LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

APLICAÇÃO DE MODELOS PREDITIVOS PARA O SETOR ALIMENTAR: UM ESTUDO COMPARATIVO

LEONARDO LOURENÇO DE ALMEIDA

NOVEMBRO - 2020

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM
ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO
RELATÓRIO DE ESTÁGIO

APLICAÇÃO DE MODELOS PREDITIVOS PARA O SETOR ALIMENTAR: UM ESTUDO COMPARATIVO

LEONARDO LOURENÇO DE ALMEIDA

ORIENTAÇÃO:

PROFESSOR DOUTOR JOÃO NICOLAU

NOVEMBRO – 2020

Agradecimentos

Em primeiro lugar agradecer ao meu orientador, Professor Doutor João Nicolau, pois sem ele o presente trabalho não seria possível, estando presente em todas as fases no decorrer do estágio, mas também pelo apoio concedido no desenvolvimento deste relatório;

Agradeço igualmente ao Eng. Luís Batista responsável do projeto na Gallo Worldwide, por todo apoio, disponibilidade e dedicação prestados no decorrer do estágio;

Aos professores e alunos da Universidade de Aveiro e do ISEG, que contribuíram durante para o meu percurso académico e na minha aprendizagem;

Ao Adriano Pinho por toda a ajuda incondicional, paciência e preocupação;

Aos meus amigos, em particular à Beatriz Alvadia, Beatriz Rocha, Kevin Pais, Liliana Ribeiro e ao Ludgero Glórias pelo vosso companheirismo, paciência, conselhos e ajudas;

E por fim, à minha família, em especial aos meus pais, por tudo o que me ensinaram, por todos os valores que me incutiram e por sempre terem apoiado as minhas escolhas.

Resumo

Na sociedade atual a inovação surge como um papel cada vez mais preponderante nas empresas. O presente relatório surge no âmbito de um estágio curricular desenvolvido numa empresa líder a nível mundial no comércio grossista de azeites, com o principal objetivo de encontrar um modelo capaz de prever os preços das suas mercadorias. Para tal, foram analisadas várias metodologias, fazendo uma junção entre modelos tradicionais e mais inovadores e recentes. Sendo por isso, analisados os modelos ARIMA; ARIMAX; VAR como modelos mais tradicionais, em contradição às redes neuronais artificiais do tipo MLP; GMDH. Para o estudo de caso foram utilizados os dados dos três azeites de mais interesse para a empresa, distribuídos por dois conjuntos temporais diferentes, permitindo assim a análise do impacto da dimensão da amostra nas previsões. Estudou-se o impacto de variáveis independentes (nomeadamente meteorológicas, macroeconómicas, entre outras que afetam a produção da azeitona), têm nos preços de compra do azeite.

Os resultados apontam para um melhor desempenho do modelo VAR em todos os grupos de dados em análise, obtendo assim as melhores previsões dentro do conjunto de modelos. Destaca-se ainda, a preferência de modelos mais tradicionais quando a série tem um menor comprimento temporal, e uma melhor eficácia das redes neuronais em conjuntos de dados mais elevados, destacando ainda a preferência da rede do tipo GMDH face à rede MLP. Conclui-se ainda, que dentro do vasto conjunto de variáveis em análise, é uma variável binária que influencia a produção (safra), a que possuí maior impacto nas previsões.

Palavras Chave:

Séries Temporais; Previsão; Redes Neuronais Artificiais; Azeite; ARIMA; ARIMAX; VAR; MLP; GMDH.

Abstract

In today's society, innovation appears as an increasingly prevalent role in companies. This report comes as a part of a curricular internship developed at a world leader in the wholesale of olive oil with the main objective of finding a model capable of predicting the prices of its goods. To this end, several methodologies were analyzed, making a junction between traditional and more innovative and recent models. Therefore, the ARIMA models were analyzed; ARIMAX; VAR as more traditional models, in contradiction to artificial neural networks of the MLP type; GMDH. For the case study, data from the three olive oils of most interest to the company was used, distributed over two different time sets. Thus, allowing the analysis of the impact of the sample size on the forecasts. The impact of independent variables (namely meteorological, macroeconomic, among others that affect olive production) was studied on the purchase prices of olive oil.

The results point to a better performance of the VAR model in all groups of data under analysis, thus obtaining the best forecasts within the set of models. Also, noteworthy is the preference for more traditional models when the series has a shorter time length, and a better efficiency of neural networks in higher data sets, also highlighting the preference of the GMDH type network over the MLP network. It is also concluded that, within the vast set of variables under analysis, it is a binary variable that influences production (*safra*), which has the greatest impact on forecasts.

Keywords:

Time Series; Forecast; Artificial Neural Networks; Olive oil; ARIMA; ARIMAX; VAR; MLP; GMDH.

Índice

ÍNDICE DE FIGURAS

Corpo do Texto

Anexos

ÍNDICE DE TABELAS

Corpo do Texto

Anexos

GLOSSÁRIO

ACF *Autocorrelation Function*

ADF *Augmented Dickey-Fuller*

AIC *Akaike Information Criteria*

AICc *Akaike's Information Criterion corrected*

ANFIS *Adaptive neuro fuzzy inference system*

ANN *Artificial neural network*

AR *Autoregressive*

ARIMA *Autoregressive integrated moving average*

ARIMAX *Autoregressive Integrated Moving Average with Explanatory Variable*

ARMA *Autoregressive–moving-average*

BIC *Bayesian Information Criteria*

CRISPP-DM *Cross Industry Standard Process for Data Mining*

EVI Azeite Extra Virgem

GARCH *Autoregressive conditional heteroskedasticity*

GMDH *Group method of data handling*

GTS *General-to-specific*

GUI *Graphical user interface*

HQ *Hannan–Quinn information criterion*

IA Inteligência Artificial

IFAPA *Investigación y Formación Agraria y Pesquera*

KDD *Knowledge Discovery and Data Mining*

LAM Azeite Lampante

LR *Likelihood-ratio test*

LSTM *Long short-term memory*

MA *Moving average*

MAPE *Mean Absolute Percentage Error*

ML *Machine Learning*

MLP *Multilayer perceptron*

MSE *Mean Squared Error*

OMS Organização Mundial da Saúde

PACF *Partial autocorrelation function*

RBF *Radial basis function*

RMSE *Root Mean Squared Error*

SEMMA *Sample, Explore, Modify, Model and Assess*

STL *Seasonal and Trend decomposition using Loess*

SUR *Seemingly unrelated regressions*

TFM Trabalho final de Mestrado

VAR *Vector autoregression*

VIR Azeite Virgem

"A oliveira e o azeite estão profundamente ligados aos povos do Mediterrâneo, da alimentação à arte e à religião."

Reis (2014, p.7)

1. Introdução

1.1 Enquadramento

Na sociedade atual, marcada pela competitividade, a capacidade de inovação e adaptação permite às empresas valorizarem-se face à concorrência. Sendo as instituições de ensino superior, motores de inovação, a cooperação Universidades-Industrias é um fator importante para o desenvolvimento económico de um país e promoção de novo conhecimento (Pereira, 2004).

O estágio realizado e descrito no contexto deste Trabalho Final de Mestrado (TFM), teve início de uma parceria realizada entre o Instituto Superior de Economia e Gestão – Universidade de Lisboa, no âmbito do mestrado em Econometria Aplicada e Previsão, e a empresa Gallo Worldwide, Lda.

A Gallo Worldwide é uma *joint-venture* entre a Unilever[1] e a Sociedade Francisco Manuel dos Santos, fundada em Portugal em 1919. É detentora da marca Gallo, primeira marca portuguesa de azeite, e quarta no ranking mundial, presente em mais de 40 países. Especializou-se no comércio por grosso de azeite, óleos e gorduras alimentares, estando há mais de 100 anos presente no mercado, assente num paralelismo entre a inovação e a tradição.

Deste modo, propôs-se a criação de um modelo preditivo que permitisse estimar os preços das mercadorias mensais e semanais, dos três tipos de azeite mais utilizados pela empresa. Este projeto foi desenvolvido em *home office*, durante aproximadamente 6 meses, entre novembro de 2019 e abril de 2020. Ressalve-se desde já a abrangência do tema, e de que forma uma boa estimação dos indicadores, nomeadamente o preço de compra das matérias-primas e de venda do produto final, pode permitir a redução de custos, maximizando os lucros (Doganis et al., 2006).

O ato de prever, por mais confiável que seja, tem sempre alguma incerteza associada. Sendo, então, o objetivo primordial deste trabalho encontrar o modelo que melhor se ajuste à previsão, reduzindo a sua incerteza. Dentro do vasto universo de modelos de previsão, o presente relatório irá debruçar-se nos modelos baseados em Machine Learning (ML), e comparando-os com os modelos tradicionais ARIMA e VAR.

[1] A Unilever é uma das principais empresas de bens de consumo a nível mundial, criando e vendendo cerca de 400 marcas em mais de 190 países.

Técnicas de ML e inteligência artificial (IA) oferecem benefícios significativos aos tomadores de decisão em termos de novas abordagens de modelação e previsão de dados, sendo prova disso, os cerca de 28 mil milhões de dólares anuais esperados a serem gastos globalmente por firmas do setor financeiro em inteligência artificial até 2021 (Aziz et al., 2019).

As últimas décadas foram caracterizadas pela aceleração (velocidade) na produção de dados com maior variedade e volume, os três "Vs" que qualificam o que é popularmente conhecido como *Big Data*, sendo este o fator propício à aplicação de técnicas de ML, justificando o aumento da sua popularidade. Simultaneamente, o avanço tecnológico tem impulsionado esta popularidade, devido às melhorias significativas de poder de processamento computacional e armazenamento de dados (Coenen, 2011), que permitem atenuar um dos flagelos do ML, o elevado tempo de treino e a capacidade computacional necessária.

1.2 Motivação

A principal motivação para a realização deste estágio surge devido ao ambiente em que se insere, isto quer dizer, trata-se de um desafio aliciante, realizado em ambiente laboral, numa das maiores empresas a nível nacional e mundial do seu setor - o que, desde logo, é uma mais-valia para uma primeira experiência a nível profissional. Por outro lado, é um tema com uma enorme abrangência que pode ser utilizado e aplicado a inúmeras situações e que poderá dar um forte contributo para impulsionar o crescimento da empresa.

1.3 Problema

O desafio proposto consiste na construção de um modelo que permita prever o preço das mercadorias, semanal e mensalmente. Para além disso, a previsão deverá ser feita para os três tipos de azeite principais que a empresa comercializa, nomeadamente, o Azeite Extra Virgem (EVI), Virgem (VIR), Lampante (LAM). A dimensão temporal deverá também ser suficiente para que, ao dia 15 de cada mês, seja possível prever o mês seguinte, no horizonte semanal e mensal.

1.4 Objetivos

Os objetivos inicialmente traçados pela empresa baseavam-se no desenvolvimento do melhor modelo de previsão, obtendo este através da comparação entre modelos, não só os presentes neste estudo, como os desenvolvidos em simultâneo por outro aluno a quem foi proposto o mesmo estágio e problemática, tendo este abordado o problema através de modelos SUR[2] e GARCH[3].

Foi pedido ainda o desenvolvimento de uma medida de erro que se baseie na variação das séries, comparando se a variação da previsão corresponde à verdadeira variação da série de preços, tendo em conta um critério de margem negocial definido pela empresa.

Por outro lado, os resultados devolvidos pelo modelo deverão ser de fácil compreensão, e de fácil implementação, num *software* de acesso livre. Permitindo um baixo custo de aprendizagem aos diversos utilizadores.

Assim sendo, o principal objetivo foi construir um modelo realista e flexível, que fosse capaz de fazer uma previsão eficaz e eficiente para todos os tipos de azeite, tendo em especial consideração o problema da variação proposto pela empresa.

1.5 Estrutura do TFM

O presente relatório encontra-se estruturado da seguinte forma: O capítulo 1 diz respeito à apresentação da empresa alvo deste TFM, tal como faz uma breve introdução da mesma e o seu enquadramento. Apresenta igualmente, os problemas, as motivações e os objetivos a realizar ao longo do estágio e da redação da TFM. No capítulo 2, são apresentados os conceitos teóricos fundamentais à compreensão do problema e as principais motivações que levaram à seleção de cada modelo, através da revisão de trabalhos relacionados, sendo introduzida ainda a revisão da literatura relevante sobre a aplicação de diferentes modelos na previsão; No capítulo 3, são apresentadas as principais tecnologias e metodologias aplicadas em trabalhos desta natureza, sendo ainda descritas as principais bibliotecas adotadas na realização deste projeto; O capítulo 4 é constituído pela aplicação prática de cada uma das etapas que compõe a metodologia CRISP-DM. Por fim, no capítulo 5 são apresentadas as conclusões do trabalho realizado e as recomendações futuras do trabalho e da investigação.

[2] *Seemingly Unrelated Regressions*
[3] *Generalized Autoregressive Conditional Heteroskedasticity*

2. Revisão da Literatura

Num mercado em constante expansão, onde constantemente surgem novos produtos e hábitos, o azeite tem se solidificado como um produto com cada vez mais importância no setor comercial. No ano de 2019, os dados provisórios do *International Olive Council*[4], apontam para que no mundo a produção de azeite tenha atingido as 3.217.500 toneladas, onde destas, a europa representa cerca de 70%, sendo Espanha o país produtor número um, com uma produção de 1.789.900t. de azeite, enquanto que em Portugal a produção rondará as 100.316t. (IOC, 2019).

Os agentes envolvidos no setor estão especialmente interessados na previsão do preço do azeite, pois uma previsão precisa poderá levar a um aumento global dos benefícios (Pérez-Godoy et al., 2010). De facto, para empresas que procuram um bom desempenho das vendas, o conhecimento prévio do preço das mercadorias é, sem dúvida, uma mais-valia. Empresas grossistas precisam de manter o equilíbrio de stocks, entre responder à procura dos clientes, e controlar os custos operacionais. Não manter em stock unidades suficientes do seu produto para atender à procura, poderá levar à perda de vendas para um concorrente (Pankratz, 1983). No entanto, possuir um stock elevado que permite satisfazer a procura dos clientes o tempo todo, pode resultar em excesso de stock, levando a questões como aumento dos custos de manutenção, ou a "*inventory writedowns*"[5], que geram uma redução de margens de lucro (Lu et al., 2012). É na melhoria das margens de lucro que este estudo se foca, uma boa previsão do preço das mercadorias permite saber quando comprar, aumentando os stocks, para evitar uma futura subida de preços, ou saber quando adiar a compra para esperar por uma descida de preços, tendo em consideração os stocks existentes, permitindo um novo controle dos stocks, onde se tem em consideração o preço das mercadorias.

Para prever séries temporais financeiras, é necessário captar os comportamentos e características da mesma, para assim, se formular um modelo matemático que melhor se adeque e represente a série temporal. Alguns dos comportamentos típicos de séries temporais são resultantes da influência de diversos fatores, como alterações macroeconómicas, evolução tecnológica, mudanças ambientais, entre outros fenómenos

[4] *International Olive Council* (IOC) é um organismo internacional que representa todos os 17 países membros que, dentro da Instituição, produzem azeitona e azeite.
5 *inventory writedowns* ou perdas por imparidade, acontecem quando existe uma desvalorização dos inventários, ou seja, o valor de venda atual das mercadorias passou a ser inferior ao valor de aquisição.

imprevisíveis (Bouzada, 2012). Geralmente essas características inerentes às séries temporais podem ser decompostas em quatro componentes principais, sendo elas tendência, sazonalidade, ciclo e erro. No entanto, a separação entre a tendência e os componentes cíclicos não é fácil de realizar e normalmente, são combinados numa componente conjunta, denominada por ciclo de tendência (Damrongkulkamjorn & Churueang, 2005).

Existe uma grande quantidade de modelos de previsão que conseguem obter bons resultados a partir de séries temporais. Através da literatura verificou-se que para os preços de mercadorias, em especial destaque o setor agrícola e da olivicultura, essa abundância não é exceção. Assim a figura 1, pretende demonstrar alguns artigos onde foram aplicados diversos modelos ao longo dos tempos por vários autores para os diferentes tipos de mercadorias.

Dentro do vasto leque de modelos existentes serão selecionados cinco, divididos sob duas premissas, a existência de uma relação linear entre os valores futuros de uma série temporal e os valores atuais e passados, obtendo-se assim modelos lineares (2.1.1); e os modelos não lineares, onde não existe linearidade entre as variáveis (2.1.2).

Um dos modelos mais conhecidos é o ARIMA - *Autoregressive Integrated Moving Average* (Box & Jenkins, 1976), sendo amplamente utilizado, possui uma grande aplicabilidade em problemas de previsão de séries temporais, fornecendo previsões precisas em curtos períodos de tempo, com relativa facilidade de implementação (Torbat et al., 2018). Sendo, portanto, o primeiro modelo a ser considerado e que servirá de ponto de partida para a seleção dos modelos seguintes. Apesar da sua ampla popularidade, este modelo apresenta limitações, sendo a principal, a forma linear do modelo, assumindo uma estrutura de correlação linear entre os valores das séries temporais e, portanto, nenhum padrão não linear pode ser capturado pelo modelo ARIMA (Zhang, 2003).

Como visto, alguns comportamentos típicos de séries temporais são resultantes da influência de diversos fatores, por isso é frequente a necessidade de incorporar uma ou mais séries temporais para prever o valor de outra série, esta é outra limitação que se impõe ao modelo ARIMA, que apenas explora o relacionamento entre a variável prevista e o seu passado, excluindo qualquer relacionamento entre a variável em estudo e outra variável económica (Spencer, 1993). Face a esse problema, Box & Tiao (1975) introduziram uma modificação ao modelo ARIMA, transformando-o num modelo multivariado, sendo vulgarmente conhecido por ARIMAX - *Autoregressive Integrated*

Moving Average with Explanatory Variable, este passa por uma alternativa linear ao modelo ARIMA, permitindo incluir o estudo dos modelos multivariados, ou seja, a inclusão de novas variáveis. O modelo ARIMAX não surge com tanta frequência quando se trata de previsões no campo económico, no entanto, mostra-se "bom" em previsões de variáveis afetadas por outras variáveis (Anggraeni et al., 2017).

O problema com o modelo ARIMAX surge, quando é necessário possuir os valores futuros das variáveis exógenas, para realizar a previsão da variável dependente (Zhao et al., 2016), valores estes geralmente desconhecidos, e de relativo interesse em obter previsões dos mesmos. Sims (1980) desenvolveu um modelo multivariado capaz de prever conjuntamente todas as variáveis não fazendo distinção entre exógenas e dependentes, o VAR - *Vector Autoregression*, permite capturar as correlações intertemporais de uma seleção de variáveis económicas.

Geralmente, os modelos anteriores podem fornecer bons resultados de previsão quando a série em estudo é linear ou quase linear. No entanto, séries onde exista muita não linearidade, o desempenho da previsão é menor, devido a estes modelos econométricos serem construídos sobre suposições de linearidade, não conseguindo capturar os padrões não-lineares ocultos (Yu et al., 2008). Desde as últimas décadas do século XX que o interesse por técnicas de computação leve (redes neuronais) para a modelação e previsão tem vindo a aumentar. Modelos baseados em inteligência artificial têm vindo a superar os modelos baseados em estatística, na previsão de séries temporais (Do & Yen, 2019), e em comparação com os modelos econométricos tradicionais, métodos baseados em técnicas de computação leve fornecem um maior grau de robustez e capacidade de prever a volatilidade (Haidar et al., 2008).

As ANN's - *Artificial Neural Network,* ganharam destaque dentro do conjunto de *machine learning*, como modelos alternativos para previsões económicas e financeiras, sendo capazes de executar modelações não lineares sem um conhecimento *à priori* sobre as relações entre as variáveis de entrada e saída (Zhang et al., 1998). Dentro das ANN's a rede *multilayer percepton* (MLP) criada por Rosenblatt (1958) é considerada uma das mais importantes, tendo como principal vantagem a sua arquitetura. Uma vantagem ainda associada às ANN's é, segundo Jha & Sinha (2013), a sua forma funcional, flexível e universal, não havendo a necessidade de especificar um modelo particular para um determinado conjunto de dados. Contudo Baron (1994), afirma que as ANN's têm várias

desvantagens, como serem incapazes de explicar as suas decisões, agindo como "caixas negras" nas quais não se sabe o porquê de chegarem a um determinado resultado.

O modelo GMDH - *Group Method of Data Handling* (Ivakhnenko, 1971), é uma derivação das ANN's, inserida no ramo das ANN's *feedfoward*. É considerada uma rede neuronal polinomial capaz de quebrar o conceito de "caixa negra", característico das redes neuronais (Schneider & Steiner, 2005). O GMDH foi comprovado como um método eficaz em campos, como *Data Mining*, previsão, otimização e reconhecimento de padrões (Teng et al., 2014), sendo que Do & Yen (2019) consideram o GMDH como um bom modelo para servir de ferramenta na previsão do preço de mercadorias, capaz de superar as previsões apresentadas por outros tipos de redes neuronais, e.g. ANFIS[6], ANN[7] e LSTM[8].

2.1 Modelos de Previsão em Séries Temporais

2.1.1 Modelos Lineares

Modelos lineares são muito flexíveis e por isso amplamente utilizados em diversas áreas do saber como ciências físicas, engenharia, ciências sociais e finanças (Faraway, 2015). São modelos simples e de fácil compreensão, que apresentam uma relação entre variáveis linear nos parâmetros, o que implica que a variação de um parâmetro seja independente da variação de outro. Por exemplo, uma expressão da forma $E(Y|x) = \alpha + \beta x^2$, é um modelo linear em $\alpha\ e\ \beta$, mas o modelo $E(Y|x) = \alpha e^{\beta x}$, não é um modelo linear em $\alpha\ e\ \beta$.

A linearidade pode ainda ser definida através da média, onde segundo Nicolau (2012), um modelo $y_t = \mu_t + u_t$, onde u_t são os erros e $\mu_t = g(y_{t-1} \dots y_{t-p}; u_{t-1}, \dots, u_{t-p})$ é a média condiciona, é considerado linear na média, se a função g é linear nos seus argumentos. Ou seja, a especificação $\mu_t = \phi y_{t-1} + \theta u_{t-1}$ é linear, enquanto que, $\mu_t = \phi y_{t-1}^2$ é não linear na média.

[6] *Adaptive neuro fuzzy inference system*
[7] ANN do tipo e *multilayer perceptron*, com duas camadas ocultas
[8] *Long short-term memory*

2.1.1.1 *Autoregressive Integrated Moving Average*

Os modelos ARIMA, vulgarmente conhecidos por modelos *Box-Jenkins*, analisam séries temporais estocásticas univariadas. Usando a terminologia usual de *Box e Jenkins*, uma notação parcimoniosa dos modelos ARIMA, com termos sazonais, pode ser escrita na forma:

$$\text{ARIMA}\ (\mathrm{p}, \mathrm{d}, \mathrm{q})(\mathrm{P}, \mathrm{D}, \mathrm{Q})_s, \tag{1}$$

onde p, P correspondem ao número de parâmetros da parte autorregressiva (AR), não sazonais e sazonais, respetivamente. Os parâmetros d, D, representam o número de diferenciações, não sazonais e sazonais, que são necessárias para transformar as séries não estacionárias em séries estacionárias. s, representa o período sazonal. Por fim, os parâmetros q, Q determinam o número de médias móveis (MA), não sazonais e sazonais, a ser utilizado.

Deste modo, a aplicação do modelo ARIMA é uma abordagem iterativa de modelação em 3 estádios – identificação, estimação e diagnóstico, para finalmente aplicar a previsão.

Na fase de identificação, deve-se realizar um exame visual ao gráfico da série temporal, permitindo observar os comportamentos da série temporal, como por vezes a estacionaridade, valores extremos, valores ausentes, etc. Deve-se verificar nesta etapa a estacionaridade da série, não só através da análise gráfica, mas também com o auxílio de um teste formal de raiz unitária, exemplo o teste *augmented Dickey–Fuller* (ADF), que permite assim encontrar o valor da diferenciação. A hipótese nula deste teste assume que a série tem uma raiz unitária. Convencionou-se que para o restante TFM não se rejeita as hipóteses nulas, se o valor de probabilidade (*p-value*), associado à componente estatística do teste, for superior a 0,05. Uma vez encontrado, o valor que permite atingir a estacionaridade, o próximo passo é identificar os parâmetros do modelo, ou seja, ordens AR e MA, examinando a função autocorrelação (ACF) e a função autocorrelação parcial (PACF).

Na etapa de estimação, é estimado cada um dos modelos provisórios atribuindo várias ordens p, P e q, Q, devendo a seleção inicial destas ordens ter em consideração a análise da PACF e ACF respetivamente. Os modelos estimados são comparados através dos critérios de informação de Akaike (AIC), Schwarz-Bayesian (BIC) e *Akaike's Information Criterion corrected* (AICc), onde o melhor modelo é segundo Ramos et al.

(2015), aquele cuja a combinação das ordens p, P e q, Q minimizem os critérios de informação AIC e AICc.

Por vezes, nem sempre é fácil seguir este procedimento e encontrar os valores dos parâmetros do modelo. Deste modo é útil recorrer-se a algumas ferramentas já desenvolvidas e presentes em *softwares* informáticos, como o *R Core Team* (2019), que fazem essa estimação de forma rápida e automática, devolvendo o modelo que melhor se ajusta aos dados (Hyndman & Khandakar, 2008).

No estádio diagnóstico, é examinada a qualidade do ajuste do modelo. Esta avaliação ocorre através da análise dos resíduos. Idealmente estes devem atender às suposições de ruído branco, ou seja, não serem auto-correlacionados. Sendo para isso, realizado o teste *Ljung-Box* que permite testar se as primeiras k auto correlações dos resíduos são diferentes do que seria esperado num processo de ruído branco. A hipótese nula assume, que as primeiras k auto correlações são nulas. Foram ainda analisados os gráficos ACF e PACF dos resíduos, de modo a identificar a existência de padrões com elevado valor de probabilidade estatística.

Se as condições de ruído branco não forem constatadas, ou seja, se o modelo "falhar" no teste *Ljung-Box*, ou se existirem picos nos gráficos ACF e PACF dos resíduos fora dos limites ($\pm 2/\sqrt{n}$), é necessário ajustar-se um novo modelo, até que os pressupostos sejam satisfatórios.

2.1.1.2 *Autoregressive Integrated Moving Average with Explanatory Variable*

A metodologia de Box-Tiao, comumente chamada de ARIMAX, representa uma extensão do modelo ARIMA, através da adição de entradas exógenas (X), tornando assim o ARIMAX num modelo multivariado (Camelo et al., 2018). Sendo essa a principal diferença entre os dois modelos, o ARIMAX continua a possuir os parâmetros autorregressivos e de médias móveis, acrescentando a introdução de variáveis exógenas, que devem ser bem identificadas, para evitar a inclusão de variáveis irrelevantes (Bennett et al., 2014).

De forma semelhante ao modelo ARIMA, o ARIMAX pode ser escrito na forma parcimoniosa, com termos sazonais:

$$\text{ARIMAX}\ (\mathrm{p,d,q,r})(\mathrm{P,D,Q})_s, \quad (2)$$

onde p, d, q, P, D, Q, s tem o mesmo significado que no modelo ARIMA visto anteriormente na equação (1), adicionando apenas r, que representa o número de variáveis exógenas (Camelo et al., 2018).

Harvey (1990) trata o problema de modelação do ARIMAX como uma extensão da modelação ARIMA, uma vez que os distúrbios são gerados por um processo ARMA[9](p, q). Assim a aplicação do modelo ARIMAX tem uma abordagem equivalente à modelação em 3 estádios do modelo ARIMA – identificação, estimação, diagnóstico e finalmente, previsão. Cumprindo os passos anteriormente relatados, de identificação das ordens $(p, d, q)(P, D, Q)$, e ainda os mesmos critérios de seleção, os critérios de informação AIC, AICc e BIC. É ainda aplicado ao modelo selecionado os mesmos testes de diagnóstico aos resíduos.

Tal como no ARIMA, é possível recorrer ao uso do *software R*, fazendo este a estimação de forma automática, devolvendo os parâmetros $(p, d, q)(P, D, Q)s$ a considerar, devendo o modelo formado "cumprir" os testes da fase diagnostico.

2.1.1.3 *Vector Autoregression*

Os modelos de vetores autoregressivos (VAR) ganharam popularidade através de Sims (1980), num estudo onde analisou a previsão da taxa de crescimento da atividade real da economia, sendo que as referências técnicas definitivas para os modelos VAR são atribuídas a Lütkepohl (1991). É um dos modelos mais bem-sucedidos, flexíveis e fácil de implementar, na análise de séries temporais multivariadas. O modelo VAR provou ser especialmente útil para descrever o comportamento dinâmico de séries temporais económicas e financeiras para previsão, fornecendo muitas vezes, previsões superiores às apresentadas pelos modelos univariados (Zivot & Wang, 2003). Sendo essa capacidade de previsão, uma das principais vantagens na utilização dos vetores autoregressivos, especialmente previsões de curto prazo, assim vários autores como Tsay (2010), Campbell et al. (1997), Hamilton (1994) utilizam os modelos VAR na previsão de séries temporais financeiras.

[9] Autoregressive Moving Average (Box & Jenkins, 1976)

Este modelo surge da necessidade de se desenvolverem modelos com menores restrições, que tratassem todas as variáveis económicas como variáveis endógenas. Assim, o VAR ao tratar todas as variáveis de forma simétrica, sem a implementação de quaisquer restrições quanto à independência e dependência entre elas, permite descrever cada uma das variáveis endógenas no sistema como uma função dos valores desfasados de todas as variáveis endógenas (Caiado, 2002).

O modelo VAR pode ser expresso através da fórmula matemática de ordem p ou, simplesmente, VAR(p) dada por:

$$Y_t = A_0 + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \varepsilon_t \quad (3)$$

onde $Y_t = (Y_{1t}, \dots, Y_{kt})'$ é um vetor $p \times 1$ de k variáveis endógenas, A_0 é um vetor $p \times 1$ de termos independentes, $A_1, \dots, A_p$ são as matrizes de coeficientes no desfasamento de y, e $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{kt})'$ é um vetor $p \times 1$ de erro aleatório (Caiado, 2002).

Para a implementação de modelos VAR, deve-se ter em consideração o cumprimento das hipóteses, de estacionaridade do vetor de variáveis endógenas, o vetor de erros aleatórios deve ser ruído branco, não autocorrelacionado, com média zero e variância constante $\varepsilon_t \approx N(0, \sigma), Cov(\varepsilon_t, \varepsilon_j) = 0 \vee i \neq j$. Só então se pode proceder à identificação da ordem p do modelo VAR(p), geralmente recorrendo ao método dos critérios de informação, que consiste em ajustar sequencialmente modelos autoregressivos vetoriais de ordens $p = 0, \dots, p_{max}$, e escolher o valor p que minimize os critérios de seleção, sendo os critérios de informação mais utilizados os critérios AIC, BIC e Hannan-Quinn (HQ) (Zivot & Wang, 2003). Krolzig (2001) recomenda ainda o uso do método *general-to-specific* (GTS), demonstrando que este obtém bons resultados na seleção da ordem p, podendo este ser usado como um complemento ao método dos critérios de informação. Este método recorre a testes de hipóteses, habitualmente com estatística da Razão de Verossimilhanças (LR), para testar desde uma ordem máxima p_{max} até $p_{max} - i$, $\vee i = 0, \dots p_{max}$, a hipótese nula $H_0: A_M = 0$, sendo que, a não rejeição da hipótese nula leva ao teste do valor i de ordem seguinte, enquanto que a rejeição desta hipótese leva à escolha da ordem p, sendo $\hat{p}_{GTS} = p_{max} - i$.

2.1.2 Modelos Não Lineares

Uma forma simples de introduzir modelos não lineares, apresentada por Nicolau (2012), consiste em apresentar a não linearidade através dos momentos condicionais. Considerando o modelo:

$$y_t = \mu_t + u_t \qquad u_t = \sigma_t \varepsilon_t \tag{4}$$

onde $u_t = g(y_{t-1}, y_{t-2}, \dots, y_{t-p}; u_{t-1}, u_{t-2}, \dots, u_{t-q})$ é a média condicional de y_t, ε_t é um ruído branco e $\sigma_t = h\left(y_{t-1}, y_{t-2}, \dots, y_{t-\tilde{p}}; u_{t-1}, u_{t-2}, \dots, u_{t-\tilde{q}}\right) > 0$ é a variância condicional de y_t. O modelo é não linear na média no caso da função g ser não linear dos seus argumentos[10]. Trata-se de um modelo não linear na variância, no caso de σ_t não ser constante ao longo do tempo, uma vez que, neste caso, o processo $\{u_t\}$, definido em (4), é não linear por ser um processo multiplicativo (Nicolau, 2012).

Enquanto que a maioria dos modelos lineares é conhecido, essencialmente definidos através da representação ARMA, o número de especificações não lineares é virtualmente infinito (Nicolau, 2012). É nesta infinidade, que as redes neuronais se enquadram.

2.1.2.1 *Artificial Neural Network* - MLP

Máquinas inteligentes, com um grande número de elementos simples, são assunto de pesquisa desde meados do século XIX, mas só no final dos anos 50, que um campo conhecido como redes neuronais começou a ganhar destaque e a evoluir (Harvey, 1994).

As ANN's são modelos capazes de extrair as regularidades, reconhecer padrões e detetar relações em conjuntos de dados aparentemente desconexos (Cavalheiro et al., 2011), assim sendo, têm uma boa capacidade para prever sistemas não lineares, o que os torna relevante para prever séries temporais financeiras, isto é, preços de mercadorias, que geralmente apresentam características de não linearidade (Do & Yen, 2019).

Estas têm a capacidade de poderem ser aplicadas a diversos tipos de dados, isto inclui, dados com ruído, ou dados que apresentem componentes típicas de séries temporais como tendência, sazonalidade ou ciclo. No entanto, estudos têm vindo a demonstrar que quando se aplica séries com algum processamento estas conseguem obter melhores resultados. É o caso de Zhang & Qi (2005), que demostram no seu estudo que

[10] Uma função é não linear se não for uma função linear afim, i.e. se não verificar a relação $f(x_1, \dots, x_n) = a_0 + a_1 x_1 + \cdots + a_n x_n$, onde $a_i \in \mathbb{R}$.

ANN's aplicadas a dados corrigidos de tendência e sazonalidade conseguem obter resultados significativamente mais precisos.

As ANN's são inspiradas no funcionamento do cérebro humano, tentando replicar o comportamento das redes neuronais biológicas. Haykin (2008) afirma que estas se assemelham ao cérebro em dois aspetos:

1º- O conhecimento é adquirido pela rede a partir do seu ambiente através de um processo de aprendizagem.

2º- Forças de conexão entre neurónios, conhecidas como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido.

Ao analisar um neurónio biológico, apresentado na figura 2 (A), observa-se que e consiste numa única célula capaz de realizar um processamento simples. Cada neurónio é então, estimulado através de ligações provenientes de outros neurónios, chamadas sinapses, produzindo um sinal elétrico que se propaga ao longo do axónio até outros neurónios. A representação de um neurónio artificial, figura 2 (B), baseia-se na generalidade dos casos, neste modelo simplificado de funcionamento dos neurónios biológicos, criado por Rosenblatt (1958) também conhecido como *perceptron*.

Em termos matemáticos podemos descrever um neurónio artificial k, figura 2 (B), através do seguinte conjunto de equações:

$$u_k = \sum_{j=1}^{m} w_{kj} * x_j \tag{5}$$

$$y_k = \varphi(u_k + b_k) \qquad v_k = u_k + b_k \tag{6}$$

onde $x_1, x_2, \dots, x_m$, são o impulso de entrada, a quais é atribuído um respetivo peso sináptico $w_{k1}, w_{k2}, \dots, w_{km}$, do neurónio k. u_k , é a saída após o combinador linear ligar os *inputs* aos respetivos pesos, ao qual um valor de viés[11] b_k é adicionado, obtendo assim a combinação de entradas v_k, a que será aplicada a função de ativação φ gerando a saída y_k, posteriormente a ser aplicada como entrada dos neurónios das camadas seguintes (Haykin, 2008).

Em termos gerais qualquer função diferenciável pode ser utilizada como função de ativação, mas na prática, apenas um conjunto limitado de funções de ativação, que

[11] O viés é uma constante (geralmente 1), que tem como função fornecer a cada nó um valor que permita maior flexibilidade de aprendizagem

apresentem um "bom comportamento[12]", são utilizadas (Zhang et al., 1998), sendo estas apresentadas na tabela 1. As redes podem possuir funções de ativação diferentes para diferentes neurónios, na mesma camada ou em camadas diferentes no entanto, quase todas as ANN usam a mesma função de ativação, especialmente para neurónios localizados mesma camada (Zhang et al., 1998). Para a análise de séries temporais é regularmente escolhida a função de ativação sigmóide (Jha & Sinha, 2013).

A estrutura de uma ANN é então constituída por um conjunto de vários neurónios artificiais, que desenvolvem ligações entre si. Essas ligações entre os neurónios podem ser essencialmente divididas em duas classes: Redes *feedfoward* e as redes recorrentes (Silva, 2015). Numa rede do tipo *feedfoward* os neurónios de cada camada apenas estão conectados com os neurónios das camadas seguintes, movendo-se as informações apenas num sentido, para a frente. Dentro desta classe de redes incluem-se vários modelos como *Multi-Layer Perceptron* (MLP), redes com funções de base radial (RBF) e redes polinomiais GMDH (Cavalheiro et al., 2011). Nas redes recorrentes os neurónios contêm ligações para trás e/ou ligações dentro da mesma camada, tornando-as potencialmente instáveis e levando a um aumento do tempo de convergência (Silva, 2015).

Uma ANN multicamada, é composta tipicamente por uma camada de entrada que contém, tantos neurónios/nós de origem quantas as variáveis de *input* (onde nenhuma computação é realizada), uma ou mais camadas ocultas, também conhecidas por neurónios intermédios, e uma camada de saída, cujo número de neurónios corresponde à dimensão do *output* (Haykin, 2008). Um exemplo de uma rede multicamada do tipo *feedfoward*, com uma camada oculta pode ser observado na figura 3.

Após a análise da estrutura de uma ANN, é visível que um dos passos essenciais na sua construção passa pela determinação do número de camadas ocultas e o número de neurónios em cada camada. Wanas et al. (1998) consideram que o melhor desempenho de uma rede neuronal é alcançado quando o número de camadas ocultas é fornecido por $\log(T)$, onde T representa o número de dados de treino. Já Cybenko (1989) defende que não é necessário a utilização de mais de duas camadas oculas, complementado por Thomas et al. (2017) que demonstram que a utilização de duas camadas ocultas é geralmente preferível, na medida que, consegue obter melhor desempenho que apenas uma camada oculta. Em relação ao número de neurónios em cada camada, um número

[12] Zhang et al. (1998) definem como funções "bem comportadas", as funções que sejam limitadas, monotonicamente crescentes e diferenciáveis.

muito elevado de neurónios nas camadas ocultas pode vir a provocar problemas de perda de generalização da rede, ou seja, *overfitting*. Não existe um valor pré-definido sendo a melhor alternativa por tentativa e erro, no entanto, Heaton (2008), sugere que o número de neurónios ocultos deve corresponder a 2/3 do tamanho da camada de entrada somado ao tamanho da camada de saída, nunca podendo ultrapassar o dobro do tamanho da camada de entrada, e estando contido ao intervalo entre o tamanho da camada de entrada e o tamanho da camada de saída.

Para encontrar os pesos sinápticos, w_{km}, uma ANN tem de passar pelo processo de aprendizagem. Primeiramente os dados devem ser divididos em dois subconjuntos, o primeiro subconjunto são os dados de treino que devem incluir 60-90% dos dados totais, e serão utilizados especificamente para o processo de aprendizagem. No outro conjunto encontram-se os dados de teste, os restantes 10-40% dos dados totais, que têm por finalidade verificar se a capacidade de generalização da rede está dentro dos níveis aceitáveis, validando assim a topologia desenvolvida (da Silva et al., 2017).

A aprendizagem de uma ANN do ponto de vista da otimização, é equivalente a minimizar uma função de erro global (Møller, 1993), através da atualização constante dos pesos sinápticos. Em tarefas de previsão, este processo segue, por norma, o paradigma da aprendizagem supervisionada. No entanto, existe ainda outro paradigma principal no aprendizado das ANN's, a aprendizagem não-supervisionada, sendo a sua maior finalidade tarefas de descrição (Gama et al., 2017). A aprendizagem supervisionada tem por base a presença de um supervisor externo, que fornece o conjunto de entradas e saídas, associando a saída a um conjunto de valores de entrada, permitindo avaliar a saída obtida com a saída desejada, informando assim a rede sobre o erro da resposta atual.

É, no tipo de algoritmo que cada modelo usa no processo de aprendizagem dos pesos, que as redes neuronais estudadas no presente TFM mais diferem. Um algoritmo é um processo sequencial finito em que se estipula, com generalidade e sem restrições, as ações executáveis que procuram uma solução para um determinado problema.

O *Multi Layer Perceptron* (MLP) é um dos modelos simplificados criados por Rosenblatt (1958), e é segundo Yilmaz & Kaynar (2011), a arquitetura de rede mais popular e amplamente usada, nas mais diversas áreas de estudo. Este modelo segue a estrutura típica de uma ANN, multicamada, do tipo *feedfoward*, descrito anteriormente.

A popularidade do modelo MLP deve-se em parte ao processo de aprendizagem por este adotado. O MLP recorre ao algoritmo *Backpropagation*, onde o erro da resposta atual é propagado para a camada anterior que ajusta os pesos sinápticos, propagando-o sucessivamente até à camada de entrada, repetindo este processo até o erro ser inferior a um valor limite pré-determinado (Hecht-Nielsen, 1992).

2.1.2.2 *Group Method of Data Handling*

A história da rede neural do tipo GMDH tem origem no final da década de 60 e início dos anos 70. Inicialmente, Alexey Ivakhnenko, em 1966, introduziu um polinómio, que é o algoritmo básico do GMDH (Dag & Yozgatligil, 2016a). Ainda Ivakhnenko, (1970, 1971) introduziu o método de auto-organização heurística, que no ano seguinte se veio a tornar a teoria base do algoritmo GMDH. Mais tarde Kondo (1998) propôs a criação da rede neural do tipo GMDH, na qual o algoritmo funciona de acordo com o método de auto-organização heurística.

De uma forma geral, o GMDH, decompõe sistemas complexos em subconjuntos menores, agrupando-os em várias camadas, permitindo manipular os *inputs*, por meio de combinações, para seguidamente serem enviados aos neurónios da rede. A rede polinomial é construída através do algoritmo GMDH, por um processo de aprendizagem supervisionada, onde os neurónios avaliam um número limitado de entradas, através de uma função de ativação polinomial, gerando uma saída, que servirá como entrada nos neurónios da camada seguinte (rede do tipo *feedfoward*) (Cavalheiro et al., 2011). Estando este procedimento diretamente relacionado com as redes neuronais biológicas, onde primeiramente lidam com um número limitado de entradas de cada vez, resumindo as informações de entrada, para depois as transmitir para níveis mais elevados de raciocínio. Liu et al. (2000) consideram o GMDH como uma classe especial das ANN's, que pode ser utilizado efetivamente como um previsor para estimar *outputs* de sistemas complexos.

Kondo & Ueno, (2006) afirmam que o algoritmo GMDH permite organizar automaticamente as redes neurais multicamadas, onde os parâmetros a definir numa ANN típica, como o número de neurónios em cada camada oculta, o número de camadas ocultas e as variáveis de entrada relevantes, são determinados de forma automática, de maneira a minimizar o critério de erro de previsão definido como Critério de Informação de Akaike

(AIC). As ANN's convencionais, nomeadamente a MLP, não têm esta capacidade de identificação estrutural da arquitetura de rede, devido à não exclusividade dos pesos de conexão, levando a que, a técnica de minimização do AIC não possa ser aplicada para a determinar a melhor arquitetura da rede (Hagiwara et al., 1993). Devido a este procedimento auto organizável, (Schneider & Steiner, 2005) consideram-se os algoritmos da rede GMDH mais poderosos quando comparados com modelos tradicionais estatísticos.

2.1.2.2.1 O algoritmo

Nas redes GMDH são desenvolvidos, entre outros, o algoritmo do mesmo nome, algoritmo GMDH (Cavalheiro et al., 2011). São técnicas de modelação que aprendem as relações entre as variáveis, que na ótica de séries temporais, corresponde à aprendizagem da relação entre os desfasamentos. Depois de aprender as relações, o algoritmo GMDH seleciona de forma automática o percurso a seguir (Dag & Yozgatligil, 2016a).

Se uma série tiver como variáveis de *input* $x_1, x_2, \dots, x_m$ (i.e., a série temporal desfasada a ser regredida) e *output* y. A relação pode ser descrita como:

$$y = f(x_1, x_2, \dots, x_m) \tag{7}$$

onde f representa o polinómio desenvolvido por Ivakhnenko em 1966 e é dado por:

$$y = a + \sum_{i=1}^{m} b_i \cdot x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} \cdot x_i \cdot x_j \cdot x_k + \dots \tag{8}$$

onde m é o número de variáveis, $a, b, c, d, \dots$ são os coeficientes das variáveis no polinómio, ou seja, os pesos (Dag & Yozgatligil, 2016a). Em geral, os termos são usados no cálculo de termos quadrados. O exemplo da fórmula de cálculo dos termos quadráticos, para duas variáveis é obtido através da equação (9):

$$y = a + \sum_{i=1}^{m} b_i \cdot x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} \cdot x_i \cdot x_j \tag{9}$$

onde m = 2, o que requer que a especificação estime 6 coeficientes em cada modelo. O algoritmo GMDH considera todas as combinações de pares p de séries temporais desfasadas. Onde cada combinação é inserida em cada neurónio, e através das duas entradas, é obtida a saída desejada. O algoritmo GMDH é um sistema de camadas em que existem neurónios e o seu número em cada camada é definido pelo número de *inputs*

como salienta Valença (2005, p. 26, apud Cavalheiro et al., 2011), onde o número de neurónios é determinado pela combinação, $h = \boldsymbol{C}_2^m$, ou seja, $h = m(m-1)/2$.

A arquitetura do algoritmo GMDH é ilustrada na figura 4, possuindo este exemplo retirado de Dag & Yozgatligil (2016), quatro variáveis de *input* e três camadas. Como o número de *inputs* é quatro (m=4), o número de neurónios calculado para a primeira camada é seis. Os coeficientes da equação (9) são estimados em cada neurónio, e usando esses coeficientes estimados juntamente com os *inputs* em cada neurónio, é prevista a saída desejável. De acordo com um critério externo escolhido, e.g. erro quadrático médio (MSE), p neurónios são selecionados e os neurónios $h - p$ são eliminados da rede. Na figura 4, quatro neurónios são selecionados enquanto dois são eliminados da rede. As saídas obtidas dos neurónios selecionados tornam-se as entradas da próxima camada. Esse processo, representado na figura 5, continua até à última camada, onde apenas um neurónio é selecionado, correspondendo a saída deste, ao valor previsto para a série temporal em causa (Do & Yen, 2019).

3. Principais metodologias e tecnologias

3.1 Principais metodologias adotadas em projetos de *Data Mining*

Sempre que necessário a aplicação de diferentes técnicas de *Data Mining* a uma problemática, deve recorrer-se a uma sequência de processos uniformizada (metodologia), que seja comum, independentemente do problema a considerar.

Existem diversas metodologias que se podem adotar, mas para Azevedo & Santos (2008) as metodologias mais populares são:

- *Sample Explore Modify Model and Assess* (SEMMA), desenvolvido por o *SAS Institute;*
- *Knowledge Discovery and Data Mining* (KDD), desenvolvido por Fayyad et al. (1996);
- *Cross Industry Standard Process for Data Mining* (CRISP-DM), desenvolvido por um consorcio composto por *DaimlerChryrler, SPSS and NCR.*

Dentro deste conjunto, a metodologia CRISP-DM destaca-se como a mais utilizada, possuindo a vantagem de ter presente o objetivo do negócio e não só aspetos técnicos (Azevedo & Santos, 2008). A CRISP-DM ajusta-se a qualquer problema,

independentemente da indústria e assenta sobre seis fases distintas, que podem ser vistas como um ciclo, sendo assim possível recuar e avançar entre os diferentes estádios, como se pode observar na figura 6, sendo estes, a compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e execução.

No presente TFM adotar-se-á a metodologia CRISP-DM e posteriormente, analisar-se-á cada estádio desta. Na etapa inicial, compreensão do negócio, procurar-se-á abordar o projeto do ponto de vista da empresa, identificando os fatores que possam ter influência no seu decorrer, as suas necessidades, e os objetivos a alcançar com a implementação. A segunda etapa, compreensão dos dados, consistirá na recolha de dados e na sua exploração, verificando-se ainda, a qualidade dos mesmos. Em seguida no terceiro estádio, preparação dos dados, serão aplicados todos os processos de tratamentos de dados, para futuramente serem usados nas etapas de construção dos modelos. Algumas das principais tarefas realizadas nesta etapa são: limpeza e transformação dos dados, seleção de atributos, criação de novas variáveis, entre outras. Na fase de modelação, quarta etapa, são aplicados os modelos escolhidos, aos dados anteriormente selecionados.

No caso deste TFM serão modelos anteriormente descritos: ARIMA, ARIMAX, VAR, redes neuronais do tipo MLP, e GMHD. Depois da modelação, e já com os modelos estimados, segue-se a quinta etapa, avaliação, onde tal como o nome indica, é o momento de avaliar os modelos, fazendo uma análise crítica dos resultados, através de critérios definidos à priori. No caso dos resultados obtidos não forem satisfatórios, deve-se rever todo o processo até então efetuado, procurando determinar a existência de algum fator ou tarefa importante que não tenha sido considerado. Por fim, na etapa de execução, será planeada a melhor forma de utilização dos modelos, onde o conhecimento adquirido deverá ser organizado e apresentado de forma a que o cliente o consiga utilizar, devendo ainda serem criados mecanismos de monitorização e manutenção dos modelos construídos.

3.2 Principais Tecnologias Utilizadas

O R (R Core Team, 2019) é uma linguagem de programação voltada para manipulação, análise e visualização de dados. Dentro das várias vantagens desta aplicação, destacam-se algumas como a gratuitidade; a existência de uma interface gráfica (GUI) de simples acesso e "*user friendly*", o RStudio (RStudio Team, 2020), também de acesso gratuito; dispõe de um alargado reportório de *packages*, contendo variadas funções, como a capacidade de importar e exportar dados de outras aplicações com relativa facilidade, entre outras.

De facto, este *software* apresenta vários *packages* previamente implementados, desenvolvidos por diversos autores, que não são mais que funções que servirão de auxílio na construção do modelo. Dentro deste conjunto, destaca-se o uso neste TFM de *packages* como o *xlsx* (Dragulescu & Arendt, 2020) e o *readxl* (Wickham et al., 2019), que permitem importar e exportar ficheiros do Microsoft Excel para o R. Existem ainda *packages* como *sweep, tidyquant, timetk* (Dancho & Vaughan, 2020a, 2020b, 2020c) normalmente utilizados no auxílio e interpretação de séries temporais. Numa perspetiva mais direcionada para a aplicação de modelos, a *package forecast* (Hyndman et al., 2020) destaca-se por ser uma ferramenta útil para a aplicação de modelos ARIMA e ARIMAX, sendo ainda utilizada a *package TSA* (Chan & Ripley, 2020) uma vez que possui funções especificas para modelos ARIMAX. No caso do modelo VAR utilizam-se os *packages MTS* e *vars* (Tsay & Wood, 2018; Pfaff & Stigler, 2018) que se destacam pela sua eficácia; enquanto que nas redes neuronais o *nnfor* (Kourentzes, 2019) é um *package* especialmente desenvolvido para redes do tipo MLP, capaz de verificar a estacionaridade e sazonalidade de uma série temporal, e aplicar as transformações necessárias de forma automática, tendo por base o *package neuralnet* (Fritsch et al., 2019), um dos mais famosos em termos de treino de ANN's; por fim para o modelo GMDH o *package gmdh* (Dag & Yozgatligil, 2016b) permite o ajuste deste tipo de rede, a séries temporais univariadas através de um comando simplificado, auxiliando-se em *packages* como *RSNNS* (Bergmeir et al., 2019) e *geospt* (Melo et al., 2015), sendo ainda utilizado o *package GMDHreg* (Tilve, 2020) que permite a introdução de variáveis independentes, no treino de redes do tipo GMDH.

3.3 Avaliação dos Modelos de Previsão

Para comparar a qualidade das previsões dos modelos, a fim de escolher o modelo com o melhor desempenho, é necessário avaliar individualmente cada um dos modelos estimados. Não sendo uma tarefa simples, deve ser realizada tendo em consideração a avaliação de vários métodos, como, por exemplo os erros de previsão.

O primeiro passo, em problemas de previsão, geralmente é fixar um instante de tempo t_1, que permite dividir o período de observação da série em dois subperíodos (Nicolau, 2012, p.210): dados de treino (*in-sample estimation period*) – com o qual os modelos vão aprender, e ajustar-se na fase de modelação, geralmente de 1 a t_1; dados de teste (*out-of sample forecast*) – que têm por finalidade averiguar o erro associado ao modelo aprendido. Os dois conjuntos têm necessariamente de ser disjuntos, caso contrário, o modelo produziria previsões demasiado otimistas, uma vez que estaria já ajustando os dados de teste.

Neste TFM, apenas foram considerados os dados de teste para a obtenção dos erros de previsão, uma vez que o interesse do TFM reside na verificação do desempenho dos modelos na previsão de dados futuros e não na forma como estes se ajustaram aos dados de treino. Assim, e uma vez que o objetivo passa pela criação de um modelo de previsão a 4 passos, deve-se recorrer à chamada previsão recursiva (*recursive forecasting*), que consiste na alteração do t_1, acrescentando mais observações no período da estimação (dados de treino), a fim de obter várias iterações onde se realizará sempre uma previsão a 4 passos permitindo obter uma amostra de valores observados e valores previstos (Nicolau, 2012, p.217).

Considerando Y_t como o valor observado no momento t, $\hat{Y}_t$ o valor previsto para o mesmo instante, e s o número de previsões *out-of-sample*, pode definir-se os critérios de erro de previsão mais utilizados através das seguintes equações:

Raiz do Erro Quadrático Médio (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(Y_t - \hat{Y}_t)^2} \tag{10}$$

Erro Percentual Absoluto Médio (MAPE):

$$MAPE = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100 \quad (11)$$

É de notar que o RMSE é obtido numa escala diferente, penalizando fortemente os erros maiores, assim sendo, mesmo que a grande maioria das previsões sejam boas, o RMSE pode ser alto desde que exista uma previsão má ou muito má. O mesmo não acontece no MAPE que tem uma abordagem menos severa e se a grande maioria das previsões está correta então o MAPE irá ser relativamente baixo (Nicolau, 2012, p. 211).

Foi ainda proposto pela empresa a criação de uma medida de erro que se ajustasse às suas necessidades. O pedido consistia na criação de um método que prioriza as mudanças da tendência, ou seja, se a variação da previsão corresponde à verdadeira variação. Este deve ainda ter em consideração as margens negociais, definidas pela empresa, que neste TFM será representada por letras (a, b, f) devido a questões de confidencialidade. Todo o método foi criado em Microsoft Excel devido à simplicidade de utilização e à interface simples.

Inicialmente, é determinado através do valor o valor fixo f, quais as margens negociais a aplicar, ou seja, se os valores se encontrarem acima de f, aplicam-se as margens a, caso contrário aplicam-se as margens b. Após definida a margem a aplicar é definida se a "tendência" é crescente, decrescente ou constante, através da variação de Y_t, ΔY_t, onde $\Delta Y_t = Y_t - Y_{t-1}$. O processo é aplicado aos valores observados Y_t, e repetido para os valores previstos $\hat{Y}_t$, obtendo assim respetivamente V_t^o e V_t^p. Todo processo é apresentado no fluxograma da figura 7, aplicado aos valores observados.

O próximo passo consiste em verificar se $V_t^o = V_t^p$, transformando numa variável binária, onde 1 representa os acertos e 0 os erros. Após o passo anterior basta calcular a percentagem de acertos ou de erros, no caso deste TFM foi calculada a percentagem de acertos, nomeando esta percentagem como "acertos".

4.Aplicação da metodologia CRISP-DM

4.1 Compreensão do Negócio

O problema apresentado neste TFM diz respeito a uma empresa líder no setor do azeite. A compreensão do negócio, é fundamental para obter um modelo que reflita as suas necessidades. Desta forma, inicialmente foram realizadas reuniões para perceber o funcionamento da empresa, as suas necessidades, e o comportamento das oliveiras. Estas reuniões decorreram com um especialista no setor agrícola, que passou algumas considerações sobre as oliveiras e a produção de azeite, como os fatores climatéricos, a alternância[13] das oliveiras, entre outros fatores que influenciam a produção dos olivais. No entanto, sendo o principal foco as previsões dos preços do azeite, a empresa apresentou ainda as variáveis usadas até então nas suas previsões, e os modelos que aplicavam para o efeito.

As variáveis em estudo EVI, VIR, LAM, são os três tipos de azeite de maior interesse para este estudo. Os tipos de azeites presentes encontram-se todos na categoria de azeites virgens, que são aqueles obtidos unicamente do fruto da oliveira, através de processos físicos ou mecânicos, em condições que não alteram o produto, e diferenciam-se através do nível de acidez, sendo o azeite extra virgem o que possui um menor grau de acidez, e o azeite lampante aquele com um maior grau. Assim como derivam da mesma categoria, e como poderá ser constatado nos tópicos seguintes, não existe grande diferença entre o comportamento dos preços dos três tipos.

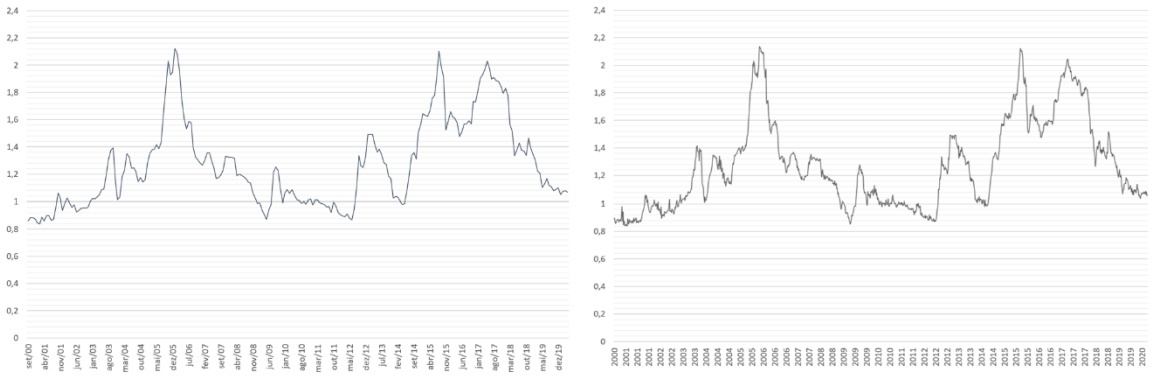
4.2 Compreensão dos Dados

Para a realização deste projeto, a empresa forneceu total acesso aos dados disponíveis e utilizados até ao momento, disponibilizando ainda as fontes de acesso relevantes para os dados meteorológicos. Após uma análise inicial, constatou-se a existência de dados agrupados mensal e semanalmente desde 2000, para os três tipos de azeite a analisar. Destacou-se ainda a falta de dados anteriores a 2014 para as restantes variáveis, tendo sido o acesso fornecido através de resumos anuais, e transcritos para um ficheiro Excel já existente, completando assim, todas as séries temporais desde 2000; foi ainda adicionado ao ficheiro existente os dados relativos ao clima.

[13] "A alternância na produção, designada usualmente por safra e contrassafra, é caracterizada por após um ano de boa produção, a safra, se seguir um ano com pouca produção, a contrassafra" (Pereira, 2017).

4.2.1 Análise Exploratória

Por questões de confidencialidade, no que se segue, os dados foram alterados, mantendo refletido o comportamento das séries temporais. Devido à forte semelhança comportamental das variáveis em estudo, irá ser dado destaque apenas aos gráficos e figuras da variável EVI, que serão apresentados através das figuras A, sendo que os restantes gráficos e figuras, relacionados às variáveis VIR e LAM serão apresentados a par nos anexos como figuras B. Neste estudo, são apresentadas as séries em dados mensais, com observações de setembro de 2000 a abril de 2020 (236 observações), sendo as mesmas variáveis apresentadas em dados semanais, com observações desde a semana 34 do ano de 2000 até à semana 16 de 2020 (1023 observações). Na figura 8-A é representada a evolução gráfica dos preços do azeite extra virgem, considerando os períodos temporais em análise.



*Os valores dos preços foram alterados, mas refletem o comportamento das séries

Figura 8-A: Preço do Azeite Extra Virgem por tonelada desde 2000, mensal e semanalmente, da esquerda para a direita respetivamente;

Fonte: Elaboração própria

As séries mensais e semanais apresentam também um comportamento semelhante, o que era expectável, já que representam o preço do mesmo produto, diferenciadas pela oscilação entre semanas, mais evidentes que as oscilações mensais.

A série é caracterizada pela crescente tendência inicial, que dura até ao final de 2005, podendo ser justificada pela incapacidade de resposta ao aumento do consumo, provocado pela OMS (Organização Mundial da Saúde) após declarar os benefícios para a saúde do azeite (Reis, 2014). Em 2005 e 2006 verificam-se acréscimos significativos das áreas plantadas, emergindo os olivais intensivos conduzidos por tecnologia moderna,

permitindo um aumento da produção, retornando os preços para valores mais reduzidos, tendo estes sofrido posteriormente uma forte quebra com a crise de 2008, e estendendo-se até 2009, seguindo-se um ligeiro aumento, no entanto, continuando com uma tendência de baixa de preços durante mais 3 anos até ao final de 2012, onde baixa de produção levou a um aumento dos preços (Butler, 2012), no entanto, a produção de 2013/2014 foi elevada levando novamente a uma queda de preços (Butler, 2013). De seguida motivado pela especulação causada pela antecipação de um período de seca verificou-se de novo uma tendência crescente até à colheita de 2015/2016 (Ridley, 2015). As variações subsequentes foram motivadas pela produção, onde fracas produções provocaram um aumento dos preços, enquanto maiores produções levam a um efeito de baixa de preços que decorre até a atualidade (Hernandez, 2017, 2019; Vasilopoulos, 2018).

4.2.2 Decomposição por Sazonalidade e Tendência usando Loess

Através da figura 8-A não é evidente a presença de sazonalidade, para tal recorreu-se à decomposição o *Seasonal and Trend decomposition using Loess* (STL) (Cleveland et al., 1990), no entanto, este tipo de decomposição apenas funciona para modelos aditivos, e uma vez que existe suspeita de presença de sazonalidade, aplicou-se o logaritmo natural ao modelo de forma a passar de um modelo multiplicativo para um modelo aditivo. Os resultados obtidos estão presentes na figura 9-A.

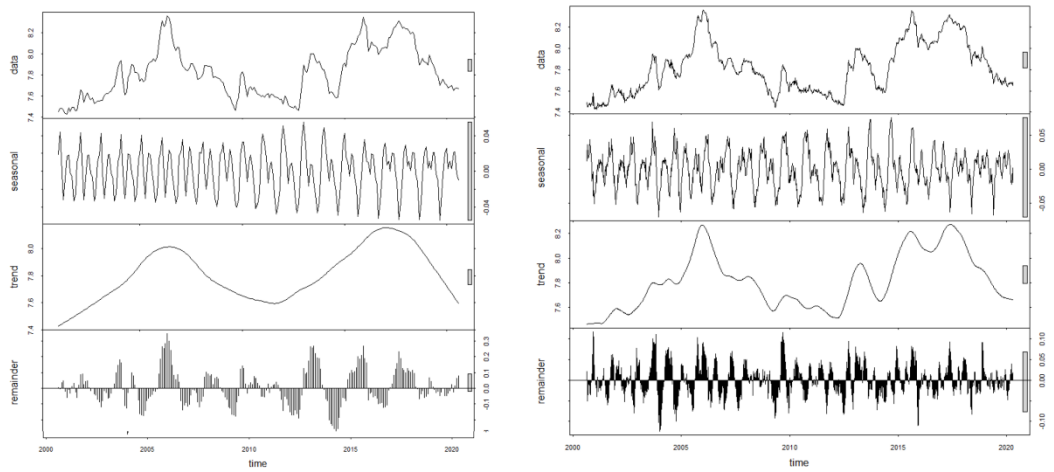


Figura 9-A: Decomposição das séries temporais nas três componentes principais: sazonalidade, tendência e parte aleatória, mensal e semanalmente, da esquerda para a direita respetivamente; Fonte: Elaboração própria

Na figura 9-A (e restantes azeites 9-B) é prescritível as variações da tendência descritas anteriormente em 4.2.1, comprovando a não estacionaridade da série, confirmada ainda pelos gráficos da figura 10, onde mostra que as ACF (*Autocorrelation*

Function) atenuam lentamente para as determinadas séries de preços do azeite virgem extra. No entanto, a principal vantagem desta decomposição foi a evidência da sazonalidade anual, que está presente em ambos os períodos temporais em estudo.

4.3 Preparação dos Dados

Nesta fase dividem-se os dados em dois subconjuntos, como explicado no capítulo 3.3, tendo esta divisão sido realizada de acordo com o critério definido por da Silva et al. (2017), explicado no capítulo (2.1.2), para que assim seja logo propenso para o uso das redes neuronais, tendo sido definido 80% para os dados de aprendizagem/dados de treino, sendo os restantes 20% os dados de teste, que correspondem a aproximadamente aos últimos 4 anos (em termos mensais e semanais), sendo esta divisão conjugada com o método de *recursive forecasting*, devido à previsão fixa em quatro passos. Por outro lado, e como neste TFM se pretende fazer o estudo de modelos multivariados, procurou-se verificar a importância de algumas variáveis que possam ter influência nos preços do azeite.

4.3.1 Variáveis Explicativas

A conselho dos especialistas do negócio, e agrícolas, foram testadas várias variáveis explicativas, apresentadas de seguida. Apesar das séries de preços do azeite serem idênticas em ambos os períodos temporais, as variáveis explicativas existentes, são diferentes para os modelos mensais e semanais; no entanto, as variáveis relativas ao clima são comuns em ambos os casos. Também nas séries que se seguem, os dados foram alterados, mantendo refletido o seu comportamento, devido a questões de confidencialidade.

4.3.1.1 Variáveis mensais

Em termos mensais as variáveis exclusivas analisadas foram escolhidas através da recomendação da empresa em junção ao estudo de Rivera et al. (2011) que utiliza as *salidas* (figura 11-a), importações (figura 11-b) e as existências (figura 11-d) para prever o preço do azeite no curto e médio prazo, sendo ainda analisada a produção (figura 11-c) devido à constante presença como justificação das flutuações do preço em diversos artigos. Na análise gráfica da figura 11, é possível constatar que a variável *salidas*, que é a junção do consumo, perdas e exportações, figura 11 (a), tem uma tendência crescente, com oscilações que vão em sentido inverso às oscilações dos preços, ou seja, o aumento do preço leva à diminuição das *salidas* e vice-versa. Pelo contrário, as importações

realizadas têm oscilações que vão de encontro com o aumento dos preços, figura 11 (b), no entanto, no último ano tiveram um comportamento contrário, como evidenciado na figura, têm uma tendência crescente, sendo que eram inexistentes no início da análise, nos finais de 2000.

A variável produção tem uma forte componente sazonal, que pode ser facilmente observada na figura 11 (c), esta deriva do facto de a colheita e produção de azeite apenas ocorrer nos meses de novembro a maio, atingindo o seu pico nos mês de dezembro e janeiro, sendo que nos restantes meses não existe qualquer produção. Nesta variável é ainda possível observar o fenómeno de safra e contrassafra relatado anteriormente, sendo também visível a atenuação deste após a industrialização dos cultivos de 2005/2006. A variável existências diz respeito à quantidade de azeite existente e armazenado pelos produtores, sendo esta uma variável com elevada sazonalidade, que também depende da época de produção, onde pela análise gráfica da figura 11 (d) não parece ter impacto no preço do azeite.

4.3.1.2 Variáveis semanais

No que diz respeito aos dados semanais, estes também possuem variáveis exclusivas, não disponíveis em termos mensais. Assim para os modelos multivariados mensais, as variáveis analisadas serão as toneladas compradas nessa semana de azeite, extra virgem (TEVI), virgem (TVIR), Lampante (TLAM), figuras 12 (a), (b), (c), respetivamente; e a média das quantidades transacionadas por operação, em toneladas (MTOPE), figura 12 (d). As séries em causa ostentam uma forte oscilação, aparentando terem uma relação negativa com os preços, ou seja, um aumento dos preços leva a uma diminuição das toneladas adquiridas e do número de transações.

4.3.1.3 Variáveis Comuns

No presente TFM foi ainda investigada a questão relativa às variáveis climatéricas, seguindo assim a sugestão de Pérez-Godoy et al. (2010) de incluir variáveis meteorológicas, sendo as variáveis selecionadas a velocidade do vento, humidade, temperatura e precipitação, oriundas da sugestão dos especialistas na área. Estas variáveis serão uma média das várias regiões de interesse da empresa, que se agruparam em dados mensais e semanais. Na figura 13, estão presentes as variáveis em causa, obtidas do IFAPA (*Investigación y Formación Agraria y Pesquera*), sendo que, as séries ostentam uma elevada sazonalidade, algo expectável, dado se tratar de séries meteorológicas, onde

apenas a precipitação média figura 13 (d), aparenta ter uma relação com os preços do azeite, sendo que a diminuição da precipitação leva a um aumento do preço e vice-versa.

Também às variáveis explicativas foram realizados testes de estacionaridade, e ainda nesta etapa, realizados todos os processos de diferenciação às variáveis que necessitavam para alcançarem a estacionaridade, sendo que apenas a produção e a temperatura média mensal se encontravam estacionarias, não necessitando de nenhum processo de transformação.

Por fim, será criada uma variável *dummy* de forma a representar a alternância na produção, permitindo assim analisar o conceito de safra e contrassafra que afeta as oliveiras. O valor 1 irá simbolizar os anos de safra (geralmente anos par), enquanto que 0 representa os anos de contrassafra (anos ímpares).

4.4 Modelação

Nesta etapa foram aplicados os modelos propostos sobre o conjunto de dados de treino de forma a obter as melhores especificações, tendo em consideração as respetivas metodologias e as suas diferentes restrições, para poder prosseguir com a comparação.

Começou-se por aplicar os modelos, aos preços do azeite extra virgem, sendo posteriormente realizado o mesmo processo para os restantes azeites. Numa fase inicial, esta aplicação começou pelos modelos univariados, ou seja, sem a inclusão de variáveis independentes.

4.4.1 Modelos Univariados

- ARIMA (seção 2.1.1.1): A modelação do ARIMA foi realizada através da estimativa automática dos parâmetros $(p, d, q)(P, D, Q)_s$ pela função *auto.arima* do *package forecast*, obtendo assim as estimativas a considerar inicialmente. Seguida da análise dos resíduos do modelo fornecido, onde são ajustados os parâmetros $(p, d, q)(P, D, Q)$, de forma a cumprir com todos os testes diagnósticos estipulados.

- ANN-MLP (seção 2.1.2.1): Na construção da rede neuronal do tipo MLP, foi utilizada a função *mlp* do *package nnfor*. Este *package* identifica de forma automática o número de neurónios/nós da camada de entrada, que no caso de previsões uni variadas, são os *lags* da série temporal em causa. Quanto ao número de camadas ocultas foram testadas redes com 2 camadas ocultas e com apenas 1, uma vez que não existe total concordância sobre o tema. O número de neurónios em cada camada

é determinado de acordo com a sugestão de Heaton (2008), ou seja 2/3 do tamanho da camada de entrada somado ao tamanho da camada de saída.

- GMDH (seção 2.1.2.2): Para as redes do tipo GMDH foi utilizada a função *fcast* do *package GMDH*, que permite fazer previsões de curto prazo para uma série temporal uni variada, tendo sido apenas preciso definir o número de observações a serem previstas, que no caso são 4.

4.4.2 Análise Multivariada

No que diz respeito aos modelos multivariados, o processo começou pela análise de uma matriz de correlação entre as diferentes variáveis analisadas anteriormente, observada na figura 14. É possível, observar uma forte correlação entre algumas variáveis, nomeadamente meteorológicas, nos diversos tipos de azeite, o que pode indicar à partida a irrelevância de algumas destas. Em cada modelo, foram ainda sendo sucessivamente adicionadas as diversas variáveis explicativas, a fim de verificar se os resultados melhoravam ou não com a sua adição.

- VAR (seção 2.1.1.3): O modelo VAR foi estimado recorrendo aos *packages vars* e MTS. Começou-se por encontrar a ordem p através do método GTS, complementado com o método dos critérios de informação AIC, BIC e HQ, obtendo uma ordem de $p = 1$. Para a seleção das variáveis no caso mensal foram executadas todas as combinações possíveis entre as variáveis (2045 combinações), avaliadas pelo critério de informação AIC, obtendo como melhor combinação as importações mais a variável safra. O mesmo foi realizado para o caso semanal onde se obteve a temperatura média e a variável safra como a combinação com o menor valor de AIC.

- ARIMAX (seção 2.1.1.2): Para o modelo ARIMAX são estimadas as ordens $(\mathrm{p},\mathrm{d},\mathrm{q})(\mathrm{P},\mathrm{D},\mathrm{Q})_s$ através do mesmo processo e funções utilizadas no modelo ARIMA, no entanto os ajustes das ordens $(\mathrm{p},\mathrm{d},\mathrm{q})(\mathrm{P},\mathrm{D},\mathrm{Q})$ foi realizado com recurso à função ARIMAX do *package* TSA, que está otimizado para modelos do tipo ARIMAX. Neste modelo foram pré-selecionadas variáveis e respetivos desfasamentos, através dos modelos aplicados pela empresa, tendo sido estudado o impacto da introdução/remoção das variáveis em análise e de novos desfasamentos. Ao contrário do modelo VAR não foram estudadas todas as combinações possíveis,

mas foram estudadas combinações tendo em consideração as combinações do modelo VAR, sendo selecionadas as que possuem um menor valor de AIC.

- ANN-MLP (seção 2.1.2.1): Utilizando a função *mlp* do *package nnfor* é ainda possível adicionar variáveis independentes como nós da camada de entrada, ajudando assim na construção da rede. Desta forma foram selecionados o conjunto de variáveis com melhor resultado, uma vez que a criação de redes neuronais é um processo computacionalmente dispendioso, principalmente em amostras elevadas, como o caso das séries semanais.

- GMDH (seção 2.1.2.2): Para a construção da ANN do tipo GMDH recorrendo a variáveis independentes, foi necessário encontrar um *package* diferente do anteriormente usado, que permitisse a inclusão de novas variáveis. Assim a função *gmdh.combi*, do *package GMDHreg* permite a inclusão de até 4 variáveis independentes sem se tornar muito dispendioso e demorado em termos computacionais, tendo de ser definido o algoritmo a ser utilizado, no caso o algoritmo de Ivakhnenko. Assim novamente, os conjuntos de variáveis testados são reduzidos devido à capacidade computacional, sendo no caso mensal, apenas o que obteve melhores resultados, e no caso semanal devido à falta de consenso os dois conjuntos que obtiveram melhores resultados {EVIt-5; Mtopet-6;Tmedt-6;Safrat} e {VIRt-5;LAMt-5;Mtopet-6; Safrat}.

4.5 Avaliação

A avaliação dos algoritmos foi realizada através da abordagem *recursive forecasting* (seção 3.3), em que para o processo de comparação, predefiniu-se uma *seed* de modo a garantir que os resultados obtidos não variem de cada vez que se corresse o algoritmo no mesmo conjunto de teste. Assim foram avaliadas as previsões a 4 passos para os vários modelos e conjuntos de variáveis, e avaliados através dos métodos RMSE, MAPE, e do método criado para a empresa Acertos, anteriormente especificados (secção 3.3). Os resultados obtidos encontram-se distribuídos por modelos, na tabela 2 à tabela 6.

Para que fosse possível avaliar todos os modelos de forma igualitária, através da abordagem *recursive forecasting*, foi desenvolvido em *R* uma janela de comandos com

recurso à função *for*, que permite o retreino dos modelos e a previsão a 4 passos em cada momento.

Da análise da tabela 4 podemos retirar que no modelo VAR, o conjunto que de variáveis com um menor AIC em termos mensais, corresponde ao conjunto de variáveis com melhores resultados, importações e safra. No entanto, o mesmo não acontece nos dados semanais, onde o conjunto com melhores resultados é aquele que também apresenta melhores resultados no modelo ARIMAX, o que contém os preços dos restantes azeites, as toneladas compradas nessa semana do azeite em causa, a temperatura média, a média das quantidades transacionadas por operação, e a *dummy* safra.

No caso das redes neuronais, do tipo MLP os casos univariados obtiveram na generalidade melhores resultados, à exceção dos casos semanais do azeite virgem e lampante, onde as variáveis safra e média das quantidades transacionadas por operação, conseguiram obter um RMSE menor, sendo que o MAPE continuou a ser menor para os casos univariados. O mesmo acontece para a rede do tipo GMDH onde o caso univariado se destacou em quase todos os casos, à exceção do azeite lampante semanal, que a mesma combinação de que o modelo MLP teve melhor resultados na medida RMSE, continuando o MAPE sendo preferível para o caso univariado. Esta preferência pelo univariado pode dever-se a diversos fatores, sendo um deles o estudo das variáveis independentes nestes modelos não ser tão aprofundado, devido às limitações computacionais, e por isso os conjuntos em causa, apesar de serem os mais relevantes nos restantes modelos, puderem não ser os mais adequados para as redes em causa. Outro fator que pode ter influência é o facto de os *packages* utilizados tanto no caso MLP, o *packages nnfor*, como no caso do GMDH, o *packages GMDH*, estarem otimizados para a construção e previsão de séries univariadas, sendo mesmo impossível utilizar variáveis independentes no *packages GMDH*, podendo ter assim tido um certo impacto no momento da previsão quando comparado com as redes multivariadas.

De modo geral é fácil observar através da tabela 7, que é o modelo VAR que obtém melhores resultados em qualquer um dos casos em estudo. Tal pode dever-se ao estudo de todas as combinações de variáveis, que permitiu encontrar o melhor conjunto, mas também à elevada essa capacidade de previsão, que o modelo VAR possui.

Tabela 7 – Comparação do desempenho da previsão dos modelos ARIMA, ARIMAX, VAR, MLP e GMDH.

Série temporal	Periodicidade	Modelo	RMSE
Extra Virgem	Mensal	ARIMA (1,1,1)x(2,0,2)12	332,70
		ARIMAX (2,1,1)x(0,0,1)12 (ΔEVIt-5;ΔImpt-6;Safrat)	315,22
		VAR(1) ΔImpt;Safrat	**264,69**
		MLP (10) Univariada	386,01
		GMDH Univariada	346,74
	Semanal	ARIMA (2,1,1)x(1,0,0)52	141,69
		ARIMAX (1,1,1)x(1,0,1)52 ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTEVIt-6;ΔMtopet-6;ΔTmedt-6;Safrat	140,71
		VAR(1) ΔVIRt;ΔLAMt;ΔTEVIt;ΔMtopet;ΔTmedt;Safrat	**116,16**
		MLP (10) Univariada	141,97
		GMDH Univariada	139,76
Virgem	Mensal	ARIMA (1,1,1)x(2,0,2)12	364,16
		ARIMAX (1,1,1)x(1,0,1)12 ΔVIRt-5;ΔImpt-6;Safrat	355,50
		VAR(1) ΔImpt;Safrat	**306,75**
		MLP (10) Univariada	431,36
		GMHD Univariada	395,60
	Semanal	ARIMA (2,1,2)x(1,0,0)52	139,81
		ARIMAX (1,1,1)x(0,0,1)52 ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTVIRt-6;ΔMtopet-6;ΔTmedt-6;Safrat	140,23
		VAR(1) ΔEVIt;ΔLAMt;ΔTVIRt;ΔMtopet;ΔTmedt;Safrat	**114,78**
		MLP (12;12) ΔMtopet-6;Safrat	140,43
		GMHD Univariada	139,16
Lampante	Mensal	ARIMA (4,1,4)x(1,1,2)12	374,61
		ARIMAX (2,1,1)x(1,0,1)12 ΔLAMt-5;ΔImpt-6;Safrat	374,76
		VAR(1) ΔImpt;Safrat	**312,52**
		MLP (10;10) Univariada	385,91
		GMHD Univariada	405,12
	Semanal	ARIMA (3,1,3)x(1,0,1)52	145,18
		ARIMAX (2,1,1)x(0,0,1)52 ΔLAMt-5;ΔTmedt-6;Safrat	145,63
		VAR(1) ΔEVIt;ΔVIRt;ΔTLAMt;ΔMtopet;ΔTmedt;Safrat	**122,52**
		MLP (12;12) ΔMtopet-6;Safrat	140,19
		GMHD ΔLAMt-5;ΔMtopet-6;Safrat	166,12

Fonte: Elaboração própria

"Retirando" o modelo VAR, e analisando os restantes modelos, podemos verificar que é o modelo ARIMAX quem obtém melhores resultados em termos mensais para o azeite extra virgem e virgem, e o ARIMA para o azeite lampante. Tal pode derivar do facto de as redes neuronais necessitarem de um maior conjunto de dados para realizar um treino de forma eficaz e obter resultados mais satisfatórios, como demostrado por Raudys & Jain (1991). Resultados estes que se verificam nos modelos semanais onde o GMDH se destaca para os azeites extra virgem e virgem, e a rede do tipo MLP obteve melhores resultados para o azeite lampante.

Numa avaliação através do método de erro da variação (“acertos”) proposto pela empresa, podemos observar, que tal método não se demonstra suficientemente eficaz. O método em causa obteve melhores resultados, maior taxa de acertos, nos modelos que apresentaram pior desempenho nos métodos RMSE e MAPE. Tal pode estar relacionado ao facto de a margem de lucro negocial ser um efeito atenuador que permite que modelos com sobre ajustamento (*overfitting*), obtenham melhor resultados, pois a sua falta de generalização e o seu comportamento ser uma “cópia” do comportamento passado aumenta o número de “constantes” que levam a acertos, face aos restantes modelos que têm efetivamente melhores resultados nas restantes medidas de erros de previsão.

Assim esta medida não pode ser usada como uma medida comparativa, podendo ser adotada como um indicador de eficácia do modelo aquando este aponta uma mudança na sua previsão.

4.6 Execução

Iniciando a última etapa, a execução, e após concretizada a escolha do modelo “vencedor”, o modelo VAR, procedeu-se à automatização do processo, para que o custo de aprendizagem do utilizador final fosse o menor possível. Foi ainda criada uma base de dados meteorológica em Excel, com o processo de seleção de dados e respetivos links de acesso, para que o processo de atualização das variáveis fosse o mais simples possível.

Com base nessa necessidade de criar um ficheiro com baixo custo de aprendizagem de *R*, os modelos foram desenvolvidos em R *Markdown* (Allaire et al., 2020), uma vez que, o formato ajuda o utilizador , recomendando a instalação de todas as *packages* necessárias, apenas necessita de um comando para realizar o procedimento, e ainda permite transformar as análises em documentos, relatórios, apresentações entre outros. O conjunto de comandos necessários para a obtenção das previsões até 4 passos, foi escrito de maneira a que não precise de qualquer edição, apenas a introdução da localização da base de dados.

Foi adicionalmente escrito a lista de comando, também de forma automatizada, para os restantes modelos, no intuito da empresa puder ao longo do tempo analisar o comportamento dos restantes modelos, uma vez em conjuntos de dados maiores as redes obtiveram melhores resultados, e assim não ficar “presa” ao melhor modelo atual.

5. Conclusão

No campo da previsão estatística, a inovação é algo fundamental, pois permite a que cada vez mais os erros sejam menores e melhores previsões possam ser feitas. No entanto, tal como a Gallo procura um equilíbrio entre tradição e inovação, também no campo da previsão se deve ter atenção a essa premissa.

O presente relatório teve como principal objetivo o desenvolvimento de um modelo preditivo, que permitisse uma estimação fiável do preço de compra das mercadorias, baseando-se em preços mensais e semanais. Deste modo, as previsões obtidas deveriam ser o mais realista possível, uma vez que uma previsão acima do esperado poderá levar a uma compra de stocks desnecessária e a um aumento de custos de manutenção, e uma previsão abaixo do esperado poderá levar a um adiamento da compra e a uma perda de lucros para a empresa. Por outro lado, procurou-se incluir diversas variáveis, para analisar o seu impacto nos preços e permitissem explicar as previsões.

Assim sendo, procurou-se comparar dentro de 5 metodologias diferentes qual delas obteria o melhor modelo, sendo as metodologias estudadas ARIMA, ARIMAX, VAR, MLP, GMDH, todas elas aplicadas aos três tipos de azeite e aos 2 períodos temporais em causa.

Começa-se por destacar o impacto da variável safra, que de entre o conjunto de variáveis em análise foi a que possuiu maior impacto de relevância. O mesmo não aconteceu com a variável produção o que leva a deduzir que a variável safra não está a captar os anos de elevada produção, uma vez que o fenómeno inclusive está atenuado pela agricultura industrial, mas sim a captar a especulação que continua presente no mercado de azeite à volta dos anos de safra e de contrassafra.

É ainda importante destacar que o modelo tradicional VAR obteve melhores resultados para todos os conjuntos de dados em análise, destacando-se assim dos modelos de redes neuronais, que geralmente obtêm melhores resultados. No entanto, ressalva-se que o presente relatório vai de encontro com a generalidade da literatura sobre a matéria em causa, que demonstra que em grandes amostras, as redes neuronais obtêm melhores resultados que os modelos tradicionais ARIMA (Zhang et al., 1998). Ainda ressalvar que em quatro dos seis conjuntos de dados em causa, a rede GMDH obteve melhores resultados que a rede do tipo MLP, indo de encontro com a conclusão apresentada Do & Yen (2019).

No entanto, limitações podem ser apresentadas ao estudo, nomeadamente a combinação de variáveis independentes usadas nas redes neuronais ter derivado dos modelos lineares, podendo não ser, neste caso específico, o mais adequado. Outra limitação, surgiu do elevado custo de aprendizagem das ANN, que dificultou o cumprimento dos prazos definidos pela empresa, e levou a uma menor análise em termos de variáveis e outros fatores externos. Por fim, a base de dados em análise diz respeito a um setor de negócio específico, o mercado de azeite, o que compromete a generalização dos resultados. De notar ainda que a pandemia de covid-19, e todos os efeitos económicos e sociais dela provenientes, podem ainda vir a interferir nas séries em estudo no presente TFM alterando assim a eficácia dos modelos propostos.

Como a principal proposta de trabalho futuro, destaca-se a investigação de modelos híbridos, que permite assim a combinação entre a inovação e a tradição. Conceito estudado por Zhang (2003) entre ANN de vários tipos e o modelo ARIMA, que pode ser alargado a mais modelos tradicionais, neste caso, ao modelo VAR. Já em termos dos modelos VAR, sugere-se estudar a possibilidade dos componentes das séries temporais multivariadas em causa serem cointegrados, que nesse caso, o uso do Modelo Vetorial de Correção de Erros (VEC) é o mais recomendado. Propõe-se ainda a utilização de métodos combinados, exemplo VAR-GMDH seguindo o exemplo de Yefimenko (2018), que no estudo em causa permite a junção do tradicional com a inovação.

Bibliografia

Allaire, J.J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. & Iannone, R. (2020). *rmarkdown: Dynamic Documents for R*.

Anggraeni, W., Andri, K.B., Sumaryanto & Mahananto, F. (2017). The Performance of ARIMAX Model and Vector Autoregressive (VAR) Model in Forecasting Strategic Commodity Price in Indonesia. *Procedia Computer Science*. 124. p.pp. 189–196.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European conf. data mining (p.pp. 182–185).

Aziz, S., Dowling, M.M., Hammami, H. & Piepenbrink, A. (2019). Machine Learning in Finance: A Topic Modeling Approach. *SSRN Electronic Journal*.

Baron, R. (1994). Knowledge extraction from neural networks: A survey, in: Report no. 94-17, Laboratoire de l'Informatique du Paralle´lisme, Ecole Normale Supe´rieure de Lyon

Bennett, C., Stewart, R.A. & Lu, J. (2014). Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*. 7 (5). p.pp. 2938–2960.

Bergmeir, C., Benítez, J.M., Zell, A., Mache, N., Mamier, G., Vogt, M., Döring, S., Hübner, R., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Reczko, M., Riedmiller, M., Seemann, M., Ritt, M., DeCoster, J., Biedermann, J., Danz, J., Wehrfritz, C., Kursawe, P. & El-Ama, A. (2019). RSNNS: Neural Networks using the Stuttgart Neural Network Simulator (SNNS). R package version 0.4-12. *https://CRAN.R-project.org/package=RSNNS*.

Bouzada, M. (2012). Aprendendo Decomposição Clássica: Tutorial para um Método de Análise de Séries Temporais. *TAC – Tecnologias de Administração e Contabilidade*. 2 (1). p.pp. 1–18.

Box, G.E.P. & Jenkins, G.M. (1976). *Time series analysis : forecasting and control*. San Francisco: Holden-Day.

Box, G.E.P. & Tiao, G.C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*. 70 (349). p.p. 79.

Butler, J. (2012). *Producer Prices Climb Again in Spain*. [Online]. 26 Dezembro 2012. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/europe/olive-oil-prices-climb-again/31729. [Acedido: 12 Setembro 2020].

Butler, J. (2013). *Spanish Harvest Marked by High Volumes, Low Prices - Olive Oil Times*. [Online]. 30 Dezembro 2013. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/europe/spanish-harvest-marked-high-volumes-low-prices/37747. [Acedido: 12 Setembro 2020].

Caiado, J. (2002). Modelos VAR, Taxas de Juro e Inflação. *in: Literacia e Estatística Actas do X Congresso da Sociedade Portuguesa de Estatística*. p.pp. 215–228.

Camelo, H. do N., Lucio, P.S., Leal Junior, J.B.V., Carvalho, P.C.M. de & Santos, D. von G. dos (2018). Innovative hybrid models for forecasting time series applied in wind generation based on the combination of time series models with artificial neural networks. *Energy*. 151. p.pp. 347–357.

Campbell, J.Y., Lo, A.W. & MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ.

Cavalheiro, E.A., Vieira, K.M. & Ceretta, P.S. (2011). Aplicação De Redes Neurais Polinomais Gmdh Na Previsão Do Índice Ibovespa. *CAP Accounting and Management - B4*. 5 (5). p.pp. 40–47.

Chan, K.-S. & Ripley, B. (2020). TSA: Time Series Analysis. R package version 1.3. *https://CRAN.R-project.org/package=TSA*.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*. 6 (1). p.pp. 3–73.

Coenen, F. (2011). Data mining: Past, present and future. *Knowledge Engineering Review*. 26 (1). p.pp. 25–29.

Cybenko, G. (1989). Continuous Value Neural Networks with Two Hidden Layers Are Sufficien. *Mathematics of Control, Signals, and Systems*. 2. p.pp. 303–314.

Dag, O. & Yozgatligil, C. (2016a). GMDH: An R package for short term forecasting via GMDH-type neural network algorithms. *R Journal*. 8 (1). p.pp. 379–386.

Dag, O. & Yozgatligil, C. (2016b). GMDH: Short Term Forecasting via GMDH-Type Neural Network Algorithms. R package version 1.6. *https://CRAN.R-project.org/package=GMDH*.

Damrongkulkamjorn, P. & Churueang, P. (2005). Monthly energy forecasting using decomposition method with application of seasonal ARIMA. Em: *7th International Power Engineering Conference, IPEC2005*. 2005, pp. 1–229.

Dancho, M. & Vaughan, D. (2020a). sweep: Tidy Tools for Forecasting. R package version 0.2.3. *https://CRAN.R-project.org/package=sweep*.

Dancho, M. & Vaughan, D. (2020b). tidyquant: Tidy Quantitative Financial Analysis. R package version 1.0.1. *https://CRAN.R-project.org/package=tidyquant*.

Dancho, M. & Vaughan, D. (2020c). timetk: A Tool Kit for Working with Time Series in R. R package version 2.2.1. *https://CRAN.R-project.org/package=timetk*.

Do, Q.H. & Yen, T.T.H. (2019). Predicting primary commodity prices in the international market: An application of group method of data handling neural network. *Journal of Management Information and Decision Science*. 22 (4). p.pp. 471–482.

Doganis, P., Alexandridis, A., Patrinos, P. & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*. 75 (2). p.pp. 196–204.

Dragulescu, A. & Arendt, C. (2020). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.6.3. *https://CRAN.R-project.org/package=xlsx*.

Faraway, J.J. (2015). *Linear models with R*. 2ºedition. Florida: Chapman & Hall/CRC.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. 17 (3). p.pp. 37–54.

Fritsch, S., Guenther, F., Wright, M.N., Suling, M. & Mueller, S.M. (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. *https://CRAN.R-project.org/package=neuralnet*.

Gama, J., Carvalho, A.P. de L., Faceli, K., Lorena, A.C. & Oliveira, M. (2017). *Extração de Conhecimento de Dados*. 3ª. Edições Sílabo (ed.). Lisboa.

Hagiwara, K., Toda, N. & Usui, S. (1993). On the problem of applying AIC to determine the structure of a layered feedforward neural network. Em: *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*. 1993, pp. 2263–2266.

Haidar, I., Kulkarni, S. & Pan, H. (2008). Forecasting model for crude oil prices based on artificial neural networks. *ISSNIP 2008 - Proceedings of the 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. p.pp. 103–108.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

Harvey, A. (1990). *The Econometric Analysis of Time Series*. 2ª. MA MIT Press (ed.). Cambridge.

Harvey, R.L. (1994). *Neural network principles*. Prentice-Hall, New York.

Haykin, S. (2008). *Neural Networks and Learning Machines*. Englewood Cliffs. NJ: Prentice-Hall.

Heaton, J. (2008). *Introduction to Neural Networks for Java*. 2.ª Ed. K. Smith & WordsRU.com (eds.). Heaton Research, Inc.

Hecht-Nielsen, R. (1992). Theory of the Backpropagation. Em: *Neural Networks for Perception*. Academic Press, pp. 65–93.

Hernandez, E. (2019). *Higher Production in Spain Leads to Lower Prices*. [Online]. 7 Março 2019. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/higher-production-in-spain-leads-to-lower-prices/67194. [Acedido: 12 Setembro 2020].

Hernandez, E. (2017). *Spanish Olive Oil Prices Increase with Lower Productoin* . [Online]. 3 Novembro 2017. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/spanish-olive-oil-prices-increase-lower-productoin/59630. [Acedido: 12 Setembro 2020].

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y. & Zhou, Z. (2020). forecast: forecasting functions for time series and linear models. R package version 8.12. *http://CRAN.R-project.org/package=forecast*.

Hyndman, R.J. & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*. [Online]. 27 (3). p.p. 22. Available from: http://www.jstatsoft.org/v27/i03/paper.

IOC (2019). International Olive Council. *Newsletter No: 144 DECEMBER 2019*. [Online]. Available from: https://www.internationaloliveoil.org/wp-content/uploads/2019/12/NEWSLETTER_144_ENGLISH.pdf.

Ivakhnenko, A.G. (1970). Heuristic self-organization in problems of engineering cybernetics. *Automatica*. 6 (2). p.pp. 207–219.

Ivakhnenko, A.G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man and Cybernetics*. 1 (4). p.pp. 364–378.

Jha, G.K. & Sinha, K. (2013). Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System. *Agricultural Economics Research Review*. [Online]. 26 (2). p.pp. 229--239. Available from: http://ageconsearch.umn.edu/bitstream/162150/2/8-GK-Jha.pdf.

Kondo, T. (1998). GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem. Em: *Proceedings of the 37th SICE Annual Conference. International Session Papers*. 1998, pp. 1143–1148.

Kondo, T. & Ueno, J. (2006). Revised GMDH-type Neural Network Algorithm with a Feedback Loop Identifying Sigmoid Function Neural Network. *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*. 2 (5). p.pp. 985–996.

Kourentzes, N. (2019). nnfor: Time Series Forecasting with Neural Networks. R package version 0.9.6. *https://CRAN.R-project.org/package=nnfor to link to this page.*

Krolzig, H.-M. (2001). *General-to-Specific Reductions of Vector Autoregressive Processes*.

Liu, H.S., Lee, B.Y. & Tarng, Y.S. (2000). In-process prediction of corner wear in drilling operations. *Journal of Materials Processing Technology*. 101 (1). p.pp. 152–158.

Lu, C.J., Lee, T.S. & Lian, C.M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. *Decision Support Systems*. 54 (1). p.pp. 584–596.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg.

Melo, C., Santacruz, A. & Melo, O. (2015). geospt: Geostatistical Analysis and Design of Optimal Spatial Sampling Networks. R package version 1.0-2. *https://CRAN.R-project.org/package=geospt*.

Møller, M.F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*. 6 (4). p.pp. 525–533.

Nicolau, J. (2012). *Modelação de Séries Temporais Financeiras*. Coimbra: Almedina. Coleção Econômicas II. n. 18.

Pankratz, A. (1983). *Forecasting With Univariate Box- Jenkins Models: concepts and cases.* John Wiley & Sons, New York. New York. USA.

Pereira, C. (2004). A cooperação Universidade-Indústria: o caso da Universidade do Minho. Em: *Cadernos de Geografia*. Coimbra, pp. 283–293.

Pereira, C.Á. (2017). *Caracterização da Fenologia de 5 Cultivares de Oliveiras Tradicionais Portuguesas.* Tese de Mestrado em Agricultura Sustentável. Escola Superior Agrária de Elvas - IPPortalegre. Ciências Agrárias. Elvas.

Pérez-Godoy, M.D., Pérez-Recuerda, P., Frías, M.P., Rivera, A.J., Carmona, C.J. & Parras, M. (2010). CO2RBFN for short and medium term forecasting of the extra-virgin olive oil price. *Studies in Computational Intelligence*. 284. p.pp. 113–125.

Petersen, R. (2018). *6 essential steps to the data mining process - BarnRaisers, LLC*. 1 Outubro 2018. Measurement and ROI. Available from: https://barnraisersllc.com/ 2018/10/01/data-mining-process-essential-steps/. [Acedido: 31 Agosto 2020].

Pfaff, B. & Stigler, M. (2018). vars: VAR Modelling. R package version 1.5-3. *https://CRAN.R-project.org/package=vars*.

R Core Team (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

Ramos, P., Santos, N. & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*. 34. p.pp. 151–163.

Raudys, S.J. & Jain, A.K. (1991). Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 13 (3) p.pp. 252–264.

Reis, P. (2014). *O Olival em Portugal, Tecnologias e relação com o desenvolvimento rural*. Lisboa: Animar-associacão portuguesa para o desenvolvimento local.

Ridley, E. (2015). *Spanish Olive Oil Prices Surge* . [Online]. 22 Outubro 2015. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/europe/spanish-olive-oil-prices-surge/49315. [Acedido: 12 Setembro 2020].

Rivera, A.J., Pérez-Recuerda, P., Pérez-Godoy, M.D., del Jesús, M.J., Frías, M.P. & Parras, M. (2011). A study on the medium-term forecasting using exogenous variable selection of the extra-virgin olive oil with soft computing methods. *Applied Intelligence*. 34 (3). p.pp. 331–346.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. 65 (6) p.pp. 386–408.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. [Online]. Available from: http://www.rstudio.com/.

Schneider, S. & Steiner, M. (2005). Conditional Asset Pricing - Predicting Time Varying Beta-Factors with Group Method of Data Handling Methods. *SSRN Electronic Journal*. p.p. 27.

Silva, A.M. (2015). *Modelos Preditivos Aplicados ao Retalho*. Porto: Faculdade de Economia do Porto.

da Silva, I.N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L.H.B., dos Reis Alves, S.F., da Silva, I.N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L.H.B. & dos Reis Alves, S.F. (2017). Artificial Neural Network Architectures and Training Processes. Em: *Artificial Neural Networks*. Springer International Publishing, pp. 21–28.

Sims, C.A. (1980). Macroeconomics and Reality. *Econometrica*. 48 (1). p.pp. 1–48.

Spencer, D.E. (1993). Developing a Bayesian vector autoregression forecasting model. *International Journal of Forecasting*. 9 (3). p.pp. 407–421.

Teng, G., He, C. & Gu, X. (2014). Response model based on weighted bagging GMDH. *Soft Computing*. 18 (12). p.pp. 2471–2484.

Thomas, A.J., Petridis, M., Walters, S.D., Gheytassi, S.M. & Morgan, R.E. (2017). Two hidden layers are usually better than one. *Communications in Computer and Information Science*. 744. p.pp. 279–290.

Tilve, M.V. (2020). GMDHreg: Regression using GMDH Algorithms. R package version 0.2.1. *https://CRAN.R-project.org/package=GMDHreg*.

Torbat, S., Khashei, M. & Bijari, M. (2018). A hybrid probabilistic fuzzy ARIMA model for consumption forecasting in commodity markets. *Economic Analysis and Policy*. 58. p.pp. 22–31.

Tsay, R. (2010). *Analsis of Financial Time Series*. 3.ª Ed. Chicago, IL: JOHN WILEY & SONS, INC.

Tsay, R.S. & Wood, D. (2018). MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 1.0. *https://CRAN.R-project.org/package=MTS*.

Vasilopoulos, C. (2018). *Olive Oil Prices Slip in European Markets* . [Online]. 26 Março 2018. Olive Oil Times. Available from: https://www.oliveoiltimes.com/business/europe/olive-oil-prices-slip-in-european-markets/62564. [Acedido: 12 Setembro 2020].

Wanas, N., Auda, G., Kamel, M.S. & Karray, F. (1998). On the optimal number of hidden nodes in a neural network. Em: *Canadian Conference on Electrical and Computer Engineering*. 1998, pp. 918–921.

Wickham, H., Bryan, J., RStudio, Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., Hoerl, D. & Miller, E. (2019). readxl: Read Excel files. R package version 1.3.1. *https://CRAN.R-project.org/package=readxl*.

Yefimenko, S. (2018). Building vector autoregressive models using COMBI GMDH with recurrent-and-parallel computations. Em: *Advances in Intelligent Systems and Computing*. 2018, Springer Verlag, pp. 601–613.

Yilmaz, I. & Kaynar, O. (2011). Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert Systems with Applications*. 38 (5). p.pp. 5958–5966.

Yu, L., Wang, S. & Lai, K.K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*. 30 (5). p.pp. 2623–2635.

Zhang, G., Patuwo, B.E. & Hu, M.Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*. 14 (1). p.pp. 35–62.

Zhang, G.P. & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*. 160 (2). p.pp. 501–514.

Zhang, P.G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 50. p.pp. 159–175.

Zhao, E., Zhao, J., Liu, L., Su, Z. & An, N. (2016). Hybrid wind speed prediction based on a self-adaptive ARIMAX model with an exogenous WRF simulation. *Energies*. 9 (1). p.pp. 1–20.

Zivot, E. & Wang, J. (2003). Vector Autoregressive Models for Multivariate Time Series. Em: *Modeling Financial Time Series with S-Plus®*. pp. 385–429.

Anexos A - Figuras

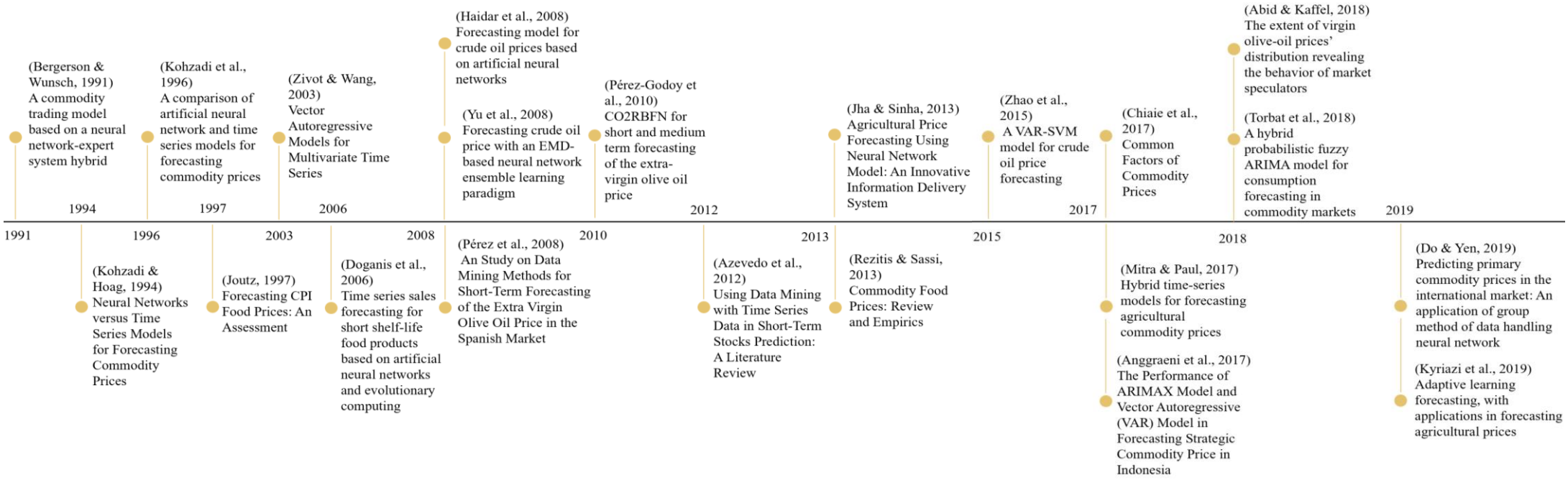


Figura 1: Artigos publicados entre 1991 e 2019 sobre a previsão dos preços de mercadorias, com especial destaque no setor agrícola e da olivicultura.

Fonte: Elaboração própria

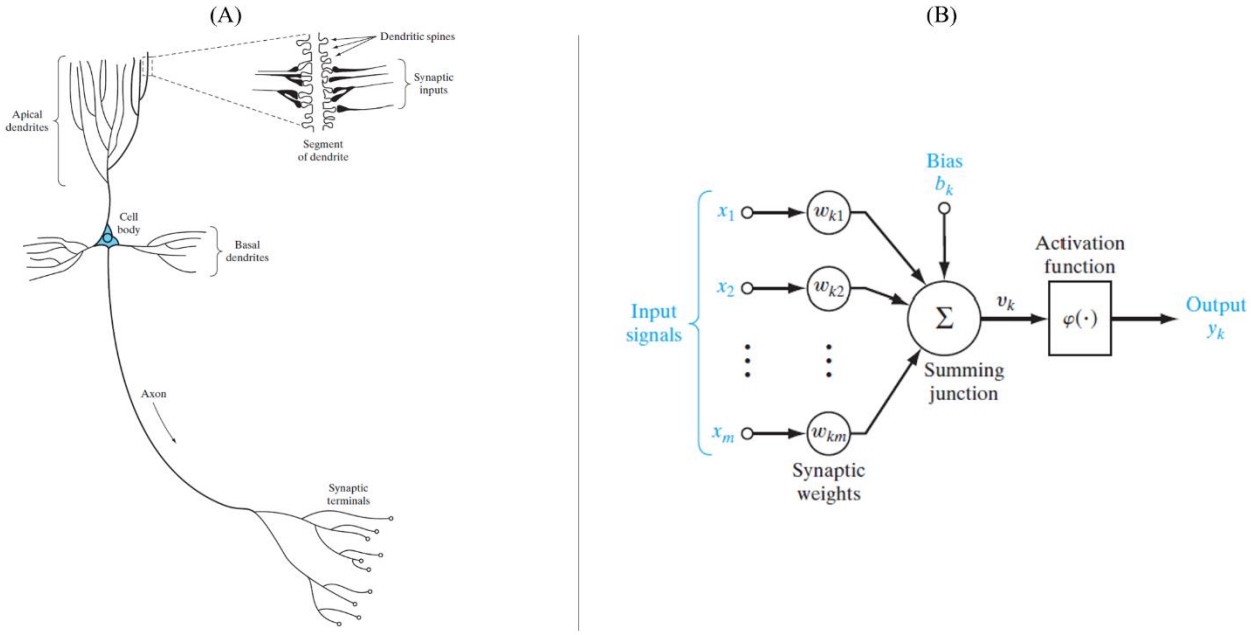


Figura 2: Modelo de um neurónio: biológico, a célula piramidal (A); artificial não linear (B) (Haykin, 2008).

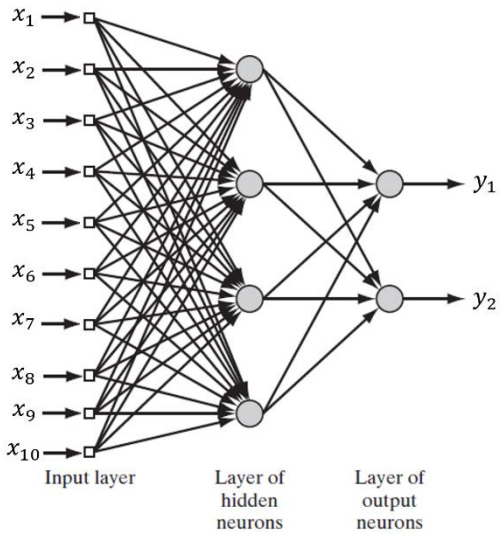


Figura 3: ANN do tipo *feedforward* totalmente conectada, com uma camada oculta e uma camada de saída;
(Haykin, 2008) (Adaptado).

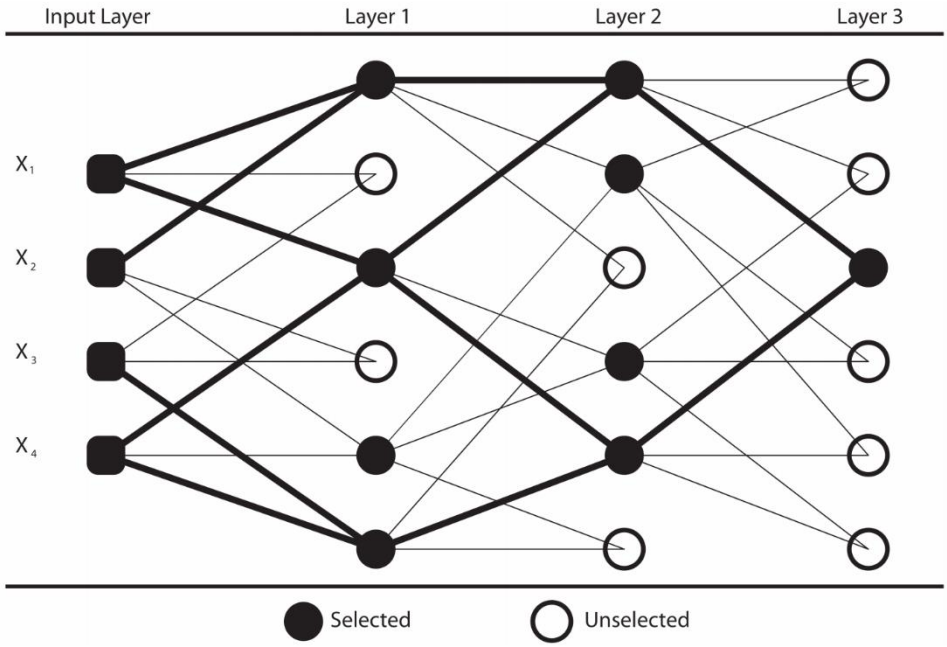


Figura 4: Arquitetura do algoritmo GMDH; (Dag & Yozgatligil, 2016a).

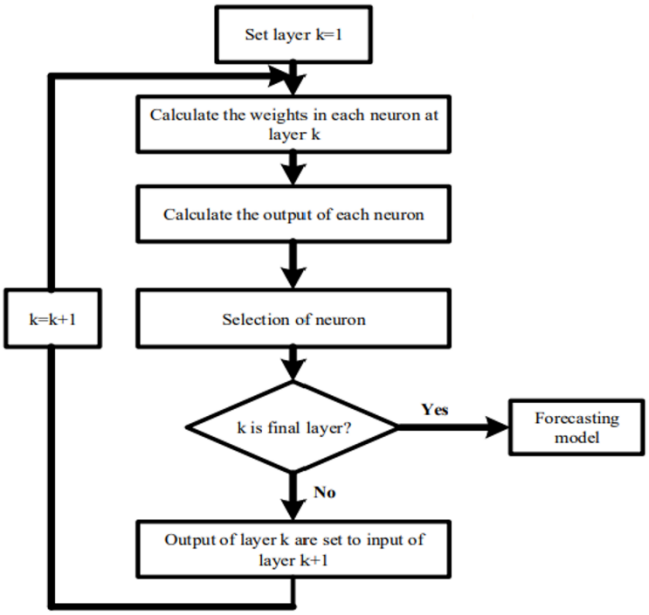


Figura 5: Fluxograma do algoritmo GMDH; (Do & Yen, 2019) (Adaptado).

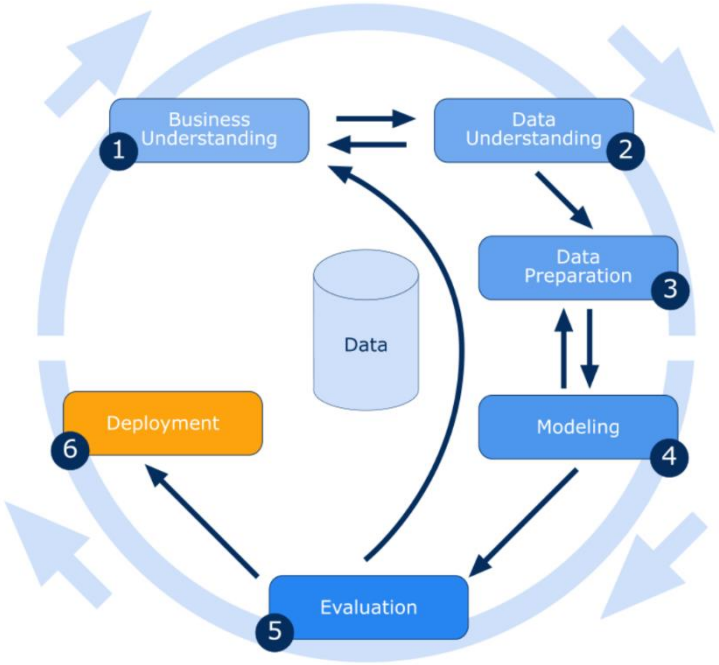


Figura 6: Fases da Metodologia CRISP-DM; (Petersen, 2018).

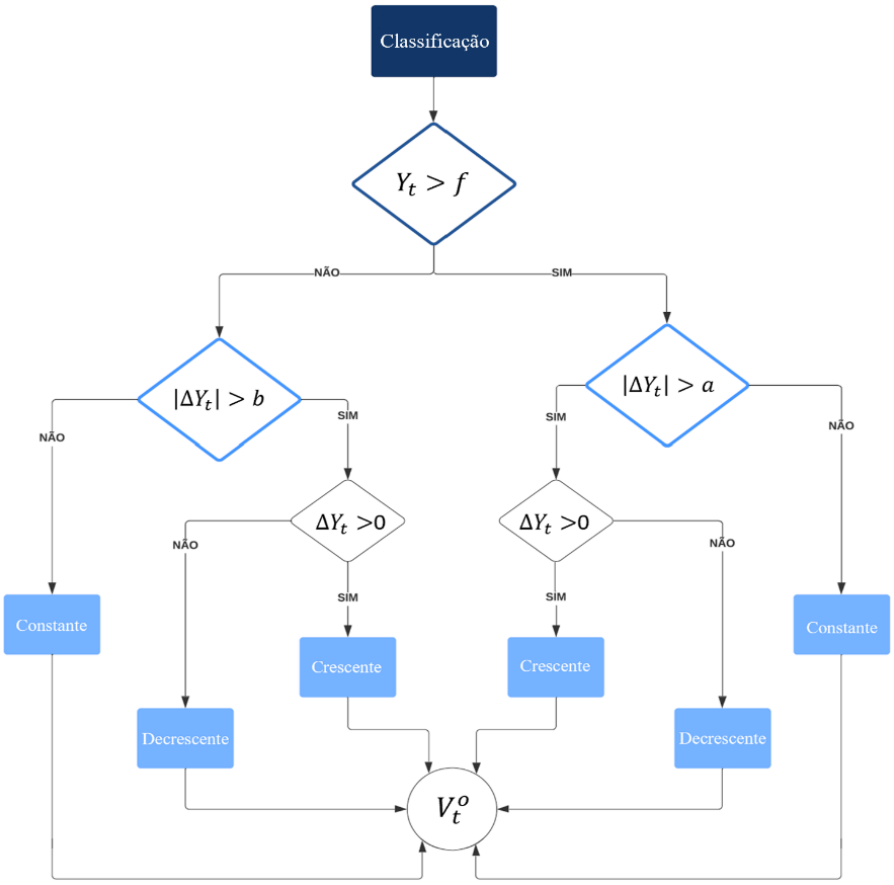
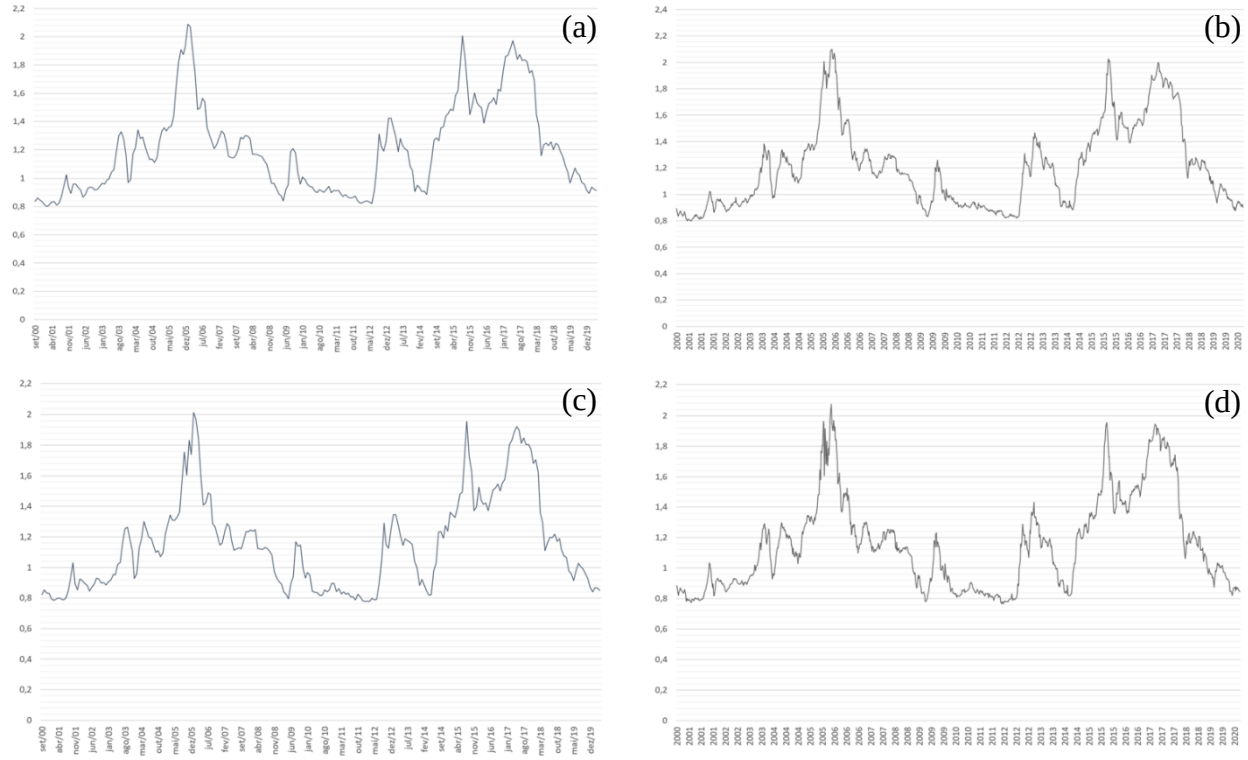


Figura 7: Fluxograma do processo de classificação da "tendência", valores observados;

Fonte: Elaboração própria



*Os valores dos preços foram alterados, mas refletem o comportamento das séries

Figura 8-B: Preços dos Azeite Virgem e Lampante por tonelada desde 2000, de cima para baixo, da esquerda para a direita, respetivamente; (a)VIR mensalmente, (b)VIR semanalmente, (c)LAM mensalmente, (d)LAM semanalmente;

Fonte: Elaboração própria

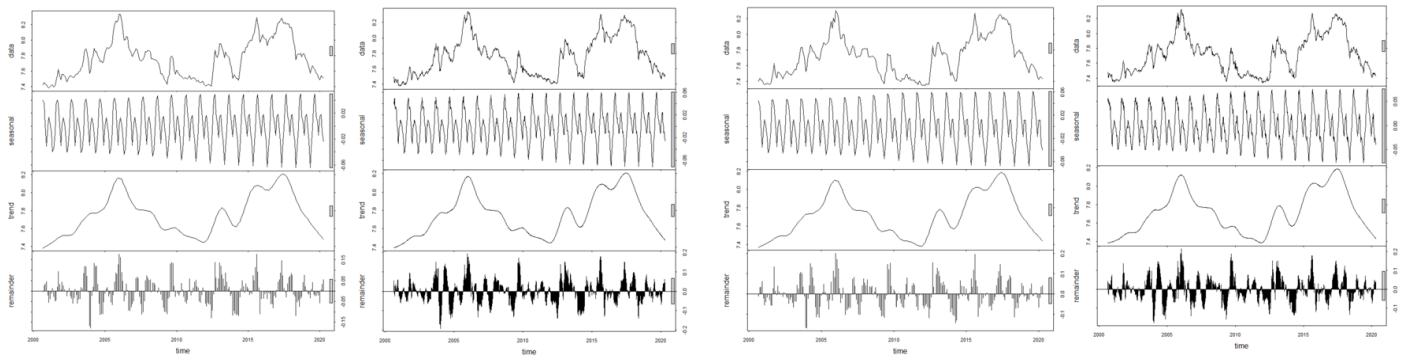


Figura 9-B: Decomposição das séries temporais VIR e LAM nas três componentes principais: sazonalidade, tendência e parte aleatória, ordenados por VIR mensal e semanal LAM mensal e semanal da esquerda para a direita, respetivamente;

Fonte: Elaboração própria

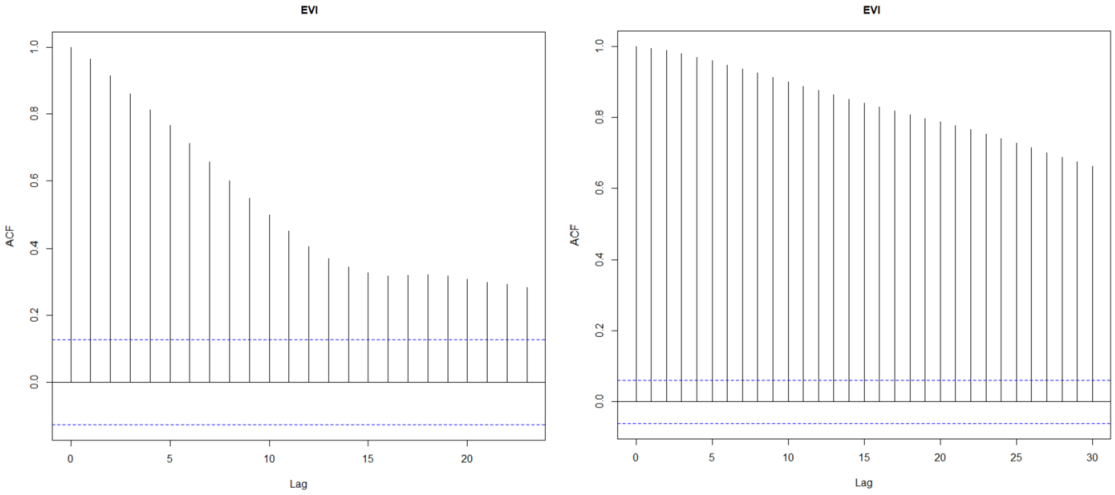


Figura 10: Gráficos das ACF dos preços do azeite extra virgem mensal e semanal, da esquerda para a direita respetivamente;

Fonte: Elaboração própria.

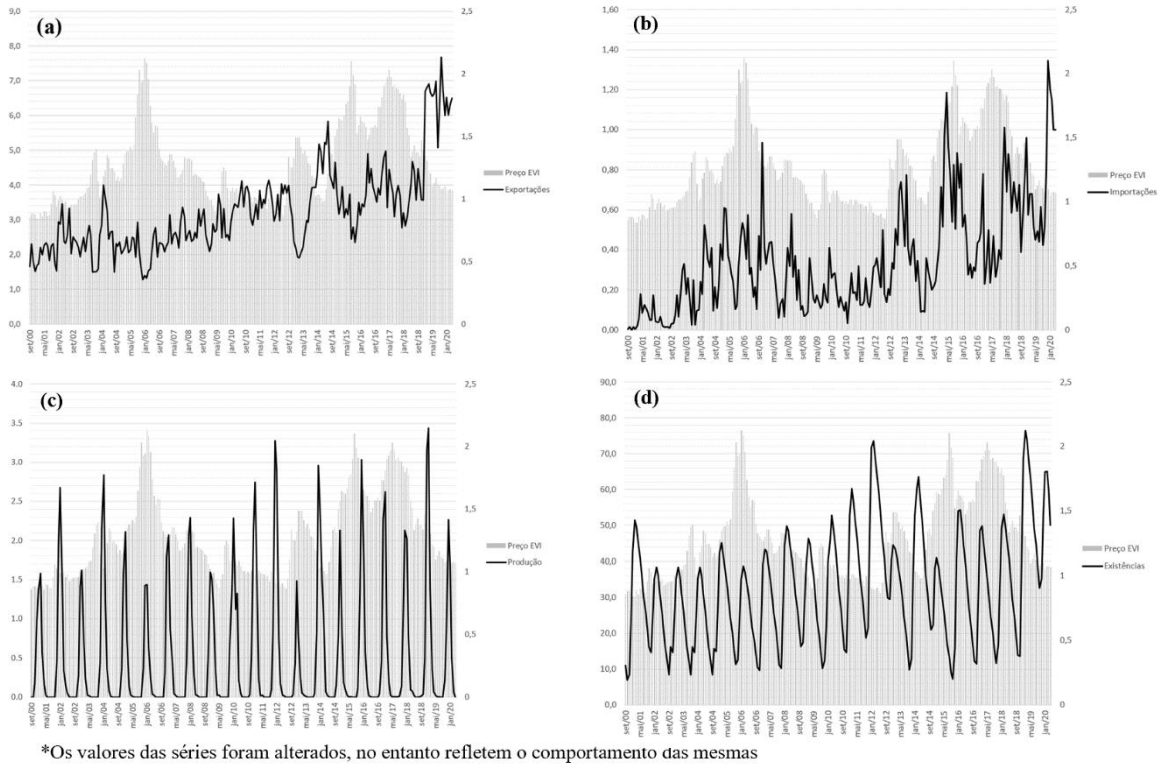
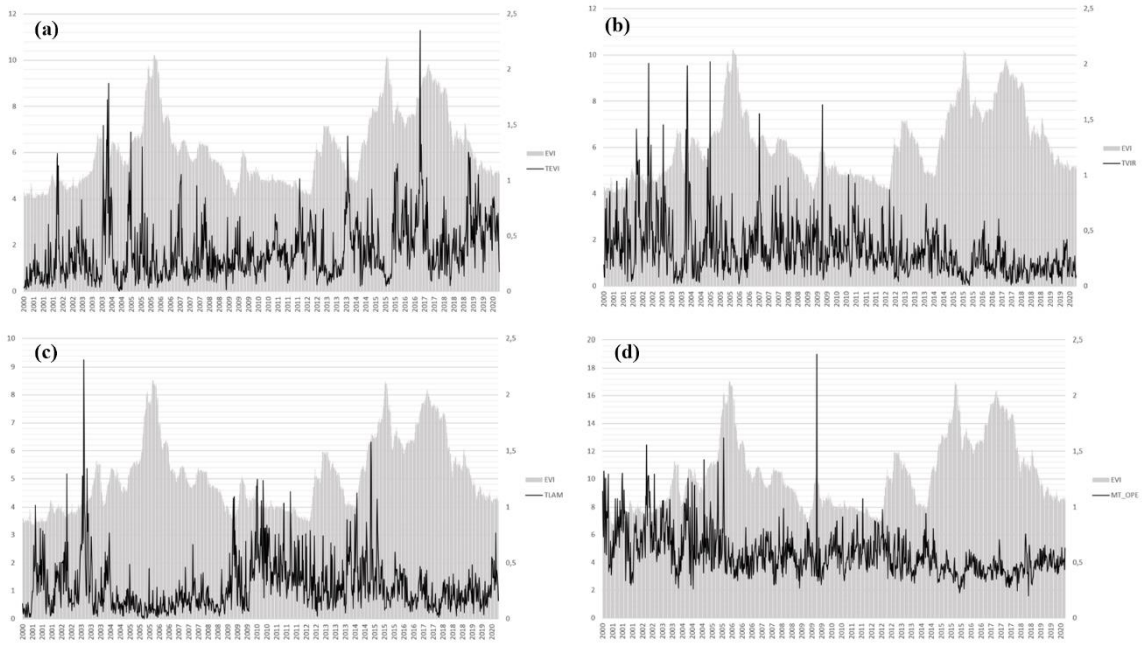


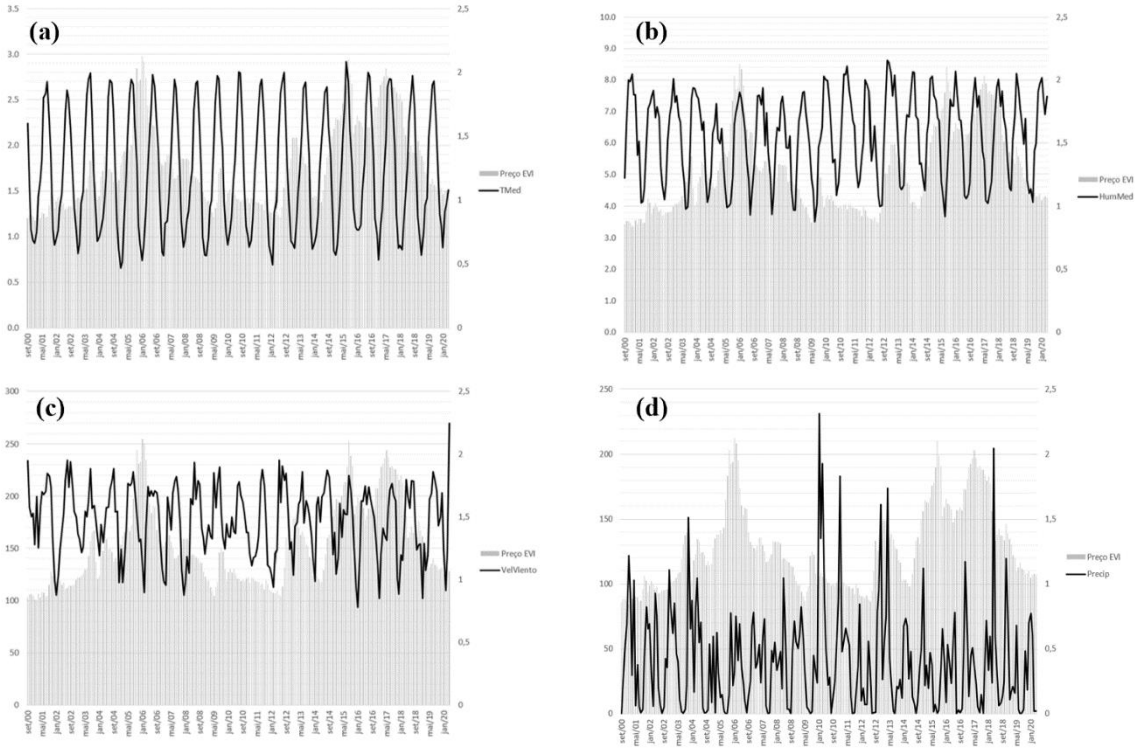
Figura 11: Preços do azeite mensal extra virgem versus: (a) exportações; (b) importações; (c) produção; (d) existências;

Fonte: Elaboração própria.



*Os valores das séries foram alterados, no entanto refletem o comportamento das mesmas

Figura 12: Preços do azeite semanal extra virgem versus: (a) toneladas adquiridas de azeite EVI; (b) toneladas adquiridas de azeite VIR; (c) toneladas adquiridas de azeite LAM; (d) número total de operações realizadas;
Fonte: Elaboração própria



*Os valores da série "preço do azeite extra virgem" foram alterados, no entanto refletem o comportamento da mesma

Figura 13: Preços do azeite semanal extra virgem versus: (a) temperatura média; (b) humidade média; (c) velocidade do vento média; (d) precipitação média;
Fonte: Elaboração própria

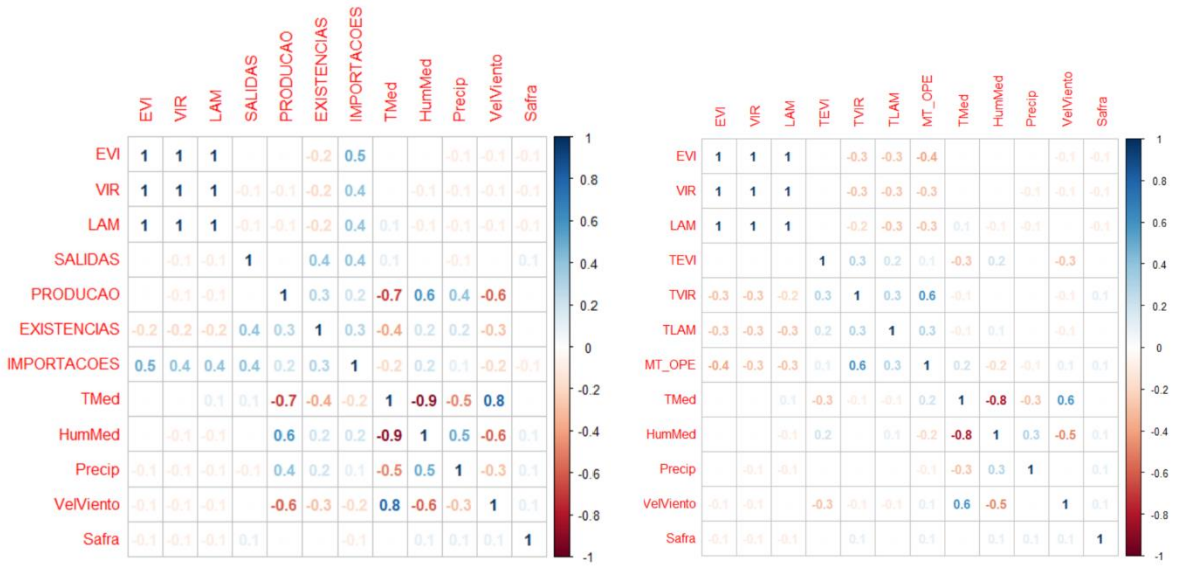


Figura 14: Correlação entre as diferentes variáveis utilizadas na construção dos modelos, mensais e semanais da esquerda para a direita, respetivamente.

Fonte: Elaboração própria

Anexos B - Tabelas

Tabela 1 Principais funções de ativação utilizadas em ANN's

Função	Formula matemática
Função Sigmóide (logística)	$\varphi(v) = (1 + exp(-v))^{-1}$
Função tangente hiperbólica	$\varphi(v) = (\exp(v) - \exp(-v))/(\exp(v) + \exp(-v))$
Função seno ou Função cosseno	$\varphi(v) = sin(v)$ ou $\varphi(v) = cos(v)$
Função linear	$\varphi(v) = v$

Tabela 1: Principais funções de ativação utilizadas em ANN's; (Zhang et al., 1998).

Tabela 2 Desempenho da previsão do modelo ARIMA

Série temporal	Modelo		RMSE	MAPE (%)	Acertos (%)
	Periodicidade	ARIMA			
Extra Virgem	Mensal	$(1,1,1)x(2,0,2)_{12}$	332,70	8,60%	26,09
	Semanal	$(2,1,1)x(1,0,0)_{52}$	141,69	3,72%	38,24
Virgem	Mensal	$(1,1,1)x(2,0,2)_{12}$	364,16	10,18%	28,26
	Semanal	$(2,1,2)x(1,0,0)_{52}$	139,81	4,00%	36,27
Lampante	Mensal	$(4,1,4)x(1,1,2)_{12}$	374,61	11,02%	36,96
	Semanal	$(3,1,3)x(1,0,1)_{52}$	145,18	4,18%	43,14

Tabela 2: Desempenho da previsão do modelo ARIMA para os períodos de maio 2016 a abril 2020.
Fonte: Elaboração própria

Tabela 3 Desempenho da previsão dos modelos ARIMAX

Série temporal	Modelo		RMSE	MAPE (%)	Acertos (%)
	Periodicidade/ Ordens	Variáveis Independentes			
Extra Virgem	Mensal (2,1,1)x(0,0,1)12	Todas*	327,44	8,450%	45,65
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;Tmedt-6;ΔImpt-6;Safrat	319,09	8,312%	34,78
		ΔEVIt-5;ΔImpt-6;Safrat	**315,22**	**8,074%**	39,13
		ΔImpt-6;Safrat	329,12	8,623%	34,78
	Semanal (1,1,1)x(1,0,1)52	Todas¨	141,24	3,627%	45,59
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTEVIt-6;ΔMtopet-6;ΔTmedt-6;Safrat	**140,71**	3,633%	44,61
		ΔEVIt-5;ΔTmedt-6;Safrat	140,95	**3,618%**	45,10
Virgem	Mensal (1,1,1)x(1,0,1)12	Todas*	364,13	10,100%	60,87
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;Tmedt-6;ΔImpt-6;Safrat	356,75	9,970%	50,00
		ΔVIRt-5;ΔImpt-6;Safrat	**355,50**	**9,800%**	47,83
		ΔImpt-6;Safrat	366,07	10,140%	50,00
	Semanal (1,1,1)x(0,0,1)52	Todas¨	140,51	3,897%	50,00
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTVIRt-6;ΔMtopet-6;ΔTmedt-6;Safrat	**140,23**	3,914%	49,51
		ΔVIRt-5;ΔTmedt-6;Safrat	140,54	**3,900%**	48,04
Lampante	Mensal (2,1,1)x(1,0,1)12	Todas*	376,42	11,620%	47,83
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;Tmedt-6;ΔImpt-6;Safrat	375,41	11,370%	39,13
		ΔLAMt-5;ΔImpt-6;Safrat	**374,76**	**10,380%**	41,30
		ΔImpt-6;Safrat	381,53	11,440%	30,56
	Semanal (2,1,1)x(0,0,1)52	Todas¨	145,80	**4,080%**	42,16
		ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTLAMt-6;ΔMtopet-6;ΔTmedt-6;Safrat	145,67	4,109%	43,14
		ΔLAMt-5;ΔTmedt-6;Safrat	**145,63**	4,120%	42,16

*ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔSalidast-6;Prodt-6;ΔExist-6;ΔImpt-6;Tmedt-6;ΔHmedt-6;ΔChuvat-6;ΔVentot-6;Safrat

¨ΔEVIt-5;ΔVIRt-5;ΔLAMt-5;ΔTevit-6;ΔTvirt-6;ΔTlamt-6;ΔMtopet-6;ΔTmedt-6;ΔHmedt-6;ΔChuvat-6;ΔVentot-6;Safrat

Tabela 3: Desempenho da previsão dos modelos ARIMAX para os períodos de maio 2016 a abril 2020.

Fonte: Elaboração própria

Tabela 4 Desempenho da previsão dos modelos VAR (1)

Série temporal	Modelo		RMSE	MAPE (%)	Acertos (%)
	Periodicidade	Variáveis Independentes			
Extra Virgem	Mensal	Todas*	369,66	10,399%	34,78
		ΔVIRt;ΔLAMt;Tmedt;ΔImpt;Safrat	271,84	7,219%	32,61
		ΔImpt;Safrat	**264,69**	**6,861%**	28,26
	Semanal	Todas¨	116,20	**2,979%**	39,41
		ΔVIRt;ΔLAMt;ΔTEVIt;ΔMtopet;ΔTmedt;Safrat	**116,16**	3,045%	43,35
		ΔTmedt;Safrat	118,09	3,088%	47,78
Virgem	Mensal	Todas*	417,51	12,759%	43,48
		ΔEVIt;ΔLAMt;Tmedt;ΔImpt;Safrat	307,42	8,648%	36,96
		ΔImpt;Safrat	**306,75**	**8,630%**	32,61
	Semanal	Todas¨	115,81	3,313%	45,81
		ΔEVIt;ΔLAMt;ΔTVIRt;ΔMtopet;ΔTmedt;Safrat	**114,78**	**3,235%**	45,32
		ΔTmedt;Safrat	116,77	3,280%	45,81
Lampante	Mensal	Todas*	436,20	14,431%	32,61
		ΔEVIt;ΔVIRt;Tmedt;ΔImpt;Safrat	320,12	9,605%	36,96
		ΔImpt;Safrat	**312,52**	**9,130%**	41,30
	Semanal	Todas¨	122,74	3,690%	32,02
		ΔEVIt;ΔVIRt;ΔTLAMt;ΔMtopet;ΔTmedt;Safrat	**122,52**	3,604%	35,47
		ΔTmedt;Safrat	124,18	**3,571%**	42,86

*ΔEVIt;ΔVIRt;ΔLAMt;ΔSalidast;Prodt;ΔExist;ΔImpt;Tmedt;ΔHmedt;ΔChuvat;ΔVentot;Safrat

¨ΔEVIt;ΔVIRt;ΔLAMt;ΔTevit;ΔTvirt;ΔTlamt;ΔMtopet;ΔTmedt;ΔHmedt;ΔChuvat;ΔVentot;Safrat

Tabela 4: Desempenho da previsão dos modelos VAR para os períodos de maio 2016 a abril 2020.

Fonte: Elaboração própria

Tabela 5 Desempenho da previsão dos modelos ANN - MLP

Série temporal	Modelo		Rede		RMSE	MAPE (%)	Acertos (%)
	Periodicidade	Variáveis Independentes	Camadas ocultas	Número de Neurónio			
Extra Virgem	Mensal	Univariada	1	10	**386,01**	10,520%	30,43
		Univariada	2	(10;10)	399,34	**9,950%**	43,48
		$\Delta Imp_{t-6};Safra_t$	1	12	438,22	12,210%	39,13
	Semanal	Univariada	1	10	**141,97**	4,000%	31,86
		Univariada	2	(10;10)	145,50	**3,800%**	40,20
		$\Delta Mtope_{t-6};Safra_t$	1	12	146,13	3,850%	33,82
Virgem	Mensal	Univariada	1	10	**431,36**	**12,470%**	39,13
		Univariada	2	(10;10)	439,40	13,440%	28,26
		$\Delta Imp_{t-6};Safra_t$	1	12	477,02	13,868%	39,13
	Semanal	Univariada	1	10	148,44	4,212%	29,41
		Univariada	2	(10;10)	144,34	3,972%	39,22
		$\Delta Mtope_{t-6};Safra_t$	2	(12;12)	**140,43**	**3,950%**	37,25
Lampante	Mensal	Univariada	1	10	413,54	11,785%	43,48
		Univariada	2	(10;10)	**385,91**	**10,410%**	39,13
		$\Delta Imp_{t-6};Safra_t$	2	(12;12)	417,91	12,850%	45,65
	Semanal	Univariada	1	10	149,38	4,248%	42,16
		Univariada	2	(10;10)	142,04	3,999%	39,22
		$\Delta Mtope_{t-6};Safra_t$	2	(12;12)	**140,19**	**3,891%**	32,84

Tabela 5: Desempenho da previsão dos modelos MLP para os períodos de maio 2016 a abril 2020.

Fonte: Elaboração própria

Tabela 6 Desempenho da previsão dos modelos GMDH

Série temporal	Modelo		RMSE	MAPE (%)	Acertos (%)
	Periodicidade	Variáveis Independentes			
Extra Virgem	Mensal	Univariada	**346,74**	**8,675%**	28,26
		$\Delta Imp_{t-6};Safra_t$	385,81	10,090%	34,78
	Semanal	Univariada	**139,76**	**3,562%**	36,27
		$\Delta VIR_{t-5};\Delta LAM_{t-5};\Delta Mtope_{t-6};Safra_t$	214,80	5,722%	43,63
		$\Delta EVI_{t-5};\Delta Mtope_{t-6};Safra_t$	158,87	4,043%	49,02
Virgem	Mensal	Univariada	**380,55**	**10,419%**	34,78
		$\Delta Imp_{t-6};Safra_t$	438,92	12,790%	26,09
	Semanal	Univariada	**139,16**	**3,963%**	35,78
		$\Delta EVI_{t-5};\Delta LAM_{t-5};\Delta Mtope_{t-6};Safra_t$	215,47	6,271%	48,53
		$\Delta VIR_{t-5};\Delta Mtope_{t-6};Safra_t$	162,24	4,594%	52,45
Lampante	Mensal	Univariada	**405,12**	**11,600%**	39,13
		$\Delta Imp_{t-6};Safra_t$	451,43	14,080%	52,17
	Semanal	Univariada	174,21	**4,678%**	47,55
		$\Delta EVI_{t-5};\Delta VIR_{t-5};\Delta Mtope_{t-6};Safra_t$	233,31	6,985%	42,65
		$\Delta LAM_{t-5};\Delta Mtope_{t-6};Safra_t$	**166,12**	4,700%	54,41

Tabela 6: Desempenho da previsão dos modelos GMDH para os períodos de maio 2016 a abril 2020.

Fonte: Elaboração própria