Faculdade de Engenharia da Universidade do Porto



Previsão de Investimentos com Base em Informação Esparsa

João Pedro Espírito Santo de Almeida

VERSÃO FINAL

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

Orientador: José Nuno Moura Marques Fidalgo
Coorientador: Pedro Miguel Ferreira Macedo

Setembro de 2020

© João Pedro Espírito Santo de Almeida, 2020

Resumo

O estudo descrito neste documento foi inspirado na obrigatoriedade da elaboração de um Plano de Desenvolvimento e Investimento da Rede de Distribuição (PDIRD) por parte do Operador da Rede de Distribuição (ORD) prevista no regulamento vigente relativo ao setor elétrico, o art.º 41º do Decreto-Lei n.º 29/2006, de 15 de Fevereiro que estabelece os princípios gerais relativos à organização e ao funcionamento do sistema elétrico nacional.

Neste plano é necessário estimar os investimentos obrigatórios (IO) a realizar na rede nacional de distribuição (RND) por um período temporal de cinco anos sendo este plano renovado de dois em dois anos. O IO recai sobre o ORD que tem a responsabilidade de responder aos pedidos de ligação, como por exemplo, novas habitações, novos parques eólicos, indústrias, etc. Desta forma, o IO depende de diversos fatores, tais como a saúde económica do país, o ritmo da construção civil, o crescimento da carga, etc.

Uma das características mais marcantes deste problema é a existência de uma grande quantidade de fatores que podem influenciar o IO, mas simultaneamente uma grande escassez de exemplos, o que constitui um grande desafio em termos de obtenção de modelos de estimação precisos e robustos.

O estudo aqui apresentado surge no seguimento de um realizado anteriormente 'Estimação de Investimentos na Rede de Distribuição' e baseia-se nos mesmos dados, mas visa explorar um diferente modelo de previsão (LASSO), o qual é capaz de integrar a gestão da complexidade do modelo no processo de regressão.

Vão ser apresentados e discutidos os resultados obtidos para avaliar o método LASSO em comparação com outros modelos de regressão e com o método desenvolvido pelo INESCTEC, atualmente utilizado pela EDP.

O principal objetivo é comparar ambas as abordagens em termos de desempenho e de complexidade, sendo que uma das principais dificuldades para obter um bom modelo de previsões é definir o critério objetivo que melhor se enquadra com o problema sob análise.

Abstract

The study described in this document was inspired by the mandatory elaboration of a Distribution Network Development and Investment Plan (DNDIP) by the Distribution Network Operator (DNO) provided for the current regulation on the electricity sector, article 41 of Decree Law no. 29/2006, of 15 February, which establishes the general principles relating to the organization and operation of the national electrical system.

In this plan, it is necessary to estimate the required investments (RI) to be carried out in the national distribution network (NDN) for a period of five years, with this plan being renewed every two years. THE RI is implicated in DNO program, which has the responsibility to respond to connection requests, such as new housing, new wind farms, industries, etc. Therefore, the RI depends on several factors, such as the country's economic health, the pace of construction, the growth of electric charge, etc.

One of the most striking characteristics of this problem is the existence of a large number of factors that can influence the RI, but at the same time a great scarcity of examples, which constitutes a great challenge in terms of obtaining accurate and robust estimation models.

The study present here follows on from a previously conducted 'Estimation of Investments in the Distribution Network' and is based on the same data, but aims to explore a different forecasting model (LASSO), which is capable of integrating complexity management model in the regression process.

The results obtained to evaluate the LASSO method will be present and discussed I comparison with other regression models and with the method developed by INESCTEC, currently used by the DNO.

The main objective is to compare both approaches in terms of performance and complexity, and one of the main difficulties in obtaining a good forecasting model is to define the objective criterion that best fits the problem under analysis.

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores.

Ao professor Doutor José Nuno Fidalgo por toda a paciência, orientação, disponibilidade e compreensão. Por toda a revisão cuidada do trabalho elaborado e brevidade de resposta.

Ao Engº Pedro Macedo por estar sempre disponível e pronto a ajudar, e por toda a tranquilidade e motivação. Sem as sugestões e diretrizes de ambos, nada disto teria sido possível.

Em seguida, um agradecimento especial à minha família e amigos pelo apoio durante todo este período e em todos as fases importantes da minha vida.

Neste grupo, um especial obrigado aos meus pais por serem o meu suporte de vida, pela confiança, e por todos os valores e conhecimento que me transmitem diariamente; à minha irmã por toda a disponibilidade e motivação; também a ela e ao meu cunhado por me presentearem com um afilhado incrível que todos os dias me alegra sempre com um sorriso.

Por fim, mas não menos especial, gostaria de agradecer à Mariana pela compreensão e por todo o apoio, tanto nesta fase como em todos os momentos partilhados.

O meu especial obrigado,
João Almeida

“To the optimist, the glass is half full. To the pessimist, the glass is half empty. To the engineer, the glass is twice as big as it needs to be”

Unknown

Índice

Resumo ..... **i**

Abstract ..... **iii**

Agradecimentos ..... **v**

Índice ..... **ix**

Lista de figuras ..... **xi**

Lista de tabelas ..... **xii**

Abreviaturas e Símbolos ..... **xiii**

Capítulo 1 ..... **1**

Introdução 1
1.1 - Enquadramento 2
1.2 - Objetivos 6
1.3 - Estrutura do Documento 6

Capítulo 2 ..... **8**

Metodologia 8
2.1 - Caracterização Geral da Informação 9
2.2 - LASSO com recurso ao Excel 13
2.3 – LASSO com recurso ao R 15

Capítulo 3 ..... **21**

Resultados 21
3.1 - Modelos obtidos sem recurso ao LASSO 22
3.2 - Modelos obtidos usando o LASSO com recurso ao R segundo as mesmas variáveis de 3.1. 25
3.3 – Comparação dos resultados dos pontos 3.1. e 3.2. 28
3.4 – LASSO com Excel considerando todas as variáveis 29
3.5 – LASSO com o R considerando todas as variáveis 32
3.6 – Comparação de todos os modelos obtidos 33

Capítulo 4 ..... **36**

Conclusões e Desenvolvimentos Futuros 36
4.1 - Conclusões Finais 36
4.2 - Desenvolvimentos Futuros 37

Referências 39

Lista de figuras

Figura 1.1- Exemplo de um tipo de modelo simples. 3

Figura 1.2- Exemplo de um tipo de modelo complexo. 3

Figura 1.3- Relação entre a complexidade do modelo e o erro, sendo descrito o comportamento da tendência, da variância e do erro total associado. 4

Figura 2.1- Gráfico do comportamento de IO ao longo do tempo com recurso a modelos simples e complexos. 9

Figura 2.2- Relação linear entre IO e IC 11

Figura 2.3- Relação linear entre IO e NNC 12

Figura 2.4- Relação linear entre IO e NNF 12

Figura 2.5- Fluxograma da elaboração do modelo LASSO de forma manual 14

Figura 2.6- Gráfico que mostra a variação do erro quadrático em função do logaritmo de lambda . 17

Figura 3.1- Gráfico do IO ao longo dos anos para os modelos resultantes das equações numerados acima. 24

Figura 3.2- Gráfico do IO ao longo dos anos para os modelos do LASSO obtidos com recurso ao R .. 26

Figura 3.3- Representação gráfica das médias dos melhores modelos obtidos no ponto 3.1. e 3.2. 28

Figura 3.4- Modelos criados com recurso ao LASSO no Excel de forma manual relativos aos anos de treino de 2003 a 2016. 30

Figura 3.5- Modelos criados com recurso ao LASSO no Excel de forma manual relativos aos anos de treino de 2003 a 2015. 30

Figura 3.6- Média dos modelos gerados pelo LASSO considerando para teste 2016/2017 e só 2017. 32

Figura 3.7-- Evolução de diferentes modelos obtidos para a previsão do IO ao longo do tempo 33

Lista de tabelas

Tabela 2.1- Valores normalizados das variáveis. 10

Tabela 2.2- Valores normalizados das variáveis dependentes para os anos de previsões 13

Tabela 2.3- Tabela dos coeficientes do LASSO no caso em que os dados de entrada estão compreendidos entre 2003 e 2016 sendo apenas o ano de 2017 para teste. 18

Tabela 3.1- Erro de treino e de teste dos modelos ilustrados na figura 3.1. 24

Tabela 3.2- Erro de treino e de teste dos modelos ilustrados na figura 3.2. 26

Tabela 3.3- Erros de treino e de teste associados aos modelos ilustrados na figura 3.4 e 3.5. 31

Tabela 3.4- Erro de treino e teste relativo à média dos melhores modelos obtidos no ponto 3.2. e no ponto 3.4. 32

Tabela 3.5- Erro (%) dos modelos associados à curva do gráfico 3.7. 34

Abreviaturas e Símbolos

CNCS	Construção de Novos Cabos Subterrâneos
CNLA	Construção de Novas Linhas Aéreas
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
NNC	Número de Novas Construções
NCL	Número de Clientes na rede
NNF	Número de Novos Fogos
NPPT	Nova Potência Instalada na rede, em Postos de Transformação
ORD	Operador da Rede de Distribuição
PDIRD	Plano de Desenvolvimento e Investimento na Rede de Distribuição
PIB	Produto Interno Bruto
TDes	Taxa de Desemprego
FEPICOP	Federação Portuguesa da Indústria da Construção e Obras Públicas

Capítulo 1

Introdução

A dissertação descrita neste documento foi desenvolvida no âmbito da disciplina Dissertação do Mestrado Integrado em Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto. Descreve um trabalho de investigação cujo objetivo final é a elaboração de um modelo de previsões para realizar a estimação dos Investimentos Obrigatórios (IO) necessários a realizar na rede de distribuição por parte do Operador da Rede de Distribuição (ORD).

A necessidade da realização deste estudo resulta da obrigatoriedade de elaboração de um Plano de Desenvolvimento e Investimento da Rede de Distribuição (PDIRD), por parte do Operador da Rede de Distribuição (ORD), prevista no regulamento vigente relativo ao setor elétrico: art.º 41º do Decreto-Lei n.º 29/2006, de 15 de Fevereiro [1] que estabelece os princípios gerais relativos à organização e ao funcionamento do sistema elétrico nacional.

Neste plano vai ser necessário incluir os investimentos obrigatórios (IO) a realizar na rede nacional de distribuição (RND). O IO resulta da responsabilidade que o ORD tem em responder aos pedidos de ligação, tais como, por exemplo, novas habitações, novos parques eólicos, indústrias, etc. O reforço da rede a montante devido aos novos pedidos de ligação ou incrementos de potência também faz parte do IO. Se for necessário aumentar a capacidade do transformador de um posto de transformação (PT) devido a novos pedidos, este reforço também faz parte do IO. Desta forma, o IO depende de diversos fatores, tais como a saúde económica do país, o ritmo da construção civil, o crescimento da carga, etc. Vários destes fatores podem ser caracterizados por diversos indicadores. Por exemplo, a taxa de inflação, o produto interno bruto e o índice de preços no consumidor são alguns dos indicadores que caracterizam a saúde económica do país. Isto significa que existe uma quantidade considerável de variáveis com potencial impacto no IO.

No entanto, sendo o histórico de IO disponibilizado anualmente, apenas desde 2002, é notório o reduzido número de exemplos. A conjugação destes dois factos (grande número de variáveis, por um lado, e pequeno número de exemplos) constitui um dos principais desafios da estimação do IO.

Neste documento, vai ser exposta uma metodologia (LASSO) diferente da atualmente utilizada pelo ORD que visa gerar um modelo de previsões de IO simples e, ao mesmo tempo, eficiente.

No capítulo introdutório, é apresentado um breve enquadramento do estudo em questão, onde será abordado a elaboração deste método, os principais objetivos desta investigação bem como a estrutura de todo o documento.

1.1 - Enquadramento

Perante o problema "como elaborar um modelo de previsões" é necessário perceber qual o modelo mais adequado a adotar para a resolução do mesmo.

O IO, variável dependente do que se pretende prever, pode depender de forma não-linear das variáveis independentes, pode depender de variáveis que vão influenciar o seu desempenho mas que poderão ser redundantes, e pode também depender de fatores ocasionais cuja caraterização é bastante complexa e nem revelar melhorias no desempenho do modelo. Refere-se como exemplo o facto de se realizar ou não eleições no presente ano. É impossível prever os acontecimentos futuros com precisão absoluta, pois existem sempre fatores inesperados e efeitos difíceis de delinear. Esta imprevisibilidade é usualmente denominada de ruído.

O valor real a prever pode ser descrito pela seguinte função:

$$Valor\ real = Modelo\ de\ previsão + Ruído \tag{1.1}$$

$$\equiv Y = f(X) + \varepsilon \tag{1.2}$$

sendo que, o objetivo principal é aproximar ao máximo o modelo de previsão do valor real, tendo em conta que é impossível serem iguais pois raramente o nível de ruído será desprezável. Podem existir casos pontuais onde os valores são muito próximos do real, mas admite-se que, em termos gerais, haja sempre algum afastamento. Em termos estatísticos, o mais relevante é a análise comportamental geral e não as situações pontuais.

Para determinar o modelo que melhor se adequa ao problema, é necessário verificar se [2]:

- Os valores previstos fazem sentido;
- Visualizar a evolução do modelo;
- Realizar avaliações numéricas de modo a analisar os modelos;
- Comparar modelos.

Numa primeira abordagem ao problema, é possível optar por um tipo de modelo simples, como por exemplo uma regressão linear, ou por um modelo mais complexo, como por exemplo uma regressão não-linear (polinomial, exponencial, etc.).

As imagens de seguida apresentadas, facilitam a compreensão dos fatores que permitem a escolha de um modelo em detrimento do outro.

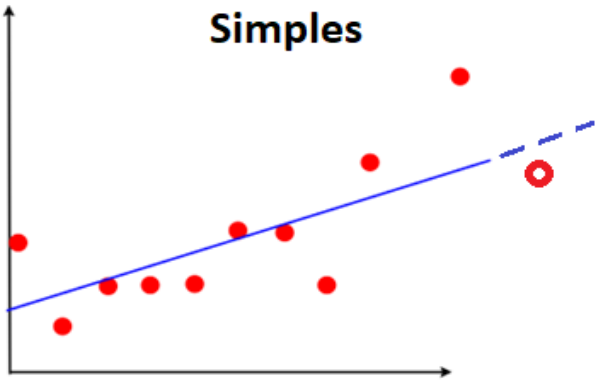


Figura 1.1- Exemplo de um tipo de modelo simples.

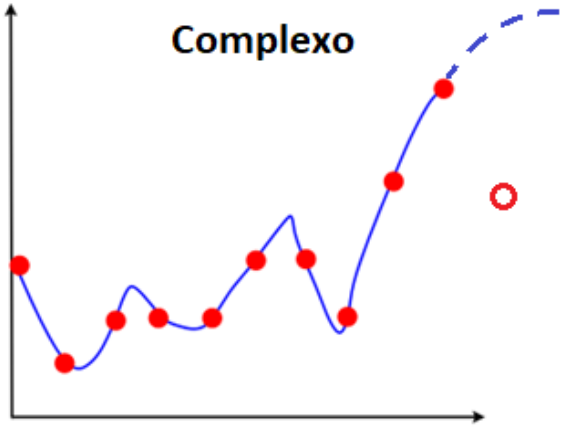


Figura 1.2- Exemplo de um tipo de modelo complexo.

A linha traçada descreve o modelo de previsão gerado, considerando os círculos a vermelho os dados de treino e a circunferência a vermelho os dados de teste. Na figura 1.1 observa-se uma regressão linear e na figura 2.2 uma regressão polinomial.

Na escolha de um modelo de previsão, existem dois conceitos a destacar:

- Variância: que pode ser interpretada como uma medida das flutuações nos dados. Elevada variância implica um maior afastamento do modelo em relação ao valor real e baixa variância diz que os valores reais estão próximos dos valores esperados. A variância pode estar relacionada com a sobre-adaptação do modelo (overfitting).
- Viés (termo em inglês "bias"): corresponde a um desvio sistemático, eventualmente causado por problemas na aplicação do algoritmo de treino. Um elevado Viés pode fazer com que o modelo não identifique relações relevantes entre entradas e saídas. O Viés também pode traduzir-se numa sub-adaptação do modelo (underfitting).

A variância é um termo abrangente que pode ser utilizado em situações distintas. Neste caso específico, refere-se à variância entre os erros de treino e de teste. Ou seja, o erro de treino vai ter um valor e o erro de teste vai ter outro distinto. Considera-se a variância entre os erros, a flutuação que ocorre quando se transita de treino a teste.

Analisando-se agora os gráficos acima ilustrados, o modelo simples vai ter um elevado viés pois os pontos estão afastados do modelo e uma baixa variância visto os valores obtidos no modelo não se encontrarem dispersos. Quanto ao modelo polinomial, pode-se observar um baixíssimo viés (praticamente zero, neste caso específico pois todos os pontos a vermelho fazem parte do modelo) e uma variância elevada, o que pode provocar um erro de teste elevado, desfasado e superior ao erro de treino.

As principais soluções para colmatar o elevado viés é o aumento do número de variáveis ou aumentar a complexidade do modelo. No caso da variância, as principais abordagens para conseguir a sua diminuição é aumentar o número de dados de treino ou através de técnicas de regularização [3].

Neste estudo, os dados são limitados e as variáveis foram selecionadas e disponibilizadas à partida. Para efeitos comparativos, as condições iniciais vão ser as mesmas.

A imagem abaixo apresentada, figura 2.1., ilustra a relação do modelo considerando a variância e a tendência.

Um modelo ideal é aquele que tem uma baixa variância e uma baixa tendência [4], contudo nenhum destes casos se enquadra nestas condições e torna-se necessário perceber o seu comportamento de modo a fazer uma seleção.

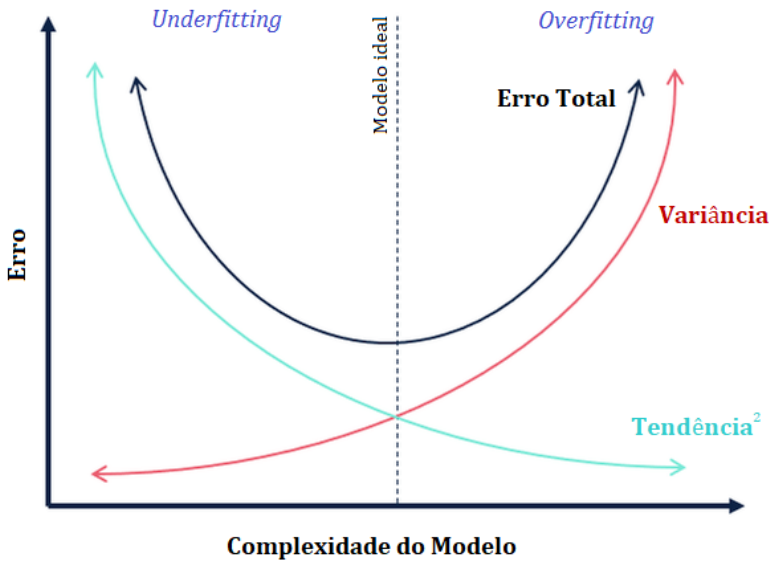


Figura 1.3- Relação entre a complexidade do modelo e o erro, sendo descrito o comportamento da tendência, da variância e do erro total associado.

Um dos grandes problemas no que diz respeito a modelos de previsões (com muitas entradas e poucos exemplos) é que estes apresentam valores de erro de treino muito baixos, mas elevados erros de teste. Isto acontece, pois, existe uma sobre adaptação dos dados (*Overfitting*). O *Overfitting* acontece quando um modelo se ajusta eficazmente a um conjunto de dados históricos, mas não mostra capacidade para prever novos resultados. O modelo adapta-se demasiado aos dados de treino e, por consequência, ao ruído que a ele está associado, provocando erros de teste elevados [5].

No caso do modelo complexo, regressão polinomial, vai ocorrer *Overfitting*. O cenário obtido representa situações onde o modelo fica muito próximo da realidade, mas também situações em que este falha por muito o valor real, resultando assim num modelo inconsistente.

No caso do modelo simples, regressão linear, verifica-se *Underfitting*. O *Underfitting* encontra-se no extremo oposto do Overfitting e acontece quando o modelo não se adapta bem aos dados de treino, muitas vezes por incapacidade do próprio modelo, como acontece na figura 1.1: sendo um modelo linear, nunca poderia adaptar-se adequadamente a uma curva claramente afastada dum modelo linear (reta) [6].

Neste estudo vai ser considerada uma regressão linear, onde existirá sempre um erro associado ao valor de saída, mas que resulte em previsões aceitáveis e de menor oscilação em comparação com modelos complexos não-lineares (por exemplo, polinomial), onde previsões de adaptação ao treino podem ser muito boas, mas também divergir consideravelmente. Outro motivo para usar modelos lineares reside na sua dependência de um menor número de parâmetros. Enquanto uma relação linear simples y=mx+b, apenas existem 2 parâmetros a determinar (m e b), num polinómio de grau 4, seria necessário determinar mais parâmetros. Dado que o número de exemplos de treino é bastante limitado, convém reduzir também o número de parâmetros a determinar.

Uma regressão linear pode ser simples ou múltipla. A simples é referente a uma relação direta entre a variável dependente (*output*) e apenas uma variável independente (*input*). A múltipla tem o mesmo raciocínio que a simples, mas com múltiplas variáveis independentes. Neste estudo, vai ser considerada uma regressão linear múltipla pois vamos prever IO tendo em consideração mais do que uma variável de entrada.

Existem várias metodologias e critérios de aproximação para concretizar a regressão linear. Um dos critérios mais conhecido é o dos Mínimos Quadrados, cujo objetivo é determinar os parâmetros de regressão que minimizam a soma dos quadrados dos desvios.
As metodologias podem conter ou não restrições relativamente às variáveis ou aos parâmetros. No caso em que existem restrições ou condições que penalizam a complexidade do modelo, diz-se que existe Regularização.

As metodologias de regressão *Ridge* e LASSO usam a regularização, processo onde é adicionada informação para prevenir o *Overfitting*, recorrendo a penalizações. A diferença entre estas duas metodologias é a forma como a penalização é aplicada.

No caso de uma regressão *Ridge*, a penalização é aplicada no somatório dos coeficientes ao quadrado [7]. Já a regressão LASSO aplica a penalização no valor absoluto dos coeficientes [8].

Isto pode ser observado analisando as funções objetivo em cada um dos casos.

Função objetivo *Ridge*:

$$\mathrm{RSS}\ (\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (1.2)$$

Função objetivo LASSO:

$$\text{RSS}\,(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.3}$$

Onde,

→ $\text{RSS}\,(\beta) = \sum_{i=1}^{n} (\varepsilon_i)^2$, "*Residual Sum of Squares*", Soma do Quadrado dos Resíduos; Soma dos erros quadráticos da diferença entre cada ponto do modelo e o valor real;

→ β_j - Valor do coeficiente em relação à variável;

→ λ – Coeficiente de penalização.

Em ambas as situações, à medida que o λ aumenta, o modelo perde flexibilidade e isso implica que vai existir um aumento do viés e uma diminuição da variância.

A principal diferença entre estas duas metodologias é o fato de a *Ridge* considerar sempre todas as variáveis, ou seja, por muito grande que a penalização seja, o valor do coeficiente associado pode ser muito baixo, mas nunca é zero absoluto, o que faz com que a variável tenha de ser considerada para obter a previsão. Com o LASSO, o coeficiente associado à variável vai tomar mesmo o valor zero e isso faz com que a variável deixe de ser considerada para a previsão. Desta forma, as variáveis redundantes não vão ser consideradas na obtenção do modelo de previsão.

O modelo que se está a desenvolver para prever o IO tem um elevado número de variáveis independentes comparativamente ao reduzido número de anos de histórico disponível para fazer esta previsão.

Posto isto, o LASSO mostra ser uma metodologia que se enquadra no problema a analisar.

1.2 - Objetivos

O principal objetivo do trabalho consiste em desenvolver modelos de previsões, que permitam aos ORD obter estimações de IO por um período temporal de cinco anos sendo que os dois primeiros anos são o principal foco.

De forma a desenvolver o principal objetivo deste trabalho, é essencial definir a metodologia a utilizar para obter um modelo de previsões que seja eficiente e ao mesmo tempo simples.

Será feita uma análise comparativa dessa mesma metodologia utilizada de duas formas distintas: manual ou computacional.

Outro desafio exposto é a dificuldade que existe em definir o critério de seleção de um bom modelo de previsões e a sua relação com os diferentes tipos de metodologias abordadas.

1.3 - Estrutura do Documento

Este documento apresenta um estudo desenvolvido no âmbito da temática Previsão de Investimentos com Base em Informação Esparsa, encontrando-se dividido em 4 capítulos.

No capítulo 1 é feito um enquadramento geral ao tema, onde são enumerados os objetivos e são expostos os conceitos gerais associados à metodologia adotada para realização deste estudo.

No capítulo 2 desenvolve-se a metodologia do trabalho de uma forma explicativa, apresentando duas formas distintas de aplicar o mesmo método: uma manual e outra computacional.

No capítulo 3 expõem-se os resultados obtidos sob a forma de gráficos explicativos e onde é feita uma análise comparativa entre os diversos modelos.

Por fim, no capítulo 4, são apresentadas as conclusões principais e sugeridas propostas para desenvolvimentos futuros.

Capítulo 2

Metodologia

A previsão de IO é considerada um problema atual e, como tal, surgiu a necessidade da realização de estudos de investigação de forma a dar resposta a esta situação. Dada a novidade do tema, torna-se desafiante obter referências bibliográficas alusivas ao mesmo que apresentem metodologias previamente definidas, testadas e aplicáveis, o que representa uma limitação no desenvolvimento deste tipo de estudos.

Este estudo foi motivado por dois fatores. Primeiro, até 2006/2007, o crescimento da carga do sistema foi bastante estável; a partir daí estabilizou, chegando mesmo a decrescer com a crise económica que teve início em 2009. Segundo, e talvez mais determinante no IO, o índice de construção já vinha a decair desde o início da década, atingiu mínimos históricos também na altura da crise. A metodologia tradicionalmente usada pela EDP para estimação do IO não estava preparada para este tipo de situações e começou a ter desempenhos inadequados. Em 2014, a EDP Distribuição solicitou a colaboração do INESC TEC, que desenvolveu uma metodologia que permite prever os IO necessários a incluir no PDIRD [9]. Esta ferramenta é atualmente utilizada para as previsões do operador da rede de distribuição.

O projeto aqui apresentado, surge no seguimento do anterior 'Estimação de Investimentos na Rede de Distribuição' [9], baseia-se nos mesmos dados, e visa explorar um diferente modelo de previsão, sendo neste caso a regressão LASSO.

O gráfico seguinte ilustra a diversidade de modelos obtidos numa abordagem inicial ao problema.

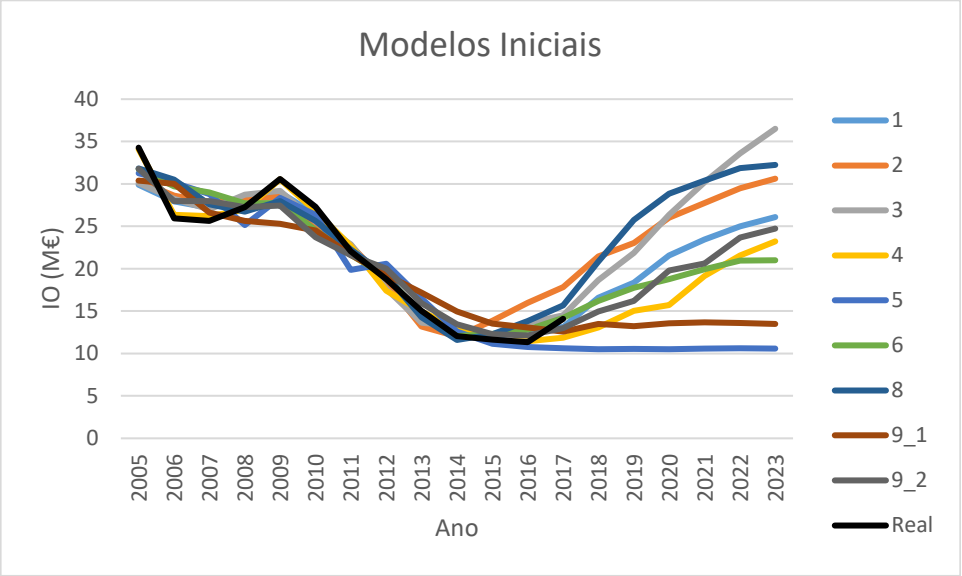


Figura 2.1- Gráfico do comportamento de IO ao longo do tempo com recurso a modelos simples e complexos.

O gráfico apresentado, que vai ser aprofundado na secção 3.1., é referente aos modelos que foram inicialmente desenvolvidos para prever o IO. Como podemos observar, estes modelos revelam uma grande amplitude e divergência de previsões entre si, apesar de se adaptarem bastante bem nos anos de treino (até 2015). Isto acontece porque existe uma infinidade de modelo capazes de se adaptar "perfeitamente" aos dados de treino. O problema é que a complexidade destes modelos, que lhes dá esta capacidade de adaptação, também é o que provoca a enorme divergência de previsões (após 2015). Desta forma, a nova metodologia vai fazer com que estas "vassouras" (divergência de modelos) desapareçam e que os modelos resultantes sejam similares.

O LASSO é um tipo de regressão linear que tenta prever o melhor modelo para um menor número de variáveis segundo critérios de penalização. Este método, ao procurar fazer um balanço entre o melhor resultado possível e o número mínimo de variáveis do qual o mesmo depende, torna o modelo final simples e ao mesmo tempo preciso [10], resultando num menor número de parâmetros de entrada comparativamente a um modelo tradicional.

A regressão LASSO foi realizada recorrendo a duas abordagens distintas, apresentadas na secção 2.2. e 2.3.

2.1- Caracterização Geral da Informação

Quando se pretende obter um modelo de estimação é necessário perceber quais as variáveis a considerar de entre as que se encontram disponíveis, o que são e qual o seu comportamento ao longo do tempo.

Os valores apresentados referentes a IO sofreram uma transformação linear devido à confidencialidade. No entanto, esta transformação em nada influência os modelos resultantes.

Os dados utilizados neste estudo foram todos facultados pelo INESC TEC, atualmente encarregue por realizar as previsões dos IO nos projetos de colaboração com a EDP Distribuição.

Dados de entrada:

- Índice de Construção (IC);
- Nº de Novas Construções (NNC) (nº);
- Nº de Novos Fogos Construídos (NNF) (nº);
- Índice de Harmonização de Preços ao Consumidor (IHPC);
- Produto Interno Bruto (PIB) (M€);
- Taxa de Desemprego (TDes) (%);
- Energia Consumida na rede (Econs) (Gwh);
- Nº de Clientes na rede (NCL)
- Nova Potência Instalada na rede, em termos de PT (NPPT) (kVA);
- Construção de Novas Linhas Aéreas (CNLA) (m);
- Construção de Novos Cabos Subterrâneos (CNCS) (m).

Saída:

- Investimento Obrigatório (IO) (M€).

A tabela que se segue contém todas as variáveis independentes já normalizadas, sendo que a variável dependente sofreu apenas uma transformação linear, de modo a mostrar o comportamento histórico dos dados mantendo a sua confidencialidade.

Tabela 2.1- Valores normalizados das variáveis.

Ano	IO	IC	NNC	NNF	IHPC	PIB	TDes	Econs	NºCl	PT	LA	CS
2003	36.64	2	2	2	1.93	1	1	1	1	1.80	2	1.66
2004	30.10	1.87	1.90	1.92	1.77	1.18	1.04	1.27	1.1	2	1.67	2
2005	34.29	1.82	1.86	1.88	1.68	1.37	1.15	1.55	1.4	1.84	1.39	1.95
2006	25.92	1.76	1.79	1.87	1.89	1.59	1.16	1.73	1.6	1.81	1.50	1.83
2007	25.61	1.68	1.72	1.78	1.75	1.87	1.19	1.90	1.8	1.65	1.39	1.70
2008	27.25	1.64	1.55	1.51	1.80	1.97	1.15	1.85	1.8	1.81	1.29	1.23
2009	30.57	1.59	1.34	1.27	1	1.87	1.36	1.81	1.9	1.47	1.41	1.77
2010	27.27	1.48	1.29	1.24	1.51	2	1.51	2	2	1.48	1.79	1.27
2011	22.07	1.37	1.20	1.14	2	1.89	1.61	1.85	1.97	1.44	1.43	1.42
2012	18.75	1.26	1.10	1.06	1.82	1.66	1.93	1.64	1.86	1.21	1.28	1.12
2013	15.02	1.15	1.02	1.01	1.30	1.71	2	1.55	1.81	1.25	1.22	1.11
2014	12.07	1.05	1	1	1.17	1.80	1.74	1.54	1.83	1.14	1.03	1
2015	11.64	1	1.01	1.02	1.32	2.00	1.58	1.60	1.88	1	1	1.04
2016	11.34	0.99	1.05	1.06	1.35	2.16	1.42	1.63	1.98	1.34	1.00	1.04
2017	14.09	0.97	1.06	1.07	1.54	2.38	1.20	1.61	2.09	1.41	1.04	1.06

Esta normalização foi feita considerando os anos de teste de 2003 a 2015 e a forma como foi realizada pode ser observada na secção 2.2., resultando da equação (2.4).

Nesta secção expõe-se uma análise mais detalha a três variáveis: IC, NNC e NNF, pois, como vai ser possível observar no ponto 3.4., estas são as variáveis que revelam maior correlação com o IO.

Para uma melhor perceção da relação entre estas variáveis e o IO, recorre-se ao coeficiente de determinação, também conhecido como R^2, que permite perceber o nível de relação linear entre duas variáveis neste caso. Ou seja, através de um modelo estatístico linear generalizado é possível obter uma relação entre duas variáveis que resulta num valor de R^2, $0 < R^2 < 1$, sendo que quanto mais próximo o valor se encontra do 1, maior é a razão explicativa para conseguir obter uma variável a partir da outra [11].

A expressão que permite o cálculo de R^2 é a seguinte:

$$R^2 = 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - f_i)^2} \tag{2.1}$$

Onde,

→ $\sum_i (y_i - \bar{y})^2$ - Erro quadrático total

→ $\sum_i (y_i - f_i)^2 = \sum_i (\varepsilon_i)^2$- Erro quadrático residual

Recorrendo a uma análise gráfica para demonstrar o R^2 referente às três variáveis mencionadas anteriormente.

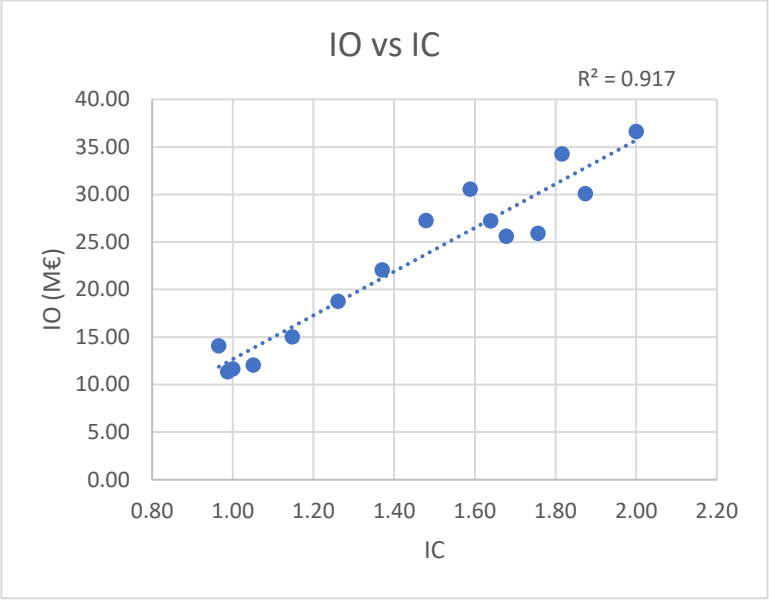


Figura 2.2- Relação linear entre IO e IC

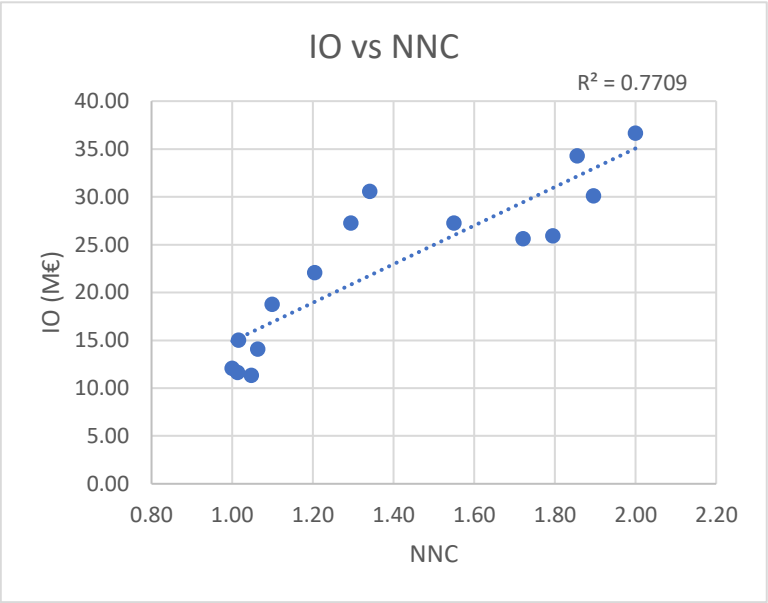


Figura 2.3- Relação linear entre IO e NNC

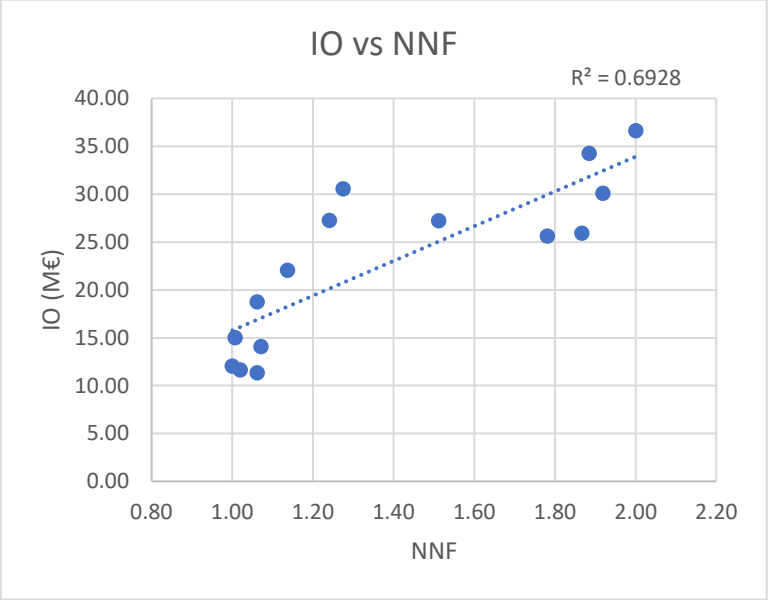


Figura 2.4- Relação linear entre IO e NNF

Como se pode observar no canto superior de cada gráfico encontra-se o valor de R^2 para cada situação. Esta relação foi calculada com os valores das variáveis normalizadas.

Os valores de R^2 para cada situação foram os seguintes:

- IC $\rightarrow R^2 = 0.917$;
- NNC $\rightarrow R^2 = 0.7709$;
- NNF $\rightarrow R^2 = 0.6928$.

O valor de R^2 representa uma medida de ajuste da variável dependente, IO, que é justificável de prever tendo em consideração a variável independente em questão através de um modelo estatístico linear generalizado.

É notório o destaque que o IC tem no que diz respeito ao valor obtido, que é quase de 92%, o que mostra que existe uma forte relação linear entre estas grandezas.

No caso de NNC e NNF, o coeficiente de determinação vai apresentar valores significativamente mais baixos.

Na tabela 2.1 foram expostos os valores utilizados para a realização do treino e teste do modelo. No entanto, também vão ser consideradas projeções das variáveis independentes para estimar o valor de IO nos seis anos seguintes ao último de teste.

Os dados utilizados nestas projeções foram facultados pelo INESC TEC e resultam de um estudo realizado anteriormente.

Dentro de todos os dados facultados e tendo em conta a análise que foi feita ao IC, NNC e NNF, é necessário fazer algumas referências alusivas aos dados projetados.

Os valores de IC nos dois primeiros anos (2018 e 2019) resultam de projeções realizadas pela Federação Portuguesa da Indústria da Construção e Obras Públicas (FEPICOP) entre outras notícias, sendo que os quatro anos seguintes foram projetados pelo INESC TEC.

Quanto a NNC e NNF, ambos foram sempre projetados pelo INESC TEC e a sua projeção é feita com base no IC.

Das restantes variáveis, LA e CS também foram previstas pelo INESC TEC sendo que todas as outras resultaram de projeções feitas por entidades certificadas.

A seguinte tabela serve para fundamentar alguns dos modelos que vão ser explorados no capítulo seguinte e para uma melhor perceção do seu comportamento.

Tabela 2.2- Valores normalizados das variáveis dependentes para os anos de previsões

Ano	IC	NNC	NNF	IHPC	PIB	TDes	Econs	NºCl	PT	LA	CS
2018	0.99	1.08	1.08	1.55	2.59	1.12	1.64	2.18	1.42	1.10	1.07
2019	1.01	1.09	1.09	1.59	2.80	1.05	1.66	2.28	1.44	1.11	1.08
2020	1.03	1.10	1.09	1.65	3.03	1.03	1.71	2.37	1.44	1.15	1.09
2021	1.04	1.10	1.09	1.66	3.23	0.98	1.73	2.46	1.43	1.12	1.09
2022	1.04	1.10	1.09	1.74	3.41	0.98	1.76	2.52	1.42	1.15	1.08
2023	1.05	1.10	1.10	1.76	3.64	0.94	1.79	2.58	1.41	1.13	1.08

2.2 - LASSO com recurso ao Excel

Numa primeira fase, recorreu-se ao Excel e de forma manual implementou-se o LASSO, explorando-se assim possíveis reações do mesmo. A ideia foi conhecer o método pormenorizadamente e adquirir sensibilidade para o problema em questão.

Com recurso à expressão (2.2), que compõe uma regressão linear, foi possível, após normalização das variáveis, perceber os valores que os seus coeficientes poderiam assumir.

$$Y = \alpha_1 * x_1 + \alpha_2 * x_2 + .. + \alpha_n * x_n + b \qquad (2.2)$$

Através da função Solver do Excel, execução do método dos mínimos quadrados, é possível obter os valores dos coeficientes. Segundo esta análise, é notória a diferença entre as variáveis relativamente ao seu grau de importância.

Sendo assim, foi testada a seguinte hipótese:

- Colocação do coeficiente da variável com menor grau de importância a 0 e análise do erro, este caso foi repetido até restarem apenas duas variáveis;

O fluxograma que se segue, descreve o processo realizado no Excel de forma a obter os modelos de previsão.

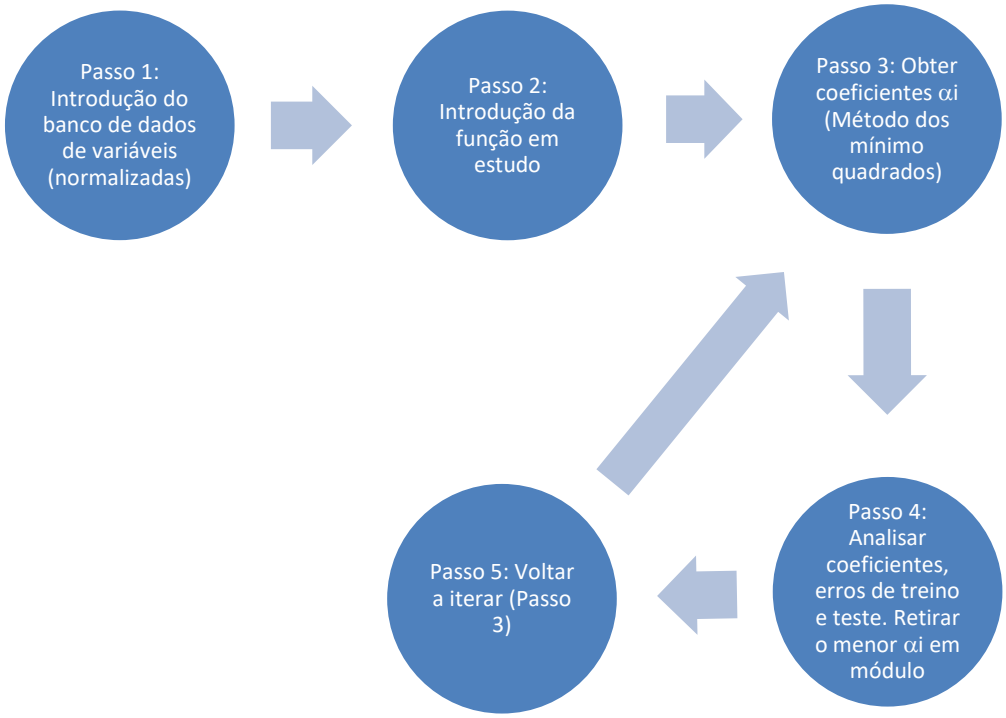


Figura 2.5- Fluxograma da elaboração do modelo LASSO de forma manual

Uma das fórmulas utilizadas no passo 2 foi a seguinte:

$$Inv_n = Ener_n \times \alpha_1 + PIB_n \times \alpha_2 + T.infl_n \times \alpha_3 + T.des_n \times \alpha_4 + I.Cons_n \times \alpha_5 + k_1 \quad (2.3)$$

Onde:

- n - Ano considerado;
- α_x - Respetivo coeficiente associado às diferentes variáveis;
- k_1 – Constante a determinar.

De forma a clarificar o processo para a obtenção do LASSO a partir do Excel, vai-se analisar a figura 2.5.

No Passo 1, procede-se à normalização de todas as variáveis independentes sendo que a mesma foi conseguida da seguinte forma:

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} + 1 \tag{2.4}$$

No caso da variável independente, a sua normalização foi feita apenas dividindo o seu valor original por um milhão.

Na fórmula pode observar-se o "+ 1" devido à necessidade de alterar o range de valores resultantes para valores superiores a 1, ou seja, $X_{normalizado} \in [1; 2]$ pois os valores entre 0 e 1 podiam conduzir a resultados indesejados. Por exemplo, se um modelo usar a variação ΔX, em vez do próprio valor, existiria o risco de ΔX tender para infinito, caso no ano anterior X=0.

No Passo 2, vai-se introduzir a fórmula (2.3) para se obter o valor de IO com base no resultado do Solver.

O Passo 3, é onde se estabelece o critério para definir o modelo. Onde o resultado de IO vai sofrer a influência do objetivo definido no Solver, que consistiu na minimização do somatório do erro quadrático associado aos valores obtidos no treino, método dos mínimos quadrados.

O somatório erro quadrático (SEQ) vai ser obtido:

$$\text{SEQ (IO)} = \sum_{i=1}^{N} (IO_i^{Real} - IO_i^{Prev})^2 \tag{2.5}$$

Onde N é referente ao número de anos de treino.

Na sequência, vai-se chegar ao Passo 4, no qual se analisa os valores dos coeficientes associados a cada uma das variáveis. Os módulos dos coeficientes são comparados e é eliminada a variável a que corresponde o coeficiente de menor valor.

Este processo é cíclico até se atingir o critério definido para paragem ou até restar apenas uma variável.

Neste método também foram realizados testes implementando uma função que limitava o somatório dos valores dos coeficientes obtidos, forçando assim o Solver a definir alguns deles com valores muito próximos de zero.

No entanto, nos modelos que vão ser abordados na secção 3, com recurso ao Excel, optou-se por não aplicar nenhuma limitação para o modelo ter uma total liberdade de seleção.

2.3 – LASSO com recurso ao R

Uma segunda abordagem foi realizada com recurso a uma ferramenta computacional autónoma já existente. Neste caso específico foi utilizado o Rstudio "um software livre de ambiente de desenvolvimento integrado para o R, uma linguagem de programação para gráficos e cálculos estatísticos" [12].

O 'R' apresenta diversos packages (bibliotecas) criadas para diversos efeitos, algumas delas correspondem exatamente aquilo que se pretende realizar, sendo que o package 'glmnet' é o que melhor se adequa ao problema em estudo nesta dissertação. Recorrendo à documentação

deste package verifica-se a seguinte descrição: “Extremely efficient procedures for fitting the entire LASSO or elastic-net regularization path for linear regression, logistic and multinominal models, …” [13].

Dentro deste package é possível obter várias funções. A que foi utilizada para prever o modelo foi ‘cv.glmnet’ que faz validação cruzada, produz gráficos e retorna o valor de lambda. Esta função requer parâmetros e dados de entrada, dos quais dois deles são todos os valores das variáveis de entrada (inputs) e a variável que estamos a prever (output) [13]. Os restantes foram assumidos como padrão exceto o alfa que é o parâmetro que define o tipo de modelo. Como já é conhecido, no modelo utilizado (regressão LASSO), o alfa toma o valor 1. Caso se mantivesse como ‘default’, a função teria liberdade para usar um modelo de regressão diferente do LASSO. No entanto, esse não é o objetivo pretendido. Esta função retorna diferentes valores de lambda, que é o coeficiente de penalização da complexidade do modelo, e esse mesmo lambda vai ter os coeficientes associados às variáveis para criar o modelo de previsão.

Utilizando, o ‘cv.glmnet’ e não o ‘glmnet’, a seleção de lambdas é feita de forma automática através da validação cruzada, cross-validation (cv), indicando assim os lambdas que a função acha mais vantajoso considerar. Estes lambdas estarão compreendidos entre o lambda mínimo (lambda.min) e o lambda mínimo com um erro padrão associado (lambda.1se).

O lambda mínimo é o que corresponde ao caso com o menor número de variáveis e concomitantemente o melhor erro de treino associado, ou seja, para conseguir diminuir o erro era necessário aumentar o número de variáveis.

O lambda.1se é o lambda que o programa considera como o melhor caso. Este lambda.1se tem um erro padrão de desvio do lambda.min a ir para o ponto onde o número de variáveis é igual. Caso contrário os modelos resultantes tinham diferentes quantidades de variáveis consideradas. Este erro padrão é obtido pelo quociente do desvio padrão e a raiz quadrada do número de exemplos considerados para treino.

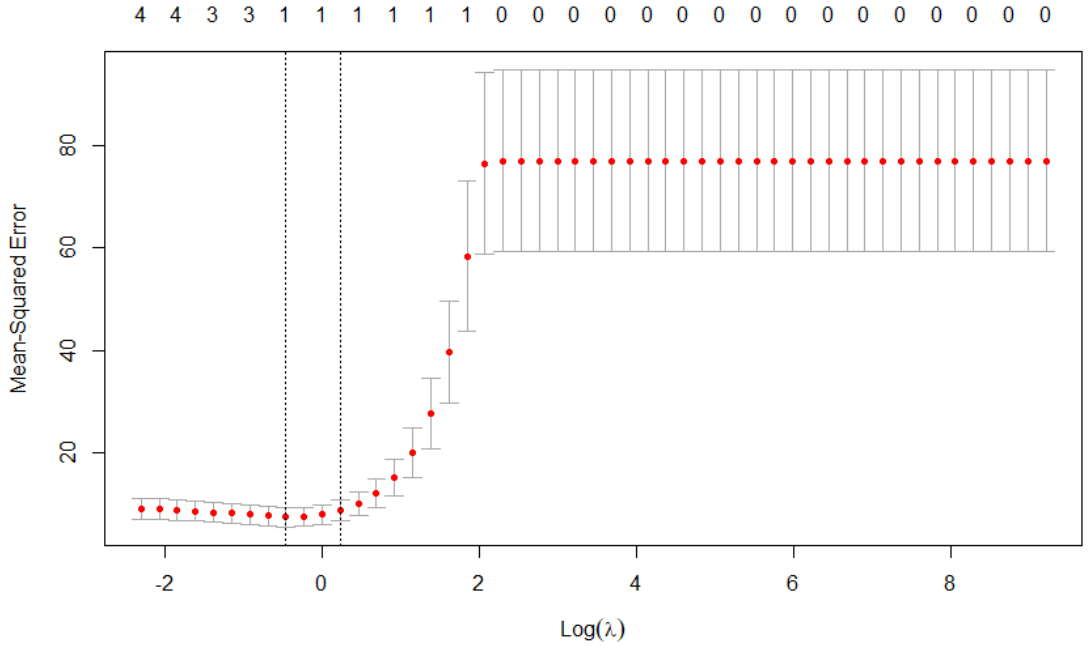


Figura 2.6- Gráfico que mostra a variação do erro quadrático em função do logaritmo de lambda

Na figura 2.6 é possível observar os pontos lambda.min e o lambda.1se, apresentados por linhas a tracejado. Sendo o ponto mais à esquerda o lambda.min que apresenta um menor erro quadrático e o mais à direita o lambda.1se. Os outros pontos são referentes a diferentes lambdas considerados, sendo os valores presentes no topo do gráfico indicativos do número de variáveis consideradas nos modelos.

Depois de executada a função 'cv.glmnet' é necessário utilizar outra função para extrair os coeficientes gerados. Essa função é 'coef' onde vai ser chamada a função cv.glmnet e o lambda para o qual se quer extrair o modelo [14].

A tabela seguinte demonstra alguns lambdas gerados e os seus respetivos modelos. Estes valores são referentes aos modelos que estão ilustrados na figura 3.2 e que irão ser discutidos na secção 3.2.

Tabela 2.3- Tabela dos coeficientes do LASSO no caso em que os dados de entrada estão compreendidos entre 2003 e 2016 sendo apenas o ano de 2017 para teste.

	Variáveis						Erro (%)	
Lambdas	IC	IHPC	PIB	TDes	Econs	Ordenada na origem	Treino	Teste
2	17,93	0	0	0	0	-3,10	13,71	2,48
1,58	19,19	0	0	0	0	-4,97	12,38	2,02
1,26 (lambda.1se)	22,13	0	0	0	0	-9,31	11,32	5,60
1	21,00	0	0	0	0	-7,63	10,48	8,44
0,79	21,62	0	0	0	0	-8,57	9,82	10,70
0,63 (lambda.min)	20,20	0	0	0	0	-6,45	9,29	12,50
0,5	22,52	-1,6e-7	0	0	4,3e-9	-9,90	8,90	13,92
0,4	23,10	-0,50	0	0	0,01	-9,99	8,75	15,82

A diminuição do lambda significa o aumento da complexidade do modelo, ou seja, mais variáveis independentes. Neste sentido, percebe-se que o erro de treino (penúltima coluna) seja tendencialmente decrescente com a diminuição de lambda, ao passo que o erro de teste é tendencialmente crescente. É possível observar que na transição de lambda de 0,63 (lambda.min) para 0,5, o programa obtém melhores valores de erro de treino, mas sofre um incremento no número de variáveis a considerar, sendo os seus valores quase desprezáveis em comparação com o valor atribuído a IC. No entanto, esse valor deixa de ser 0 e passa a ser considerado. Como a diferença no erro de treino não é suficientemente significativa para ser considerada, o programa vai considerar o 0,63 como o lambda.min. De notar também que o lambda.min apresenta um menor erro de teste que o lambda 0,5.

Foram também analisados casos retirando algumas variáveis do grupo em análise pois não se mostraram relevantes.

Este estudo inicial foi complementado com os que se descrevem a seguir. Os dados de entrada e de saída são conhecidos desde o ano de 2002 até ao ano de 2017. Posto isto, foram feitas diversas análises e sorteado o ano de teste deste período de dados para perceber se as variáveis mais importantes se mantinham ou não.

Inicialmente, as análises foram feitas com o ano de 2002, que se veio a revelar um ano atípico e desta forma, também por ser no início dos dados, foi desconsiderado. Existem outros anos, como 2005 e 2009 (anos de eleições) e 2016 (outlier). No entanto, para uma melhor robustez no método, foram considerados, pois, em termos práticos existirão condições fora do padrão e os exemplos históricos também já são reduzidos e é necessário aproveitar esses dados. O estudo que vai ser exposto na secção 3 foi realizado com início no ano de 2003 e a acabar em 2015 e 2016. Não foi feito até 2017 pois é necessário deixar sempre algum ano para teste

de forma a perceber se o modelo simulado com os anos anteriores apresenta resultados próximos do real.

Outra situação analisada foi considerando desde o ano de 2003 até 2014 para teste. Contudo esse caso também vai acabar por ser desconsiderado pois observa-se que o modelo resultante vai apresentar um erro de treino consideravelmente maior devido à falta de dados de entrada. Nestas condições, a sub-adaptação vai corromper o modelo obtido em demasia. Como o número de anos em análise já é um número baixo à partida, é preciso ter especial cuidado para não descredibilizar o modelo devido à falta de exemplos.

Após análise das diferentes variações dos dados de entrada, os casos que prevaleceram foram:

- Treino com os anos de 2003 a 2015 e teste com 2016 e 2017;
- Treino com os anos de 2003 a 2016 e teste com 2017.

Capítulo 3

Resultados

Neste capítulo vão ser apresentados e discutidos os resultados obtidos para analisar o método do LASSO. O principal objetivo é comparar as abordagens em termos de desempenho e de complexidade.

As próximas secções descrevem os seguintes modelos testados:

3.1- Modelos obtidos sem recurso ao LASSO. Este estudo foi realizado numa abordagem inicial ao problema e foi feito para perceber o comportamento de diversos modelos com complexidades distintas, sem recurso ao LASSO. Nesta situação é percetível um dos grandes problemas quando se pretende ultrapassar: a grande divergência de estimativas. Estas "vassouras" revelam a dificuldade que existe em definir um critério que resulte num bom modelo de previsões dada a possibilidade de aparecerem modelos com diferentes tendências.

3.2- Modelos obtidos usando o LASSO com recurso ao R segundo as mesmas condições de 3.1. Esta análise visa ilustrar que o LASSO revela ser uma metodologia adequada para eliminar estas vassouras e que os modelos obtidos com recurso a esta metodologia revelam um comportamento mais homogéneo entre si.

3.3- Comparação dos pontos 3.1 e 3.2. É feita uma primeira análise comparativa entre estes dois pontos onde os modelos obtidos são resultado de um banco de dados de apenas cinco das variáveis independentes.

3.4- LASSO com Excel com todas as variáveis. Expõe-se a metodologia LASSO de uma forma mais simples. Recorre-se ao Excel para auxiliar neste processo onde a seleção das variáveis para o modelo é feita através de "forçar" a sua saída. Desta forma, é possível obter modelos que coincidem com o princípio do LASSO onde a sua implementação é feita de forma manual.

3.5- LASSO com o R considerando todas as variáveis. Este caso é similar ao 3.2 sendo que o que varia é o banco de variáveis disponíveis para formar o modelo. Em 3.2 tinha-se cinco variáveis e em 3.5 tem-se onze variáveis (todas as variáveis). Vai-se analisar as diferenças entre os dois casos e perceber a reação nos modelos obtidos com LASSO ao existir um incremento de variáveis.

3.6- Comparação de todos os modelos obtidos. Neste ponto é feita a analise comparativa de todos os melhores modelos obtidos. Todos os modelos têm características próprias e o modelo exposto para cada caso é o reflexo da média dos melhores casos dentro dessas condições. Também se expõe o comportamento futuro do modelo atualmente utilizado pela EDP.

Todos os modelos descritos nesta secção vão ter como base dois cenários:

- Treino desde 2003 até 2015 e teste 2016 e 2017;
- Treino desde 2003 até 2016 e teste 2017.

Após 2017, a avaliação dos modelos é maioritariamente especulação. No entanto, todas as variáveis dependentes utilizados pós 2017 foram projeções que resultaram de um estudo feito pelo INESCTEC, explicação detalhada no fim da secção 2.1. Também, todos os modelos apresentados nas figuras (curvas a preto) são referentes ao modelo real.

3.1- Modelos obtidos sem recurso ao LASSO

Os modelos apresentados nesta secção demonstram a dificuldade que existe em criar um modelo de raiz. Considerando que o problema aqui apresentado é recente, vai ser necessário fazer uma abordagem inicial vasta de modo a perceber qual o melhor caminho a seguir. Neste caso, pretende-se perceber se o caminho "LASSO" apresenta potencial para resolver o problema em questão.

Nestes exemplos, para além de regressões lineares, também vão ser considerados outros modelos de regressão. São exemplos que ilustram cinco variáveis que estão descritas mais à frente onde serão explorados os seus erros de treino e teste associados para tentar perceber se o modelo em questão apresenta valores que o tornem um modelo a considerar.

Os modelos de regressão testados foram os seguintes:

(1) $$Inv_n = {Ener_{n-1}}^{\alpha 1} \times {PIB_{n-1}}^{\alpha 2} \times {T.infl_{n-1}}^{\alpha 3} \times {T.des_{n-1}}^{\alpha 4} \times {I.Cons_{n-1}}^{\alpha 5} \times k_1 + k_2 \quad (3.1)$$

(2) $$Inv_n = Ener_{n-1} \times \alpha_1 + PIB_{n-1} \times \alpha_2 + T.infl_{n-1} \times \alpha_3 + T.des_{n-1} \times \alpha_4 + I.Cons_{n-1} \times \alpha_5 \times k_1 + k_2 \quad (3.2)$$

(3) $$Inv_n = (Ener_{n-1} \times \alpha_1 + PIB_{n-1} \times \alpha_2 + T.infl_{n-1} \times \alpha_3 + T.des_{n-1} \times \alpha_4 + I.Cons_{n-1} \times \alpha_5) \times k_1 + k_2 \quad (3.3)$$

(4) $$Inv_n = {Ener_{n-1}}^{\alpha 1} \times {PIB_{n-1}}^{\alpha 2} \times {T.infl_{n-1}}^{\alpha 3} \times {T.des_{n-1}}^{\alpha 4} \times {I.Cons_{n-2}}^{\alpha 5} \times k_1 + k_2 \quad (3.4)$$

(5) $$Inv_n = {Ener_{n-2}}^{\alpha 1} \times {PIB_n}^{\alpha 2} \times {T.infl_{n-2}}^{\alpha 3} \times {T.des_n}^{\alpha 4} \times {I.Cons_{n-2}}^{\alpha 5} \times k_1 + k_2 \quad (3.5)$$

(6) $$Inv_n = {Ener_{n-1}}^{\alpha 1} \times {T.des_{n-1}}^{\alpha 2} \times {I.Cons_{n-2}}^{\alpha 3} \times k_1 + k_2 \quad (3.6)$$

(7) $$Inv_n = {Ener_{n-1}}^{\alpha 1} \times {T.des_{n-1}}^{\alpha 2} \times {I.Cons_{n-2}}^{\alpha 3} \times \frac{\sum_{n-2}^{n} Inv}{3} \times k_1 + k_2 \quad (3.7)$$

(8) $$Inv_n = {Ener_{n-1}}^{\alpha 1} \times {T.des_{n-1}}^{\alpha 2} \times {I.Cons_{n-2}}^{\alpha 3} \times {Inv_{n-1}}^{\alpha 4} \times k_1 + k_2 \quad (3.8)$$

(9) $$Inv_n = Ener \times \alpha_1 + PIB_{n-1} \times \alpha_2 + I.Cons_{n-1} \times \alpha_3 + Inv_{n-1} \times \alpha_4 + k_1 \times k_2 \quad (3.9)$$

Onde:

- *n* -> ano em estudo
- *Inv* -> investimento
- *Ener* -> energia consumida
- *PIB* -> produto interno bruto
- *T.infl* -> taxa de inflação
- *T.des* -> taxa de desemprego
- *I.Cons* -> índice de construção
- *k´s e α's -> variáveis a estimar pelo modelo*

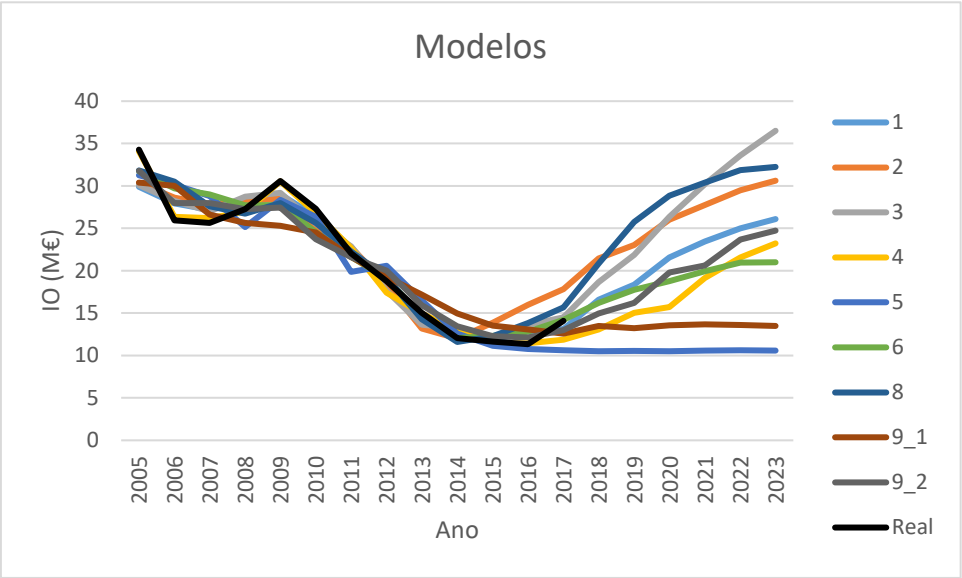


Figura 3.1- Gráfico do IO ao longo dos anos para os modelos resultantes das equações numerados acima.

Na figura acima apresentada, é possível observar a representação gráfica dos modelos obtidos. Consideremos que, todos os modelos no gráfico estão a ser iniciados em 2005 e não em 2003 pois o treino foi iniciado em 2003, mas algumas funções vão buscar variáveis referentes a dois anos anteriores ao que se está a analisar e desta forma o primeiro IO a ser previsto é o de 2005. De notar também, que o caso do modelo (7) não foi exposto pois a sua curva apresenta um crescimento exponencial excessivo e os erros de teste associados são superior a 1000%. Na legenda, é possível observar uma curva "9_1" e "9_2". Isto representa duas abordagens diferentes consideradas nesse modelo. Como o investimento no ano anterior também é como uma variável de entrada a considerar, um dos ensaios usa o valor real e outro usa o valor obtido na iteração anterior, respetivamente. No modelo (8), essa abordagem também foi feita, mas o que aqui está expresso é apenas o que recorre ao valor de IO já conhecido e não às previsões do próprio modelo pois esse, tal como aconteceu com o modelo (7), apresenta valores que não fazem sentido ser considerados.

Após obtida a curva de cada modelo para valores futuros, é importante observar o erro associado a cada um destes modelos e perceber se são modelos promissores ou não.

Tabela 3.1- Erro de treino e de teste dos modelos ilustrados na figura 3.1.

Erro (%)	Treino	Teste
1	5.75	10.25
2	7.54	33.75
3	6.66	11.44
4	2.49	8.31
5	8.55	14.91
6	6.23	6.90

8	6.04	16.55
9_1	11.16	13.12
9_2	7.01	7.36

As curvas representadas no gráfico e os valores obtidos na tabela são a média das duas curvas resultantes com os anos de treino de 2003 a 2015 e de 2003 a 2016 para o modelo em questão. Os valores de teste apresentados são valores considerando 2016 e 2017. Na maioria dos modelos, caso o período de teste abarcasse mais anos pós 2017, o erro de teste associado iria aumentar substancialmente

Analisando, em detalhe, três casos que apresentam melhores erros, todos com erros de treino e de teste inferiores a 10%.

Modelo (4): Observando a curva, primeiro, o ano de 2017 revela um grande erro, contudo como o ano de 2016 tem um valor praticamente idêntico ao real, vai compensar e diminuir assim o erro de teste. O comportamento do modelo mostra que o IO até 2019 tem um crescimento lento e inicia um crescimento evidente em 2020. Este desfecho não se enquadra com o previsto pelos índices socioeconómicos pois o que seria de esperar era uma estabilização e um crescimento progressivo. Isto acontece, pois, este modelo atribui bastante peso ao PIB e como podemos observar na tabela das variáveis normalizadas (tabela 2.2), o PIB aumento de forma bastante acentuada.

Modelo (6): Este modelo apresenta uma curva que revela crescimento seguido de uma estabilização. Aqui não se observa nenhum pico pois como podemos observar na expressão do modelo, o PIB não é considerado.

Modelo (9_2): Neste modelo também é percetível o impacto do PIB e analisando a curva ao longo dos anos, o crescimento associado é acentuado e em degraus, o que compromete a robustez do modelo.

Estes exemplos ilustram a dificuldade de obter modelos estáveis, num subsistema com muitas variáveis relativamente ao número de exemplos.

Os erros de treino e de teste são os valores mais importantes a considerar quando se está a criar um modelo de previsões, mas também é importante analisar o comportamento do modelo ao longo dos anos seguintes ao teste, tentar justificar as variações e perceber se faz sentido ser dessa forma.

3.2 - Modelos obtidos usando o LASSO com recurso ao R segundo as mesmas variáveis de 3.1.

Este cenário foi explorado para perceber se os modelos obtidos com o LASSO resultavam em modelos mais uniformes e semelhantes entre si, ao contrário do que se apresenta no ponto 3.1. O banco de dados considerado para este caso foi o mesmo no ponto anterior, ou seja, o programa teve disponível cinco variáveis que podia selecionar para definir o modelo. Recorreu-

se ao R por ser um processo mais rápido e de maior fiabilidade, onde a seleção de variáveis é feita de forma automática.

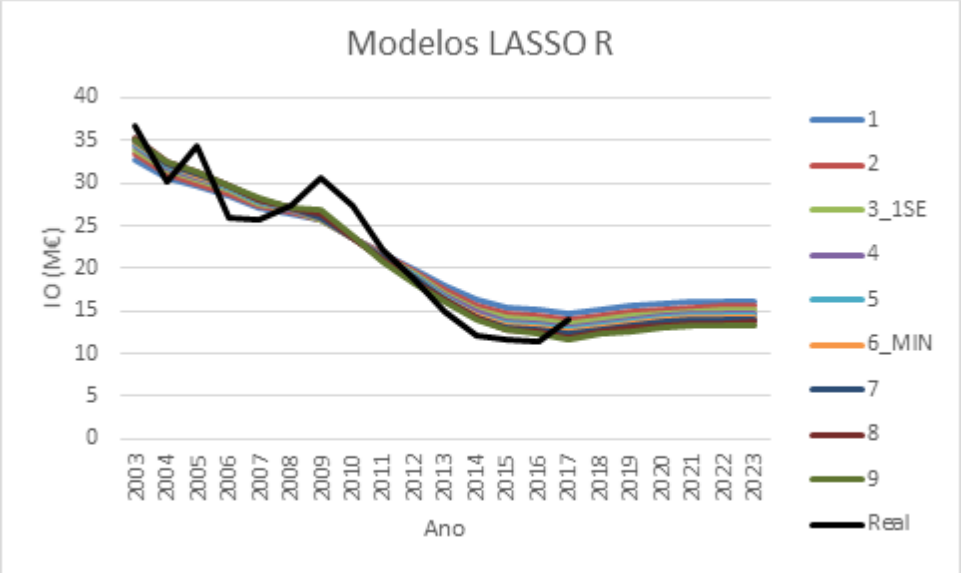


Figura 3.2- Gráfico do IO ao longo dos anos para os modelos do LASSO obtidos com recurso ao R

Através da observação do gráfico 3.2, dá para concluir que os modelos obtidos com o LASSO são mais homogéneos, isto é, mais coerentes entre si. A curva que descrevem varia em valor numérico, mas mantém uma evolução similar entre os casos. O modelo a ter em consideração para análise de previsões vai ser a média dos quatro modelos compreendidos pelo 3_SE, 4, 5 e 6_MIN pois, como podemos observar na tabela abaixo, esses vão ser os modelos que vão apresentar erros mais aceitáveis considerando o balanço entre o treino e o teste. Também convém reforçar que esses são os modelos que a própria biblioteca do LASSO defende como melhores.

Tabela 3.2- Erro de treino e de teste dos modelos ilustrados na figura 3.2.

Erro (%)	Treino	Teste
1	13.50	11.87
2	12.26	8.84
3_SE	11.27	9.68
4	10.49	10.95
5	9.87	11.95
6_MIN	9.37	12.75
7	9.00	13.39
8	8.86	14.54
9	8.65	15.83

Os modelos representados nesta tabela, são modelos que dependem apenas de uma variável, o IC. Esses mesmos modelos vão apresentar desde o 1 até ao 6_MIN valores de IC com

tendências crescentes sendo que o que contrapõe este aumento é também um aumento no valor negativo da ordenada na origem.

Só no modelo 7, 8 e 9, após passar o mínimo, é que obtemos modelos com duas ou mais variáveis. Podemos observar que, com o acréscimo de variáveis, o erro de treino vai melhorar, mas isso não se reflete no erro de teste. Quando isto ocorre é porque o coeficiente associado ao IC vai aumentar e a outra variável que vai aparecer tem um valor inicialmente muito baixo e que tenta ajustar o equilíbrio entre o IC e a ordenada na origem. Isto vai refletir uma maior adaptação ao treino, mas que não se vai manifestar no teste.

Segundo a biblioteca 'glmnet' utilizada para gerar este modelo a partir do LASSO, existe um modelo que apresenta o menor erro de treino com o menor número de variáveis. No entanto, o programa recomenda a utilização de um outro modelo que sofre um desvio padrão em relação ao ponto onde o mínimo foi atingido. É sobre o valor de IO gerado pelo modelo que este desvio é calculado e não sobre as variáveis. Posto isto, vai ser considerado um modelo que representa a média dos quatro modelos desde o 3 até ao 6. Não esquecer que o LASSO atribui uma forte penalização aquando do aumento do número de variáveis e que a função utilizada aplica validação cruzada o que faz com que os modelos resultantes sejam semelhantes.

3.3 – Comparação dos resultados dos pontos 3.1. e 3.2.

Esta secção compara uma média calculada a partir de modelos resultantes do LASSO (ponto 3.2.) com uma média de diversificado de modelos (ponto 3.1.).

Recordar:

- O ponto 3.1. foi feito com o intuito de perceber a dificuldade que existe em criar um modelo a partir do zero.
- O ponto 3.2. apresenta modelos utilizando o LASSO onde os resultados são fruto de apenas uma variável.

Ao considerarmos um caso com cinco variáveis e tendo o LASSO tendência para penalizar o aumento do número de variáveis, era de esperar um modelo com um número mínimo de variáveis o que pode implicar um erro elevado associado. Por outro lado, esse mesmo modelo é um modelo simples pois a sua dependência é de uma variável.

Vamos agora calcular a média dos 3 melhores modelos do ponto 3.1. que foram analisados em detalhe nesse mesmo ponto e a média dos 4 modelos resultantes do LASSO também como foi explicado em 3.2. O número de modelos a ser considerado na média não é o mesmo, mas se observarmos os resultados do LASSO, podíamos apenas considerar a média dos modelos lambda.min com o lambda.1se. Os outros dois modelos considerados encontram-se com valores intermédios entre estes dois e o comportamento das suas curvas é idêntico. Por isso, optou-se por considerar estes 4 modelos para uma melhor precisão dos valores gerados pela média dos modelos.

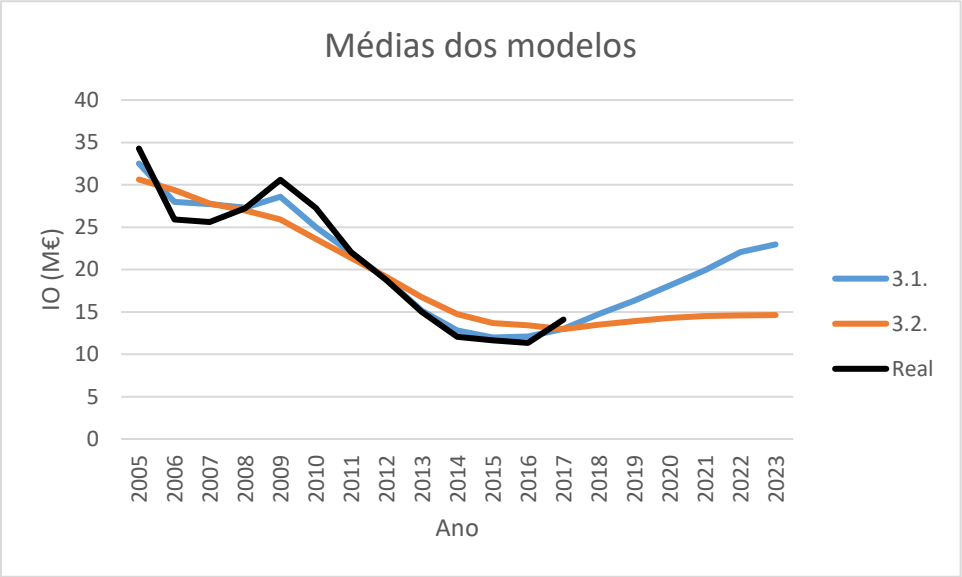


Figura 3.3- Representação gráfica das médias dos melhores modelos obtidos no ponto 3.1. e 3.2.

Observando as curvas do gráfico, a que melhor se parece enquadrar com a curva real é a 3.1. Os erros de treino e de teste associados a este modelo são inferiores aos gerados pelo LASSO (3.2.). Contudo, como explicado no ponto 3.1., os modelos que fundamentam esta média

aqui representada são modelos que isoladamente apresentam algumas falhas, no sentido do que é suposto ser a evolução da curva. No caso do LASSO, esta média de modelos é feita com base em modelos que só consideram a variável IC e a ordenada na origem sendo que o que se observa na sua curva, é o comportamento dessa variável ao longo do tempo. Esta média de modelos apresenta uma tendência que tende a estabilizar nos 15 (M€) e que apresenta uma tendência crescente desde o último ano de teste até à estabilização muito suave.

Desta forma, é de ter em consideração que o modelo 3.1. apresenta valores muito bons de erro de treino e de teste, mas o seu comportamento a nível de crescimento é comprometedor. Observando o gráfico, é visível uma evolução acentuada, pois dois das três curvas que compõem este modelo sofrem uma grande influência do PIB. Este comportamento é comprometedor e demonstra que o modelo se encontra demasiado preso às variáveis explicativas.

Quanto ao LASSO, é um modelo que não apresenta o melhor resultado, mas por outro lado não é de descartar. Ou seja, o modelo aqui expresso tem como base apenas uma variável e, por isso, é natural existir um erro associado, pois estamos a falar de um IO que tem diversos fatores a considerar. Para um modelo simplista, iniciado apenas com cinco variáveis de entrada e resultando num modelo que é uma relação direta a uma dessas variáveis, os valores que apresenta são relevantes e faz sentido aprofundar o modelo.

Os modelos que resultam em 3.1. são modelos que consideram IC, mas que a variável que mais se destaca é o PIB. O LASSO mostra algo completamente diferente pois diz que ao considerar apenas uma variável, a mais importante de considerar é o IC, o que mostra um resultado algo distinto.

O LASSO mostra ser um método de regressão que simplifica bastante a análise ao utilizador. É um modelo simples de traçar e minimalista. Desta forma, as variáveis que daí resultam são aquelas que vão ter uma maior relação com o output em causa. Nesse sentido, é um bom método a utilizar numa abordagem inicial a um problema de previsões.

3.4 – LASSO com Excel considerando todas as variáveis

Os valores apresentados foram obtidos partindo de uma regressão linear, explicado no ponto 2.2, tal como todo o processo que foi realizado. Após realizados os ciclos iterativos até à obtenção de apenas três variáveis independentes, a situação que se obtinha era sempre a mesma. O IC, NNC e NNF prevaleciam. Na grande maioria dos casos, os valores dos coeficientes associados a cada uma delas foi semelhante, o que dificulta a decisão de qual variável retirar para obter um modelo apenas com duas variáveis. Por isso, optou-se por, após obtenção das três variáveis com maior grau de importância, realizar uma combinação dessas mesmas três variáveis e gerar assim três modelos de duas variáveis em vez de apenas um.

Mesmo em abordagens que não vão ser aqui expostas, como anos de teste de 2003 a 2014 ou 2004 a 2015, o resultado é sempre o mesmo: IC, NNC e NNF.

Para se conseguir obter modelos diferentes, tratando-se este trabalho de um estudo e tendo em consideração o resultado anterior, foi adotada uma estratégia que consiste em "forçar" a saída de duas das variáveis importantes. Optou-se por retirar o NNC e o NNF pois os seus valores normalizados são parecidos com os de IC, o seu crescimento semelhante, as suas projeções foram realizadas com base em IC e o coeficiente de determinação, R^2, associado a IC é superior, de forma significativa, como foi demonstrado na secção 2.1. Daí, optar-se pela permanência de IC em prol de NNC e NNF.

Considerou-se também o facto de o LASSO, realizado através do R, resultar sempre em modelos onde IC está presente e é a variável predominante.

Os gráficos que vão ser agora apresentados são referentes aos dois ensaios realizados com anos de treino a começar em 2003 e a terminar em 2015 ou 2016.

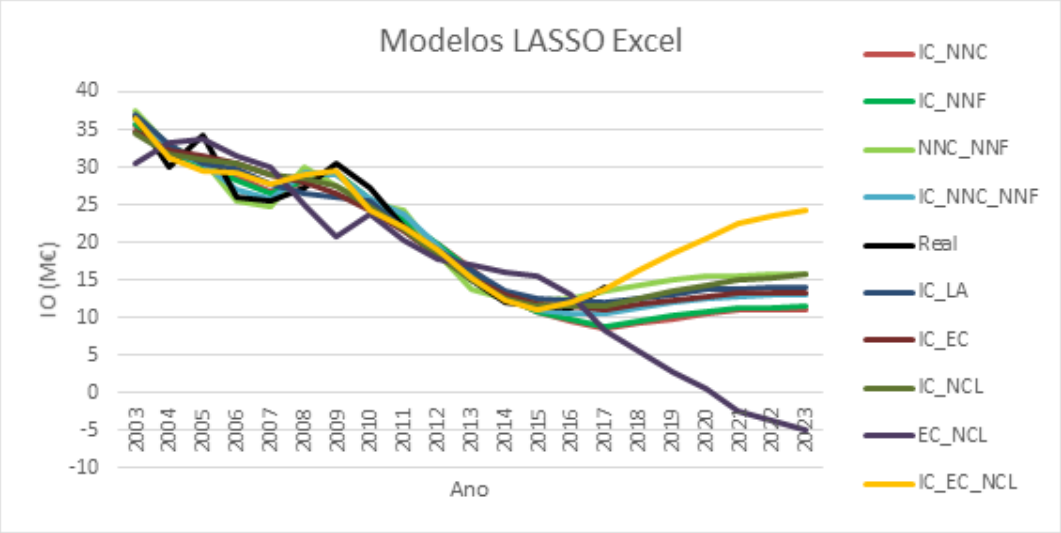


Figura 3.4- Modelos criados com recurso ao LASSO no Excel de forma manual relativos aos anos de treino de 2003 a 2016.

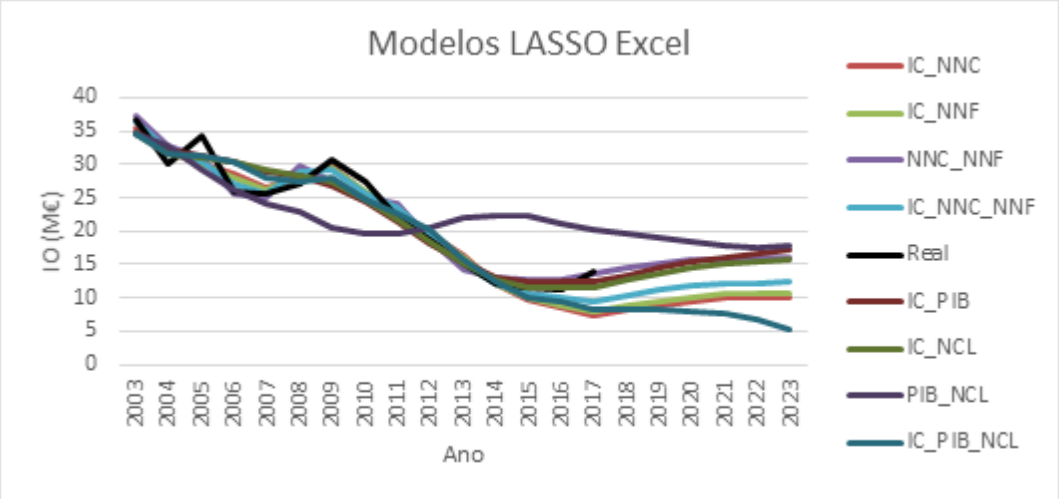


Figura 3.5- Modelos criados com recurso ao LASSO no Excel de forma manual relativos aos anos de treino de 2003 a 2015.

Observando os dois gráficos, onde é possível ver a curva dos modelos ao longo dos anos, existem alguns modelos que chamam à atenção por apresentarem curvas com picos e variações

acentuadas. De notar que alguns desses casos, menos coerentes, resultam aquando da criação do modelo considerando duas variáveis que não sejam IC, NNC ou NNF. Os modelos que apresentem curvas decrescentes muito acentuadas vão ser desconsiderados pois, atendendo ao comportamento das variáveis explicativas, a tendência é que este apresente crescimento.

Tabela 3.3- Erros de treino e de teste associados aos modelos ilustrados na figura 3.4 e 3.5.

03_16	Treino	Teste		03_15	Treino	Teste
IC_NNC	7.15	40.01		**IC_NNC**	6.19	36.62
IC_NNF	6.71	37.79		**IC_NNF**	5.53	32.43
NNC_NNF	6.50	4.23		**NNC_NNF**	6.33	7.52
IC_NNC_NNF	4.93	26.04		**IC_NNC_NNF**	4.82	22.15
IC_LA	7.54	15.36		**IC_LA**	7.05	11.13
IC_EC	6.97	21.37		**IC_PIB**	7.78	10.80
IC_NCL	6.22	18.72		**IC_NCL**	6.51	10.64
EC_NCL	16.30	40.52		**PIB_NCL**	27.18	65.42
IC_EC_NCL	5.45	2.60		**IC_PIB_NCL**	7.30	28.26

Os modelos que se destacam são os que apresentam um erro de teste inferior a 10%, que é o caso de NNC_NNF e IC_EC_NCL. Este último, como podemos observar pelo traçado da curva na figura 3.4 a amarelo, vai-se comportar de forma diferente da maioria dos modelos. A curva apresenta um crescimento muito acentuado após o teste e deve-se ao facto de o modelo depender de NCL e EC que como podemos constatar nos dados normalizados, tabela 2.2., secção 2.1., são valores que vão estar sempre a aumentar sendo que no caso específico de NCL, vai passar o limite histórico em todos os anos de projeções. De realçar que este modelo apenas aparece em uma das duas condições de treino. Quanto ao NNC_NNF, este parece ser um modelo promissor pois aparece nos dois testes e em ambos apresenta baixos valores de erro de treino e teste. Contudo, é importante ter-se em consideração que aquando da realização do LASSO com 3 variáveis IC_NNC_NNF, os resultados dos coeficientes associados foram iguais, mas com valores opostos para NNC e NNF. Se verificarmos o valor normalizado das variáveis, podemos observar que estes valores são idênticos. Ou seja, estas duas variáveis anularam-se e foi praticamente IC que traçou o modelo. De notar que, IC, NNC e NNF apresentam comportamentos idênticos em termos de histórico, após analisar as variáveis normalizadas.

O modelo NNC_NNF apresenta bons valores de erros e uma curva com uma tendência promissora. Este modelo vai ser analisado de forma mais detalhada na secção 3.6.

3.5- LASSO com o R considerando todas as variáveis

Este caso tem a mesma base que a situação referida no ponto 3.2, sendo que a principal diferença é o banco de variáveis disponível. Nesta situação, os modelos obtidos resultam de um banco de onze variáveis (todas as variáveis).

Em 3.2., ao serem atribuídas 5 variáveis de entrada o LASSO retorna um modelo baseado em apenas uma variável, IC. Nesta situação, são utilizadas 11 variáveis de entrada e o melhor modelo que o LASSO retorna é baseado em duas variáveis, IC e LA sendo que a que apresenta um maior peso é o IC.

Mais à frente, vai ser detalhado o valor destas variáveis e feita uma análise comparativa com outros modelos. Agora vai-se observar as curvas resultantes dos quatro modelos que foram considerados segundo estes critérios.

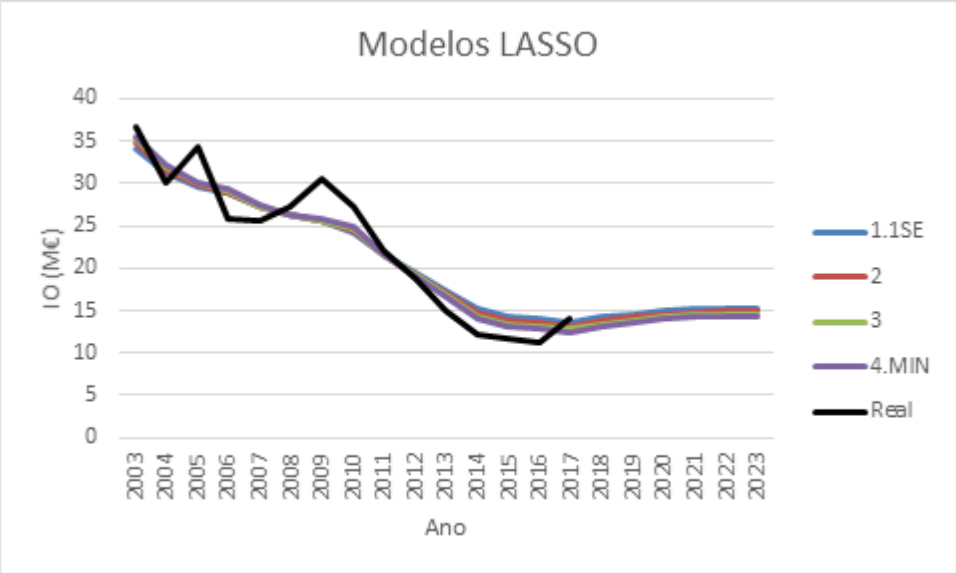


Figura 3.6- Média dos modelos gerados pelo LASSO considerando para teste 2016/2017 e só 2017.

O comportamento aqui representado é muito idêntico ao do ponto 3.2. pois o IC vai ser a variável que vai ter quase 90% de preponderância sobre o modelo. Com estes resultados, podemos concluir que o LASSO, das variáveis que foram adicionadas, apenas considerou como uma mais valia para adicionar ao modelo o LA. Isso vai ser visível analisando os erros associados a estes dois casos.

Tabela 3.4- Erro de treino e teste relativo à média dos melhores modelos obtidos no ponto 3.2. e no ponto 3.4.

Erro (%)	Treino	Teste
3.2.	10.00	13.22
3.4.	9.68	12.67

Observando os valores dos erros, percebe-se que a variação do erro nos dois casos não é muito acentuada e isso acontece devido ao peso que IC tem no caso 3.4..

Posto isto, o LASSO apresenta resultados promissores no que diz respeito ao delineamento das variáveis com maior impacto e quais se devem considerar. O modelo resultante revela um aumento da performance muito pequeno, mas é importante referir que esse aumento podia nem ser possível de obter. Desta forma, pode observar-se o impacto e a relevância de IC na previsão de IO.

Contudo, dadas as características deste tipo de problema de previsões, a estratégia seguida neste estudo parece ser uma alternativa viável, dado ser de fácil utilização e por ser eficiente.

3.6- Comparação de todos os modelos obtidos

Aqui vão ser expostos e analisados os melhores modelos dos tópicos abordados anteriormente e adicionado também o modelo de previsões que é atualmente utilizado pelo operador de rede de distribuição.

No que remete ao capítulo 3.1., o modelo que vai ser selecionado é a média dos 3 melhores modelos que foram apresentados.

Referente ao ponto 3.2. foi feita a média dos quatro modelos como explicado nesse mesmo ponto.

No ponto 3.3., só um modelo vai ser selecionado pois os resultados que apresenta são bastante mais promissores que os outros, esse modelo é o que foi criado utilizando NNC e NNF.

No ponto 3.4., tal como em 3.2., o modelo aqui colocado é a média dos quatro modelos que retornam utilizando a biblioteca do R referente ao LASSO.

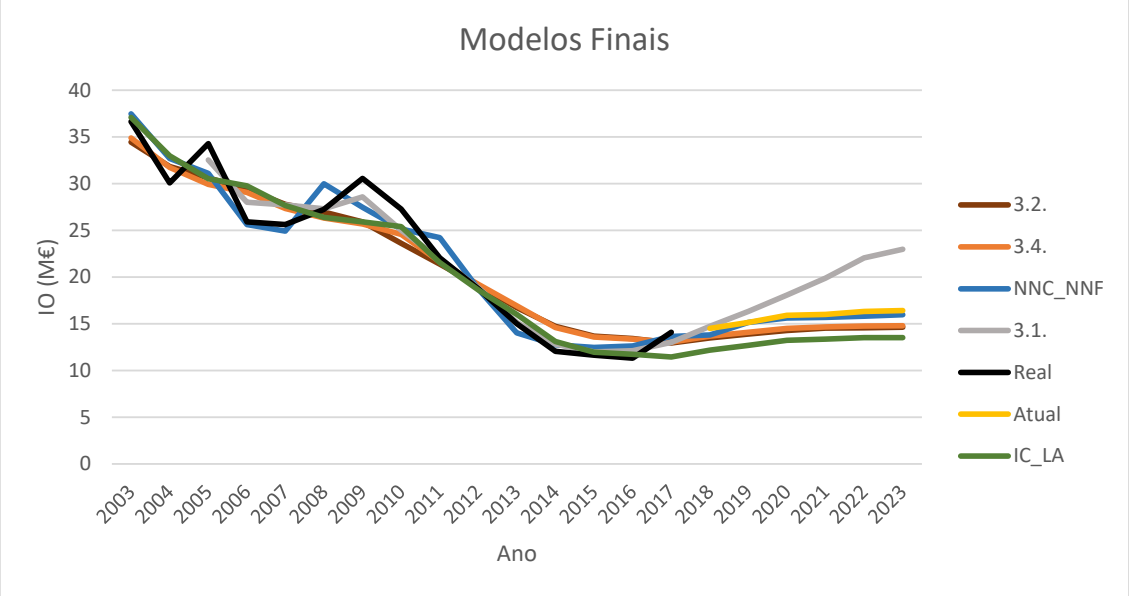


Figura 3.7-- Evolução de diferentes modelos obtidos para a previsão do IO ao longo do tempo

Os modelos 3.2. e 3.4. são modelos resultantes da regressão LASSO com recurso ao R.

O modelo NNC_NNF é um modelo que aparece com a implementação do LASSO no Excel de forma manual, como foi explicado anteriormente.

O IC_LA é um modelo que foi aqui adicionado para perceber o que resulta aquando da utilização de apenas estas duas variáveis para a criação de um modelo como aconteceu com o caso de NNC_NNF, com o auxílio do Solver do Excel. Dado que este é o modelo que a biblioteca do R retorna, o objetivo foi perceber o comportamento e os valores que os coeficientes destas variáveis tomam em comparação com os que foram gerados automaticamente.

O modelo 'Atual', como o nome indica, é o modelo que é atualmente utilizado para a previsão dos valores do IO.

Tabela 3.5- Erro (%) dos modelos associados à curva do gráfico 3.7.

Erro %	Treino	Teste
3.2.	10.00	13.22
3.4.	9.68	12.67
NNC_NNF	6.27	7.33
3.1.	4.30	7.27
IC_LA	7.00	11.09
Atual	7.41	9.09

Observando o gráfico, a primeira curva que sobressai é a 3.1. pois revela um comportamento diferente de todas as outras. O que levou a questionar se este modelo é um bom exemplo a considerar. Por outro lado, observando a tabela dos erros (%), este mesmo modelo é o que apresenta melhores valores de erro (valores mais baixos). Contudo o seu crescimento aparenta ser demasiado acentuado em relação ao que se espera. É um modelo que suscita dúvidas e que define uma curva com uma tendência diferente de todas as outras. Este modelo revela a dificuldade existente em definir um critério que resulte num bom modelo de previsão dada a possibilidade de aparecerem modelos com tendências divergentes. Note-se que poderiam ter sido definidos critérios relativos à evolução futura, de modo a filtrar os casos com divergência acentuada face à tendência média.

Os dois modelos, 3.2. e 3.4., são muito semelhantes. É possível observar um melhoramento quando se adiciona mais variáveis, mas é uma percentagem mínima no erro e o comportamento da curva é muito semelhante, tanto que as curvas parecem sobrepostas no gráfico. Os erros resultantes destes modelos são os que apresentam os resultados mais altos, o que leva a questionar se são modelos viáveis em comparação aos outros. Contudo, o comportamento da curva apresentada é muito semelhante ao modelo atual sendo que o que varia é o valor absoluto do modelo atual ser mais alto ao longo dos anos. É importante reforçar que se espera que o IO tenha uma tendência crescente nos anos posteriores ao teste e, nesse sentido, este modelo segue essa tendência mas parte de um valor que, devido a que resulta das suas previsões de teste, está abaixo do real e então o erro associado a essa diferença vai-se manter com o decorrer do tempo.

O modelo IC_LA foi um modelo criado para perceber a diferença obtida no modelo gerado com as variáveis que o R selecionou (caso 3.4) mas implementado no Excel. Pode-se observar

na tabela que os erros associados a este modelo são mais promissores que os gerados pelo R mas ao interpretar o gráfico, é percetível que a curva descrita pelos dois modelos é idêntica. Ou seja, o seu comportamento futuro é semelhante e isto faz todo o sentido visto estarmos a falar das mesmas variáveis, o valor do coeficiente atribuído a cada uma delas nos dois casos ter sido diferente. Este modelo parece resultar em previsões de IO um pouco abaixo do que seria de esperar tendo em consideração o modelo atual e sendo este o modelo que revela o menor valor absoluto ao longo de todos os anos de previsão. No entanto, a verificação destas conclusões apenas seria possível com a utilização dos dados reais nos anos subsequentes ao Teste.

O modelo NNC_NNF, feito de forma manual utilizando o Solver do Excel segundo o princípio do LASSO, apresenta um resultado muito idêntico ao utilizado atualmente sendo as suas curvas muito semelhantes. É um modelo que se revela bastante promissor, tanto a nível comportamental (evolução da curva no gráfico) como na análise numérica (erros associados ao modelo), sendo o que apresenta os melhores valores. No entanto, a metodologia LASSO ao ser implementada de forma manual, ou seja, a seleção das variáveis ter sido feito com base na análise do utilizador e "forçada" a saída daquela que se considerou menos relevante, faz com que este modelo não revele tanta fiabilidade como um modelo LASSO que recorreu exclusivamente ao R, onde todo o processo foi feito de forma automática e computacional.

Capítulo 4

Conclusões e Desenvolvimentos Futuros

Neste capítulo final são expressas as principais conclusões a extrair da pesquisa e do estudo desenvolvido, bem como os diversos obstáculos sentidos na sua execução. Para serem considerados em trabalhos futuros, na última secção, são enunciados um conjunto de sugestões possíveis para melhorar as estimações e os processos. Ainda nesta última secção são propostos vários complementos e modificações ao programa desenvolvido.

4.1 - Conclusões Finais

O principal objetivo de desenvolver modelos de previsões que permitam ao ORD obter estimações de IO por um período temporal de cinco anos, sendo o principal foco os primeiros dois anos, foi cumprido.

O LASSO revelou ser uma metodologia interessante na abordagem a este problema pois devido à limitação do histórico de dados de entrada e ao extenso número de variáveis independentes, é uma mais valia retirar as variáveis redundantes. É necessário ter especial atenção ao balanço eficiência-simplicidade visto a relação entre elas ser inversa. No entanto, considerando os modelos obtidos com o LASSO comparativamente ao modelo atualmente usado pela EDP, estes não são de descartar no que diz respeito ao valor de IO previsto e são modelos com um menor número de variáveis. O processo por trás da criação de um modelo, devido ao elevado número de variáveis, pode ser muito exaustivo e o LASSO revela-se uma solução para esse problema. Isto torna também o modelo gerado, de uma maior facilidade de análise. No que diz respeito à eficiência, não se revela ser o mais preciso pois o erro de treino é superior a outro tipo de modelos. Contudo, essa falta de precisão está associada à simplicidade do modelo. O LASSO revela ser uma metodologia promissora, pelo menos numa abordagem inicial a um problema de previsões, permitindo obter a preponderância das variáveis importantes e a descredibilização das redundantes.

A principal dificuldade para obter um bom modelo de previsões é definir o critério objetivo que melhor se enquadra com a metodologia a abordar. Uma forma de definir esse critério é

através da relação variância-tendência de forma a tentar minimizar ao máximo ambas e conseguir um equilíbrio. Desta forma, o LASSO também revela ser uma metodologia a explorar, dado tratar-se de um projeto que implica uma renovação ao longo do tempo. Ao existir a necessidade de definir um novo modelo, com a adição de novos dados, vai ser compensada uma das falhas desta metodologia que é a eventual sub-adaptação.

4.2- Desenvolvimentos Futuros

Visto ser um estudo relativamente recente, existem ainda diversas metodologias por testar. O facto de o histórico ser reduzido faz com que algumas metodologias fiquem limitadas, mas com o decorrer dos anos, a quantidade de dados disponíveis é cada vez maior e vai permitir outro tipo de abordagens.

Para tentar colmatar a dificuldade na seleção de bons modelos, pode ser realizado um estudo que em vez de recair sobre o modelo resultante, assente sobre a forma como definir um bom modelo dada a situação particular deste problema. A título de exemplo, sugere-se a análise se o impacto é maior caso a estimação seja acima ou abaixo do valor real pois ambas as situações requerem respostas diferentes.

Quanto a metodologias, seria interessante utilizar um modelo de previsões entre o Ridge e o LASSO, visto realizar à mesma uma penalização às variáveis e colocá-las a zero. No entanto, não vai ser tão “radical” como o LASSO nem tão “benevolente” com o Ridge, resultando assim num modelo que provavelmente vá perder um pouco de simplicidade em prol da eficiência.

Referências

[1] Decreto-Lei n.º 29/2006 de 15 de fevereiro. Diário da República n.º 33, Série I-A de 2006-02-15. Ministério da Economia e da Inovação. Lisboa.

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

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

[4] Ng, A.Y., Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization. [https://www.coursera.org/learn/deep-neural-network]

[5] Hawkins, D. M., "The problema of overfitting", Journal of chemical information and computer sciences. 2004

[6] Everitt, B.S. e Skrondal, A., Cambridge Dictionary of Statistics, Cambridge University Press. 2010

[7] Gruber, M., Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators, CRC Press. 1998

[8] Tibshirani, R., "Regression Shrinkage and Selection via the lasso", Journal of Royal Statistical Society. Serie B (methodological). 1996

[9] Macedo, P. M. F., Estimação de Investimentos na Rede de Distribuição, Dissertação realizada no âmbito do Mestrado Integrado em Engenharia Electrotécnica e de Computadores Major Energia, Faculdade de Engenharia da Universidade do Porto. 2014

[10] Rose J. e Dierker Lisa, Data Analysis and Interpretation Specialization. Wesleyan University [https://www.coursera.org/specializations/data-analysis]

[11] Fidalgo, N. e Monteiro, C. Aula 2 - Modelos de regressão. Técnicas para Previsão, Faculdade de Engenharia da Universidade do Porto, 2013.

[12] "Rstudio". Disponível em https://rstudio.com/products/rstudio. Acesso em 12/04/2020

[13] Friedman J., Hastie T., Tibshirani R., Narasimhan B., Tay K., Simon N. e Qian J., Package 'glmnet'. R package. 2020 [https://cran.r-project.org/web/packages/glmnet/glmnet.pdf]

[14] Hastie, T. e Qian J., An Introduction to glmnet. R package. 2016 [https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf]

[15] Paradis, E., R for Beginners. Montpellier University. 2015 [https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf]

[16] R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2005 [http://www.R-project.org]

[17] Machlis, S., Beginner's guide to R: Introduction. 2017 [https://www.computerworld.com/article/2497143/business-intelligence-beginner-s-guide-to-r-introduction.html]

[18] Peng, R. D., R Programming. Johns Hopkins University. 2020 [https://www.coursera.org/learn/r-programming]

[19] Ismay, C. e Kim, A. Y., Statistical Inference via Data Science: A ModernDive into R and the Tidyverse. CRC Press. 2019

[20] Torres-Reyna, O., Introduction to RStudio. Princeton University. 2013 [https://dss.princeton.edu/training/RStudio101.pdf]

[21] James G., Witten D., Hastie T. e Tibshirani R., An Introduction to Statistical Learning: with Applications in R. Springer. 2013