Previsão Eleitoral para a Assembleia da República Portuguesa

Diamantino Azevedo, Luís Correia, Graça Gaspar
LabMAg, Departamento de Informática,
Faculdade de Ciências, Universidade de Lisboa, Portugal
dsazevedo@gmail.com, {luis.correia, gg}@di.fc.ul.pt

Resumo

Em trabalho anterior utilizaram-se técnicas de *Data Mining* para predizer resultados eleitorais, sem utilizar sondagens, recorrendo a variáveis socioeconómicas, disponíveis publicamente sobre Portugal, no período abrangido pelas treze eleições para a Assembleia da República, entre 1974 e 2009. No entanto, o espectro político considerado nesse trabalho não abrange os 100% dos votos expressos, mas apenas os quatro partidos com assento parlamentar regular desde 1975 cuja votação atinge cerca de 84%. Na abordagem anteriormente adoptada, cada um dos quatro partidos tradicionais foi tratado separadamente, resultando em previsões independentes. Neste artigo analisa-se a extensão desse trabalho à previsão do intervalo restante dos resultados eleitorais e a sua utilização para garantir a restrição de que a percentagem total de votos expressos soma 100%. Os resultados mostraram que os métodos anteriormente aplicados permitem obter previsões com resultados de qualidade similar para o conjunto das forças partidárias complementares.

palavras-chave: correlação, assembleia da república, previsão eleitoral, variáveis socioeconómicas.

Abstract

In earlier work, Data Mining techniques using a large number of public available socioeconomic variables about Portugal were used to predict election results between 1974 and 2009 (thirteen elections) for the Parliament of the Portuguese Republic. However, the political spectrum considered in that work does not cover 100% of the votes cast, but only the four parties with regular parliamentary seat since 1975, whose voting reaches about 84%. In the approach previously adopted, each of the four traditional parties was treated separately resulting in independent forecasts. In this paper we analyse the extension of that work to the forecasting of the remaining interval of the election results and its use to ensure the restriction that the total percentage of votes cast adds up to 100%. The results showed that the methods previously applied allow making predictions with similar quality for the remaining partisan forces.

keywords: correlation, parliament, electoral forecast, socioeconomic variables

1. Introdução

Portugal gere-se pela Constituição da República Portuguesa, documento proveniente da Assembleia Constituinte de 1976, que sofreu a última revisão em 2005. Nela são enunciados os múltiplos procedimentos eleitorais que regulam a democracia em Portugal, nomeadamente as eleições para a Assembleia da República (AR). Com base nesses resultados eleitorais forma-se o Governo, pois, são eles que sustentam a nomeação do Primeiro-ministro e a consequente formação do Gabinete Governativo.

Recentemente, foram apresentados exercícios de previsão eleitoral [Azevedo, 2012] [Azevedo, et al., 2012] que utilizaram 654 variáveis Socioeconómicas e Ambientais (SE), obtidas de repositórios públicos (http://www.pordata.pt), para prever os resultados das eleições para a Assembleia da República.

Os modelos obtidos permitiram a previsão individual da percentagem de votos expressos nos quatro partidos mais antigos com assento parlamentar (http://www.cne.pt), Partido Socialista (PS), Partido Social Democrata (PSD), Partido Popular (PP) e Partido Comunista Português (PCP), os únicos que obtiveram resultados que permitindo eleger deputados para aquela Assembleia em todas as eleições realizadas entre 1974 e 2009.

A escolha deste indicador específico, a percentagem de votos expressos, prende-se com o facto de as alternativas, "número de deputados", ou "número de votos expressos", serem muito dependentes de factores externos, como, por exemplo, o método de atribuição de deputados e as taxas de abstenção.

A proposta baseava-se em composições lineares entre um conjunto reduzido de variáveis SE, seleccionadas automaticamente por correlação com os valores das votações obtidos nas eleições decorridas desde 1975. O conjunto desses valores partidários, que foram sujeitos a um tratamento semelhante mas independente e simultâneo, não considerava a votação noutros partidos nem o intervalo entre a soma dos resultados daqueles quatro partidos e os 100%. Na realidade, nunca se verificou, em qualquer fase daquela proposta, o cumprimento deste limite.

Pretende-se então aplicar o mesmo método, considerando também os votos noutros partidos e a restrição de 100% para o total.

Objectivos e algumas referências relevantes serão apresentados na próxima secção. Na terceira secção iremos explicar e descrever algumas das metodologias utilizadas. Na quarta, apresentaremos os resultados e, na quinta, apontaremos as conclusões.

2. Objetivos

Previsões políticas e eleitorais são relativamente frequentes em democracia [van der Brug, et al., 2007], seja utilizando sondagens ou inquéritos, geralmente aplicando amostragens [Magalhães, et al., 2011], quer por métodos que aplicam sensibilidade social e política, (*educated guess*) [Nadeau, et al., 1995], ou ainda por técnicas mais elaboradas, divididas em duas grandes famílias: a mais comum, em que se faz essa predição com base em desempenhos e por vezes conhecida como abordagem explicativa [Aguiar-Conraria, et al., 2008] [Caleiro, 2012]; e outra, em que se pretende prever

resultados utilizando variáveis disponíveis a maior ou menor distância dos eventos, designada frequentemente por abordagem preditiva.

A abordagem explicativa, frequente desde meados do seculo passado [Key Jr., 1966] [Fiorina, 1981], assenta prioritariamente no pressuposto de que as eleições servem como julgamento pelos eleitores do desempenho do governo. Os modelos propostos nesta abordagem usam variáveis escolhidas heuristicamente com cariz político ou social que se reconhecem ser de alguma influência em eleições. Estão nesta categoria o Produto Interno Bruto, o Desemprego e a Inflação. Complementarmente, utilizam-se outras variáveis, nomeadamente a existência de conflitos institucionais, a partilha de ideologia politica com o Presidente da República, reeleições, recandidaturas do líder no partido incumbente, etc. que assumem habitualmente formatos binários. Recentemente, alguns trabalhos em Portugal [Magalhães, 2011] e em Espanha [Aguiar-Conraria, et al., 2011] utilizaram variações destas formulações.

Esta previsão explicativa identifica geralmente o partido "vencedor" por oposição a um outro "perdedor". Consequentemente, é frequente analisarem-se poucos partidos do espectro político, pois a quantidade daqueles que possuem realmente possibilidades de vencer as eleições é reduzida, não sendo, por isso, considerados todos os concorrentes.

A abordagem preditiva propõe efectuar previsões sem procurar efeitos explicativos nas variáveis, ou informação utilizada, e procura prever resultados eleitorais, apenas com base em análises estatísticas automáticas. Das informações disponíveis, utiliza aquelas que melhor contribuem para aquele objectivo. Em oposição à filtragem prévia das variáveis assente em critérios heurísticos pré-determinados, é utilizada uma filtragem assente numa selecção regulada pelo objectivo. Na nossa proposta essa filtragem é baseada na correlação entre a variável alvo e as variáveis SE disponíveis.

A consequência inerente à abordagem explicativa, uma explicação qualitativa dos factos ocorridos, é substituída pela previsão quantitativa dos resultados eleitorais, seleccionando e filtrando as variáveis que melhor se comportam. Sendo assim, é possível utilizar a mesma técnica preditiva com partidos cujos resultados não lhe permitem concorrer ao poder sem ser por intermédio de coligações.

Além disso, aplicando a mesma metodologia, é possível obter uma previsão sobre a composição geral dos resultados eleitorais abrangendo a totalidade dos partidos. Na sequência do trabalho anterior, adopta-se neste artigo uma abordagem preditiva para previsão do remanescente da percentagem dos votos expressos.

Desde as primeiras eleições em 1975, além dos quatro partidos tradicionais, nenhuma das restantes forças partidárias se apresentou a todos os treze actos eleitorais realizados (até 2009). Este facto impossibilitou um processamento individual similar ao anterior e fez surgir a hipótese de prever a soma do conjunto dos votos expressos para o total dos restantes partidos concorrentes às eleições.

Consequentemente, definiu-se um novo valor a prever que consiste na diferença entre a percentagem total de votos nos quatro partidos tradicionais e o total dos 100%. Neste trabalho esse valor foi designado por "5p", como se de um quinto partido se tratasse.

3. Metodologia

3.1. As variáveis SE

Utilizaram-se os dados presentes no *website* www.pordata.pt que a Fundação Francisco Manuel dos Santos dedica à disponibilização de dados estatísticos sobre Portugal.

Dele obteve-se um conjunto de 654 variáveis, divididas em Temas. Indica-se para cada tema o número de variáveis disponíveis à data da recolha que cumpriam os requisitos terem valores anuais, desde 1974 até 2009, sem valores omissos:

- Ciência e tecnologia – 2;
- Contas do Estado – 29;
- Contas Nacionais – 38;
- Cultura – 17;
- Educação – 132;
- Emprego e mercado de trabalho – 72;
- Empresas – 9;
- Justiça – 9;
- População – 202;
- Protecção Social – 32;
- Rendimento e despesas familiares – 19;
- Saúde – 6;
- Território e ambiente – 87

.

3.2. Previsão incluindo um "quinto partido"

A anterior abordagem considerava apenas os quatros partidos tradicionais. Ao estender esta previsão ao quinto partido, esta abordagem pode ser confrontada agora com a seguinte questão processual:

"Como respeitar o limite de 100% a que o tipo de resultados eleitorais utilizados obriga?"

Sendo o valor real da "votação" no quinto partido a diferença entre a soma dos valores partidários individuais e o limite de 100%, uma solução simplificada poderia passar pela utilização da soma dos valores de previsão dos resultados eleitorais dos quatro partidos e pela identificação da diferença com a previsão do "5p". Obviamente, esta aproximação implicava o respeito absoluto pelo limite. Note-se porém que, caso as previsões dos quatro partidos sejam previsões por excesso, adotado esta solução, obter-se-ia uma votação negativa para o "5p".

No entanto, nada indica que não possam ser considerados os métodos anteriormente propostos, reutilizando-os para prever o resultado eleitoral para o "5p", como se de um partido real se tratasse. Essa foi a via adoptada, sabendo que o eventual problema da diferença para os 100% teria de ser resolvido.

3.3. Agregações

Tal como no trabalho anterior [Azevedo, 2012], foi necessário agregar valores de anos consecutivos, uma vez que cada variavel SE possuia 35 valores anuais, existindo apenas treze actos eleitorais no mesmo espaço de tempo, entre 1974 e 2009.

As formas de agregação de valores definidos levam em conta a semântica do problema, combinando os valores dos anos entre eleições. Para tal foram definidos três períodos de agregação:

- MF (memória fixa) que engloba os últimos quatro anos antes de cada acto eleitoral;
- MC (memória curta) que abrange os anos desde o último ato eleitoral, sendo diferente da anterior sempre que existe eleições antecipadas
- ML (memória longa) que engloba todos os anos desde 1974.

3.4. Fórmulas de agregação

Os diversos tipos de agregações resultam da utilização de uma forma particular de combinação dos valores:

- Média;
- MédiaW (Média em que cada um dos anos anteriores é ponderado de forma inversamente proporcional à sua distancia ao ano da eleição);
- MédiaW2 (Média em que cada um dos anos anteriores é ponderado de forma inversamente proporcional ao quadrado da sua distancia ao ano da eleição);
- Mediana;
- Declive (da recta de regressão do atributo face a um período temporal);
- Diferença (entre o primeiro e o último valor do período temporal).

Estas são aplicadas sobre um dos períodos de influência anteriormente descritos, isto é, o intervalo de anos abrangido naquilo a que chamamos as diferentes memórias.

Nas fórmulas a seguir apresentadas são utilizadas as seguintes variáveis:

v_i – valor da variável SE a agregar, no ano i;
n – ano da eleição corrente;
j – número de anos entre a eleição corrente e a anterior;
$\mathcal{M}$ – tipo de memória, ou período de anos a agregar, com valores possíveis: MF, ML ou MC;
$I_{\mathcal{M}}$ – primeiro ano incluído na agregação, onde $\mathcal{M}$ representa um dos tipos de memória acima referidos, sendo:

$$I_{\mathcal{M}} = \begin{cases} n\text{-}4 \text{ se } \mathcal{M}=MF \\ 1974 \text{ se } \mathcal{M}=ML \\ n\text{-}j \text{ se } \mathcal{M}=MC \end{cases}$$

$$\mathcal{M}\text{Média} = \frac{\sum_{i=I_{\mathcal{M}}}^{n-1} v_i}{(n-1)-I_{\mathcal{M}}+1} \quad (1)$$

$$\mathcal{M}\text{MédiaW} = \frac{\sum_{i=I_{\mathcal{M}}}^{n-1} \frac{v_i}{(n-1)-i+1}}{(n-1)-I_{\mathcal{M}}+1} \quad (2)$$

$$\mathfrak{M}\text{MédiaW2} = \frac{\sum_{i=I_{\mathfrak{M}}}^{n-1} \frac{v_i}{((n-1)-i+1)^2}}{(n-1)-I_{\mathfrak{M}}+1} \quad (3)$$

$$\mathfrak{M}\text{Mediana} = \text{Mediana}(V_{n-1};\cdots;V_{I_{\mathfrak{M}}}) \quad (4)$$

$$\mathfrak{M}\text{Declive} = \frac{\sum_{i=I_{\mathfrak{M}}}^{n-1}(v_i-\bar{v})(i-\bar{i})}{\sum_{i=I_M}^{n-1}(i-\bar{i})^2} \quad (5)$$

$$\mathfrak{M}\text{Diferença} = V_{n-1}-V_{I_{\mathfrak{M}}} \quad (6)$$

Considerou-se como forma adicional de agregação, a utilização do valor da variável SE num único ano, o ano anterior à eleição corrente, designada por MA1, (Memória Actual):

$$MA1 = v_{n-1} \quad (7)$$

Cada um dos esquemas de fórmulas (1) a (6) é aplicado aos três tipos de memória, gerando dezoito agregações diferentes a que se junta a (7), totalizando dezanove.

3.5. Normalização

As amplitudes dos valores das variáveis socioeconómicas originais eram muito diversas, indo, frequentemente, desde décimas a milhões.
Por isso foram convertidas, incluindo as variáveis dependentes [Marquardt, 1980], utilizando-se a normalização "média/desvio padrão" pelo que qualquer uma passou a ser representada por um atributo, com média igual a zero e desvio padrão igual a um.

3.6. Correlação

3.6.1. Das variáveis SE com a variável eleitoral

Partindo dos dados SE originais, as agregações aplicadas geraram um grupo alargado de atributos que as representavam segundo os conceitos descritos. Obviamente que alguns teriam mais influência ou seriam mais determinantes para a previsão do resultado partidário que outros, o que implicou a utilização de um filtro para as extrair. Utilizou-se a correlação como forma de filtragem.

Essa filtragem foi realizada em duas etapas. A primeira seleccionou os dez atributos com maiores correlações absolutas em cada uma das dezanove agregações. Com isso obteve-se um grupo de 190 atributos.

A etapa seguinte seleccionou, do grupo anterior, vários conjuntos com um número reduzido de atributos. Alguns testes iniciais [Azevedo, et al., 2012] indicaram que, de entre conjuntos com dimensões de dez, cinco e um único atributo, o que apresentava previsões mais próximas dos valores da variável alvo era o de cinco atributos. Deste

modo, utilizou-se prioritariamente esta dimensão, mantendo-se presentes, no entanto, um conjunto de dez e outro de um, cujos resultados finais vieram a confirmar os testes preliminares sobre a escolha dimensional.

Nesta fase é importante referir que, enquanto no trabalho original [Azevedo, et al., 2012] o processo foi efectuado paralela e simultaneamente para os quatro principáis partidos, agora trabalhou-se unicamente o designado por "5p".

3.6.2. Das Variáveis SE entre si

Nos conjuntos partidários de dez atributos considerados no trabalho anterior relativo aos quatro partidos tradicionais, a menor correlação absoluta, com a variável eleitoral, é de 0,80 e no conjunto referente a este restante intervalo partidário é de 0,77. Num conjunto de atributos com elevada correlação com um outro, podem existir atributos com elevada correlação entre si. Isso poderia interferir com o desempenho de alguns métodos. Por essa razão foram seleccionados dois outros conjuntos para os quais se estabeleceu um valor limite de correlação.
Para o primeiro, e escolhendo sempre atributos com o máximo de correlação possível com a variável objectivo, foram escolhidos os que não apresentassem mais do que 0,80 de correlação entre si

Para o segundo, foram seleccionados cinco atributos de modo a que abrangessem equidistantemente a totalidade do intervalo dos valores das correlações, presentes no grupo dos 190.

Deste modo geraram-se seis conjuntos de atributos:

- Três conjuntos de cinco atributos:
  - Um, constituído pelos atributos com correlação máxima com a variável alvo, a que se deu o nome de "CMax(5)";
  - Outro com o limite de 0,8 de correlação entre os atributos, chamado "CS";
  - Um outro que utilizava o intervalo de correlação disponível no conjunto dos 190, chamado "CL";
- Um com os dez atributos com maiores correlações com a variável alvo, denominado "CMax(10)";
- Um outro com apenas o atributo de maior correlação, chamado "CMax(1)".
- E um em que se utilizavam as componentes principais que explicassem pelo menos 95% da variância do conjunto das 190, a que se chamou "ACP.95".

3.7. Métodos utilizados

Aqueles conjuntos alimentaram os seguintes métodos:

- Máquinas de Vectores de Suporte, (SVM) [Cortes, et al., 1995];
- Regressão Linear, (RLM) [Draper, et al., 1998]:
- Perceptrão Multicamada (MLP) [Zilouchian, et al., 2001]:
- Método dos Gradientes, (Grad), [Microsoft, 2011].

Utilizou-se a validação cruzada (VC) como método de avaliação, aplicando-se a variante *Leave-one-out (VC-LOO)* e como medida de ajuste dos modelos obtidos, o *root mean squared error (RMSE)*.

3.7.1. Regressão Linear
A técnica conhecida como Regressão Linear (RL) (Draper, et al., 1998), é utilizada para modelar uma variável alvo como função linear de outras.

3.7.2. Máquinas de vetores de suporte
As Máquinas de Vetores de Suporte (*Support Vector Machines*, SVM), (Cortes & Vapnik, 1995) são uma coleção de métodos supervisionados de aprendizagem não probabilística usada para classificação e regressão. Permitem aproximar funções não lineares à custa de uma função de transformação designada por Kernel.

3.7.3. Redes Neuronais
O termo Redes Neuronais refere-se a uma colecção de modelos, inspirados no funcionamento dos neurónios, que através de um processamento em camadas transformam um problema não linear num linear (Almeida, 1999).

3.7.4. Método de Gradientes
É um modelo de optimização genérico (Kantarovitch & Vulich, 1938), (Dantzig, 1947) que foi aplicado neste trabalho para determinar os coeficientes de uma combinação linear dos atributos para minimizar o RMSE.

3.8. Parametrização dos modelos
Os ficheiros contendo os diversos conjuntos de atributos foram submetidos aos métodos acima referidos implementados no programa *WEKA* v.3.7.5. Em cada um desses métodos, através de ajustes experimentais sucessivos dos respectivos parâmetros, procurou-se minimizar o valor de *RMSE* dos seis conjuntos de atributos considerados.

O Método dos Gradientes foi aplicado utilizando o suplemento *Excel*, *"Premium Solver for Education"*, opção *"Standard GRG nonlinear"*. Este método foi aplicado unicamente ao ficheiro contendo o conjunto de atributos CMax(5).

4. Resultados

4.1. Previsão dos resultados eleitorais, incluindo o conjunto dos partidos não tradicionais ("5p")
No Quadro 1 mostram-se os melhores resultados obtidos anteriormente, [Azevedo, 2012] para os quatro partidos tradicionais e os novos resultados apurados para o conjunto dos outros votos expressos, "5p". Optou-se por repetir os dados referentes aos partidos tradicionais, recuperados do trabalho anterior, por razões de clareza textual.

A coluna "Método.Atributos" apresenta, em cada partido, os três pares de método e conjunto de atributos que obtiveram erro menor, apresentado na coluna "RMSE". A coluna "p" indica as previsões efetuadas para a eleição de 2009 utilizando o modelo aprendido com as doze primeiras instâncias segundo o modelo proposto. A coluna "v" indica o valor real dessa eleição, |v-p| a diferença absoluta entre "p" e "v". Por seu lado a coluna "s" representa o intervalo de previsão indicado pela Eurosondagem numa sondagem, à "boca das urnas", para a mesma eleição.

Quadro 1- Melhores "Método.Atributos" obtidos com treino usando as primeiras doze instâncias. Valores reais e previsão, utilizando a 13ª instância e dados da Eurosondagem, e diferenças para a 13ª eleição (2009).

	Metodo.Atributos	RMSE	p	v	|v-p|	s
PCP	Grad.CMax(5)	**0.005**	**7.72%**	**7.86%**	0.14%	**6.50%**
	MLP.ACP.95	**0.007**	**6.50%**		1.36%	**a**
	SVM.ACP.95	**0.008**	**6.90%**		0.96%	**8.70%**
PP	SVM.CS	**0.011**	**8.00%**	**10.43%**	2.43%	**7.70%**
	MLP.CS	**0.011**	**7.23%**		3.20%	**a**
	Grad.CMax(5)	**0.013**	**8.66%**		1.77%	**9.90%**
PSD	Grad.CMax(5)	**0.023**	**27.58%**	**29.11%**	1.53%	**26.90%**
	SVM.CMax(5)	**0.031**	**26.47%**		2.64%	**a**
	SVM.CS	**0.031**	**29.95%**		0.84%	**30.70%**
PS	SVM.CMax(5)	**0.026**	**41.09%**	**36.56%**	4.53%	**36.20%**
	Grad.CMax(5)	**0.027**	**40.02%**		3.46%	**a**
	MLP.ACP.95	**0.030**	**45.15%**		8.59%	**40.40%**
5p	Grad.Cmax(5)	**0.012**	**18.40%**	**16.04%**	2.36%	**10.30%**
	SVM.CMax(5)	**0.019**	**21.80%**		5.76%	**a**
	RLM.CMax(10)	**0.022**	**18.16%**		2.12%	**22.70%**

As linhas que correspondem ao valor "5p" foram as obtidas de acordo com a metodologia seguida no presente trabalho, sendo que os valores da coluna "s" correspondem à diferença para os 100% das somas das previsões mínimas e máximas partidárias apresentadas pela Eurosondagem.

Os valores de erro obtido para o intervalo "5p" são similares aos erros obtidos para os restantes quatro partidos tradicionais o que vem confirmar a hipótese colocada inicialmente para a forma de estimar os resultados do "quinto partido".

4.2. Os conjuntos e modelos preferenciais

No Quadro 1, realça-se que os conjuntos de atributos que permitem obter melhores resultados são os "CMax(5)" e "CS". A ocorrência de conjuntos com dimensão diferente de cinco é marginal, isto porque os conjuntos "CS" são também, tal como já se disse, de dimensão cinco.

Verifica-se também que o Método dos Gradientes aparece no "5p" na primeira posição, tal como acontece em quase todos os outros partidos. Quando não é assim, aparece sempre colocado nos três melhores resultados. Uma análise mais rigorosa da comparação do desempenho dos métodos passaria por testes estatísticos que a dimensão reduzida dos dados dificulta mas que poderá ser alvo de trabalho futuro.

4.3. Ajuste de previsões

Tal como foi expresso, é necessário analisar a soma das previsões para os diferentes partidos de forma a ajustar essas previsões ao valor total de 100%.

O Quadro 2 resume os valores de previsão da 13ª instância aplicando o modelo aprendido com as doze primeiras.

Quadro 2 - Valores de previsão para a 13ª instância dos cinco partidos, primeiro utilizando o apenas o conjunto Grad.CMax(5) e seguidamente por posição.

	Grad.Cmax(5)	1º classif	2º classif	3º classif
PCP	7.72%	7.72%	6.50%	6.90%
PP	8.66%	8.00%	7.23%	8.66%
PSD	27.58%	27.58%	26.47%	29.95%
PS	40.02%	41.09%	40.02%	45.15%
5p	18.40%	18.40%	21.80%	18.16%
Total	**102.37%**	**102.78%**	**102.01%**	**108.82%**

A primeira coluna utiliza o par "Grad.CMax(5)", onde a soma das previsões assume o valor de 102,37%, cerca de 2,37 pontos percentuais (pp) acima do limite. As colunas seguintes apresentam respectivamente a previsão de acordo com o par método/atributos colocado em 1º, 2º e 3º lugar por RMSE no quadro 1.

Quadro 3 - Os valores de previsão, para os 5 partidos obtido pelo o conjunto Grad.CMax(5) corrigidos para a soma limite, coluna **C,** e o valor do "5p", calculado pela diferença para os 100%, da soma das previsões sem correcção dos quatro partidos, coluna **n/C**.

	C	n/C
PCP	7.54%	7.72%
PP	8.46%	8.66%
PSD	26.94%	27.58%
PS	39.09%	40.02%
5p	17.97%	*16.03%*
	100.00%	100.00%

No Quadro 3 comparam-se os valores de previsão corrigida para 100% com a tentativa, referida na secção 3b, de estimar directamente o resultado do quinto partido pela diferença entre 100% e a soma das previsões dos quatro restantes.

Note-se que que os valores apresentados em qualquer destas duas aproximações obteve valores aceitáveis e sempre dentro do intervalo de valores apresentados pela Eurosondagem.

4.5. Atributos relevantes para o quinto partido

Mesmo assumindo que os atributos não foram escolhidos em função da sua relevância sociopolítica, não devemos deixar de olhar os resultados da escolha não assistida efectuada automaticamente pela filtragem.

Para o quinto partido, o conjunto CMax(5), o de melhores resultados em termos de menor RMSE, continha cinco atributos derivados de diferentes agregações das seguintes variáveis:

- (%) Óbitos por algumas causas de morte - Doenças do aparelho circulatório
- (%) Óbitos por algumas causas de morte - Doenças do aparelho digestivo
- Precipitação total - Beja
- Doutoramentos realizados em Portugal ou no estrangeiro e reconhecidos por universidades portuguesas: Realizados no estrangeiro - Total
- Doutoramentos realizados no estrangeiro e reconhecidos por universidades portuguesas: Total (sendo o 4º e 5º atributo são idênticos, embora tenham sido recolhidos de diferentes fontes de dados).

Como perspectivámos inicialmente e de acordo com a diferença fundamental entre a abordagem que propomos e a abordagem explicativa, não é dado como certo que a filtragem efetuada selecione variáveis que, para além da sua contribuição preditiva tenham influência substantiva na variável alvo. Na verdade, é frequente o resultado da selecção automática de variáveis detectar correlações espúrias e pode ser isso que está a acontecer neste caso.

5. Conclusões

A reutilização dos processamentos e métodos preditivos utlizados em trabalhos anteriores, no intervalo restante entre a percentagem de votação dos quatro partidos tradicionais concorrentes às eleições para a Assembleia da Republica de Portugal e o limite de 100%, selecciona modelos semelhantes e apresenta valores de RMSE, obtidos por VC/LOO, de grandeza equivalente.

Este resultado é interessante porque complementa os resultados anteriores e mostra que uma quantidade variável de partidos de tendências diversas tem uma percentagem de votação relacionável através de "*Data Mining*" com variáveis socioeconómicas.

Estes modelos preferenciais são lineares e aplicados sobre conjuntos de cinco atributos obtidos por redução de frequência de variáveis SE. As dezanove agregações utilizadas não esgotam as opções possíveis, mas pretendiam mimetizar alguns dos efeitos e influências nos eleitores, sendo posteriormente filtradas por majoração da correlação absoluta com os resultados eleitorais partidários.

É possível, obtendo erros de ajuste razoavelmente pequenos, prever resultados eleitorais em percentagem de votos expressos, para os quatro partidos tradicionais, PCP, PS, PSD, PP e para a restante votação, partindo de variáveis SE publicamente disponíveis.

As variáveis de origem seleccionadas, sendo expressões de variadas características Socioeconómicas Portuguesas, são representativas da situação social, económica e financeira do país. No entanto, porque não foram escolhidas por nenhum tipo de relevância mas apenas por filtragem não supervisionada, podem não conter significado explicativo relevante para as opções políticas individuais.

Anteriormente, uma das conclusões obtidas apontava para uma variação dimensional semelhante entre os resultados partidários e os erros obtidos pelos testes aos modelos. Os resultados deste estudo parecem suportar essa conclusão.

A soma dos valores de previsão obtidos para todos os partidos, os quatro tradicionais e o quinto, representando este o intervalo restante, é, antes de qualquer correcção, ligeiramente superior ao limite de 100% para os primeiros classificados por RMSE em VC-LOO, utilizando as doze primeiras instancias. Essa diferença é próxima de 2 pp, salvo para o último dos resultados apresentados onde é ligeiramente superior a 8%. Se por um lado este desajuste é marginal, podendo ser aceite se for estabelecida a usual margem de erro de 5%, a sua normalização corrige os valores, permitindo o respeito daquele limite sem afectar a qualidade da previsão de todos os resultados para a Eleição da Assembleia da Republica Portuguesa de 2009.

Finalmente, utilizando o mesmo processo de cálculo do intervalo que atribuímos ao quinto partido, a diferença entre a soma dos valores de percentagem de votação nos quatro partidos tradicionais e o valor limite de 100%, às previsões, isto é, calculando a diferença entre os valores de previsão obtido pelo modelo de menor RMSE e aquele mesmo limite, obtêm-se valores que são, também eles, interessantes do ponto de vista preditivo.

BIBLIOGRAFIA

Aguiar-Conraria, Luís e Magalhães, Pedro. 2008. *Uma Previsão dos Resultados das eleições Legislativas de 2009*.

Aguiar-Conraria, Luis, Magalhaes, Pedro C. e Lewis-Beck, Michael S. 2011. *Forecasting Spanish Elections.* Braga : Núcleo de Investigação em Politicas Económicas.

Almeida, Luis F.G. 1999. *Redes Neuronais Temporais.* Lisboa: UNOVA.

Assembleia Constituinte 1976. 2005. *Constituição da Républica Portuguesa - VII Revisão Constitucional.* Lisboa : s.n..

Azevedo, Diamantino. 2012. *Eleições para a Assembleia da Républica e as variações socioeconomicas em Portugal.* Lisboa : Dissertação de Mestrado - DEIO/DI - FCUL.

Azevedo, Diamantino, Correia, Luis e Gaspar, Graça. 2012. *Dados socioeconómicos são bons preditores de resultados eleitorais para a Assembleia da República?* Almada : Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, pp. 468-471.

Caleiro, António. 2012. *Do especifico para o geral nos modelos de ciclos eleitorais*.

Cortes, C. e Vapnik, V. 1995. Support Vector Networks. *Machine Learning*, pp. 237-297.

Dantzig, D. Van. 1947. On the principles of intuitionistic and afirmative mathematics. *Indagationes Mathmaticae*, pp. 429-440, 506-517.

Draper, Norman R e Smith, Harry. 1998. *Applied Regression Analysis.* s.l. : Wiley & Sons.

Draper, Norman R. e Smith, Harry. 1998. *Applied Regression Analysis.* Third. Canadá : John Wiley & Sons, Inc..

Fiorina, M. 1981. *Retrospective voting in American National Elections.* Nwe Haven : Yale University Press.

Gauss, Carl Friedrich. 1809. *Theorie der Bewegung der Himmelskörper, die die Sonne in Kegelschnitten umkreisen.* [trad.] C. H. Davis.

Hippert, Henrique S., Pedreira, Carlos E. e Souza, Reinaldo C. 2001. Neural Networks for Short-Term Load Forecasting: A review and Evaluation. *IEEE Transactions on power systems.* February de 2001, Vol. 16.

Kantarovitch, L. e Vulich, B. 1938. Sur la représentation de opérations linéares. Leningrad : s.n., Vol. 5, pp. 119-165.

Key Jr., V.O. 1966. The resposible electorate: Rationality in Presidential Voting. *The resposible electorate.* s.l. : Harvard University Press.

Magalhães, Pedro. 2011. June 2011 Portuguese Parliamentary Elections: Pre-Election Report. *Election Reports.* [Online] 4 de Maio de 2011. http://themonkeycage.org/blog/category/election-reports/.

Magalhães, Pedro, Aguiar-Conraria, Luís e Pereira, Miguel M. 2011. *As sondagens e os resultados eleitorais em Portugal.* s.l. : Sociedade Portuguesa de Estatistica.

Marquardt, Donald W. 1980. You Should Standardize the Predictor Variables in Your Regression Models. *Journal of the American Statistical Association.* Vol. 75, pp. 87-91.

McCulloch, W. S. e Pitts, W. 1943. *A logical calculus of the ideias immanent in nervous activity.* s.l. : Bull Math. Biophysics.

Microsoft. 2011. Solver Uses Generalized Reduced Gradient Algorithm. *http://support.microsoft.com.* [Online] 19 de Setembro de 2011. [Citação: 28 de Novembro de 2012.] http://support.microsoft.com/kb/82890.

Mitofsky, Warren J. 1998. *Was 1996 a worse year for pols than 1948?* s.l. : Public Opinion Quarterly. pp. 230-249.

Nadeau, Richard e Neimi, Richard G. 1995. *Educated Guesses the Process of answering factual knowledge questions in surveys.* s.l. : The American Association for Public Opinion Research.

van der Brug, Wouter, van der Eijk, Cees e Franklin, Mark. 2007. *The Economy and the Vote.* Amesterdam : Cambridge.

Veiga, Francisco J. e Veiga, Linda G. 2004. *The Determinants of vote intentions im Portugal.* Braga : NIPE.

Wikipédia. Eleições legislativas portuguesas de 2009. *http://pt.wikipedia.org.* [Online] [Citação: 28 de Novembro de 2012.] http://pt.wikipedia.org/wiki/Elei%C3%A7%C3%B5es_legislativas_portuguesas_de_2009.

Witten, Ian H., Frank, Eibe e Hall, Mark A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques.* s.l. : Morgan Kaufmann.

Wolfe, Phil e Frank, Marguerite. 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly.* Vol. 3, pp. 95-110.

Zilouchian, Ali e Jamshidi, Mohammad. 2001. *Intelligent Control Systems Using Soft Computing Methodologies.* Boca Raton : CRC Press, Inc.

Diamantino S. Azevedo é Licenciado em Farmácia e Mestre em Gestão de Informação pela Universidade de Lisboa, (UL), onde é colaborador na unidade de investigação Laboratório de Modelação de Agentes (LabMAg). Profissionalmente colaborou em várias áreas entre as quais se destaca o desenho do actual sistema estatístico utilizado pelo Observatório do Turismo de Lisboa. Os seus interesses de investigação são *Data Mining* e *Forecasting*.

Luís M.P. Correia (http://www.di.fc.ul.pt/~lcorreia) é professor associado, com agregação, da Universidade Lisboa (UL), no Departamento de Informática da Faculdade de Ciências, departamento a que preside, desde setembro de 2012. De 2004 a 2012 coordenou a unidade de investigação Laboratório de Modelação de Agentes (LabMAg) da UL. Os seus interesses de investigação são vida artificial, robôs móveis, algoritmos evolucionários e auto-organização em sistemas multi-agentes.

Graça Gaspar (http://www.di.fc.ul.pt/~gg) é professora auxiliar no Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (UL) e membro da unidade de investigação Laboratório de Modelação de Agentes (LabMAg) da UL. É doutorada em Ciência da Computação pela Faculdade de Ciências de Lisboa. Os seus interesses de investigação são modelação de agentes cognitivos, aprendizagem automática e raciocínio na web semântica.