electionsBR: R Functions to Download and Clean Brazilian Electoral Data

Fernando Meireles[1]
Denisson Silva[2]
Beatriz Costa[3]

In this note, we present a software package in R that automatically download and clean Brazilian electoral data, directly from the Brazilian Superior Electoral Court (TSE) website. The package provides 12 main functions that can be used to extract information on candidates' backgrounds, electoral coalitions' composition, and electoral results from both local and federal elections in Brazil since 1996. At the same time, these functions correct several common problems found in this data, such as encoding issues, different number of rows and columns per base, and incorrect informations. Thus, the package contributes to the political science community by facilitating the access to electoral data in Brazil and by increasing research transparency. We also illustrate the package's main potentials with a step-by-step tutorial on how it works.

Introdução

Dados eleitorais são partes fundamentais da pesquisa em Ciência Política. Em particular, informações sobre os resultados eleitorais constituem material empírico necessário para testar diversas teorias, bem como para refinar e formular outras novas. Mas, apesar desta importância, acessar este tipo de informação nem sempre é fácil. Na política comparada, há tempos é reconhecida a dificuldade de obter informações confiáveis e consistentes sobre as eleições em diversos países, especialmente em democracias recentes (Caramani, 2000; Powell & Tucker, 2014); em outros casos, estas informações são obtidas individualmente por pesquisadores a partir de fontes secundárias, o que compromete a transparência das pesquisas que as utilizam e acaba, em última instância, dificultando o acúmulo de conhecimento na área.

No Brasil, problemas como estes não são raros. Até os anos 2000, a obtenção dos resultados das eleições para todos os cargos era algo trabalhoso. Pesquisadores de várias áreas precisavam obter pessoalmente cópias dos documentos nos Tribunais

1 Doutorando em Ciência Política (UFMG). fernando.meireles@ufrgs.br.
2 Doutorando em Ciência Política (UFMG). denissonsilva@ufmg.br.
3 Mestranda em Ciência Política (UFMG). bea.s.costa@gmail.com.

Eleitorais, tanto nacional quanto estaduais, e extrair manualmente os dados necessários – sujeitos aos mais diversos erros de imputação e codificação. Como consequência direta disto, livros com estas informações eleitorais tornaram-se populares na área, suscitando inclusive revisões de teorias aceitas construídas em cima de outros dados (Dos Santos, 2002; Nicolau, 1998). Com a ampliação do acesso à internet, alguns desses dados também começaram a ser disponibilizados em *websites* especializados[4].

Em 2009[5], o Tribunal Superior Eleitoral (TSE) criou o Repositório de Dados Eleitorais[6], que disponibiliza os dados eleitorais brasileiros com ampla abrangência temporal. Mas, ao mesmo tempo em que isto facilitou o acesso a esses dados, também criou problemas. Em particular, as bases de dados do TSE são disponibilizadas de forma fragmentada em arquivos de texto delimitados por caracteres, o que dificulta a sua obtenção e tratamento; além disso, muitos destes apresentam problemas de formatação (número diferente de variáveis e observações, codificações diferentes para uma mesma variável, entre outros); são constantemente atualizados; e, – o que constitui o maior problema para difusão destas informações –, as bases não possuem documentação em inglês, barrando o acesso a estes dados pela comunidade científica internacional.

Nesta nota, introduzimos uma ferramenta para contornar estes problemas: o *software package electionsBR* para o ambiente de programação R. Inspirado em outros softwares recentemente desenvolvidos para auxiliar a pesquisa em Ciência Política (Blackwell, 2014; Carroll, Lewis, Lo, Poole, & Rosenthal, 2013; Ho, Imai, King, & Stuart, 2007; Honaker, King, Blackwell, & others, 2011), nosso pacote oferece um conjunto de funções que permitem baixar, remover erros e agregar milhares de informações eleitorais, – que incluem resultados eleitorais e de apuração, dados sobre o *background* dos(as) candidatos(as) a todos os cargos eletivos disponíveis no Brasil e sobre as

4 Dois exemplos são: http://jaironicolau.iesp.uerj.br/ e http://www3.ucam.edu.br/leex/. Acesso em: 06/10/2016. O primeiro não está mais ativo, enquanto que o segundo, ainda que ativo, está sem atualizações desde 2006.

5 Disponível em: http://www.conjur.com.br/2013-ago-24/tse-coloca-disposicao-banco-dados-detalhes-eleicoes-1950. Acesso em: 06/10/2016.

6 Disponível em: http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais. Acesso em: 07/10/2016.

legendas –, diretamente do *website* do TSE, acompanhados de extensa documentação em inglês. Deste modo, o pacote permite automatizar a tarefa de coletar dados eleitorais no Brasil e atualizá-los conforme o TSE corrige suas bases, facilitando, portanto, a replicação e a transparência das pesquisas que os utilizam.

No que segue, introduzimos o pacote electionsBR para, nas seções subsequentes, oferecer um passo a passo de como ele pode ser usado para obter diversos dados eleitorais do Brasil.

electionsBR package

Para lidar com os problemas descritos anteriormente na obtenção de dados eleitorais no Brasil, o *electionsBR* reúne um conjunto de funções, escritos na linguagem R, que automatiza a tarefa de obter os dados eleitorais do TSE. Para tanto, o pacote realiza três tarefas principais[7]:

1. Conecta o R ao Repositório de Dados Eleitorais do TSE, extraindo especificamente as informações eleitorais desejadas pelo usuário; neste ponto, o pacote funciona como um *web crawler*, acessando de forma automática o conteúdo requerido pelo usuário;
2. Baixa os arquivos necessários para construir as bases de dados, realizando as operações necessárias para tanto, que, entre outras, incluem alocar espaço na memória virtual do computador, criar um diretório temporário para armazenar as informações, abrir os arquivos compactados do *website* do TSE e carregar os arquivos brutos com os dados;
3. Por fim, ele organiza todas estas informações, unindo os diversos arquivos brutos de cada consulta e os excluindo da memória física do computador após a operação, além de padronizar todas as variáveis que são retornadas, corrigir problemas de *encoding* e preencher como *missing* informações ausentes; ao fim, o usuário tem uma base completa, documentada e pronta para uso com as informações solicitadas.

---

7 Para mais detalhes sobre os procedimentos empregados pelo pacote, é possível acessar o *source code* dele em: https://github.com/silvadenisson/electionsBR. Acesso em: 06/10/2016.

O pacote está disponível para download pelo *Comprehensive R Archive Network* (CRAN)[8], que é o principal repositório de pacotes para o ambiente de programação R, mantido pelo *The R Foundation*. Por conta disto, ele pode ser acessado de qualquer lugar do mundo gratuitamente; adicionalmente, o CRAN testa os pacotes submetidos a ele, garantindo, assim, compatibilidade com os principais sistemas operacionais disponíveis.

De forma geral, o pacote está organizado em seis grupos principais de funções, cada um relacionado a um tipo específico de dado eleitoral. Estes grupos são: (1) candidaturas; (2) detalhes de apuração no nível dos municípios/zonas eleitorais; (3) legendas; (4) partidos no nível dos municípios/zonas eleitorais; (5), resultados eleitorais no nível dos municípios/zonas eleitorais; e, (6), filiação partidária e perfil de eleitores(as). Para cada grupo, duas funções específicas estão disponíveis, uma para eleições nacionais e, a outra, para eleições municipais – à exceção do grupo seis. O Quadro 1, abaixo, descreve cada uma das funções, seus códigos para chamada no R e os dados que elas coletam.

Quadro 1 – Descrição das funções contidas no electionsBR

Função	Código de Chamada	Descrição
Candidaturas	candidate_fed(year) candidate_local(year)	Dados sobre as candidaturas
Apuração	details_mun_zone_fed(year) details_mun_zone_local(year)	Dados sobre a apuração eleitoral
Legendas	legend_fed(year) legend_local(year)	Dados sobre as legendas que disputaram eleições
Partidos	party_mun_zone_fed(year) party_mun_zone_local(year)	Dados sobre os partidos que disputaram eleições no nível dos municípios/zona eleitoral
Resultados	vote_mun_zone_fed(year) vote_mun_zone_local(year)	Resultados eleitorais no nível dos municípios/zona eleitora
Eleitores	voter_affiliation(year) voter_profile(year)	Dados sobre os eleitores

Fonte: Elaboração própria.

Como se pode ver pelo Quadro 1, o nome de cada função procura descrever, de forma condensada, o dado extraído do TSE, conforme a documentação mantida pelo

8 Disponível em: https://cran.r-project.org/package=electionsBR. Acesso em: 06/10/2016.

órgão. Por exemplo, para obtermos informações sobre os candidatos que disputaram as eleições nacionais em 1998, a função usada é a `candidate_fed`, onde `fed` refere-se às eleições federais; para as eleições municipais, a função correspondente é a `candidate_mun`. Além dos sufixos `mun` e `fed`, que referem-se a eleições municipais e federais, respectivamente, `mun_zone`, no meio do código de chamada de algumas funções, refere-se à unidade de agregação dos dados: `party_mun_zone_fed`, deste modo, retorna dados eleitorais sobre os partidos que disputaram eleições federais (sufixo `fed`) agregados ao nível município/zona eleitoral – a unidade de agregação mais comum no repositório do TSE[9]. Por fim, os prefixos no código de chamada de cada função indicam o grupo ao qual elas pertencem (e.g., `legend` indica que o grupo é o de dados de legenda, `party`, que o grupo de dados é o sobre os partidos, e assim por diante). O Quadro 2 resume a lógica por detrás no código de chamada de cada função disponível no pacote.

Quadro 2 – Organização do pacote electionsBR

Prefixo	Indica o grupo ao qual a função pertence: `candidate_`, `detail_`, `legend`, `party_`, e `voter_`
Sufixo	Indica o tipo de eleição: `_mun`, `_fed`

Fonte: Elaboração própria.

Organizado desta forma, o pacote abarca os principais dados eleitorais disponibilizados pelo TSE, mantendo seus níveis de agregação originais. Os dados acessíveis, além disso, cobrem o período que vai de 1994 a 2016 – período ampliado regularmente pelos autores para incorporar novas eleições. Para cada uma das funções, é esperado do usuário apenas fornecer o ano da eleição de interesse (argumento `year`) e, em poucos segundos, a base de dados é compilada (em formato `data.frame`, onde colunas são variáveis e, por sua vez, linhas são observações).

Feita esta descrição da organização interna do pacote, ilustramos na sequência como usá-lo para obter dados eleitorais do TSE.

9 Para as funções que não possuem indicação explícita, a unidade varia de acordo com o grupo ao qual a função pertence: no caso de dados sobre partidos, eles são a unidade de agregação; o mesmo vale para candidatos e legendas.

Como o pacote funciona?

Como todo *software package* que não faz parte do *core* de uma linguagem, é necessário instalar o `electionsBR` para poder utilizar suas funções. A versão estável do pacote pode ser instalada diretamente pelo console do R via CRAN com o código[10]:

```
install.packages("electionsBR")
```

Para instalar as versões *beta* do pacote, que incluem atualizações ainda não publicadas oficialmente no CRAN, basta usar o seguinte código:

```
if(!require("devtools")) install.packages("devtools")
devtools::install_github("silvadenisson/electionsBR")
```

Com o pacote devidamente instalado[11], basta apenas carregá-lo na sessão do R através da função `library`:

```
library(electionsBR)
```

Feito isso, as funções contidas no pacote podem ser acessadas na sessão pelos seus códigos de chamada. Todas elas requerem apenas um argumento, `year`, que deve ser passado como `integer` à função, indicando o ano da eleição de interesse. Os anos disponíveis para as eleições são federais são: 1998, 2002, 2006, 2010 e 2014. Para as eleições municipais, estão disponíveis: 1996, 2000, 2004, 2008, 2012 e 2016. Excepcionalmente para as funções de extração dos dados sobre eleitores, os períodos são diferentes: para os dados de filiação, não é necessário passar o argumento ano, apenas o estado e a sigla do partido (a coleta é feita a partir dos dados mais recentes do TSE); para o perfil dos eleitores(as), os dados estão disponíveis para todas as eleições, tanto federais quanto municipais, desde 1994. Caso o usuário digite um ano inválido, o pacote retornará um erro (e.g., *Please, check the documentation and try again*)[12].

---

10 O pacote requer uma versão do R igual ou superior a 2.10. Para verificar a lista completa de requisitos e os testes de adequação em vários sistemas operacionais, ver a página do pacote no CRAN: https://cran.r-project.org/package=electionsBR. Acesso em: 06/10/2016.

11 Note-se que o `electionsBR` depende, ele próprio, de outros pacotes para funcionar. Estes pacotes, que devem estar disponíveis previamente, são: `dplyr`. Para informações técnicas sobre as dependências do pacote, ver a documentação no CRAN (Disponível em: https://cran.r-project.org/package=electionsBR. Acesso em: 06/10/2016.).

12 Além disso, o TSE atualiza com frequência suas bases de dados. No momento em que escrevemos esta nota, por exemplo, dados para as eleições de 1996 e 2000 estão sendo revistos e podem, por esta

A título de ilustração, podemos baixar todos os dados sobre as candidaturas na eleição federal de 2010 com a função `candidate_fed` da seguinte forma:

```
# Salvamos os dados no objeto 'candidaturas'
candidaturas <- candidate_fed(2010)
## Processing the data…Done
```

Em instantes, a função coleta as informações solicitadas (além de retornar mensagens indicando o início e o fim do processo). O objeto resultado é um `data.frame` que contém 22576 observações e 43 variáveis – livre de erros. Podemos ver o nome dessas variáveis disponíveis com a função `names`:

```
names(candidaturas)
##  [1] "DATA_GERACAO"                  "HORA_GERACAO"
##  [3] "ANO_ELEICAO"                   "NUM_TURNO"
##  [5] "DESCRICAO_ELEICAO"             "SIGLA_UF"
##  [7] "SIGLA_UE"                      "DESCRICAO_UE"
##  [9] "CODIGO_CARGO"                  "DESCRICAO_CARGO"
## [11] "NOME_CANDIDATO"                "SEQUENCIAL_CANDIDATO"
## [13] "NUMERO_CANDIDATO"              "CPF_CANDIDATO"
## [15] "NOME_URNA_CANDIDATO"           "COD_SITUACAO_CANDIDATURA"
## [17] "DES_SITUACAO_CANDIDATURA"      "NUMERO_PARTIDO"
## [19] "SIGLA_PARTIDO"                 "NOME_PARTIDO"
## [21] "CODIGO_LEGENDA"                "SIGLA_LEGENDA"
## [23] "COMPOSICAO_LEGENDA"            "NOME_COLIGACAO"
## [25] "CODIGO_OCUPACAO"               "DESCRICAO_OCUPACAO"
## [27] "DATA_NASCIMENTO"               "NUM_TITULO_ELEITORAL_CANDIDATO"
## [29] "IDADE_DATA_ELEICAO"            "CODIGO_SEXO"
## [31] "DESCRICAO_SEXO"                "COD_GRAU_INSTRUCAO"
## [33] "DESCRICAO_GRAU_INSTRUCAO"      "CODIGO_ESTADO_CIVIL"
## [35] "DESCRICAO_ESTADO_CIVIL"        "CODIGO_NACIONALIDADE"
## [37] "DESCRICAO_NACIONALIDADE"       "SIGLA_UF_NASCIMENTO"
## [39] "CODIGO_MUNICIPIO_NASCIMENTO"   "NOME_MUNICIPIO_NASCIMENTO"
```

razão, estar incompletos.

```
## [41] "DESPESA_MAX_CAMPANHA"           "COD_SIT_TOT_TURNO"
## [43] "DESC_SIT_TOT_TURNO"
```

A maioria das funções do pacote extrai informações que são comuns nas bases do TSE, como `DATA_GERACAO` e `HORA_GERACAO`, que indicam a data e a hora em que o arquivo bruto com as informações foi compilado pelo TSE. Além destas, `ANO_ELEICAO` e `SIGLA_UF` indicam o ano e o estado aos quais às demais informações se referem. Para acessar a estrutura completa das bases retornadas por cada função, basta acessar a documentação delas com o código de chamada precedida de um ponto de interrogação (ou como argumento da função `help`). Para a função `candidate_fed`, por exemplo, bastaria executar uma das seguintes linhas:

```
?candidate_fed
help(candidate_fed)
```

Com isso, é possível saber que `CPF_CANDIDATO` indica o número de CPF do(a) candidato(a), ou que `NUMERO_PARTIDO` indica o número do partido pelo qual o(a) candidato(a) concorreu.

Todas as funções disponíveis também possuem um argumento `logical` opcional, `ascii`, que, por padrão, é `FALSE`, e que serve para manter ou não a acentuação nos campos textuais. Isto evita, por exemplo, que computadores sem suporte para tipos de *encoding*[13] cosigam ler adequadamento as informações coletadas. O código abaixo usa esse argumento para coletar os mesmos dados já extraídos, agora sem acentuação:

```
# Salvamos os dados no objeto 'candidaturas´, sem acentos
candidaturas <- candidate_fed(2010, ascii = TRUE)
## Processing the data...Done
```

Como analisar as informações obtidas pelo pacote?

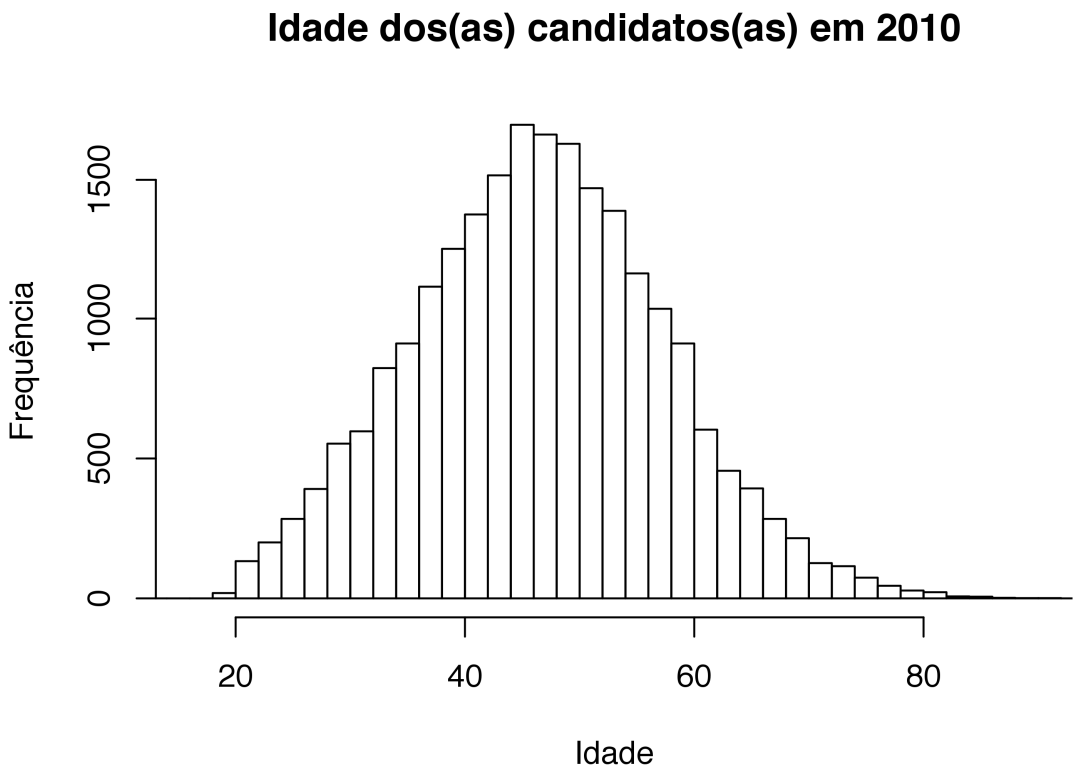
Como cada função retorna um `data.frame`, todos os dados podem ser analisados normalmente com outras ferramentas disponíveis no ambiente R. Podemos, por exemplo, calcular a média de idade de todos(as) os(as) candidatos(as) que concorreram em 2010 com a função `mean`:

13 *Encondig* pode ser entendido como o conjunto de caracteres suportados por um sistema.

```
mean(candidaturas$IDADE_DATA_ELEICAO, na.rm = T)
## [1] 47.74352
```

Ou criar um histograma com a distribuição destas idades:

```
hist(candidaturas$IDADE_DATA_ELEICAO,
     breaks = 1000,
     xlim = c(16, 90),
     main = "Idade dos(as) candidatos(as) em 2010",
     xlab = "Idade",
     ylab = "Frequência")
```



Principais vantagens do pacote

Velocidade

Comparado ao processo manual de ir ao *website* do Repositório do TSE, baixar, descompactar, carregar um a um, unir e limpar os dados para evitar erros, o `electionsBR` é extremamente rápido. De forma mais concreta, num processador i7 de

2.4Ghz e 8gb de memória RAM, o processo de baixar a base de candidaturas nas eleições federais de 2010, que usamos no exemplo anterior, tomou 3.9 segundos para ser concluído:

```
## Unit: seconds
##           expr      min       lq     mean   median       uq      max neval
##  candidate_fed 3.943167 3.943167 3.943167 3.943167 3.943167 3.943167     1
```

O tempo total, entretanto, variará conforme a conexão com a *internet*, o tamanho da base requerida (via de regra, as de eleições municipais são mais pesadas) e a resposta do servidor do TSE.

Transparência e replicação

O acesso aos dados do TSE por meio de um *software package* em código aberto permite, entre outras coisas, maior transparência no uso dos dados. Como a extração e limpeza das informações é feita de forma automática, erros humanos na manipulação (como erros de imputação, adoção de procedimentos inadequados, falta de registro nas decisões tomadas, etc.) são evitados. Na prática, isto significa que análises inteiras podem ser feitas e compartilhadas com base no pacote, reduzindo enormemente o tempo gasto com uma análise prévia da consistência dos dados.

Da mesma forma, qualquer análise realizada a partir de dados eleitorais do TSE pode ser replicada, ou atualizada, sem a necessidade de se compartilhar a base de dados original: basta incluir no código de replicação chamadas às funções do pacote para coletar os dados. Para uso acadêmico, também é possível gerar as informações bibliográficas do pacote (que fazem referência a esta nota) com o código:

```
citation("electionsBR")
```

Documentação

Além das vantagens anteriores, por vir acompanhado de extensa documentação em inglês, criada com base na documentação original disponbilizada em português, o pacote amplia o acesso aos dados eleitorais brasileiros – especialmente para a comunidade internacional. Independentemente disso, esta documentação também

possibilita simplesmente verificar de forma rápida a natureza das informações coletadas durante uma análise.

Considerações finais

Como procuramos mostrar nesta nota, obter dados eleitorais não é algo fácil, mas desse esforço depende parte importante das pesquisas na Ciência Política e nas Ciências Sociais, de forma mais ampla. Nossa principal contribuição, neste sentido, foi introduzir uma ferramenta para auxiliar pesquisadores na área a obterem este tipo de dado de forma fácil, rápida e confiável – no caso, dados eleitorais brasileiros, para todos os cargos e cobrindo um período de duas décadas, diretamente do Repositório de Dados Eleitorais do TSE. Adicionalmente, oferecemos indicações básicas de como instalar, carregar e usar o pacote para obter estes dados.

Referências

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, *22*(2), 169–182. SPM-PMSAPSA.

Caramani, D. (2000). *Elections in western europe since 1815: Electoral results by constituencies*. Macmillan.

Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., & Rosenthal, H. (2013). The structure of utility in spatial models of voting. *American Journal of Political Science*, *57*(4), 1008–1028. Wiley Online Library.

Dos Santos, W. G. (2002). *Votos e partidos: Almanaque de dados eleitorais: Brasil e outros países*. FGV Editora.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236. SPM-PMSAPSA.

Honaker, J., King, G., Blackwell, M., & others. (2011). Amelia ii: A program for missing data. *Journal of statistical software*, *45*(7), 1–47.

Nicolau, J. M. (1998). *Dados eleitorais do brasil, 1982-1996*. Editora Revan.

Powell, E. N., & Tucker, J. A. (2014). Revisiting electoral volatility in post-communist countries: New data, new results and new approaches. *British Journal of Political Science*, *44*(01), 123–147. Cambridge Univ Press.