

Bus&City: Dados de Transporte Coletivo Urbano Enriquecidos com Informações Criminais e de Relevo

Vinícius Gabriel de Jesus Almeida¹, Thais R. M. Braga Silva¹, Fabrício A. Silva¹

¹Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPED-LAB)
Universidade Federal de Viçosa, Florestal, MG, Brasil

{vinicius.jesus, thais.braga, fabricio.asilva}@ufv.br

Abstract. *This study presents Bus&City, an enriched dataset for bus routes in the cities of São Paulo and Curitiba, incorporating crime and topography information. Using government sources and data from INPE, the dataset aims to facilitate the development of works that improve the displacement of individuals through Context-Aware Mobility. The bus routes were structured into graphs, allowing for detailed analyses and adjustments based on safety and topography. The data is publicly available to support both researchers and public managers in creating safer and more efficient routes.*

Resumo. *Este estudo apresenta Bus&City, um conjunto de dados enriquecidos para rotas de ônibus nas cidades de São Paulo e Curitiba, incorporando informações sobre criminalidade e relevo. Utilizando fontes governamentais e dados do INPE, o dataset visa facilitar o desenvolvimento de trabalhos para a melhoria do deslocamento de indivíduos através da Mobilidade Ciente de Contexto. As rotas de ônibus foram estruturadas em grafos, permitindo análises detalhadas e ajustes baseados em segurança e topografia. Os dados são disponibilizados publicamente para apoiar tanto pesquisadores quanto gestores públicos na criação de rotas mais seguras e eficientes.*

1. Introdução

A *Mobilidade Urbana* refere-se à facilidade com que indivíduos, como pedestres, ciclistas, motoristas de veículos particulares e usuários de transporte público, se deslocam nas cidades [Aragão 2006]. Esse conceito tem sido fundamental nos últimos anos para a realização de diversas atividades sociais e econômicas, tais como aquelas ligadas a educação, trabalho e lazer [Magagnin and da Silva 2008]. No entanto, com a complexidade crescente dos grandes centros urbanos, inúmeros desafios relacionados ao deslocamento das pessoas têm sido identificados.

Um dos fundamentos para aprimorar a mobilidade dentro das cidades é a chamada *Mobilidade Ciente de Contexto*, a qual se baseia na utilização de um ou mais tipos de dados sobre as entidades envolvidas, chamados contextos, para melhorar os fluxos de deslocamento de veículos, motoristas e passageiros, tanto de transportes privados quanto públicos [Vahdat-Nejad et al. 2016]. Um dos serviços associados à *Mobilidade Ciente de Contexto* é a Construção de Rotas Contextuais, o que envolve o uso de dados contextuais, combinando uma ou mais fontes de informações, com o objetivo de encontrar trajetórias específicas, que atendam a determinados requisitos. Dessa forma, os usuários podem

escolher, com base no contexto, o tipo de rota que desejam seguir, seja ela a mais curta, a menos congestionada, a mais cênica, a mais segura, entre outras.

Entre os diversos contextos possíveis, um dos tipos que tem atraído bastante atenção na literatura sobre *Mobilidade Urbana* nos últimos anos são as rotas seguras [Santos et al. 2017]. Essas rotas evitam áreas com altos índices de criminalidade. Ou seja, a partir de dados sobre crimes, posteriormente convertidos por uma estratégia para identificar regiões inseguras, o algoritmo de recomendação de rotas seria capaz de compreender esse contexto e, assim, traçar um trajeto que se afaste dessas áreas [Ladeira et al. 2019]. No entanto, os estudos na área têm se concentrado apenas em veículos particulares, negligenciando os transportes coletivos urbanos [Almeida et al. 2022]. Um dos motivos para isso é a dificuldade e a falta de disponibilidade de dados enriquecidos com informações contextuais sobre os transportes coletivos urbanos.

O objetivo deste trabalho é disponibilizar um conjunto de dados de rotas de ônibus de duas grandes cidades do Brasil, São Paulo e Curitiba, enriquecidos com dados criminais das mesmas, i.e., informações sobre as localidades que mais concentram registros de ocorrência de crimes, de acordo com as autoridades policiais. Para complementar, os dados de relevo das cidades também foram adicionados, uma vez que este é um outro contexto que afeta o deslocamento de passageiros entre pontos de parada de ônibus. Este *dataset*, chamado de Bus&City, foi disponibilizado em uma estrutura de grafo, um para cada cidade, em que cada ponto de parada de cada linha de ônibus é representado como um vértice enriquecido, e as arestas entre estes vértices representam vias do município, utilizadas no trajeto de cada linha. Com esse conjunto de dados, tanto pesquisadores quanto gestores públicos podem avaliar e ajustar rotas de ônibus visando a melhoria da qualidade de vida da população.

O restante deste trabalho está organizado da seguinte forma: na seção 2 são apresentados os trabalhos relacionados encontrados. Em seguida, o processo de coleta e preparação dos dados é discutido na seção 3. Possíveis aplicações do conjunto de dados são apresentadas na seção 4. Por fim, as considerações finais, com instruções sobre a forma de obtenção da base de dados Bus&City, são apresentadas na seção 5.

2. Trabalhos Relacionados

Em [Scabora et al. 2019] é apresentado G-FranC, um dataset de atividades criminais mapeadas como uma rede complexa em um sistema de gerenciamento de banco de dados relacional. Neste trabalho, a malha viária da cidade de São Francisco, na Califórnia/EUA, é representada como um grafo e os crimes, obtidos por meio de uma base de dados policial, mapeados nos vértices. As arestas são ponderadas utilizando distâncias como peso. Os dados foram disponibilizados pelos autores em tabelas relacionais PostgreSQL, juntamente com *scripts* utilizados para a geração dos esquemas necessários.

PolRoute-DS [Sá et al. 2021] é uma base de dados que viabiliza a testagem de possíveis rotas a serem utilizadas por policiais em centros urbanos. O dataset une o mapa da malha viária da cidade, no caso São Paulo/SP, representada como um conjunto de vértices e segmentos de via com comprimento de 150 a 200 metros, a dados sobre criminalidade obtidos de fontes policiais. Os crimes são mapeados em arestas e o resultado é armazenado em um banco de dados relacional, cujo esquema é composto por sete tabelas. Por fim, além de estatísticas sobre o dataset, os autores descreveram um estudo de caso

sobre a criação de rotas utilizando uma estratégia ponderada, visando fazer com que as viaturas cubram áreas com crimes de maior ameaça à vida.

Utilizando APIs específicas disponibilizadas por quatro cidades, quais sejam Rio de Janeiro/RJ, Curitiba/PR, São Paulo/SP e Brasília/DF, além de informações sobre suas linhas de ônibus, disponíveis na plataforma Moovit, os autores em [Melo et al. 2023] apresentam o dataset BRBus, composto por dados geoespaciais dos veículos coletivos destes centros urbanos. Os dados das diferentes localidades consideradas foram consolidados, limpos e padronizados. Em seguida, os autores enriqueceram o BRBus com informações de outras fontes, armazenando e disponibilizando a base por meio de arquivos CSV. As aplicações do dataset vão do monitoramento de parâmetros de cada cidade, até a comparação dos resultados entre cidades distintas, por meio de análise exploratória.

Com o objetivo de melhorar a qualidade dos dados disponibilizados pelas autoridades policiais da cidade de São Paulo/SP, os autores em [Freitas et al. 2023] apresentam a base de dados SPSafe. Nela, estão consolidados, de forma padronizada e sem inconsistências, os dados criminais referentes aos boletins de ocorrência registrados na cidade no período de 2003 a 2022. Os dados finais foram armazenados e disponibilizados pelos autores em arquivos de formato CSV e JSON, por serem formatos abertos, facilmente integráveis a APIs, bancos de dados relacionais ou modelos de aprendizado de máquina.

Os trabalhos de [Scabora et al. 2019], [Sá et al. 2021] e [Freitas et al. 2023], de forma similar a este, fazem uso de dados de criminalidade obtidos por meio de fontes policiais, na elaboração do dataset. Os dois primeiros, em particular, mapeiam estes dados, sobre um grafo que representa a malha viária de uma cidade, enquanto o último apenas trata de melhorar a qualidade das informações já oferecidas pelas autoridades. Nenhum destes trabalhos associa a utilização destas informações criminais ao sistema de transporte público das cidades relacionadas. Por outro lado, [Melo et al. 2023] apresenta uma base de dados ligada às linhas de ônibus de quatro cidades. Entretanto, o foco está em oferecer a possibilidade de análises geoespaciais sobre os veículos coletivos enquanto os mesmos percorrem seus trajetos, não envolvendo assim nenhum tipo de informação sobre segurança pública. Por fim, nenhum dos datasets apresentados associa dados sobre relevo ao mapa viário das localidades consideradas.

[Ladeira et al. 2019] propõem uma solução para a construção de rotas seguras para veículos particulares, baseada em clusterização com o algoritmo *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) ¹. Esse estudo faz uso dos *datasets* da Here API ² para obter as rotas mais rápidas dos veículos e do *Open Data Portal* ³ de Chicago para acessar os dados criminais da cidade. Entretanto, após o processamento dessas informações, nenhuma das bases derivadas foi disponibilizada para uso.

Outro trabalho relevante é o de [de Souza and Villas 2020], que apresenta um sistema completo de planejamento de rotas para veículos particulares com base em previsões futuras do fluxo da cidade, considerando fatores urbanos como mobilidade e criminalidade. Os autores utilizaram, em sua solução, uma rede neural *Long Short Term Memory*

¹<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

²<https://developer.here.com/develop/rest-api>

³Chicago Data Portal: <https://data.cityofchicago.org/>

(LSTM) e, para a clusterização dos crimes, a função de densidade Gaussiana (KDE). Os dados de mobilidade e segurança pública também foram obtidos do *Open Data Portal* em conjunto com o *OpenStreetMap*⁴. No entanto, após o processamento, nenhuma das bases geradas foi disponibilizada.

Por outro lado, diferentemente dos trabalhos apresentados, o foco da solução desenvolvida neste estudo está em um *dataset* contendo as rotas de ônibus, enriquecido com dados de criminalidade e relevo. Além disso, este trabalho disponibiliza o *dataset*, ao contrário dos demais.

3. Construção do Dataset

O dataset apresentado neste trabalho foi construído por meio de módulos ETL (*Extract, Transform and Load*). A Figura 1 ilustra os componentes de cada um desses módulos, suas respectivas entradas e saídas e a interligação entre eles. Os detalhes sobre como cada módulo foi implementado serão apresentados na sequência.

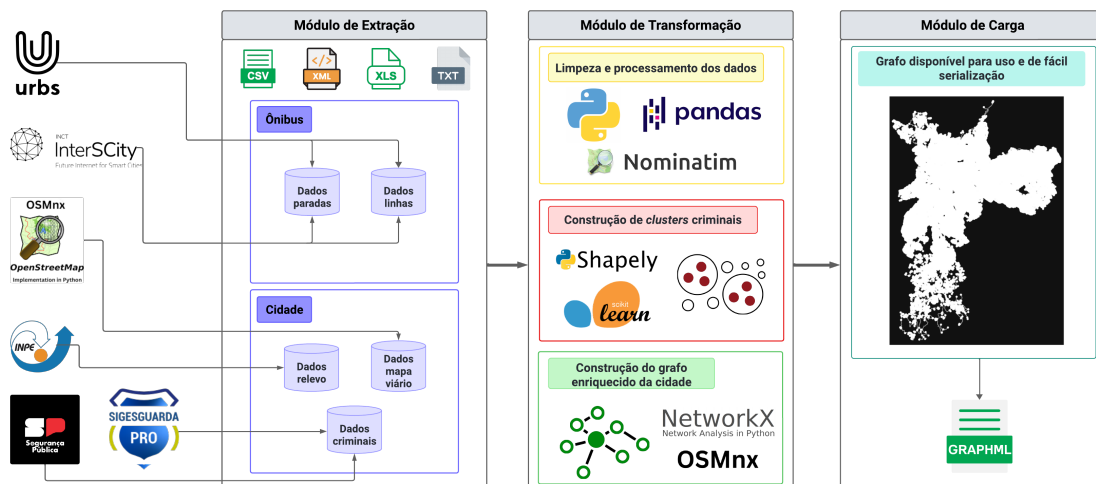


Figura 1. Fluxograma de etapas de processamento dos dados.

3.1. Extração

Os dados necessários para a construção do dataset Bus&City, para além do grafo que representa a malha viária das cidades, são: dados criminais, dados da malha rodoviária municipal e dados de relevo. A seguir está descrita a forma de obtenção de cada um deles.

3.1.1. Dados criminais

Para construir bases de dados criminais confiáveis, optou-se sempre por fontes de dados governamentais, através de portais de transparência e canais de comunicação oficiais. Além disso, como o processo de coleta começou por volta do início de 2022, o ano escolhido para a coleta completa foi 2019, pois esse não contava com interferências causadas pela pandemia de Covid-19.

⁴<https://www.openstreetmap.org/>

Para a cidade de São Paulo, os dados criminais foram obtidos através da plataforma **SSP**⁵, responsável por todas as polícias em todo o estado. Devido ao fato do dataset estar relacionado a pedestres e trânsito, dentre as categorias de registros criminais presentes na plataforma, as que mais se mostraram coerentes, foram: furto e roubo de veículos e furto e roubo de celulares, totalizando 48 arquivos, sendo um para cada mês de 2019, para cada uma das quatro categorias mencionadas.

Já para Curitiba, a coleta foi feita através da plataforma oficial da cidade **SiGes-Guarda**⁶, na qual diferentemente de São Paulo, os registros criminais de Curitiba estão agrupados em um único arquivo, que vem sendo incrementado desde 2009. Uma das principais dificuldades encontradas ao procurar dados criminais oficiais de Curitiba foi encontrá-los com anotações de geolocalização. Dentre todas as bases criminais analisadas, a informação mais próxima da localização esperada foi o logradouro mais adjacente à ocorrência.

3.1.2. Dados da malha rodoviária

Começando por São Paulo, os dados referentes às linhas de ônibus foram obtidos de forma gratuita e aberta através da plataforma **InterSCity**⁷. Para a cidade de Curitiba, os dados também foram obtidos de forma gratuita, através da plataforma **URBS**⁸. O acesso se deu através de uma solicitação via e-mail para que pudéssemos utilizar a API da plataforma, e assim, obter todas as informações de linhas de ônibus e pontos de paradas necessárias.

3.1.3. Dados de relevo

Para se obter os dados de relevo das cidades de São Paulo e Curitiba, a base de dados geomorfométrica do Brasil, disponibilizada publicamente pelo INPE (Instituto Nacional de Pesquisas Espaciais) em **TOPODATA**⁹ foi utilizada.

3.2. Transformação

O dataset **Bus&City** é composto pela malha viária das cidades envolvidas, enriquecida com informações sobre as linhas de ônibus das mesmas, polígonos criminais mapeados e dados sobre o relevo municipal. Para representar as vias das cidades, foi utilizada a bibli-

⁵Secretaria de Estado da Segurança Pública de São Paulo. <http://www.ssp.sp.gov.br/transparenciassp/Consulta2022.aspx>

⁶SiGesGuarda - Sistema de Gestão da Guarda Municipal de Curitiba. Base contendo os dados das ocorrências atendidas pela Guarda Municipal de Curitiba. <https://www.curitiba.pr.gov.br/dadosabertos/busca/?termo=guarda>

⁷O InterSCity é um projeto de pesquisa colaborativo sediado pelo Instituto Nacional de Ciência e Tecnologia (INCT) da Internet do Futuro para Cidades Inteligentes. https://interscity.org/open_data/

⁸A Urbanização de Curitiba S/A é uma empresa de economia mista que administra equipamentos e espaços públicos, além do sistema de transporte da cidade de Curitiba. <https://www.urbs.curitiba.pr.gov.br/>

⁹TOPODATA - Banco de Dados Geomorfométricos do Brasil. <http://www.dsr.inpe.br/topodata/acesso.php>

oteca OSMnx¹⁰, que se baseia na consolidada biblioteca NetworkX¹¹ para a construção dos grafos, além de enriquecer os vértices e arestas do grafo com dados do OpenStreetMap¹², que é um projeto de mapeamento colaborativo para criar um mapa livre e editável do mundo. De acordo com o OSMnx, os vértices são esquinas das cidades e as arestas as vias das mesmas.

Alguns desafios foram encontrados ao manipular as bases de dados brutas, além das conversões e ajustes necessários para adaptar esses dados limpos à estrutura de grafo do OSMnx utilizada. Com isso, a seguir, serão apresentadas as principais manipulações feitas sobre as bases de dados, os novos *datasets* obtidos pós processamento, e a integração desses com os grafos construídos através da biblioteca OSMnx.

3.2.1. Dados criminais

Em relação aos dados criminais, para São Paulo, o primeiro passo para tratá-los foi fazer a limpeza dos dados vazios, crimes pertencentes às regiões metropolitanas, e colunas desnecessárias para nosso modelo, considerando como informação válida apenas linhas contendo informações do crime, data completa da ocorrência, endereço e latitude e longitude. Além disso, também foi necessário aplicar filtros para manter apenas crimes com rubricas relacionadas ao contexto de transporte público, envolvendo situações relacionadas ao trânsito, mas também aos passageiros. Com isso, ao final de todo esse processo, todos os arquivos foram unificados em um único documento, a fim de facilitar a manipulação, com 212.079 registros criminais geolocalizados.

Já para Curitiba, para cada crime registrado, a geolocalização atribuída à ele foi obtida através da conversão do logradouro para as coordenadas mais próximas. Para fazer este mapeamento foi utilizada a ferramenta **Nominatim**¹³. Em seguida, fez-se necessário selecionar e identificar manualmente as categorias dos crimes de acordo com as descrições das rubricas, filtrando os crimes descritos que tivessem alguma relação mínima com o contexto de trânsito e mobilidade de pedestres e, conseqüentemente, de passageiros. Desta forma, o resultado final do dataset limpo conta com 12.448 registros criminais geolocalizados.

Apesar dos crimes já estarem limpos e com suas respectivas localizações, cada um deles de forma isolada não é capaz de nos dizer quais regiões concentram os maiores números de crimes e assim, conseqüentemente, podem ser consideradas menos seguras. Para isso, foi utilizado o conceito de clusterização. O algoritmo utilizado foi o *Density Based Spatial Clustering of Application with Noise* (DBSCAN)¹⁴. Este algoritmo é interessante pois constrói o agrupamento com base na proximidade dos pontos vizinhos que estão sendo avaliados, além de permitir escolher também um limiar mínimo de ocor-

¹⁰**OSMnx**: Biblioteca de criação e manipulação de grafos. <https://osmnx.readthedocs.io/en/stable/>

¹¹**NetworkX**: Biblioteca em Python para estudar grafos e redes. <https://networkx.org/>

¹²**OpenStreetMap**: Projeto de produção colaborativa de dados geo-espaciais abertos. <https://www.openstreetmap.org/>

¹³Esta é a interface de depuração para o mecanismo de pesquisa usado no site do OpenStreetMap. <https://nominatim.openstreetmap.org/>

¹⁴**DBSCAN**: Método clássico de agrupamento baseado em densidade. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

rências necessárias para que o *cluster* resultante seja válido. Com isso, o agrupamento gerado fica na posição original inicial, onde os pontos selecionados estão localizados. Ao aplicar o algoritmo sobre os dados das duas cidades, utilizando os parâmetros a seguir, que foram obtidos empiricamente, os resultados foram: para $v = 14$ e $\varepsilon = 85\text{m}$ em São Paulo, foram obtidos 2.028 *clusters* criminais, e com $v = 3$ e $\varepsilon = 250\text{m}$ para Curitiba, 360 *clusters* puderam ser identificados.

Entretanto, apesar da construção dos *clusters*, esses ainda não possuem o formato compatível de visualização e utilização necessária, que neste caso, equivale à polígonos. Desta forma, mais um passo é aplicado para que seja possível obtê-las. A conversão utilizada para gerar os polígonos é conhecida como **Convex Hull**¹⁵, que basicamente consiste em pegar um conjunto de pontos clusterizados e, através do menor número possível de pontos externos, gerar um polígono convexo que contenha todos os outros pontos dentro. Com isso, ao finalizar este passo, o número de polígonos disponíveis passa a ser 1.961 para São Paulo, e 136 para Curitiba.

3.2.2. Dados da malha rodoviária

Os dados de ônibus da cidade de São Paulo são representados através de dois arquivos. O primeiro com as linhas de ônibus (2.069), e seus respectivos pontos de paradas, ordenados. O segundo (complementar), possui os pontos de paradas e suas respectivas coordenadas. Desta forma, foi necessário apenas mapear os pontos de paradas para os vértices mais próximos do grafo utilizado para representar a malha viária da cidade. Para Curitiba, os dados possuíam um formato diferente. Apesar de também existir um arquivo com informações sobre as linhas de ônibus (676), esse não possuía dados sobre seus pontos de paradas. Com isso, foi necessário acessar os arquivos de pontos de paradas de cada linha, para então conseguir contruí-las. Ao final deste processo, foi observado que a cidade de São Paulo conta com 42 pontos de paradas por linha, em média, e Curitiba com 27.

3.2.3. Dados de relevo

Como os dados de relevo são disponibilizados através de quadrículas, que representam um recorte de uma dada área, foi necessário identificar apenas as regiões de interesse e utilizá-las. Assim como para os pontos de paradas, as informações de relevo também são representadas por coordenadas, junto do valor da altitude, em metros. Com isso, bastou identificar os vértices mais próximos de cada referência de relevo disponível e enriquecê-los com a altitude. Estes dados são essenciais para enriquecer os vértices das cidades, pois são capazes de possibilitar a identificação das inclinações das vias que os conectam. Desta forma, é possível determinar os níveis de aclive que estão contidos nas rotas construídas, e avaliar a viabilidade dessas de acordo com os critérios dos usuários.

3.3. Carga

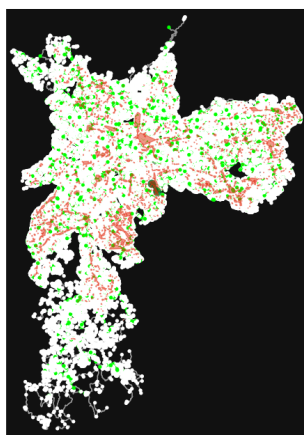
Apesar da biblioteca OSMnx disponibilizar uma forma de salvar a estrutura do grafo através de um arquivo *graphml*, semelhante ao padrão XML, também é possível extrair

¹⁵Convex Hull: A convex hull is the smallest convex polygon containing all the given points. <https://www.geeksforgeeks.org/convex-hull-using-divide-and-conquer-algorithm/>

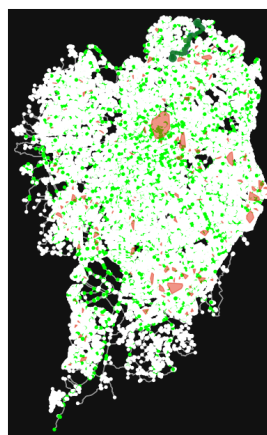
apenas os vértices do grafo e salvá-los como CSV, JSON, Parquet, dentre outras opções. Esses vértices contêm inúmeras informações úteis dentro de si, além de todas as informações extras acrescentadas ao longo do processo de transformação (Seção Subseção 3.2). Contudo, neste trabalho, utilizou-se o formato *graphml*, pois ele é capaz de serializar o grafo em seu formato original por meio de um único arquivo, que é de fácil desserialização, prático para gerenciar e replicar, uma vez que não requer configurações de ambientes de bancos de dados.

3.4. Dicionário de Dados

Para facilitar a organização dos dados, foi criado e disponibilizado um projeto no GitHub ¹⁶, que contém os arquivos principais, em *graphml*, para as cidades de São Paulo e Curitiba. Esse projeto conta também com arquivos auxiliares com exemplos de como configurar o ambiente, criar instâncias dos grafos da maneira correta, e de fato utilizar os dados enriquecidos. A Figura 2 conta com um exemplo visual de como estão os grafos, sendo os pontos de paradas sinalizados através de pontos verdes, e os polígonos criminais, em vermelho.



(a) Grafo da cidade de São Paulo.



(b) Grafo da cidade de Curitiba.

Figura 2. Grafo enriquecido das duas cidades.

Embora o grafo contenha informações tanto nas arestas quanto nos vértices, todas as informações adicionais usadas para enriquecê-lo foram armazenadas nos vértices. Apesar dessa escolha, é importante destacar que tanto os vértices quanto as arestas são representados de forma semelhante pela ferramenta OSMnx. Ambos possuem identificadores específicos, seguidos de dicionários de dados com as informações pertinentes a cada grupo. No caso das arestas, é possível encontrar informações como os vértices que as conectam, o nome da via, a velocidade máxima permitida, o comprimento, se a via é de mão única ou não, entre outras.

¹⁶Repositório do projeto NESPeD UFV - SBBB-DSW-2024: <https://github.com/NESPEDUFV/sbbd-dsw-2024>

Para os vértices, que receberam informações extras, é necessário um detalhamento adicional. Sendo assim, com o objetivo de facilitar a compreensão dessas informações, foi criado um dicionário de dados com legendas (Tabela 1), que conta com um exemplo de vértice com todas as informações adicionais possíveis.

Atributo / Valor	Significado
y: -23.57628	Latitude do vértice mapeado pela ferramenta OSMnx.
x: -46.69560	Longitude do vértice mapeado pela ferramenta OSMnx.
ponto_parada: True	Indicador se o vértice é do tipo embarque/desembarque.
latlong_original: (-23.57637, -46.69542)	Geolocalização original do ponto de parada.
distancia_vertice_mais_proximo: 21.29	Distância da geolocalização original para o vértice mapeado.
altitude_latlong_mais_proxima: (-23.5766, -46.6958)	Geolocalização da altitude mais próxima.
altitude_distancia_mais_proxima: 39.6545	Distância da geolocalização mais próxima de altitude.
altitude_metros_mais_proxima: 724.0	Altura de referência da altitude mais próxima.
dentro_poligonos: [{ "id_poligono": 600, "poligono_criminal": "POLYGON ((-46.695 -23.576, ...))", "cod_setor_GeoCenso_2010": 355030862000163, "poligono_setor": "POLYGON ((-46.694 -23.580, ...))", "classe_social": "4 - Classe Alta" }]	Lista de polígonos que contêm o vértice. Esses polígonos contam com sua forma geométrica, sua classe social e sua referência para o dado oficial do IBGE ¹⁷ .
dicionario_poligonos_mais_proximos: { "raio_85_vizinhos_14": [{ "id_poligono": 600, "distancia_m": 22.37, "dentro_poligono": true, "poligono_criminal": "POLYGON ((-46.695 -23.576, ...))", "cod_setor_GeoCenso_2010": 355030862000163, "poligono_setor": "POLYGON ((-46.694 -23.580, ...))", "classe_social": "4 - Classe Alta" }], ...] }	Lista com top M polígonos mais próximos do vértice, com base em uma configuração de clusterização ϵ e v . É importante para mapear os principais polígonos próximos, e contribuir na análise socioeconômica dos resultados, por possuir informações como classe social.

Tabela 1. Informações disponíveis nos vértices dos grafos.

4. Aplicações do Dataset

Por se tratar de um *dataset* rico em informações de contexto sobre as cidades, algumas alternativas de uso são possíveis.

Uma possibilidade de aplicação é o uso dos dados criminais e de relevo, combinados ou não, para gerar rotas para motoristas ou pedestres. Desta forma, se torna possível que esses usuários obtenham rotas de acordo com a sua necessidade, sejam elas evitando aclives acentuados para reduzir o esforço físico ou consumo de combustível do veículo; ou então desviando de regiões marcadas como inseguras, devido à incidência de *clusters* criminais. Além de também ser possível combinar as duas informações, caso seja coerente para o contexto aplicado.

Uma outra possibilidade de aplicação é no contexto de planejamento de soluções para as cidades. Como os *clusters* criminais estão distribuídos pela cidade, é possível fazer análises de onde estão os principais focos de ocorrência. Isso, associado às informações

¹⁷Instituto Brasileiro de Geografia e Estatística: <https://www.ibge.gov.br/>

de pontos de paradas, possibilita a identificação de regiões candidatas a um aumento no número de policiamento, a fim de aumentar a segurança daqueles passageiros que esperam seus ônibus nos pontos de paradas, por exemplo.

Recentemente, esses dados foram utilizados no trabalho [de Jesus Almeida et al. 2024] com o foco em rotas alternativas para veículos coletivos. As rotas alternativas permitem que coletivos como ônibus, por exemplo, se aproveitem de contextos criminais para evitar regiões inseguras, e que pedestres também se beneficiem disso, combinado com a questão de diminuição de esforço físico para chegar aos pontos de paradas.

5. Considerações Finais

Este trabalho apresentou Bus&City, um conjunto de dados importante para estudos sobre mobilidade urbana. Os dados, disponibilizados publicamente ¹⁸, contemplam as rotas de ônibus das cidades de São Paulo e Curitiba, enriquecidas com dados criminais e de relevo dessas cidades. Com esses dados, pesquisadores terão facilidade em desenvolver projetos voltados para mobilidade ciente de contexto e sistemas de transportes inteligentes. Além disso, gestores públicos poderão utilizar os dados para criar modelos para melhorar as rotas de ônibus das suas cidades.

Como trabalhos futuros, espera-se contemplar mais cidades do Brasil, além de enriquecer a base com outras fontes de dados, como informações contextuais sobre trânsito e fluxo de mobilidade dos pedestres. Isso permitirá eventualmente enriquecer o grafo com novos tipos de dados. Além disso, fazer uso de automatizações utilizando *Large Language Models* (LLMs) para classificar as categorias das rubricas dos crimes, pode ser uma abordagem interessante para facilitar a escalabilidade do trabalho para mais cidades.

Agradecimentos

Os autores agradecem pelo auxílio financeiro fornecido pela CAPES para a realização deste trabalho.

Referências

- Almeida, V., Silva, T., and Silva, F. (2022). Se for, vá na paz: Construindo rotas seguras para veículos coletivos urbanos. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 140–153, Porto Alegre, RS, Brazil. SBC.
- Aragão, J. (2006). Curso de gestão integrada da mobilidade urbana, módulo 01 - política nacional de mobilidade urbana. *Secretaria Nacional de Transporte e Mobilidade Urbana do Ministério das Cidades e Centro de Formação de Recursos Humanos em Transportes da Universidade de Brasília, Brasília, DF*.
- de Jesus Almeida, V. G., Silva, T. R. M. B., and Silva, F. A. (2024). Um método centrado em dados para melhorar a segurança de transportes coletivos municipais. In *Anais do XXXIV Simpósio Brasileiro de Bancos de Dados*, Porto Alegre, RS, Brazil. SBC.
- de Souza, A. and Villas, L. (2020). Vem tranquilo: Rotas eficientes baseado na dinâmica urbana futura com deep learning e computação de borda. In *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 351–364, Porto Alegre, RS, Brazil. SBC.

¹⁸<https://github.com/NESPEDUFV/sbbd-dsw-2024>

- Freitas, J. B., Clarindo, J. P., and Aguiar, C. (2023). Spsafe: um dataset sobre dados de criminalidade no estado de são paulo. In *Anais do V Dataset Showcase Workshop*, pages 48–57, Porto Alegre, RS, Brasil. SBC.
- Ladeira, L., Souza, A., Pereira, G., Silva, T. H., and Villas, L. (2019). Serviço de sugestão de rotas seguras para veículos. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 608–621, Porto Alegre, RS, Brasil. SBC.
- Magagnin, R. C. and da Silva, A. N. R. (2008). A percepção do especialista sobre o tema mobilidade urbana. *Transportes*, 16(1).
- Melo, R., Vasconcelos, F., Silva, R., Santos, P., Ramos, V., and Coutinho, F. (2023). Brbus - construindo um dataset para monitoramento geoespacial dos ônibus de cidades brasileiras. In *Anais do V Dataset Showcase Workshop*, pages 25–35, Porto Alegre, RS, Brasil. SBC.
- Santos, F. A., Rodrigues, D. O., Silva, T. H., Loureiro, A. A. F., and Villas, L. A. (2017). Rotas veiculares cientes de contexto: Arcabouço e análise usando dados oficiais e sensoriados por usuários sobre crimes. In *Anais do XXII Workshop de Gerência e Operação de Redes e Serviços*, Porto Alegre, RS, Brasil. SBC.
- Scabora, L., Spadon, G., Rodrigues, L. S., Cazzolato, M. T., Araújo, M. V. S., Sousa, E. P. M., Traina, A. J. M., Rodrigues-Jr, J. F., and Traina-Jr, C. (2019). G-franc: A dataset of criminal activities mapped as a complex network in a relational dbms. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2019 Companion*, pages 366–376, Fortaleza, CE, Brazil, October 7-10, 2019. SBC.
- Sá, B., Muller, G., Banni, M., Santos, W., Lage, M., Rosseti, I., Frota, Y., and de Oliveira, D. (2021). Polroute-ds: um dataset de dados criminais para geração de rotas de patrulhamento policial. In *Anais do III Dataset Showcase Workshop*, pages 117–127, Porto Alegre, RS, Brasil. SBC.
- Vahdat-Nejad, H., Ramazani, A., Mohammadi, T., and Mansoor, W. (2016). A survey on context-aware vehicular network applications. *Vehicular Communications*, 3:43–57.