CandiDATA: um *dataset* para análise das eleições no Brasil

Felipe F. Vasconcelos[1], João V. S. Tavares[1], Murilo U. Ribeiro[1], Fabio J. Coutinho[1], João Paulo Clarindo[2]

[1]Instituto de Computação (IC)
Universidade Federal de Alagoas (UFAL)
57072-900 – Maceió – AL – Brazil

[2]Instituto de Ciências Matemáticas e Computação (ICMC)
Universidade de São Paulo (USP)
13566-590 – São Carlos – SP – Brazil

{ffv,fabio,mugr,jvst}@ic.ufal.br, jpcsantos@usp.br

Abstract. *The Superior Electoral Court (TSE), responsible for Brazilian elections, provides data about candidates and parties related to the elections since 1933. These data constitute an important collection in order to serve as a basis for works in several research areas. However, there are many inconsistencies in data provided by the TSE, such as missing fields and non-standard fields, making analysis and integration with external databases difficult. This work presents CandiDATA, a standardized dataset of Brazilian elections. CandiDATA is available in open format and covers the election period between 1945 and 2020, including additional information beyond the TSE data.*

Resumo. *O Tribunal Superior Eleitoral (TSE), responsável pelas eleições no Brasil, disponibiliza dados acerca de candidatos e partidos que disputaram o pleito eleitoral desde 1933. Esses dados constituem um importante acervo, sendo utilizados como base por trabalhos em diferentes áreas de pesquisa. Entretanto, os dados disponibilizados pelo TSE possuem diversos problemas relacionados à falta de padronização e inconsistência, dificultando a análise e a integração com bases externas. Este trabalho apresenta CandiDATA, um* dataset *padronizado com dados das eleições brasileiras. CandiDATA é disponibilizado em formato aberto e abrange o período eleitoral entre 1945 e 2020, incluindo informações adicionais, além dos dados do TSE.*

1. Introdução

As nações democráticas existentes no mundo contemporâneo encontram-se organizadas em diferentes modelos e estruturas que refletem características próprias de cada sociedade. Avaliar o grau de maturidade da democracia de um país é um processo complexo, pois necessita considerar aspectos distintos. A revista *The Economist* estabelece um índice para examinar o estado da democracia em 167 países a partir de cinco categorias: processo eleitoral e pluralismo; funcionamento do governo; participação política; cultura política e liberdades civis. O Brasil ocupa a posição 49 neste ranking com a pontuação geral de 6,92. Entretanto, quando as categorias são verificadas de modo isolado, as pontuações podem ser bem distintas, tal como quando se observa o quesito "processo eleitoral e pluralismo" que recebeu o *score* de 9,58, figurando como a segunda melhor pontuação dentre todos os países avaliados. [Economist 2021]

De fato, o processo eleitoral brasileiro mantido pela Justiça Eleitoral conta com o reconhecimento de organismos internacionais, sendo muito conhecido pela celeridade nas apurações de resultados, visto que as eleições são realizadas integralmente com urnas eletrônicas desde o ano 2000. Além disso, nas últimas eleições, mais da metade dos eleitores foram identificados por biometria, fator importante para evitar possíveis ações fraudulentas como, por exemplo, o voto em nome de terceiros já falecidos. Portanto, verifica-se que o processo de informatização das eleições brasileiras tem evoluído consistentemente desde 1985 quando houve a implantação do cadastro eleitoral informatizado [Tribunal Superior Eleitoral 2016].

A informatização do sistema eleitoral brasileiro impulsiona o processo de geração e coleta de dados, tais como: perfil de candidatos e eleitores, apuração de resultados, prestação de contas, pesquisas eleitorais, etc [TSE 2020]. Neste contexto, o portal do Tribunal Superior Eleitoral (TSE) disponibiliza uma compilação de informações brutas das eleições incluindo, por exemplo, os resultados das eleições que ocorreram no Brasil desde o ano de 1933. Este material constitui um importante acervo para análises estatísticas e pesquisas de acadêmicos, jornalistas, sociólogos, cientistas políticos e demais interessados. Todavia, a existência de diferentes problemas dificultam seriamente a manipulação e análise dos dados, tais como: a ausência de padrões, a incompletude de dados relevantes, a presença de inconsistências na documentação, o uso de formatos distintos para a representação de dados idênticos, a utilização de padrões e formatos pouco conhecidos e a fragmentação de informação em numerosos arquivos. Esses problemas adicionam maior complexidade à manipulação dos dados, dificultando a análise, principalmente, por usuários que não são da área da computação ou possuem pouco conhecimento em programação.

Este trabalho apresenta CandiDATA, um *dataset* construído para analisar dados de eleições brasileiras do período de 1945 a 2020. O *dataset* foi gerado a partir de dados obtidos do portal do TSE os quais foram submetidos à padronização, limpeza e transformação. CandiDATA também foi enriquecido a partir a inferência de dados visando suprir informações ausentes em anos anteriores a 1996. Desta forma, campos nulos, mal-formatados e inconsistentes foram removidos enquanto novos campos foram adicionados, facilitando a integração e auxiliando no processo de tomada de decisão por gestores, pesquisadores e demais interessados. CandiDATA está disponível gratuitamente em formato aberto e pode ser lido por ferramentas de análise e linguagens de programação para sua utilização em diferentes áreas e ramos da sociedade.

Este documento encontra-se organizado da seguinte maneira: a Seção 2 discute os trabalhos relacionados; a Seção 3 descreve os dados originais e relata problemas encontrados na base de dados do TSE; a Seção 4 descreve uma visão geral da construção do *dataset* CandiDATA; a Seção 5 discute sua aplicabilidade e apresenta exemplos de uso; e, por fim, a Seção 6, reúne as considerações finais acerca do trabalho, incluindo, suas limitações e trabalhos futuros, além do endereço de disponibilização do *dataset*.

2. Trabalhos Relacionados

Diversos trabalhos utilizam como fonte de pesquisa o *Repositório de Dados Eleitorais* disponibilizado no portal do TSE [Araújo 2009] [Speck and Speck 2012] [Barbosa et al. 2009].

Camargo et al. (2016) buscam encontrar padrões nos perfis de candidatos a vereador em municípios do Rio Grande do Sul, a partir de dados do TSE referentes às eleições do ano de 2012. Inicialmente, os autores ressaltam que os dados do TSE precisaram ser convertidos para arquivos do tipo *Comma-Separated Values* (CSV), seguido de alterações para a padronização de atributos. Posteriormente, aplicaram técnicas de mineração de dados para a identificação de padrões, utilizando o algoritmo J48. Os resultados apresentados apontam como fatores relevantes para a eleição no cargo de vereador: a carreira política; a idade; o grau de escolarização e o gênero do candidato.

Filho and Pappa (2015) desenvolveram uma ferramenta para a caracterização demográfica de usuários do Twitter. A partir da análise de mensagens postadas e de informações do perfil, são inferidos o gênero, a idade e a classe social do usuário. Para demonstrar o uso da ferramenta, os autores utilizaram dados de eleitores obtidos junto ao repositório do TSE, de modo a construir uma distribuição demográfica real para compará-la com a distribuição demográfica construída a partir dos dados de usuários do Twitter.

CEPESP (2020) desenvolveu a plataforma CepespData com o objetivo de facilitar o acesso integrado às bases de dados disponibilizadas pelo TSE em seu repositório. Os dados mantidos na plataforma podem ser obtidos diretamente no site da CepespData [1] ou através de *Application Programming Interfaces* (APIs) que podem ser acessadas através de linguagens de programação como R e Python. O trabalho supera alguns problemas encontrados no repositório do TSE, entretanto, os dados são restritos ao período de 1998 a 2018.

Diferentemente, CandiDATA reúne dados de 1945 a 2020, permitindo análises mais abrangentes e com uma perspectiva histórica. Cabe ressaltar que os dados mais antigos apresentam um número maior de incompatibilidades em relação a formatos inadequados, ausência de padrões e informações relevantes.

3. Dados do portal do TSE

CandiDATA foi construído a partir de dados obtidos no portal do Tribunal Superior Eleitoral[2], os quais são organizados em diferentes categorias, como Candidatos, Resultados, entre outras. Cada categoria está dividida por ano, sendo cada ano subdividido por unidade federativa. Os dados disponibilizados cobrem os períodos eleitorais entre os anos de 1933 a 2020. Esta organização fragmentada dos dados demanda maior esforço na exploração dos conteúdos. Por exemplo, faz-se necessário acessar várias partes diferentes do site para reunir informações sobre um determinado candidato.

O TSE disponibiliza os dados em arquivos compactados no formato ZIP, que contêm, em sua grande maioria, arquivos de texto, sem cabeçalhos, com delimitação em ponto-e-vírgula e codificação `Latin1`. Entretanto, dados das eleições mais recentes, a partir de 2012, encontram-se armazenados em arquivos CSV, com o mesmo delimitador dos anos anteriores, mas com cabeçalhos. Também são disponibilizados arquivos PDF intitulados "leia-me", que descrevem o dicionário dos dados.

[1] `https://cepespdata.io/`

[2] `https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1`

3.1. Problemas encontrados nos dados publicados pelo TSE

Conforme mencionado anteriormente, a análise dos dados publicados no portal do TSE exige um esforço considerável devido à presença de diversos problemas tais como: inconsistência em dados e documentações, ausência de padrões e formatos conhecidos.

A seguir, são descritos em detalhes os principais problemas encontrados nos dados do TSE durante a construção do CadiDATA.

- Não apresentam regularidade quanto ao **formato dos seus arquivos**, visto que até 2010 os arquivos encontram-se em formato TXT desprovidos de cabeçalho e a partir do ano de 2012 estão em CSV com cabeçalho.

- Presença de **inconsistência na descrição dos campos apresentados nos dicionários**, por exemplo, existem nomes distintos para o mesmo atributo.

- Representação dos conteúdos em latin1, podendo ocasionar erros durante o carregamento desses dados por **não utilizarem o padrão internacional**.

- **Datas de nascimento** com falhas de padronização, incluindo formatos distintos tais como `dd/MM/yyyy`, `ddMMyy` ou `dd/MM/yy`.

- Atributos relacionados a **códigos de municípios** não utilizam o padrão de identificação de cidades do Instituto Brasileiro de Geografia e Estatística (IGBE).

- **Falta de padronização no código de título do eleitor**, dado que o padrão atual é de 12 dígitos e encontram-se dados representados por padrões antigos com menos de 12 dígitos.

- Ausência do campo **gênero** do candidato nos anos de 1945 até 1994.

- **Falta padronização dos nomes dos atributos**, observando-se que cada ano tem sua formatação própria dos atributos.

- O TSE apresenta **duas representações de campos nulos**: `#NULO#` e `#NE#` para *string*, como também `-1` e `-3` para inteiros.

4. Construção do *Dataset*

A Figura 1 ilustra as etapas do desenvolvimento do CandiDATA, que consiste em três módulos: (i) padronização – responsável pela padronização dos dados para campos nulos, datas de nascimento incorretas, etc; (ii) transformação – que converte códigos de cidade para o formato determinado pelo IBGE, adição do Código Brasileiro de Ocupações (CBO) e inferência de gênero; e (iii) consolidação – onde os dados são dispostos na forma descrita no dicionário.

Estes módulos foram construídos na linguagem Python, por meio da utilização da biblioteca de análise de dados Pandas[3] e da biblioteca NumPy[4]. A escolha dessa linguagem foi baseada na facilidade de manipulação de dados e arquivos CSV, além da

[3] `https://pandas.pydata.org/`
[4] `https://numpy.org/`

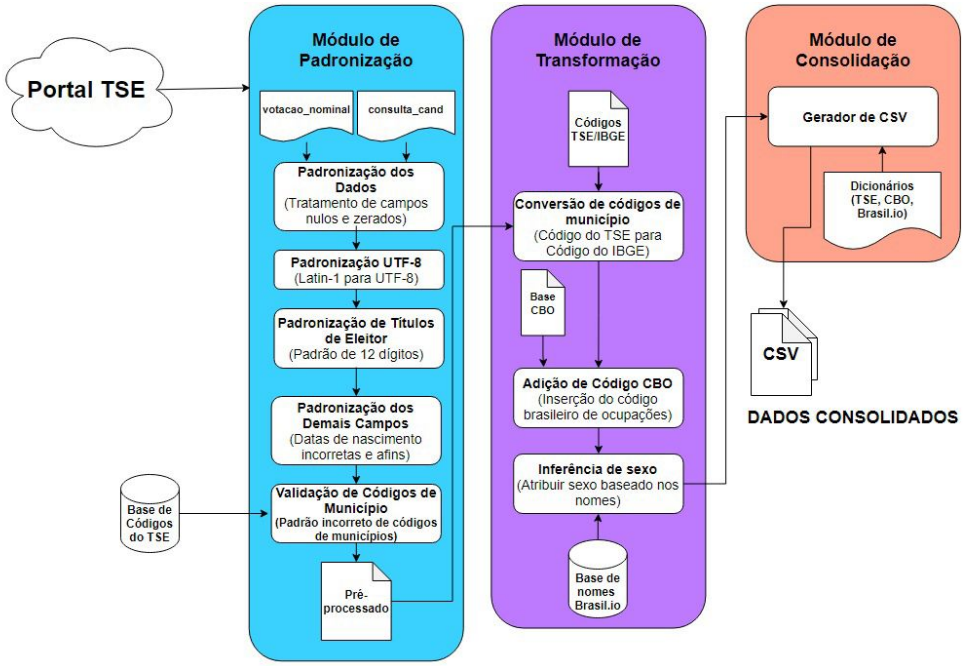

Figura 1. Fluxo de execução do desenvolvimento do CandiDATA

existência de uma vasta coleção de bibliotecas consolidadas nessa área. Nas próximas seções, são discutidas os procedimentos realizados em cada módulo.

4.1. Módulo de Padronização

Conforme descrito na Seção 3.1, foram encontrados diversos problemas relacionados à falta de padronização nos dados do TSE. Logo, a etapa inicial de construção do CandiDATA consiste no desenvolvimento de um módulo exclusivo para a padronização de dados relevantes. O foco desse módulo está em realizar tratamentos para formatação e codificação UTF-8 dos seguintes dados:

- **Titulo de Eleitor**: Todos os códigos de título de eleitor foram padronizados para o formato atual, de doze dígitos, por meio da adição de zeros à esquerda do número.
- **Campos nulos**: Os campos nulos foram padronizados, atribuindo valores `NULO` para *strings* e `-1` para números inteiros.
- **Data de nascimento**: As datas foram padronizadas no formato `yyyy-MM-dd`.
- **Nomes**: Nomes e outros campos do tipo *string* foram codificados para a especificação UTF-8, utilizando a biblioteca Unidecode[5].

4.2. Modulo de Transformação

O módulo de transformação foi a parte responsável pela correção de informações inconsistentes e adição das ausentes nas bases do TSE. Dentre os trabalhos realizados, destacam-se a adição tanto do código IBGE dos municípios como do CBO, além da inferência do gênero dos candidatos a partir de seu nome.

[5]`https://pypi.org/project/Unidecode/`

A conversão dos códigos dos municípios se fez necessária para uma melhor padronização do *dataset*, visto que a base do TSE utiliza-se de um código próprio para identificar cada município. Logo, foi utilizado o padrão de códigos de cidades disponibilizadas pelo IBGE, facilitando a integração com bases externas, incluindo as bases de dados sociais e econômicas acerca de cada município disponibilizadas pelo IBGE. Essa conversão foi feita com a ajuda de um *dataset* de conversão de códigos TSE/IBGE [6]

Em seguida, foram inferidas as ocupações dos candidatos, utilizando um *dataset*[7] que contém dados relativos ao Código Brasileiro de Ocupação (CBO). Esses códigos, padronizados pelo Ministério do Trabalho [MTE 2020], são utilizados em outras bases, como por exemplo, bases relacionadas à saúde pública [Clarindo et al. 2019], que também utilizam códigos CBO para identificar a ocupação de pacientes. Dados inválidos sobre ocupação ou onde ocorreram ambiguidades, foram atribuídos valores nulos.

Pelo fato dos dados do TSE não apresentarem informações sobre o gênero do candidato, foi realizada a inferência a partir da utilização de uma base de nomes pública[8], comparando o primeiro nome com o gênero listado no *dataset* de nomes. Caso não haja nomes equivalentes neste *dataset*, é atribuído um valor nulo.

4.3. Módulo de Consolidação

Por fim, o módulo de consolidação tem como função principal a transformação do arquivo final em formato CSV, que é um formato aberto, regulamentado pela *Internet Engineering Task Force* (IEFT), consistindo de um arquivo textual com formato tabular, que utiliza vírgulas como delimitadores [Shafranovich 2005]. Também ocorreu a inserção dos cabeçalhos nos arquivos, de modo que todas as edições das eleições passaram a ter os campos padronizados em um único formato.

4.4. Dicionário de dados

A partir dos dicionários de dados disponibilizados pelo TSE, foi construído um dicionário padronizado para CandiDATA apresentado na Tabela 1, o qual inclui atributos adicionais, como NR_IDADE_DATA_POSSE (idade do candidato), DT_NASCIMENTO (data de nascimento do candidato), NR_TITULO_ELEITORAL_CANDIDATO (título eleitoral do candidato), CD_GENERO (código do gênero), DS_GENERO (gênero do candidato), CD_CBO (código da ocupação do candidato), CD_MUN_IBGE (código do município baseado no IBGE) e os campos padronizados descritos na seção 4.2.

5. Aplicações do *CandiDATA*

Um diferencial importante do CandiDATA está na amplitude do período em que as informações são disponibilizadas e da inclusão de campos adicionais referentes a gênero, ocupações e título de eleitor dos candidatos. Além disso, outro aspecto relevante é a viabilidade de integração com bases externas, por meio da utilização de campos padronizados. Portanto, CandiDATA pode ser utilizado por gestores e pesquisadores interessados nas eleições brasileiras, bem como por profissionais da área de TI que busquem dados abertos para diferentes finalidades.

[6] https://github.com/betafcc/Municipios-Brasileiros-TSE
[7] https://github.com/datasets-br/cbo
[8] https://brasil.io/dataset/genero-nomes/nomes/

Tabela 1. Dicionário de dados do CandiDATA, onde os campos destacados em vermelho referem-se a campos não-existentes nas bases originais do TSE

VARIÁVEL	DESCRIÇÃO
DT_GERACAO	Data de extração dos dados para geração dos arquivos pelo TSE.
HH_GERACAO	Hora de extração dos dados para geração dos arquivos pelo TSE.
ANO_ELEICAO	Ano da eleição.
CD_TIPO_ELEICAO	Código do tipo de eleição.
NM_TIPO_ELEICAO	Nome do tipo de eleição.
NR_TURNO	Número do turno da eleição.
CD_ELEICAO	Código da eleição.
DS_ELEICAO	Descrição da eleição.
DT_ELEICAO	Data em que ocorreu a eleição.
TP_ABRANGENCIA	Tipo de abrangência da eleição.
SG_UF	Sigla da Unidade da Federação em que ocorreu a eleição.
SG_UE	Sigla da Unidade Eleitoral do candidato.
NM_UE	Nome de Unidade Eleitoral do candidato.
CD_MUNICIPIO	Código TSE do município onde ocorreu a eleição.
NM_MUNICIPIO	Nome do município onde ocorreu a eleição.
NR_ZONA	Número da Zona Eleitoral onde ocorreu a eleição
CD_CARGO	Código do cargo do candidato.
DS_CARGO	Descrição do cargo do candidato.
NR_IDADE_DATA_POSSE	Idade do candidato.
DT_NASCIMENTO	Data de nascimento
SQ_CANDIDATO	Número sequencial do candidato gerado internamente pelos sistemas eleitorais.
NR_CANDIDATO	Número do candidato na urna.
NR_TITULO_ELEITORAL_CANDIDATO	Título eleitoral do candidato.
NM_CANDIDATO	Nome completo do candidato.
NM_URNA_CANDIDATO	Nome de urna do candidato
NM_SOCIAL_CANDIDATO	Nome social do candidato.
CD_GENERO	Código do gênero do candidato.
DS_GENERO	Gênero do candidato.
CD_SITUACAO_CANDIDATURA	Código da situação do registro de candidatura do candidato.
DS_SITUACAO_CANDIDATURA	Descrição da situação do registro de candidatura do candidato.
CD_DETALHE_SITUACAO_CAND	Código dos detalhes da situação do registro de candidatura do candidato.
DS_DETALHE_SITUACAO_CAND	Descrição dos detalhes da situação do registro de candidatura do candidato.
TP_AGREMIACAO	Tipo de agremiação.
NR_PARTIDO	Número do partido do candidato.
SG_PARTIDO	Sigla do partido do candidato.
NM_PARTIDO	Nome do partido do candidato.
SQ_COLIGACAO	Sequencial da coligação gerado pela Justiça Eleitoral, da qual o candidato pertence.
NM_COLIGACAO	Nome da coligação da qual o candidato pertence.
DS_COMPOSICAO_COLIGACAO	Descrição da composição da coligação da qual o candidato pertence.
CD_SIT_TOT_TURNO	Código da situação de totalização do candidato naquele turno.
DS_SIT_TOT_TURNO	Descrição da situação de totalização do candidato naquele turno.
ST_VOTO_EM_TRANSITO	O indicativo se o quantitativo de votos se refere ao voto em trânsito.
QT_VOTOS_NOMINAIS	Quantidade total de votos nominais do candidato naquele município e zona.
CD_CBO	Código da ocupação do candidato no Código Brasileiro de Ocupações (CBO).
CD_MUN_IBGE	Código do município baseado no IBGE.

Conforme visto na seção 2, existem trabalhos que abordam inúmeros aspectos sobre as eleições brasileiras, como a participação feminina, a representatividade limitada e a falta de transparência. Diante disso, CandiDATA tem o potencial de promover diversas análises com perspectiva histórica como, por exemplo, a construção de gráficos mostrando o crescimento da participação das mulheres nas eleições no decorrer dos anos. Além disso, é possível estabelecer comparativos e correlações entre diferentes características dos candidatos tais como gênero, taxa de escolaridade, profissão, faixa etária, filiação partidária, etc.

Assim, com auxílio do CandiDATA pode-se realizar diversas consultas que gerem informações relevantes para a sociedade. A título de demonstração, as seções 5.1 e 5.2 apresentam exemplos de uso do *dataset*.

5.1. Participação feminina na política

Utilizando o *dataset* CandiDATA foi possível analisar a participação feminina no pleito eleitoral ao longo dos últimos anos. A Figura 2 demonstra a evolução da participação das mulheres como candidatas nas eleições a cada ano. Verifica-se um aumento significativo da participação feminina a partir do ano de 2008.

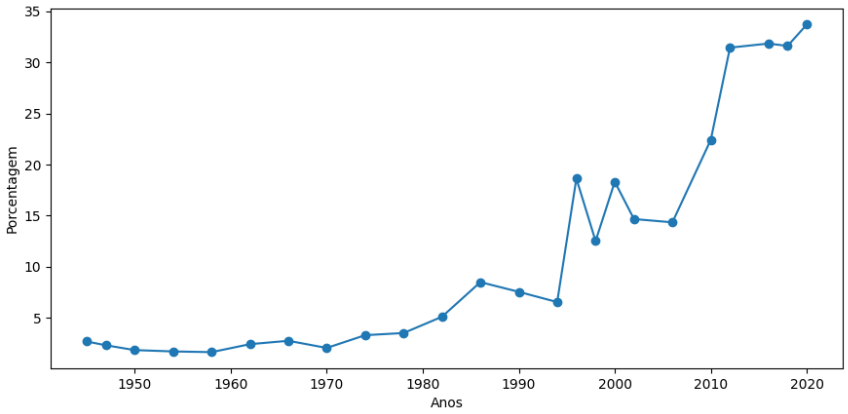

Figura 2. Porcentagem da participação feminina nas eleições

5.2. Profissões com maior participação política em 2020

Na Figura 3 é possível verificar as profissões com maior participação entre os candidatos do pleito eleitoral do ano de 2020. O perfil de agricultor foi o mais popular dentre os candidatos das eleições municipais de 2020.

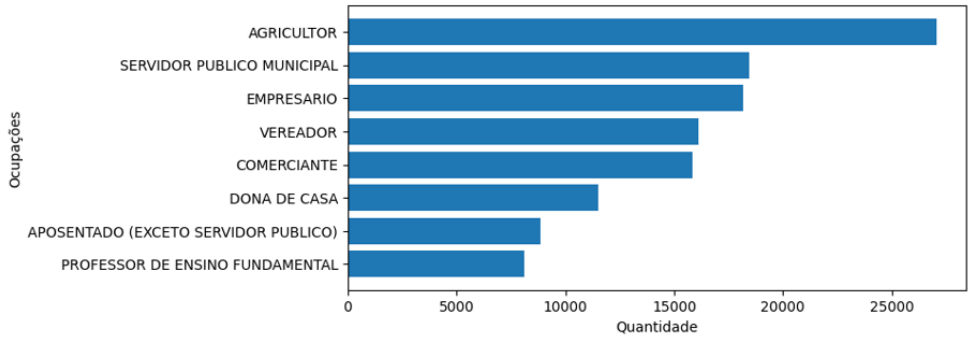

Figura 3. Ocupações dos candidatos nas eleições de 2020

6. Considerações finais

Este artigo apresentou o *dataset* CandiDATA, que disponibiliza um conjunto de dados sobre as eleições brasileiras ocorridas entre 1945 e 2020. Os dados foram padronizados e incluem campos adicionais que auxiliam pesquisadores e gestores em análises e na possível integração com bases externas.

Algumas limitações são encontradas em CandiDATA como, por exemplo, casos onde não foi possível inferir o genêro do candidato a partir da base de nomes, atribuindo-se, portanto, o valor nulo para essas situações.

Em relação aos trabalhos futuros, pretende-se expandir o *dataset* para a inclusão de campos relacionados a dados de partidos políticos bem como prestações de contas de eleições e declaração de bens dos candidatos. Além disso, novos formatos de arquivos, como JavaScript Object Notation (JSON), de uso comum em sistemas de gerenciamento de banco de dados não-relacionais, também serão disponibilizados.

O *dataset* encontra-se disponível no endereço `https://github.com/felipeVsc/CandiDATA` sendo dividido por ano e por Unidade Federativa e seguindo o padrão de nome: `votacao_candidato_munzona_ano_uf`. Neste repositório também está disponível um dicionário dos dados contendo a descrição de cada campo.

Referências

Araújo, C. (2009). Gênero e acesso ao poder legislativo no brasil: as cotas entre as instituições e a cultura. *Revista Brasileira de Ciência Política*, 2:23.

Barbosa, C. d. F., Cavalcanti, V. R. S., et al. (2009). Mulher na política: uma avaliação contemporânea. *SEMOC-Semana de Mobilização Científica-Segurança: A paz é o fruto da justiça*.

Camargo, A., Silva, R., Amaral, E., Heinen, M., and Pereira, F. (2016). Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do rio grande do sul. *Revista Brasileira de Computação Aplicada*, 8(1):64–73.

CEPESP, F. (2020). Cepespdata - banco de dados políticos. `https://empregabrasil.mte.gov.br/76/cbo/`. [Online; acesso em ago. 12].

Clarindo, J. P., Fontes, W., and Coutinho, F. (2019). Qualisus: um dataset sobre dados da saúde pública no brasil. In *Proceedings of the Sattelite Events of 34h Brazilian Symposium on Databases*.

Economist, T. (2021). Global democracy has a very bad year.

Filho, R. M. and Pappa, G. L. (2015). pg. 137-142. In *PROCEEDINGS OF SATELLITE EVENTS OF THE 30TH BRAZILIAN SYMPOSIUM ON DATABASES*, page 1–320.

MTE (2020). Classificação brasileira de ocupações. `https://empregabrasil.mte.gov.br/76/cbo/`. [Online; acesso em ago. 11].

Shafranovich, Y. (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180, RFC Editor.

Speck, B. and Speck, B. (2012). A filiação partidária no brasil. uma análise exploratória a partir de dados do tse e do wvs. In *Anais do 8º encontro da Associação Brasileira de Ciência Política*, page 21.

Tribunal Superior Eleitoral (2016). Urna eletrônica : 20 anos a favor da democracia. *Tribunal Superior Eleitoral*, pages 1–41.

TSE (2020). Repositório de dados eleitorais. `https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1`. [Online; acesso em ago. 11].