SPSafe: um *dataset* sobre dados de criminalidade no estado de São Paulo

Juliana Bernardes Freitas, João Paulo Clarindo, Cristina D. Aguiar

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

{julianabfreitas, jpcsantos}@usp.br, cdac@icmc.usp.br

Abstract. *Public safety has faced many challenges in recent years due to increased criminality. As a result, the Secretariat of Public Security of the State of São Paulo, Brazil, provides data on incident reports that managers and researchers can use for analysis. However, these data present several problems related to standardization, consistency, and heterogeneity, difficulting their use and integration with other databases. In this paper, we introduce SPSafe, a standardized dataset of incident reports registered between 2003 and 2022 in São Paulo, aiming to help public safety managers and researchers to use these data. SPSafe adds quality aspects to the data, such as eliminating inconsistencies and providing availability in open formats.*

Resumo. *A segurança pública tem enfrentado uma série de desafios com o aumento da criminalidade nos últimos anos. Com isso, a Secretaria de Segurança Pública do estado de São Paulo, Brasil, disponibiliza dados relativos aos boletins de ocorrências, que podem ser utilizados para análise por parte de gestores e pesquisadores. Entretanto, esses dados apresentam diversos problemas relacionados à padronização, consistência e heterogeneidade, dificultando o seu uso e integração com outras bases. Visando auxiliar gestores e pesquisadores na área de segurança pública, este artigo introduz SPSafe, um* dataset *padronizado de boletins de ocorrência registrados entre o ano de 2003 e 2022 no estado de São Paulo. SPSafe agrega aspectos de qualidade aos dados, como a eliminação de inconsistências e a disponibilização em formatos abertos.*

1. Introdução

Nos últimos anos, a segurança pública no Brasil enfrenta uma série de desafios complexos e persistentes, com o aumento nos índices em crimes como violência patrimonial (roubos, assaltos, furtos), homicídios e tráfico de drogas [Cerqueira et al. 2020]. No estado de São Paulo, por exemplo, de acordo com dados da Secretaria de Segurança Pública (SSP/SP), houve um aumento de crimes como roubos e furtos em 2022 em comparação com o período pré-pandemia da COVID-19 [Saleme 2023]. Com isso, foram criadas inúmeras iniciativas por parte do poder público para a disponibilização de dados relacionados a este tema nos níveis municipais, estaduais e federais [BRASIL 2022].

Os dados relacionados à segurança pública podem ser utilizados por gestores, pesquisadores e população em geral para identificar padrões e auxiliar na tomada de decisão, definindo novas políticas públicas para combater a criminalidade. Por exemplo, um gestor

pode identificar áreas com maior incidência de roubos, e com base nas informações obtidas por estes dados, direcionar recursos para essas regiões. Com isso, é fundamental que os dados sejam disponibilizados abertamente. A SSP/SP oferece dados relacionados aos Boletins de Ocorrência (BOs)[1], disponíveis mensalmente, por diversas categorias de crimes, como roubos e furtos de veículos e celulares, homicídios, feminicídios e latrocínios. O órgão ainda divulga trimestralmente índices relacionados a esses crimes.

A disponibilização de dados abertos governamentais é um desafio sob o ponto de vista técnico, sendo associado à implantação, aos princípios e aos formatos [Macedo and Lemos 2021]. Tim Berners-Lee propôs um sistema de *ranking* (ou modelo cinco estrelas) para definir níveis de acessibilidade para dados abertos [W3C 2013]. De acordo com esse *ranking*, os dados disponibilizados pela SSP/SP enquadram-se no nível dois, com dados publicados em formatos proprietários .xls e .xlsx, que são planilhas do Microsoft Excel.

Entretanto, mesmo disponibilizados em formatos conhecidos pelo público, os dados da SSP/SP possuem inúmeros problemas. Por exemplo, codificações de arquivos distintas, falta de padronização no nome dos campos, má formatação dos campos, campos de data e hora inconsistentes, falta de padronização em campos relacionados aos municípios e fragmentação dos dados em diversos arquivos. Esses problemas dificultam a manipulação e a análise dos dados. Isso significa que a utilização desses dados implica em uma maior complexidade para gestores, pesquisadores e interessados criarem soluções para a tomada de decisão com base nestes dados, devida à alta complexidade no processo de Extração, Transformação, e Carga (do inglês *Extract, Transform and Load*, ETL).

Logo, visando auxiliar pesquisadores e gestores que atuam na área da segurança pública, este trabalho introduz SPSafe, um *dataset* construído a partir de dados de BOs registrados entre o ano de 2003 e 2022 no estado de São Paulo. No processo de ETL da geração do *dataset*, foram aplicados diversos aspectos de melhoria aos dados, incluindo padronização, supressão de dados inválidos, eliminação de inconsistências, garantia de integração com outras bases, e mapeamento para os formatos abertos *Comma-Separated Values* (CSV) e *JavaScript Object Notation* (JSON). Com isso, SPSafe pode ser facilmente persistido em Sistemas Gerenciadores de Banco de Dados (SGBDs) relacionais e não-relacionais. Ele também pode ser utilizado em diversas aplicações, como *data warehousing*, aprendizagem de máquina, análise quantitativa e qualitativa, e na criação de índices estatísticos relacionados à temática de segurança pública. SPSafe está disponível gratuitamente para *download*, com os dados agrupados por ano.

Este artigo está estruturado da seguinte forma. Na seção 2 são discutidos trabalhos relacionados. Na seção 3 são descritos os dados disponibilizados pela SSP/SP e seus problemas. Na seção 4 são descritos os passos executados para a construção do SPSafe. Na seção 5 são descritas possíveis aplicações para o *dataset* e são apresentados exemplos de consultas que utilizam estes dados. O artigo é concluído na seção 6 com as considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos utilizam os dados de BOs oferecidos pela SSP/SP em inúmeros contextos, como para verificar o perfil de mulheres vítimas de violência durante a pandemia

[1] `http://www.ssp.sp.gov.br/transparenciassp/`

da COVID-19 [Evangelista et al. 2022], identificar padrões relacionados ao suicídio de pessoas idosas [Gianvecchio and Jorge 2023] e investigar a relação entre o desarmamento e as taxas de crime [Dutra et al. 2023]. Nesta seção, são discutidos trabalhos relacionados que realizam processos de ETL visando auxiliar na tomada de decisão.

Sá et al. (2021) apresentam PolRoute-DS, um *dataset* baseado nos dados de BOs registrados no estado de São Paulo entre 2003 e 2019. O objetivo é auxiliar na geração de rotas de patrulhamento policial. PolRoute-DS é baseado em um *data warehouse* de trajetória que possui tabelas de dimensão relacionadas aos segmentos de vias, sendo que conjuntos destes segmentos formam grafos que podem ser utilizados para definir rotas para patrulhamento policial. O *dataset* é composto por tabelas de dimensão e de fatos. O fato é o total de ocorrências por crime, por segmento de uma via. Os autores disponibilizam o *download* do *dataset* livremente.

Neto e Panhan (2020) propõem um modelo para geração de mapas, utilizando tecnologias gratuitas para o processo de ETL. Todo o processo de ETL é detalhado, identificando inúmeros problemas de padronização e qualidade dos dados. Os autores utilizam dados de BOs realizados no estado de São Paulo referentes ao ano de 2018 para verificar a incidência no roubo de veículos na cidade de Bragança Paulista. Os dados resultantes não são disponibilizados livremente.

Diferentemente destas soluções, SPSafe reúne dados de BOs registrados no estado de São Paulo entre 2003 e 2022, com foco na padronização e melhoria de campos existentes, e a adição de novos campos. SPSafe é disponibilizado em um formato que pode ser utilizado em diversos campos de pesquisa e aplicações, podendo ser utilizado desde uma análise quantitativa até como fonte inicial para criação de ambientes de *data warehousing* espacial. SPSafe encontra-se disponibilizado abertamente para *download*.

3. Dados da SSP/SP

O SPSafe foi construído a partir dos dados de BOs registrados no estado de São Paulo, os quais estão disponíveis no Portal de Transparência da SSP/SP. A Tabela 1 mostra os dados disponíveis neste portal, divididos em categorias, e com períodos distintos. A depender da categoria, os dados estão agrupados por mês e por ano.

Tabela 1. Categorias utilizadas para agrupamento de dados no Portal da SSP/SP, com os períodos disponíveis

Categoria do crime	Período
Feminicídio	2015 a 2022
Furto de celular	2010 a 2022
Furto de veículos	2003 a 2022
Homicídio	2017 a 2022
Latrocínio	2018 a 2022
Lesão corporal seguida de morte	2016 a 2022
Morte decorrente de intervenção policial	2013 a 2022
Morte suspeita	2013 a 2022
Roubo de celular	2010 a 2022
Roubo de veículos	2003 a 2022

Há uma variabilidade no formato dos arquivos de dados, sendo disponibilizados nos formatos desenvolvidos para Microsoft Excel, .xls, utilizado em versões anteriores ao Excel 2007, e .xslx, utilizado a partir do Excel 2007. O portal ainda disponibiliza

um dicionário de dados, contendo uma descrição breve sobre cada campo presente nestes arquivos.

3.1. Problemas Encontrados

Durante o processo de construção do SPSafe, foram detectados inúmeros problemas nos dados disponibilizados pela SSP/SP, como falta de padronização, inconsistências e variabilidade no formato dos dados. A seguir, são descritos em detalhes os problemas encontrados.

- **Diferentes formatos de arquivos**. Os arquivos referentes aos dados de BOs relacionados aos furtos e roubos de celulares e veículos são disponibilizados em .xls, enquanto os BOs de outros crimes são disponibilizadas em .xlsx.
- **Grande quantidade de arquivos**. Os dados são agrupados por categorias e por meses e anos, o que pode dificultar a análise quantitativa por outros atributos, como municípios.
- **Codificações de arquivos distintas**. Alguns arquivos estão codificados em UTF-8, enquanto outros em UTF-16, o que pode implicar em erros no carregamento e visualização dos dados.
- **Falta de padronização no nome dos campos**. O nome de mesmo campo varia dependendo do arquivo e, normalmente, esses nomes são diferentes do que consta no dicionário de dados.
- **Má formatação dos campos**. Alguns campos apresentam espaços em branco em quantidade excessiva, caracteres especiais e palavras acentuadas de forma incorreta, ocasionando erros de integração.
- **Campos de datas inconsistentes**. As data de nascimento e de ocorrências possuem formatos distintos como dd/MM/yyyy e yyyy-MM-dd HH:mm:ss. Também existem datas inválidas, como às referentes ao ano de 1899, por exemplo.
- **Falta de padronização nos nomes de municípios**. Municípios são nomeados sem padronização (como `SÃO PAULO`, `S. PAULO`, `SAO PAULO`), além de não haver campos relacionados aos códigos de municípios determinados pelo Instituto Brasileiro de Geografia e Estatística (IBGE).
- **Campos nulos**. Os campos referentes ao período da ocorrência e ao município possuem valor nulo em registros que têm informações como horário da ocorrência e dados de latitude e longitude.
- **Campos de horários inconsistentes** A grande maioria das ocorrências estão no formato 12h, porém sem indicação se é a.m. ou p.m..

4. Construção do SPSafe

A Figura 1 ilustra as etapas para a criação do SPSafe, que contém três módulos: (i) extração, no qual os dados são extraídos do Portal da SSP/SP utilizando técnicas *web scraping*; (ii) transformação, no qual os dados são padronizados e integrados com outras fontes, e (iii) carga, no qual os dados são dispostos na forma descrita no dicionário de dados. Estes módulos foram construídos utilizando a linguagem de programação Python. Nas seções 4.1 a 4.3 são discutidos os procedimentos e as tecnologias utilizadas em cada módulo, respectivamente. Detalhes sobre o dicionário são descritos na seção 4.4.

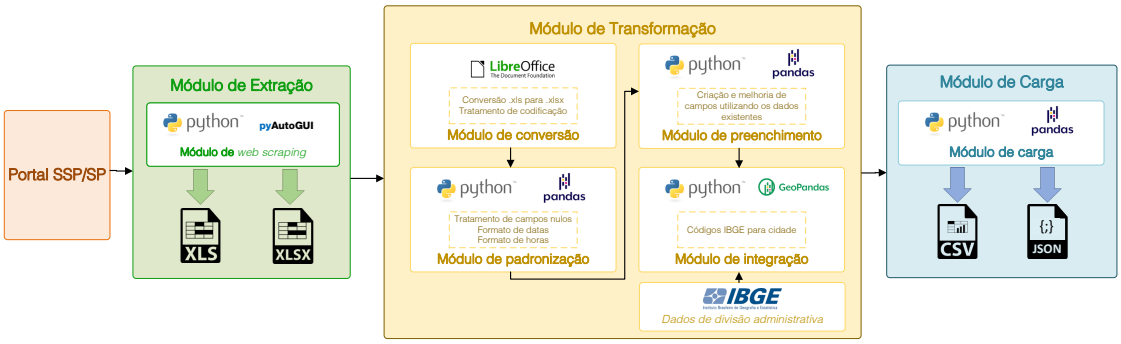

Figura 1. Fluxo de execução para a criação do SPSafe

4.1. Módulo de Extração

O módulo de extração é responsável por extrair os dados do portal da SSP/SP. Desde que não existe uma *Application Programming Interface* (API) disponível publicamente para a extração destes dados, foi implementado um *web scraping* utilizando a biblioteca PyAutoGUI[2] para essa finalidade. Essa biblioteca permite automatizar ações de teclado e *mouse* em outras aplicações, como um navegador *web*. Foram extraídos dados de 798 arquivos referentes aos anos de 2003 e 2022, visto que os do ano de 2023 estavam incompletos. Dentre esses arquivos, 6 estavam no formato .xlsx com codificação UTF-8 e subdivididos em abas por ano, e os outros 792 arquivos estavam no formato .xls com codificação UTF-16, agrupados por mês e ano. O volume extraído foi de 8GB.

4.2. Módulo de Transformação

O módulo de transformação é responsável pela conversão, padronização, preenchimento e integração dos dados. O primeiro modulo é o **módulo de conversão**, que consiste em uniformizar os formatos dos arquivos de dados para prover consistência nas operações de transformações seguintes. Para isso, optou-se em converter todos os arquivos .xls para .xlsx utilizando a ferramenta `soffice` disponível no pacote LibreOffice[3]. Na conversão, também foi uniformizada a codificação UTF-8.

Em seguida, os dados convertidos são enviados para o **módulo de padronização**, no qual, utilizando a biblioteca Pandas[4], ocorreram as seguintes transformações: (i) remoção de espaços em branco (*trimming*), acentos e caracteres especiais; (ii) uniformização dos campos no formato `string` em caixa alta; (iii) padronização de datas no formato yyyy-MM-dd e (iv) conversão de horários do formato 12h para o formato 24h, onde foi utilizado o campo PERIODO_OCORRENCIA. Por exemplo, se a hora da ocorrência estiver registrada como 02:47, e o período de ocorrência for “à tarde”, soma-se doze horas a esse horário, obtendo o horario correto, 14:47.

Após a padronização, os dados passam pelo **módulo de preenchimento**, em que campos nulos foram preenchidos a partir dos dados existentes. Os campos preenchidos são:

[2]`https://pyautogui.readthedocs.io/en/latest/`
[3]`https://libreoffice.org/`
[4]`https://pandas.pydata.org/`

- **PERIODO_OCORRENCIA**. Este campo pode ser preenchido verificando o campo HORA_OCORRENCIA. Se a hora de ocorrência for entre 0h e 5h, é atribuído o valor `DE MADRUGADA`; se for entre 05h e 12h, `PELA MANHA`; se for entre 12h e 18h, `A TARDE`; se for entre 18h e 0h, `A NOITE`. Caso o campo HORA_OCORRENCIA seja nulo, o campo de período de ocorrência é preenchido com o valor `EM HORA INCERTA`.
- **MARCA_VEICULO** e **MODELO_VEICULO**. Para facilitar consultas analíticas referentes aos roubos e furtos de veículos, dois novos campos foram criados, contendo dados sobre a marca e o modelo do veículo separadamente.

Por fim, os dados são encaminhados para o **módulo de integração**, no qual os dados da localização da ocorrência foram integrados com os dados de divisão territorial disponibilizados pelo IBGE[5]. Casos de inconsistência no campo original CIDADE foram resolvidos da seguinte forma. Utilizando dados de latitude e longitude, a cidade da ocorrência foi inferida a partir de uma junção espacial de *containment* entre o ponto e o polígono correspondente à divisão territorial do município. Para isso, foi utilizada a biblioteca GeoPandas[6]. Como resultado, foi criado um campo adicional COD_IBGE, que associa a cidade encontrada na junção espacial com um código padronizado pelo IBGE, com sete dígitos.

4.3. Módulo de Carga

O módulo de carga é responsável por converter os *dataframes* Pandas em arquivos nos seguintes formatos: (i) CSV, gerando arquivos de texto com formato tabular que utiliza vírgulas como separadores; e (ii) JSON, gerando arquivos em formato semi-estruturado utilizado amplamente por APIs e SGBDs não-relacionais. Estes formatos são regulamentados pela *Internet Engineering Task Force* (IEFT), garantindo a escala três de cinco no *ranking* de Berners-Lee [W3C 2013]. A conversão foi feita utilizando funções nativas do Python. Devido ao grande volume, os dados foram agrupados por ano. Os dados foram carregados com os campos conforme o dicionário de dados disponibilizado.

4.4. Dicionário de dados

Utilizando os dados disponibilizados pela SSP/SP, foi criado um dicionário de dados para o SPSafe, apresentado na Tabela 2. Os campos destacados em azul referem-se aos seguintes campos adicionais: CODIGO_BOLETIM (junção do número de boletim com o ano do boletim), COD_IBGE (código IBGE da cidade da ocorrência), PONTO_CRIME (ponto geográfico da ocorrência), MARCA_VEICULO (marca do veículo envolvido na ocorrência) e MODELO_VEICULO (modelo do veículo envolvido na ocorrência). Campos destacados em verde referem-se aos atributos que sofreram uma melhora significativa na etapa de transformação dos dados, como o campo CIDADE.

5. Aplicações do SPSafe

O *dataset* SPSafe pode ser utilizado por profissionais da segurança pública, pesquisadores, gestores e analistas de dados. Devido à padronização dos dados, é possível integrar

[5]`https://www.ibge.gov.br/geociencias/todos-os-produtos-geociencias.html`

[6]`https://geopandas.org/`

Tabela 2. Dicionário de dados do SPSafe, onde os campos destacados em azul referem-se a campos não-existentes nas bases originais da SSP/SP, enquanto os campos destacados em verde referem-se a campos que sofreram melhorias significativas no processamento de transformação

Dados gerais do boletim de ocorrência	
Nome do Campo	**Descrição do Campo**
NUM_BO	Número do boletim de ocorrência
ANO_BO	Ano do boletim de ocorrência
CODIGO_BOLETIM	Junção do número de boletim com o ano do boletim separados por '/'
NATUREZA_APURADA	Tipo de crime cometido
DATA_OCORRENCIA	Data em que o crime ocorreu
HORA_OCORRENCIA	Hora em que o crime ocorreu
PERIODO_OCORRENCIA	Período do dia em que o crime ocorreu
CIDADE	Cidade em que o crime ocorreu
COD_IBGE	Código da cidade em que o crime ocorreu, no formato IBGE
LOGRADOURO	Via em que o crime ocorreu
NUMERO_LOGRADOURO	Número que identifica em uma via o local do crime
BAIRRO	Bairro em que o crime ocorreu
UF	Sigla da unidade federativa em que o crime ocorreu
TIPO_LOCAL	Tipo de local em que o crime ocorreu
LATITUDE	Latitude do ponto em que o crime ocorreu
LONGITUDE	Longitude do ponto em que o crime ocorreu
PONTO_CRIME	Ponto geográfico em que o crime ocorreu, no formato WKT
DELEGACIA_ELABORACAO	Delegacia em que o boletim de ocorrência foi elaborado
DEPARTAMENTO_ELABORACAO	Departamento em que o boletim de ocorrência foi elaborado
SECCIONAL_ELABORACAO	Seccional em que o boletim de ocorrência foi elaborado
Dados sobre a pessoa envolvida na ocorrência	
Nome do Campo	**Descrição do Campo**
TIPO_PESSOA	Indica se a pessoa é a vítima ou autora do crime
GENERO_PESSOA	Gênero
IDADE_PESSOA	Idade
DATA_NASCIMENTO_PESSOA	Data de nascimento
COR_PELE	Cor de pele
PROFISSAO	Profissão
Dados sobre o veículo envolvido na ocorrência, caso seja aplicável	
Nome do Campo	**Descrição do Campo**
PLACA_VEICULO	Placa do veículo
UF_VEICULO	Unidade federativa do emplacamento
CIDADE_VEICULO	Cidade do emplacamento
COR_VEICULO	Cor do veículo
MARCA_VEICULO	Marca do veículo
MODELO_VEICULO	Modelo do veículo
ANO_FABRICACAO	Ano de fabricação do veículo
ANO_MODELO	Ano do modelo do veículo
TIPO_VEICULO	Tipo de veículo envolvido
Dados sobre o telefone celular envolvido na ocorrência, caso seja aplicável	
Nome do Campo	**Descrição do Campo**
MARCA_CELULAR	Marca do celular
QUANT_CELULAR	Quantidade de celulares
Outras informações sobre o boletim de ocorrência	
Nome do Campo	**Descrição do Campo**
BO_INICIADO	Data e hora em que o BO foi iniciado
BO_EMITIDO	Data e hora em que o BO foi concluído
DATA_HORA_ELABORACAO	Data e hora de elaboração do BO
DATA_COMUNICACAO	Data em que o BO foi comunicado a delegacia
BO_AUTORIA	Responsável pela realização do BO
FLAGRANTE	Indica que se trata de uma situação de flagrante
EXAME	Responsável pelo exame de corpo
SOLUCAO	Tipo de solução dada ao crime
ESPECIE	Espécie de patrimônio envolvido no crime
STATUS	Status do crime
FLAG_VITIMA_FATAL	Indica se houve fatalidades
DESDOBRAMENTO	Desdobramento do caso

o *dataset* com bases externas, possibilitando a tomada de decisão em áreas relacionadas à segurança pública, como índices de distribuição de renda, moradia, transporte público e educação em uma cidade. SPSafe ainda pode ser utilizado como base para *datasets* de treinamento e teste para algoritmos de aprendizagem de máquina para prever, por exemplo, a incidência de crimes em uma região.

Nas seções 5.1 e 5.2 são detalhados dois exemplos de consultas que podem ser feitas utilizando SPSafe, sendo os resultados plotados em gráficos e mapas.

5.1. Quantidade de roubos e furtos no estado de São Paulo agrupados por ano

Nesse exemplo, é calculada a quantidade de ocorrências de roubos e furtos em todo o estado de São Paulo no período entre 2018 e 2022, agrupadas pelo ano de registro do boletim. Esses números podem ser relevantes no entendimento de quais eventos anuais afetam ou não a quantidade de crimes no estado. O resultado dessa consulta pode ser visualizado na Figura 2.

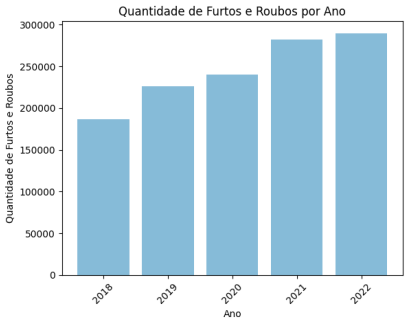

Figura 2. Gráfico de barras da quantidade total de Boletins de Ocorrência relacionados aos roubos e furtos no estado de São Paulo por ano (2018-2022)

5.2. Distribuição da quantidade de BO por município do estado de São Paulo

Nessa consulta, é calculada a quantidade de ocorrências por município no estado de São Paulo no período entre 2003 e 2022. Devido à diferença na quantidade de ocorrências entre os municípios, os dados foram normalizados de acordo com a população do município para melhor visualização. O resultado, ilustrado na Figura 3, com destaque na Região Metropolitana do estado, evidencia uma grande concentração de crimes na cidade de São Paulo, e em cidades da Baixada Santista. No mapa, destaca-se uma maior ocorrência de crimes em cidades com menor densidade populacional, como Miracatu e Boituva. Essa análise pode ser relevante na definição de políticas públicas que priorizem combater a criminalidade em cidades com maiores índices de criminalidade.

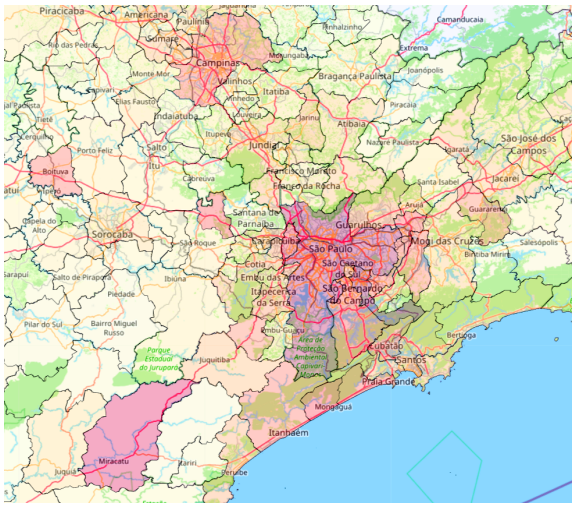

Figura 3. Mapa de cores indicando o número de BOs no estado de São Paulo, com destaque na Região Metropolitana, no período entre 2003 e 2022, normalizadas conforme a população do município.

6. Conclusão

Este artigo apresentou SPSafe, um *dataset* construído a partir de dados de boletins de ocorrência disponibilizados pela Secretaria de Segurança Pública do estado de São Paulo. Para a criação do SPSafe, foi desenvolvido um fluxo de execução, consistindo de extração, transformação e carga. Na extração, foram utilizadas técnicas de *web scraping*. Na transformação, os dados foram convertidos, padronizados, preenchidos em caso de campos nulos e integrados com dados de outras fontes. Finalmente, na carga, os dados foram disponibilizados por meio de arquivos de formato aberto (CSV e JSON). Como resultado, SPSafe pode ser integrado facilmente às ferramentas para tomada de decisão, SGBDs relacionais e não-relacionais e *dashboards* interativos.

O objetivo do SPSafe é auxiliar gestores e pesquisadores que atuam na área de segurança pública, disponibilizando dados padronizados e com garantia de integração com outras bases de dados, no formato aberto. Um exemplo de uso de SPSafe é no auxílio para a criação de um ambiente de *data warehousing* espacial para auxiliar um gestor de segurança pública na tomada de decisão estratégica. Em [Freitas et al. 2023], é introduzido um esquema-estrela baseado em dados de criminalidade, e nos estudos de caso, SPSafe foi utilizado para a execução de consultas analíticas espaciais.

SPSafe encontra-se disponível em `https://github.com/julianabfreitas/SPSafe`, nos formatos CSV e JSON. Os dados encontram-se agrupados por ano. A documentação relativa aos campos e tabelas adicionais para integração, como divisão territorial dos municípios do estado de São Paulo, também são disponibilizadas.

Existem algumas limitações no *dataset* devido à não-disponibilidade de dados de boletins de ocorrência por parte da Secretaria de Segurança Pública do estado de São Paulo. Por exemplo, dados de latrocínios são disponibilizados somente a partir de 2010. Como resultado, algumas consultas quantitativas relacionadas ao período entre 2003 e 2018 podem se mostrar inconsistentes em certos agrupamentos, como por regiões. Além disso, há campos nulos que não podem ser inferidos a partir de outros dados disponíveis, como o *status* de crimes.

Em trabalhos futuros, pretende-se expandir SPSafe a partir da inclusão de dados de outras fontes. Uma primeira fonte é o Sistema de Informações sobre Mortalidade do SUS (SIMSUS), a qual identifica as causas de óbitos, incluindo as decorrentes de violência [Clarindo et al. 2019]. Outra fonte a ser integrada refere-se ao Sistema de Informações do Departamento Penitenciário Nacional (SISDEPEN), que lista dados sobre prisões[7]. A partir da integração dessas duas fontes, é possível realizar outras análises, como os feminicídios antes de 2015.

Agradecimentos

Este trabalho foi apoiado pela Universidade de São Paulo, a partir do Programa Unificado de Bolsas de estudo para apoio à formação de estudantes de graduação (PUB-USP), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

[7] `https://www.gov.br/senappen/pt-br/servicos/sisdepen`

Referências

BRASIL (2022). Plano de Dados Abertos 2022-2024. Technical report, Ministério da Justiça e Segurança Pública, Brasília.

Cerqueira, D., Bueno, S., Palmieri Alves, P., Sergio de Lima, R., R. A. da Silva, E., Ferreira, H., Pimentel, A., Barros, B., Marques, D., Pacheco, D., de Oliveira Accioly Lins, G., dos Reis Lino, I., Sobral, I., Figueiredo, I., Martins, J., Chacon Armstrong, K., and da Silva Figueiredo, T. (2020). Atlas da Violência 2020. *Relatório Institucional*, pages 1–91.

Clarindo, J. P., Fontes, W. d. S., and Coutinho, F. (2019). QualiSUS: um dataset sobre dados da Saúde Pública no Brasil. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2019 Companion*, pages 418–428, Fortaleza, CE. SBC.

Dutra, G. J., Silva, L. A., Oliveira, P. R., and Lírio, V. S. (2023). A relação entre o desarmamento e as taxas de crime no estado de São Paulo. *Revista Econômica do Nordeste*, 1(1):49–66.

Evangelista, D. N., Mazzu-Nascimento, T., Rodríguez-Martín, D., Negri, M., Lisboa, U. P. d. S., Sousa, A. S., Abubakar, O., and Aciole, G. G. (2022). Violência contra a mulher no estado de São Paulo: o perfil das vítimas durante a pandeia da COVID-19. *Hygeia - Revista Brasileira de Geografia Médica e da Saúde*, 18.

Freitas, J. B., Clarindo, J. P., and Aguiar, C. D. (2023). Ambiente de data warehousing espacial para tomada de decisão sobre dados de crimes. In *XXXVIII Simpósio Brasileiro de Banco de Dados: Workshop de Trabalhos de Alunos da Graduação (WTAG), SBBD 2023 Companion*, Belo Horizonte. SBC.

Gianvecchio, V. A. P. and Jorge, M. H. P. d. M. (2023). Estudo do suicídio na população idosa do Estado de São Paulo, Brasil, segundo dados da segurança pública. *Debates em Psiquiatria*, 13:1–20.

Macedo, D. F. and Lemos, D. L. d. S. (2021). Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. *AtoZ: novas práticas em informação e conhecimento*, 10(2):14.

Neto, A. G. V. and Panhan, A. M. (2020). Modelo para geração de mapas criminais: Uma análise a partir de estudo de caso de dados abertos. *Revista do Instituto Brasileiro de Segurança Pública (RIBSP) - ISSN 2595-2153*, 2(5):171–185.

Sá, B. C., Muller, G., Banni, M., Santos, W., Lage, M., Rosseti, I., Frota, Y., and Oliveira, D. d. (2021). PolRoute-DS: um Dataset de Dados Criminais para Geração de Rotas de Patrulhamento Policial. *Anais do Dataset Showcase Workshop (DSW)*, pages 117–127.

Saleme, I. (2023). São Paulo tem aumento de roubos e furtos em 2022 e alta de crimes no fim do ano. In *CNN Brasil. Disponível em: https://www.cnnbrasil.com.br/nacional/sao-paulo-tem-aumento-de-roubos-e-furtos-em-2022-e-alta-de-crimes-no-fim-do-ano/*, São Paulo.

W3C (2013). Linked Data Glossary. Technical report, W3C.