensiva sobre as temáticas relacionadas com o problema a solucionar, bem como a exploração das soluções HEM já implementadas. No segundo semestre foram analisados e caracterizados os dados provenientes de clientes reais, implementadas as soluções estudadas, efetuados testes e obtidos resultados com o objetivo da determinação da configuração adequada a uma previsão eficiente individual e em bloco e do desenvolvimento de uma adaptação dinâmica às rotinas dos consumidores residenciais. O resultado obtido deste trabalho foi uma configuração com base no modelo SARIMA cujo valor médio de MAPE é de cerca de 33.8%, 34.1% e 20.7%, para as granularidades horária, octa-horária e diária, respetivamente, que está presentemente a ser posto em produção num sistema que irá analisar 10 000 habitações. O modelo desenvolvido neste trabalho será utilizado como validação para novas iterações com sucessivas melhorias do mesmo.

KeyWords: ARMA, ARIMA, Clustering, Consumo de eletricidade, CUSUM, Deteção de change-points, Diebold-Mariano, DTW, Holt-Winters, Inferência Bayesiana, K-Means, Métodos de previsão, MSE, SARIMA, SARIMAX, Série temporal

Abstract

Electricity consumption forecasts are an essential part of an efficient management of smart grids, demand response programs and household consumption. This is due to the fact that information about the future state of the electric system aids in the planning of energy production allowing the producer a more efficient energy generation approach. A end-consumer will also benefit from this information as it will allow him to identify electricity consumption patterns that, with the additional information of future dynamic electricity prices, will promote a more efficient management of electricity consumption which may result in a cheaper electricity bill.

This internship, at Whitesmith, will contribute to the Unplugg platform, namely with the development of a forecasting model, clustering of time series and change-point detection. The state of the art shows that the models ARMA, ARIMA, SARIMA and Holt-Winters have proved to be very efficient in short to medium term forecasting. Forecasting of residential electricity consumption will be the main subject of this report and the aforementioned models will be compared to determine the best model for this kind of problem. Factors that influence forecasting performance, such as the sizes of historic and forecasting windows, the incorporation of exogenous variables of environment conditions, filters, transformations and historic types will be studied in this project. The clustering of normalized and non-normalized timeseries will also be addressed as it will allow bulk forecasting of time series and the segmentation of types of clients by their consumption patterns. Therefore an extensive investigation has been made regarding the context of Home Energy Management, in which this internship is integrated, as well as the state of the art. Many solutions were implemented, compared and combined in order to determine the best configuration to create an efficient forecasting model that adapts to the residential consumer's dynamics. A dataset with real consumption data was the focal point of this study. It follows that the best configuration is based on a SARIMA model with the average MAPE value of about 33.8 %, 34.1 % and 20.7 % for hourly, eight-hourly and daily data, respectively. This configuration is currently being used by a system that will evaluate 10 000 dwellings. The model developed in this study will be used as validation for new iterations with successive improvements of the system.

KeyWords: ARMA, ARIMA, Bayesian inference, Change-points detection, Clustering, CUSUM, Diebold-Mariano, DTW, Electricity consumption, Forecasting methods, Holt-Winters, K-Means, MSE, SARIMA, SARIMAX, Time series

Conteúdo

1	Introdução	15
1.1	Contexto	18
1.1.1	Sistemas de <i>Home Energy Management</i> (HEM)	18
1.1.2	<i>Smart Homes</i>	18
1.1.3	<i>Smart meters</i> , <i>power meters</i> e <i>smart plugs</i>	18
1.1.4	Sensores e atuadores	19
1.1.5	<i>Smart Grid</i>	19
1.2	Definição do problema	20
1.2.1	Previsão do consumo agregado	20
1.2.2	Janela de previsão	21
1.2.3	Clustering de séries temporais	21
1.2.4	Alterações nos padrões	21
1.2.5	Unplugg	22
2	Estado da arte	25
2.1	Previsão do consumo energético	25
2.2	Clustering de séries temporais	26
2.3	Deteção de pontos de mudança	27
2.4	Desagregação do consumo	27
2.5	Reconhecimento de atividades	28
3	Conceitos teóricos	29
3.1	Séries temporais	29
3.2	Análise de séries temporais	29
3.3	Caracterização de séries temporais	30
3.3.1	Séries estacionárias <i>vs</i> séries não estacionárias	32
3.4	Modelos de séries temporais	35
3.4.1	Modelos não-sazonais	35
3.4.1.1	ARMA	36
3.4.1.2	ARIMA	38
3.4.2	Modelos Sazonais	39
3.4.2.1	Sazonalidade	39
3.4.2.2	SARIMA	40
3.4.2.3	Suavização Exponencial	41
3.5	Clustering	42
3.5.1	K-Means	42
3.5.2	Dinamic Time Warping	43
3.6	Análise de Change-points	44
3.6.1	Hipótese	45
3.6.2	Deteção	46
3.6.2.1	CUSUM e MSE	46
3.6.2.2	Inferência Bayesiana	48
3.7	Avaliação da previsão	51

4	Procedimentos Experimentais	53
5	Resultados	59
5.1	Pré-processamento	59
5.2	Caracterização do conjunto de dados	62
5.2.1	Consumo de clientes mistos	62
5.2.2	Consumo doméstico	65
5.3	Análise de séries temporais	67
5.3.1	Sazonalidade	67
5.3.2	Seleção do modelo	68
5.3.3	Escolha das janelas de histórico e previsão	71
5.3.3.1	Consumo horário	71
5.3.3.2	Consumo octa-horário	72
5.3.3.3	Consumo diário	73
5.3.4	Influência de vários fatores na previsão	74
5.3.4.1	Temperatura e humidade como variáveis exógenas	74
5.3.4.2	Tipos de histórico	77
5.3.4.3	Filtros	79
5.3.4.4	Transformações	80
5.4	Clustering	83
5.4.1	Clustering de séries temporais não-normalizadas	83
5.4.2	Clustering de séries temporais normalizadas	89
5.5	Pontos de mudança	93
6	Conclusões	97
7	Trabalho futuro	99
	Apêndices	105
A	Clustering de séries de consumo normalizadas	105
A.1	Clustering de séries normalizadas de consumo octa-horário	105
A.2	Clustering de séries normalizadas de consumo diário	106

Lista de abreviaturas

ACF	Função de auto-correlação (<i>Autocorrelation Function</i>)
ACVF	Função de auto-covariância (<i>Autocovariance Function</i>)
ADF	Dickey-Fuller ampliado (<i>Augmented Dickey-Fuller</i>)
AIC	Critério de informação de Akaike (<i>Akaike Information Criterion</i>)
AR	Auto-regressivo (<i>Auto-regressive</i>)
ARMA	Média móvel auto-regressiva (<i>Auto-Regressive Moving Average</i>)
ARIMA	Média móvel integrada auto-regressiva (<i>Autoregressive Integrated Moving Average</i>)
ARIMAX	Média móvel integrada auto-regressiva com variáveis exógenas
CUSUM	Soma cumulativa (<i>Cumulative Sum</i>)
DR	Procura e resposta à procura (<i>Demand Response</i>)
DTW	Distância baseada em alinhamento temporal dinâmico (<i>Dynamic time warping distance</i>)
EDP	Energias de Portugal
FIM	Medida da Independência Funcional (<i>Functional Independence Measure</i>)
HEM	Gestão da energia doméstica (<i>Home energy management</i>)
HVAC	Ventilação, aquecimento e ar-condicionado (<i>Heating ventilation and air-conditioning</i>)
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LPT	Teste estatístico de rácio de verosimilhança (<i>Likelihood-ratio procedure test</i>)
MA	Média móvel (<i>Moving average</i>)
MAE	Erro absoluto médio (<i>Mean Absolute Error</i>)
MASE	Erro absoluto médio escalado (<i>Mean Absolute Scaled Error</i>)
MAPE	Erro percentual absoluto médio (<i>Mean Absolute Percentage Error</i>)
MAD	Desvio absoluto médio (<i>Mean Absolute Deviation</i>)
MSE	Erro quadrático médio (<i>Mean Square Error</i>)
NILD	Deteção de cargas não intrusiva (<i>Non-intrusive load detection</i>)
NILI	Identificação de cargas não intrusiva (<i>Non-intrusive load identification</i>)
PACF	Função parcial de auto-correlação (<i>Partial autocorrelation function</i>)
RMSE	Raiz do erro quadrático médio (<i>Root mean square error</i>)
SARIMA	Média móvel integrada auto-regressiva sazonal (<i>Seasonal Autoregressive Integrated Moving Average</i>)
SARIMAX	Média móvel integrada auto-regressiva sazonal com variáveis exógenas
WS	Whitesmith

Lista de Figuras

1	Exemplos de correlogramas.[1]	38
2	Correlograma de uma série temporal.	40
3	Grelha utilizada no algoritmo DTW.[2]	44
4	Comparação entre os cálculos da distância Euclideana e DTW.[2]	44
5	Exemplos de séries temporais com diferentes tipos de <i>change-points</i> .[3]	45
6	Soma cumulativa de uma série temporal com a indicação da linha de base ($y = 0$) e do máximo absoluto.	47
7	<i>Change-Points</i> detetados pelos estimadores CUSUM e MSE.	47
8	Distribuições das variáveis λ_1 , λ_2 e τ	49
9	Sobreposição do consumo esperado e do consumo real, com o <i>change-point</i> detetado assinalado a amarelo.	50
10	Exemplo de uma série temporal com as granularidades horária, octa-horária e diária de consumo energético.	59
11	Exemplo de limites STD e MAD na remoção de <i>outliers</i> de um série temporal. . . .	60
12	Preenchimento de dois valores em falta de uma série temporal de consumo energético de granularidade 20 minutos.	61
13	Gráfico circular de categorias.	62
14	Histograma das potências contratadas para todos os tipos num conjunto de 1000 consumidores.	63
15	Histogramas da média e desvio padrão do consumo energético num conjunto de 1000 consumidores.	64
16	Histograma das potências contratadas de um conjunto de 100 consumidores domésticos. .	65
17	Histogramas da média e desvio padrão de um conjunto de 100 consumidores domésticos para séries temporais com e sem <i>outliers</i>	66
18	Exemplo da previsão de uma série temporal utilizando os modelos ARMA, ARIMA, SARIMA, Holt-Winters aditivo e Holt-Winters multiplicativo.	68
19	Sobreposição dos valores de temperatura e humidade para vários distritos na primeira semana do mês de Março de 2016.	74
20	Histogramas da correlação entre a temperatura e o consumo energético de 100 séries temporais de consumo doméstico.	75
21	Histogramas da correlação entre a humidade e o consumo energético de 100 séries temporais de consumo doméstico.	75
22	Sobreposição dos dois tipos de histórico.	77
23	Sobreposição de uma série temporal de consumo horário à qual foram aplicados filtros com a série original.	79
24	Apresentação de uma série de consumo energético horário e resultado da aplicação de transformações na série.	80
25	Valores de \mathcal{E} para o consumo horário não-normalizado para cada valor de k	83
26	Centróides de <i>clusters</i> para um conjunto de 100 séries temporais de consumo horário doméstico.	84
27	Valores de \mathcal{E} para o consumo octa-horário não-normalizado para cada valor de k . . .	85
28	Centróides de <i>clusters</i> para um conjunto de 100 séries temporais de consumo octa-horário doméstico.	86
29	Valores de \mathcal{E} para o consumo diário não-normalizado para cada valor de k	87

30	Centróides de <i>clusters</i> para um conjunto de 100 séries temporais de consumo diário doméstico.	88
31	Valores de \mathcal{E} para o consumo horário não-normalizado para cada valor de k	89
32	Centróides dos <i>clusters</i> para um conjunto de 100 séries normalizadas de consumo horário doméstico.	90
33	<i>Clusters</i> do consumo horário normalizado com destaque do centróide.	91
34	<i>Change-points</i> detetadas na 1 ^a semana.	93
35	<i>Change-points</i> detetados na 2 ^a semana.	93
36	<i>Change-points</i> detetados na 3 ^a semana.	94
37	<i>Change-points</i> detetados na 4 ^a semana.	94
38	<i>Change-points</i> detetados na 5 ^a semana.	95
39	Previsões com histórico prévio ao primeiro <i>change-point</i> , incompleto entre <i>change-points</i> e histórico total entre <i>change-points</i>	96
40	Valores de \mathcal{E} para o consumo octa-horário não-normalizado para cada valor de k . . .	105
41	<i>Clustering</i> de um conjunto de 100 séries normalizadas de consumo octa-horário doméstico.	106
42	Valores de \mathcal{E} para o consumo diário não-normalizado para cada valor de k	106
43	<i>Clustering</i> de um conjunto de 100 séries normalizadas de consumo diário doméstico. .	107

Lista de Tabelas

1	Análise das ACF e PACF.[4]	37
2	Valores médios das métricas dos vários modelos na previsão de 100 séries temporais de consumo doméstico.	69
3	Ordenação dos modelos pelo número de vezes cujo resultado do Diebold-Mariano indica que a previsão é superior à dos outros modelos e valor desta frequência.	69
4	Ordenação dos modelos pela frequência em que o valor da respetiva métrica é inferior ao dos outros modelos.	70
5	Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade horária.	71
6	Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade octa-horária.	72
7	Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade diária.	73
8	Correlação entre temperatura e humidade para cada distrito.	76
9	Ordenação das variáveis exógenas ou ausência destas, pela frequência em que o valor da respetiva métrica é inferior ao das outras previsões.	76
10	Ordenação das variáveis exógenas pelo número de vezes cujo resultado do Diebold-Mariano indica que a previsão é superior à da utilização de outras variáveis exógenas ou da sua não utilização para efeitos de previsão.	76
11	Ordenação do tipo de histórico pela frequência em que o valor da respetiva métrica é inferior ao do outro tipo de histórico.	77
12	Ordenação dos tipos de histórico por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à do outro tipo de histórico e indicação da frequência.	78
13	Ordenação dos filtros pela frequência em que o valor da respetiva métrica é inferior ao dos outros filtros.	79
14	Ordenação dos filtros por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à dos outros filtros e indicação da frequência.	79
15	Ordenação das transformações pela frequência em que o valor da respetiva métrica é inferior ao dos outros filtros.	80
16	Ordenação dos filtros por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à dos outros filtros e indicação da frequência.	81
17	Valor das métricas de erro para a previsão individual horária <i>vs</i> previsão baseada em centróides.	84
18	Valor das métricas de erro para a previsão individual octa-horária <i>vs</i> previsão baseada em centróides.	86
19	Valor das métricas de erro para a previsão individual diária <i>vs</i> previsão baseada em centróides.	88

1 Introdução

O estágio, do qual se apresenta agora o relatório, foi levado a cabo na Whitesmith (WS), uma empresa de consultadoria focada no desenvolvimento de produtos de *software* e *hardware*, sediada no Instituto Pedro Nunes, Coimbra.

O trabalho a ser exposto neste relatório enquadra-se no contexto da plataforma Unplugg. Esta plataforma, a ser desenvolvida pela WS, tem o objetivo de retirar informações úteis no sentido da redução do consumo energético. Deste modo, suporta um sistema de gestão energética residencial, acessível através de uma página *web*, em que o processamento correspondente à gestão e análise dos dados é realizada na *cloud*. Neste *website* é possível inserir dados de consumo energético manualmente ou, caso o utilizador possua *smart meters* ou *smart plugs*, associar dispositivos à plataforma para que dados de consumo sejam recolhidos automaticamente. A partir dos dados de consumo disponibilizados é possível retirar informações úteis. Os modelos implementados, que constituem a componente de inteligência, analisam os dados e fornecem informações que irão motivar os utilizadores para uma gestão mais eficiente da sua energia elétrica. Neste contexto, a plataforma inicial suportava modelos da desagregação do consumo, deteção de *standby*, entre outros.

O estágio terá como objetivo contribuir para o desenvolvimento de modelos inteligentes da plataforma Unplugg. Este projeto enquadra-se no âmbito dos sistemas de *Home Energy Management*, que é uma área relativamente recente, sendo necessário discernir quais os projetos mais relevantes, quais os mais importantes para a empresa e, ainda, aqueles que serão passíveis de desenvolvimento no tempo disponível. De acordo com os objetivos atuais e planos futuros da WS para a plataforma Unplugg, avalia-se como prioritária a abordagem de dois aspetos: a elaboração de um modelo de previsão baseado no histórico do consumo energético.

A eletricidade é um recurso indispensável para as economias, a nível global, nacional ou local. Um consumo responsável, eficiente e bem planeado tornou-se uma necessidade para um bom retorno monetário, estabelecendo um conjunto de restrições essencial para delinear qualquer nova infraestrutura. De forma a otimizar a sua utilização, por razões de caráter ambiental ou por um esforço de manutenção de preços competitivos, os serviços de fornecimento estão constantemente a tentar ajustar a produção de energia à procura efetiva. Tradicionalmente, esta compatibilização foi levada a cabo a nível agregado (país ou regiões extensas) mas, recentemente, o desenvolvimento de *Smart Grids* abriu a porta para o controlo desagregado e otimização da produção de energia em ambientes restritos, como cidades ou parques industriais, e para a identificação de padrões de utilizadores em escalas menores, como seja a utilização doméstica. No entanto, enquanto a procura global é relativamente fácil de prever e compreender, graças à conjugação de um grande número de elementos em todas as regiões, a previsão para elementos desagregados é muito mais difícil uma vez que são muito maiores as variações e as incertezas[5].

Embora haja uma variedade de fontes de onde podem ser obtidos dados de energia, analisar esses dados é uma tarefa intrinsecamente difícil. Neste contexto, a previsão é incontornável e, quaisquer que sejam as circunstâncias ou horizontes de tempo envolvidas, indispensável num planeamento adequado e eficaz.

A previsibilidade de um evento ou uma quantidade depende de vários fatores[6], incluindo:

- a quantidade de dados que estão disponíveis;
- quão bem estão identificados e entendidos os fatores que influenciam o fenómeno;
- se as previsões podem afetar o comportamento do consumidor.

As previsões de consumo energético ou, mais especificamente, de consumo de eletricidade podem ser altamente precisas, na medida em que todas as três condições são geralmente satisfeitas. Em primeiro lugar, existe uma boa ideia sobre os fatores que contribuem para o consumo. Este é promovido por condições ambientais, com efeitos menores para aspetos de calendário, como feriados, e com impacto, por vezes substantivo, das condições económicas. Desde que haja um histórico suficiente de dados de consumo e condições relevantes é, em princípio, possível desenvolver um bom modelo com previsões precisas.

Existe, por vezes, a conceção errada de que as previsões não são passíveis de efetuar sobre ambientes em mudança. No entanto, um bom modelo de previsão capta também o comportamento dinâmico dos fenómenos sobre observação.

A previsão tem de se adaptar a situações muito diferenciadas, que incluem diferentes janelas de previsão baseadas em diferentes dados históricos, com padrões específicos como sejam, sazonalidades, tendências, possibilidade de alterações abruptas, entre muitos outros aspetos. Os métodos de previsão também apresentam diferentes características e graus de complexidade, desde um simples algoritmo baseado na observação mais recente (previsão naïve) até estruturas complexas constituídas por combinações de modelos. O resultado da previsão é muito relevante no contexto da gestão eficiente do consumo energético, uma vez que disponibiliza informação sobre o futuro, permitindo um planeamento atempado não só da produção de energia como do seu consumo com, nomeadamente, a redução da energia gasta por aparelhos em *standby*. Torna-se, também, necessário recorrer a ferramentas capazes de descobrir padrões dentro das grandes quantidades de informação que são recolhidas, como por exemplo o *clustering*. O *clustering* é a principal técnica utilizada para a partição de dados em grupos com base em propriedades internas inerentes aos dados, quando o conhecimento *a priori* não está disponibilizado. Outro aspeto que deve ser abordado tem a ver com a deteção de alterações nesses mesmos padrões, que ocorrem nos chamados *change-points*. No contexto do consumo energético doméstico, estas alterações podem estar relacionadas com mudanças na rotina, avarias ou aquisições de aparelhos elétricos, períodos de férias, etc. Todas estas alterações e as suas consequências alteram o paradigma, levando à necessidade de refazer o modelo de previsão. A análise de todos estes aspetos leva à possibilidade de criação de serviços em torno do consumo doméstico, promovendo uma gestão mais eficiente do consumo energético através de sugestões ou do controlo automático de aparelhos elétricos. Do ponto de vista da empresa fornecedora, a previsão em bloco dos seus clientes permite a gestão eficiente da distribuição e produção de energia, enquanto que o *clustering* permitirá a organização de clientes por tipo de padrão de consumo. Esta análise não é trivial devido à sua forte componente humana, que injeta muita variabilidade no consumo devido a situações de crise económica, alterações da rotina e devido aos dados que possuem granularidades reduzidas, o que limita a capacidade de previsão.

O objetivo deste projeto de estágio “Intelligence for Energy Efficiency and Comfort in the Smart Home Domain” foi definido como a investigação e implementação de métodos de previsão do consumo energético doméstico, bem como o desenvolvimento de outras ferramentas necessárias para poder disponibilizar estas funcionalidades como “Machine Learning as a Service”, integrado no projeto Unplugg. Para tal, foram definidos como componentes deste serviço a criação de modelos de previsão, o *clustering* de séries temporais de consumo e a deteção de *change-points*. Desenvolve-se assim, nos modelos de previsão, capacidades de adaptação dinâmica às mudanças de consumo, mesmo aquelas de carácter mais abrupto.

Na Secção 1, **Introdução**, é efetuado o enquadramento do trabalho na área dos sistemas de *Home Energy Management*, apresentando-se termos relacionados, para além da definição da problemática associada à previsão do consumo energético, do *clustering* de séries temporais, e da deteção

de alterações nos padrões de consumo. Na Secção 2, **Estado da arte**, é apresentada a revisão da literatura sobre as soluções HEM consideradas relevantes. Na Secção 3, **Conceitos teóricos**, expõe-se os conceitos básicos de séries temporais, modelos de previsão, algoritmos de *clustering*, detecção de *change-points* e métodos utilizados para a avaliação dos resultados. necessários à resolução do problema em questão. As experiências a partir das quais foram obtidos os resultados estão descritas na Secção 4, **Procedimento Experimental**. Na Secção 5, **Resultados**, é apresentado o tratamento de dados realizado, os testes efetuados para o estudo dos diversos componentes, e a interpretação dos mesmos. Na Secção 6, **Conclusões**, são apresentadas as conclusões finais baseadas na interpretação realizada na secção anterior. A Secção 7, **Trabalho Futuro** procura indicar aspetos interessantes a explorar para um futuro desenvolvimento do trabalho.

1.1 Contexto

A área em que se enquadra este trabalho, *Home Energy Management*, interseja-se com várias temáticas relevantes como sejam as *smart homes*, *smart meters* e *smart grids*, pelo que estas serão abordadas ao longo desta secção.

1.1.1 Sistemas de *Home Energy Management* (HEM)

Os sistemas de *Home Energy Management* (HEM) são instalados ao nível residencial, com o objetivo de monitorizar e gerir o consumo elétrico. Atuam no sentido de aumentar a eficiência energética e reduzir os custos do consumo. Estes sistemas são, normalmente, constituídos por contadores inteligentes para a recolha de dados, sistemas de análise de dados para a obtenção de informações úteis e sistemas de automação para atuar em função das mesmas. Unplugg[7], Watt-is[8], Plotwatt[9], Bidgely[10] e Opower[11] são soluções HEM baseadas em *software*, que fornecem sugestões e informações úteis sobre o consumo energético. Eragy[12], AlertMe[13] e UFO Power Center[14] são exemplos de sistemas HEM que se baseiam na automatização da gestão do consumo.

De entre as soluções HEM destacam-se a consciencialização do consumidor de energia, a deteção de *standby* e ocupação, o reconhecimento do estado de dispositivos, a previsão e desagregação do consumo, a deteção de anomalias no consumo energético de dispositivos, a automatização do conforto, a comparação de consumos e eficiência energética entre utilizadores, a simulação de tarifas de energia elétrica e a elaboração de recomendações para eficiência energética.

As soluções mais relevantes estão descritas em maior detalhe na Secção 2, com especial ênfase para os modelos de previsão do consumo energético.

1.1.2 *Smart Homes*

Os sistemas HEM são especialmente relevantes para *Smart Homes*. Estas residências caracterizam-se por se encontrarem equipadas com dispositivos conectados, que podem ser controlados remotamente, tais como termostatos inteligentes, *smart plugs*, *smart bulbs*, sensores de luminosidade, temperatura, entre outros. Estes dispositivos procedem à recolha de informação sobre o estado da habitação, permitindo a automatização de uma gestão conducente a melhorias da eficiência na realização de atividades na casa, do conforto dos seus ocupantes e a superiores poupanças de energia.

1.1.3 *Smart meters, power meters e smart plugs*

Com o objetivo de gerir o consumo energético, são instalados nas *Smart Homes* dois tipos de dispositivos: *power meters* ou *smart meters* e *smart plugs*. Os *smart meters* são dispositivos, semelhantes aos contadores elétricos convencionais instalados no mesmo local, que permitem a leitura e envio automático dos dados à fornecedora, sem a necessidade de intervenção humana. Esta recolha proporciona leituras mais frequentes e cobranças sobre o valor real, em oposição às estimativas normalmente efetuadas uma vez que, embora a cobrança seja feita mensalmente, a leitura não o é. Os *power meters* diferem dos *smart meters* no sentido em que os dados recolhidos pelos dispositivos são geralmente destinados ao uso local, pelo que raramente são enviados para a fornecedora. As *smart plugs* efetuam a monitorização do consumo individual de cada tomada. Na generalidade, ambos os dispositivos apresentam, em tempo quase real, os dados do consumo em *displays* integrados ou através do acesso a páginas *web*, para dispositivos com a capacidade de

conexão à rede. Destacam-se o Current Cost[15] e o TED5000[16] como soluções de *power meters* e o Cloogy[17] como uma solução mista de *power meters* e *smart plugs*.

1.1.4 Sensores e atuadores

Para além dos dispositivos referidos anteriormente, podem também ser instalados termostatos inteligentes. Estes dispositivos permitem, por meio de sensores de temperatura, humidade e presença e através de atuadores ligados a sistemas HVAC (sistemas de aquecimento, ventilação e ar-condicionado), gerir automaticamente *set points* (valor-alvo de temperatura definida no termostato), criar horários de funcionamento, regular a temperatura e a humidade e poupar energia no processo. Os termostatos inteligentes mais utilizados são o Nest[18], o Ecobee[19], o Tendril[20] e o Radio Thermostat[21].

Para além do controlo dos sistemas HVAC existem também sistemas inteligentes de controlo da iluminação. Estes, ao retirarem a informação de sensores de luminosidade e presença e ao identificarem a localização geográfica e hora atual (para determinar o nascer e pôr do sol), permitem gerir eficientemente as *smart bulbs*, ou lâmpadas inteligentes, tanto ao nível da intensidade luminosa como ao nível da cor. Um exemplo destes sistemas é o LIFX[22].

Os sistemas HEM constituem o núcleo integrante de todos estes sistemas. Recolhem e analisam minuciosamente a informação de todos os sensores existentes, produzindo sugestões e identificando as ações necessárias para os atuadores. Através da automatização da gestão do consumo, será possível uma maior eficiência energética, com o mínimo de incómodo para o utilizador.

1.1.5 Smart Grid

A rede elétrica tem como função transportar e distribuir a energia, fazendo com que esta chegue ao consumidor final. Este processo é bastante complexo devido à existência de grandes oscilações, quer na oferta quer na procura. Registe-se ainda que o consumo tende a aumentar, à medida que surgem novos dispositivos elétricos. Esta imprevisibilidade pode causar instabilidade na rede, fundamentalmente durante picos de procura (concentração com grande intensidade da procura num curto intervalo de tempo) provocando, no pior dos casos, a necessidade de produção de energia extra, o que envolve a ligação de novas centrais. Esta situação extrema reflete-se em custos bastante elevados para os produtores e distribuidores de energia elétrica.

Com o objetivo de superar esta problemática, surge a rede inteligente, *smart grid*. Esta consiste numa camada que, aplicada sobre a rede de distribuição, lhe confere a inteligência necessária para uma melhor gestão da produção e distribuição da energia elétrica, tornando estas mais eficientes (económica e energeticamente), confiáveis e sustentáveis.

As *smart grids* proporcionam uma melhor gestão dos fluxos de energia através de um canal de comunicação bidirecional entre o consumidor final e a fornecedora. Este processo é operacionalizado através da instalação de *smart meters*, que não só registam e comunicam o consumo elétrico à fornecedora como também recebem informações, tais como o custo da utilização atual e informações relativas a programas *demand response*. Estes programas podem ser definidos como um conjunto de ações a ser tomadas pelo utilizador, em resposta a estados particulares no sistema elétrico [23].

Os referidos programas promovem a alteração intencional dos padrões de consumo do utilizador final, motivada por incentivos, taxas dinâmicas ou taxas fixas, como por exemplo a aplicação de uma taxa mais reduzida durante o período noturno. Estes programas permitem às fornecedoras proporcionar serviços com maior estabilidade e qualidade, através da diminuição da frequência e

impacto dos picos e vales na procura de energia. O consumidor, ao gerir o seu consumo (*demand side management*) previamente, durante e após períodos de picos previstos pela fornecedora, reduzindo ou reagendando o consumo de energia, estará a beneficiar de custos reduzidos no consumo de eletricidade e a aumentar a eficiência energética da sua residência.

Estas redes inteligentes promovem a redução dos custos de produção, o aumento da eficiência energética e uma gestão mais eficiente das fontes de energia, facilitando a incorporação de energia elétrica variável, característica das fontes renováveis. Daí que se possa afirmar que “as redes inteligentes são, sem dúvida, o futuro da distribuição de energia eléctrica em Portugal”[24]. A sua flexibilidade é, também, facilitadora da aplicação de taxas variáveis.

O primeiro projeto de rede elétrica inteligente em Portugal surgiu na cidade da Évora, em 2010, com a instalação de 31000 contadores inteligentes, através do projeto *InovGrid* implementado pela EDP, Energias de Portugal[25].

A previsão do consumo energético é uma necessidade para a gestão eficiente das *smart grids*, dos programas de *demand response* e da gestão da eficiência energética de uma habitação. Isto deve-se ao facto de que, ao obter informação sobre o eventual estado futuro do sistema elétrico, é possível tomar medidas preventivas que permitem, ao produtor de energia, gerir melhor a produção. Numa outra perspetiva e no ponto de vista do consumidor, é possível identificar padrões de consumo e planear o seu consumo energético eficientemente, beneficiando de menores custos.

1.2 Definição do problema

O problema a abordar neste trabalho consiste na criação de um modelo de previsão do consumo energético residencial a integrar na plataforma Unplugg. Para além da definição do método de abordagem, é necessário determinar qual a janela de previsão, isto é, o intervalo de tempo, no futuro, para o qual será efetuada a previsão.

Relativamente ao enquadramento deste modelo na plataforma, há a considerar que um elevado número de utilizadores implica a necessidade da construção de vários modelos, um para cada utilizador ou dispositivo. Desta forma é fundamental, para tornar o sistema sustentável, a redução do número de modelos a concretizar. Uma abordagem para a resolução deste problema passa pelo agrupamento de consumos semelhantes podendo, assim, a cada grupo aplicar-se a mesma parametrização. Outro obstáculo prende-se com a dinâmica resultante de alterações na rotina e/ou aquisição de novos dispositivos elétricos, por parte dos utilizadores. Estas mudanças no padrão de consumo constituem alterações que levarão à necessidade de re-parametrização dos modelos, a partir desse ponto de mudança.

1.2.1 Previsão do consumo agregado

Nesta tarefa são comumente aplicados dois tipos de modelos com componente temporal: os que relacionam valores atuais com valores históricos e erros de previsões passadas, como é o caso dos modelos ARIMA (definidos na Secção 3.4), e aqueles cuja componente temporal é representada sob a forma de índice no eixo dos xx , presente nos modelos de regressão. A previsão será tratada através de séries temporais, descritas mais aprofundadamente na Secção 3. Como será evidenciado na Secção 3.3, esta previsão é uma tarefa complexa, uma vez que envolve um elevado número de passos para a construção de cada modelo.

1.2.2 Janela de previsão

O consumo elétrico é influenciado por vários fatores de âmbito social, económico, ambiental e contextual, levando à criação de modelos de previsão cada vez mais complexos, de forma a incorporar estes componentes.

Os modelos podem, de um modo geral, ser classificados em uni-variados ou multi-variados. Caso, por exemplo, se utilize apenas o valor do consumo energético, trata-se de uma aplicação uni-variada. Já se forem utilizadas variáveis para descrever o consumo e as condições ambientais, o método torna-se multi-variado.

A escolha dos métodos de previsão passa também pelas definições do período de estimativa:

- A longo prazo, de um a cerca de vinte anos.
- A médio prazo, de um mês a várias semanas.
- A curto prazo, entre uma hora, um dia e algumas semanas. Este é o período mais relevante para o presente trabalho uma vez que permite, ao nível da fornecedora, prevenir sobrecargas e, ao nível do consumidor final, planear o consumo.

A curto prazo, a análise de séries temporais uni-variadas para a previsão é limitante pois ignora fatores como o dia da semana (segunda, terça, ...), o tipo de dia (dia de semana, fim-de-semana e feriado) ou as condições ambientais (temperatura, humidade, ...). No entanto, a médio e longo prazo, o impacto dos fatores externos são reduzidos pois a sua influência é diluída pelo grande intervalo temporal e uma previsão baseada apenas no histórico do consumo pode ser efetuada.

Outro aspeto a considerar na escolha do período de previsão é a origem do consumo de dados. Como exemplo refira-se a maior previsibilidade a curto prazo do consumo energético ao nível de cada dispositivo, individualmente, do que do consumo total de uma residência.

1.2.3 Clustering de séries temporais

Os padrões de consumo energético diferem muito de residência para residência, o que implica a determinação de parâmetros individuais para cada uma. Este aspeto pode dever-se às diferentes características da habitação, como o tipo de isolamento e número e tipo de dispositivos elétricos presentes, e pelos residentes, como número de habitantes e diferentes rotinas. Estima-se que o universo de consumidores de energia elétrica se aproxime do valor de 6 milhões em Portugal[26], pelo que o agrupamento de séries temporais é considerada uma componente essencial ao serviço proporcionado pela plataforma Unplugg. O objetivo do *clustering* de séries temporais de consumo energético é identificar séries semelhantes, por forma a permitir descobrir séries que possam ser modeladas com a mesma parametrização. Permite, também, tipificar comportamentos. Uma das componentes mais importantes desta tarefa é a escolha da função de semelhança. A elaboração de *clusters* pode ser dividida, essencialmente, em três abordagens baseadas em: características, dados não processados e modelos[27].

1.2.4 Alterações nos padrões

Um *change-point* consiste no instante temporal que marca uma mudança brusca no comportamento de uma série temporal, resultando numa alteração dos parâmetros do modelo, a partir desse ponto. O estudo de *change-points* passa pela deteção da existência de uma ou mais mudanças e subsequente estimação da sua localização.

Estas mudanças de parâmetros podem resultar de alterações na média, variância, ou ambas, ou ainda estar associadas a modelos de regressão linear. Para identificar a localização de *um change-point* na média ou em modelos de regressão pode ser utilizado o método da razão de verossimilhança[28] através do teste estatístico de rácio de verossimilhança, LPT (do inglês *likelihood-ratio procedure test*). Para a deteção de um *change-point* na variância[29] pode ser utilizado o teste estatístico baseado na soma cumulativa, CUSUM (do inglês *cumulative sum*).

1.2.5 Unplugg

A Unplugg foi construída para permitir “Machine Learning as a Service”, isto é, a disponibilização de algoritmos de análises de dados e *machine learning* a clientes.

Essencialmente os objetivos desta plataforma são:

- Permitir a clientes o envio de dados devidamente catalogados (de energia consumida, produção de energia elétrica, entre outras) para o sistema da empresa, através de uma Interface de Programação de Aplicações (API do inglês *Application Programming Interface*), que serão armazenados em bases de dados para o efeito. São estes os dados objeto de análise.
- Colocar-se como fornecedor de algoritmos de análise de dados e *machine learning* com foco em fontes de dados relacionados com a energia.
- Proceder à aplicação vários modelos aos dados, consoante o resultado desejado. Estes modelos incluem mas não são limitados à eliminação do consumo energético de aparelhos em *stand-by* e à desagregação do consumo em dispositivos elétricos.

O projeto de estágio aqui apresentado procura desenvolver este último tópico, contribuindo com modelos adicionais como a previsão do consumo agregado, o *clustering* de séries temporais e a deteção de alterações nos padrões do consumo. Os requisitos definidos para o projeto são o desenvolvimento de uma biblioteca de previsão de séries temporais do consumo energético, a incluir na plataforma. Esta biblioteca será constituída por um conjunto de métodos que recebem como argumentos séries temporais, com ou sem a inclusão de variáveis exógenas associadas. A biblioteca permitirá a definição de horizontes temporais de previsão e a utilização de séries temporais com diferentes granularidades. O sistema, no limite, terá de comportar grandes quantidades de dados, pelo que o número de previsões a efetuar deverá ser o menor possível. Desta forma, será ainda necessário desenvolver funcionalidades adicionais que complementem o modelo de previsão. Outros requisitos serão, portanto, o *clustering* de séries temporais para o reconhecimento de padrões típicos no consumo, visando reduzir o custo computacional e evitar a previsão sobre séries semelhantes e a determinação dos pontos temporais (*change-points*) a partir dos quais será necessário proceder à atualização dos modelos de previsão para manter a respetiva qualidade. A integração dos modelos de previsão, *clustering* e deteção de *change-points* na plataforma implica a utilização da linguagem de programação *Python*, incluindo as bibliotecas *pandas* e *pymc*. Estas bibliotecas permitem respetivamente a manipulação fácil e robusta de dados relacionais ou catalogados, como sejam séries temporais *dataframes*, e a implementação de modelos estatísticos Bayesianos. Inclui-se também a possibilidade de *port* de outras linguagens como o R, que apresenta uma grande coleção de ferramentas dedicadas à previsão em séries temporais.

Um exemplo da utilização da Unplugg seria a criação de uma plataforma com base nas informações obtidas da análise dos dados enviados. Esta plataforma poderia apresentar informação

relevante ao utilizador como o histórico de consumo e o consumo atual, as tarifas que mais se adequam a este tipo de consumo, a previsão do consumo, entre outras, para possibilitar uma gestão do seu consumo energético e produção doméstica, de forma a planejar gastos, reduzir consumo em alturas de tarifas mais elevadas, ou mover cargas para outras alturas em que não existam picos da procura que exijam, por parte da produtora, recorrer a métodos de produção de energia menos eficientes e mais poluidoras.

2 Estado da arte

Os sistemas de *Home Energy Management*, como descrito na Secção 1.1.1, são sistemas residenciais de monitorização, análise e controlo do consumo energético, no sentido da redução de energia elétrica consumida, automatização do conforto e poupança nos custos dos consumos totais. Para tal, estes sistemas apresentam uma multiplicidade de soluções. Ao longo desta secção será apresentado o estado da arte sobre as soluções selecionadas por serem consideradas as mais relevantes.

2.1 Previsão do consumo energético

A previsão é uma tarefa complexa. No caso da previsão através de séries temporais, para obter a sazonalidade numa previsão de uma hora é necessário ter em conta não só o valor do consumo da hora correspondente no dia anterior como também da hora correspondente no mesmo dia da semana anterior. Existem vários modelos que podem ser utilizados na previsão, dos quais se destacam as redes neurais, o método de Holt-Winters, o método ARMA (do inglês *Auto-Regressive Moving Average*), o método ARIMA (do inglês *Autoregressive Integrated Moving Average*), o ARIMA sazonal/não sazonal ou vetorial. Os modelos podem, de um modo geral, ser classificados em uni-variados ou multi-variados. Caso, por exemplo, se utilize apenas o consumo em MWh, trata-se de uma aplicação uni-variada. Já se forem utilizadas variáveis para descrever o consumo e procura, o método torna-se multi-variado.

Citando-se um exemplo [30], para uma previsão foram utilizados três algoritmos de *machine learning* (*naive Bayes*, *k-nearest neighbor* e *support vector machines*) e um algoritmo de previsão baseado em séries temporais (um modelo ARMA de ordem (1,1)). Os algoritmos de *machine learning* devolvem uma classe de previsão do consumo energético, enquanto que o modelo ARMA devolve um valor contínuo. Esta abordagem revela que a previsão do consumo individual de cada dispositivo permite uma maior qualidade de previsão a curto prazo enquanto que a previsão do consumo total da habitação é mais previsível a prazos mais longos.

Outra aproximação semelhante de análise do consumo energético [31] foi testada através da modelação da série temporal por um processo de ARMA que se revelou bastante rigoroso na previsão do consumo para sistemas de grandes dimensões.

Noutra abordagem, recorreu-se a dados exógenos ao problema, neste caso dados ambientais para além do consumo total. O reconhecimento de padrões foi efetuado através de um modelo de *Artificial Neural Network* com entradas como a temperatura exterior, rácio de humidade e energia consumida. Este modelo proporcionou a previsão do consumo energético, a curto prazo, na última hora com um erro médio reduzido durante um ano completo de simulação [32].

Para a previsão do consumo energético total para simulação de programas *demand response* na Suécia, foram propostas várias abordagens para esquemas do consumo agregado *top-down* (do lado da fornecedora) e *bottom-up* (do lado do utilizador) para a previsão, a curto e longo prazo [33]. Para a previsão a curto prazo foi utilizada uma rede neuronal com três camadas, treinada com o histórico do consumo e dados ambientais, capaz de prever o consumo energético horário para o dia seguinte. Para a previsão a longo prazo foi utilizado o modelo *Markov-chain* que parte dos mesmos dados da previsão a curto prazo, tendo-lhes sido adicionados parâmetros econométricos, obtidos por meio de um questionário, e ainda o sinal de *demand response* do processador de DR¹.

Noutra abordagem [34], procurou elaborar-se modelos que melhor se ajustassem ao consumo energético residencial. Para tal foram estimados os modelos ARMA e ARIMA e testados ambos os

¹DR do inglês *Demand Response*.

modelos, para várias janelas de previsão, de forma a descobrir o par modelo/janela mais adequado, considerando o critério de informação de Akaike[35] (AIC do inglês *Akaike Information Criterion*), para o modelo ARMA, e o erro médio quadrático (RMSE do inglês *Root Mean Square Error*), para o modelo ARIMA, na análise da qualidade de previsão. Conclui-se que o modelo ARIMA obteve melhores resultados para a seleção do período de previsão para as ordens mensal e trimestral, enquanto que o modelo ARMA obteve melhores resultados para as ordens diária e semanal. Os períodos selecionados foram 28 dias, 5 semanas, 6 meses e 2 trimestres.

Com o objetivo de identificar as variáveis que influenciam o consumo energético da zona residencial de baixa de Brisbane, Queensland, e desenvolver a previsão do dia seguinte, foi criado um modelo ARIMAX² e uma rede neuronal [36]. Para o treino da rede neuronal e criação do modelo de previsão foram utilizados os seguintes dados: voltagem, corrente e fator de potência do transformador que distribui eletricidade para essa zona, para além da humidade e temperatura, durante o mesmo período. A rede neuronal obteve ligeiramente melhores resultados na precisão da previsão. No entanto o modelo de redes neuronais modela melhor flutuações suaves enquanto que o ARIMAX modela melhor os picos de procura. Devido a este facto, foi implementado um modelo híbrido que atingiu precisões compreendidas entre 74% e 84%, para o consumo total do próximo dia.

Para a previsão do consumo elétrico doméstico e não doméstico da Nova Zelândia[37], foram criados seis modelos: três modelos baseados em curvas de crescimento (Logistic, Harvey Logistic, Harvey), um modelo que combina os fatores económicos e demográficos com uma curva de crescimento, o ARIMA e o modelo VAL. Os resultados mostraram que o ARIMA é o melhor modelo para a previsão a curto prazo, o VAL para médio prazo e o Harvey para o longo prazo.

2.2 Clustering de séries temporais

Para a determinação de *clusters* de séries temporais semelhantes é necessário, primeiro, definir qual a função de verosimilhança ou de distância que se irá utilizar para determinar o nível de semelhança entre estas. Destacam-se a distância Euclideana entre as expansões autorregressivas [38], a correlação de coeficientes e distâncias relacionadas de Pearson ou mesmo *dynamic time warping distance* (DTW). Os algoritmos de *clustering* podem facilmente ser escolhidos de entre vários disponíveis, que incluem *fuzzy C-Means*, *agglomerative hierarchical*, *K-Means* ou *K-Medoids* ou até *relocation clustering procedures*[27].

No âmbito das séries temporais do consumo energético de edifícios, foi elaborada uma análise a diversas funções de verosimilhança, com o objetivo de construir um modelo para a validação de *clusters*, o método *clustered-vector balance* [39]. Das funções analisadas destacam-se a distância Euclideana para *features* contínuas, a correlação de Pearson na análise de tendências e evoluções, e a distância de Mahalanobis, que pode ser considerada uma evolução da distância Euclideana, para ter em conta a correlação dos dados através da integração da matriz de covariância.

Uma abordagem[40] para o *clustering* de séries temporais no âmbito do consumo energético passou pela elaboração da curva correspondente à média do consumo energético ao longo de uma estação, utilizando depois o algoritmo *K-Means* para agrupar séries, minimizando a distância Euclidiana de elementos do grupo. O algoritmo *K-Means* é um processo iterativo de particionamento que procura agrupar as curvas em conjuntos de forma a minimizar a distância Euclidiana de cada elemento do grupo com o centro do grupo.

Também no âmbito dos programas *demand response*, outra abordagem[41] passou pela utilização de *clustering* por particionamento, *K-Means clustering*, na descoberta de grupos de clientes cujo

²ARIMAX do inglês *Autoregressive Integrated Moving Average with eXogenous variables*

perfil energético, em períodos de 24 horas, seja semelhante. As funções de semelhança utilizadas foram a diferença dos quadrados, a distância Euclideana e a distância geométrica média.

2.3 Detecção de pontos de mudança

De entre as soluções encontradas para este problema, destacam-se o rácio de semelhança (*likelihood-ratio procedure*), a *Bayes solution*, o método das somas cumulativas (CUSUM), o critério informacional e o método de transformação de ondulas (*wavelet transformation*).

Baseado no método de somas cumulativas (CUSUM), uma outra abordagem considerou para análise N painéis, cada um com T observações, verificando se a média se manteve constante durante o período da observação, ou se sofreu alteração num momento indefinido [42].

Uma outra proposta, também baseada no método CUSUM [29], compreende uma extensão da abordagem referida anteriormente. Nela, a validação foi efetuada através da simulação de *Monte-Carlo*, com a conclusão de que o método é eficiente na deteção de alterações na variância. De forma a integrar a presença de autocorrelação no modelo, foi criada uma função de verosimilhança através da modificação da abordagem informacional [3]. Este método foi utilizado para a deteção de alterações climáticas e permite a deteção de alterações na variação de magnitude superior ao desvio padrão da série.

2.4 Desagregação do consumo

A desagregação do consumo energético consiste na identificação do consumo individual de cada equipamento, através da análise do consumo total da residência. A *Non-intrusive load detection* (NILD) ou *non-intrusive load identification* (NILI) é uma forma não intrusiva de efetuar a desagregação, partindo apenas da informação do consumo de uma *smart plug*, *smart meter* ou dispositivos ligados ao contador de energia.

Uma abordagem para a desagregação envolve a comparação dos consumos de cada habitação com comportamentos padrão presentes numa base de dados previamente desenvolvida e, posteriormente, sobre esses resultados a aplicação de análises estatísticas para identificação das melhores comparações [43]. Uma alternativa para a mesma tarefa passa pelo *sparse coding* ou seja, uma família de métodos não-supervisionados, aplicados em dados de aprendizagem, com conjuntos de bases *overcomplete*, de modo a representar os dados eficientemente [44]. O objetivo deste método é o de encontrar um conjunto de vetores de base, de tal modo que seja possível representar o vetor de entrada como uma combinação linear dos vetores de base [45]. Foi também desenvolvido um algoritmo [46] que permite detetar os dispositivos consumidores, baseando-se em alterações nas potências ativa/real e reativa ³ da energia consumida. Através de rotinas de pré-processamento, os estados *on/off* ou estados múltiplos de um mesmo dispositivo são agregados de forma a melhorar a individualização dos consumidores. Este algoritmo parte do princípio de que a soma de todas as variações no consumo de energia de um dispositivo, durante um ciclo de operação (*off-on-off*), é nula.

³A potência ativa corresponde ao valor de energia consumida num determinado intervalo de tempo ($P = U \times I \times \cos(\phi)$ em Watt). A potência reativa é a potência não útil, ou seja, a potência que circula de forma oscilante nas instalações mas não é consumida por nenhum recetor ($P = U \times I \times \sin(\phi)$).

2.5 Reconhecimento de atividades

Certos autores [47] definem atividades como estruturas complexas recursivas, cujo reconhecimento é efetuado como uma tarefa de *data mining*. Em contraste com este tipo de propostas em que os modelos surgem a partir dos dados, outras partem de um conhecimento *a priori*, baseado num cenário específico, sendo os modelos de atividade mapeados nos dados, o que leva a modelos sobreajustados (*overfitting*). Numa implementação [48] opta-se por uma definição de atividades baseadas, por exemplo, em índices de independência funcional para o ser humano, às quais são associados dispositivos elétricos. Dentre estes índices citem-se *Barthel Index* e *Functional Independence Measure* (FIM), que são medidas conhecidas de atividades da vida diária. Uma vez detetada uma operação de um dispositivo elétrico, identifica-se uma tarefa, cuja referência é indicada pela ontologia.

Por vezes assume-se a existência da correlação entre uma atividade específica e o consumo de energia e que, desta forma, é possível prever o consumo energético baseado no reconhecimento da atividade que está a ser efetuada [49]. Técnicas de *machine learning* são utilizadas para explorar esta relação e para a construção de modelos de previsão.

Para o mapeamento de atividades em classes que indiquem o valor de energia consumida relativa a cada atividade executada, foram utilizados e comparados três classificadores de *machine learning*: *Bayesian Belief Network*, *Support vector machine* e uma rede neuronal.

Uma rede neuronal [50] previamente treinada pode constituir o modelo utilizado na deteção da atividade, tendo como entrada dados obtidos pelos sensores. Os sensores utilizados foram, nomeadamente, de infravermelhos, de temperatura, de humidade e monitores de energia em cada dispositivo. Estes definem o contexto real do utilizador. Padrões de consumos energéticos não compatíveis com contexto da atividade registada são indicadores de dispositivos que estão a desperdiçar energia.

3 Conceitos teóricos

Tendo em conta o mencionado na definição do problema relativamente aos métodos mais utilizados para a previsão do consumo agregado, a análise de séries temporais, é apresentada de seguida uma revisão dos principais conceitos [51].

A previsão através da análise de séries temporais requer a execução dos seguintes passos:

- Visualizar o gráfico da série para examinar as suas características principais, e verificar a existência, nomeadamente, de tendência, de componentes de sazonalidade, de alterações bruscas no comportamento da série, e de *outliers* (ver mais em detalhe à frente, na Secção 3.2).
- Remover a tendência e os componentes de sazonalidade, de forma a obter resíduos estacionários. Para tal pode ser necessário aplicar transformações aos dados da série (ver Secção 3.3.1).
- Escolher um modelo que se ajuste aos resíduos, através de uma série de estatísticas amostrais, como a função de auto-correlação (Secção 3.4).
- Efetuar a previsão da série temporal através da previsão dos resíduos e invertendo quaisquer transformações que tenham sido aplicadas em passos anteriores, de forma a obter previsões para a série original $\{X_t\}$

3.1 Séries temporais

Uma série temporal consiste num conjunto de observações x_t , feitas sequencialmente em cada instante t , ao longo do tempo. Uma série temporal discreta no tempo é aquela cujas observações são realizadas num conjunto discreto de instantes T_0 . Uma série temporal contínua é obtida através de uma amostragem, contínua ao longo de um determinado intervalo de tempo. As séries temporais a serem analisadas ao longo deste trabalho serão do tipo discreto, uma vez que o objeto de estudo serão variáveis discretas recolhidas em intervalos fixos como é o caso do consumo energético de uma casa e da temperatura ambiente.

3.2 Análise de séries temporais

A modelação de uma série temporal requer a descrição de quatro componentes: tendência, sazonalidade, ciclicidade e aleatoriedade. A modelação será, então, função destes componentes.

Nesse sentido, a análise de séries temporais tem como objeto a identificação das seguintes características:

- **Tendência:** refere-se à existência de uma certa inclinação, em que os valores aumentam ou diminuem ao longo do tempo;
- **Sazonalidade:** prende-se com a existência de padrões repetidos regularmente e está, normalmente, relacionada com as estações do ano, meses, dias da semana, entre outros;

- **Presença de *outliers*:** correspondem a valores que diferem demasiado do padrão normal da série, podendo constituir erros pontuais de medição ou valores anormalmente diferentes. A presença de *outliers* numa série temporal tem influência negativa na qualidade da previsão obtida pelo modelo da série. Isto é devido à existência de uma relação de dependência entre valores adjacentes a *outliers*;
- **Presença de ciclos ou períodos não relacionados com a sazonalidade:** são repetições de valores que ocorrem em intervalos de tempo bem definidos.

Outras características passíveis de análise são a estabilidade da variância, a existência de alterações súbitas no comportamento da série e a aleatoriedade.

O modelo clássico de decomposição permite a representação dos dados através da expressão

$$X_t = m_t + s_t + Y_t, \quad (1)$$

onde m_t é uma função que varia lentamente e que contempla a componente da tendência, s_t é uma função de período conhecido, d , correspondente à componente sazonal e Y_t corresponde à componente estacionária do ruído aleatório. O componente cíclico, que não tem um período fixo, encontra-se incluído no termo tendência, sendo designado por tendência/cíclico.

3.3 Caracterização de séries temporais

A caracterização do tipo de série temporal, com tendência, sazonal, cíclica ou irregular é, assim, um aspeto de grande importância. Pode ser levado a cabo por inspeção visual ou recorrendo a testes, como sejam o Augmented Dickey–Fuller (ADF) e o Kwiatkowski – Phillips – Schmidt – Shin (KPSS) [35].

Se a série contiver muito ruído aleatório será necessário proceder à sua filtragem, recorrendo a técnicas de *smoothing*, com métodos exponenciais ou de média móvel. Os métodos de *smoothing* permitem também fazer a análise de tendência e sazonalidade (Holt-Winters). Esta análise é feita com funções de auto-correlação (ACF⁴ e PACF⁵).

A primeira questão que deve ser respondida numa análise recorrendo à ACF e PACF tem a ver com a série ser ou não aleatória. Tal leva-se à prática recorrendo-se a testes apropriados. A análise visual permite também obter informação sobre *outliers*, valores em falta e quebras estruturais nos dados para além de verificar se um modelo não-linear é mais apropriado para os dados.

Função de auto-correlação (ACF)

Seja $\{X_t\}$ uma série temporal univariada $E(X_t^2) < \infty$, a função da média de $\{X_t\}$ é

$$\mu_X(t) = E(X_t) \quad (2)$$

e a função de covariância de $\{X_t\}$ é

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))] \quad (3)$$

⁴ACF do inglês *Autocorrelation function*.

⁵PACF do inglês *Partial autocorrelation function*.

sendo r e s quaisquer instantes da série temporal. A função $\gamma_X(\cdot)$ é referida como a função de auto-covariância e $\gamma_X(h)$ como o seu valor na *lag*⁶ h , em que $h = |r - s|$.

Seja $\{X_t\}$ uma série temporal estacionária (conceito aprofundado na Secção 3.3.1), a função de auto-covariância (ACVF)⁷ de $\{X_t\}$ na *lag* h é

$$\gamma_X(h) = Cov(X_{t+h}, X_t). \quad (4)$$

A função de auto-correlação (ACF) de $\{X_t\}$ na *lag* h é

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t). \quad (5)$$

As propriedades lineares de covariâncias são facilmente verificáveis: se $EX^2 < \infty, EY^2 < \infty, E_Z^2 < \infty$ e a, b e c são constantes reais. Nesse caso

$$Cov(aX + bY + c, Z) = aCov(X, Z) + bCov(Y, Z). \quad (6)$$

A função de auto-correlação é utilizada para modelos mas, no entanto, o objeto inicial de estudo serão os dados observados $\{x_1, x_2, \dots, x_n\}$. Para determinar o nível da dependência nos dados e selecionar um modelo que tenha tal em conta, é necessário recorrer à função de auto-correlação amostral (*sample ACF*)⁸. Considerando os dados como valores efetivos da série estacionária $\{X_t\}$, então a *sample ACF* irá proporcionar uma estimativa do valor da ACF de $\{X_t\}$. Esta estimativa será utilizada para determinar o modelo apropriado a utilizar na representação da dependência dos dados.

Sendo x_1, \dots, x_n as observações de uma série temporal, a média da amostra de x_1, \dots, x_n é obtida através de

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t. \quad (7)$$

A função de auto-covariância amostral é, então,

$$\hat{\gamma}(h) := n^{-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n. \quad (8)$$

A função de auto-correlação amostral pode apresentar-se como

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n. \quad (9)$$

⁶O termo *lag* refere-se aos valores históricos.

⁷ACVF do inglês *Autocovariance Function*.

⁸*sample ACF* do inglês *Sample auto-correlation function*.

Função parcial de auto-correlação (PACF)

A função parcial de auto-correlação (PACF) de um processo ARMA $\{X_t\}$ (descrito mais em pormenor na Secção 3.4) é a função $\alpha(\cdot)$ definida pelas equações

$$\alpha(0) = 1 \quad (10)$$

e

$$\alpha(h) = \phi_{hh}, \quad h \geq 1, \quad (11)$$

onde ϕ_{hh} é o último componente de

$$\phi_h = \Gamma_j^{-1} \gamma_h, \quad (12)$$

com $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$, e $\gamma_h = [\gamma(1), \gamma(2), \dots, \gamma(h)]'$.

3.3.1 Séries estacionárias vs séries não estacionárias

Numa série temporal estacionária os dados, ao longo do tempo, mantêm a sua média, variância e estrutura de auto-correlação. Por oposição, numa série temporal não estacionária os dados apresentam uma tendência (positiva ou negativa) e/ou ciclo ou sazonalidade. $\{X_t\}$ é considerada fracamente estacionária se

- (i) $\mu_X(t)$ for independente de t e
- (ii) $\gamma_X(t+h, t)$ for independente de t para cada h

Por outro lado, diz-se que a estacionariedade forte de uma série temporal $\{X_t, t = 0, \pm 1, \dots\}$ é definida pela condição de (X_1, \dots, X_n) e $(X_{1+h}, \dots, X_{n+h})$ possuírem a mesma distribuição conjunta para todos os número inteiros h e $n > 0$.

Uma série estacionária corresponde a um processo estocástico em que a distribuição de probabilidade não se modifica por translação temporal. Menos estritamente, basta que o processo tenha um primeiro momento e uma auto-covariância que não varie com o tempo.

As séries temporais não estacionárias deverão ser objeto de uma transformação prévia, por forma a torná-las em séries estacionárias. O objetivo destas transformações não será a melhoria da previsão, mas sim a eliminação da tendência, para que se possa efetuar o teste de existência de sazonalidade. Caso exista, a eliminação da tendência permite a análise da componente sazonal sem o efeito da primeira.

Séries temporais com tendência e sem sazonalidade

Na ausência de sazonalidade o modelo pode ser definido por

$$X_t = m_t + Y_t, \quad t = 1, \dots, n \quad (13)$$

onde $E[Y_t] = 0$.

Para estimação da tendência em séries temporais com ausência de sazonalidade podem aplicar-se diversas abordagens:

(a) Alisamento (*smoothing*) com um filtro finito de média móvel

Seja q um inteiro não negativo correspondente ao número de pontos considerado de cada lado do ponto central e considerando a média móvel de duplo lado

$$W_t = (2q + 1)^{-1} \sum_{j=-q}^q X_{t-j} \quad (14)$$

do processo $\{X_t\}$ definido em (13). Então para $q + 1 \leq t \leq n - q$,

$$W_t = (2q + 1)^{-1} \sum_{j=-q}^q m_{t-j} + (2q + 1)^{-1} \sum_{j=-q}^q Y_{t-j} \approx m_t, \quad (15)$$

assumindo que m_t é aproximadamente linear ao longo do intervalo $[t - q, t + q]$ e que a média dos termos de erro ao longo do intervalo é próximo de zero. A média móvel proporciona, então, as estimativas

$$\hat{m}_t = (2q + 1)^{-1} \sum_{j=-q}^q X_{t-j}, \quad q + 1 \leq t \leq n - q \quad (16)$$

(b) Alisamento (*smoothing*) exponencial

Para qualquer valor fixo de $\alpha \in [0, 1]$, o alisamento exponencial $\hat{m}_t, t = 1, \dots, n$ é definido pelas recursividades (*recursions*)

$$\hat{m}_t = \alpha X_t + (1 - \alpha) \hat{m}_{t-1}, \quad t = 2, \dots, n \quad (17)$$

e

$$\hat{m}_1 = X_1 \quad (18)$$

(c) Alisamento (*smoothing*) por eliminação de componentes de alta frequência

Através da eliminação de componentes de alta frequência da sua expansão da série de Fourier, é possível suavizar a série temporal.

(d) Ajuste polinomial

Uma tendência do tipo $m_t = a_0 + a_1 t + a_2 t^2$ pode ser ajustada aos dados através da seleção dos parâmetros a_0, a_1 e a_2 , de forma a minimizar a soma dos quadrados, $\sum_{t=1}^n (x_t - m_t)^2$. O método dos mínimos quadráticos também pode ser utilizado para estimar tendências polinomiais de maior ordem.

No que diz respeito à eliminação da tendência, esta pode ser processada através da diferenciação (*differencing*). Define-se o operador de diferença de *lag-1* ∇ como

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t, \quad (19)$$

onde B é o operador de deslocamento no sentido inverso,

$$BX_t = X_{t-1} \quad (20)$$

As potências dos operadores B e ∇ são definidas, por exemplo, como $B^j(X_t) = X_{t-j}$ e $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$, $j \geq 1$, com $\nabla^0(X_t) = X_t$

Séries temporais com tendência e sazonalidade

Na presença de tendência e sazonalidade o modelo pode ser definido pelo modelo clássico (Equação (1)) de decomposição, com $EY_t = 0$, $s_{t+d} = s_t$, e $\sum_{j=1}^d s_j = 0$.

Para estimação dos componente de tendência e sazonalidade, pode aplicar-se a abordagem que a seguir se indica. A tendência é primeiramente estimada através da aplicação do filtro de média móvel para eliminar a componente sazonal e para "abafar" o ruído. Se o período d for par, como por exemplo $d = 2q$, então usa-se

$$\hat{m}_t = (0.5x_{t-q} + x_{t-q+1} + \dots + x_{t+q-1} + 0.5x_{t+q})/d, \quad q < t \leq n - q. \quad (21)$$

Se o período for ímpar, como por exemplo $d = 2q + 1$, então usa-se a média móvel. O segundo passo consiste na estimação da componente sazonal. Para cada $k = 1, \dots, d$ calcula-se a média w_k dos desvios $\{(x_{k+jd} - \hat{m}_k + jd), q < k + jd \leq n - q\}$. Uma vez que a soma média os desvios não é necessariamente zero, a componente s_k é estimada por

$$\hat{s}_k = w_k - d^{-1} \sum_{i=1}^d w_i, \quad k = 1, \dots, d \quad (22)$$

e $\hat{s}_k = \hat{s}_{k-d}$, $k > d$

Após retirar a componente de sazonalidade estimada,

$$d_t = x_t - \hat{s}_t, \quad t = 1, \dots, n \quad (23)$$

a tendência é estimada novamente a partir dos dados sem componente de sazonalidade $\{d_t\}$ utilizando um dos métodos descritos anteriormente. Após a estimação da componente de sazonalidade e a re-estimação da componente de tendência, a componente estimada do ruído é dada por

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad t = 1, \dots, n \quad (24)$$

Para eliminação das componentes de tendência e sazonalidade aplicadas a dados com sazonalidade de período d define-se um operador de diferença de *lag-d* ∇_d como

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t. \quad (25)$$

Aplicando o operador ∇_d ao modelo

$$X_t = m_t + s_t + Y_t, \quad (26)$$

onde $\{s_t\}$ tem período d , obtém-se

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d}, \quad (27)$$

que proporciona uma decomposição da diferença $\nabla_d X_t$ numa componente de tendência ($m_t - m_{t-d}$) e uma componente de ruído ($Y_t - y_{t-d}$). A componente de tendência pode ser eliminada aplicando uma potência do operador ∇ .

Após atingir a estacionariedade da série, é possível modelar a sequência do ruído, ou seja, os resíduos obtidos através da diferenciação dos dados ou da estimação e posterior eliminação dos componentes de tendência e sazonalidade. Se não existir dependência entre os resíduos, é possível considerar as observações como variáveis aleatórias independentes e não será necessário efetuar nenhum outro processo de modelagem. Por outro lado, se existir dependência, será necessário construir um modelo de série temporal estacionária mais complexo para o ruído, de forma a ter a dependência em conta.

De referir que a média móvel e o *smoothing* espectral são métodos não paramétricos para a estimação e não para a construção do modelo, enquanto que a construção do modelo pode ser efetuada

1. ajustando uma tendência polinomial (através dos mínimos quadrados), subtraindo a tendência ajustada e encontrando o modelo que seja adequado aos resíduos

ou

2. eliminando a tendência por diferenciação (*differencing*) e encontrando, posteriormente, um modelo estacionário à série diferenciada.

Um dos cuidados a ter com as transformações é a aplicação de transformações inversas às efetuadas sobre os valores obtidos na previsão, uma vez que os modelos encontrados para séries temporais transformadas não modelam a série original. Apenas após a aplicação das transformações inversas, os valores obtidos se tornarão válidos na previsão do modelo original.

3.4 Modelos de séries temporais

Um modelo de serie temporal para os dados observados $\{x_t\}$ é a especificação das distribuições associadas de uma sequência de variáveis aleatórias $\{X_t\}$, onde $\{x_t\}$ é tida como uma realização.

Existem vários modelos que podem ser utilizados para a previsão, dos quais se destacam o ARMA⁹, o ARIMA¹⁰ e o ARIMAX¹¹, o SARIMA¹², o SARIMAX¹³ e os métodos de *smoothing* exponencial.

3.4.1 Modelos não-sazonais

Os modelos não-sazonais considerados neste trabalho foram os modelos ARMA e o ARIMA, que serão descritos nas seguintes secções.

⁹ARMA do inglês *Auto-Regressive Moving Average*.

¹⁰ARIMA do inglês *Autoregressive Integrated Moving Average*.

¹¹ARIMAX do inglês *Autoregressive Integrated Moving Average with eXogenous variables*.

¹²SARIMA do inglês *Seasonal Autoregressive Integrated Moving Average*.

¹³SARIMAX do inglês *Seasonal Autoregressive Integrated Moving Average with eXogenous variables*.

3.4.1.1 ARMA

Diz-se que um processo linear tem a representação

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad (28)$$

para qualquer t , $\{Z_t\} \sim WN(0, \sigma^2)$, onde WN se refere a *white noise*¹⁴ (designação inglesa para ruído aleatório com um espectro uniforme de frequências que cobrem uma grande gama) e $\{\psi_j\}$ é uma sequência de restrições com $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

$\{X_t\}$ é um processo ARMA(p, q) se $\{X_t\}$ for estacionária e para todo o t

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}, \quad (29)$$

onde $\{Z_t\} \sim WN(0, \sigma^2)$ e os polinômios $(1 - \phi_1 z - \dots - \phi_p z^p)$ e $1 + \theta_1 z + \dots + \theta_q z^q$ não têm fatores comuns. O processo pode ser reescrito, de forma concisa, uma vez que $z^j X(t) = X(t-j) = B^j X_t$,

$$\phi(B)X_t = \theta(B)Z_t, \quad (30)$$

onde $\phi(\cdot)$ e $\theta(\cdot)$ são, respetivamente, os polinômios de grau p e de grau q

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \quad (31)$$

e

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q \quad (32)$$

sendo B o operador de deslocamento no sentido inverso ($B^j X_t = X_{t-j}$, $B^j Z_t = Z_{t-j}$, $j = 0 \pm 1, \dots$)

Uma série temporal $\{X_t\}$ é dita um processo auto-regressivo de ordem p (ou AR(p)) se $\theta(z) \equiv 1$, e um processo de média móvel de ordem q (ou MA(q)) se $\phi(z) \equiv 1$.

Um processo causal AR(p) é definido por

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t, \quad Z_t \sim WN(0, \sigma^2), \quad (33)$$

onde $\{Z_t\} \sim WN(0, \sigma^2)$, $|\phi| < 1$ e Z_t não está correlacionado com X_s para cada $s < t$.

$\{X_t\}$ é um processo de média móvel, MA(q), de ordem q se,

$$X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}, \quad (34)$$

onde $\{Z_t\} \sim WN(0, \sigma^2)$ e $\theta_1, \dots, \theta_q$ são constantes.

Resumindo, o ARMA(p, q) é um modelo uni-variado que combina os modelos AR e MA. AR, componente auto-regressiva, consiste numa regressão linear do valor atual em relação a um ou mais valores da série. O valor p constitui a ordem do modelo AR e especifica a maior ordem do parâmetro auto-regressivo. Se o valor de p for definido como 2, por exemplo, ambos os parâmetros auto-regressivos *lag* 1 e *lag* 2 serão incluídos no modelo.

¹⁴Se $\{X_t\}$ for uma sequência de variáveis aleatórias não correlacionadas, com média zero e variância σ^2 , então X_t é uma série estacionária referida como ruído branco (*white noise*). A sequência é indicada através da notação $X_t \sim WN(0, \sigma^2)$.

MA, componente da média móvel, consiste numa regressão linear do valor atual em relação a *white noise* ou choques aleatórios de um ou mais valores da série. Assume-se que os choques aleatórios, em cada ponto, pertencerão à mesma distribuição, geralmente normal, com localização em zero e escala constante. Estes choques aleatórios são propagados para os valores futuros da série temporal.

O valor q constitui a ordem do modelo MA e especifica a maior ordem do parâmetro da média móvel. Se o valor de q for definido como 2, por exemplo, ambos os parâmetros *lag* 1 e *lag* 2 de média móvel serão incluídos no modelo.

A determinação das ordens de p e q do ARMA são efetuadas pela análise visual das funções ACF e PACF ou pelo teste de estimação de vários modelos para cada valor de p e q e discriminando cada modelo através do valor dos critérios de informação, como sejam o método linear dos mínimos quadrados ou a estimativa da probabilidade máxima.

As funções ACF e PACF devem ser analisadas conjuntamente, uma vez que apresentam padrões que apenas podem ser interpretados conjuntamente. O gráfico de um par de ACF/PACF é chamado de correlograma. A interpretação dos correlogramas é apresentada na Tabela 1, enquanto que exemplos são apresentados na Figura 1.

Tabela 1: Análise das ACF e PACF.[4]

Modelo	ACF	PACF
AR(p)	diminui gradualmente	anula-se após p lags
MA(q)	anula-se após q lags	diminui gradualmente
ARMA(p,q)	diminui gradualmente	diminui gradualmente

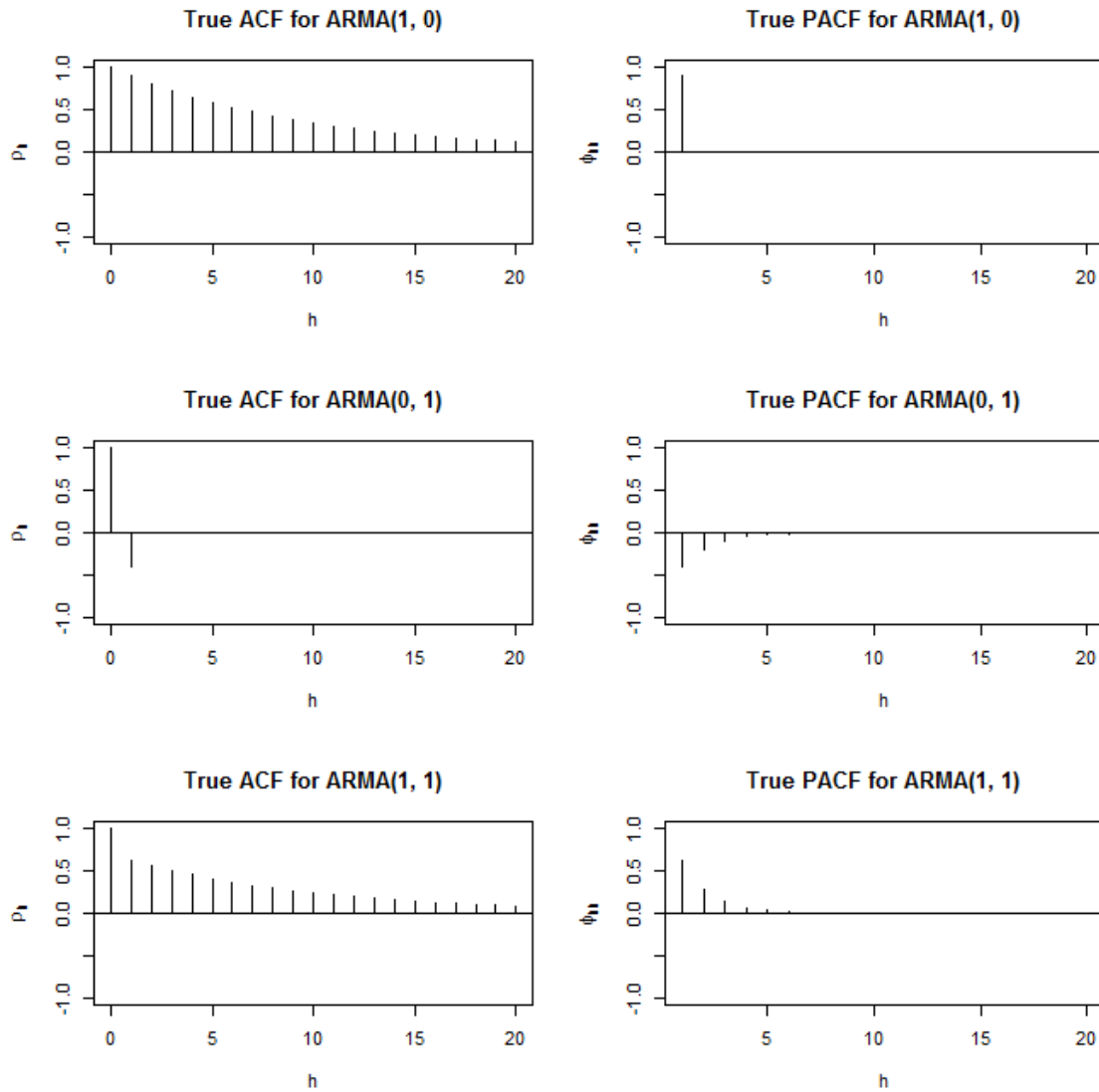


Figura 1: Exemplos de correlogramas.[1]

3.4.1.2 ARIMA

Este modelo é semelhante ao descrito anteriormente, sendo a sua definição efetuada através da adição de mais um parâmetro, d , correspondente ao número de diferenças a realizar para tornar a série estacionária. No modelo ARIMA(p,d,q), para um d inteiro não negativo a expressão que traduz o processo é

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad Z_t \sim WN(0, \sigma^2) \quad (35)$$

onde $\phi(z)$ e $\theta(z)$ são polinômios de grau p e q , respetivamente, e $\phi(z) \neq 0$ para $|z| \leq 1$.

Uma previsão com qualidade requer a determinação da ordem adequada para o modelo ARIMA(p , d , q), de forma a capturar o perfil dinâmico dos dados. Esta determinação é efetuada através da observação visual da série e da análise das funções de auto-correlação (ACF) e de auto-correlação parcial (PACF). O teste de Box-Jenkins é utilizado para o efeito e tem os seguintes passos:

- A identificação da estacionariedade para determinação da ordem de diferenciação, d , que pode ser efetuada através do método gráfico, apresentando os dados em função do tempo e os valores de ACF (caso ACF não diminua até zero ou se encontre em decrescimento lento sugere que a série não seja estacionária).
- A caracterização da componente sazonal. A sazonalidade pode ser identificada através de uma observação visual, da análise da função ACF ou da análise da densidade espectral.
- A estimação dos parâmetros p e q de forma semelhante à feita para o modelo ARMA.

3.4.2 Modelos Sazonais

Os modelos sazonais escolhidos para a análise deste trabalho incluíram dois tipos de modelos: baseados na média móvel e auto-regressão, para o qual foram considerados os modelos SARIMA e SARIMAX, e baseados no alisamento exponencial, para o qual foram considerados os modelos Holt-Winters aditivo e Holt-Winters multiplicativo.

3.4.2.1 Sazonalidade

Os modelos sazonais de previsão exigem a introdução de um parâmetro relativo ao período que corresponde a um conjunto de pontos que tem um padrão semelhante ao conjunto anterior com essa mesma duração. Este parâmetro denomina-se de período de sazonalidade ou apenas sazonalidade. É, por conseguinte, necessário identificar esse parâmetro. A sazonalidade de uma série temporal pode ser identificada através da análise visual do respetivo correlograma ou, mais especificamente, do gráfico da função de auto-correlação (ACF). Se uma série for sazonal, observa-se um comportamento, normalmente, sinusoidal com amortecimento ao longo do tempo. Identificando os máximos locais e observando a decalagem a que correspondem, é possível identificar o valor da sazonalidade. Na Figura 2 é possível observar que o máximo local relativo à sazonalidade de valor 24 para uma série temporal de granularidade horária.

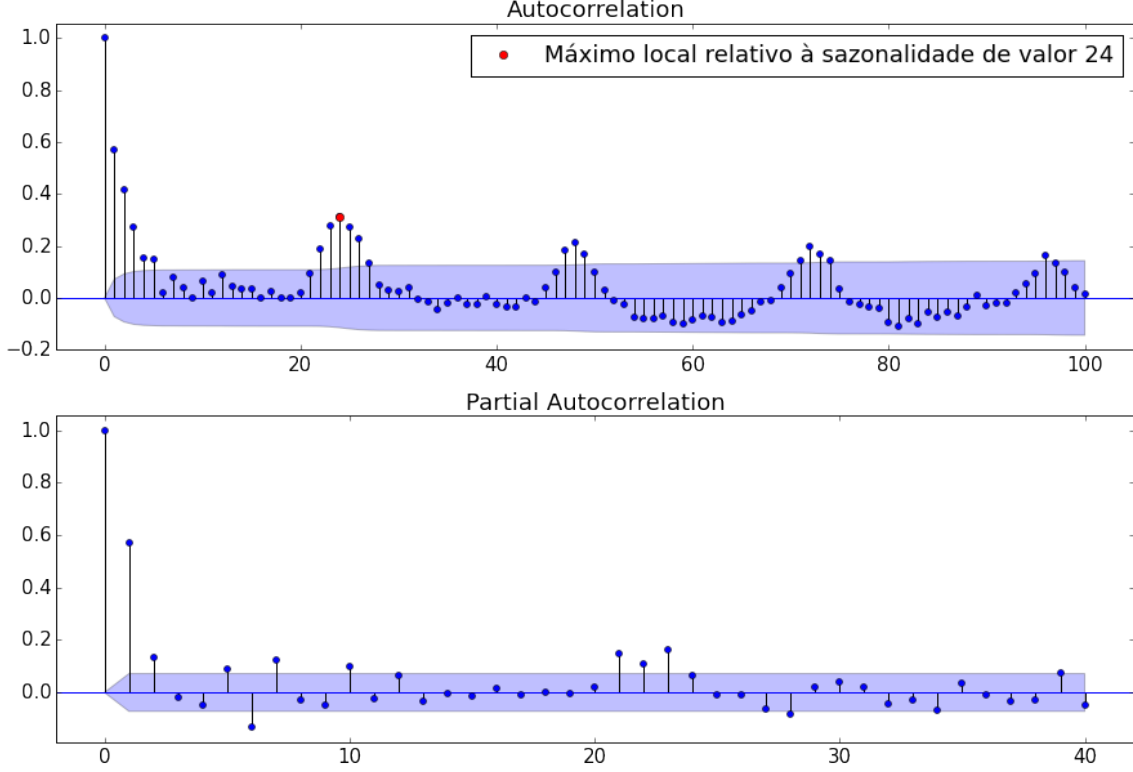


Figura 2: Correlograma de uma série temporal.

3.4.2.2 SARIMA

Ao modelo ARIMA pode ser acrescentada uma componente sazonal, constituída por termos semelhantes aos da componente não-sazonal, mas diferindo desta por envolver observações com decaimento temporal correspondente ao período da sazonalidade s . Dá-se, assim, origem ao modelo SARIMA usualmente representado por

$$\text{SARIMA}(p, d, q)(P, D, Q)_s.$$

Esta integração da sazonalidade no modelo passa pela multiplicação das componentes sazonais AR e MA da seguinte forma [51]

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2) \quad (36)$$

em que

$$Y_t = (1 - B)^d(1 - B^s)^D X_t, \quad (37)$$

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p, \quad (38)$$

$$\Phi(z) = 1 - \Phi_1 z - \dots - \Phi_P z^P, \quad (39)$$

$$\theta(z) = 1 - \theta_1 z + \dots + \theta_q z^q, \quad (40)$$

$$\theta(z) = 1 - \Theta_1 z + \dots + \Theta_Q z^Q. \quad (41)$$

O modelo SARIMAX estende o modelo SARIMA incluindo variáveis exógenas, de carácter adicional, para melhor explicar o comportamento da variável endógena e visando a melhoria do desempenho da previsão. O modelo SARIMAX é representado por

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t + \mu(B)X_t, \quad \{Z_t\} \sim WN(0, \sigma^2), \quad (42)$$

onde X_t é a variável exógena e μ representa o vetor dos coeficientes associados.

3.4.2.3 Suavização Exponencial

As previsões por métodos de *smoothing* exponencial, ou suavização exponencial, baseiam-se em valores obtidos através de médias ponderadas de observações passadas. Os pesos associados a estas observações sofrem um decaimento exponencial de acordo com o aumento da antiguidade destas, sendo o maior peso atribuído à observação histórica mais recente e o menor à mais antiga.

Holt-Winters é um método sazonal de previsão obtido através da combinação de equações de *smoothing*. Este método existe em duas variações: Holt-Winters aditivo e Holt-Winters multiplicativo. Ambos os modelos são constituídos pela combinação de componentes de nível, l_t , tendência, b_t , e sazonalidade, s_t , com os parâmetros de *smoothing* α , β^* e γ .

O modelo Holt-Winters aditivo é representado pela equação[6]

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t-m+h_m^+}, \quad (43)$$

com m sendo a sazonalidade e

$$h_m^+ = \lfloor (h-1) \bmod m \rfloor^{15} + 1, \quad (44)$$

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}), \quad (45)$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1}, \quad (46)$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m}. \quad (47)$$

Por sua vez, o modelo Holt-Winters multiplicativo é representado pela equação

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t-m+h_m^+}, \quad (48)$$

com

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(l_{t-1} + b_{t-1}), \quad (49)$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1}, \quad (50)$$

¹⁵Designa-se por $\lfloor u \rfloor$ o maior número inteiro não superior a u .

$$s_t = \gamma \frac{y_t}{(l_{t-1} + b_{t-1})} + (1 - \gamma)s_{t-m}. \quad (51)$$

Os parâmetros α , β^* e γ variam entre 0 e 1 e são obtidos através da seleção de valores que minimizam a soma dos erros quadráticos, SSE, com

$$SSE = \sum_{t=1}^T (y_t - \hat{y}_{t|t-1})^2.$$

3.5 Clustering

Os algoritmos de *clustering* são formas de organizar ou agrupar conjuntos de objetos de acordo com o respetivo nível de semelhança que esses objetos apresentam entre si. O *clustering* de séries temporais apresenta problemas adicionais que não existem no âmbito do *clustering* tradicional de vetores de características. Tal deve-se à associação de uma componente temporal aos valores que não pode ser ignorada aquando do agrupamento de séries temporais, o que torna esta problemática pouco trivial. Um exemplo é a heterogeneidade das rotinas dos consumidores, dos aparelhos elétricos associada à mutabilidade dinâmica do dia a dia de cada um, que origina séries temporais de consumo energético muito diferenciadas. Neste contexto, a distância Euclideana entre pontos terá pouca relevância como medida de semelhança, uma vez que dificilmente os valores de consumo de dois indivíduos diferentes num dado momento serão semelhantes. A análise dos padrões de consumo, por outro lado, será mais significativa para a determinação da semelhança de séries temporais de consumo. Desta forma, é necessário adaptar os métodos de *clustering* a este tipo de circunstâncias, tendo em consideração não só o algoritmo mas também a métrica de distância. Raramente existe conhecimento *a priori* sobre os *clusters* de um certo conjunto de séries temporais o que leva à utilização de algoritmos de técnicas não supervisionadas. Seguidamente será descrito um algoritmo adequado utilizado neste trabalho e a métrica, respetivamente.

3.5.1 K-Means

K-Means é um algoritmo não supervisionado de *clustering*, muito utilizado na literatura para séries temporais, que consiste na repartição de um conjunto de séries por k *clusters*. O algoritmo estabelece uma correspondência entre séries e *clusters*, minimizando a distância de cada série ao centro de massa do *cluster*, centróide, a que pertencem. Assim, o algoritmo K-means procura minimizar a função:

$$\sum_{j=1}^k \sum_{i=1}^n D(x_i^{(j)} - c_j), \quad (52)$$

onde k é o número de *clusters*, j o número de séries temporais, $x_i^{(j)}$ a série temporal de índice i que pertence ao *cluster* j e c_j o centróide do *cluster* j . O algoritmo começa pela inicialização dos k centróides, um para cada *cluster*. Em cada iteração é atribuído a cada série temporal o *cluster* mais próximo e são recalculados os centróides através da média de todas as séries temporais de cada *cluster*. O algoritmo termina caso se atinja o número de iterações máximo estabelecido ou caso não tenha existido alteração nos *cluster*, relativamente à iteração anterior.

A inicialização dos centróides afeta o desempenho do K-Means pelo que, embora convirja, pode não encontrar sempre a solução ótima. Existem várias formas de inicializar os centróides, constituindo a mais simples numa escolha aleatória de séries temporais do conjunto, sendo esta utilizada para o desenvolvimento deste projeto.

Este algoritmo de *clustering* exige a definição de um valor k especificando o número de *clusters* a formar. Uma vez que não existe conhecimento *a priori* sobre este valor, foi utilizada uma métrica adequada para a escolha do valor k . Assim procede-se ao cálculo do valor da distância inter-*cluster* (Γ_v) e da distância intra-*cluster* (Λ_v) para cada execução do algoritmo com um k específico. Estas distâncias são dadas por[39]

$$\Lambda_v = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} e(p_i^{(j)}, p_0^{(j)}) \quad (53)$$

$$\Gamma_v = \frac{1}{k(k-1)} \sum_{j=1}^k \frac{n_j}{n} \sum_{l \neq j}^{n_j} e(p_0^{(j)}, p_0^{(l)}) \quad (54)$$

Uma vez que estes valores originam distribuições diferentes, é necessário efetuar uma *z-transform* prévia para cada um seguida do cálculo de

$$\mathcal{E}_v(\chi) = \Gamma_{v_z} - \Lambda_{v_z}, \quad (55)$$

onde se utiliza uma notação óbvia. Será selecionado o valor de k que maximiza $\varepsilon_v(\chi)$.

3.5.2 Dinamic Time Warping

Dinamic Time Warping[2] (DTW) é uma métrica que permite a análise de padrões de séries temporais. Caracteriza-se pela sua não linearidade e capacidade de alinhamento elástico. O DTW é baseado num algoritmo que procura alinhar duas séries temporais, distorcendo a sua referência temporal de forma a obter uma correspondência ótima. Esta correspondência assenta no caminho que une o elemento (n, m) ao elemento $(1, 1)$ da matriz representada na Figura 3 e que corresponda ao menor valor da célula (n, m) . A grelha representada tem um tamanho precisamente $n \times m$, sendo n o número de pontos da série A e m o número de pontos da série B , no cálculo da distância entre as séries A e B . A grelha é preenchida, definindo-se o valor da primeira célula $(1,1)$ como

$$g(1, 1) = d(y_1, x_1), \quad (56)$$

onde $d(y_1, x_1)$ é a diferença entre os valores das séries temporais A e B no instante i . As restantes células são calculadas através de

$$g(i, j) = d(i, j) + \min\{g(i, j-1), g(i-1, j-1), g(i-1, j)\}. \quad (57)$$

Após preenchida na totalidade, o valor da distância DTW entre as séries é obtido através de

$$\text{DTW}(A, B) = \frac{g(n, m)}{n + m}. \quad (58)$$

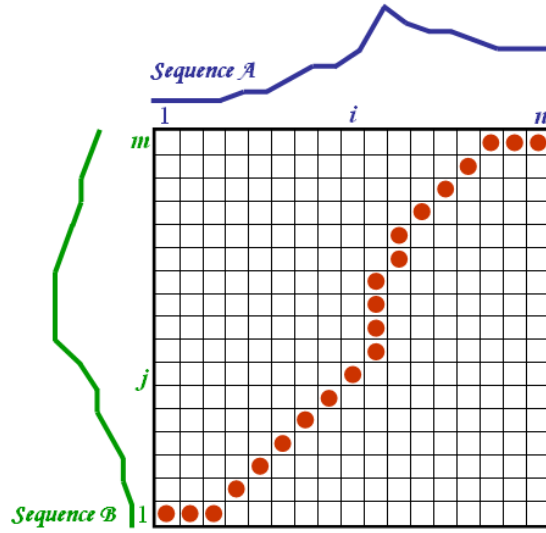


Figura 3: Grelha utilizada no algoritmo DTW.[2]

A Figura 4a representa a forma como uma série de distâncias comuns, como a Euclideana, alinham pontos (i a i) de séries diferentes que resultaram em valores de pouca similaridade para determinadas séries temporais. O cálculo da distância DTW, por outro lado, distorce os índices temporais de forma a encontrar índices vizinhos do índice i de uma série temporal que mais se aproxime do índice i da série temporal com que se está a comparar, como é possível observar na Figura 4b.

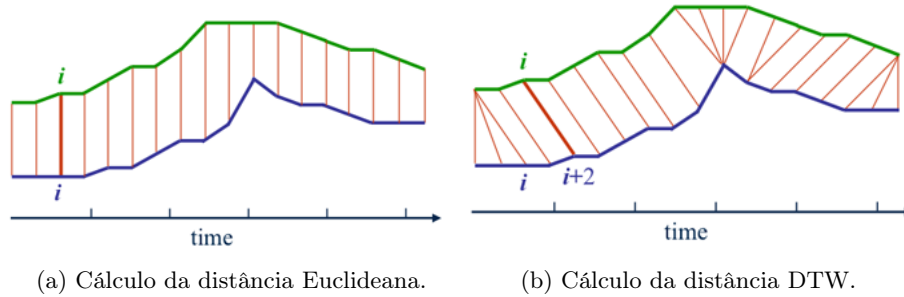


Figura 4: Comparação entre os cálculos da distância Euclideana e DTW.[2]

3.6 Análise de Change-points

Change-points, ou pontos de mudança, são índices temporais que dividem dois intervalos com características diferentes, correspondentes a instantes a partir dos quais ocorreu uma mudança no comportamento da variável sob inspeção. Os *change-points* podem representar alterações na média,

variância, em modelos de regressão linear ou ainda em combinações das características anteriores. A análise de *change-points* procura estudar alterações bruscas no comportamento das séries temporais, tendo como objetivo final a sua deteção, isto é, localizar o momento em que ocorreu a mudança e, também, quantificar o grau de certeza da sua ocorrência. A Figura 5 apresenta exemplos de séries temporais com vários tipos de *change-points*: na média (a), na variância (b), na média e na variância (c), na interceção de um modelo de regressão linear (d) e na interceção de um modelo de regressão linear e na tendência (e). Em (f) não existe nenhum *change-point* embora exista uma forte auto-correlação positiva.

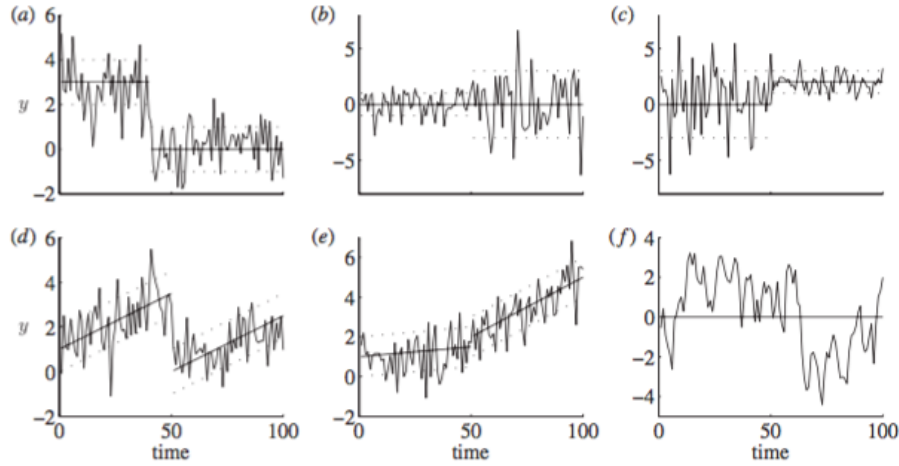


Figura 5: Exemplos de séries temporais com diferentes tipos de *change-points*. [3]

A maioria dos *change-points* são, no entanto, difíceis de detetar apenas pela observação visual da série temporal, pelo que serão apresentadas, de seguida, técnicas mais efetivas.

3.6.1 Hipótese

Previamente à estimação da localização de um *change-point* é necessário discernir se efetivamente ocorreu. À ocorrência de uma mudança pode, assim, ser atribuído um nível de confiança de ter ocorrido uma mudança. Caso este valor esteja acima de 90-95% é possível admitir que esta esteja presente. Uma forma de determinar este valor passa por efetuar uma análise da soma cumulativa (CUSUM) [52] da série temporal e de séries criadas a partir da original por *bootstrapping*. O teste de hipótese começa pela inicialização da primeira soma cumulativa, S_0 com o valor 0 e pelo cálculo recursivo das restantes através de

$$S_i = S_{i-1} + (X_i - \bar{X}), \quad (59)$$

onde S_{i-1} é a soma cumulativa anterior, X_i o valor no índice i e \bar{X} a média dos valores da série temporal. O gráfico deste conjunto de somas começa, portanto, em $S_0 = 0$, passa por um máximo absoluto e retorna a zero, $S_N = 0$. Após o cálculo das somas cumulativas obtém-se o estimador de

magnitude da mudança através de

$$S_{diff} = S_{max} - S_{min}, \quad (60)$$

com S_{max} sendo a soma cumulativa de maior valor e S_{min} a de menor valor. O valor S_{diff} da série original é comparado com valores calculados a partir de *bootstraps* da série original, contando-se o número de vezes em que S_{diff} é superior a $S_{diff}^{(i)}$ do *bootstrap* i . Dividindo este número pelo número total de *bootstraps*, que deve ser elevado para um teste significativo, obtêm-se o nível de confiança,

$$\text{Nível de confiança} = 100 \times \frac{X}{N} \% \quad (61)$$

A análise do gráfico CUSUM permite também tirar conclusões sobre a existência ou não de um ou mais *change-points*. Um gráfico com rampa pronunciada revela a existência de um *change-point*, sendo o valor de S_{diff} elevado, o que significa que terá existido uma alteração significativa no comportamento da série, contrariamente ao que se passa com um gráfico relativamente plano.

A deteção de *change-points* efetivos torna-se mais complicada devido à natureza sazonal, no caso do consumo de eletricidade. Será de prever que serão detetados *change-points* à noite, devido a uma redução brusca do consumo doméstico mas, no entanto, este facto repete-se todas as noites e não corresponde a uma mudança de rotina. O mesmo sucede ao longo da semana, sendo possível observar uma redução geral do consumo no fim-de-semana delimitada por dois *change-points*.

3.6.2 Deteção

Existem vários métodos de deteção de *change-points*. Os estimadores da sua localização baseados na média são os mais relevantes para o consumo doméstico de eletricidade, uma vez que a típica mudança de rotina de um consumidor resulta numa alteração do valor deste parâmetro. De entre os estimadores de mudanças na média destacam-se o relativo a somas cumulativas (CUSUM) e o relativo ao erro quadrático médio (MSE), pela sua simplicidade e o estimador Bayesiano, pela sua robustez.

3.6.2.1 CUSUM e MSE

O estimador CUSUM analisa as diferenças entre os pontos e a média que correspondem às somas cumulativas, sendo o índice da soma cumulativa de valor mais afastado de zero (Figura 6), o correspondente à localização do *change-point*. A estimação da localização será obtida através do valor de m que satisfaz

$$|S_m| = \max_{i=0, \dots, N} |S_i|. \quad (62)$$

O estimador MSE separa, em cada momento m , a série temporal em duas partes e compara os erros quadráticos médios antes e depois, sendo a localização do *change-point* o m que minimize o valor de MSE,

$$MSE(m) = \sum_{i=1}^m (X_i - \bar{X}_1)^2 + \sum_{i=m+1}^N (X_i - \bar{X}_2)^2, \quad (63)$$

onde

$$\bar{X}_1 = \frac{\sum_{i=1}^m X_i}{m} \quad \text{e} \quad \bar{X}_2 = \frac{\sum_{i=m+1}^N X_i}{N - m} \quad (64)$$

Na Figura 7 é apresentado um exemplo da detecção de *change-points* com os estimadores CUSUM e MSE.

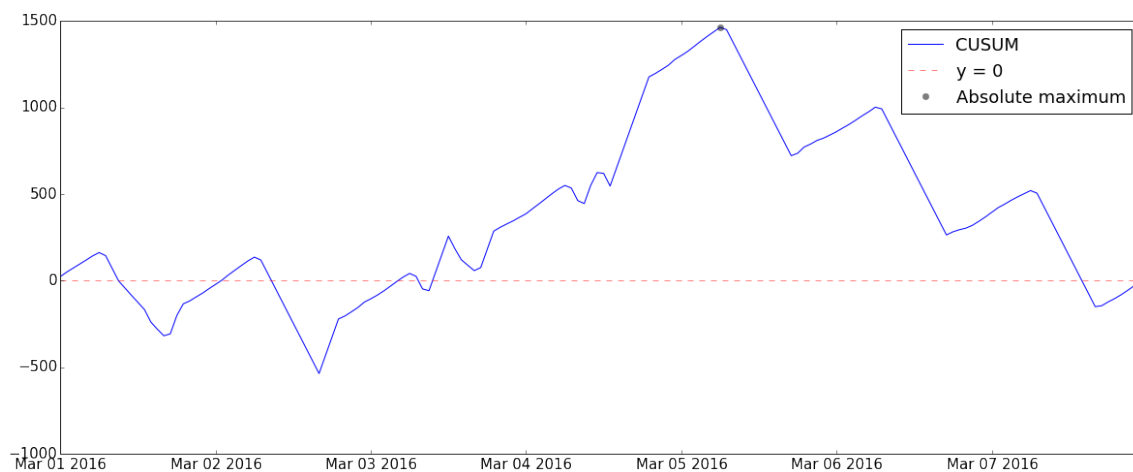


Figura 6: Soma cumulativa de uma série temporal com a indicação da linha de base ($y = 0$) e do máximo absoluto.

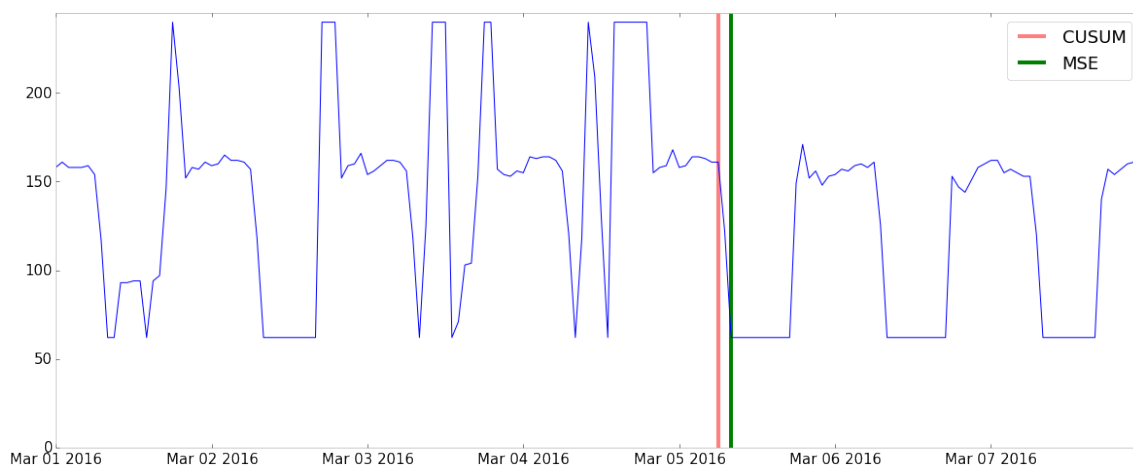


Figura 7: *Change-Points* detetados pelos estimadores CUSUM e MSE.

3.6.2.2 Inferência Bayesiana

Outra forma de detetar *change-points* passa pela inferência Bayesiana. Este método tem por base a atribuição de probabilidades a valores de uma variável aleatória.

Para a deteção de *change-points* por inferência Bayesiana começa-se por modelar o problema, ou seja, elaborar as distribuições de probabilidade das variáveis aleatórias. O consumo energético em Watts representa uma variável inteira discreta podendo ser representada por

$$C_i \sim \text{Poisson}(\lambda),$$

sendo C_i o consumo para a hora i . Na deteção de *change-points* assume-se que a distribuição da variável do consumo prévia ao momento do *change-point*, τ seja diferente da distribuição imediatamente após, o que significa que existirão dois λ tais que

$$\lambda = \begin{cases} \lambda_1 & \text{se } t < \tau \\ \lambda_2 & \text{se } t \geq \tau \end{cases}. \quad (65)$$

A modelação dos λ , variáveis contínuas, pode ser efetuada através da distribuição exponencial:

$$\lambda_1 \sim \text{Exp}(\alpha)$$

$$\lambda_2 \sim \text{Exp}(\alpha).$$

O parâmetro α denomina-se de hiper-parâmetro por ser progenitor de outro parâmetro, neste caso λ . Este parâmetro é normalmente iniciado como o inverso da média dos valores de consumo da série a analisar, uma vez que o valor esperado de uma distribuição exponencial é o inverso de λ . A última variável a modelar corresponde à localização do *change-point* no espaço temporal. Uma vez que os intervalos de amostragem das séries de consumo constituem intervalos regulares e completos, pode-se modelar τ com a distribuição discreta uniforme,

$$\tau \sim \text{DiscreteUniform}(1, N)$$

por este ser um valor discreto (índice temporal que corresponderá a uma certa hora de um certo dia), que varia entre a hora inicial e o número total de horas da série temporal, N . Nesta distribuição, uma vez que não existe nenhuma informação sobre onde o *change-point* possa estar localizado, atribui-se um valor de crença *a priori*, uniforme para todos os valores desta distribuição,

$$P(\tau = k) = \frac{1}{N} \quad (66)$$

O algoritmo que permite estimar λ_1 , λ_2 e τ utiliza estimativas *a priori* destes parâmetros, sendo obtidas amostras através do método Markov-Chain Monte Carlo. Este método está descrito em detalhe, por exemplo em [53].

Na Figura 8 surge um exemplo da distribuição de cada uma das três variáveis, λ_1 (consumo esperado antes do *change-point*), λ_2 (consumo esperado depois do *change-point*) e τ (localização do *change-point*).

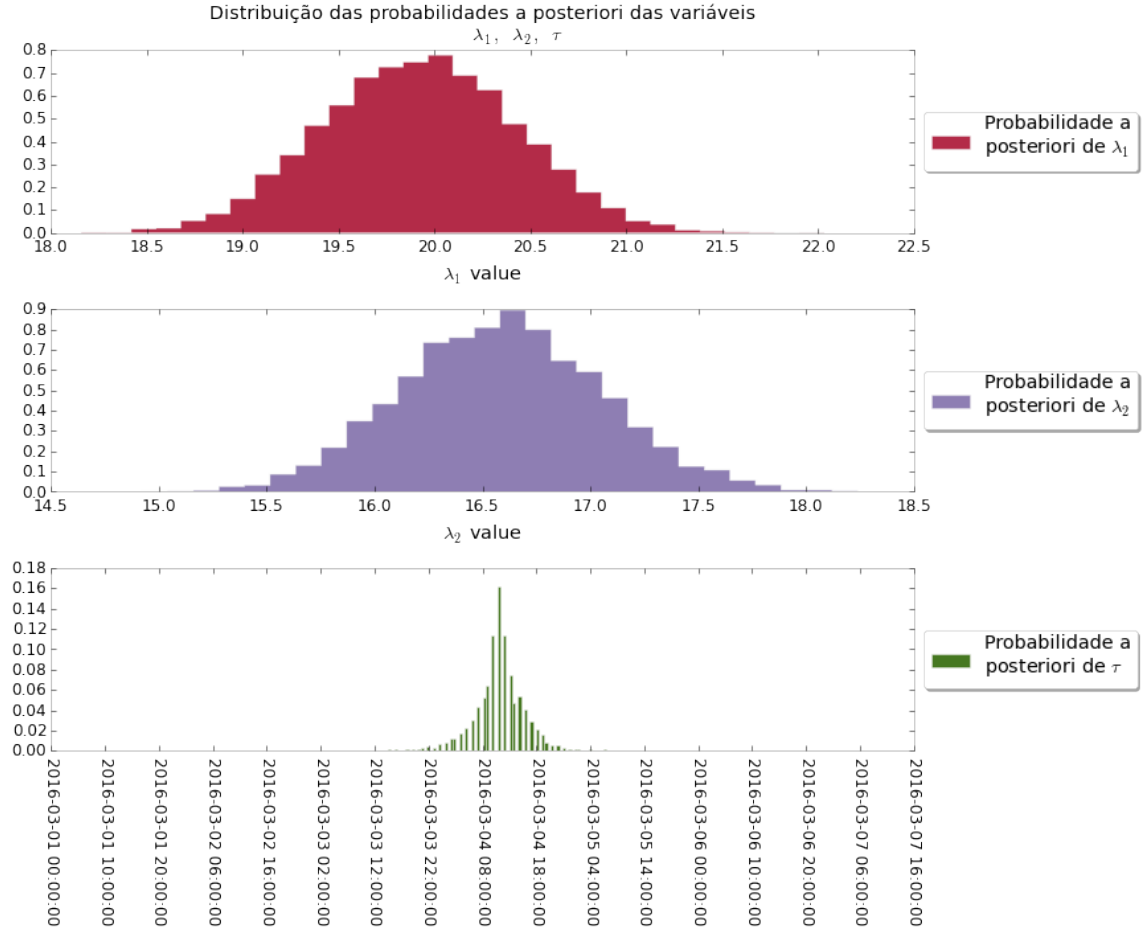


Figura 8: Distribuições das variáveis λ_1 , λ_2 e τ .

O valor esperado de uma dada distribuição de probabilidades refere-se ao valor médio para qual a variável tende, com o aumento das amostras disponíveis. Este será o indicador a analisar para a existência de um *change-point* na média, uma vez que o valor esperado do consumo irá ser diferente previamente e após o *change-point*. Sabendo que o valor esperado de uma distribuição de Poisson é equivalente ao parâmetro λ dessa distribuição, será necessário obter os λ para cada instante t . Apresenta-se na Figura 9 um exemplo de detecção de *change-points* por inferência Bayesiana.

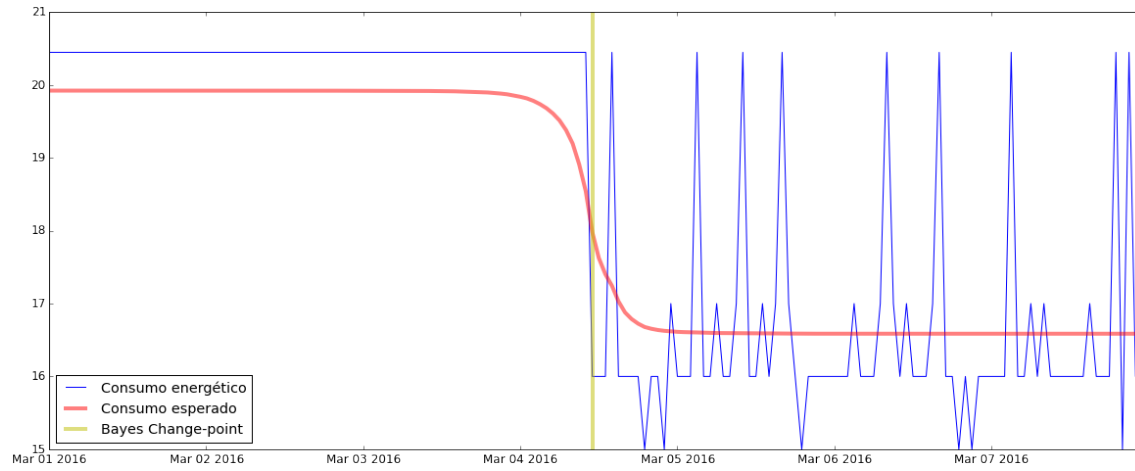


Figura 9: Sobreposição do consumo esperado e do consumo real, com o *change-point* detetado assinalado a amarelo.

3.7 Avaliação da previsão

A comparação das várias previsões, produto de diversas configurações e parâmetros dos modelos de previsão, foi efetuada através da análise de testes e métricas relevantes. No que respeita aos primeiros, utilizou-se o teste Diebold-Mariano[54], cuja hipótese nula é que não existe diferença entre as funções de perda

$$\frac{1}{m} \sum_{t=n+1}^{n+m} [(Y_t - \hat{Y}_t^{(1)})^2 - (Y_t - \hat{Y}_t^{(2)})^2], \quad (67)$$

onde n corresponde ao índice do ponto fixo, m ao número de pontos na janela de previsão, Y_t ao valor da função original no índice de previsão t e $\hat{Y}_t^{(i)}$ ao valor da previsão do modelo i no índice de previsão t . Caso este valor seja inferior a zero e o teste t de *Student* indicar um valor de $p/2$ ¹⁶ inferior a 0.05¹⁷, a hipótese nula é rejeitada e conclui-se que a previsão do primeiro modelo ($\hat{Y}^{(1)}$) terá sido melhor que a do segundo modelo, caso contrário aceita-se a hipótese nula e nada se pode concluir. Considera-se que terá capacidade de previsão significativamente superior caso o valor de p seja inferior a 0.05. O teste t de *Student* é efetuado sobre

$$d = (Y_t - \hat{Y}_t^{(1)})^2 - (Y_t - \hat{Y}_t^{(2)})^2 \quad (68)$$

com 0 como o valor esperado na hipótese nula. Para a realização deste teste, por exemplo na escolha de um modelo de previsão, são primeiro ordenadas as previsões dos modelos a comparar pelo valor crescente da média dos erros de previsão quadráticos (MSPE), obtidos por

$$MSPE = \frac{1}{m} \sum_{t=n+1}^{n+m} (Y_t - \hat{Y}_t^{(1)})^2. \quad (69)$$

Após esta ordenação, são apenas comparados as primeiras duas previsões com menor erro, uma vez que a previsão que é potencialmente mais precisa é a primeira (e nunca poderá ser considerada pior que as restantes). Caso não seja estabelecida como aquela que é significativamente melhor em comparação com a seguinte, não existirá nenhum modelo cuja previsão seja efetivamente superior a todos os restantes.

Numa outra perspetiva, a comparação direta de métricas, sem atribuir significado estatístico, recorreu ao erro absoluto médio ou MAE (do inglês *Mean Absolute Error*), raiz do erro quadrático médio ou RMSE (do inglês *Root Mean Squared Error*), erro médio absoluto em percentagem ou MAPE (do inglês *Mean Absolute Percentage Error*), erro médio absoluto escalado ou MASE (do inglês *Mean absolute scaled error*) e à métrica de distância com deformação dinâmica temporal ou DTW (do inglês *Dinamic Time Warping*), abaixo definidos:

$$MAE = \text{Avg}(|\text{orig} - \text{pred}|) \quad (70)$$

$$RMSE = \sqrt{\text{Avg}((\text{orig} - \text{pred})^2)} \quad (71)$$

¹⁶O valor de p é dividido por 2 de forma a representar o valor de um teste de um cauda necessário para o problema em questão.

¹⁷Foi utilizado o intervalo de confiança de uma cauda de 95%.

$$\text{MAPE} = 100 \times \frac{\text{Avg}(|orig - pred|)}{\text{Avg}(orig)} \quad (72)$$

$$\text{MASE} = \text{Avg}\left(\frac{|orig - pred|}{Q}\right), \quad (73)$$

onde

$$Q = \frac{1}{\text{len}(\text{pred}) - 1} \times \sum_{i=1}^{\text{len}(\text{train})} \text{train}_i \quad (74)$$

Para algumas experiências foi utilizada uma combinação dos dois tipos de avaliação, com o teste de Diebold-Mariano a ser utilizado para obter uma confirmação mais rigorosa das conclusões obtidas das métricas. Por proporcionar só por si pouca informação (raras são as vezes que certas configurações de previsão serão significativamente superiores às restantes) o teste Diebold-Mariano foi sempre utilizado em paralelo com uma avaliação por métricas.

4 Procedimentos Experimentais

Para o desenvolvimento deste projeto efetuou-se o estudo das séries temporais de consumo presentes neste conjunto de dados. Foram selecionadas 1000 séries temporais de consumo para a caracterização do conjunto de dados sem qualquer modificação ou pré-processamento. A caracterização consistiu na identificação das categorias existentes nessa amostra de séries temporais, bem como no número de séries temporais que se incluem em cada categoria, na análise dos valores extremos de consumo energético, por meio do cálculo dos mínimos e máximos de cada categoria, na distribuição da potência contratada, e na distribuição do consumo médio e respetivo desvio padrão. Uma vez que se irá restringir o âmbito deste estudo ao consumo doméstico, foram selecionadas 100 séries temporais de dados de consumo pertencentes a essa categoria, que sofreram uma caracterização semelhante à do conjunto de 1000 séries de tipos de clientes mistos, com a adição da comparação do consumo médio e desvio padrão de séries temporais com e sem *outliers* por meio de um histograma. Estas séries, após serem pré-processadas para a eliminação de *outliers*, preenchimento de valores em falta e alteração da granularidade para as três consideradas neste trabalho (horária, octa-horária e diária), serão utilizadas para a elaboração dos modelos de previsão, *clustering* e deteção de *change-points*.

Qualquer previsão depende de uma série de fatores relacionados com o modelo de previsão, a série utilizada para ajustar os parâmetros e ainda a janela de previsão. Estes fatores não são independentes entre si, sendo necessário efetuar um compromisso de forma a ser possível elaborar uma sequência de experiências. Esta sequência teve origem num conjunto de suposições baseadas na intuição e em conclusões do estado da arte. Foi, então, definida uma sucessão de experiências a realizar de forma a obter resultados, progredindo sequencialmente para a escolha dos parâmetros que otimizam a previsão. Estas experiências procuraram definir a melhor composição dos seguintes fatores

- sazonalidade,
- modelo de previsão,
- tamanho da janela de histórico,
- tamanho da janela de previsão,
- introdução ou não de variáveis exógenas relevantes,
- tipo de histórico,
- filtros e transformações ao histórico.

O estado da arte sugere vários modelos eficientes para a previsão, com especial ênfase para os modelos sazonais. Estes modelos exigem a determinação de um parâmetro de sazonalidade para a sua utilização. Para tal será efetuada a determinação do período de sazonalidade para cada uma das granularidades horária, octa-horária e diária. Previamente à sua comparação, será efetuado um estudo da qualidade de previsão de dois modelos não sazonais, ARMA e ARIMA, e três modelos sazonais, SARIMA, Holt-Winters aditivo e Holt-Winters multiplicativo, através da análise dos valores médios das métricas de previsão, para o conjunto de 100 séries do consumo doméstico. Seguidamente, serão sujeitos a testes de capacidade e comparados, de forma a obter o

modelo com melhor capacidade a ser utilizado para as experiências subsequentes. Esta comparação será constituída pela avaliação das métricas, determinando o número de vezes em que cada modelo possui um valor de erro inferior aos restantes e através do teste estatístico de Diebold-Mariano.

Para a seleção da janela de histórico e previsão, foram testados pares de tamanhos de janela de histórico e de previsão baseados na sazonalidade detetada. De forma a reduzir a quantidade de comparações a realizar, foram escolhidos três tamanhos de janelas de histórico. As janelas de histórico utilizadas foram 3, 4 e 5 períodos de sazonalidade para a granularidade horária e octa-horária (com uma margem acrescentada de 15 pontos para a primeira janela de histórico e 5 para as restantes) e 6, 7 e 8 (com uma margem acrescentada de 5 pontos para todas) para a diária. A escolha das janelas de histórico baseou-se num estudo do mínimo de amostras necessárias para modelos sazonais de previsão [55], que indica que serão necessários tamanhos de histórico de $m + 5$ valores para o método Holt-Winters e $p + d + q + P + Q + mD$ para o SARIMA, onde p , d , q , P , D e Q se referem às ordens do modelo e m ao período de sazonalidade (representado por Saz nas seguintes considerações). Uma vez que as ordens máximas para o SARIMA estão definidas como 3 e o valor mínimo para este modelo é sempre superior ao do mínimo para o Holt-Winters, o mínimo de valores a conter no histórico seria

- Para a granularidade horária:

$$p + d + q + P + Q + \text{Saz}_{\text{horária}} \times D = 3 + 3 + 3 + 3 + 3 + 3 \times 24 = 3 \times \text{Saz}_{\text{horária}} + 15$$

- Para a granularidade octa-horária:

$$p + d + q + P + Q + \text{Saz}_{\text{octa-horária}} \times D = 3 + 3 + 3 + 3 + 3 + 3 \times 21 = 3 \times \text{Saz}_{\text{octa-horária}} + 15$$

- Para a granularidade diária:

$$p + d + q + P + Q + \text{Saz}_{\text{diária}} \times D = 3 + 3 + 3 + 3 + 3 + 3 \times 7 = 5 \times \text{Saz}_{\text{diária}} + 1$$

As quatro janelas de previsão escolhidas baseiam-se em frações da sazonalidade: 1/6, 1/3, 1/2 e 1, uma vez que janelas de previsão superiores à do período de sazonalidade implicam a integração de pontos previstos para o cálculo da componente de sazonalidade dos modelos o que origina uma propagação do erro. De notar que o tamanho da janela de previsão é um número inteiro pelo que será utilizado o inteiro mais próximo do valor da fração pelo limite inferior. De forma a efetuar testes consistentes, realizados a partir de alguma informação comum, foi definido um ponto fixo, para todas as séries e para todos os pares de janelas de histórico e previsão, que assinala o fim da janela histórico e o início da janela de previsão. É a partir deste ponto fixo que se seleciona o número de pontos anteriores que constituirão o histórico e os pontos a partir do qual incidirá a previsão.

Vários fatores como a introdução de variáveis exógenas, aplicação de filtros e transformações nas séries temporais e, ainda, a alteração do tipo de histórico, podem influenciar a qualidade de previsão e serão o alvo das análises que se seguem. A granularidade utilizada para as seguintes experiências foi a do consumo horário, com a janela de histórico e previsão selecionada acima, para o modelo com melhor desempenho, o SARIMA, com a sazonalidade de valor 24.

Dos fatores que influenciam a previsão, o mais interessante é a incorporação de variáveis exógenas na previsão do consumo energético, como a temperatura e a humidade. A presença de vários

dispositivos que gerem as condições de uma habitação, como sejam aparelhos de ar-condicionado e desumidificadores, procuram contrariar efeitos atmosféricos exteriores, proporcionando o maior conforto aos habitantes. Desta forma, é possível assumir que existirá uma certa ligação entre aumentos ou diminuições no consumo energético e as respostas destes dispositivos às temperaturas ou níveis de humidade devolvidos pelos seus sensores. Estas alterações são reveladas no consumo energético, podendo existir uma correlação entre consumo e fatores ambientais. Para verificar se a introdução destas variáveis exógenas melhora a previsão de uma mesma série temporal, foram utilizados dados recolhidos de temperatura e humidade para o intervalo de tempo da série temporal. Uma vez que nem todos os distritos possuem estações meteorológicas, certos valores foram obtidos a partir da estação meteorológica mais próxima. Todos estes dados foram recolhidos do *site* UndergroundWeather[56]. De forma a interpretar melhor a correlação entre as variáveis exógenas e o consumo, para o conjunto de séries temporais de consumo doméstico, serão elaborados histogramas dos valores da correlação, separando-os em valores positivos e valores negativos. Será, ainda, calculada, para cada distrito, a correlação que as variáveis exógenas possuem entre si. A análise das métricas de erro e o teste Diebold-Mariano foram utilizados para a determinação da contribuição de variáveis exógenas na melhoria da qualidade de previsão.

Os históricos utilizados até ao momento consistiram na sequência correta e correspondente à da data em que os valores foram efetivamente recolhidos. No entanto, a análise de séries temporais de consumos domésticos reais revela que existe uma certa repetição da rotina para dias da semana iguais, o que sugere que a previsão possa ser melhorada com um histórico modificado, constituído apenas pelo mesmo dia da semana que aquele que se deseja prever. Um histórico deste tipo para a previsão de um sábado seria constituído apenas por valores do consumo em sábados anteriores ao da previsão. Uma vez que os modelos utilizados neste projeto apenas são compatíveis com séries temporais sem interrupções (que neste caso seriam todos os dias da semana que não são sábado) foi mantida a hora do dia do *timestamp* mas foi alterado o dia, de forma a corresponderem às datas de, por exemplo, se se quisesse utilizar como histórico os 7 sábados anteriores, as datas seriam aquelas de uma semana sequencial correspondente à semana anterior ao início da previsão. Para este teste foram criados históricos de igual tamanho: histórico de um mesmo tipo de dia da semana e histórico sequencial. De forma semelhante às experiências anteriores, também as previsões originadas através destes dois tipos de histórico foram comparadas através da análise das métricas de erro e da aplicação do teste Diebold-Mariano.

Os filtros também podem influenciar a eficiência da previsão, no sentido em que permitem reduzir o ruído de uma certa série temporal, podendo melhorar a previsão da série real através do ajuste dos modelos a uma série filtrada. Para este estudo utilizaram-se os filtros Baxter-King bandpass, Hodrick-Prescott e Holt-Winters e efetuaram-se previsões para cada uma das séries com e sem filtros.

De forma semelhante aos filtros, as transformações também podem melhorar a previsão, no sentido em que representam séries temporais mais fáceis de modelar por parte dos modelos de previsão. As transformações utilizadas neste estudo foram: Box-Cox, logaritmização (Log) e aplicação da raiz quadrada (Sqrt). Os modelos são ajustados a partir de um histórico transformado, do qual resulta uma previsão transformada. Estas transformações são reversíveis, pelo que, para obter a previsão efetiva, é apenas necessário aplicar a transformação inversa.

Uma questão importante, tratada no projeto, deriva do facto de o serviço de previsão vir a ser

potencialmente utilizado por um conjunto de clientes numeroso e heterogéneo. Este aspeto leva à necessidade de abordar formas de permitir a escalabilidade e ter em atenção a volatilidade associada à atividade humana. Embora cada pessoa tenha a sua rotina, é possível identificar rotinas típicas em que cada um se enquadrará. Explorando este aspeto, será possível identificar grupos de séries temporais e fazer uma previsão geral para cada grupo, sem o afastamento significativo relativo à previsão de cada série individualmente considerada. Um *cluster* de séries temporais reúne séries temporais semelhantes constituindo, o seu centróide, a série temporal do consumo correspondente a uma rotina específica. A previsão com base no centróide permite reduzir o número de séries temporais a prever, uma vez que o resultado se aplicaria a todas as séries temporais desse *cluster*. Foram comparados os erros da previsão do centróide em relação a cada série temporal com os erros obtidos pela previsão de cada série individual. O algoritmo de *cluster* escolhido foi o K-Means, com a função de semelhança DTW. Previamente ao *clustering* através do algoritmo K-Means, procedeu-se à escolha do número de *clusters* ótimo, entre 2 e 10, recorrendo à maximização do parâmetro \mathcal{E} constante na Equação 55. De forma a abordar o *clustering* para a previsão em bloco foram utilizadas séries temporais sem qualquer alteração. Uma vez que o objetivo para o *clustering* não-normalizado seria o da previsão em bloco de séries temporais, os conjuntos das séries temporais a agrupar terão o tamanho do histórico definido anteriormente. Será efetuada a previsão do tamanho do horizonte temporal selecionado em testes anteriores associado à granularidade das séries com a sazonalidade respetiva, através do modelo SARIMA. Foram efetuados *clusters* para cada uma das granularidades das séries temporais de consumo doméstico. Ao utilizarem-se séries não-normalizadas pretendeu-se fazer uma comparação em que tanto os detalhes da sequência temporal como o valor médio do consumo são levados em conta. Ao longo desta secção serão comparados os erros médios de previsão de 100 séries temporais com previsões baseadas nos centróides a que pertencem (previsão em bloco) e com previsões individuais. Por outro lado, para a elaboração da tipificação do consumidor através do reconhecimento de padrões semelhantes, foi necessária a aplicação de uma normalização a todas as séries do conjunto previamente ao *clustering*. O resultado deste *clustering* será a visualização e análise dos padrões encontrados, cuja representação é efetuada pelo centróide de cada *cluster*.

Existe uma certa dinâmica no dia-a-dia das pessoas que dificulta a previsão em bloco descrita acima. É esta volatilidade associada a alterações de rotina, à compra de novos dispositivos, entre outros fatores, que altera o paradigma da série a prever. Uma vez que os modelos de previsão são muito dependentes da série temporal, esta alteração brusca ou gradual nos padrões do consumo energético pode tornar obsoletos os modelos construídos anteriormente, por não se adaptarem à nova realidade. A deteção de *change-points* é, por conseguinte, imperativa por permitir a identificação desses momentos de alteração e permitir manter a eficiência da previsão, atualizando o modelo de previsão. Existem vários métodos para a deteção dos diversos tipos de alterações em séries temporais. Foi selecionada a deteção de *change-points* na média, pois os eventos de alteração típicos de um consumo residencial são abrangidos por esta categoria. Os métodos analisados foram a estimação pela soma cumulativa (CUSUM), pela média dos erros quadráticos (MSE) e pela inferência Bayesiana. De forma a estudar a capacidade de identificação destes métodos foi efetuada a deteção de *change-points* com uma janela de 7 dias de um consumo horário, que se move de dia em dia no período de um mês. O impacto dos *change-points* nos modelos apresentados será evidenciado por um exemplo constituído por três previsões com diferentes históricos delimitados por *change-points* que indicam a diferença de consumo entre os dias da semana e um fim-de-semana. A primeira previsão será efetuada com um histórico com limite superior na localização do primeiro *change-point* (início do fim-de-semana) para a previsão de um sábado, a segunda previsão utiliza

um histórico de um sábado (com limite inferior no primeiro *change-point*) para prever um domingo e a última previsão utiliza o histórico do fim-de-semana (delimitado pelos dois *change-points*) para prever uma segunda-feira.

5 Resultados

O desenvolvimento deste projeto tem como objetivo a análise de séries temporais de consumo doméstico real, a comparação das várias componentes da previsão, incluindo o tipo de modelo, histórico e janela de previsão, e a agrupamento das séries temporais. Obter-se-á, assim, um conjunto de parâmetros visando a maximização do desempenho da previsão. Com esse fim, compararam-se modelos de previsão, analisaram-se janelas de histórico e de previsão, focando cada uma das granularidades (horária, octa-horária e diária), testaram-se vários tipos de histórico, filtros e transformações e efetuaram-se análises a *clusters* de séries temporais. Apresentam-se a seguir os resultados fruto das experiências descritas na Secção Procedimentos Experimentais.

5.1 Pré-processamento

Os modelos de previsão de séries temporais assumem que estas sejam constituídas por uma sequência regular com dados correspondentes e valores com significado. Para assegurar estas condições é necessário proceder ao tratamento de *outliers* e ao preenchimento dos valores em falta. No contexto do problema, não se irá recorrer a uma granularidade inferior a uma hora, tendo sido selecionadas as granularidades horária, octa-horária e diária. O consumo horário permitirá identificar padrões ao longo do dia, enquanto que o consumo octa-horário identificará os comportamentos nas diversas partes do dia: 0 a 8 horas para a noite, 8 a 16 para a concentração da atividade laboral, 16 a 24 para as atividades ao fim do dia. O consumo diário identificará os padrões dos diferentes dias da semana, e a sua alteração ao longo de cada mês. A re-amostragem das séries temporais é efetuada através da substituição dos valores individuais, obtidos a cada 15 minutos pela sua média ao longo do período considerado, consistindo a Figura10 num exemplo disso.

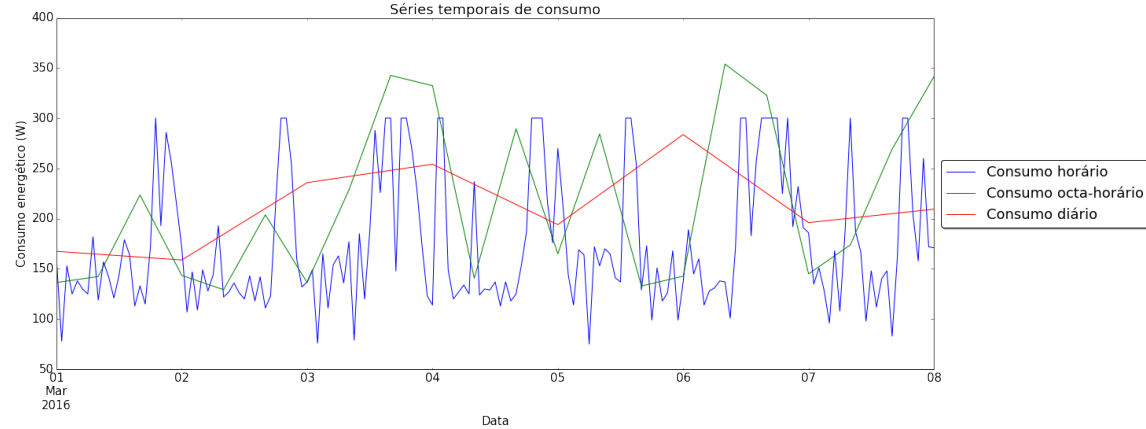


Figura 10: Exemplo de uma série temporal com as granularidades horária, octa-horária e diária de consumo energético.

Os valores de consumo são considerados *outliers* caso se encontrem fora dos intervalos

$$[\text{Média} - 3 \times \text{STD}, \text{Média} + 3 \times \text{STD}]$$

ou

$$[\text{Média} - 3 \times \text{MAD}, \text{Média} + 3 \times \text{MAD}],$$

com STD (do inglês *Standard Deviation*) sendo o valor do desvio padrão da série e MAD (do inglês *Mean Absolute Deviation*) o valor do desvio padrão absoluto. Um exemplo destes intervalos para uma série temporal está ilustrado na Figura 11. Para dados com distribuições normais MAD e STD relacionam-se por $\text{STD} = 1.253 * \text{MAD}$. Será utilizado ao longo do trabalho a definição de *outlier* por MAD e não por STD, uma vez que o primeiro tem tendência a ser mais restritivo. Os *outliers* identificados são substituídos pelo valor limite mais relevante do intervalo definido acima.

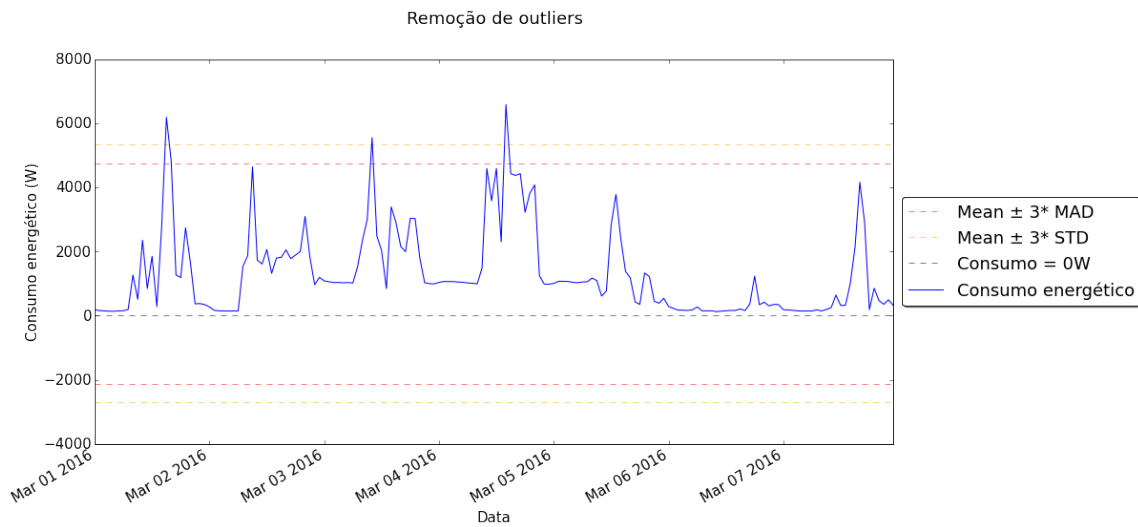


Figura 11: Exemplo de limites STD e MAD na remoção de *outliers* de um série temporal.

Os valores em falta são preenchidos pela média dos x pontos anteriores. O valor de x utilizado para cada uma das granularidades de consumo horário, octa-horário e diário corresponde ao seu período de sazonalidade (Figura 12). Caso não seja possível obter uma sequência de pontos anteriores do tamanho desejado, o valor é substituído pela média da série temporal. Foi aplicado um limiar de 10% de valores em falta a partir do qual a série não será utilizada.

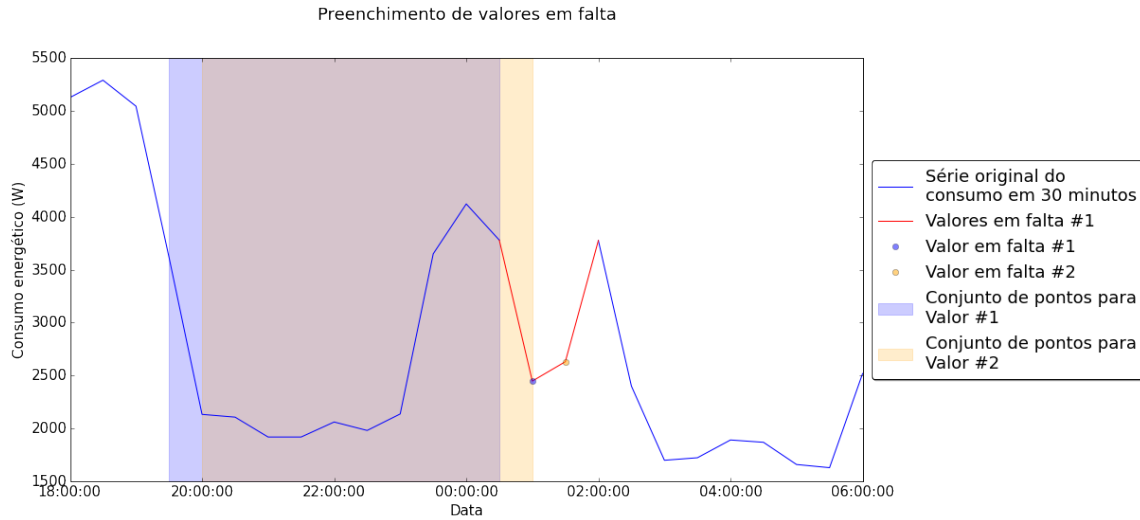


Figura 12: Preenchimento de dois valores em falta de uma série temporal de consumo energético de granularidade 20 minutos.

Várias séries temporais apresentam apenas um padrão de consumo linear, mantendo o mesmo valor de consumo ao longo do período da amostra. Este tipo de consumo é detetado através da análise do desvio-padrão que será muito reduzido. Séries temporais com estas características ($std \simeq 0$) não serão, também, utilizadas para análises futuras.

5.2 Caracterização do conjunto de dados

No contexto da atividade e parecerias da empresa, foi disponibilizado um conjunto de dados de consumo. Estes dados foram recolhidos ao longo do ano 2015 e início do ano 2016, em intervalos de 15 minutos, e constituem uma coleção de consumos reais recolhidos por contadores energéticos inteligentes instalados em vários pontos do país. Os dados são heterogêneos, possuem um volume significativo e são de elevada relevância para o objetivo deste projeto. Cada valor de consumo é apresentado em Watts, sendo a granularidade de 15 minutos. A natureza destes dados permite construir modelos que se adequam à realidade do consumo energético doméstico e permite obter resultados, identificar problemas e ultrapassar obstáculos que não surgiriam com dados de natureza mais controlada.

5.2.1 Consumo de clientes mistos

Foram selecionadas 1000 séries temporais correspondentes a cerca de 4 meses de consumo energético. Estas séries dividem-se em quatro grupos diferentes, de acordo com o tipo de consumidor: hotelaria, indústria, domésticos e “outros”. Nesta amostra de 1000 consumidores, a maior parte (46%) são consumidores domésticos, como é possível observar no gráfico da Figura 13.

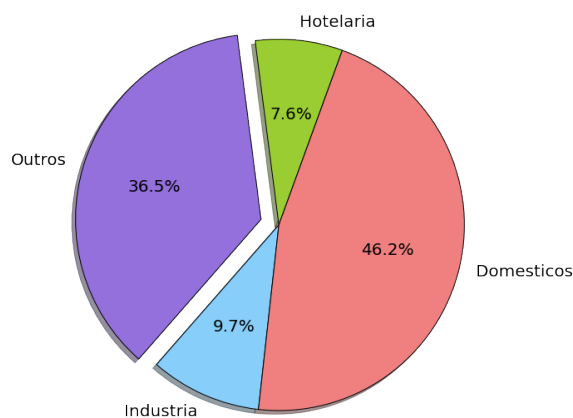


Figura 13: Gráfico circular de categorias.

É possível ainda organizar as séries temporais de acordo com a potência contratada (1kVA \sim 1000W), como está representada na Figura 14. De uma forma geral [57], o escalão 1.15 kVA é apropriado a uma instalação constituída apenas por pouca iluminação e um aparelho de fraca potência, o escalão 3.45 kVA permite iluminação e poucos aparelhos de baixa potência ligados em simultâneo ou um aparelho de maior potência, como uma máquina de lavar, o escalão 6.9 kVA permite dois aparelhos de potência moderada, o escalão 10.35 kVA permite quatro aparelhos de potência moderada e os escalões superiores a este permitem numerosos aparelhos de potência elevada em funcionamento simultâneo. Uma análise preliminar de cada conjunto revela que os mínimos e máximos de consumo energético variam entre os 0 W e os 40 000W, sendo a hotelaria

caracterizada por um menor valor máximo (24308 W) sendo o maior máximo (39692 W) encontrado no consumo industrial.

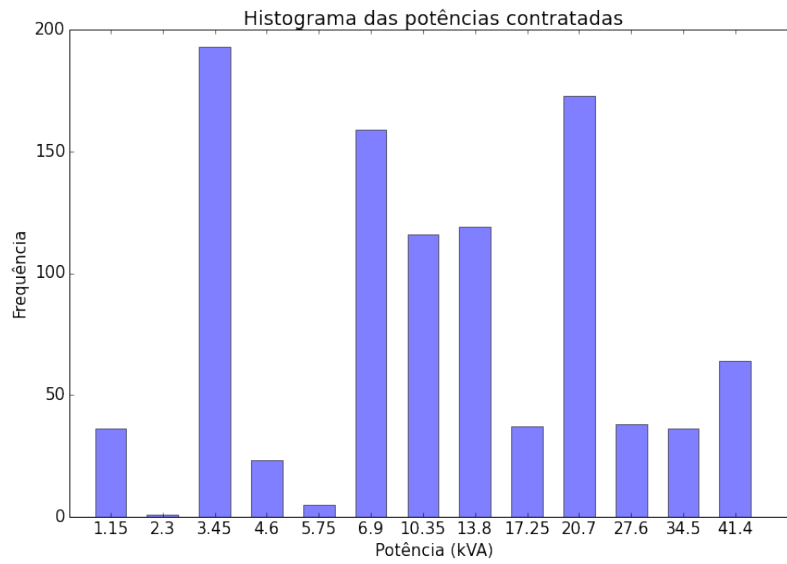
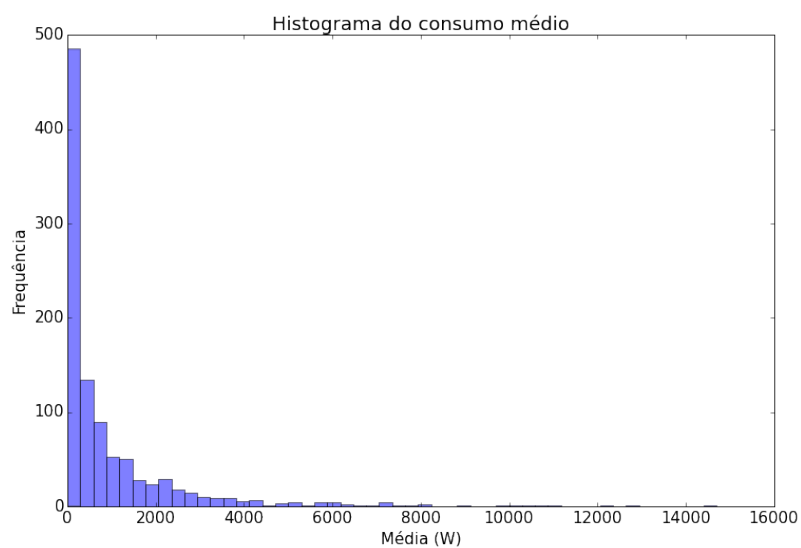
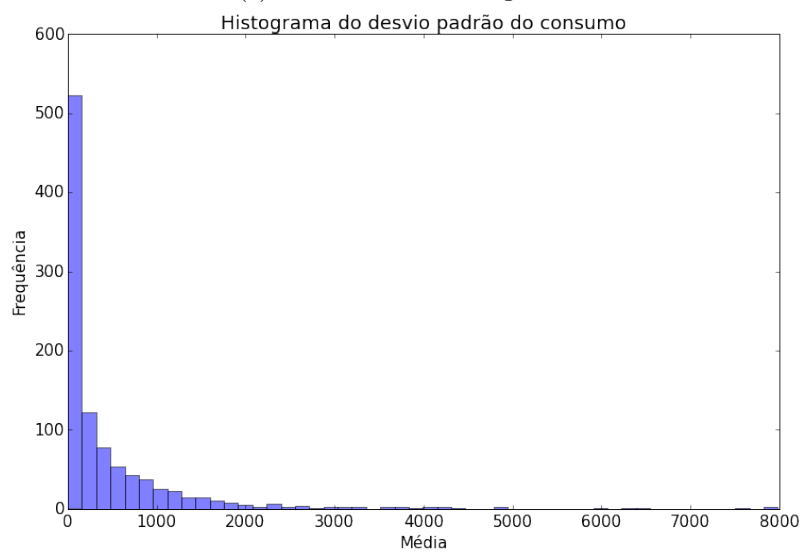


Figura 14: Histograma das potências contratadas para todos os tipos num conjunto de 1000 consumidores.

O histograma do consumo médio (Figura 15a) revela uma concentração elevada de médias abaixo dos 2000W, o que pode ser justificado pela porção elevada de consumidores domésticos e de hotelaria presente na amostra em relação ao consumidor industrial.



(a) Consumo médio energético.



(b) Desvio padrão.

Figura 15: Histogramas da média e desvio padrão do consumo energético num conjunto de 1000 consumidores.

5.2.2 Consumo doméstico

Analisando apenas as potências contratadas pela categoria do consumo doméstico, é possível elaborar o histograma da Figura 16. A maior parte das potências contratadas por consumidores domésticos encontra-se abaixo do escalão 4.6 kVA. Este facto reflete-se no consumo médio que não excede os 3000 W.

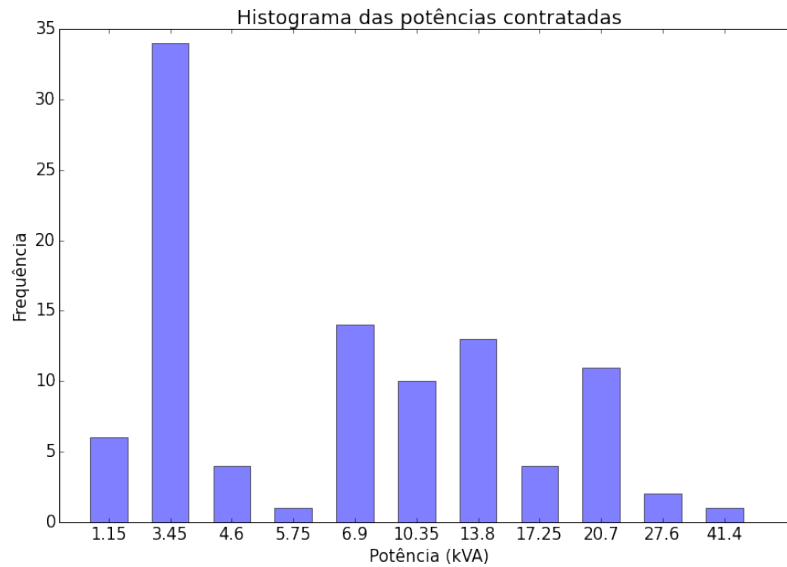
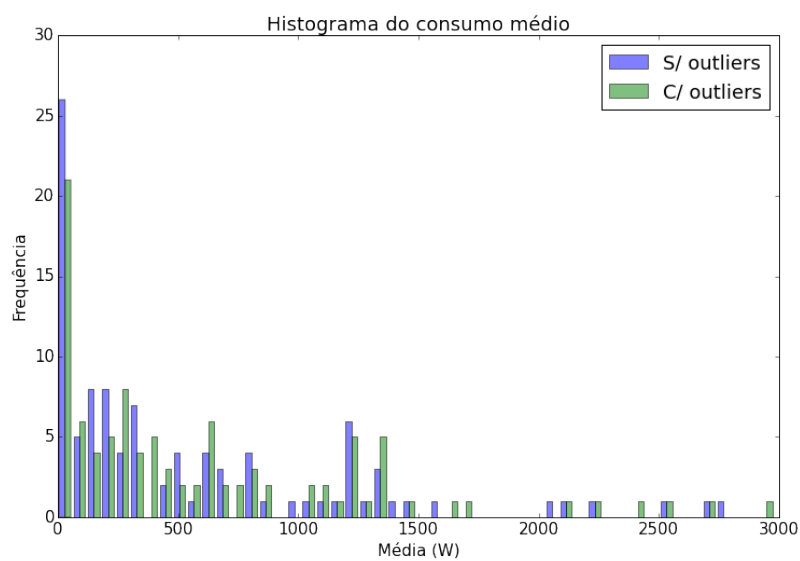
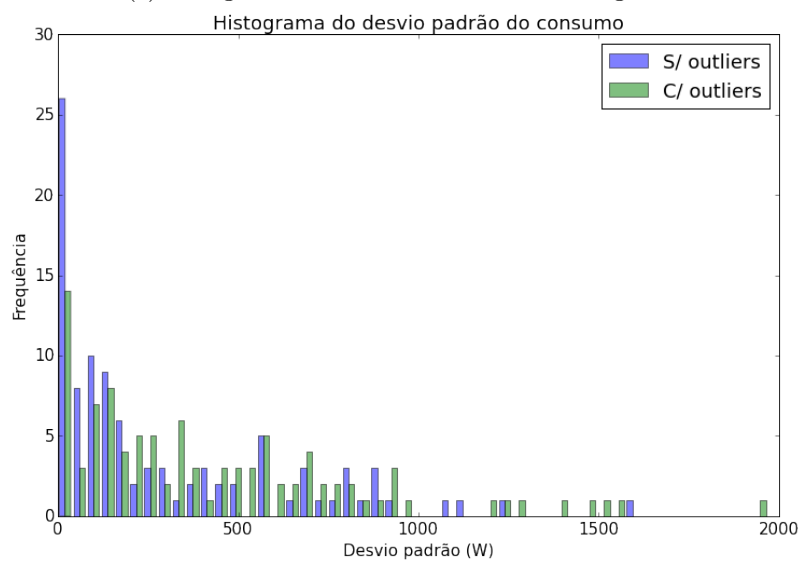


Figura 16: Histograma das potências contratadas de um conjunto de 100 consumidores domésticos.

Todas as séries temporais deste conjunto contêm valores de consumo durante um período de 4 meses e sofreram uma amostragem de granularidade horária, octa-horária e diária. Foram ainda retirados *outliers* e preenchidos valores em falta. A verificação da influência do tratamento de *outliers* nas séries temporais de consumo doméstico foi efetuada através da comparação das distribuições das médias e dos desvios-padrão do consumo energético doméstico. A Figura 17 apresenta o histograma da média e desvio padrão para as séries temporais de consumo doméstico com e sem *outliers*, onde é possível observar, para os dados sem *outliers*, uma regressão para valores mais reduzidos em relação aos dados com *outliers*. Conclui-se, assim, que os pontos discrepantes influenciavam significativamente os dados pelo que se optou pela sua remoção.



(a) Histogramas das médias do consumo energético.



(b) Histograma dos desvios padrão do consumo energético.

Figura 17: Histogramas da média e desvio padrão de um conjunto de 100 consumidores domésticos para séries temporais com e sem *outliers*.

5.3 Análise de séries temporais

A análise de séries temporais terá como objetivo identificar os melhores parâmetros para o modelo de previsão. Estas determinações são incrementais, sendo a conclusão da experiência anterior utilizada para a seguinte e assim sucessivamente, até serem recolhidas todas as configurações com melhor sucesso para compor o modelo de previsão final. Os parâmetros objeto de estudo são o período de sazonalidade para cada uma das granularidades, o modelo de previsão de entre ARMA, ARIMA, SARIMA, Holt-Winters aditivo e Holt-Winters multiplicativo, o tamanho das janelas de histórico e previsão para cada uma das granularidades, a incorporação de variáveis exógenas, tipos de histórico e a aplicação de filtros e transformações.

5.3.1 Sazonalidade

O estado da arte e a experiência pessoal das rotinas sugerem a existência de duas sazonalidades: a diária e a semanal. É de esperar, então, que a sazonalidade de um consumo horário seja de 24 pontos (24 horas), ou seja um padrão de consumo repete-se diariamente. Para o consumo octa-horário, tendo por base o mesmo princípio, seria de esperar uma sazonalidade de 3, correspondendo à noite, período de máxima atividade laboral e período de atividade mais reduzida correspondente ao intervalo de um dia que se repete para o seguinte. O consumo diário evoluirá ao longo da semana. Pelo estudo dos correlogramas identificaram-se os seguintes períodos de sazonalidade:

- **Consumo horário:** 24 períodos de 1 hora
- **Consumo octa-horário:** 21 (7 dias) períodos de 8 horas
- **Consumo diário:** 7 períodos de 1 dia

Ao contrário da hipótese de partida o consumo octa-horário apresenta uma sazonalidade que se pode interpretar como uma repetição semanal dos padrões de consumo.

5.3.2 Seleção do modelo

Os modelos utilizados para o seguinte estudo foram ARMA, ARIMA, SARIMA, Holt-Winters aditivo (HW ADD) e Holt-Winters multiplicativo (HW MULT). Na Figura 18 apresenta-se um exemplo de previsão de 12 pontos de uma série de consumo horário efetuada por estes modelos. Na parte superior surge o histórico utilizado para a modelação, bem como uma representação desta para os modelos ARMA e ARIMA e ainda a incorporação das previsões. Na parte inferior apresenta apenas os últimos quatro pontos do histórico e da modelação dos modelos ARMA e ARIMA e as previsões para todos os modelos.

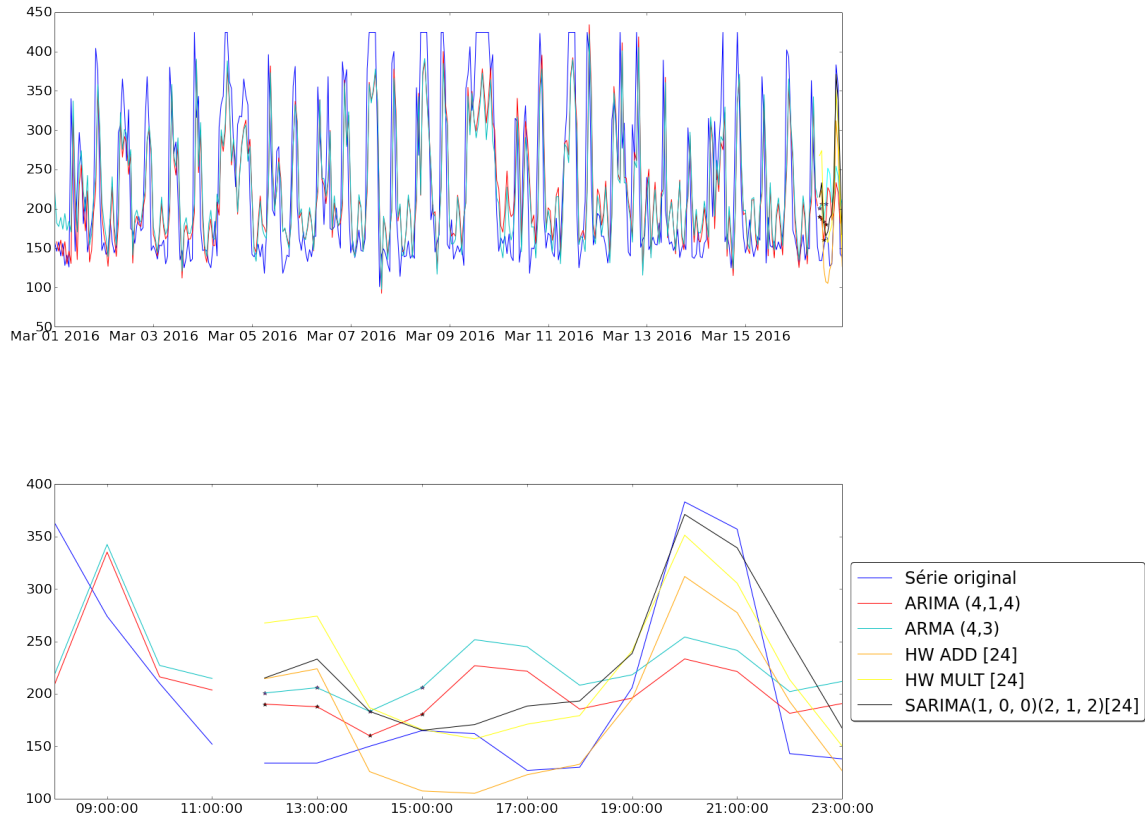


Figura 18: Exemplo da previsão de uma série temporal utilizando os modelos ARMA, ARIMA, SARIMA, Holt-Winters aditivo e Holt-Winters multiplicativo.

Recorde-se que estes modelos estão a ser aplicados sobre a previsão de séries individuais, de consumo doméstico. Trata-se pois de um tipo associado a uma grande variabilidade, sem os efeitos de compensação inerentes à agregação de um grande número de consumidores, como é, normalmente, efetuado por produtoras e distribuidoras de energia. Apesar disso, os erros percentuais ou relativos com os diversos modelos são relativamente baixos, apontando para frações tipicamente abaixo de 60%. Com a inclusão do tratamento da sazonalidade e/ou tendência, os erros decrescem para níveis de 40% ou até menor (Tabela 2). Estes valores demonstram claramente a viabilidade de utilização

dos modelos de previsão considerados e sugerem a utilização específica do SARIMA como modelo mais adequado. Mais à frente serão detalhados os aspetos conducentes à escolha do modelo mais adequado.

Modelos / Métricas	RMSE	MAPE	DTW
ARMA	224.0	53.6	1776.5
ARIMA	243.6	58.8	2026.8
HW ADD	169.8	42.5	1116.5
HW MULT	174.9	40.6	1128.4
SARIMA	143.6	35.8	992.4

Tabela 2: Valores médios das métricas dos vários modelos na previsão de 100 séries temporais de consumo doméstico.

A Tabela 3 ilustra o número de vezes que cada modelo apresentou uma capacidade de previsão significativamente superior à dos outros modelos. A Tabela 4 apresenta, de forma ordenada e entre parênteses, o número de vezes em que cada modelo revelou ter o mínimo de uma determinada métrica em relação aos restantes modelos. Os testes Diebold-Mariano, aplicados sobre as 100 séries de consumo doméstico alvo de análise, apresentaram os resultados constantes da Tabela 3. Salienta-se a partir dessa tabela o facto de que em apenas 11 das séries foi possível identificar um método com superior capacidade de previsão. Nestas situações o método SARIMA revelou-se como aquele que mais frequentemente apresenta uma melhor capacidade. Optando diretamente pelo estabelecimento da frequência em que as métricas de erro apresentam valores inferiores de acordo com uma maior capacidade de previsão, obtêm-se os resultados constantes da Tabela 4. Naturalmente que desta maneira não surgem resultados ao qual possa ser atribuído significado estatístico. Em contrapartida, todas as experiências apresentam um método "ganhador". Verifica-se que a utilização direta das métricas confirma os resultados obtidos com base no teste de Diebold-Mariano. Note-se que este último é adaptado à comparação de capacidades de previsão de modelos e não, propriamente, uma comparação genérica do respetivo desempenho.

Modelo	Resultado Diebold-Mariano
SARIMA	7
ARIMA	2
ARMA	1
HW MULT	1
HW ADD	0

Tabela 3: Ordenação dos modelos pelo número de vezes cujo resultado do Diebold-Mariano indica que a previsão é superior à dos outros modelos e valor desta frequência.

Métrica	Modelos
RMSE	SARIMA (59), HW ADD (19), ARMA (9), HW MULT (3), ARIMA (2)
MAPE	SARIMA (56), ARMA (19), HW ADD (11), HW MULT (4), ARIMA (2)
DTW	SARIMA (38), HW ADD (33), ARMA (11), HW MULT (8), ARIMA (2)

Tabela 4: Ordenação dos modelos pela frequência em que o valor da respetiva métrica é inferior ao dos outros modelos.

5.3.3 Escolha das janelas de histórico e previsão

O modelo SARIMA revelou a melhor capacidade de previsão, pelo que será este o selecionado para os testes subsequentes.

5.3.3.1 Consumo horário

Para o consumo horário, são apresentadas na Tabela 5 os resultados para vários pares de janelas e para a sazonalidade 24 ($Saz = 24$). Pela observação da Tabela 5 conclui-se que uma janela de histórico $5 \times Saz + 15$ permite resultados interessantes com uma janela de previsão $Saz/3$. Esta apresenta erros não muito superiores à janela de previsão $Saz/6$, mas duplica o horizonte de previsão. Aumentando mais este último, os erros parecem tornar-se excessivos e o mesmo sucede caso se proceda à redução da janela de histórico.

Histórico	Previsão			
$3 \times Saz + 15$	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 87.8	RMSE: 103.4	RMSE: 139.5	RMSE: 160.6
	MAPE: 39.2	MAPE: 34.2	MAPE: 36.3	MAPE: 36.3
	DTW: 363.3	DTW: 823.1	DTW: 1330.5	DTW: 2783.7
$4 \times Saz + 5$	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 129.1	RMSE: 143.4	RMSE: 201.4	RMSE: 248.6
	MAPE: 50.8	MAPE: 54.5	MAPE: 56.3	MAPE: 53.9
	DTW: 476.7	DTW: 981.8	DTW: 1896.8	DTW: 4101.1
$5 \times Saz + 5$	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 107.2	RMSE: 114.4	RMSE: 161.9	RMSE: 190.2
	MAPE: 33.3	MAPE: 33.8	MAPE: 36.0	MAPE: 35.6
	DTW: 401.9	DTW: 804.2	DTW: 1509.8	DTW: 3437.3

Tabela 5: Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade horária.

5.3.3.2 Consumo octa-horário

Para o consumo octa-horário, são apresentadas na Tabela 6 as janelas com o menor valor para cada uma das métricas e para a sazonalidade 21 ($Saz = 21$). Refira-se que nem sempre foi possível proceder ao cálculo do MAPE, facto já previamente referido por outros autores[58]. Verifica-se neste caso, também, que o período de histórico deve ser maximizado, sendo os indicadores relativamente consistentes em que se pode estender a previsão até pelo menos $Saz/2$ (Tabela 6).

Histórico	Previsão			
3*Saz+15	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 1754.8	RMSE: 1768.4	RMSE: 1650.1	RMSE: 2232.3
	MAPE: inf	MAPE: inf	MAPE: inf	MAPE: inf
	DTW: 5302.9	DTW: 13788.0	DTW: 18654.7	DTW: 40594.9
4*Saz+5	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 255.6	RMSE: 278.1	RMSE: 279.1	RMSE: 315.5
	MAPE: 51.1	MAPE: 48.3	MAPE: 46.2	MAPE: 48.2
	DTW: 761.4	DTW: 1906.7	DTW: 2577.5	DTW: 5455.8
5*Saz+5	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 199.4	RMSE: 209.3	RMSE: 204.9	RMSE: 227.6
	MAPE: 39.8	MAPE: 35.0	MAPE: 34.1	MAPE: 34.6
	DTW: 596.3	DTW: 1452.6	DTW: 1957.7	DTW: 4235.0

Tabela 6: Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade octa-horária.

5.3.3.3 Consumo diário

Para o consumo diário, são apresentadas na Tabela 7 as janelas com o menor valor e para cada uma das métricas para a sazonalidade 7 ($Saz = 7$). Na generalidade, pode afirmar-se que a maior capacidade de previsão parece surgir para uma janela de histórico máxima e uma janela de previsão mínima.

Histórico	Previsão			
6*Saz+5	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 84.9	RMSE: 118.4	RMSE: 137.3	RMSE: 164.7
	MAPE: 22.0	MAPE: 23.6	MAPE: 24.4	MAPE: 28.6
	DTW: 84.9	DTW: 272.5	DTW: 422.5	DTW: 1045.4
7*Saz+5	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 85.5	RMSE: 122.5	RMSE: 137.6	RMSE: 178.2
	MAPE: 21.9	MAPE: 24.4	MAPE: 24.2	MAPE: 29.3
	DTW: 85.5	DTW: 282.9	DTW: 415.0	DTW: 1057.1
8*Saz+5	Saz/6	Saz/3	Saz/2	Saz
	RMSE: 82.8	RMSE: 113.4	RMSE: 124.3	RMSE: 165.3
	MAPE: 20.7	MAPE: 22.6	MAPE: 23.1	MAPE: 27.8
	DTW: 82.8	DTW: 262.1	DTW: 392.5	DTW: 1003.2

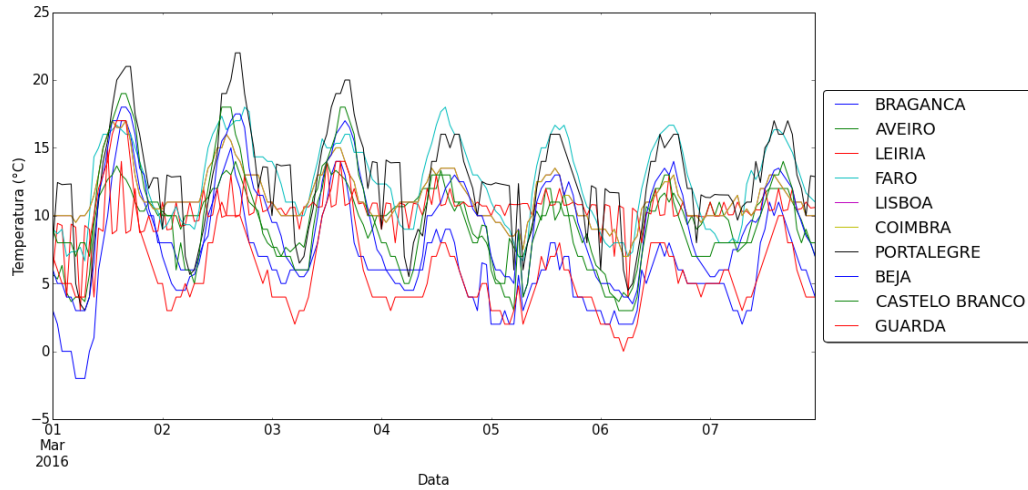
Tabela 7: Valores das métricas de erro de previsão para cada um dos pares janelas de histórico e previsão para a granularidade diária.

5.3.4 Influência de vários fatores na previsão

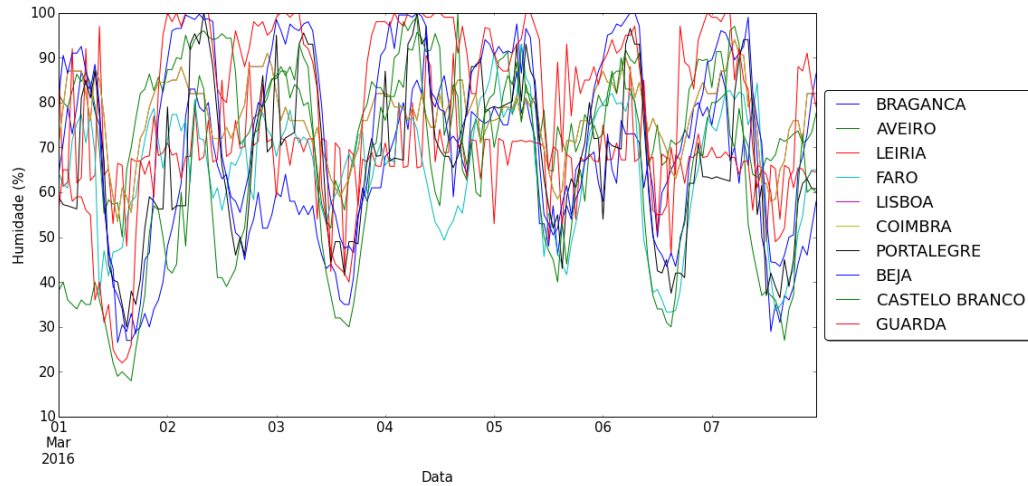
Nesta secção serão apresentados os resultados relativos à influência da incorporação de variáveis exógenas, da utilização de um tipo de histórico diferente e da aplicação de filtros e transformações.

5.3.4.1 Temperatura e humidade como variáveis exógenas

Na Figura 19 são apresentadas as séries temporais de temperatura (Figura 19a) e humidade (Figura 19b) para um conjunto de distritos portugueses.



(a) Valores de temperatura.



(b) Valores de humidade.

Figura 19: Sobreposição dos valores de temperatura e humidade para vários distritos na primeira semana do mês de Março de 2016.

De forma a interpretar melhor a correlação entre as variáveis exógenas e o consumo, para o conjunto de séries temporais de consumo doméstico, foram elaborados histogramas dos valores da correlação, separando-os em valores positivos e valores negativos. Como é possível observar nas Figuras 20 e 21, verifica-se uma menor prevalência dos valores negativos de correlação. Verifica-se também uma clara predominância de valores que não excedem 0.3 tanto para a temperatura como para a humidade.

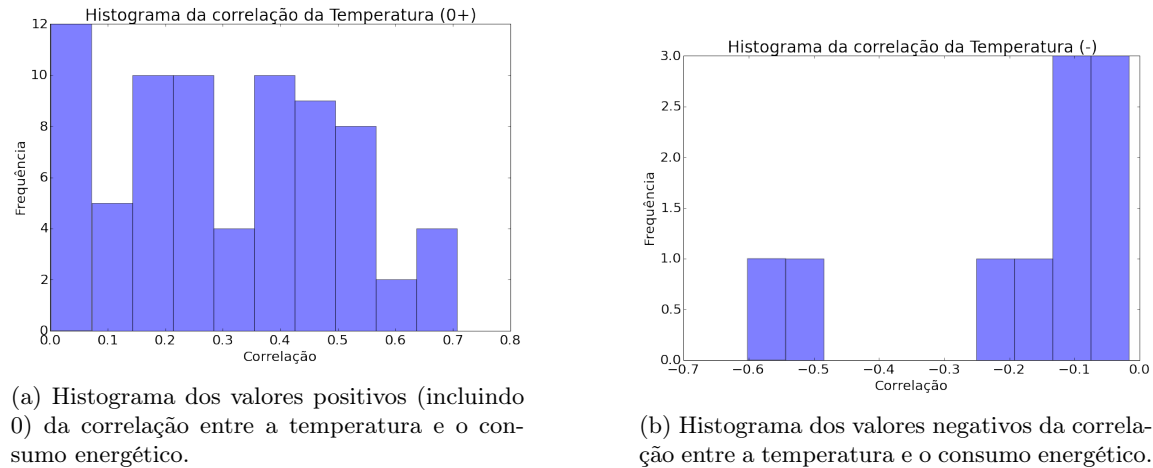


Figura 20: Histogramas da correlação entre a temperatura e o consumo energético de 100 séries temporais de consumo doméstico.

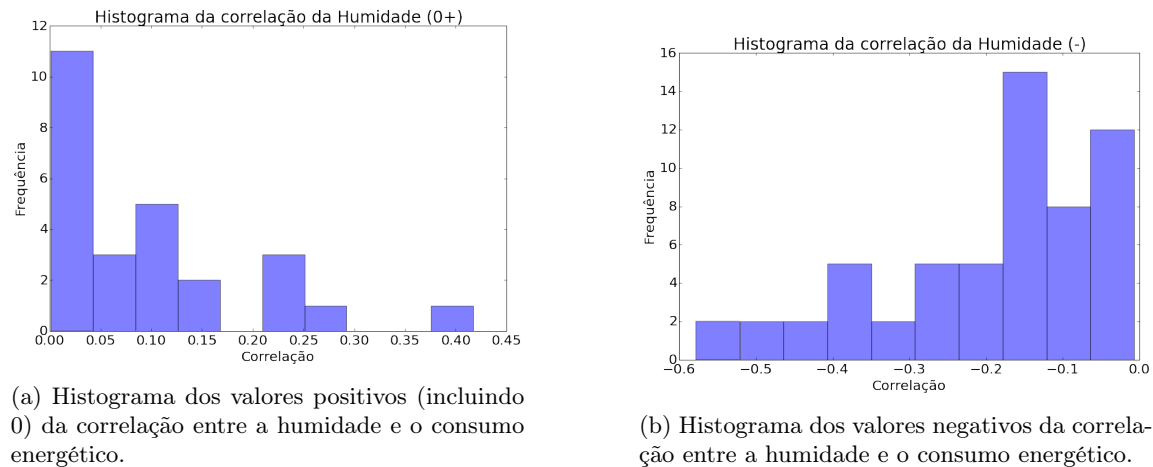


Figura 21: Histogramas da correlação entre a humidade e o consumo energético de 100 séries temporais de consumo doméstico.

Na Tabela 8 está apresentada a correlação entre as variáveis exógenas para cada distrito representativo. Verifica-se uma correlação negativa genericamente elevada entre a temperatura e a humidade em cada um dos distritos.

Distrito	Correlação
Faro	-0.46837
Bragança	-0.70401
Coimbra	-0.60196
Guarda	-0.59264
Portalegre	-0.67279
Leiria	-0.32955
Castelo Branco	-0.68720
Lisboa	-0.60196
Aveiro	-0.38844
Beja	-0.75725

Tabela 8: Correlação entre temperatura e humidade para cada distrito.

Os resultados da comparação entre a previsão sem variáveis exógenas e com as variáveis exógenas de temperatura e humidade estão apresentados na Tabela 9. No que diz respeito à capacidade de previsão verifica-se que, recorrendo a métricas, em cerca de 50% dos casos a maior capacidade de previsão é apresentada pelo método SARIMA, em oposição ao SARIMAX que apresenta cerca 25% de melhores previsões para cada uma das variáveis. Recorreu-se, ainda, ao teste de Diebold-Mariano para identificar possíveis melhorias e, se existirem, determinar se são significativas. Os resultados não são tão drásticos com base no teste Diebold-Mariano (Tabela 10) mas continuam a apontar para um pior desempenho pela introdução da variável exógena.

Métrica	Variáveis
RMSE	Sem (45), Hum (31), Temp (24)
MAPE	Sem (48), Temp (26), Hum (26)
DTW	Sem (49), Temp (27), Hum (24)

Tabela 9: Ordenação das variáveis exógenas ou ausência destas, pela frequência em que o valor da respetiva métrica é inferior ao das outras previsões.

Variável exógena	Resultado Diebold-Mariano
Sem	14
Hum	12
Temp	11

Tabela 10: Ordenação das variáveis exógenas pelo número de vezes cujo resultado do Diebold-Mariano indica que a previsão é superior à da utilização de outras variáveis exógenas ou da sua não utilização para efeitos de previsão.

5.3.4.2 Tipos de histórico

Os dois históricos a comparar na presente secção serão o histórico real, cujos instantes de amostragem correspondem à realidade, e o histórico constituído apenas pelo mesmo dia da semana que aquele a prever, cujos instantes de amostragem estão modificados para aparentarem um histórico sequencial no mesmo período que o outro tipo de histórico. Para o segundo tipo de histórico foram recolhidos dias ao longo das semanas prévias à da previsão. Na Figura 22 encontram-se sobrepostos esses dois tipos de históricos. De notar que as séries apresentadas incluem o dia a prever, pelo que é possível observar que o consumo para o último dia apresenta um padrão igual para ambos os históricos.

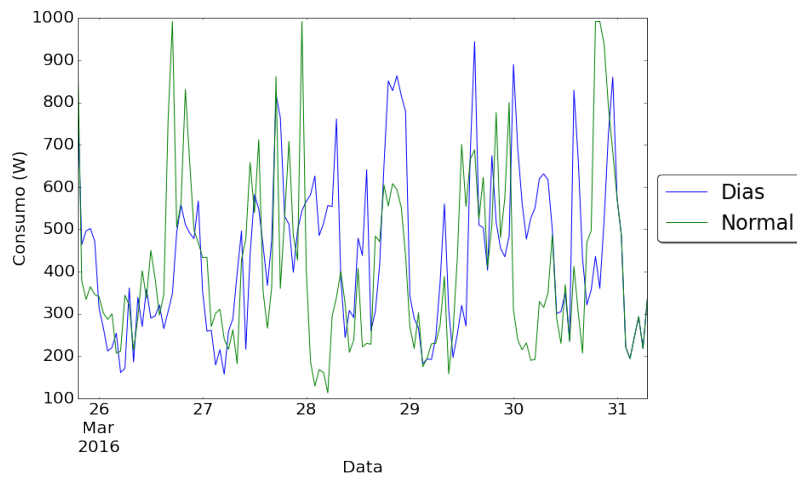


Figura 22: Sobreposição dos dois tipos de histórico.

Os resultados do estudo da influência do tipo de histórico na previsão encontram-se apresentados na Tabela 11. Da observação desta tabela verifica-se que, ao contrário do que se poderia esperar, a previsão recorrendo a toda a sequência semanal apresenta melhor grau de previsão do que aquela em que se recorre a sequências sobre o mesmo dia da semana. Estas observações são confirmadas pelos resultados do teste de Diebold-Mariana, Tabela 12. Este resultado poderá ter a ver com o facto de que uma evolução mais gradual permite uma adaptação do modelo, que não se consegue com hiatos de uma semana.

Métrica	Tipo de histórico
RMSE	Normal (78), Dias (22)
MAPE	Normal (73), Dias (27)
DTW	Normal (74), Dias (27)

Tabela 11: Ordenação do tipo de histórico pela frequência em que o valor da respetiva métrica é inferior ao do outro tipo de histórico.

Tipo de histórico	Resultado Diebold-Mariano
Normal	18
Mesmo tipo de dias	4

Tabela 12: Ordenação dos tipos de histórico por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à do outro tipo de histórico e indicação da frequência.

5.3.4.3 Filtros

Um exemplo de uma série temporal original sobreposta com as séries temporais a que foram aplicados os filtros Baxter-King bandpass, Hodrick-Prescott e Holt-Winters apresenta-se na Figura 23. Os resultados da Tabela 13 indicam que o filtro normal apresenta a maior taxa de sucesso. Tal observação é confirmada pelos resultados dos testes Diebold-Mariano, embora a ordenação dos filtros não seja exatamente coincidente. Os resultados deste estudo apresentam-se na Tabela 14.

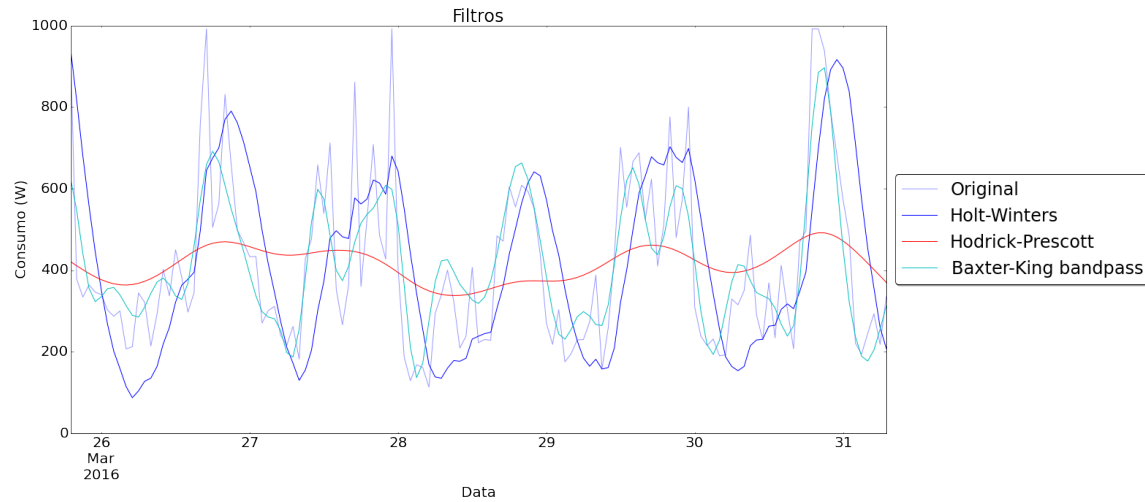


Figura 23: Sobreposição de uma série temporal de consumo horário à qual foram aplicados filtros com a série original.

Métrica	Filtro
RMSE	Baxter-King bandpass (38), Normal (37), Hodrick-Prescott(25), Holt-Winters (0)
MAPE	Normal (63), Baxter-King bandpass (19), Hodrick-Prescott (11), Holt-Winters (7)
DTW	Baxter-King bandpass (58), Hodrick-Prescott (29), Normal (13), Holt-Winters (0)

Tabela 13: Ordenação dos filtros pela frequência em que o valor da respetiva métrica é inferior ao dos outros filtros.

Filtro	Resultado Diebold-Mariano
Normal	9
Hodrick-Prescott	8
Baxter-King bandpass	6
Holt-Winters	0

Tabela 14: Ordenação dos filtros por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à dos outros filtros e indicação da frequência.

5.3.4.4 Transformações

As transformações estudadas nesta secção foram a logarítmica (Log), raiz quadrática (Sqrt) e Box-Cox e encontram-se representadas na Figura 24.

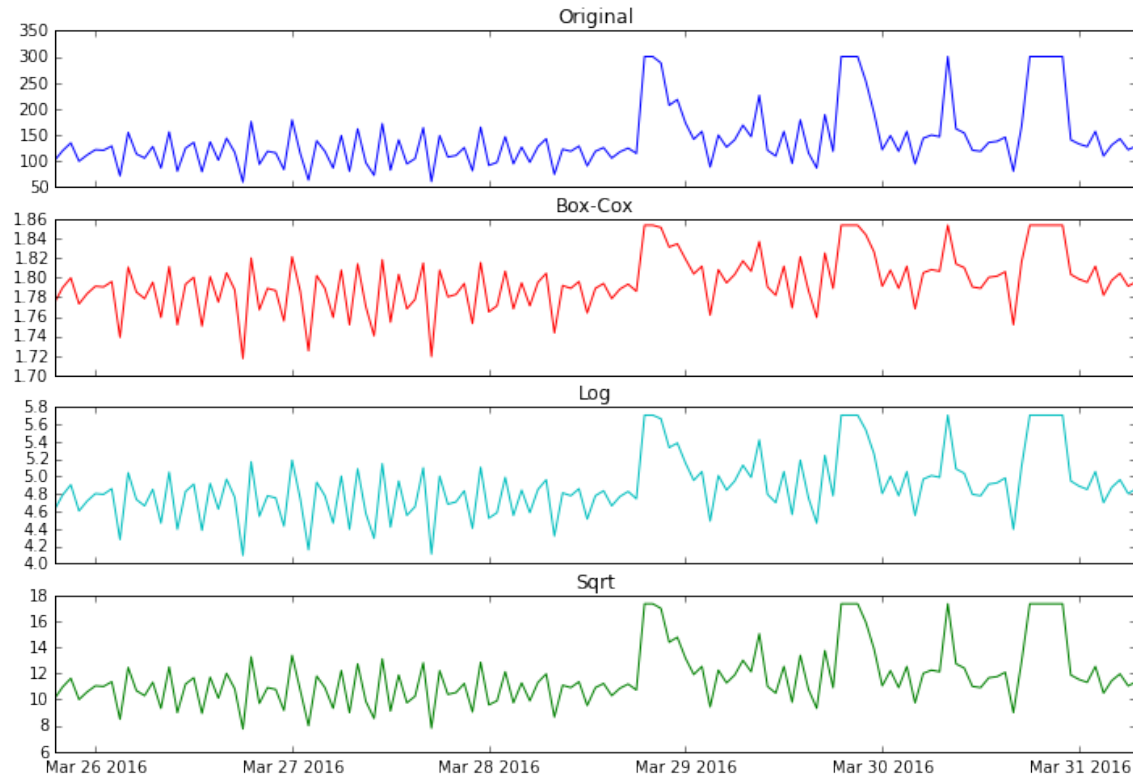


Figura 24: Apresentação de uma série de consumo energético horário e resultado da aplicação de transformações na série.

Os resultados das previsões das séries temporais com e sem transformações apresentam-se na Tabela 15 e mostram alguma prevalência da ausência de transformação (designada na tabela por Original). Nalguns casos a logaritmização melhora a capacidade de previsão. Mais uma vez, as observações baseadas em métricas são confirmadas pelos testes de Diebold-Mariano em que foi observada uma diferença significativa (Tabela 16).

Métrica	Transformação
RMSE	Log (47), Sqrt (27), Original (19), Box-Cox (2)
MAPE	Original (40), Box-Cox (36), Log (13), Sqrt (6)
DTW	Log (50), Sqrt (29), Original (16), Box-Cox (0)

Tabela 15: Ordenação das transformações pela frequência em que o valor da respetiva métrica é inferior ao dos outros filtros.

Transformação	Resultado Diebold-Mariano
Original	21
Log	8
Box-Cox	5
Sqrt	1

Tabela 16: Ordenação dos filtros por número de vezes cujo valor resultado do Diebold-Mariano indica que a previsão é superior à dos outros filtros e indicação da frequência.

5.4 Clustering

Esta experiência visa definir as condições e determinar a exequibilidade da previsão em bloco, com o intuito de reduzir o custo computacional da previsão individual (*clustering* de séries não-normalizadas) e da relevância do *clustering* de séries de consumo normalizado para a obtenção de uma tipificação dos comportamentos de consumo.

5.4.1 Clustering de séries temporais não-normalizadas

Verificou-se que o valor máximo anteriormente fixado para k apresenta o valor ótimo do parâmetro \mathcal{E} para a granularidade horária (Figura 25). Tal não é de estranhar uma vez que existe uma grande variação do valor médio de consumo, pelo que é natural que surjam um número elevado de grupos. Este aspeto está bem representado na Figura 26 onde se verifica claramente variações quer no valor médio quer na amplitude correspondente aos centróides dos grupos identificados. Será que prever a partir do centróide é semelhante a fazer a previsão individual de cada serie do grupo?



Figura 25: Valores de \mathcal{E} para o consumo horário não-normalizado para cada valor de k .

As séries temporais correspondentes a cada centróide resultado do *cluster* de séries temporais de consumo horário apresentam-se na Figura 26.

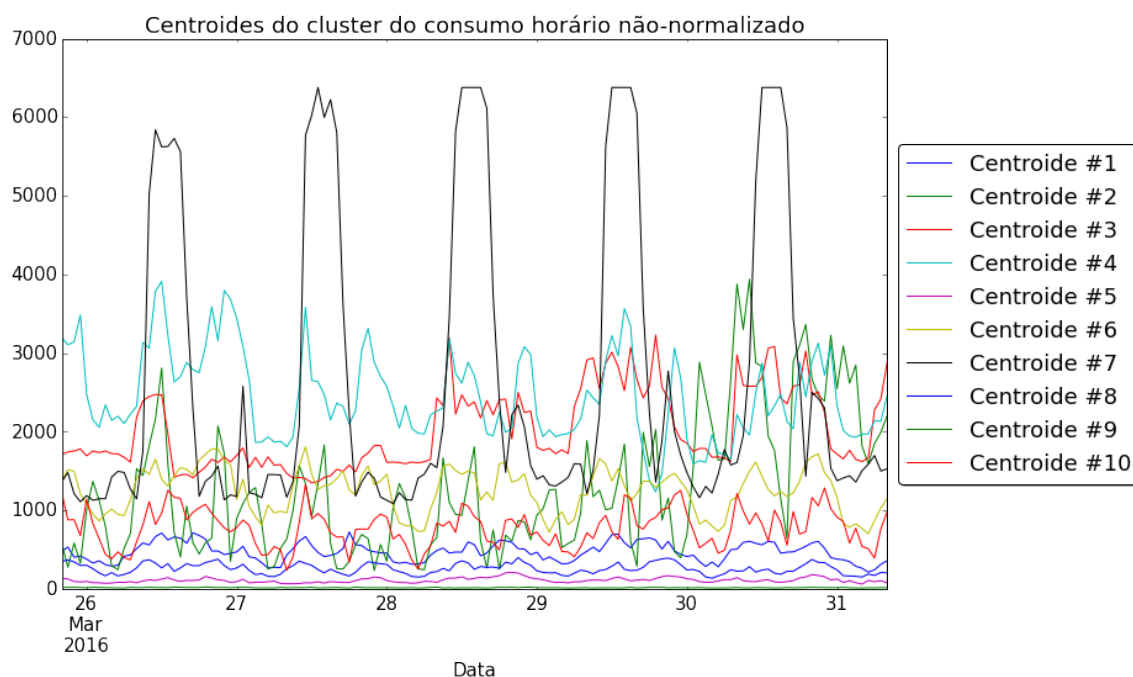


Figura 26: Centróides de *clusters* para um conjunto de 100 séries temporais de consumo horário doméstico.

Os resultados patentes na Tabela 17 revelam que, como era esperado, a previsão em bloco implica erros superiores, no entanto não serão significativamente mais elevados. É possível observar a instabilidade da métrica MAPE possivelmente devido a divisões por valores próximo de 0.

Métrica	Previsão individual	Previsão por cluster
RMSE	149.8	212.8
MAPE	7.1e+13	1.2e+14
DTW	1097.5	1722.1

Tabela 17: Valor das métricas de erro para a previsão individual horária *vs* previsão baseada em centróides.

Passando agora para a granularidade octa-horária e procedendo da mesma forma, identifica-se que o valor ótimo de agrupamentos é 7 (Figura 27). Note-se que, de facto, 2 surgiu como o valor máximo mas o nível de agrupamento parece excessivo com um número tão reduzido de agregados. Nesse sentido este valor foi descartado.



Figura 27: Valores de \mathcal{E} para o consumo octa-horário não-normalizado para cada valor de k .

As séries temporais correspondentes a cada centróide resultado do *cluster* de séries temporais de consumo octa-horário apresentam-se na Figura 28. A distinção é essencialmente feita por valor médio, revelando-se uma sobreposição limitada das séries.

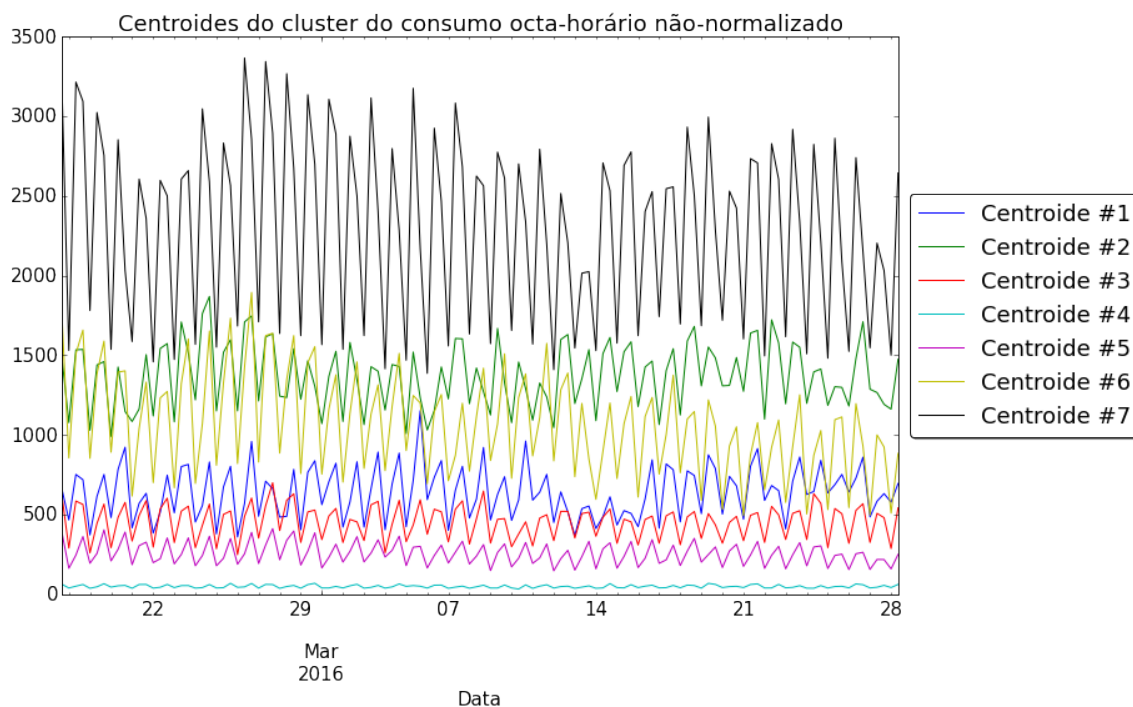


Figura 28: Centr ides de *clusters* para um conjunto de 100 s ries temporais de consumo octa-hor rio dom stico.

Novamente, e de um modo ainda mais marcado, a previs o de centr ide n o se distancia da previs o individual das s ries. Tal   patente nos erros indicados na Tabela 18

M�trica	Previs�o individual	Previs�o por cluster
RMSE	235.7	262.2
MAPE	5.3	7.6
DTW	2164.0	2561.1

Tabela 18: Valor das m tricas de erro para a previs o individual octa-hor ria *vs* previs o baseada em centr ides.

Para o consumo di rio, o n mero de grupos  timo   8 (Figura 29). Veja-se como este n mero n o parece ter uma grande depend ncia na granularidade. Mais uma vez o valor m dio parece desempenhar um papel importante na discrimina o do grupo (veja-se a Figura 30). Tamb m   not rio que se regista algum aumento do erro da previs o em bloco relativo   previs o individual (Tabela 19), em compara o com o que se observou para as outras granularidades. Apesar disso, n o s o introduzidos erros importantes com a op o pela previs o de bloco.

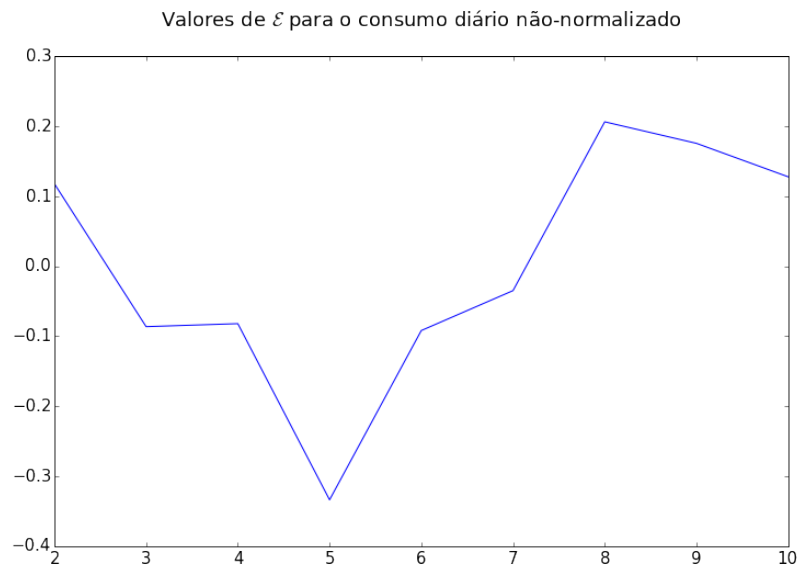


Figura 29: Valores de \mathcal{E} para o consumo diário não-normalizado para cada valor de k .

As séries temporais correspondentes a cada centróide resultado do *cluster* de séries temporais de consumo diário apresentam-se na Figura 30.

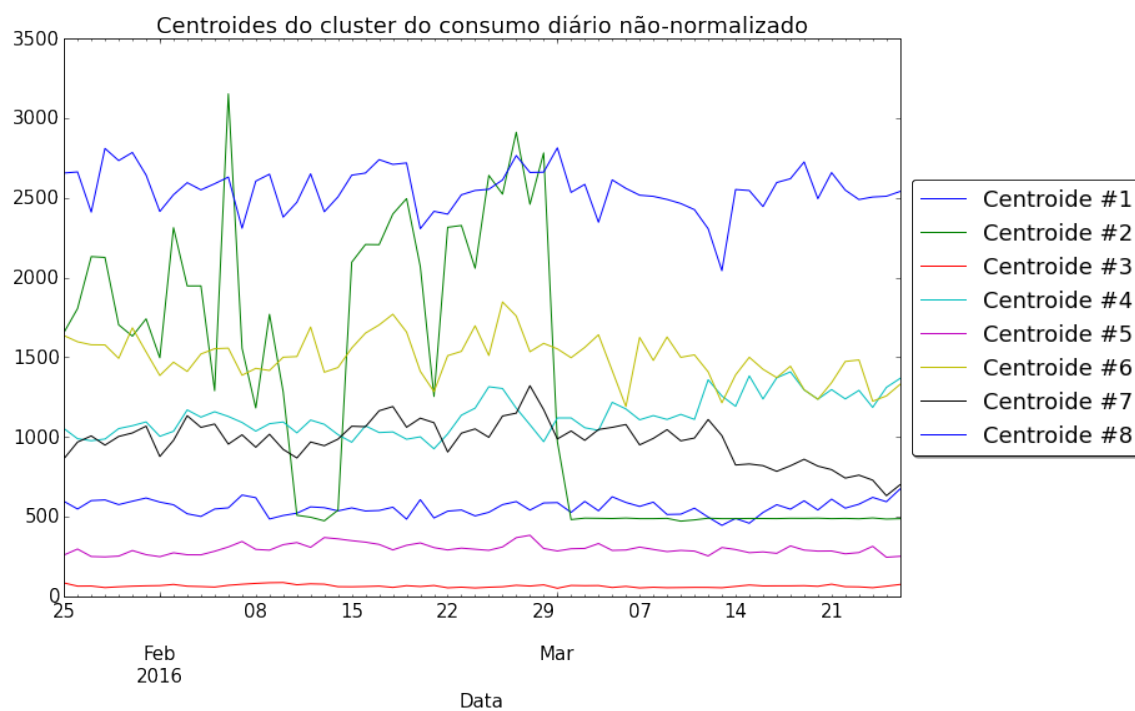


Figura 30: Centr ides de *clusters* para um conjunto de 100 s ries temporais de consumo di rio dom stico.

M�trica	Previs�o individual	Previs�o por cluster
RMSE	94.7	142.4
MAPE	18.7	42.9
DTW	94.7	142.4

Tabela 19: Valor das m tricas de erro para a previs o individual di ria *vs* previs o baseada em centr ides.

5.4.2 Clustering de séries temporais normalizadas

O *clustering* de séries temporais normalizadas que visa agrupar padrões semelhantes descartando o valor absoluto de consumo identificou um valor elevado de grupos como ótimo, quer para a granularidade horária com 10 grupos (Figura 31), octa-horária com 10 grupos, como para a diária com 9 grupos. O tipo de resultados obtido é semelhante em todas as granularidades e recorreremos à primeira para uma descrição dos resultados, encontrando-se os resultados das restantes granularidades representados nos Anexos A.1 e A.2. Juntou-se na Figura 33 a representação das séries de cada grupo, destacando-se a série correspondente ao centróide (centróides estes representados na Figura 32). Verifica-se que vários aspetos parecem ter influído na discriminação entre grupos. Entre estes, destaque-se a existência de grupos com uma periodicidade bem marcada, distinguindo-se entre si por uma maior variação percentual diferente e pelo formato da zona de maior consumo. Outros grupos caracterizam-se por serem mais irregulares, apresentado um consumo centrado em determinadas zonas do espaço temporal abrangido pela série. Este resultado é muito interessante por revelar a existência de consumos tipificados. O estudo deste aspeto implicaria uma exploração que não foi possível concretizar no espetro temporal previsto para este trabalho, podendo ser alvo de investigação futura.



Figura 31: Valores de ε para o consumo horário não-normalizado para cada valor de k .

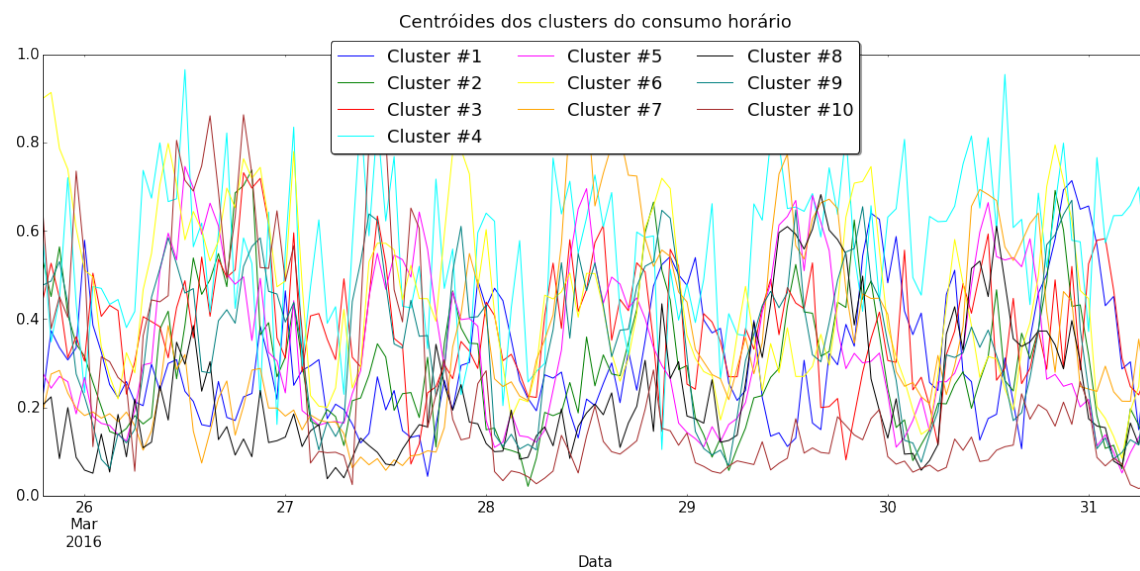
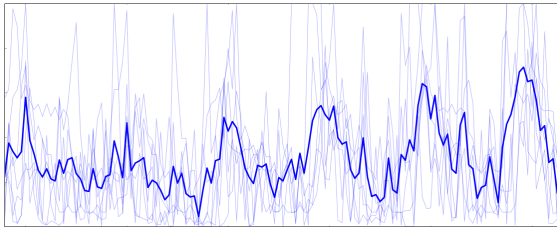
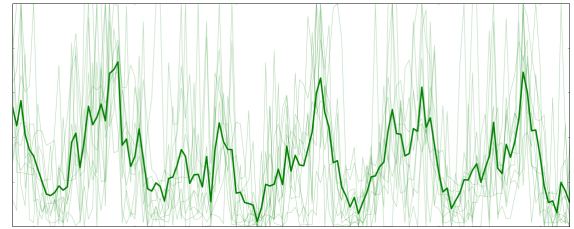


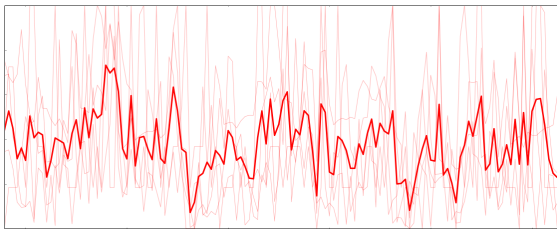
Figura 32: Centróides dos *clusters* para um conjunto de 100 séries normalizadas de consumo horário doméstico.



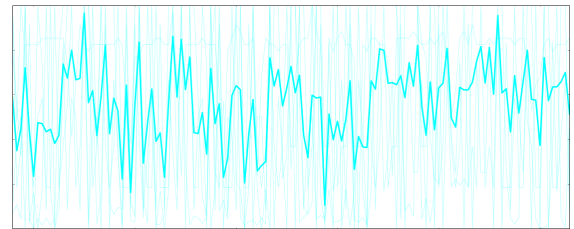
(a) Séries temporais do cluster #1



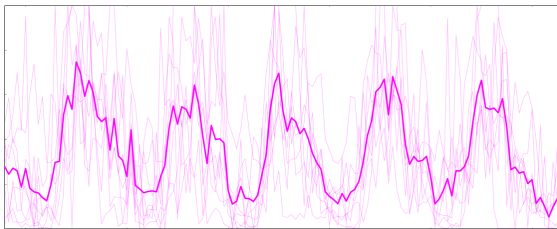
(b) Séries temporais do cluster #2



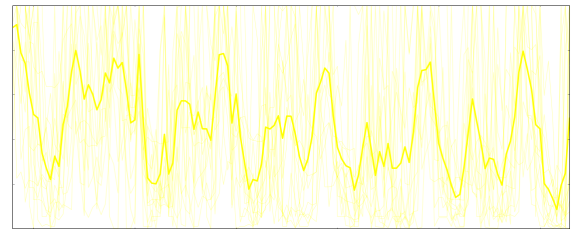
(c) Séries temporais do cluster #3



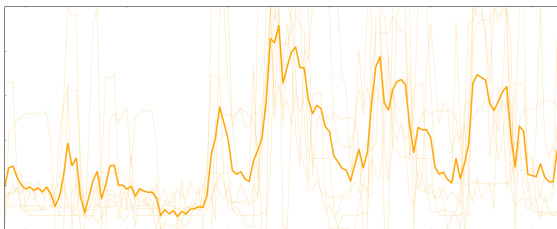
(d) Séries temporais do cluster #4



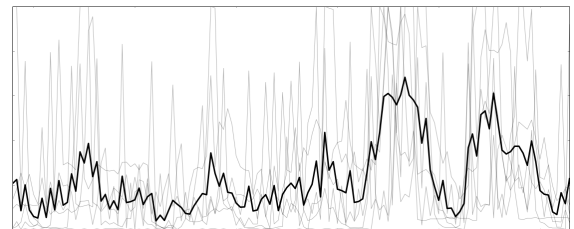
(e) Séries temporais do cluster #5



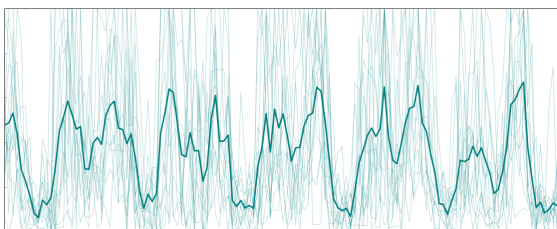
(f) Séries temporais do cluster #6



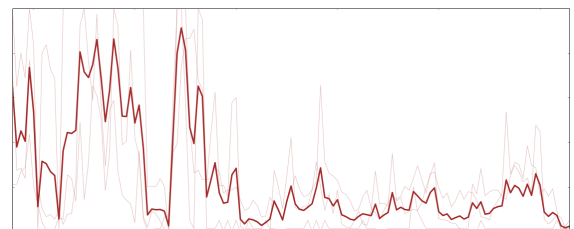
(g) Séries temporais do cluster #7



(h) Séries temporais do cluster #8



(i) Séries temporais do cluster #9



(j) Séries temporais do cluster #10

Figura 33: *Clusters* do consumo horário normalizado com destaque do centróide.

5.5 Pontos de mudança

Conclui-se que existe uma concentração elevada de *change-points* claramente relacionados com variações atribuíveis à sazonalidade. Obviamente que não se tratam de situações de relevância para os modelos, uma vez que não passam por uma evolução significativa que exija uma modificação drástica dos respetivos parâmetros. Para além disso, a sazonalidade já está introduzida nos próprios modelos e, neste trabalho, deu-se ênfase a essa introdução. No entanto surgem também, frequentemente, *change-points* que delimitam o fim-de-semana, assinalado a azul na Figura 34 e seguintes. Esta observação confirma algo que seria espetável: uma alteração de comportamento no consumo em dias não laborais relativamente àquele que se verifica durante a semana de trabalho. Estas variações parecem ser mais facilmente incorporadas no modelo com base nas maiores granularidades, recorde-se que a sazonalidade para o consumo octa-horário e diário é semanal.

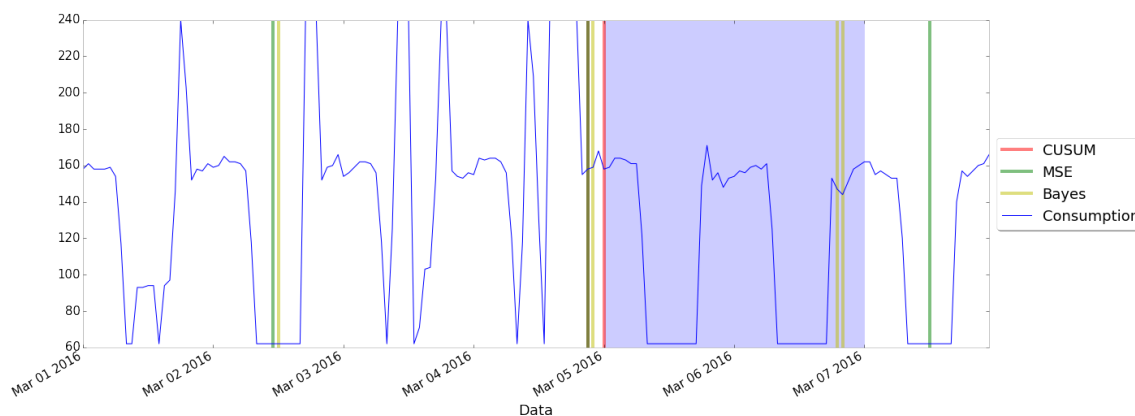


Figura 34: *Change-points* detetadas na 1ª semana.

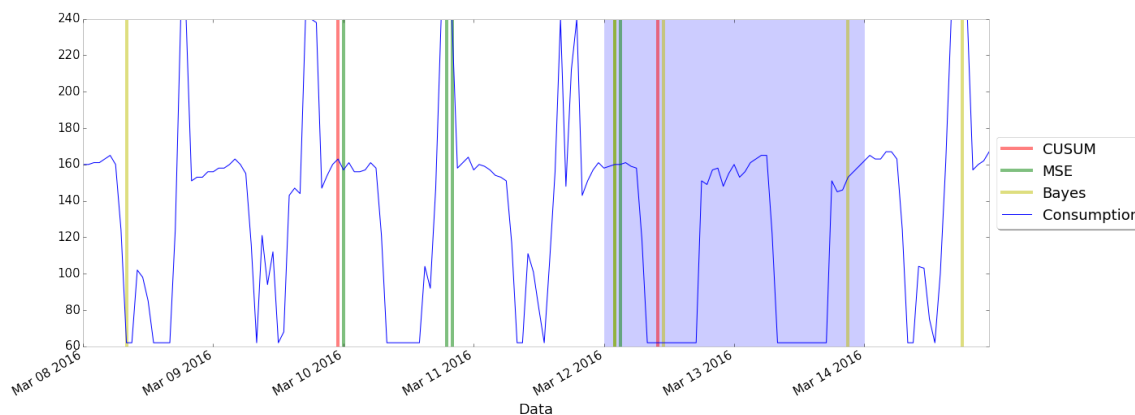


Figura 35: *Change-points* detetados na 2ª semana.

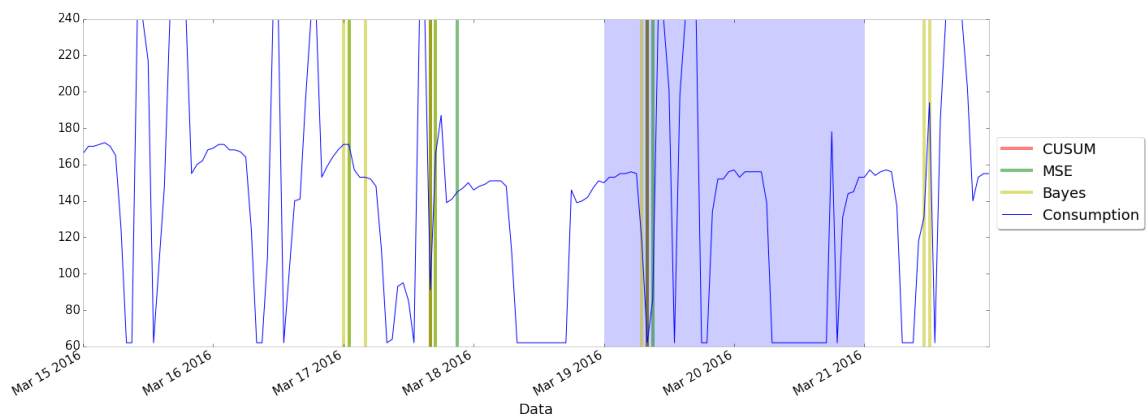


Figura 36: *Change-points* detetados na 3^a semana.

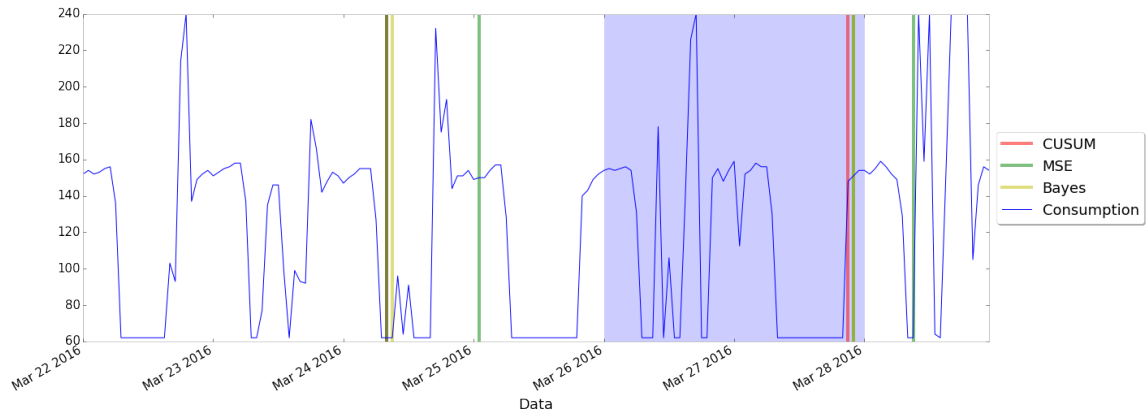


Figura 37: *Change-points* detetados na 4^a semana.

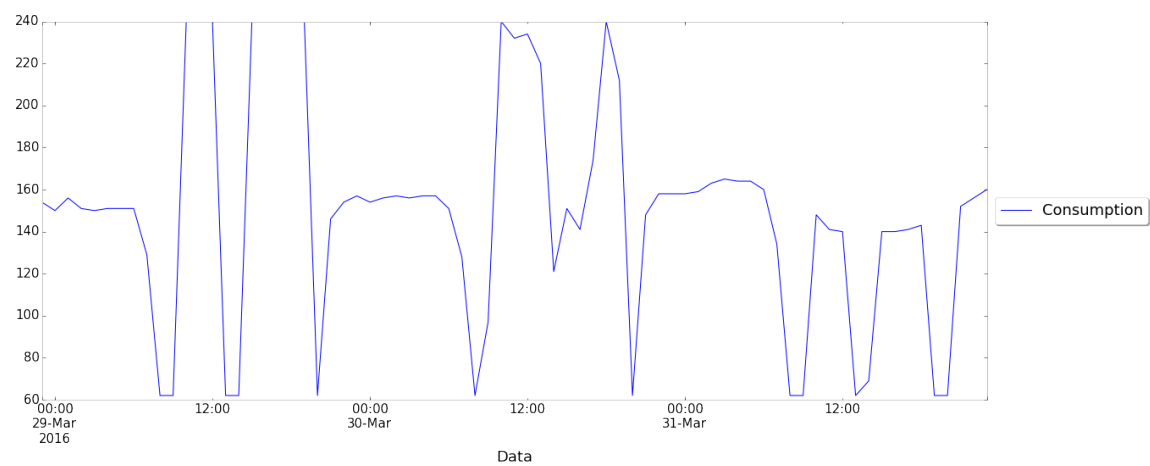


Figura 38: *Change-points* detetados na 5^a semana.

No que respeita à previsão horária esta é fortemente afetada pela existência de *change-points*, como ilustrado na Figura 39. De facto uma sazonalidade de 24 horas promove uma certa regularidade na previsão diária, pelo que o método não reage rapidamente à alteração correspondente ao fim-de-semana. Na Figura 39 é possível observar *change-points* que ocorrem imediatamente antes e imediatamente após um fim-de-semana. Variando o histórico pode promover-se alterações na capacidade de previsão. Com efeito, a previsão #1 com base num histórico de dias-da-semana (até ao primeiro *change-point*) falha claramente na previsão de um sábado. A previsão #2, com o histórico de sábado (a partir do primeiro *change-point*) para prever o domingo, minimiza já esse erro. Finalmente a previsão #3, respetiva à previsão de uma segunda-feira a partir do histórico de um fim-de-semana (entre os dois *change-points*), promove, novamente, um aumento substancial do erro.

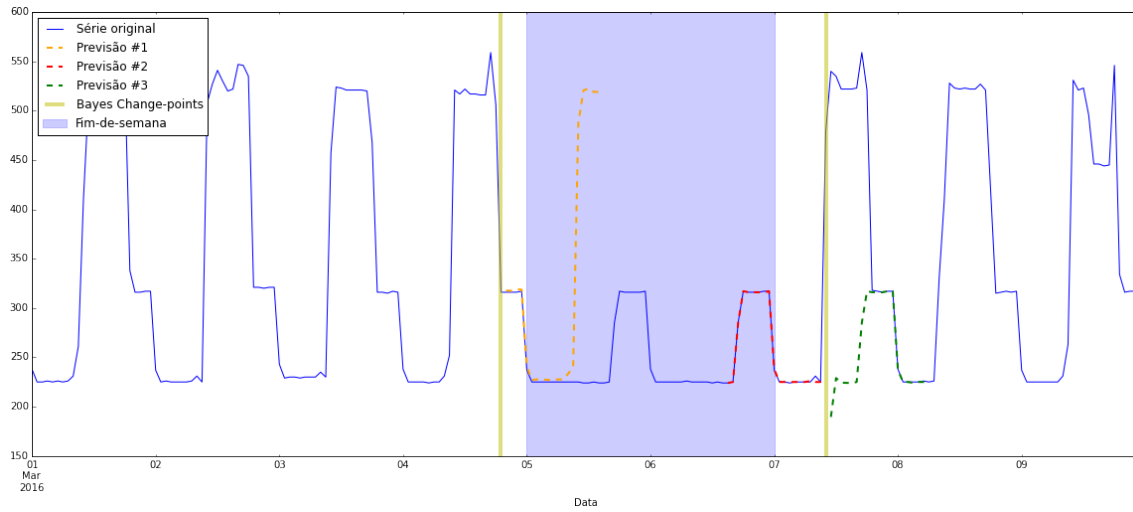


Figura 39: Previsões com histórico prévio ao primeiro *change-point*, incompleto entre *change-points* e histórico total entre *change-points*.

6 Conclusões

A análise de séries temporais permitiu confirmar ideias presentes no estado da arte bem como estabelecer características inerentes aos dados em questão, das quais se destacam:

- A sazonalidade determinada para cada uma das granularidades segue a descrita na literatura (diária e semanal), embora a granularidade octa-horária diferisse da sazonalidade esperada, a diária, tendo-se constatado que a mais apropriada seria a semanal.
- A comparação de capacidade de previsão pode ser levada a cabo quer recorrendo a métricas apropriadas, quer a testes estatísticos, normalmente com resultados concordantes entre estas duas estratégias. A utilização de conjuntos de séries temporais de volume elevado contribui para a natureza real deste tipo de estudos.
- A modelação de séries temporais de consumo de eletricidade passa claramente pela utilização de modelos de previsão que incorporem sazonalidade, destacando-se o modelo SARIMA como o modelo com maior capacidade de previsão e que foi selecionado para realizar as experiências com resultados satisfatórios.
- Neste trabalho não se verificou que a introdução de fatores exógenos aumentasse a capacidade de previsão. O mesmo se concluiu pela utilidade da aplicação de filtros e transformações.
- Ao contrário do esperado o tipo de histórico desenhado para a análise da similaridade entre dias da semana iguais para a melhoria da qualidade de previsão provou não ser claramente eficaz neste objetivo.

Os *change-points* influenciam claramente a capacidade de previsão, especialmente em situações em que não coincidem com a sazonalidade, como seja o caso dos fins-de-semana, sendo necessário efetuar modelos de previsão que tenham em conta este aspeto.

A previsão em bloco com base em *clustering* é uma alternativa interessante à previsão individual, não promovendo na generalidade uma aumento substancial do erro, contribuindo sim para a redução significativa da quantidade de previsões. No caso de estudo deste trabalho resultou numa redução de 90% das séries temporais a prever (apenas 10 para um conjunto de 100 séries temporais de consumo).

Finalmente, a utilização de granularidades diversas revelam diferentes facetas nos padrões de consumo, pelo que se torna útil recolher a informação complementar que proporcionam. Verificou-se, ainda, que a granularidade octa-horária, de utilização muito escassa ou inexistente na literatura, parece ser uma medida adequada para a divisão do consumo intra-diário, especificamente entre noite, período laboral e período pós-laboral.

O resultado obtido deste trabalho foi uma configuração com base no modelo SARIMA cujo valor médio de MAPE é de cerca de 33.8%, 34.1% e 20.7%, para as granularidades horária, octa-horária e diária, respetivamente, que está presentemente a ser posto em produção num sistema que irá analisar 10 000 habitações. O modelo desenvolvido neste trabalho será utilizado como validação para novas iterações com sucessivas melhorias do mesmo.

7 Trabalho futuro

Focando este relatório essencialmente aspetos relacionados com a previsão de consumo elétrico doméstico poderá, ainda, ser significativamente expandida a abordagem. Entre muitos outros aspetos, sugere-se o desenvolvimento de trabalho com base no período octa-horário. Uma vez que os recursos são limitados e a bateria de experiências possíveis de fazer é significativamente volumosa, foi apenas selecionada uma parte, o que não permitirá uma parametrização completa dos modelos de previsão. Um estudo mais extensivo dos fatores influenciadores da capacidade de previsão poderia passar pela utilização de outras variáveis exógenas.

A temática da deteção de *change-points* não foi explorada na sua totalidade. Vários pontos de partida para outros trabalho são:

- a identificação de *change-points* causados por padrões de sazonalidade,
- a identificação de *change-points* causados por fins-de-semana e feriados,
- a deteção de períodos de férias,
- a deteção de mau funcionamento de aparelhos elétricos e
- a deteção de mudanças efetivas na rotina.

Esta filtragem de *change-points* permitirá uma personalização mais apropriada de sugestões elaboradas por sistemas de gestão de energia doméstica.

Naturalmente que o objetivo final para o qual se deu aqui alguma contribuição tem a ver com aspetos relacionados com a otimização do consumo sujeito a restrições de conforto, economia e ambiente. Devem, assim, ser desenvolvidos modelos com base nas previsões que permitam uma automatização da gestão de aparelhos elétricos domésticos e, também, elaborações de sugestões dirigidas ao consumidor doméstico. A previsão do consumo, por exemplo, juntamente com os valores futuros das tarifas variáveis de energia elétrica poderá avisar o consumidor que, como é a rotina habitual, o consumo futuro de volume elevado irá resultar em gastos desnecessários sugerindo mover as cargas que costumam existir nesta altura para outra com tarifas menos elevadas.

Relativamente ao *clustering* de séries temporais, a previsão em bloco é de grande relevância para produtoras e distribuidoras de energia. Embora a típica abordagem utilizada por estas entidades se baseie na previsão do consumo agregado de vários clientes, será interessante explorar uma previsão, possivelmente mais precisa, sem custos computacionais significativamente superiores. Por outro lado, a tipificação dos consumidores pela padronização do seu consumo permitirá a definição de tarifas e contratos mais adequados a cada consumidor e o possível desenvolvimento de outros modelos de caracterização de consumos e consumidores, a serem desenvolvidos futuramente.

Referências

- [1] Correlogramas para AR, MA e ARMA. <http://stats.stackexchange.com/questions/153814/correlogram-and-acf-pacf-applied-to-us-index-of-unemployment-rate>. Data de acesso: 2016-01-19.
- [2] AR Fardana, S Jain, I Jovancevic, Y Suri, C Morand, and NM Robertson. Controlling a mobile robot with natural commands based on voice and gesture. 2013. <http://home.eps.hw.ac.uk/~cgb7/readinggroup/papers/RobotCommandingByVoiceAndGesture.pdf>. Data de acesso: 2016-06-10.
- [3] Claudie Beaulieu, Jie Chen, and Jorge L Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1962):1228–49, 2012.
- [4] Tabela para a seleção de modelos. <http://www.mathworks.com/help/econ/autocorrelation-and-partial-autocorrelation.html?refresh=true>. Data de acesso: 2016-01-19.
- [5] Luis Hernández, Carlos Baladrón, Javier Aguiar, Belén Carro, and Antonio Sánchez-Esguevillas. Classification and clustering of electricity demand patterns in industrial parks. *Energies*, 5(12):5215–5228, Dec 2012.
- [6] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.
- [7] Unplugg. <http://www.unplu.gg/>. Data de acesso: 2016-01-19.
- [8] Watt-is. <http://www.watt-is.com/>. Data de acesso: 2016-01-19.
- [9] PlotWatt. <https://plotwatt.com/>. Data de acesso: 2016-01-19.
- [10] Bidgely. <https://www.bidgely.com/>. Data de acesso: 2016-01-19.
- [11] Opower. <https://opower.com/>. Data de acesso: 2016-01-19.
- [12] Control4 (Eragy). <http://www.eragy.com/>. Data de acesso: 2016-01-19.
- [13] AlertMe. <http://alertme.com/>. Data de acesso: 2016-01-19.
- [14] UFO Power Center by Visible Energy, Inc. <http://www.energyufo.com/>. Data de acesso: 2016-01-19.
- [15] Current Cost. <http://www.currentcost.com/>. Data de acesso: 2016-01-19.
- [16] The Energy Detective Electricity Monitor (TED). <http://www.theenergydetective.com/>. Data de acesso: 2016-01-19.
- [17] Cloogy. <http://www.cloogy.com/en/>. Data de acesso: 2016-01-19.
- [18] Nest. <https://nest.com/>. Data de acesso: 2016-01-19.
- [19] Ecobee. <https://www.ecobee.com/>. Data de acesso: 2016-01-19.

- [20] Tendril. <https://www.tendrilinc.com/>. Data de acesso: 2016-01-19.
- [21] Radio Thermostat. <http://www.radiothermostat.com/>. Data de acesso: 2016-01-19.
- [22] LIFX. <http://www.lifx.com/>. Data de acesso: 2016-01-19.
- [23] Jacopo Torriti. *Peak Energy Demand and Demand Side Response*. Routledge, 2015.
- [24] Redes Inteligentes - EDP. <http://www.edpdistribuicao.pt/pt/rede/InovGrid/Pages/RedesInteligentes.aspx>. Data de acesso: 2016-01-19.
- [25] Rede elétrica inteligente – Wikipédia. https://pt.wikipedia.org/wiki/Rede_elétrica_inteligente. Data de acesso: 2016-01-19.
- [26] Entidade Reguladora dos Serviços Energéticos (ERSE). Caracterização da procura de energia elétrica em 2016. *ERSE, Dezembro*, page 110, 2015.
- [27] T. Warren Liao. Clustering of time series data - A survey. *Pattern Recognition*, 38(11):1857–1874, 2005.
- [28] Jie Chen and A. K. Gupta. on Change Point Detection and Estimation. *Communications in Statistics - Simulation and Computation*, 30(3):665–697, 2001.
- [29] Fuxiao Li, Zheng Tian, Yanting Xiao, and Zhanshou Chen. Variance change-point detection in panel data models. *Economics Letters*, 126:140–143, 2015.
- [30] David Lachut, Nilanjan Banerjee, and Sami Rollins. Predictability of Energy Use in Homes. *Proceedings of the International Green Computing Conference (IGCC 14)*, pages 1–10, 2014.
- [31] Taghrid Samak, Christine Morin, and David Bailey. Energy consumption models and predictions for large-scale systems. *Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium Workshops and PhD Forum, IPDPSW 2013*, pages 899–906, 2013.
- [32] Ignacio González Alonso, María Rodríguez Fernández, Juan Jacobo Peralta, and Adolfo Cortés García. A Holistic Approach to Energy Efficiency Systems through Consumption Management and Big Data Analytics. *International Journal on Advances in Software*, 6(3 and 4):261–271, 2013.
- [33] Javier Campillo, Fredrik Wallin, Daniel Torstensson, and Iana Vassileva. Energy demand model design for forecasting electricity consumption and simulating demand response scenarios in Sweden. *The Fourth International Conference on Applied Energy 2012*, 2012.
- [34] Pasapitch Chujai, Nittaya Kerdprasop, and Kittisak Kerdprasop. Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. *Proceedings of the International MultiConference of Engineers and Computer Scientists, I*, 2013.
- [35] Yongcheol Shin, Denis Kwiatkowski, Peter Schmidt, and Peter C. B. Phillips. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root : How Sure Are We That Economic Time Series Are Nonstationary? *Journal of Econometrics*, 54:159–178, 1992.
- [36] Christopher Bennett, Rodney a. Stewart, and Junwei Lu. Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*, 7(5):2938–2960, 2014.

- [37] Zaid Mohamed and Pat S Bodger. Forecasting electricity consumption: A comparison of models for new zealand. Master’s thesis, University of Canterbury. Electrical and Computer Engineering., 2004.
- [38] Domenico Piccolo. A distance measure for classifying arima models. *Journal of Time Series Analysis*, 11(2):153–164, 1990.
- [39] Félix Iglesias and Wolfgang Kastner. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. *Energies*, 6(2):579–597, 2013.
- [40] Joshua D. Rhodes, Wesley J. Cole, Charles R. Upshaw, Thomas F. Edgar, and Michael E. Webber. Clustering analysis of residential electricity demand profiles. *Applied Energy*, 135:461–471, 2014.
- [41] Alexander Lavin and Diego Klabjan. Clustering time-series energy data from smart meters. *Energy Efficiency*, 8(4):681–689, 2015.
- [42] Lajos Horváth and Marie Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.
- [43] João Barbosa. Automating Energy with the Internet of Things. Master’s thesis, Universidade de Coimbra. Departamento de Engenharia Informática, 2013.
- [44] J. Z. Kolter, Siddharth Batra, and Andrew Y. Ng. Energy Disaggregation via Discriminative Sparse Coding. *Advances in Neural Information Processing Systems*, pages 1153–1161, 2010.
- [45] Sparse Coding - Ufldl. http://ufldl.stanford.edu/wiki/index.php/Sparse_Coding. Data de acesso: 2016-01-19.
- [46] Andrei Sebastian Ardeleanu and Codrin Donciu. Nonintrusive load detection algorithm based on variations in power consumption. *EPE 2012 - Proceedings of the 2012 International Conference and Exposition on Electrical and Power Engineering*, pages 309–313, 2012.
- [47] Pietro Cottone, Salvatore Gaglio, Giuseppe Lo Re, and Marco Ortolani. User activity recognition for energy saving in smart homes. *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*, pages 1–9, 2013.
- [48] Hyun Sang Cho, Tatsuya Yamazaki, and Minsoo Hahn. AERO: Extraction of user’s activities from electric power consumption data. *IEEE Transactions on Consumer Electronics*, 56:2011–2018, 2010.
- [49] Chao Chen, Diane J. Cook, and Aaron S. Crandall. The user side of sustainability: Modeling behavior and energy usage in the home. *Pervasive and Mobile Computing*, 9(1):161–175, 2013.
- [50] Alaa Alhamoud, Felix Ruettiger, Andreas Reinhardt, Frank Englert, Daniel Burgstahler, B Doreen, Christian Gottron, and Ralf Steinmetz. SMARTENERGY . KOM : An Intelligent System for Energy Saving in Smart Home. *The 39th IEEE Conference on Local Computer Networks*, (September):685–692, 2014.
- [51] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.

- [52] Wayne A Taylor. Change-Point Analysis: A Powerful New Tool For Detecting Changes. *Analysis*, pages 1–19, 2006.
- [53] Maria L Rizzo. *Statistical computing with R*. CRC Press, 2007.
- [54] X Francis and S Roberto. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 1995.
- [55] Rob J Hyndman, Andrey V Kostenko, et al. Minimum sample size requirements for seasonal forecasting models. *Foresight*, 6(Spring):12–15, 2007.
- [56] UndergroundWeather. <https://www.wunderground.com/>. Data de acesso: 2016-04-07.
- [57] Escalões de potência. <https://www.eem.pt/pt/conteudo/clientes/contratacao/escaloes-de-potencia/>. Data de acesso: 2016-06-10.
- [58] Jan G. De Gooijer and Rob J. Hyndman. 25 Years of Time Series Forecasting. *International Journal of Forecasting*, 22(3):443–473, 2006.

Apêndices

A Clustering de séries de consumo normalizadas

A.1 Clustering de séries normalizadas de consumo octa-horário



Figura 40: Valores de \mathcal{E} para o consumo octa-horário não-normalizado para cada valor de k .

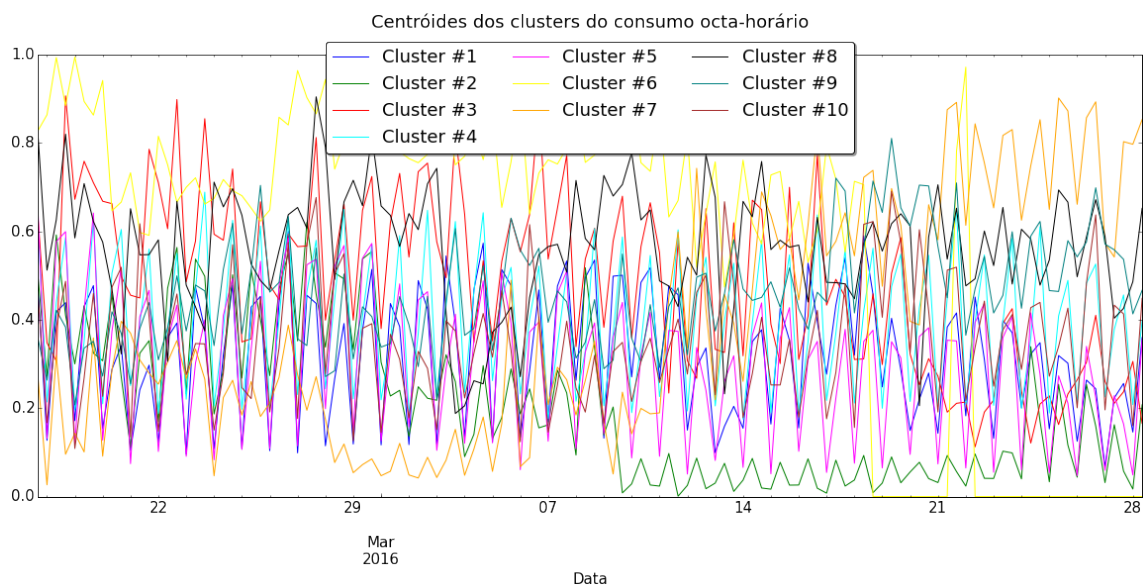


Figura 41: *Clustering* de um conjunto de 100 séries normalizadas de consumo octa-horário doméstico.

A.2 Clustering de séries normalizadas de consumo diário



Figura 42: Valores de ε para o consumo diário não-normalizado para cada valor de k .

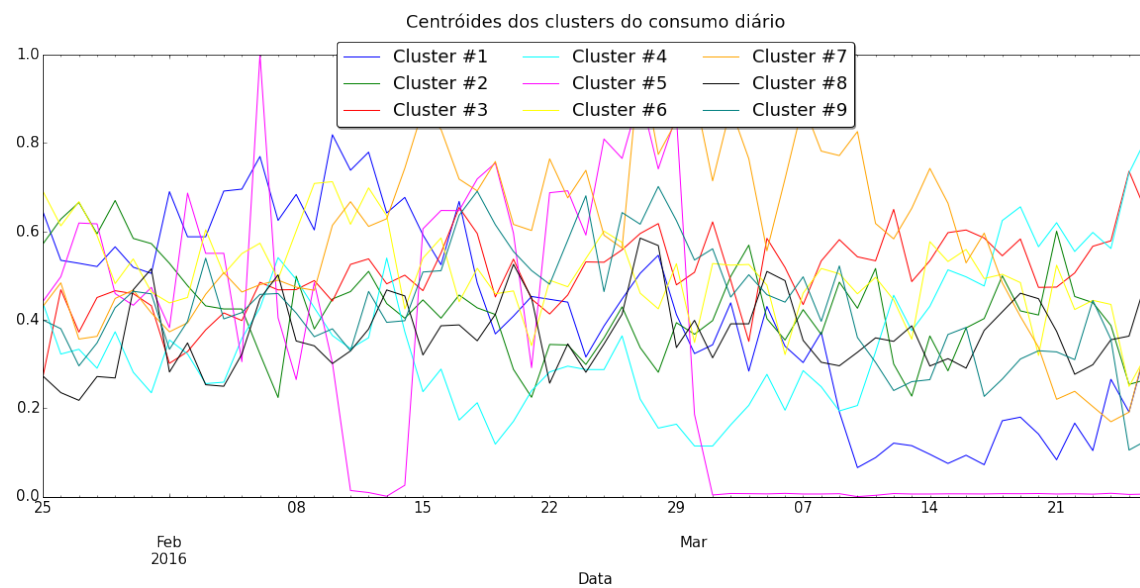


Figura 43: *Clustering* de um conjunto de 100 séries normalizadas de consumo diário doméstico.