PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia
ISSN: 1983-9456 (Impressa)
ISSN: 2317-0123 (On-line)
Editor: Fauze Najib Mattar
Sistema de avaliação: Triple Blind Review
Idiomas: Português e Inglês
Publicação: ABEP – Associação Brasileira de Empresas de Pesquisa

Modelagem e Monitoramento Preditivo das Eleições Municipais da Cidade de São Paulo por meio da Técnica de Árvores de Decisão

Predictive Modelling and Monitoring of Municipal Elections in the City of São Paulo through Decision Trees Technique

Submissão: 8 maio 2013 - Aprovação: 24 jun. 2013

Claudia Sciortino de Reina
Mestre em Estatística pela Universidade de São Paulo. Graduada em Estatística pela Universidade Federal de Pernambuco. Estatística Sênior do IBOPE Inteligência.
E-mail: claudia.reina@ibope.com.br.
Endereço: IBOPE Media - Área Ciência da Medição, 11º andar - Alameda Santos, 2101- Cerqueira César - 01419-100 - São Paulo/SP - Brasil.

Diego Monteforte Pintor
Graduado em Estatística pela Universidade de São Paulo. Estatístico Trainee do IBOPE Inteligência.
E-mail: diego.pintor@ibope.com.br.

RESUMO

Este artigo apresenta uma aplicação dos modelos supervisionados de árvores de decisão, em particular dos algoritmos Classification and Regression Tree - CART e Random Forest como preditores da intenção de voto para prefeito da cidade de São Paulo, nas eleições realizadas em 2012, a partir das características políticas, socioeconômicas e demográficas dos eleitores. Para isso, foram consideradas seis pesquisas divulgadas pelo IBOPE Inteligência durante o 1º turno[1] das eleições (entre 31/07/2012 e 01/10/2012) e quatro pesquisas divulgadas durante o 2º turno[2] (entre 09/10/2012 e 27/10/2012). A aplicação dos modelos supervisionados de árvores de decisão em ambos os turnos das eleições contribuiu para a identificação das variáveis mais relevantes para explicar a intenção de voto dos eleitores da cidade de São Paulo. Adicionalmente, permitiu descrever a evolução da intenção de voto prevista, segundo os modelos considerados, e compará-la com a evolução da intenção de voto efetivamente observada ao longo das pesquisas.

PALAVRAS-CHAVE:

Modelos supervisionados de árvores de decisão, Classification and Regression Tree - CART, Random Forest, eleições municipais, São Paulo/Brasil.

ABSTRACT

This paper presents an application of supervised decision trees models, in particular the algorithms CART (Classification and Regression Tree) and Random Forest, as predictors of voting for mayor of the city of São Paulo, in 2012 elections, from the political, socioeconomic and demographic characteristics of voters. To do this, were considered six surveys released by IBOPE Intelligence during the first round of elections (between 7/31/2012 and 10/1/2012) and four research disclosed during the 2nd round (between 10/9/2012 and 10/27/2012). The application of decision trees supervised models in both rounds of the elections contributed to the identification of the most relevant variables to explain the intention to vote of the electors of the city of São Paulo. Additionally, allowed to describe the evolution of the intention to vote, according to the models considered, and compare it with the evolution of the intention to vote effectively observed.

KEYWORDS:

Supervised models of decision trees, Classification and Regression Tree - CART, Random Forest, municipal elections, São Paulo/Brazil.

[1] Pesquisas registradas na 1ª Zona Eleitoral de São Paulo/SP sob os números: SP-00198/2012, SP-00311/2012, SP-00605/2012, SP-00835/2012, SP-01138/2012, SP-01474/2012.
[2] Pesquisas registradas na 1ª Zona Eleitoral de São Paulo/SP sob os números: SP-01852/2012, SP-01864/2012, SP-01912/2012, SP-01935/2012.

1. INTRODUÇÃO

Como mecanismo essencial de um sistema político democrático, o processo de votação é, em suas múltiplas e distintas faces, objeto constante de análise por parte de pesquisadores envolvidos com este tema.

Conforme observa Cruz (2011):

> A identificação das variáveis intervenientes do comportamento eleitoral tem sido uma das grandes preocupações da ciência política nas últimas décadas.

Este é, de fato, um dos propósitos do presente estudo, a partir do ponto de vista estatístico que será detalhado mais adiante. Aqui, contudo, o interesse recai também na investigação da possibilidade de que modelos estatísticos sejam utilizados para prever a intenção de voto, não apenas em um instante de tempo isolado, mas de maneira sequencial ao longo de um processo eleitoral.

Nesse sentido, todas as informações disponíveis (e potencialmente relevantes) nas bases de dados devem ser incorporadas às análises, a fim de obter resultados preditivos mais acurados, por mais intuitivas e naturais que possam parecer as relações de dependência encontradas entre as variáveis.

Dentro da área de mineração de dados e inteligência artificial, a classe de modelos supervisionados permite, não só classificar (segmentar) uma base de dados a partir de uma variável resposta (variável de interesse) e de variáveis independentes, como também predizer a classificação de um novo registro/observação sobre o qual se conhecem apenas os valores das variáveis independentes.

A técnica estatística de árvores de decisão constitui uma das ferramentas mais difundidas da modelagem supervisionada.

Em Rokach e Maimon (2008), pode-se encontrar uma visão mais ampla da classe de modelos supervisionados e de como as árvores de decisão se inserem nesse contexto.

O presente artigo se propõe a:

- Analisar as bases de dados das pesquisas de intenção de voto no 1º e 2º turnos das eleições para prefeito da cidade de São Paulo em 2012.
- Identificar as características políticas, socioeconômicas e demográficas mais associadas ao voto em cada um dos candidatos.
- Possibilitar a predição da intenção de voto dos entrevistados a partir de modelos supervisionados de árvores de decisão.

Esperam-se alcançar, com este trabalho, os seguintes resultados:

- Descrever quais são as variáveis mais importantes para explicar a intenção de voto ao longo do período eleitoral sob investigação.
- Analisar a evolução da intenção de voto prevista segundo os modelos considerados e compará-la com a evolução da intenção de voto efetivamente observada nas pesquisas eleitorais.

Este artigo está organizado da seguinte forma:

- A seção 2 apresenta o desenvolvimento realizado, detalhando-se a metodologia e as bases de dados utilizadas.

- Na seção 3, encontram-se os resultados da aplicação da modelagem estatística supervisionada, separadamente para cada turno das eleições municipais.
- Na seção 4, as conclusões do presente trabalho são discutidas.

O fluxo do processo realizado para atender a esses objetivos é descrito resumidamente na Figura 1.

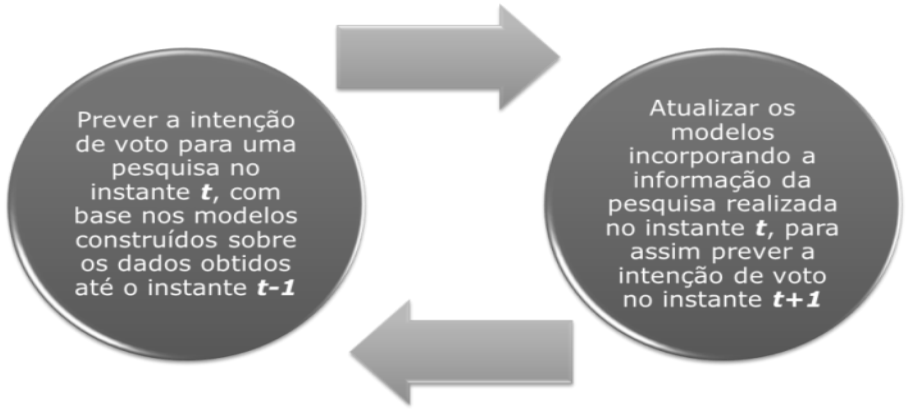

FIGURA 1
Fluxo do processo de análise.

2. DESENVOLVIMENTO

2.1 METODOLOGIA

As árvores de decisão utilizam a estratégia de **dividir para conquistar**, buscando encontrar a solução para um problema a partir de sua sucessiva decomposição em subproblemas de menores dimensões (tal qual uma regra computacional do tipo *If-Then*).

Desta forma, se consegue encontrar padrões que possam predizer eventos futuros usando uma cadeia de regras de decisão.

É importante observar que a variável resposta pode ser categórica (gerando um modelo de árvore com classificação) ou contínua (modelo de árvore com regressão), como se vê em Fonseca (1994) e Diniz e Louzada Neto (2000).

Reina et al. (2012) investigaram as várias diferenças entre os modelos classificadores com base em árvores de decisão, entre eles:

- Classification and Regression Tree -CART.
- Chi-squared Automatic Interaction Detector - CHAID.
- Exhaustive Chi-squared Automatic Interaction Detector - Exhaustive CHAID.
- Quick, Unbiased, Efficient Statistical Tree - QUEST.
- Conditional Inference Tree - CIT.
- Random Forest.

O objetivo, então, era descrever os perfis sociodemográficos dos eleitores de acordo com sua intenção de voto no 1º turno das eleições para prefeito de São Paulo em 2008, por meio desses modelos classificatórios de árvore de decisão.

Conforme abordado por Oliveira e Gadelha (2012), há inúmeros trabalhos que utilizam a teoria de árvore de decisão para determinação dos votos.

Outras técnicas também são bastante empregadas, como a regressão logística multinomial, aplicada, por exemplo, no trabalho de Andrade (2006).

No entanto, essas importantes contribuições científicas visam entender e/ou descrever a intenção de voto dos eleitores a partir de um conjunto de variáveis qualitativas, atitudinais e comportamentais desses eleitores.

Diferentemente do que já foi apresentado, o escopo do presente trabalho é ampliado no sentido de explorar o aspecto preditivo dos modelos supervisionados de árvores de decisão, indo além do aspecto descritivo das variáveis que influenciam a intenção de voto ou do perfil dos eleitores de cada candidato.

É nesse contexto que as análises estatísticas se desenvolvem a partir de uma sequência de pesquisas realizadas a certos intervalos de tempo: a cada nova pesquisa os modelos são testados, e os resultados previstos são comparados com os resultados observados.

As informações levantadas a cada pesquisa servem ao propósito de obter mais registros na base de dados, para que os modelos possam ser dinamicamente atualizados. Em tais características reside a contribuição principal do artigo para a análise estatística de pesquisas na área de opinião pública, bem como em outros setores do mercado nos quais se encontrem situações correlatas.

Tendo em vista o exposto e o caráter mais intensivo do que extensivo da presente investigação, tomou-se a decisão de utilizar, entre os diversos modelos de árvores de decisão disponíveis, dois dos mais conhecidos e empregados: os algoritmos CART e Random Forest, detalhados a seguir.

Conforme Breiman et al. (1984), entre as técnicas utilizadas na construção de classificadores com base em árvores de decisão, o algoritmo Classification and Regression Tree - CART se diferencia por sua irrestrita aplicabilidade e facilidade de entendimento frente ao fenômeno investigado.

O algoritmo Random Forest, por sua vez, consiste em retirar subamostras da amostra original e aplicar o algoritmo CART a cada uma delas (SÁ LUCAS, 2011). O algoritmo Random Forest utiliza, necessariamente, todas as variáveis explicativas introduzidas no modelo, de acordo com procedimentos aleatórios, fornecendo ao final, uma medida da importância de cada uma delas.

Essa medida, também fornecida, isoladamente, no caso do algoritmo CART, será utilizada para ranquear as variáveis de acordo com o grau de influência que exercem na explicação da intenção de voto dos eleitores. Cabe, ainda, um esclarecimento a respeito das medidas de diagnóstico dos modelos construídos.

Quando a variável resposta é categórica – caso do presente artigo, em que se quer prever a intenção de voto –, a qualidade dos modelos de árvores de decisão é avaliada comparando-se a classificação indicada pelo modelo com a verdadeira categoria da variável resposta, usualmente por meio de uma matriz conhecida como Matriz de Confusão da qual derivam medidas como as taxas de má classificação, tanto geral quanto individual. Mais detalhes sobre essas e outras medidas podem ser encontrados em Rokach e Maimon (2008).

O *software* utilizado no desenvolvimento dos modelos por meio do algoritmo CART foi o IBM SPSS Statistics versão 20, enquanto para o algoritmo Random Forest, utilizou-se o *software R* na versão 2.12.1.

2.2 BASES DE DADOS

As bases de dados analisadas correspondem a pesquisas sobre intenção de voto para prefeito da cidade de São Paulo em 2012, divulgadas pelo IBOPE Inteligência ao longo do processo eleitoral, com exceção daquelas realizadas na véspera ou no mesmo dia das eleições, quando, em geral, não se levantam outras variáveis a não ser a própria intenção de voto.

As Tabelas 1 e 2 resumem as características principais de cada pesquisa, respectivamente, para o 1º e 2º turno.

O modelo de amostragem utilizado em todas as pesquisas foi o de conglomerados em dois estágios. No primeiro estágio procedeu-se à seleção sistemática de setores censitários com probabilidade proporcional à população de 16 anos ou mais, neles residente. Dentro de cada setor censitário sorteado, foi selecionado um número fixo de eleitores segundo as cotas de sexo, grupos de idade, instrução e ramo de atividade.

TABELA 1
Aspectos gerais sobre as pesquisas divulgadas no 1º turno.

1º TURNO	PERÍODO DE CAMPO	QUANTIDADE DE ENTREVISTAS	RESPOSTAS VÁLIDAS*
1ª Pesquisa	31/07 a 02/08	805	731
2ª Pesquisa	13/08 a 15/08	805	728
3ª Pesquisa	28/08 a 30/08	1.001	911
4ª Pesquisa	10/09 a 12/09	1.001	942
5ª Pesquisa	22/09 a 24/09	1.204	1.113
6ª Pesquisa	29/09 a 01/10	1.204	1.093

*Excluindo quem não sabia ou não respondeu sobre a intenção de voto.

TABELA 2
Aspectos gerais sobre as pesquisas divulgadas no 2º turno.

2º TURNO	PERÍODO DE CAMPO	QUANTIDADE DE ENTREVISTAS	RESPOSTAS VÁLIDAS*
1ª Pesquisa	09/10 a 11/10	1.204	1.131
2ª Pesquisa	15/10 a 17/10	1.204	1.145
3ª Pesquisa	22/10 a 24/10	1.204	1.141
4ª Pesquisa	25/10 a 27/10	1.204	1.137

*Excluindo quem não sabia ou não respondeu sobre a intenção de voto.

Com relação à variável intenção de voto, optou-se por considerar a pergunta estimulada, na qual os candidatos são listados previamente para os entrevistados.

Os Quadros 1 e 2 mostram a pergunta estimulada sobre a intenção de voto, respectivamente para o 1º e 2º turnos.

QUADRO 1
Pergunta estimulada sobre a intenção de voto – 1º turno.

P. 3 (DISCO 1) SE A ELEIÇÃO PARA PREFEITO DE SÃO PAULO FOSSE HOJE E OS CANDIDATOS FOSSEM ESTES, EM QUEM O(A) SR(A) VOTARIA? (RESPOSTA ÚNICA – RU)					
01	( )	Ana Luiza	09	( )	José Serra
02	( )	Eymael	10	( )	Miguel
03	( )	Carlos Giannazi	11	( )	Paulino da Força
04	( )	Celso Russomanno	12	( )	Soninha
05	( )	Fernando Haddad	97	( )	Nenhum/Branco/Nulo
06	( )	Gabriel Chalita	98	( )	Não sabe
07	( )	Levy Fidelix	99	( )	Não respondeu
08	( )	Anaí Caproni			

QUADRO 2
Pergunta estimulada sobre a intenção de voto – 2º turno.

P. 3 EM QUEM O(A) SR(A) VOTARIA SE TIVESSE QUE ESCOLHER ENTRE: **LEIA AS ALTERNATIVAS 1 E 2. <u>FAÇA RODÍZIO</u> ENTRE OS NOMES A CADA ENTREVISTA. (RESPOSTA ÚNICA – RU)**		
01	( )	Fernando Haddad
02	( )	José Serra
07	( )	Nenhum/Branco/Nulo
08	( )	Não sabe
09	( )	Não respondeu

A evolução da intenção de voto observada ao longo das pesquisas eleitorais é mostrada nas Tabelas 3 e 4, para o 1º e 2º turnos, respectivamente.

TABELA 3
Evolução da intenção de voto observada – 1º turno.

INTENÇÃO DE VOTO OBSERVADA 1º TURNO	**PESQUISA**					
	1ª	**2ª**	**3ª**	**4ª**	**5ª**	**6ª**
Celso Russomanno	27%	29%	35%	37%	37%	36%
Fernando Haddad	7%	10%	18%	16%	19%	18%
Gabriel Chalita	5%	5%	6%	6%	8%	9%
José Serra	28%	29%	22%	20%	18%	21%
Paulinho da Força	6%	5%	2%	1%	1%	1%
Soninha	8%	6%	4%	4%	4%	4%
Outros	3%	3%	1%	1%	2%	2%
Nenhum/Branco/Nulo	16%	13%	13%	14%	10%	10%
TOTAL	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

TABELA 4
Evolução da intenção de voto observada – 2º turno.

INTENÇÃO DE VOTO OBSERVADA 2º TURNO	PESQUISA			
	1ª	2ª	3ª	4ª
Fernando Haddad	51%	52%	51%	54%
José Serra	39%	33%	38%	34%
Nenhum/Branco/Nulo	10%	15%	11%	12%
TOTAL	**100%**	**100%**	**100%**	**100%**

3. RESULTADOS

Esta seção apresenta os principais resultados da modelagem supervisionada de árvore de decisão para cada um dos turnos eleitorais, separadamente. O período investigado no 1º turno foi de 31/07/2012 a 1/10/2012 (a votação ocorreu em 7/10/2012), enquanto no 2º turno o período investigado foi de 09/10/2012 a 27/10/2012 (a votação ocorreu em 28/10/2012).

3.1 PRIMEIRO TURNO

O Quadro 3 apresenta as variáveis que foram investigadas em, pelo menos, uma das seis pesquisas consideradas no 1º turno.

Devido às diferenças entre as informações levantadas ao longo das pesquisas eleitorais para prefeito da cidade de São Paulo em 2012, foram propostas duas modelagens:

- Primeira modelagem*:* considera as variáveis comuns a todas as pesquisas realizadas no 1º turno, identificadas em amarelo claro no Quadro 3.
- Segunda modelagem: utiliza, além das anteriores, outras variáveis presentes apenas em parte das pesquisas (especificamente, na 2ª, 3ª e 5ª pesquisas), identificadas em marrom claro no Quadro 3.

Para cada modelagem de árvore de decisão proposta, seguiram-se as etapas apresentadas na Figura 1. No caso da pesquisa inicial, a "previsão" é feita com base nos próprios dados utilizados para construir os modelos.

3.1.1 PRIMEIRA MODELAGEM DO 1º TURNO

As variáveis que se mostraram mais importantes para explicar a intenção de voto dos eleitores dentro da primeira modelagem do 1º turno estão apresentadas na Tabela 5.

Podem-se observar diferenças entre os métodos utilizados: no algoritmo CART sobressaem as variáveis de rejeição ou avaliação dos governantes, enquanto no algoritmo Random Forest predominam as variáveis de caráter demográfico.

As taxas de má classificação, entretanto, são semelhantes entre os dois métodos, variando entre 50% e 60% a cada pesquisa, conforme mostra a Tabela 6.

A Figura 2 apresenta graficamente a evolução das taxas de má classificação pesquisa a pesquisa, com o tamanho da bolha representando a quantidade de observações consideradas, a qual aumenta com o processo de união das bases.

Cabe destacar que o algoritmo Random Forest exclui da análise as observações para as quais falta o valor de alguma variável explicativa (mesmo que a intenção de voto esteja registrada).

Nota-se que, para o algoritmo CART, a taxa de má classificação não diminui à medida que o número de observações na base aumenta; isso ocorre para o algoritmo Random Forest, mas o ganho de precisão não se mostra tão significativo.

Dessa forma, quando se compara a evolução da intenção de voto observada (Figura 3) com a evolução da intenção de voto prevista segundo os dois algoritmos (Figuras 4 e 5), vê-se que ambos os modelos revelam-se "insensíveis" ao crescimento do candidato Fernando Haddad nas pesquisas eleitorais para prefeito da cidade de São Paulo em 2012.

O algoritmo Random Forest consegue prever, ao menos, o declínio experimentado pelo candidato José Serra. Assim, conclui-se que as variáveis presentes na primeira modelagem (basicamente variáveis demográficas e de rejeição/avaliação dos governantes) não conseguem predizer a intenção de voto de maneira tão satisfatória.

QUADRO 3
Variáveis levantadas nas pesquisas do 1º turno.

VARIÁVEIS LEVANTADAS - 1º TURNO	PESQUISA					
	1ª	2ª	3ª	4ª	5ª	6ª
Sexo	x	x	x	x	x	x
Idade	x	x	x	x	x	x
Escolaridade	x	x	x	x	x	x
Ramo de atividade	x	x	x	x	x	x
Interesse pelas eleições	x	x	x	x	x	x
Rejeição aos candidatos	x	x	x	x	x	x
Avaliação do Prefeito Gilberto Kassab	x	x	x	x	x	x
Avaliação do Governador Geraldo Alckmin	x	x	x	x	x	x
Avaliação da Presidente Dilma Rousseff	x	x	x	x	x	x
Área que o entrevistado acredita apresentar problemas mais críticos	x	x	x	x	x	x
Renda pessoal	x	x	x	x	x	x
Renda familiar	x	x	x	x	x	x
Intenção de voto em um eventual 2º turno entre José Serra e Celso Russomanno		x	x	x	x	x
Intenção de voto em um eventual 2º turno entre José Serra e Fernando Haddad				x	x	x
Intenção de voto em um eventual 2º turno entre Fernando Haddad e Celso Russomanno				x	x	x
Conhecimento dos candidatos		x	x	x	x	
Opinião sobre os candidatos						x

Opinião quanto ao candidato vencedor		x	x	x	x	x
Partido preferido		x	x		x	x
Religião		x	x		x	x
Percepção quanto ao candidato apoiado pelo Prefeito Gilberto Kassab				x		
Percepção quanto ao candidato apoiado pelo Governador Geraldo Alckmin				x		
Percepção quanto ao candidato apoiado pela Presidente Dilma Rousseff				x		
Percepção quanto ao candidato apoiado pelo Ex-Presidente Lula				x		
Acompanhamento da propaganda eleitoral				x		x
Melhor programa				x		
Melhores propostas				x		
Zona da capital (variável controlada na amostra)	x	x	x	x	x	x

TABELA 5
Variáveis mais importantes para explicar a intenção de voto, por algoritmo– 1ª modelagem/1º turno.

IMPORTÂNCIA RELATIVA MÉDIA DAS VARIÁVEIS			
CART		**RANDOM FOREST**	
Rejeição ao candidato José Serra	20%	Idade	11%
Avaliação do Governador Geraldo Alckmin	12%	Área crítica	10%
Rejeição ao candidato Celso Russomanno	9%	Ramo de atividade	10%
Rejeição ao candidato Fernando Haddad	8%	Escolaridade	9%
Interesse pelas eleições	8%	Zona da capital	9%

TABELA 6
Evolução da taxa de má classificação dos modelos, por algoritmo – 1ª modelagem/1º turno.

PESQUISA	MÁ CLASSIFICAÇÃO CART	TAMANHO DA BASE	MÁ CLASSIFICAÇÃO RANDOM FOREST	TAMANHO DA BASE
1ª	54%	731	57%	636
2ª	53%	731	54%	636
3ª	58%	1.459	59%	1.240
4ª	53%	2.370	54%	1.948
5ª	56%	3.312	54%	2.682
6ª	56%	4.425	50%	3.568

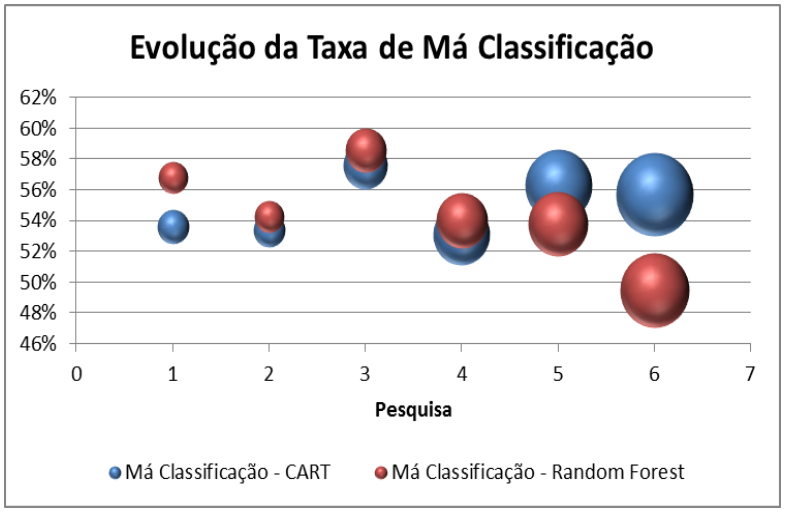

FIGURA 2
Evolução da taxa de má classificação relativamente ao tamanho da base, por algoritmo – 1ª modelagem/1º turno.

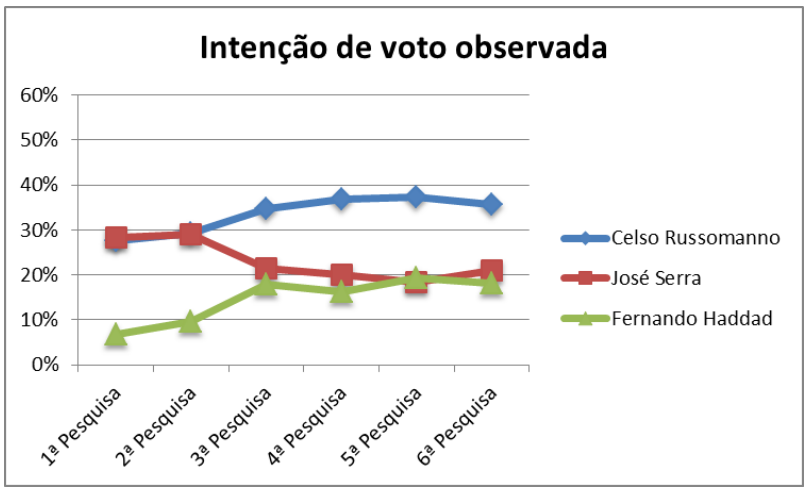

FIGURA 3
Evolução da intenção de voto observada, para os três principais candidatos – 1º turno.

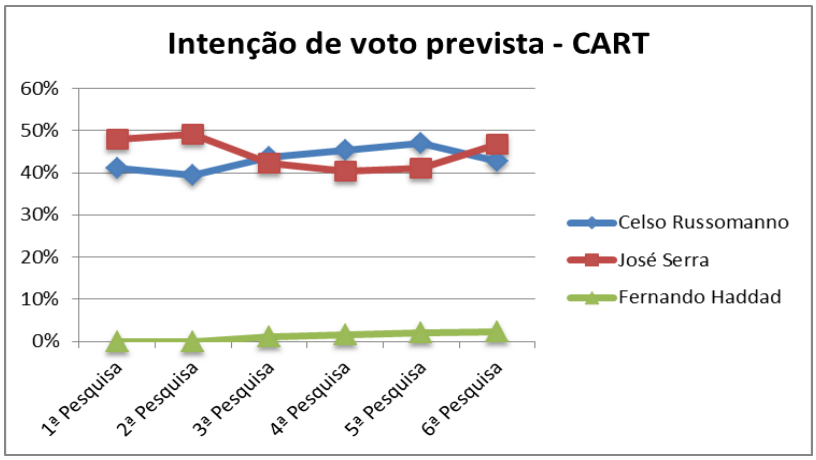

FIGURA 4
Evolução da intenção de voto prevista pelo algoritmo CART, para os três principais candidatos – 1ª modelagem/1º turno.

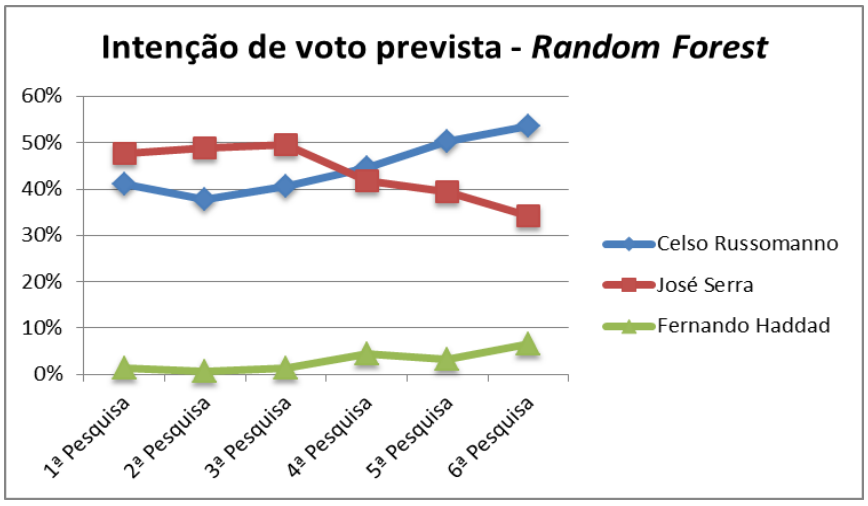

FIGURA 5
Evolução da intenção de voto prevista pelo algoritmo Random Forest, para os três principais candidatos – 1ª modelagem/1º turno.

3.1.2 SEGUNDA MODELAGEM DO 1º TURNO

As variáveis que se mostraram mais importantes para explicar a intenção de voto dos eleitores dentro da segunda modelagem do 1º turno estão apresentadas na Tabela 7. Na tabela os dois métodos concordam quanto às duas variáveis mais importantes para explicar a intenção de voto: em primeiro lugar, a intenção de voto do entrevistado em um eventual segundo turno entre os candidatos José Serra e Celso Russomanno; em segundo lugar, a opinião do entrevistado quanto ao candidato que seria o vencedor das eleições.

TABELA 7
Variáveis mais importantes para explicar a intenção de voto, por algoritmo – 2ª modelagem/1º turno.

IMPORTÂNCIA RELATIVA MÉDIA DAS VARIÁVEIS			
CART		**RANDOM FOREST**	
Intenção de voto no segundo turno	37%	Intenção de voto no segundo turno	17%
Opinião quanto ao candidato vencedor	25%	Opinião quanto ao candidato vencedor	11%
Rejeição ao candidato José Serra	8%	Área crítica	6%
Partido preferido	8%	Ramo de atividade	6%
Conhecimento do candidato Celso Russomanno	2%	Partido preferido	5%

As taxas de má classificação evoluem de acordo com a Tabela 8 e a Figura 6 (consideram-se apenas as pesquisas nas quais as variáveis da segunda modelagem estão sempre presentes).

Nesse caso, ocorre o mesmo fenômeno observado na primeira modelagem, ou seja, a taxa de má classificação dos modelos não diminui substancialmente à medida que o tamanho da base aumenta.

TABELA 8
Evolução da taxa de má classificação dos modelos, por algoritmo – 2ª modelagem/1º turno.

PESQUISA	MÁ CLASSIFICAÇÃO CART	TAMANHO DA BASE	MÁ CLASSIFICAÇÃO RANDOM FOREST	TAMANHO DA BASE
1ª				
2ª	32%	728	34%	448
3ª	30%	728	28%	448
4ª				
5ª	34%	1639	31%	998
6ª				

É importante destacar, entretanto, que a inclusão de novas variáveis de caráter predominantemente político faz com que a assertividade dos modelos cresça, passando de 50% para 70%, aproximadamente.

Desta forma, quando comparada à primeira modelagem, a segunda permite identificar, com maior intensidade, a ascensão do candidato Fernando Haddad nas pesquisas eleitorais para prefeito de São Paulo em 2012 (Figuras 7 e 8), embora não a ponto de equipará-lo ao candidato José Serra, como foi efetivamente observado até o momento analisado (Figura 9).

É possível afirmar, então, que, para o 1º turno das eleições para prefeito da cidade de São Paulo, em 2012, as variáveis presentes na segunda modelagem e ausentes na primeira (notadamente a intenção de voto em um eventual segundo turno, a opinião quanto ao candidato que sairia vencedor e o partido preferido) contribuíram para melhorar o poder preditivo dos modelos com respeito à intenção de voto dos eleitores.

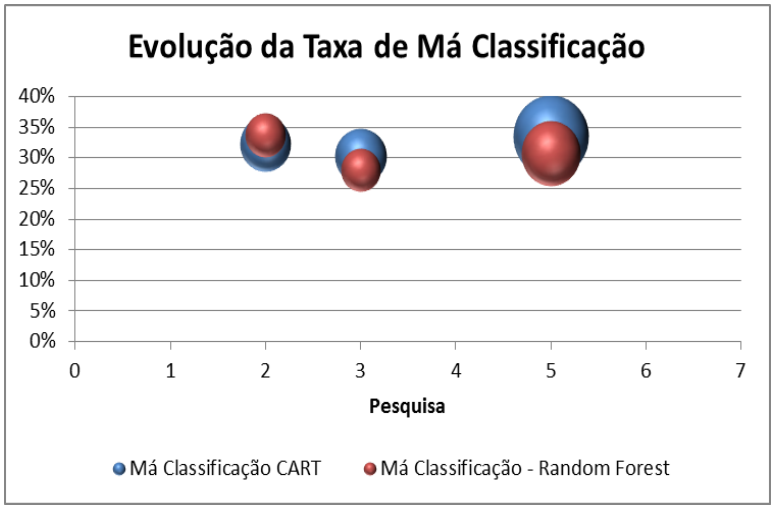

FIGURA 6
Evolução da taxa de má classificação relativamente ao tamanho da base, por algoritmo – 2ª modelagem/1º turno.

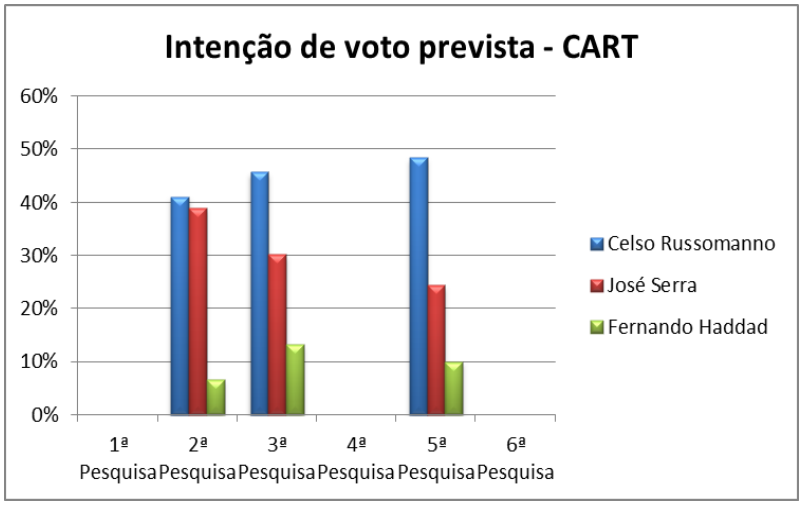

FIGURA 7
Evolução da intenção de voto prevista pelo algoritmo CART, para os três principais candidatos – 2ª modelagem/1º turno.

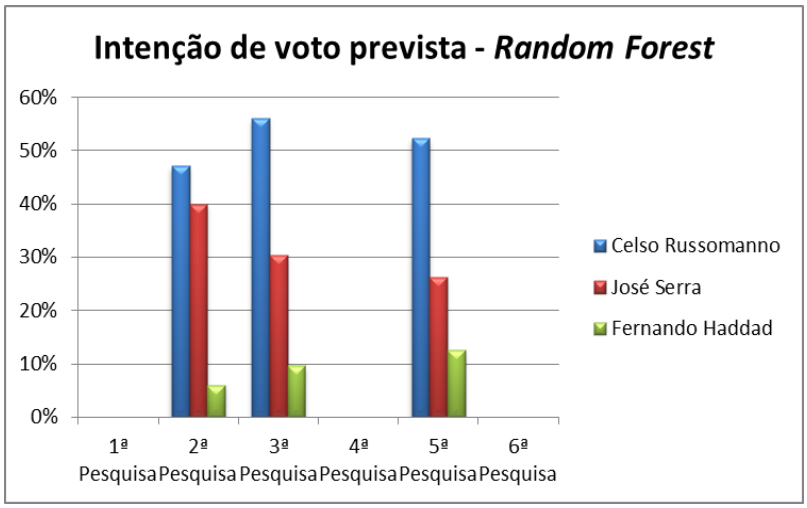

FIGURA 8
Evolução da intenção de voto prevista pelo algoritmo Random Forest, para os três principais candidatos - 2ª modelagem/1º turno.

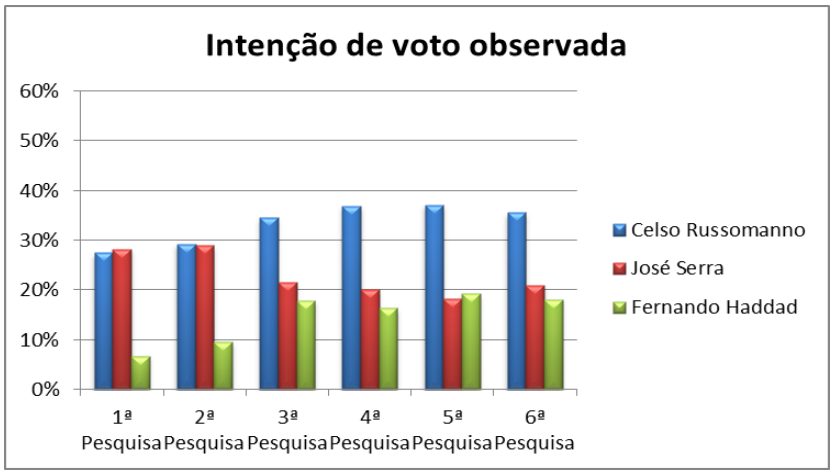

FIGURA 9
Evolução da intenção de voto observada, para os três principais candidatos - 1º turno.

3.2 SEGUNDO TURNO

A indefinição, no 1º turno, quanto ao resultado final das eleições para prefeito da cidade de São Paulo, levou os candidatos José Serra e Fernando Haddad à disputa no 2º turno, de forma que se prosseguiu com as análises dentro desse novo contexto político.

É de se destacar que as pesquisas divulgadas no 2º turno são mais homogêneas quanto às variáveis investigadas, o que levou a uma modelagem única, da qual fazem parte as variáveis destacadas em amarelo claro no Quadro 4.

As variáveis que se mostram mais importantes para explicar a intenção de voto dos eleitores no 2º turno são apresentadas na Tabela 9.

Os dois métodos coincidem quase que inteiramente, ao apontar as quatro variáveis mais relevantes no processo: em primeiro lugar, o candidato em que o entrevistado declarou ter votado no 1º turno; em segundo lugar, a opinião do entrevistado quanto ao candidato que sairia vencedor; em terceiro lugar, o partido preferido do entrevistado; e em quarto lugar, a avaliação do então Governador Geraldo Alckmin.

QUADRO 4
Variáveis levantadas nas pesquisas do 2º turno.

VARIÁVEIS LEVANTADAS - 2º TURNO	PESQUISA			
	1ª	2ª	3ª	4ª
Sexo	x	x	x	x
Idade	x	x	x	x
Escolaridade	x	x	x	x
Ramo de atividade	x	x	x	x
Interesse pelas eleições	x	x	x	x
Avaliação do Prefeito Gilberto Kassab	x	x	x	x
Avaliação do Governador Geraldo Alckmin	x	x	x	x
Avaliação da Presidente Dilma Rousseff	x	x	x	x
Renda pessoal	x	x	x	x
Renda familiar	x	x	x	x
Opinião quanto ao candidato vencedor	x	x	x	x
Partido preferido	x	x	x	x
Religião	x	x	x	x
Definição quanto ao voto	x	x	x	x
Voto no 1º turno	x	x	x	x
Opinião sobre os candidatos			x	x
Acompanhamento da propaganda eleitoral			x	
Melhor programa			x	
Melhores propostas			x	
Programa mais agressivo			x	
Zona da capital (variável controlada na amostra)	x	x	x	x

TABELA 9
Variáveis mais importantes para explicar a intenção de voto, por algoritmo – 2º turno.

IMPORTÂNCIA RELATIVA MÉDIA DAS VARIÁVEIS			
CART		**RANDOM FOREST**	
Candidato em que votou no primeiro turno	33%	Candidato em que votou no primeiro turno	26%
Opinião quanto ao candidato vencedor	22%	Opinião quanto ao candidato vencedor	23%
Partido preferido	15%	Partido preferido	12%
Avaliação do Governador Geraldo Alckmin	10%	Avaliação do Governador Geraldo Alckmin	4%
Avaliação do Prefeito Gilberto Kassab	4%	Idade	4%

As taxas de má classificação, no cenário do 2º turno, chegam a apresentar uma tendência de aumento quando as bases das pesquisas são unidas para a construção dos modelos, conforme pode ser visto na Tabela 10 e na Figura 10.

Há que se ressaltar, no entanto, como a redução no número de candidatos ocasiona uma melhora na assertividade dos modelos supervisionados de árvores de decisão nesse 2º turno: o índice de acerto varia entre 80% e 90%, pesquisa a pesquisa, com um desempenho um pouco melhor do algoritmo Random Forest.

Dessa forma, na modelagem única do 2º turno, o comportamento da intenção de voto observada (Figura 11) é mais fielmente reproduzido pelo comportamento da intenção de voto prevista pelo algoritmo CART (Figura 12) e, em especial, pelo algoritmo Random Forest (Figura 13).

TABELA 10
Evolução da taxa de má classificação dos modelos, por algoritmo – 2º turno.

PESQUISA	MÁ CLASSIFICAÇÃO CART	TAMANHO DA BASE	MÁ CLASSIFICAÇÃO RANDOM FOREST	TAMANHO DA BASE
1ª	15%	1.131	12%	853
2ª	18%	1.131	13%	853
3ª	15%	2.276	12%	1.657
4ª	18%	3.417	15%	2.467

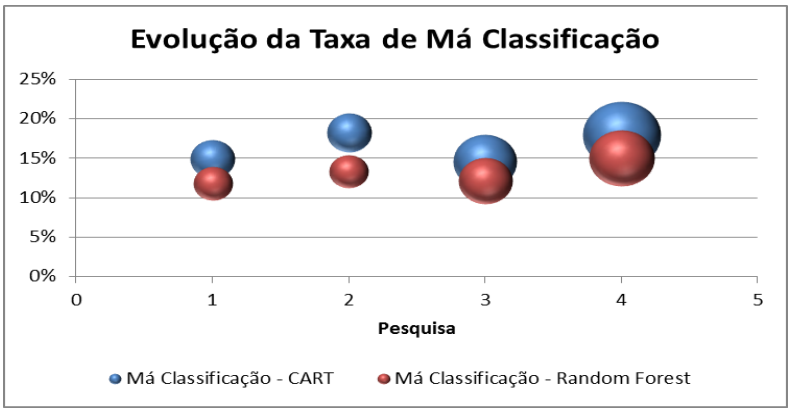

FIGURA 10
Evolução da taxa de má classificação relativamente ao tamanho da base, por algoritmo – 2º turno.

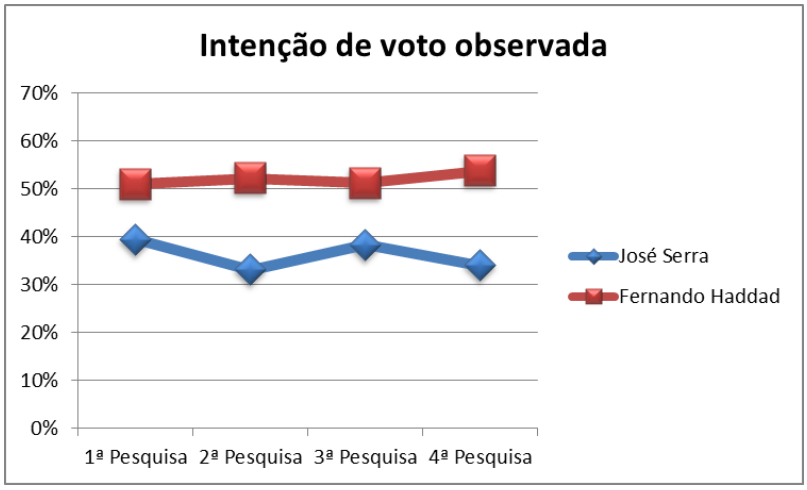

FIGURA 11
Evolução da intenção de voto observada – 2º turno.

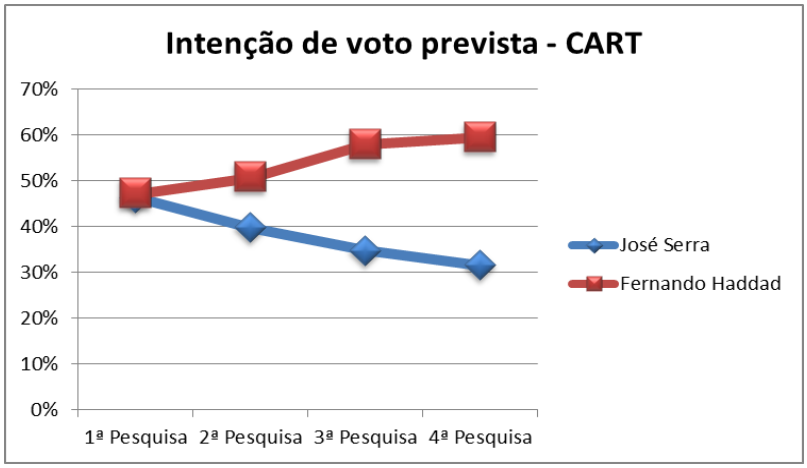

FIGURA 12
Evolução da intenção de voto prevista pelo algoritmo CART – 2º turno.

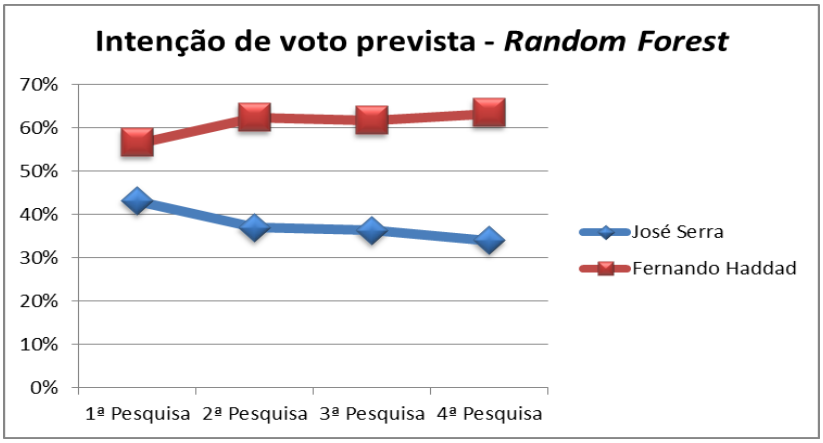

FIGURA 13
Evolução da intenção de voto prevista pelo algoritmo Random Forest – 2º turno.

4. CONCLUSÕES

O fato de a assertividade dos modelos desenvolvidos em ambos os turnos das eleições para prefeito da cidade de São Paulo não se mostrar diretamente proporcional à quantidade de observações utilizadas em sua construção indica que as variáveis explicativas consideradas (aquelas que são levantadas repetidamente a cada pesquisa eleitoral) não são suficientes para prever a intenção de voto de maneira estável e consistente ao longo do tempo (ou seja, durante o andamento do processo eleitoral).

Isso fica mais evidente ao se analisar os resultados referentes ao 1º turno, devido à maior quantidade de candidatos.

Em outras palavras, uma combinação de características políticas, socioeconômicas e demográficas que, em dado momento, está associada a determinado candidato, pode passar a estar associada ao candidato oponente dentro de um período de tempo estabelecido.

Nesse sentido, informações pontuais referentes a fatos específicos e levantadas em pesquisas isoladas, podem indicar com mais precisão, a intenção de voto dos eleitores durante o processo eleitoral. Entretanto, o próprio caráter extraordinário de tais informações inviabiliza o seu emprego como variáveis preditoras em um modelo supervisionado de árvores de decisão.

Ainda assim, ao observar os resultados das duas modelagens de árvores de decisão apresentados sobre o 1º turno, é notável o ganho de assertividade proporcionado por variáveis mais relacionadas ao contexto político (introduzidas na segunda modelagem), como a opinião dos entrevistados quanto ao candidato que sairia vencedor, o partido preferido e a intenção de voto pensando em cenários hipotéticos para o 2º turno.
Quando se analisa os resultados da modelagem única referente ao 2º turno, por sua vez, o voto declarado no 1º turno revela-se determinante para a escolha do candidato a prefeito da cidade de São Paulo em 2012.

Como observação final, é importante ressaltar que, por tratar-se de um estudo orientado a um local e período específicos, possíveis generalizações devem ser encaradas com cautela. Investigações semelhantes em outros municípios, bem como nos âmbitos estadual e federal, forneceriam um complemento ao presente trabalho, seja reforçando os aspectos aqui destacados ou trazendo novas questões para futura discussão.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, A. O. *Aplicação do modelo logístico multinomial no estudo da decisão do voto*. Dissertação. (Mestrado em Estudos Populacionais e Pesquisas Sociais). Área de Concentração em Estatística Social, apresentada à Coordenação do Mestrado da ENCE – IBGE, 2006.

BREIMAN, L.; FREIDMAN, J. H.; OLSHEN R. A.; STONE, C. J. *Classification and regression trees*. Belmont, CA: Wadworth, 1984.

CRUZ, P. A. da. *Comportamento eleitoral:* as determinantes do voto na eleição municipal de São Paulo em 2008. IV Congresso Latino Americano de Opinião Pública da WAPOR, 2011, Belo Horizonte.

DINIZ, C. A. R.; LOUZADA NETO, F. *Data Mining:* uma introdução. 14º. SINAPE da Associação Brasileira de Estatística - ABE. Caxambu – MG, 2000.

FONSECA, J. M. M. R. *Indução de árvore de decisão, histclass* – Proposta de um algoritmo não paramétrico. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Departamento de Informática. Lisboa, 1994.

OLIVEIRA, A.; GADELHA, C. *Os sentimentos dos eleitores importam para a explicação do comportamento do eleitor?* Em Debate, Belo Horizonte, v. 4, n. 4, p. 54-64, jul. 2012.

REINA, C. S. de; PINTOR, D. M.; CATALÁ, L. S.; VALFORTE, L. *Modelos supervisionados de árvores de decisão:* aplicabilidade como ferramenta para geração de conhecimento. 5. Congresso Brasileiro de Pesquisa - Mercado, Opinião e Mídia, 2012, São Paulo.

ROKACH, L.; MAIMON, O. Data Mining with Decision Trees: Theory and Applications. *Series in Machine Perception and Artificial Intelligence*, v. 69. World Scientific Pub. Co. Inc., 2008.

SÁ LUCAS, L. C. de. *Árvores, florestas e sua função como preditores:* uma aplicação na avaliação do grau de maturidade de empresas. PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia, n. 6, Março 2011.