Faculdade de Engenharia da Universidade do Porto

Planeamento de Investimentos na Rede de Distribuição com base na Técnica Spike and Slab

Hugo Francisco Rocha Costa

VERSÃO FINAL

Dissertação realizada no âmbito do
Mestrado em Engenharia Electrotécnica e de Computadores
Major Energia

Orientador: José Nuno Marques Fidalgo
Supervisor: Pedro Miguel Ferreira Macedo

julho de 2022

Resumo

Esta dissertação foi inspirada na obrigatoriedade de realizar um planeamento de investimentos, como está descrito no Plano de Desenvolvimento e Investimento da Rede de Distribuição (PDIRD), o qual consta no artº 10 do Decreto-Lei nº 15/2022, de 14 de janeiro [1], algo que se encontra a cargo do Operador da Rede de Distribuição (ORD), E-REDES em Portugal,

Este planeamento é de capital importância, pois os erros de previsão têm elevados custos quer sejam eles financeiros ou operacionais, pois, caso a previsão do investimento obrigatório seja feita por excesso, o dinheiro, por ele previsto, ficará imobilizado; caso seja feita por defeito, não haverá suporte financeiro para todas as melhorias necessárias na rede de distribuição.

O problema da estimação do investimento apresenta inúmeras variáveis, tais como a saúde económica do país, o ritmo da construção civil e o crescimento da carga, mas não conta com muitos exemplos, o que complica a criação de modelos precisos e robustos.

Para este problema, irá ser testado o desenvolvimento de modelos a partir do conceito Spike and Slab, o qual é capaz selecionar as variáveis mais relevantes em simultâneo, descartando as outras, através de um processo regressivo, bem como permite quantificar a incerteza associada à previsão realizada.

A comparação de resultados comprova que o Spike and Slab consegue produzir modelos de grande qualidade, superando a técnica de regressão LASSO (Least Absolute Shrinkage and Selection Operator), e praticamente igualando o desempenho do modelo atualmente usado pela E-REDES.

De notar que o Spike and Slab é o único, entre todos os métodos acima mencionados, que é de natureza probabilística, o que permite lidar facilmente com incerteza nos dados, além de permitir obter resultados probabilísticos, como margens de confiança, o que é uma vantagem considerável em termos de planeamento.

Abstract

This dissertation was inspired by the obligation of a Distribution Network Development and Investment Plan, which is the bridge that unites the transportation network and the consumers and aims to develop a previsions model to estimate the Mandatory Investment.

As I said, this plan is mandatory, as defined in the article 10 of the Decree Law no. 15/2022, 14th of January [1], and it's conception is the national distribution networks' job, that in the Portuguese case is E-REDES. This company has all the interest in estimating a sum of investment that allows, for example, new consumers connected to the network.

High prevision errors have financial and/or operation costs, because if the investment prediction is higher than the investment that needs to be done, the money calculated is frozen and cannot be used anywhere else. In the case that the prediction is lower than the investment needed, there won't be money to spend in all the investments required.

This problem has multiple variables, such as the economic health of the country, the construction rhythm e and growth of load, but it has almost no examples and that makes the creation of precise and robust models very hard.

To obtain this estimate, will be developed a prediction model based on the technique Spike and Slab, which selects the most relevant variables, simultaneously, through a regressive process. The results will be compared with the results obtained by other models, such as, LASSO, that has a project done before which will be mentioned during this work.

Agradecimentos

Um agradecimento especial aos meus pais que, com o seu esforço, me permitiram estudar e tornar-me na pessoa que sou hoje.

Ao professor doutor José Nuno Fidalgo e ao engenheiro Pedro Macedo por todos os conselhos que me deram ao longo deste trabalho e pelo tempo que disponibilizaram para me ajudar a cumprir esta etapa.

A todos os meus amigos que me ajudaram ao longo não só deste processo de desenvolvimento e escrita da dissertação, mas ao longo de toda a minha vida a ultrapassar as dificuldades que se iam apresentando no caminho.

A todos os mencionados, o meu obrigado,

Hugo Costa

“Não dês o peixe, dá, antes, a cana”

Unknown

Índice

Lista de figuras

Lista de tabelas

Abreviaturas

CS	Cabos Subterrâneos (novos)
LA	Linhas Aéreas (novas)
Econs	Energia Consumida
IC	Índice de Construção
IHPC	Índice de Harmonização de Preços ao Consumidor
IO	Investimento Obrigatório
LASSO	Least Absolute Shrinkage and Selection Operator
MAPE	Mean Absolute Percentage Error
NNC	Número de Novas Construções
NCl	Número de Clientes na rede
NNF	Número de Novos Fogos
ORD	Operador da Rede de Distribuição
PDIRD	Plano de Desenvolvimento e Investimento na Rede de Distribuição
PIB	Produto Interno Bruto
PT	Postos de Transformação (nova Potência Instalada)
SEQ	Somatório dos Erros Quadráticos
TDes	Taxa de Desemprego

Capítulo 1

Introdução

O consumo de energia elétrica tem vindo a aumentar, o que torna vital o investimento na rede de distribuição, que é a ponte de ligação entre a rede de transporte e as cargas.

O planeamento do investimento na rede de distribuição é de cariz obrigatório, tal como definido pelo Plano de Desenvolvimento e Investimento da Rede de Distribuição (PDIRD), o qual consta no artº 10 do Decreto-Lei nº 15/2022, de 14 de janeiro [1], e recai no Operador da Rede de Distribuição (ORD), que no caso português é a E-REDES, pelo que a mesma tem todo o interesse, em estimar o montante de investimento que permita garantir, por exemplo, que novos produtores ou consumidores possam ser ligadas à rede.

Esta dissertação foi realizada no âmbito da unidade curricular Dissertação pertencente ao Mestrado em Engenharia Eletrotécnica e de Computadores, tendo, como objetivo último, a elaboração de um modelo de previsões para estimar o Investimento Obrigatório (IO).

Para a obtenção desta estimativa, irá ser desenvolvido um modelo de previsão baseado no conceito de Spike and Slab, que é uma técnica bayesiana, capaz de selecionar as variáveis mais relevantes e, em simultâneo, através de um processo regressivo, combater a sobreadaptação, bem como fornecer a incerteza associada. Os resultados obtidos irão ser comparados com os conseguidos por outros métodos, como, por exemplo, o LASSO, sobre o qual foi realizada uma dissertação anteriormente [2].

1.1 Enquadramento

Como mencionado acima, o planeamento de investimentos obrigatórios foi impulsionado pela legislação introduzida para o efeito, sendo que este é um problema que contempla múltiplas variáveis, tais como a saúde económica do país, o ritmo da construção civil e o crescimento da carga. No entanto, o facto de não existirem muitos dados históricos, nem uma relação direta entre as múltiplas variáveis a considerar, torna complicada a estimação de modelos precisos e robustos.

Erros elevados de previsão têm custos financeiros e/ou operacionais consideráveis, pois, no caso de ser previsto um IO por excesso, o montante previsto ficará cativo, mesmo não sendo todo utilizado. Já no caso de ser estimado por defeito, significará que nem todos os

investimentos necessários poderão ser realizados ou que será necessário recorrer a financiamento urgente com taxas, normalmente, elevadas.

1.2 Objetivos

O objetivo último deste trabalho é criar um modelo capaz de responder à necessidade que o ORD tem de estimar investimentos obrigatórios, para um período de 5 anos, algo que está previsto, no PDIRD, a serem feitos na rede de distribuição de forma a suportar, por exemplo, novas ligações à rede. O modelo a implementar será baseado na técnica Spike and Slab [12].

Pretende-se que este modelo seja eficiente, pois grandes desvios do valor real acarretam problemas operacionais no ORD, o que torna as decisões da escolha dos modelos um processo minucioso.

Pretende-se também explorar a natureza probabilística da técnica Spike and Slab para caracterizar a solução em termos de margens de confiança.

1.3 Estrutura

Este documento encontra-se dividido em 5 capítulos, sendo o primeiro a introdução ao tema, a revisão bibliográfica e a importância do estudo da previsão dos Investimentos Obrigatórios na Rede de Distribuição.

No segundo capítulo é abordada a metodologia usada no trabalho, sendo em cada subcapítulo abordada uma técnica diferente.

No capítulo 3 são apresentados os principais resultados dos estudos desenvolvidos e o modelo final para previsão do IO. É também apresentada a comparação com os resultados obtidos pelo LASSO, que é uma metodologia já testada numa dissertação anterior [2]. Além disso, são efetuadas as previsões feitas com dados de anos posteriores aos fornecidos no início da dissertação, para estudo do comportamento do modelo para situações não testadas anteriormente.

Por fim, no capítulo 4, são apresentadas as conclusões e são propostos desenvolvimentos futuros.

Capítulo 2

Metodologia

A previsão do Investimento Obrigatório (IO) é um tema relativamente recente e muito específico do Sistema Elétrico Português, pelo que, não existe muita bibliografia associada ao tema, o que representa, naturalmente, um entrave na escolha quer de métodos a utilizar, quer na comparação entre métodos, por forma a ser possível determinar qual o melhor.

A necessidade de desenvolver novas abordagens para realização de estudos de estimação do IO deveu-se, principalmente, a alterações da evolução de dois fatores: o consumo e o índice de construção (IC). Quanto ao primeiro, até 2006/2007, caracterizou-se pela estabilidade de crescimento, sendo, mais tarde, marcado por uma estagnação, algo que mudou de figura com a crise económica de 2009, conduzindo mesmo a um decréscimo da carga [2]. Por outro lado, o índice de construção entrou em forte declínio, reduzindo substancialmente os pedidos de novas ligações e, consequentemente, diminuindo o IO.

Portanto, até 2006/2007, a evolução do IO era mais estável e fácil de prever, a partir dessa altura deixou de o ser, o que significa que o modelo usado pelo distribuidor deixou de ser confiável. As alterações dos fatores influentes no IO, em particular o consumo e o IC, obrigaram a repensar o modelo, no sentido de tentar aproximar as previsões aos investimentos realmente necessários.

2.1. Conceitos gerais

2.1.1 O problema da sub e sobre adaptações

Para gerar um modelo de previsões é necessário conhecer as variáveis que irão ser utilizadas, pois o modelo pode depender de:

- Forma não linear das variáveis;

- Variáveis que sejam redundantes, mesmo que tenham influência no desempenho do modelo;
- Variáveis que tenham preponderância nesse mesmo modelo, mas que sejam muito difíceis de caracterizar [3].

Qualquer modelo de regressão é dado pela soma da previsão com um erro associado a essa mesma previsão, sendo esta soma igual ao valor real.

De notar que este erro está associado à imprevisibilidade do modelo e que o objetivo é reduzir esse mesmo erro, aproximando, assim, a previsão ao valor real.

Qualquer modelo carece de verificação, sendo para isso necessária uma análise aos valores previstos, isto é, constatar se fazem ou não sentido, visualizar a evolução do modelo, realizar avaliações numéricas de modo a caracterizar a qualidade dos modelos e compará-los [3].

Para a geração de modelos é importante conhecer os conceitos de variância e viés (*bias* em inglês).

Variância é a medida das flutuações nos dados, sendo que uma elevada variância implica um maior afastamento do valor real, o que se inverte para uma baixa variância. Uma alta variância é associada a sobre-adaptação do modelo (Overfitting) [4].

Já viés corresponde a um desvio sistemático, eventualmente causado pela aplicação do modelo de teste. Um viés elevado corresponde a uma não identificação, por parte do modelo, de relações importantes entre entradas e saídas. Um elevado viés está associado a sub-adaptação do modelo (Underfitting). Um modelo simples obtido através de uma regressão linear vai ter elevado viés, pois os pontos encontram-se afastados da reta de regressão linear. No entanto, têm uma variância baixa porque os pontos não se encontram dispersos [4].

Para a correção do viés é usual considerar o aumento do número de variáveis e/ou o aumento da capacidade do modelo. Para diminuir a variância é necessário o aumento de dados de treino ou através de técnicas de regulação [5].

No caso da estimação de investimentos na rede de distribuição, os fatores influentes (variáveis de entrada), são abundantes, mas com poucos exemplos, o que pode resultar em erros de treino muito baixos, mas erros de teste altos, devido a uma sobre-adaptação dos dados (Overfitting) [2].

Overfitting ocorre quando o modelo se adapta em demasia aos dados históricos fornecidos, mas comete erros ao fazer previsões de futuro, pois o modelo adapta-se aos dados de treino e ao ruído a eles associado, provocando, assim, grandes erros de teste [4].

Quando estamos perante um modelo simples, isto é, uma regressão linear, pode ocorrer Underfitting, ou seja, o modelo não se adapta bem aos dados de treino [4].

Apesar disso, este tipo de modelo (regressão linear) é frequentemente usado, pois, apesar de existir um erro à saída, produz menos oscilações que um modelo obtido através de uma regressão não linear. Além disto, também implica o cálculo de menos parâmetros, pois para se

obter uma regressão linear simples (uma variável de entrada) apenas é necessária a obtenção de m e b, sendo m a inclinação e b a ordenada na origem [2].

Uma regressão pode ser simples ou múltipla. A simples é o estabelecer de uma relação direta entre uma entrada e uma saída. Já a múltipla tem o mesmo princípio que a simples, mas com múltiplas variáveis, sendo esse o modelo a adotar neste trabalho.

2.1.2 Técnicas Bayesianas

Para perceber o conceito de Spike and Slab é necessária a compreensão das técnicas Bayesianas, pois, este mesmo conceito é uma variante da análise Bayesiana [6].

Esta técnica utiliza probabilidades para representar partes do modelo (por exemplo, entradas) às quais estão associadas incertezas ou probabilidades [7].

Ao contrário de outras técnicas, as Bayesianas, apresentam normalmente a saída (estimativa) em forma de uma distribuição de probabilidades, que é representada por uma distribuição normal, tal como mostra a equação abaixo [13]:

$$\widehat{\boldsymbol{y}} \sim \mathbf{N}\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}, \boldsymbol{\sigma}^{\boldsymbol{2}}\boldsymbol{I}\right) \tag{1}$$

Onde:

- N representa uma distribuição normal
- β é o coeficiente do modelo
- X é a entrada
- $\beta^T X$ é a média;
- σ representa o desvio padrão.

A equação fundamental de uma regressão Bayesiana é dada por [13]:

$$\mathrm{P}\left(\beta \backslash (y, X)\right) = \frac{\mathrm{P}\left(y \backslash (\beta, X)\right) \times \mathrm{P}(\beta \backslash X)}{\mathrm{P}(y \backslash X)} \tag{2}$$

O significado dos diferentes componentes é o seguinte [8]:

- $\mathrm{P}(\beta \backslash X)$ é denominada como distribuição *a priori,* isto é, representa a informação inicial acerca dos parâmetros, sendo que nada tem a ver com os dados. Usa-se normalmente para dar indicações de peritos ou da experiência. Por exemplo, a entrada X_i tem um peso entre 40% e 60% no valor do IO.
- $\mathrm{P}(y \backslash \beta, X)$ é probabilidade de ocorrência dos dados – probabilidade de ocorrer a saída y com os parâmetros β e as entradas X – sendo este um indicador da qualidade da regressão com os parâmetros utilizados.

- $\mathrm{P}(y\backslash X)$ é uma constante normalizadora. Representa a probabilidade de ter o vetor de saída y, dada a matriz de entrada X.
- $\mathrm{P}\big(\beta\backslash(y,X)\big)$ é designada como a distribuição *à posteriori,* sendo que, esta já conta com os dados de treino (entradas e saídas).

O objetivo da utilização deste método é o conhecimento dos parâmetros β, são estes parâmetros que vão ajustar o modelo, visto que os dados são conhecidos.

Se os dados de entrada forem escassos, a distribuição a posteriori vai ser mais espalhada. À medida que o número de dados de entrada aumenta, a probabilidade, $\mathrm{P}(y\backslash\beta,X)$, sobrepõe-se à distribuição à priori, o que, com um número elevado de dados entrada, faz com que os parâmetros do modelo convirjam para os valores do critério dos mínimos quadrados [13].

2.1.3 Spike and Slab

O *Spike and Slab* é uma variante do método bayesiano e é uma forma de seleção de variáveis, pois seleciona as que são verdadeiramente importantes para o modelo, descartando as que são redundantes ou mesmo irrelevantes para o mesmo [9].

O *Spike* é a probabilidade à priori de um dado coeficiente do modelo ser zero e o *Slab* é a distribuição à priori para os valores do coeficiente da regressão [11].

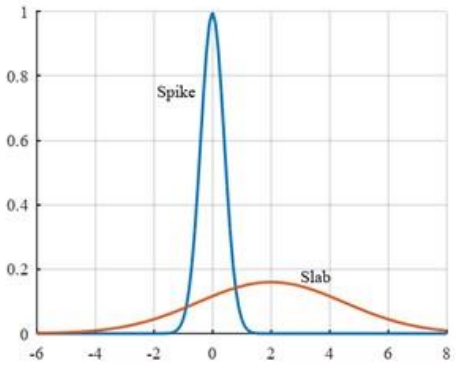

Figura 1- Exemplo de distribuições de Spike e de Slab

Considere-se um número N de variáveis de entrada, já conhecidos, associadas ao modelo, que neste caso são 11. Como temos 11 variáveis, vai ser criado um vetor de dimensão 11, em que cada elemento vale 0 ou 1. Caso não exista qualquer informação, a atribuição de 0 ou 1 será obtida através da distribuição de Bernoulli, sendo, a mesma, uma distribuição discreta no intervalo {0,1}, que tem valor 1 com a probabilidade de sucesso p e valor 0 com a probabilidade complementar q=1-p. As variáveis de entrada vão ser multiplicadas por esse vetor o que faz com que, as que são multiplicadas por 0, não constem no modelo [12].

Se a variável de entrada fizer parte do modelo é identificado um coeficiente, ß, correspondente a essa mesma variável. Este coeficiente é calculado através de uma

distribuição normal, com valor médio 0 e variância calculada por $(X^TX)^{-1}$, em que X é a matriz das variáveis de entrada [10].

Se a variável integrar o modelo temos um ß ≠ 0, já se a variável não integrar o modelo temos um ß=0.

Depois, com base nisto, é calculada a distribuição à posteriori, como indicado no subcapítulo anteior [11].

Este processo é repetido milhares de vezes utilizando a técnica de Markov Chain Monte Carlo, tendo como resultado final a indicação dos coeficientes associados a cada variável e a sua introdução ou não no modelo [12].

2.2. Dados disponibilizados

Por motivos de confidencialidade, os dados fornecidos pela E-REDES/INESC TEC já se encontram normalizados, isto é, sofreram uma transformação linear de modo a ficarem compreendidos no intervalo [1,2], sendo que, essa mesma normalização foi feita através da seguinte equação [2]:

$$X_{normalizado} = \frac{X - X_{mínimo}}{X_{máximo} - X_{mínimo}} + 1 \quad (3)$$

Onde:

- X é a variável a normalizar
- $X_{mínimo}$ é o valor mínimo de cada variável
- $X_{máximo}$ é o valor máximo de cada variável
- $X_{normalizado}$ é a variável depois de normalizada

De notar que, esta normalização, é feita somando 1, pois o valor 0, em alguma das variáveis poderia induzir erros numéricos[1] [2].

Dentro das variáveis de entrada, estão:

- Índice de Construção (IC)
- Número de Novas Construções (NNC)
- Número de Novos Fogos Contruídos (NNF)
- Índice de Harmonização de Preços ao Consumidor (IHPC)

[1] Isto acontece, por exemplo em transformações logarítmicas das variáveis.

- Produto Interno Bruto (PIB) (M€)
- Taxa de Desemprego (Tdes)
- Energia Consumida na Rede (Econs) (GWh)
- Número de Clientes na Rede (NCl)
- Nova Potência Instalada na Rede, em Postos de Transformação (PT) (kVA)
- Construção de Novas Linhas Aéreas (LA) (m)
- Construção de Novos Cabos Subterrâneos (CS) (m)

Já a variável de saída, é, simplesmente, o investimento obrigatório, ou seja, o IO.

As variáveis após a normalização são dadas por [2]:

Tabela 1 — Dados normalizados

Ano	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCl	PT	LA	CS	IO
2003	2	2	2	1.93	1	1	1	1	1.8	2	1.66	36.64
2004	1.87	1.9	1.92	1.77	1.18	1.04	1.27	1.1	2	1.67	2	30.1
2005	1.82	1.86	1.88	1.68	1.37	1.15	1.55	1.4	1.84	1.39	1.95	34.29
2006	1.76	1.79	1.87	1.89	1.59	1.16	1.73	1.6	1.81	1.5	1.83	25.92
2007	1.68	1.72	1.78	1.75	1.87	1.19	1.9	1.8	1.65	1.39	1.7	25.61
2008	1.64	1.55	1.51	1.8	1.97	1.15	1.85	1.8	1.81	1.29	1.23	27.25
2009	1.59	1.34	1.27	1	1.87	1.36	1.81	1.9	1.47	1.41	1.77	30.57
2010	1.48	1.29	1.24	1.51	2	1.51	2	2	1.48	1.79	1.27	27.27
2011	1.37	1.2	1.14	2	1.89	1.61	1.85	1.97	1.44	1.43	1.42	22.07
2012	1.26	1.1	1.06	1.82	1.66	1.93	1.64	1.86	1.21	1.28	1.12	18.75
2013	1.15	1.02	1.01	1.3	1.71	2	1.55	1.81	1.25	1.22	1.11	15.02
2014	1.05	1	1	1.17	1.8	1.74	1.54	1.83	1.14	1.03	1	12.07
2015	1	1.01	1.02	1.32	2	1.58	1.6	1.88	1	1	1.04	11.64
2016	0.99	1.05	1.06	1.35	2.16	1.42	1.63	1.98	1.34	1	1.04	11.34
2017	0.97	1.06	1.07	1.54	2.38	1.2	1.61	2.09	1.41	1.04	1.06	14.09

A evolução do IO (também esta transformada para efeitos de anonimização) apresenta-se a seguir:

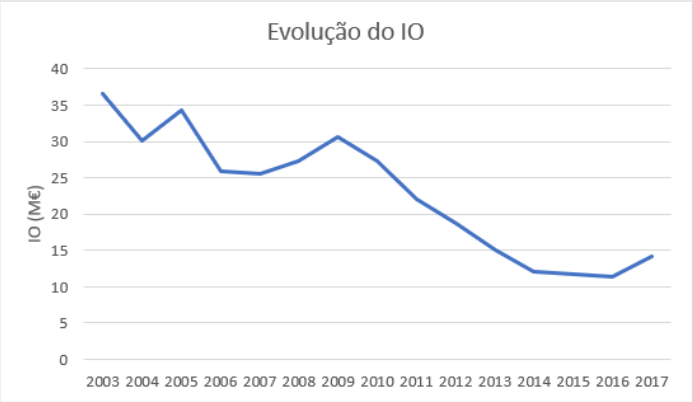

Figura 2- Evolução do IO entre 2003 e 2017

De notar que os picos de investimento, por exemplo em 2005 e 2009, acontecem em anos de eleições legislativas, algo que poderá resultar no afastamento dos modelos obtidos nestes pontos, pois as previsões das variáveis não aumentam de forma significativa para provocar uma variação brusca no IO. Claro que houve outras eleições depois destas, mas a lei mudou restringindo certas iniciativas dos autarcas, pelo que não são visíveis novos picos de IO.

2.3. Spike and Slab com recurso ao R

Para a utilização do método Spike and Slab, recorreu-se a uma implementação *Package 'BoomSpikeSlab'* [6], já disponível, no software R [7].

Para isso foi utilizada a função lm.spike, que utiliza o método abordado nesta dissertação, retornando os coeficientes integrantes da regressão linear e o somatório do erro quadrático, bem como algumas análises globais. Esta função integra um *package*, o *BoomSpikeSlab*, cuja documentação [2] se encontra disponível para download bem como a referida biblioteca.

Na função acima referida, são fornecidos os dados de entrada, o investimento obrigatório e, ainda, o número de iterações a serem realizadas pelo software.

De forma a automatizar o processo de aquisição/tratamento de dados retirados do programa e de criar uma ferramenta que possa ser utilizada no futuro, foi desenvolvido um pequeno *script* em R, tendo por base o *package* mencionado acima.

Encontra-se na figura a seguir, a explicação do código desenvolvido, o qual se encontra em anexo.

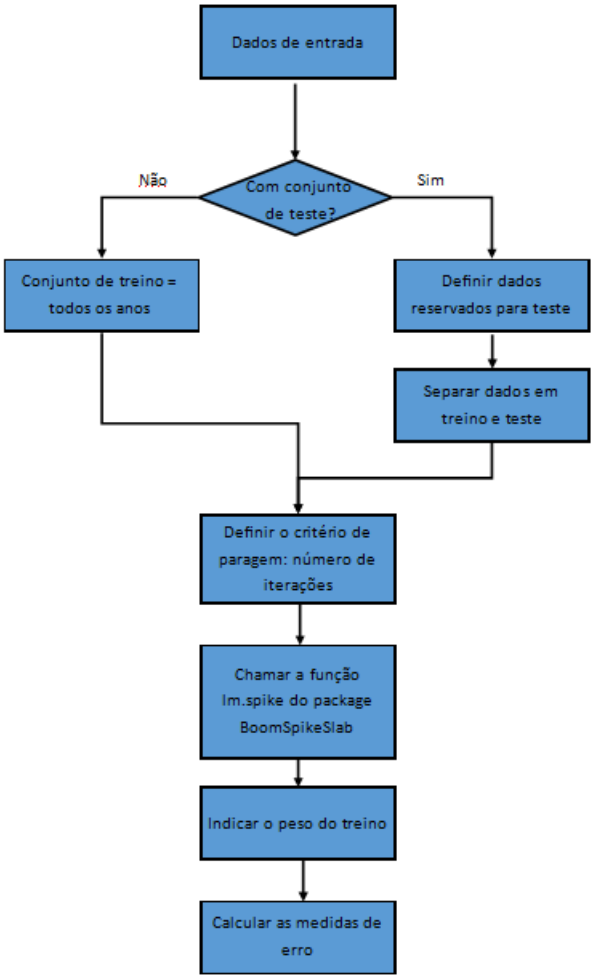

Figura 3- Fluxograma do código desenvolvido

2.3.1 Valores Médios

Numa fase inicial, estudou-se a influência de cada uma das variáveis de entrada (preditores), através do sumário gerado automaticamente pelo programa, que corresponde a medidas de erros e valores médios dos preditores calculados a partir de todos os modelos gerados. A influência das mesmas foi analisada tendo por base a média dos coeficientes a elas associado ao longo de todas as iterações do programa. Dado que, as variáveis de entrada estão normalizadas, a magnitude de cada coeficiente indicará, em princípio, a relevância do respetivo preditor.

2.3.2 Indicadores da qualidade dos modelos

Esta secção apresenta as métricas que serão usadas como critério de avaliação da qualidade dos modelos.

Nesta fase, vão ser usados todos os modelos obtidos através do software BoomSpikeSlab, ou seja, a cada iteração do programa é criado um modelo. Logo pode-se ter 500, 1000 ou 2000

modelos, pois, mais uma vez, irá ser testada a influência do número de iterações na produção de um modelo robusto e eficaz para a previsão. Além disto, mais uma vez, vai ser realizado o estudo dos diferentes cenários, isto é, modelos sem teste, com 2017 para teste e com 2016 e 2017 para teste.

Como são produzidos muitos modelos e, como visto anteriormente, a utilização não criteriosa de valores médios não produz um modelo sólido, logo é necessário escolher métodos de seleção destes mesmos modelos. No caso desta dissertação foram considerados 4 métricas de seleção dos modelos:

- Somatório dos Erros Quadráticos (SEQ);
- Erro Percentual Absoluto Médio (MAPE);
- Atribuição de Pesos aos MAPE de teste e treino
- Modelo Híbrido, que consiste em selecionar modelos através do erro absoluto anual e do SEQ, ou seja, os modelos a selecionar seriam os que tinham menor SEQ e que verificavam, simultaneamente, a condição de erro máximo, definido para cada um dos cenários, absoluto de cada ano.

De notar, ainda, que para cada um dos métodos foram selecionados 10 modelos e, que também a média dos coeficientes destes 10 modelos será alvo de análise.

Apresentam-se, a seguir, as equações utilizadas em cada uma das métricas usadas:

$$SEQ\ (IO) = \sum_{i=0}^{N} (IO_{real,i} - IO_{previsto,i})^2 \tag{4}$$

Onde:

- $IO_{real,i}$ é o IO fornecido nos dados para o ano i
- $IO_{previsto,i}$ é o IO previsto para o ano i
- N é o número de anos

$$MAPE = \frac{1}{N} \sum_{i=0}^{N} Erro_i \tag{5}$$

Onde:

- $Erro_i = \frac{|IO_{real,i} - IO_{obtido,i}|}{IO_{real,i}} * 100$
- N é o número de anos

A atribuição de pesos ao MAPE de treino e de teste é dado por:

$$Erro_{pesado} = MAPE_{teste} * \alpha + MAPE_{treino} * (1 - \alpha) \quad (6)$$

Atribuiu-se o valor de α=0,7, pois foi este o valor utilizado em dissertações anteriores.

O método híbrido é caracterizado pela escolha dos modelos que tiverem menor valor de SEQ e ao mesmo tempo não tenham um erro absoluto em cada ano superior a um determinado valor. O valor máximo do erro anual é 15%, no caso de não existir conjunto de teste, 15,5% para um ano de teste (2017) e 18% quando o conjunto de teste é composto por dois anos (2016 e 2017). Estes valores foram determinados através de experimentação, isto é, foi determinado o valor de erro máximo anual que produzia pelo menos 10 modelos utilizando os critérios acima especificados.

2.4. Obtenção da regressão linear através da média dos coeficientes dos 10 melhores modelos

De forma a agilizar o processo foi desenvolvida matematicamente a expressão da média dos 10 melhores modelos, para determinar se podia ser feita a média dos coeficientes e com esses mesmos coeficientes médios determinar o IO médio, ou se tinha de ser determinado o IO de cada modelo e só depois feita a média entre os Investimentos Obrigatórios obtidos para cada um dos modelos.

Para cada modelo, a sua regressão é dada por:

$$IO_x = a_x * IC + b_x * NNC + c_x * NNF + d_x * IHPC + e_x * PIB + f_x * Tdes + g_x * Econs + h_x * NCl + i_x * PT + j_x * LA + l_x * CS \quad (7)$$

Sendo x o modelo para o qual está a ser calculado o IO e a,b,c,d,e,f,g,h,i,j,l os coeficientes de cada modelo.

O IO médio, tendo em conta os 10 modelos é:

$$\overline{IO} = \frac{1}{N}\sum_{x=1}^{N} IO_x \quad (8)$$

Desenvolvendo,

$$\overline{IO} = \frac{1}{N}\sum_{x=1}^{N} a_x * IC + b_x * NNC + c_x * NNF + d_x * IHPC + e_x * PIB + f_x * Tdes + g_x * Econs + h_x * NCl + i_x * PT + j_x * LA + l_x * CS \quad (9)$$

$$\overline{IO} = \sum_{x=1}^{N} \frac{a_x}{N} * IC + \frac{b_x}{N} * NNC + \frac{c_x}{N} * NNF + \frac{d_x}{N} * IHPC + \frac{e_x}{N} * PIB + \frac{f_x}{N} * Tdes + \frac{g_x}{N} * Econs + \frac{h_x}{N} * NCl + \frac{i_x}{N} * PT + \frac{j_x}{N} * LA + \frac{l_x}{N} * CS \quad (10)$$

Logo e utilizando, por exemplo, o coeficiente a, associado à variável IC, tem-se:

$$\overline{IO} = \left(\frac{a_1 + a_2 + \ldots + a_N}{N}\right) * IC + \cdots \quad (11)$$

Assim sendo, fica-se com:

$$\overline{IO} = \bar{a} * IC + \bar{b} * NNC + \bar{c} * NNF + \bar{d} * IHPC + \bar{e} * PIB + \bar{f} * Tdes + \bar{g} * Econs + \bar{h} * NCl + \bar{\imath} * PT + \bar{\jmath} * LA + \bar{l} * CS \quad (12)$$

Tal como demonstrado acima, é indiferente fazer a média dos coeficientes e calcular assim o IO ou calcular o IO para cada modelo e depois fazer a média, pelo que irá ser utilizada a primeira, tornando assim o processo mais rápido.

Capítulo 3

Resultados

3.1. Valores médios dos preditores

Encontram-se a seguir os resultados obtidos, utilizando os valores médios dos preditores, sendo que neste subcapítulo são utilizados todos os modelos gerados.

Tabela 2 — Resultados obtidos com valores médios dos preditores

	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCI	PT	LA	CS	b
Sem teste	20.5	0.07	-1.15	-0.46	-0.14	-0.63	-0.13	-0.11	-0.44	0.11	0.05	-2.20
Teste 2017	21.4	-0.05	-1.32	-0.65	-0.17	-0.46	-0.03	-0.20	-0.94	0.14	0.20	-2.58
Teste 2016 e 2017	17.3	1.15	-0.43	-0.32	-0.05	-0.21	-0.06	-0.09	-0.31	0.29	0.22	-1.70

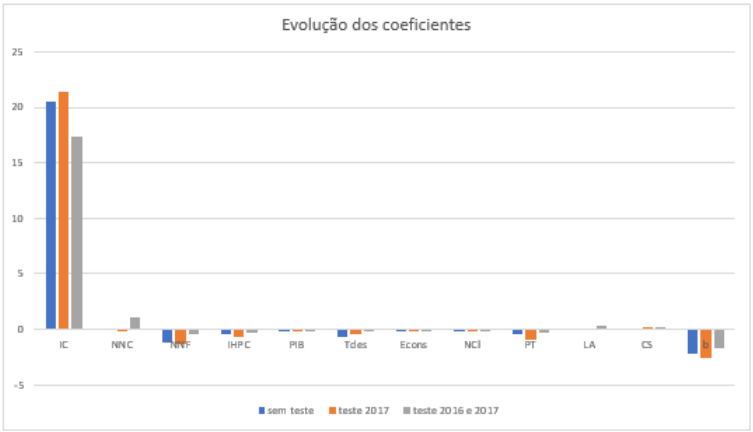

Figura 4 – Coeficientes de regressão

Tabela 3 — Desvio padrão dos coeficientes

	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCl	PT	LA	CS	b
Sem teste	7.12	5.7	4.7	1.68	0.76	1.64	0.84	0.73	2.48	1.13	1.29	4.49
Teste 2017	9.3	5.58	5.14	2.1	0.933	1.48	1.09	1.09	3.81	1.85	0.2	5.18
Teste 2016 e 2017	9.55	6.43	4.42	1.53	0.527	0.945	0.702	0.822	3.88	2.18	2.1	4.69

Tabela 4 — Desvio padrão dos coeficientes, em percentagem, dos seus valores médios

	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCl	PT	LA	CS	b
Sem teste	34.73	7744.57	408.70	366.01	560.00	259.49	623.88	643.86	559.82	1076.19	2354.01	204.09
Teste 2017	43.46	10689.66	389.39	325.58	536.21	325.27	3353.85	553.30	406.18	1360.29	100.00	200.78
Teste 2016 e 2017	55.20	559.13	1032.71	476.64	1084.36	458.74	1181.82	923.60	1247.59	756.94	967.74	275.88

A partir dos resultados acima percebe-se que o IC é a variável com maior coeficiente atribuído, o que significa que é a variável mais importante de todos os dados de entrada. Além disto, tem, também, a menor variância percentual, ou seja, apesar do seu valor mudar bastante, em módulo, ao longo das iterações, esta variação é pequena, percentualmente, quando comparada com todas as outras variações percentuais.

De notar ainda que, a presença da ordenada na origem, bem como coeficientes consideráveis do NNF, na sua globalidade, e do NNC, principalmente quando é retirada a informação de 2016 e 2017 aos dados de entrada.

A utilização dos valores médios permite, também, determinar que algumas variáveis, tais como, o LA, o CS, Econs, PIB, NCl têm uma baixa influência no modelo de previsão.

Com base nos coeficientes, apresentados na tabela 2, constroem-se os seguintes modelos:

Sem teste

$$IO = 20.5 * IC + 0.0736 * NNC - 1.15 * NNF - 0.459 * IHPC - 0.135 * PIB - 0.632 * TDes - 0.134 * Econs - 0.114 * NCl - 0.443 * PT + 0.105 * LA + 0.0548 * CS - 2.2$$

Teste 2017

$$IO = 21.4 * IC - 0.0522 * NNC - 1.32 * NNF - 0.645 * IHPC - 0.174 * PIB - 0.455 * TDes - 0.0325 * Econs - 0.197 * NCl - 0.938 * PT + 0.136 * LA + 0.202 * CS - 2.58$$

Teste 2016 e 2017

$$IO = 17.3 * IC + 1.15 * NNC - 0.428 * NNF - 0.321 * IHPC - 0.0486 * PIB - 0.206 * TDes - 0.0594 * Econs - 0.089 * NCl - 0.311 * PT + 0.288 * LA + 0.217 * CS - 1.7$$

Tabela 5 — IO obtido com valores médios dos preditores para os 3 casos

	Sem teste		Teste 2017		Teste 2016 e 2017	
Ano	**IO obtido**	**Erro (%)**	**IO obtido**	**Erro (%)**	**IO obtido**	**Erro (%)**
2003	34.25	6.52	34.29	6.41	33.70	8.03
2004	31.54	4.79	31.48	4.59	31.29	3.97
2005	30.47	11.13	30.47	11.13	30.31	11.60
2006	29.09	12.22	29.00	11.89	29.10	12.28
2007	27.56	7.61	27.51	7.40	27.67	8.03
2008	26.93	1.19	26.73	1.92	26.70	2.01
2009	26.60	13.00	26.85	12.18	26.17	14.40
2010	24.00	12.00	24.03	11.88	24.00	12.00
2011	21.59	2.18	21.50	2.58	21.81	1.16
2012	19.44	3.68	19.43	3.64	19.82	5.70
2013	17.42	15.97	17.41	15.90	17.97	19.66
2014	15.61	29.34	15.52	28.59	16.27	34.80
2015	14.62	25.58	14.49	24.47	15.42	32.46
2016	14.27	25.83	13.91	22.62	15.17	33.81
2017	13.83	1.81	13.33	5.42	14.79	5.00

Tabela 6 — Erros MAPE com valores médios dos preditores

	Sem teste		Teste 2017		Teste 2016 e 2017	
	Erro treino (%)	**Erro teste (%)**	**Erro treino (%)**	**Erro teste (%)**	**Erro treino (%)**	**Erro teste (%)**
Média	11.52	-	11.80	5.42	12.78	19.40

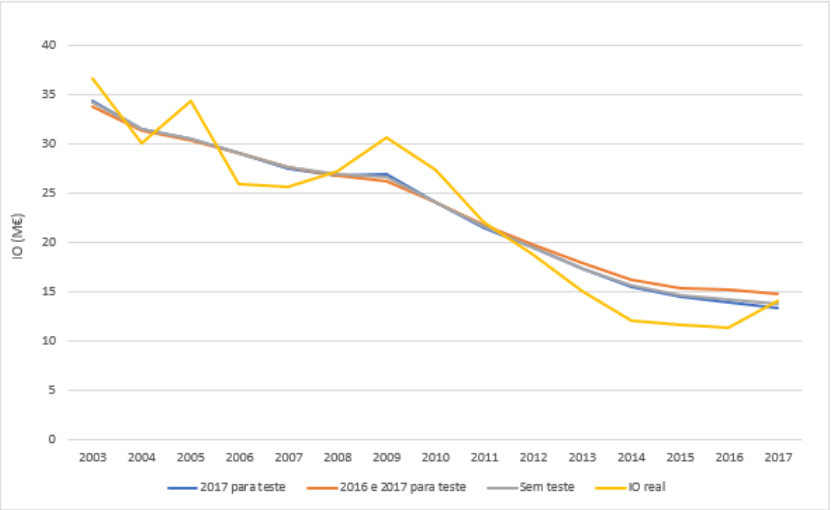

Figura 5 – Evolução das curvas do IO com valores médios dos preditores

Com valores médios dos preditores temos um aumento dos erros à medida que vamos adicionando mais anos ao teste. O objetivo desta mesma introdução de anos de teste é tornar o modelo mais robusto e ter uma forma de verificar a eficácia do mesmo, para situações em que, o mesmo, não teve adaptação aos dados utilizados no teste.

No entanto, mesmo sem anos reservados a teste obtêm-se um fraco acompanhamento da curva do IO real pelas curvas associadas aos modelos dos diferentes cenários, sendo que, se conclui que os modelos de previsão baseados em valores médios dos coeficientes produzem modelos pouco robustos, quando são usados todos os modelos gerados.

3.2. Métricas de seleção de modelos

Como são produzidos muitos modelos,(500, 1000 ou 2000) e, como visto anteriormente, a utilização de valores médios com todos estes não produz um modelo sólido, é necessário escolher métodos de seleção destes mesmos modelos, sendo que para isso vão ser utilizadas 4 métricas diferentes.

3.2.1. Critério SEQ

Neste subcapítulo vai ser abordada a escolha de 10 modelos pelo SEQ, contemplando os três cenários, ou seja, sem teste, com um ano para teste e com dois anos para teste.

Encontram-se, na tabela abaixo, a média dos coeficientes associados ao diferente número de iterações, sem conjunto de teste.

Como os 10 modelos produzem curvas muito semelhantes, irá ser considerada a média entre esses 10 modelos para apresentação de resultados e conclusões seguintes.

Tabela 7 — Média dos coeficientes e número de vezes que são usados ao longo dos 10 modelos para o diferente número de iterações sem conjunto de teste através do critério SEQ

		b	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCl	PT	LA	CS
500 iterações	Média	0	33.51	-0.18	-12.97	0.03	0	-5.14	0	0	0	0	0
	Nº de vezes usado	0	10	2	9	1	0	10	0	0	0	0	0
1000 iterações	Média	0	32.56	4.12	-16.60	0	0	-5.04	0	0	0	0.30	0
	Nº de vezes usado	0	10	2	10	0	0	10	0	0	0	1	0
2000 iterações	Média	0	33.83	4.12	-17.36	0	0	-5.37	0	0	0	0	0
	Nº de vezes usado	0	10	2	10	0	0	10	0	0	0	0	0

A partir da tabela cima é possível determinar as regressões lineares para cada número de iterações sendo as mesmas dados por:

- $IO = 33.51 * IC - 0.18 * NNC - 12.97 * NNF + 0.03 * IHPC - 5.14 * TDes$, para 500 iterações,
- $IO = 32.55 * IC + 4.12 * NNC - 16.60 * NNF - 5.04 * TDes + 0.31 * LA$, para 1000 iterações,
- $IO = 33.83 * IC + 4.11 * NNC - 17.36 * NNF - 5.37 * Tdes$, para 2000 iterações

Sendo o IO dado pela multiplicação da matriz dos dados de entrada pela matriz dos coeficientes do modelo, o mesmo é, tendo em conta a média dos coeficientes:

Tabela 8 — Investimentos obrigatórios anuais para a média dos 10 melhores modelos para o diferente número de iterações sem conjunto de teste através do critério SEQ

	IO		
Ano	**500 iterações**	**1000 iterações**	**2000 iterações**
2003	35.64	35.70	35.79
2004	32.13	32.08	32.16
2005	30.41	30.32	30.41
2006	28.50	28.22	28.21
2007	26.84	26.64	26.62
2008	29.24	29.29	29.47
2009	29.61	29.76	29.95
2010	25.57	25.83	25.74
2011	22.69	22.93	22.85
2012	18.41	18.60	18.38
2013	15.01	15.15	14.83
2014	13.13	13.23	12.93
2015	12.02	12.11	11.79
2016	11.98	12.09	11.78
2017	12.32	12.44	12.16

Tabela 9 — Comparação do SEQ para o diferente número de iterações sem conjunto de teste

	500 iterações	**1000 iterações**	**2000 iterações**
SEQ	41.42	39.51	39.41

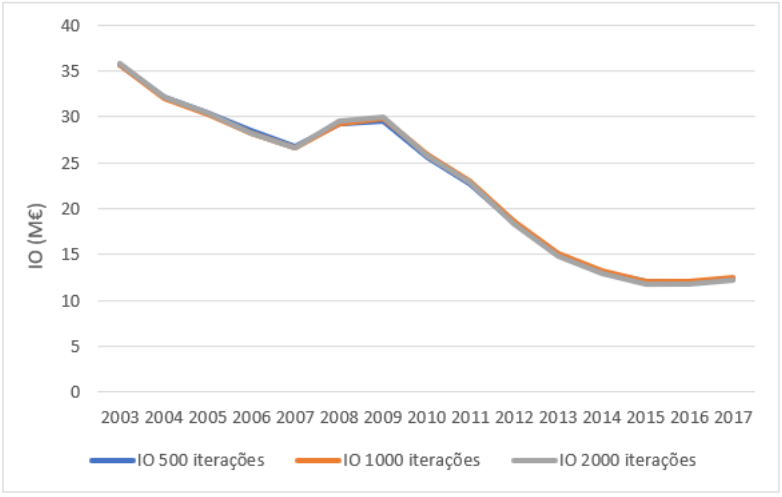

Figura 6 – Comparação das curvas do IO obtido para o diferente número de iterações sem conjunto de teste através do critério SEQ

Como demonstrado acima, o número de iterações praticamente não tem influência no modelo, pelo que, a partir daqui se vai considerar o número de iterações padrão da implementação, ou seja, 1000 iterações.

O gráfico acima demonstra que, sem conjunto de teste, este produz modelos de boa robustez.

Tal como mencionado acima, irá, a partir de agora, ser usado o valor padrão de 1000 iterações para a obtenção dos 10 melhores modelos, pois não se verifica mudança na tendência dos modelos com o aumento ou diminuição do número de iterações. Irá proceder-se então ao teste da influência dos anos de teste. Os anos de teste prendem-se com o controlo do método, ou seja, os dados relativos a esses anos são retirados dos dados de entrada do modelo, sendo, depois, estes mesmos anos usados para perceber o comportamento do modelo a dados para os quais não teve adaptação.

Tabela 10 — Média dos coeficientes e número de vezes que são usados ao longo dos 10 modelos para dois casos de conjunto de teste através do critério SEQ

		b	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCI	PT	LA	CS
Teste 2017	Média	0	34.97	0	-14.75	0	0	-5.13	-0.06	0	0	0	0
	Nº de vezes usado	0	10	0	10	0	0	10	1	0	0	0	0

Teste 2016 e 2017	**Média**	-1.3	36.88	-1.3	-14.78	0	0	-4.76	0	-0.18	0	0.08	0
	Nº de vezes usado	1	10	3	8	0	0	9	0	1	0	1	0

Sendo as suas regressões lineares dadas por:

$IO = 34.97 * IC - 14.75 * NNF - 5.13 * Tdes - 0.06 * Econs$, para 2017 como teste

$IO = 36.88 * IC - 1.3 * NNC - 14.78 * NNF - 4.76 * Tdes - 0.18 * NCl - 0.08 * PT - 1.3,$ para 2016 e 2017 como teste

O que resulta num investimento obrigatório:

Tabela 11 — Investimentos obrigatórios anuais para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério SEQ

	IO	
Ano	**Teste 2017**	**Teste 2016 e 2017**
2003	35.64	35.70
2004	32.13	32.08
2005	30.41	30.32
2006	28.50	28.22
2007	26.84	26.64
2008	29.24	29.29
2009	29.61	29.76
2010	25.57	25.83
2011	22.69	22.93
2012	18.41	18.60
2013	15.01	15.15
2014	13.13	13.23
2015	12.02	12.11
2016	11.98	12.09
2017	12.32	12.44

Tabela 12 — Erros de treino e teste para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério SEQ

	Teste 2017	Teste 2016 e 2017
Erro treino (%)	4.49	4.48
Erro teste (%)	15.58	10.78

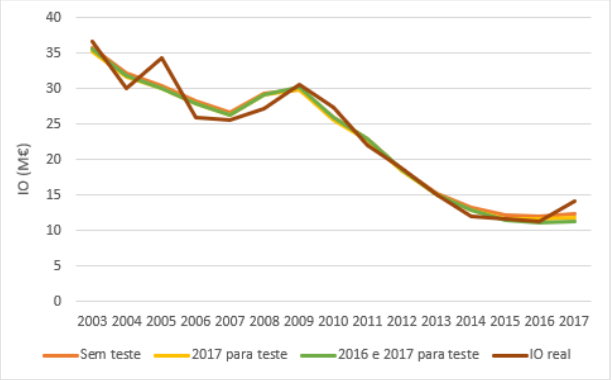

Figura 7 – Comparação entre o IO real e a média dos 10 melhores modelos para o diferente número de anos de teste através do critério SEQ

Com dois anos de teste, verifica-se que este consegue acompanhar na mesma o ano de 2016, pois não existe grande flutuação do IO real, mas em 2017 existe uma quebra da rota descendente do IO o modelo não foi capaz de acompanhar, algo que é, em parte, explicado pela pequena variação dos elementos de entrada, que não conduziu a uma variação significativa do IO.

No cômputo geral, parece um método sólido, pois consegue acompanhar bem o IO real, exceto quando existem grandes flutuações inexplicáveis no mesmo.

De notar, através da figura 7, que as curvas estão quase sobrepostas, o que demonstra que o modelo se revela eficiente quer tenha anos de teste, quer não os tenha.

3.2.2. Critério MAPE

Esta secção apresenta os resultados obtidos com o critério MAPE.

Tabela 13 — Média dos coeficientes e número de vezes que são usados ao longo dos 10 modelos para o diferente número de anos de teste através do critério MAPE

		b	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCl	PT	LA	CS
Sem teste	Média	0	35.25	-7.59	-8.25	0	-0.04	-4.94	0	0	0	0.73	0
	Nº de vezes usado	0	10	4	6	0	1	10	0	0	0	2	0
Teste 2017	Média	0	36.44	-3.87	-12.40	0	0	-5.16	-0.06	0	0	0	0
	Nº de vezes usado	0	10	2	8	0	0	10	1	0	0	0	0
Teste 2016 e 2017	Média	-1.08	37.95	-8.44	-9.20	0	0	-4.39	0	-0.18	0	0.08	0
	Nº de vezes usado	1	10	4	6	0	0	9	0	1	0	1	0

O que resulta nos modelos:

- $IO = 35.25 * IC - 7.59 * NNC - 8.25 * NNF - 0.04 * PIB - 4.94 * Tdes - 0.73 * LA$, sem conjunto de teste
- $IO = 36.44 * IC - 3.87 * NNC - 12.40 * NNF - 5.16 * Tdes - 0.06 * Econs$, para 2017 como teste
- $IO = 37.95 * IC - 8.44 * NNC - 9.20 * NNF - 4.39 * Tdes - 0.18 * NCl + 0.08 * LA - 1.08$, para 2016 e 2017 como teste

Dando origem a um IO:

Tabela 14 – Investimentos obrigatórios anuais para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério MAPE

	IO		
Ano	**Sem conjunto de teste**	**Teste 2017**	**Teste 2016 e 2017**

2003	35.30	35.12	35.14
2004	31.69	31.54	31.57
2005	29.80	29.78	29.81
2006	28.32	27.93	28.15
2007	26.54	26.24	26.36
2008	28.77	28.99	28.92
2009	29.63	29.88	30.08
2010	25.92	25.65	25.96
2011	22.80	22.73	23.00
2012	18.66	18.46	19.01
2013	15.41	15.02	15.67
2014	13.26	12.92	13.25
2015	12.01	11.64	11.78
2016	11.81	11.45	11.38
2017	12.05	11.69	11.39

Tabela 15 — Erros de treino e teste para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério MAPE

	Sem conjunto de teste	**Teste 2017**	**Teste 2016 e 2017**
Erro treino (%)	5.77	4.25	5.15
Erro teste (%)	-	17.03	9.74

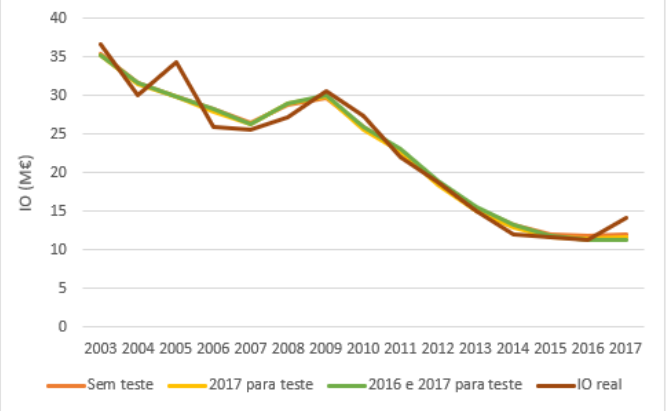

Figura 8 – Comparação entre o IO real e a média dos 10 melhores modelos para o diferente número de anos de teste através do critério MAPE

O comportamento da curva pela seleção pelo MAPE com um ano de teste, não se alterou muito, pelo que, conclui-se que o método se mantém robusto para um ano de teste, notando-se, tal como no caso anterior, um pequeno desvio na estimativa de 2017.

Mesmo retirando dois anos dos dados de entrada, o modelo continua com um bom desempenho.

De um modo geral é um método de seleção que permite, mesmo com a introdução de anos para teste, a obtenção de modelos que têm erros associados aceitáveis.

É também de destacar a coerência do modelo, isto é, independentemente dos anos reservados para teste obtêm-se modelos muito semelhantes, tal como, demonstrado na figura 8.

3.2.3. Critério de pesos aplicados ao MAPE de treino e de teste

Apresentam-se, nesta secção, os resultados obtidos pelo critério de pesos aplicados ao MAPE de treino e de teste, tal como exposto na secção 2.3.2. O peso aplicado a MAPE de treino foi de 0,3, sendo que, por isso, o peso aplicado ao MAPE de teste foi de 0,7.

Tabela 16 — Média dos coeficientes e número de vezes que são usados ao longo dos 10 modelos para o diferente número de anos de teste através do critério MAPE

		b	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCI	PT	LA	CS
Teste 2017	Média	-0.51	21.37	2.75	-4.69	0	-0.14	-3.16	-0.31	0	0	0.53	0
	Nº de vezes usado	1	10	1	3	0	1	7	1	0	0	1	0
Teste 2016 e 2017	Média	-2.21	32.77	-3.26	-9.96	0.13	0.49	-4.23	0	0	0	0	0.63
	Nº de vezes usado	1	10	2	7	1	1	9	0	0	0	0	1

Sendo as regressões lineares dadas por:

$IO = 21.37 * IC + 2.75 * NNC - 4.69 * NNF - 0.14 * PIB - 3.16 * Tdes - 0.31 * Econs + 0.53 * LA - 0.51$, para 2017 como teste

$IO = 32.77 * IC - 3.26 * NNC - 9.96 * NNF + 0.13 * IHPC + 0.49 * PIB - 4.23 * Tdes + 0.63 * CS - 2.21$, para 2016 e 2017 como teste

Resultando num IO anual:

Tabela 17 — Investimentos obrigatórios anuais para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério de pesos aplicados ao MAPE de treino e de teste

	IO	
Ano	**Teste 2017**	**Teste 2016 e 2017**
2003	35.82	34.44
2004	32.73	31.42
2005	31.12	29.89
2006	29.64	28.27
2007	27.91	26.69
2008	27.93	28.55
2009	26.83	29.29
2010	24.14	25.32
2011	21.56	22.69
2012	18.32	18.53
2013	15.75	15.34
2014	14.31	13.29
2015	13.62	12.24
2016	13.81	12.14
2017	14.05	12.43

Tabela 18 — Erros de treino e teste para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério pesos aplicados ao MAPE de treino e de teste

	Teste 2017	**Teste 2016 e 2017**
Erro treino (%)	**9.75**	**5.68**
Erro teste (%)	**0.28**	**9.42**

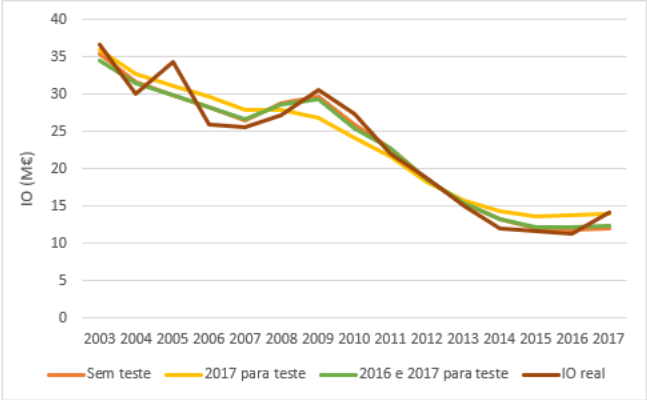

Figura 9 – Comparação entre o IO real e a média dos 10 melhores modelos para o diferente número de anos de teste através do critério pesos aplicados ao MAPE de treino e de teste

De notar que, sem conjunto de teste, a escolha por este método, é igual à feita pelo MAPE.

Num cômputo geral, o método de seleção é bastante eficaz, produzindo erros muito aceitáveis.

De realçar que, para o caso de apenas existir um ano para teste, esta seleção produz um IO de teste com um erro bastante baixo, mas acompanha o modelo de forma mais deficitária, algo comprovado pelo aumento dos erros de treino associados a cada um dos cenários.

Já os cenários sem teste e com dois anos para teste comportam-se de forma idêntica, vacilando sempre que o modelo tem pontos que se afastam da regressão linear, o que pela natureza do modelo, o impede de seguir corretamente a realidade nesses casos.

3.2.4. Critério híbrido

O critério híbrido consiste em selecionar modelos através do erro absoluto anual e do SEQ, ou seja, os modelos a selecionar seriam os que tinham menor SEQ e que verificavam, simultaneamente, a condição de erro máximo, definido para cada um dos cenários, absoluto de cada ano.

Tabela 19 — Média dos coeficientes e número de vezes que são usados ao longo dos 10 modelos para o diferente número de anos de teste através do critério híbrido

		b	IC	NNC	NNF	IHPC	PIB	Tdes	Econs	NCI	PT	LA	CS
Sem	Média	0	32.82	-8.86	-3.81	0	0	-5.12	0	0	0	0.47	0

	Nº de vezes usado	0	10	6	4	0	0	10	0	0	0	1	0
Teste 2017	**Média**	0	33.64	-1.68	-11.90	0	0	-4.88	0	0	0	0	0
	Nº de vezes usado	0	10	1	9	0	0	10	0	0	0	0	0
Teste 2016 e 2017	**Média**	-4.84	26.18	-1.91	-5.38	0.13	0	-2.74	0	0	0	2.02	0.63
	Nº de vezes usado	4	10	1	5	1	0	6	0	0	0	3	1

Sendo as regressões lineares dadas, respetivamente por:

- $IO = 32.82 * IC - 8.86 \ * NNC - 3.81 * NNF - 5.12 * Tdes + 0.47 * LA$, sem conjunto de teste
- $IO = 32.64 * IC - 1.68 \ * NNC - 11.90 * NNF - 4.88 * Tdes$, para 2017 como teste
- $IO = 26.18 * IC - 1.91 \ * NNC - 5.38 * NNF + 0.13 * IHPC - 2.74 * Tdes + 2.02 * LA + 0.63 * CS - 4.84$, para 2016 e 2017 como teste

Logo, o investimento obrigatório anual é:

Tabela 20 — Investimentos obrigatórios anuais para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério híbrido

	IO		
Ano	**Sem conjunto de teste**	**Teste 2017**	**Teste 2016 e 2017**
2003	36.14	35.23	35.53
2004	32.70	31.78	32.17
2005	30.87	30.11	30.24
2006	29.56	28.28	29.00
2007	27.69	26.63	27.12
2008	29.07	28.98	27.47

2009	29.19	29.48	27.76
2010	25.54	25.49	25.24
2011	22.43	22.64	22.22
2012	18.30	18.50	18.57
2013	15.20	15.18	15.73
2014	13.38	13.24	13.45
2015	12.38	12.09	12.43
2016	12.36	11.99	12.32
2017	12.72	12.25	12.45

Tabela 21 — Erros de treino e teste para a média dos 10 melhores modelos para o diferente número de anos de teste através do critério híbrido

	Sem conjunto de teste	**Teste 2017**	**Teste 2016 e 2017**
Erro treino (%)	6.72	**5.39**	**6.27**
Erro teste (%)	-	**13.03**	**10.16**

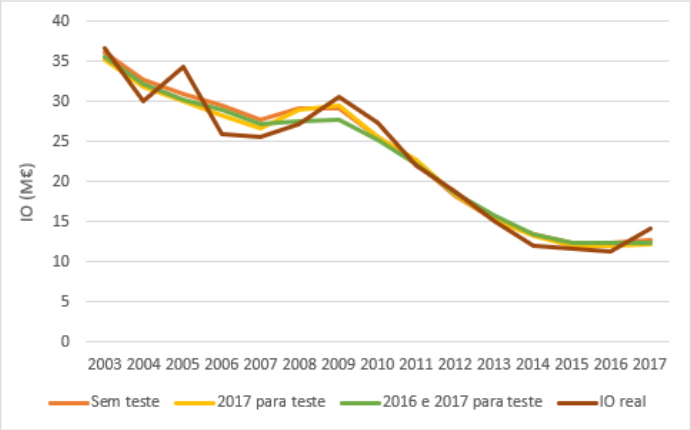

Figura 10 – Comparação entre o IO real e a média dos 10 melhores modelos para o diferente número de anos de teste através do critério híbrido

Tal como no método dos pesos, este método comporta-se de forma eficaz em todos os cenários estudados, pois tem erros relativamente constantes e limitados ao longo dos cenários que lhe vão sendo colocados.

É também percetível, através da figura 10, que quando são adicionados dois anos de teste existe uma flutuação grande do IO real que não é tão bem acompanhada pelo modelo, comparativamente quer com 1 ano de teste quer sem qualquer ano para teste.

Apesar do ponto acima referido, todos os cenários têm comportamentos bastante idênticos no seguimento do IO real, estando, como seria de esperar, limitados aquando do afastamento do IO real da sua tendência.

3.3. Comparação entre todos os critérios estudados

Irá, agora, proceder-se à análise de todos os métodos, tendo em conta o erro MAPE associado ao treino e ao teste de cada, para se poder definir qual o método mais eficaz de seleção de modelos obtidos através de iterações do Spike and Slab.

Esta comparação vai ser realizada para o número de iterações padrão, ou seja, 1000 iterações e vai ter em conta a média dos 10 melhores modelos obtidos para cada método e para cada cenário.

Tabela 22 — Comparação entre todos os métodos estudados

	Sem teste				**2017 para teste**				**2016 e 2017 para teste**			
	SEQ	**MAPE**	**Pesos**	**Híbrido**	**SEQ**	**MAPE**	**Pesos**	**Híbrido**	**SEQ**	**MAPE**	**Pesos**	**Híbrido**
Erro de treino (%)	5.78	5.77	5.77	6.72	4.49	4.25	9.75	5.39	4.48	5.15	5.68	6.27
Erro de teste (%)	-	-	-	-	15.58	17.03	0.28	13.03	10.78	9.74	9.42	10.16

Tabela 23 — Comparação entre os erros médios de todos os métodos estudados

	SEQ	**MAPE**	**Pesos**	**Híbrido**
Erro médio de treino (%)	4.92	5.06	7.07	6.13
Erro médio de teste (%)	13.18	13.39	4.85	11.60

A partir das duas tabelas acima, decidiu-se escolher, os métodos SEQ e Pesos para a derradeira análise, pois são estes modelos que têm ou o melhor erro de treino, ou o melhor erro de teste.

3.4. Modelos Finais

Tal como já concluído acima, irá ser utilizado quer o método SEQ, tal como na dissertação que serve de base para comparação [2], quer o método Pesos, sendo que o estudo vai ser realizado para dois anos de teste.

Mas desta feita irá ser utilizada apenas a média dos 4 melhores modelos, tal como feito na dissertação [2].

Considerando, primeiro, o método dos SEQ:

Tabela 24 — IO obtido pelos melhores 4 modelos recorrendo aos métodos SEQ e Pesos com 2 anos para teste

	IO	
Ano	**SEQ**	**Pesos**
2003	35.58	35.66
2004	31.99	32.05
2005	30.26	30.32
2006	28.25	28.29
2007	26.62	26.66
2008	29.39	29.54
2009	30.09	30.33
2010	25.91	26.12
2011	23.01	23.22
2012	18.68	18.90
2013	15.20	15.40
2014	13.18	13.32
2015	11.95	12.05
2016	11.83	11.90
2017	12.11	12.15

Tabela 25 — Erros de treino e teste para a média dos melhores 4 modelos recorrendo aos métodos SEQ e Pesos com 2 anos para teste

	SEQ	**Pesos**
Erro treino (%)	5.07	5.37

Erro teste (%)	9.19	9.34

A escolha de apenas 4 modelos faz cair o erro de teste em ambos os modelos, mantendo o SEQ a vantagem, mesmo que pequena, quando existem dois anos para teste. Já o erro de treino subiu no método SEQ e baixou no método pesos, mais uma vez com uma ligeira vantagem para o método SEQ.

De realçar também que em ambos os critérios só selecionam 3 variáveis e que são as mesmas em todos os casos, o que indica que estas variáveis são determinantes na produção de modelos de estimação de investimentos obrigatórios.

Posto isto, assume-se como modelo final a média dos 4 melhores modelos obtidos pelo método SEQ com 2 anos de teste, sendo a sua regressão linear dada por:

$$IO = 35.07 * IC - 14.72 * NNF - 5.13 * Tdes$$

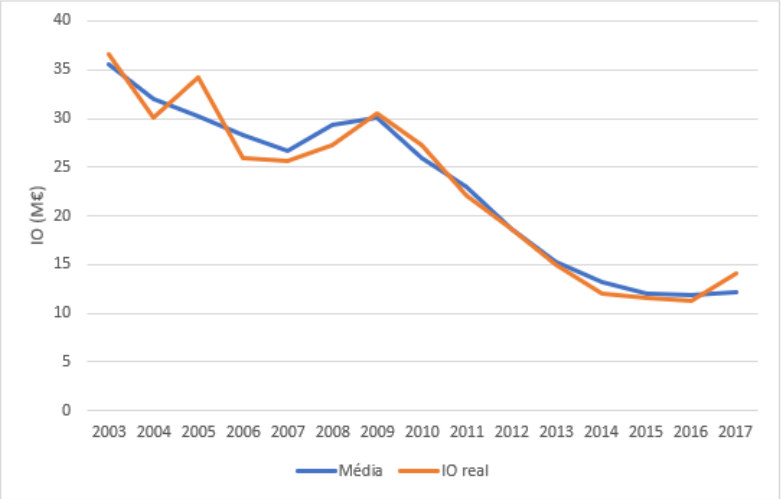

Figura 11 - Comparação entre o IO real e a média dos 4 melhores modelos recorrendo ao método SEQ com 2 anos para teste

3.5. Determinação de margens de confiança das previsões

Dado que o Spike and Slab é uma técnica probabilística é relevante calcular o intervalo de confiança a 95%, isto, representar o investimento obrigatório como um intervalo, que será válido em 95% das situações. Tal como mostra a figura a seguir, este intervalo é dado pelo valor médio (μ) ± 1,96*desvio padrão (σ).

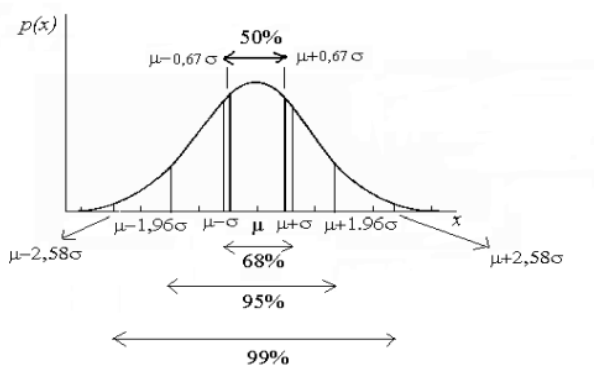

Figura 12 - Representação da distribuição normal com vários intervalos de confiança [14]

De notar que este cálculo foi feito para o método SEQ com dois anos de teste.

Tabela 26 -Intervalos do IO com um intervalo de confiança a 95%

Ano	Intervalo de confiança a 95%
2003	[33.55 ; 37.78]
2004	[30.60 ; 34.05]
2005	[29.98 ; 32.86]
2006	[27.24 ; 31.55]
2007	[25.28 ; 29.80]
2008	[25.81 ; 29.43]
2009	[24.09 ; 30.82]
2010	[22.38 ; 26.65]
2011	[19.81 ; 23.69]
2012	[17.18 ; 20.60]
2013	[14.41 ; 17.83]
2014	[11.98 ; 15.23]
2015	[10.25 ; 14.29]
2016	[9.50 ; 14.16]
2017	[8.65 ; 14.14]

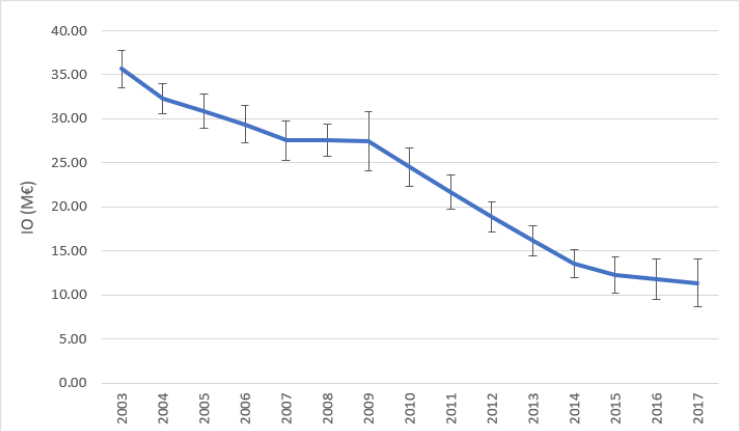

Figura 13 -IO com um intervalo de confiança a 95%

3.6. Comparação dos resultados obtidos com os obtidos pelo LASSO

Irá agora ser apresentada a tabela de resultados, realizada na dissertação que procede esta. [2].

Tabela 27 -Desempenho obtido no trabalho reportado em [2]

Erro	Treino	Teste
3.2	10.00	13.22
3.4	9.68	12.67
Atual	7.41	9.09

De notar que o modelo "Atual" é o modelo utilizado atualmente para a previsão do IO, e que foi desenvolvido no âmbito de um projeto de colaboração entre o INESC TEC e a EDP Distribuição. [2]

De forma a comparar metodologias, só podemos usar resultados obtidos utilizando todas as variáveis e com dois anos para teste.

A partir da tabela 23, percebe-se, que o Spike and Slab tem menores erros quer de treino, quer de teste que o LASSO, qualquer que seja o método escolhido, mesmo sendo apresentada, na mesma, a média dos 10 melhores modelos e não a média dos 4 melhores.

Já de acordo com a tabela 27, compreende-se que o modelo "Atual" e o modelo realizado pelo Spike and Slab são muito idênticos, existindo vantagem para este último no treino e ligeira desvantagem no teste, quando utilizados os 4 melhores modelos.

3.7. Previsões

Partindo do modelo final obtido no capítulo 3 e de dados fornecidos, à posteriori, relativos aos anos entre 2018 e 2024, foram feitas previsões de investimentos obrigatórios.

Mais uma vez, os dados, foram normalizados de igual forma, à normalização feita inicialmente para os dados até 2017, dada a confidencialidade dos mesmos.

Tabela 28 – Dados entre 2018 e 2024 normalizados

Ano	IC	NNF	Tdes
2018	1.05	1.08	1.18
2019	1.07	1.09	1.11
2020	1.08	1.09	1.09
2021	1.09	1.09	1.04
2022	1.09	1.09	1.04
2023	1.09	1.10	1.01
2024	1.10	1.10	1.00

Apresentam-se a seguir, quer o IO previsto entre 2018 e 2024, quer os valores máximos e mínimos com um intervalo de confiança a 95%.

Tabela 29 – Previsão do IO entre 2018 e 2024 e intervalo de confiança a 95%

Ano	IO	Intervalo de confiança a 95%
2018	13.68	[11.37 ; 15.98]
2019	14.22	[11.85 ; 16.60]
2020	14.57	[12.18 ; 16.96]
2021	14.65	[12.15 ; 17.16]
2022	14.73	[12.20 ; 17.25]
2023	14.86	[12.24 ; 17.47]
2024	15.11	[12.47 ; 17.76]

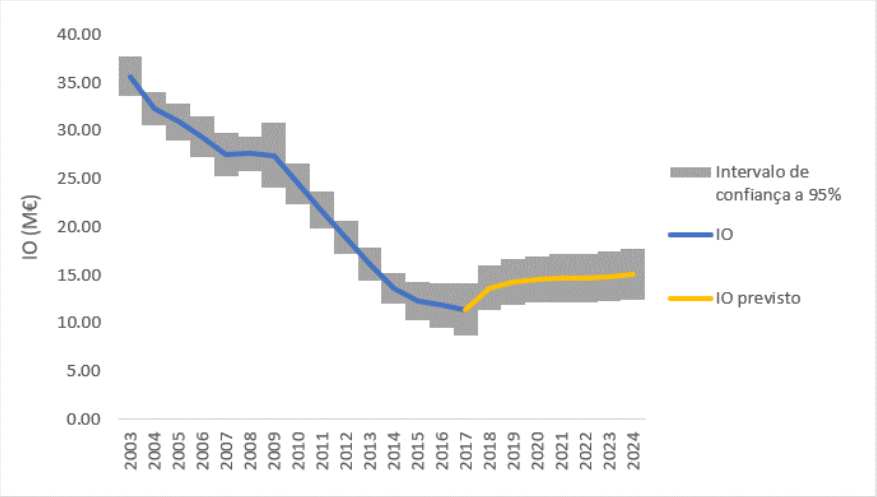

Figura 14 - Evolução do IO entre 2003 e 2017, previsões de IO e intervalo de confiança a 95% entre 2018 e 2024

De notar que, o intervalo de confiança a 95%, calculado nas tabelas 26 e 29, foi calculado utilizando os 100 melhores modelos, de forma a serem obtidos resultados com significado estatístico. Deste modo (com mais modelos), a margem de confiança obtida é mais credível e robusta.

Segundo o modelo determinado neste estudo, prevê-se um crescimento do IO ao longo dos 5 anos seguintes, sendo que só de 2018 para 2019 o crescimento será praticamente de 1 milhão de euros. Nos restantes anos o investimento irá variar entre valores na casa das centenas de milhares de euros.

É também de realçar que, todas as estimações das variáveis de entrada do problema foram calculadas no período de pré-pandemia, logo a sua eventual influência não se encontra descrita nas mesmas.

Capítulo 4

Conclusões e desenvolvimentos futuros

Este projeto tinha como objetivo último a produção de modelos para estimar o investimento obrigatório, algo que foi, sem dúvida, alcançado.

O Spike and Slab mostrou-se uma técnica robusta, principalmente no que diz respeito à seleção de variáveis, ou seja, permite conhecer quais são as variáveis que realmente têm interesse, das que são redundantes e, por isso, facilmente descartáveis num contexto de agilização do processo. No entanto, o melhor modelo foi obtido para apenas 3 variáveis, sendo que duas delas assumem um valor negativo, algo que no caso da taxa de desemprego é lógico, mas que no caso do NNF só explicado pelo facto de este ser correlacionado com o IC. Outra conclusão interessante prende-se com o facto de NNC e NNF nunca coexistirem em qualquer modelo, algo que também pode ser justificado pela correlação entre estas variáveis. Assim, quando o Spike and Slab seleciona uma destas variáveis, já não precisa de incluir a outra que terá uma informação semelhante.

Apesar de se terem obtido modelos com erros semelhantes ao modelo atualmente usado pela E-REDES, tornou-se evidente que a seleção dos mesmos tem influência no final, pois, a mesma, pode tornar este método mais ou menos interessante, pelo que foram usados vários métodos de seleção que permitiram reduzir quer os erros de treino, mas principalmente os erros de teste, que neste caso, funcionam como controlo do modelo.

Outro dos problemas encontrados foi a não adaptação dos modelos a picos ou baixas repentinas do IO, algo que é explicável pela variação insuficiente dos parâmetros de entrada que levam a uma estimação incorreta. O mesmo aconteceria se os dados de entrada tivessem picos ou baixas repentinas. Sendo assim, uma abordagem por outros métodos quer sejam eles uma regressão linear de grau 1 ou superior, poderia ter todo o interesse em ser explorada, visto que, a modelação se poderia adaptar melhor a estas mesmas alterações repentinas do IO.

Além disto, o Spike and Slab é uma técnica muito vantajosa, pois permite, para além de calcular valores individuais do IO, determinar as margens de confiança associadas a esse mesmo resultado, o que em planeamento é de capital importância, pois os valores previstos têm

sempre erros associados, algo que, pode ser, de certa forma, acautelado pela representação dos resultados não como um valor fixo, mas sim em forma de intervalo probabilístico.

Dado que, as entradas são também elas estimativas, um futuro desenvolvimento poderá ser a incorporação de incerteza nas entradas de modo obter estimativas do IO que também incluam este fator.

Posto isto, é de realçar a possível mais-valia da aplicação da técnica Spike and Slab a outros problemas de estimação, tais como, a evolução dos índices de qualidade de serviço técnica, evolução das perdas na rede ou a evolução da segurança de abastecimento.

Referências

[1] Presidência Do Conselho De Ministros, Diário da República, 1.ª série, Decreto-Lei n.º 15/2022 de 14 de janeiro," Estabelece a organização e o funcionamento do Sistema Elétrico Nacional, transpondo a Diretiva (UE) 2019/944 e a Diretiva (UE) 2018/2001", https://files.dre.pt/1s/2022/01/01000/0000300185.pdf.

[2] Almeida, J. P. E. S., Previsão de Investimentos com Base em Informação Esparsa, Dissertação realizada no âmbito do Mestrado Integrado em Engenharia Electrotécnica e de Computadores Major Energia, Faculdade de Engenharia da Universidade do Porto. 2020.

[3] Dangeti, P., Fundamentals of Statistical Modeling and Machine Learning Techniques, O'Reilly Media. 2017.

[4] Koehrsen, W., Overtting vs. Undertting: A Complete Example. 2018, https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765. Accessed: April 2022.

[5] Kolodiazhnyi K., Hands-On Machine Learning with C++, Packt Publishing. 2020

[6] Ghosh, J. K., DElampady, M., Samanta, T., An Introduction to Bayesian Analysis, Springer New York, NY, 2006.

[7] Berger, J. O., statistical Decision Theory and Bayesian Analysis, Springer, 1985

[8] Insua, D. R., Ruggeri, F., Robust Bayesian Analysis, Springer, 2000.

[9] Scott, S. L., & Varian, H. R., Predicting the present with Bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation, 5(1–2), 4–23. https://doi.org/10.1504/IJMMNO.2014.059942.

[10] Chambers, J., Software for Data Analysis, Springer New York, NY, 2008.

[11] Agarwal, A. Bayesian Variable Selection With Spike-And-Slab Priors. MSc thesis. Ohio State University, 2016.

[12] Bruinsma, W., Spike and Slab Priors. https://wesselb.github.io/assets/write-ups/Bruinsma,%20Spike%20and%20Slab%20Priors.pdf.

[13] Fidalgo, J. N. M., Macedo, P., Costa, H. Estimation of Investments with Scarce Data – comparing LASSO, Bayesian and CMLR, Faculdade de Engenharia da Universidade do Porto. 2022.

[14] Roque, A., Intervalo de Confiança para uma Média Populacional, Aula 5, Estatística II. http://sisne.org/Disciplinas/Grad/ProbEstat2/aula5.pdf

[15] Bai, R., Ročková, V., George, E., Handbook of Bayesian Variable Selection, Chapman and Hall/CRC, 2021

[16] Braun, W. J., Murdoch, D. J., A First Course in statistical programming with R, Cambridge University Press, 2021

[17] Dixon, W. J., Massey, F. J., Introduction to statistical analysis, McGraw-Hill, 1951

[18] Bryant, E. C., Statistical Analysis, Scholar Select, 2018

[19] Chen, S. M., Bauer, D. J., Belzal, W. M., Brandt, H., Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso, 2022

[20] Koch, B., Vock, D. M., Wolfson, J., Vock, L. B., Variable selection and estimation in causal inference using Bayesian spike and slab priors, Sage journals, 2020

Anexos

```
# introduzir matriz x

x <- matrix
(c(2,2,2,1.93,1,1,1,1,1.8,2,1.66,1.87,1.9,1.92,1.77,1.18,1.04,1.27,1.1,2,1.67,2,1.82,1.86,1.8
8,1.68,1.37,1.15,1.55,1.4,1.84,1.39,1.95,1.76,1.79,1.87,1.89,1.59,1.16,1.73,1.6,1.81,1.5,1.8
3,1.68,1.72,1.78,1.75,1.87,1.19,1.9,1.8,1.65,1.39,1.7,1.64,1.55,1.51,1.8,1.97,1.15,1.85,1.8,
1.81,1.29,1.23,1.59,1.34,1.27,1,1.87,1.36,1.81,1.9,1.47,1.41,1.77,1.48,1.29,1.24,1.51,2,1.5
1,2,2,1.48,1.79,1.27,1.37,1.2,1.14,2,1.89,1.61,1.85,1.97,1.44,1.43,1.42,1.26,1.1,1.06,1.82,1
.66,1.93,1.64,1.86,1.21,1.28,1.12,1.15,1.02,1.01,1.3,1.71,2,1.55,1.81,1.25,1.22,1.11,1.05,1,
1,1.17,1.8,1.74,1.54,1.83,1.14,1.03,1,1,1.01,1.02,1.32,2,1.58,1.6,1.88,1,1,1.04,0.99,1.05,1.
06,1.35,2.16,1.42,1.63,1.98,1.34,1,1.04,0.97,1.06,1.07,1.54,2.38,1.2,1.61,2.09,1.41,1.04,1.
06), nrow=15,byrow=TRUE)
colnames(x)= c("IC","NNC","NNF","IHPC","PIB","TDes","Econs","NCl","PT","LA","CS")

#introduizr matriz y

y <-
matrix(c(36.64,30.1,34.29,25.92,25.61,27.25,30.57,27.27,22.07,18.75,15.02,12.07,11.64,11.3
4,14.09))
colnames(y)= c("IO")

#introduzir o número de anos de teste

teste=0
#nº de iterações
```

```
it=1000
#média ponderada

alpha=0.3

if (teste==0){
  model <- BoomSpikeSlab::lm.spike(y~x,niter=it)
  coeficientes <- model$beta
  sse <- model$sse
  #MAPE treino
  MAPE_TREINO=0
  erro3=0
  NRMSE=0
  for (i in 1:nrow(x))
  {
    c <- coeficientes[,2:12] %*% x[i, ]
    c <- c+coeficientes[,1]
    erro <- (abs(c-y[i,])/(y[i,]))*100
    erro3= erro3 + ((c-y[i,])^2)
    MAPE_TREINO = MAPE_TREINO + (erro/nrow(x))
  }
  NRMSE_TREINO=((sqrt(erro3/nrow(x)))/mean(y))*100

  IO=matrix(NA,nrow=it,ncol=nrow(x))
  for(i in 1:nrow(x)){
     IO[,i] = coeficientes[,2:12] %*% x[i, ]+coeficientes[,1]
    }
 resultados=as.data.frame(cbind(coeficientes,IO,sse,MAPE_TREINO,NRMSE_TREINO))
  colnames(resultados)= c("b",
"IC","NNC","NNF","IHPC","PIB","TDes","Econs","NCl","PT","LA","CS","IO_2003","IO_2004","IO_2005"
,"IO_2006","IO_2007","IO_2008","IO_2009","IO_2010","IO_2011","IO_2012","IO_2013","IO_2014","I
O_2015","IO_2016","IO_2017","SSE","MAPE","NRMSE")
   resultados <- resultados[order(resultados[,28]),]
 }
 if(teste==1){
  x1=x
  x1 <- x1[-nrow(x),]
  y1=y
```

```
y1 = matrix(y1[-nrow(y),])
model <- BoomSpikeSlab::lm.spike(y1~x1,niter=it)
coeficientes <- model$beta
sse_global <- model$sse
sse_treino <- model$sse
#sse com todos os anos
  a <- coeficientes[,2:12] %*% x[nrow(x), ]
  a <- a+coeficientes[,1]
  a <- a-y[nrow(x),]
  sse_global <- sse_global+a^2
  sse_teste=a^2
  sse_ponderado=sse_treino*alpha+sse_teste*(1-alpha)
#Treino
erro2=0
NRMSE_TESTE=0
erro4=0
erro10=0
for (i in 1:nrow(x1))
{
  e <- coeficientes[,2:12] %*% x[i, ]
  e <- e+coeficientes[,1]
  erro2_tempo <- (abs(e-y[i,])/(y[i,]))*100 #erro percentual
  erro2=erro2+erro2_tempo
  erro4= erro4 + ((e-y[i,])^2)
}
NRMSE_TREINO=((sqrt(erro4/nrow(x1)))/mean(y1))*100
MAPE_TREINO = (erro2/(nrow(x1)))
 #Teste
  f <- coeficientes[,2:12] %*% x[nrow(x), ]
  f <- f+coeficientes[,1]
  erro10= ((y[15,]-f)^2)
  MAPE_TESTE <-(abs(f-y[nrow(x),])/(y[nrow(x),]))*100
  NRMSE_TESTE=((sqrt(erro10))/y[15,])*100

  IO=matrix(NA,nrow=it,ncol=nrow(x))
  for(i in 1:nrow(x)){
    IO[,i] = coeficientes[,2:12] %*% x[i, ]+coeficientes[,1]
  }
```

```
        MAPE_PONDERADO=alpha*MAPE_TREINO +(1-alpha)*MAPE_TESTE
        NRMSE_PONDERADO=alpha*NRMSE_TREINO +(1-alpha)*NRMSE_TESTE


resultados=as.data.frame(cbind(coeficientes,IO,sse_global,sse_treino,sse_teste,sse_ponderado,MAPE_TREINO,MAPE_TESTE,MAPE_PONDERADO,NRMSE_TREINO,NRMSE_TESTE,NRMSE_PONDERADO))
        colnames(resultados)= c("b",
"IC","NNC","NNF","IHPC","PIB","TDes","Econs","NCl","PT","LA","CS","IO_2003","IO_2004","IO_2005",
"IO_2006","IO_2007","IO_2008","IO_2009","IO_2010","IO_2011","IO_2012","IO_2013","IO_2014","IO_2015","IO_2016","IO_2017","SSE_GLOBAL","SSE_TREINO","SSE_TESTE","SSE_PONDERADO","MAPE_TREINO","MAPE_TESTE","MAPE_PONDERADO","NRMSE_TREINO","NRMSE_TESTE","NRMSE_PONDERADO")
        resultados <- resultados[order(resultados[,28]),]
    }
    if(teste>1){
      x1=x
      for(i in 1:teste)
      {
        x1 <- x1[-nrow(x1),]
      }
      y1=y
      for(i in 1:teste)
      {
        y1 = matrix(y1[-nrow(y1),])
      }
      model <- BoomSpikeSlab::lm.spike(y1~x1,niter=it)
      coeficientes <- model$beta
      sse_global <- model$sse
      sse_treino <- model$sse
      #sse global
      for (i in (nrow(x1)+1):nrow(x)) {
        a <- coeficientes[,2:12] %*% x[i, ]
        a <- a+coeficientes[,1]
        a <- a-y[i,]
        sse_teste=a^2
        sse_global <- sse_global+a^2
```

```
}
#Treino
MAPE_TREINO=0
erro5=0
erro6=0
for (i in 1:nrow(x1))
{
  e <- coeficientes[,2:12] %*% x[i, ]
  e <- e+coeficientes[,1]
  erro5 <- (abs(e-y[i,])/(y[i,]))*100
  erro6= erro6 + ((e-y[i,])^2)
  MAPE_TREINO = MAPE_TREINO + (erro5/(nrow(x1)))
}
NRMSE_TREINO=((sqrt(erro6/nrow(x1)))/mean(y1))*100

#Teste
MAPE_TESTE=0
NRMSE_TESTE=0
erro7=0
r=0
y2=0
for (i in (nrow(x)-teste+1):nrow(x)) {
  f <- coeficientes[,2:12] %*% x[i, ]
  f <- f+coeficientes[,1]
  h <- (abs(f-y[i,])/(y[i,]))*100
  erro7= erro7 + ((f-y[i,])^2)
  MAPE_TESTE <- MAPE_TESTE + h/(teste)
y2= y2+ y[i,]
}
NRMSE_TESTE = ((sqrt(erro7/(nrow(x)-nrow(x1))))/(y2/teste))*100

IO=matrix(NA,nrow=it,ncol=nrow(x))
for(i in 1:nrow(x)){
  IO[,i] = coeficientes[,2:12] %*% x[i, ]+coeficientes[,1]
}

MAPE_PONDERADO=alpha*MAPE_TREINO +(1-alpha)*MAPE_TESTE
NRMSE_PONDERADO=alpha*NRMSE_TREINO +(1-alpha)*NRMSE_TESTE
```

```
        sse_ponderado=sse_treino*alpha+sse_teste*(1-alpha)
resultados=as.data.frame(cbind(coeficientes,IO,sse_global,sse_treino,sse_teste,sse_ponderado,MAPE_TREINO,MAPE_TESTE,MAPE_PONDERADO,NRMSE_TREINO,NRMSE_TESTE,NRMSE_PONDERADO))
        colnames(resultados)= c("b",
"IC","NNC","NNF","IHPC","PIB","TDes","Econs","NCl","PT","LA","CS","IO_2003","IO_2004","IO_2005",
"IO_2006","IO_2007","IO_2008","IO_2009","IO_2010","IO_2011","IO_2012","IO_2013","IO_2014","IO_2015","IO_2016","IO_2017","SSE_GLOBAL","SSE_TREINO","SSE_TESTE","SSE_PONDERADO","MAPE_TREINO","MAPE_TESTE","MAPE_PONDERADO","NRMSE_TREINO","NRMSE_TESTE","NRMSE_PONDERADO")
        resultados <- resultados[order(resultados[,28]),]
    }
```