Identifying urban occurrences for cities without open data using web scraping, local news and social media: a case study in Bauru, SP

Felipe Augusto Ferreira[1], **Higor Amario de Souza**[2]

[1] Department of Computing
São Paulo State University – Bauru, SP, Brazil

[2]Department of Computer and Digital Systems Engineering, Polytechnic School
University of São Paulo – São Paulo, SP, Brazil

fa.ferreira@unesp.br, higoramario@usp.br

Abstract. *Smart cities, driven by technologies and digital platforms, use data to improve the quality of life for citizens. The lack of data on urban occurrences impairs the identification of problems and the formulation of effective public policies. The city of Bauru, for example, faces challenges in adopting initiatives aimed at data-driven urban management. This work proposes a methodology that uses local information sources, collected through web scraping, to identify urban occurrences. These occurrences are categorized and geolocated, allowing for a spatiotemporal analysis using maps and charts. The methodology supports data-driven decision-making and can be applied to any city that has local information sources. The case study in Bauru points to a large number of occurrences related to infrastructure and crime in the city.*

1. Introduction

A Smart City is one that exploits operational data to improve its services for the citizens and to use its resources more efficiently [Harrison et al. 2010]. Nowadays, the Internet has a massive amount of data available on news websites and social networks, which often reflects the problems faced by people in their daily routines. Thus, it is possible to use the data and information available on the Internet for a wide variety of purposes. This data is entered through specialized communication channels and also spontaneously by any user [Laudon and Laudon 2022].

Many cities worldwide, especially in developing countries, still face the challenges of urbanization, where people move from rural to urban areas, straining the city's resources and services like public transportation, education, electricity, water, and more. According to the [United Nations 2025], nearly half of the world's population lives in large and medium-sized cities.

Ideally, cities should have access to data originating from their ICT infrastructure, manual collection, citizen complaints, and the technical work of their departments. Moreover, cities should make open data available to hold them accountable to their inhabitants. Cities should also offer digital services and information to simplify processes and encourage spontaneous collaboration to gather insights for their own benefit. Even when data is generated and managed, there are several problems regarding its quality, such as imprecision, ambiguity, or incompleteness [Sta 2017].

Since 2011, Brazil has had the Information Access Act, which guarantees all citizens access to non-sensitive public information from its organizations at all levels: federal, state, and municipal [Brasil 2011]. Cities like São Paulo (11.4 million inhabitants[1]) provide several open data services. An example is GeoSampa[2], which contains diverse geolocated data sources, such as road infrastructure, traffic signals, land use, and citizens' complaints about urban occurrences. The city of Rio de Janeiro (6.2 million inhabitants[3]) implemented de Center of Operations and Resilience (COR-Rio), which uses many data sources to monitor and prevent accidents, to provide data services for the citizens, and for urban planning [Taveira et al. 2026].

Medium and small cities in developing countries do not always have enough data to support their public policies. For example, Bauru is a medium-sized city (380 thousand inhabitants [4]) in the state of São Paulo, Brazil. Bauru faces significant urban challenges that directly affect the quality of life of its inhabitants and the sustainable development of its community, such as urban mobility [Magagnin and da Silva 2008], housing [Krause 2020], sanitation [Gulinelli 2016], and lack of geolocated data [Dias et al. 2015]. More recently, the city launched a portal to provide geolocated data[5], but until now, the only data available is for land parcels. Bauru lacks a centralized and accessible platform that brings together comprehensive open data on urban problems in the city [Ramos 2020]. This gap in the availability of critical information can create significant obstacles to effective urban planning, resource allocation, and rapid response to emerging issues. This lack of data is a common situation in many cities similar to Bauru. Another problem is that these cities are losing local news coverage due to the competition from social networks and the high concentration of some giant news conglomerations. This phenomenon is known as "news desert" [Abernathy 2020], which can make it even more difficult to source local issues from cities.

This study presents a methodology that employs Data Science techniques (Exploratory Data Analysis–EDA–and Natural Language Processing–NLP), web scraping, and Automated Web Testing to collect, analyze, and visualize data related to urban problems. To this end, we used incident reports from news websites and information from social networks as data sources. We present a case study of the methodology applied to the municipality of Bauru. Web scraping is a technique that allows the extraction of data directly from web pages [Mitchell 2018]. We used automated tests to automatically collect data from local news websites. We categorize this data and use Data Science techniques for analysis.

By creating a system to identify and monitor urban occurrences, the goal is to provide a tool that can be used by urban planners, researchers, and other stakeholders to better understand how occurrences are handled in the municipality. This not only allows for a deeper understanding of the challenges faced by the city but also assists in decision-making and, consequently, the implementation of more effective solutions in how the city addresses its urban challenges. This problem identification methodology was designed

[1] `https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama`
[2] `https://geosampa.prefeitura.sp.gov.br`
[3] `https://cidades.ibge.gov.br/brasil/rj/rio-de-janeiro/panorama`
[4] `https://cidades.ibge.gov.br/brasil/sp/bauru/panorama`
[5] `https://bauru.geopixel.com.br/geopixelcidades3/#/home`

to be replicated in other cities, especially those that do not collect or manage their data. However, it can also be used in cities that already have public open data to complement and increase their dataset availability and their decision-making processes.

In addition to collecting data on occurrences, we also provide a tool for visualizing and analyzing these events through maps and charts. Some examples of these occurrences are fallen trees, water shortages, floods, thefts, potholes, and others. Both the data and the tool are available as open source.

In what follows, Section 2 and 3 show the background and recent related studies. Section 4 describes our methodology and also our analysis tool. Section 5 presents the results of our case study in the city of Bauru, while Section 6 presents our conclusions.

2. Background

Urban occurrences are problems and incidents that happen in the urban environment and directly impact the lives of residents and the sustainable development of cities. These urban issues result from disordered growth and a lack of adequate planning, manifesting as social and structural problems that interfere with the quality of life of the population, compromising both current and future generations [Macedo et al. 2018]. Examples of occurrences are potholes in roads, damaged sidewalks, broken traffic lights, fallen trees, traffic accidents, lack of urban cleaning, noise pollution, and crime, and so on [Lee et al. 2025]. Several smart cities can monitor these occurrences, identifying such problems and acting to reduce impacts and improve urban well-being. We used various techniques to identify urban occurrences from news and posts available on the internet to build a dataset of occurrences, using the city of Bauru as a case study. Next, we describe the techniques that allowed us to browse, collect, identify, georeference, and analyze occurrences.

Natural Language Processing (NLP) is a set of techniques for making computational use of human language [Eisenstein 2018]. It is a multidisciplinary approach that uses knowledge from linguistics, statistics, machine learning, and other fields. It involves techniques and algorithms that enable machines to understand, interpret, and generate natural language. NLP methods allow for the automatic analysis of news articles or social media posts to extract relevant information–such as identifying mentions of a particular urban problem or detecting public sentiment on a given issue.

Exploratory Data Analysis (EDA) brings together a set of techniques designed to summarize and describe, in a quantitative way, the main characteristics of a dataset. Through it, statistical measures such as mean, median, mode, percentiles, dispersion, and visualizations (graphs, tables) are used to identify patterns, trends, and distributions in the collected data. The objective is to make a (possibly large) collection of observations easier for a brain to manage and understand [Páez and Boisjoly 2022, Tukey 1977].

Geolocation analysis (or geospatial analysis) involves mapping observable features in physical space, using georeferencing tools to spatially visualize data, whether points of interest (POI), socioeconomic, geographic, etc. Unlike a purely tabular analysis, geolocation analysis enables examination of data in the context of location, proximity, density, and movement, revealing spatial patterns that would not be evident otherwise. For example, by plotting the points of occurrences on a city map, one can perceive the concentration of certain types of problems in specific neighborhoods or regions. With

geolocation analysis, public managers and researchers can direct efforts more precisely, focusing on the areas of the city that most require attention.

Automated testing tools are libraries or frameworks that execute tests to verify functionalities and ensure the quality of a software [Vincenzi et al. 2018]. Automated tests perform inputs and actions on the software and then compare the results obtained with the expected results, flagging any failures. These tools allow for the repeated running of test suites quickly and reliably, reducing the incidence of human error and increasing the coverage of tested scenarios. A well-known example is Selenium, an open-source tool widely used to automate tests in web applications. It allows for the simulation of clicks, form filling, page navigation, and other user interactions across various browsers.

3. Related work

Many cities provide data for their inhabitants, which allows anyone interested in exploring to gather insights. The city of São Paulo has the GeoSampa portal, with hundreds of geolocated open data related to the city's infrastructure, socioeconomic, and geographical features. Similarly, the Seattle GeoData[6] is the city's official geospatial data portal, providing access to information including transportation, environmental, and public safety data. These portals facilitate the access and analysis of spatial data, supporting urban planning initiatives, academic research, and the development of applications that meet community needs. Other examples are Chicago's OpenGrid [7] and Dublin's Dublinked[8].

[Agonafir et al. 2022] evaluated 10 years of complaint reports of New York's citizens to predict variables that contribute most to the occurrence of street flooding. The prediction model is based on the Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis. The most related variables are precipitation, sewer back-up, and clogged catch basins. [Bolta and Hassani 2023] used the spatiotemporal characteristics of citizen complaints from New York to predict outliers, which indicate anomalies in complaint occurrences. [Eshleman and Yang 2014] compared citizen complaints from San Francisco with a Happiness Index based on sentiment analysis from Twitter data, showing that regions of the city with higher complaint report levels are positively correlated with higher happiness. [He et al. 2024] used citizen complaints data from Beijing to identify spatiotemporal patterns of public safety. [Jiao et al. 2024] applied sentiment analysis to citizen environmental complaints (e.g., noise, air, light) from Guangzhou to study how these complaints contributed to environmental management solutions. [Lee et al. 2025] studied data from a complaint web platform, showing that higher-income neighborhoods in the city of Albany are more likely to have their complaints resolved by the government.

[Anantharam et al. 2015] used events reported on Twitter collected over 4 months in San Francisco Bay Area to extract traffic events that occurred in the city. These events were geolocated and presented on a map. [Osorio-Arjona et al. 2021] used Twitter data to identify and geolocate problems in the Madrid Metro system, identifying patterns of problems regarding places and hours with more occurrences.

[D'Andrea et al. 2018] proposed a framework to create profiles of city areas based on web data from sites and services. These data sources include city events, local news,

[6] `https://data-seattlecitygis.opendata.arcgis.com/`
[7] `https://opengrid.chicago.gov/opengrid/`
[8] `https://data.smartdublin.ie/`

traffic information, and so on. Then, they extract points of interest (POI) spread over the city. City areas are split into cells, and the k-means clustering algorithm is used to group similar areas in the city of Milan, Italy. [Bondielli et al. 2020] explored the use of online news to create city profiles, using the city of Rome as a case study. To categorize the cities, they used NLP and Support Vector Machine (SVM). The online news data was obtained from a single website RomaToday[9], over 4 years using web scraping. They classified the data into 12 macro-categories, such as Crimes, Events, Missing people, Roads and Traffic, and so on. [Gao et al. 2017] used POI data from 10 populated areas of the United States of America (e.g., New York, Los Angeles, Dallas) and Foursquare's social network data. By combining these datasets, they find the functional characteristics of such regions. [Hong et al. 2023] proposed an approach to update POI databases using logistics delivery data in the cities of Beijing and Ziyang.

[Boeing and Waddell 2017] applied web scraping to analyze rental housing in the United States of America, using data from the rental website Craigslist. They produced a geolocated dataset of 1.5 million rental listings, which can be used by urban planners and others interested in housing issues. [Dongo et al. 2021] present a framework that can collect data from both web scraping and Twitter. They conclude that web scraping is faster, but the Twitter API is more flexible. Web scraping is also more sensitive to website changes.

The main related works use datasets from large cities, whether provided by the governments themselves or by companies operating in these cities. The existence of data that indirectly addresses local issues, such as the X (former Twitter), serves as an alternative to this lack of official data. As in the studies of [Osorio-Arjona et al. 2021] and [Anantharam et al. 2015], we use social media data to identify occurrences in the medium-sized city of Bauru. Additionally, we use web scraping to collect such occurrences from local news, as done by [Bondielli et al. 2020], despite distinct purposes. The idea is that our methodology can be applied in cities without open datasets to identify occurrences related to complaints. These occurrences are geolocalized as POI (as done by [Boeing and Waddell 2017] and [Gao et al. 2017]) to enable spatiotemporal analyses like those carried out in the works of [Agonafir et al. 2022], [Bolta and Hassani 2023],[He et al. 2024], and [Lee et al. 2025]. Finally, our methodology is available as an open source tool that can be adapted to other cities. The occurrences data from Bauru is also publicly available.

4. Methodology

The methodology proposed for our study includes defining the terms of interest related to common city occurrences. Then, we select possible data sources that contain these occurrences. We proceed by collecting the data, which needs to be treated and, finally, geoprocessed for use in analyses. Figure 1 shows the steps of our methodology.

Before data collection, we defined a list of keywords to be used in the search for urban occurrences. These keywords were chosen based on the content of news websites, related work, and citizen complaint services of São Paulo (156) and New York (311). We classified the keywords into 11 categories, which are presented in Table 1 – translated from Portuguese.

[9] `www.romatoday.it`

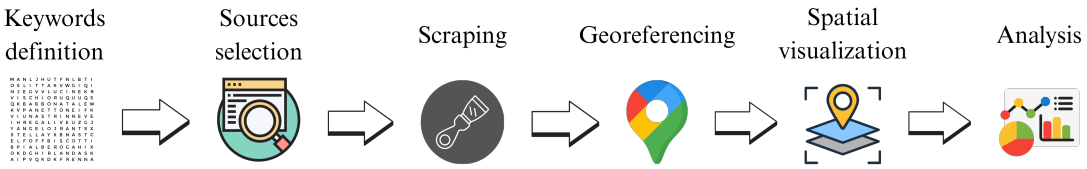

Figure 1. Steps of the proposed methodology

Table 1. Categories and keywords used to collect urban occurrences.

Category	**Keywords (before stemming)**
Climate	rain, hail, storm, wind, disaster, erosion, fall, collapse, bog, heat, cold, submerge, down
Complaints	justice, indictment, charge, report, expertise, warrant, illegal, attacked, attack, abandoned, discard, irregular, fraud, scarce, damage, failure, lack of, complaint, pane, breakdown, noise, vanish, havoc, prejudice, loss, harm
Crimes	agression, torture, stab, weapon, shot, stray bullet, attack, machete, knife, violence, crime, kidnapping, dismember, shock, criminal, offender, graffiti, victim, theft, thievery, assault, armed robbery, vandalism, vandal, gang, burglary, invasion, devious, coup, bandit, cell phone, faction, escape, penitentiary, operation, caught, arrest, police, blitz, prision, warrant, delegacy, security, camera, capture, investigation, fugitive, outlaw, jail, prisoner, inmate, jailbird, approach, denunciation, convict, suspect, riot, beat, rape, harassment, abuse, pedophilia, protective measure
Deaths and injuries	stabbed, homicide, femicide, assassin, murder, kill, iml, medical-legal institute, lethal, drown, die, dead, death, injured, injury, suicide, shot, wound, carbonize, body
Drugs and trafficking	narcotic, drug, marijuana, cocaine, crack, traffic, seizure, smuggling, contraband
Emergencies	urgency, risk, alert, emergency, relief, serious, severe
Environment	batalha river, snake, animal, mistreatment, sting, scorpion, polluted, bush, pollution, dust, smoke, fire, burn, venomous, fireman, fire-fighter
Health	hospitalized, samu, health, mosquito, infestation, dengue, zika, chikungunya, aedes aegypti, outbreak, sick, vaccine, covid, influenza, sus, hospital, upa, ubs, basic health unit, emergency care unit, pandemic, epidemic, proliferation, breeding place
Infrastructure	interruption, leak, major, supply, rotation, crater, flood, light, hole, pothole, sidewalk, traffic signal, post, tree, energy, water, sewer, manhole, asphalt, pavement, work, maintenance, repair, sanitation, rubble, rubish, trash, selective collection, reservoir, water main, well, pump, rationing, electric, wasteland, dae, cpfl, são paulo power and light company, ban, collection, break
Social problems	unemployment, homeless, hunger, precarious, neglect, resident, slum, community, housing, dwelling, vulnerable, racism, homophobia, transphobia, insult
Traffic	transport, traffic, airplane, pilot, vehicle, fall, emdurb, transit, accident, run over, collision, crash, overturn, driver, motorcyclist, truck, car, motorcycle, bus, pedestrian, cyclist, mobility, congestion, radar, signal, passenger

We searched for relevant data sources, including local and regional news websites, social media, APIs, and open data. We chose the news websites 94FM[10], Band[11], and G1[12], which are the main sources of local news in Bauru and the surrounding region. We also chose social media platform X[13], as it has an accessible API with posts about local events, including occurrences related to the categories defined in this study. We applied the NLP technique known as stemming to simplify and generalize keywords for

[10]https://www.94fm.com.br/
[11]https://www.band.com.br/band-multi/bauru-e-marilia
[12]https://g1.globo.com/sp/bauru-marilia/
[13]https://x.com/

searching the news and posts of our selected sources. We searched for news and posts between January 2023 and October 2025.

We collected the urban occurrences from the news websites using the Selenium testing framework[14], which automated the navigation on those websites to perform page scrolling and dynamic content loading. For the web scraping process, we used the Beautiful Soup[15] Python library. To collect data from X, we used the Twikit API[16]. We selected news and posts whose titles contained at least one of the defined keywords. Also, the posts must contain the word Bauru, since X is a global social media and the API does not report the location of the users.

We applied the ETL (Extract, Transform, Load) process to aggregate data from the distinct sources into a unified format. The data was processed to ensure its quality and usability. All news and posts were manually checked to remove duplicates (e.g., a same occurrence in different news), news not related to urban occurrences, and advertisements.

We used the Google Geocoding API[17] for the georeferencing process. This allowed us to relate each occurrence to a position on the map, enabling spatial analyses such as geographic distribution, identification of critical areas and territorial patterns.

The news reports and posts presented different levels of location detail: some provided the full address (e.g., street, commercial/public establishment), while others provided information about one or more neighborhoods where the occurrences took place. Thus, we made manual corrections to the georeferencing of occurrences in these cases to keep the finest grainer precision of a occurrence location. This step was necessary to ensure the accuracy and reliability of the data and to facilitate the generation of results through visualizations and statistics on the identified urban issues.

The data modeling was structured as follows. For data obtained from news websites, the following columns were created: title, subtitle, publication date, content, link, address types, coordinates, site, search terms, and addresses. The data obtained from the social network has the following fields: username, text, creation date, retweets, likes, and search term. Both datasets were stored in *CSV* files and are publicly available[18].

Based on this data from the city of Bauru, we used exploratory data analysis to identify which types of urban occurrences were most frequent and in which neighborhoods they were concentrated. We also conducted a temporal and spatial analysis of the incidents to understand their geographic distribution through the city over time. Although the analysis and data are from Bauru, this methodology can be applied to other cities and regions of interest. Finally, we created a web tool to analyze these occurrences, which we present in what follows.

4.1. Web tool for the analyses of the geolocated occurrences

We created a web application to enable the spatial analysis and visualization of occurrences, called Smart Bauru. The backend was built using the Flask[19] web microframe-

[14] `https://www.selenium.dev/`
[15] `https://beautiful-soup-4.readthedocs.io/en/latest/`
[16] `https://pypi.org/project/twikit/`
[17] `https://developers.google.com/maps/documentation/geocoding/overview`
[18] `https://zenodo.org/records/17566665`
[19] `https://flask.palletsprojects.com/en/stable/`

work to provide endpoints for data access and to manage HTTP requests between the client and the server. For data analysis, we used the Pandas[20] and Matplotlib[21] libraries, respectively, for data processing and generating charts with the results. On the frontend, we use the Leaflet.js[22] JavaScript library to generate maps with geospatial data rendering for visual analysis of urban occurrences in Bauru.

The Smart Bauru tool features a dashboard (Figure 2), visually and interactively consolidating the main indicators derived from the collected urban occurrences. This interface aims to facilitate a comprehensive understanding of the data and support decision-making by public managers, researchers, and citizens interested in the city's urban dynamics. In addition to general statistics, the dashboard offers dynamic filters by category, year, and neighborhood, allowing the user to adjust the view according to their interests. This interactivity facilitates specific analyses, such as identifying neighborhoods with increased occurrences during distinct periods or the predominance of certain types of occurrences in specific areas of the city. The web application is publicly available[23][24].

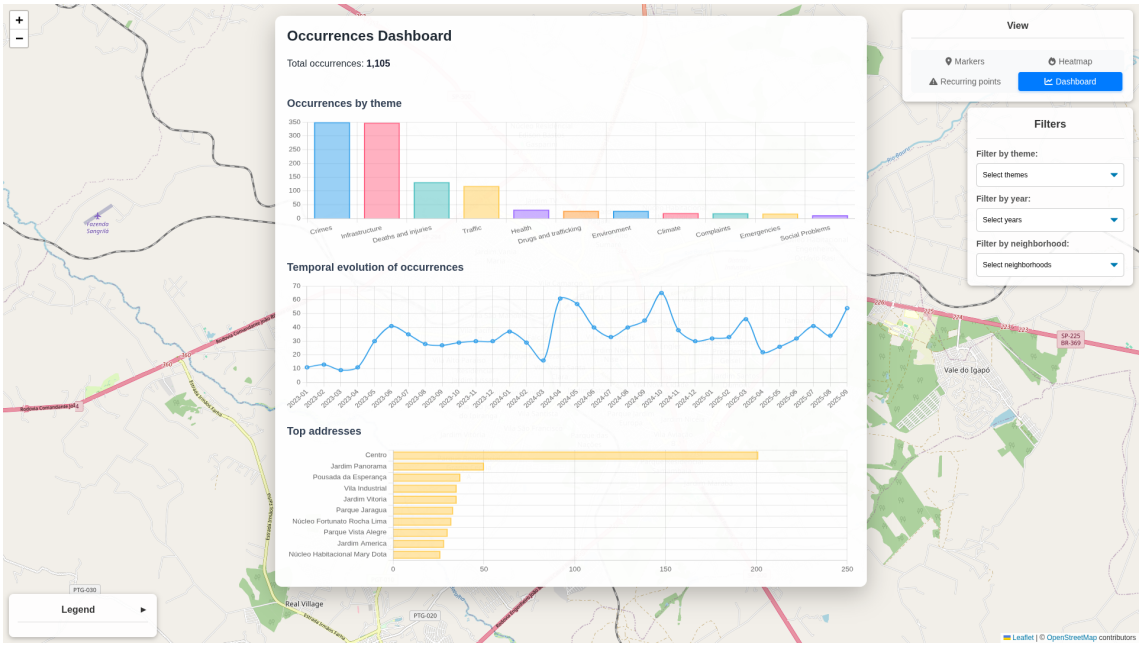

Figure 2. Smart Bauru's dashboard.

5. Results and analysis of the Bauru case study

In this section, we show several possible analyses for the data obtained from Bauru. The news dataset contains 2839 records, and the social network dataset has 184 records. Among the statistical analyses, we can see in Figure 3 the number of records from each of the news sources used. It is noted that the websites 94FM and G1 contributed the majority of the news that generated the geolocated occurrences. Figure 4 shows the distribution of

[20] https://pandas.pydata.org/
[21] https://matplotlib.org/
[22] https://leafletjs.com/
[23] https://zenodo.org/records/17566665
[24] https://github.com/f3l1pe-augusto/smart-bauru

occurrences by category, which shows a predominance of occurrences related to urban infrastructure. Figure 5 shows the number of occurrences per year, presenting its temporal evolution. This analysis allows us to observe patterns of increase or decrease in the number of records over time, highlighting how social, climatic, or structural factors can influence the frequency of reported issues. Regarding the imbalance between news and posts, the post content not always contain the users' location, which may have reduced occurrences from that source.

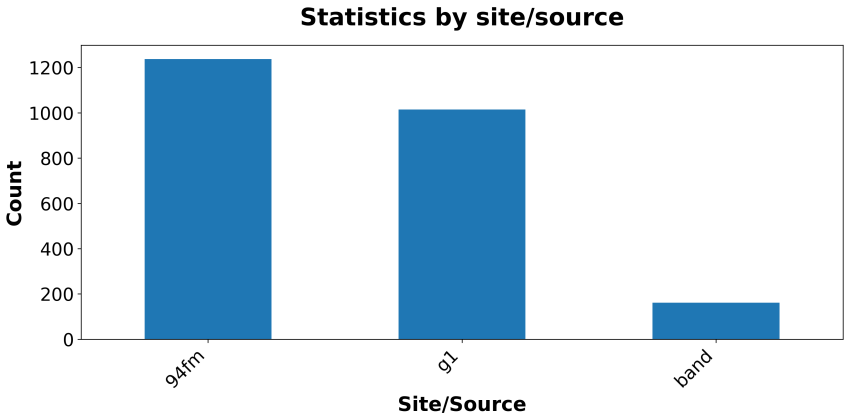

Figure 3. Number of occurrences per news website.

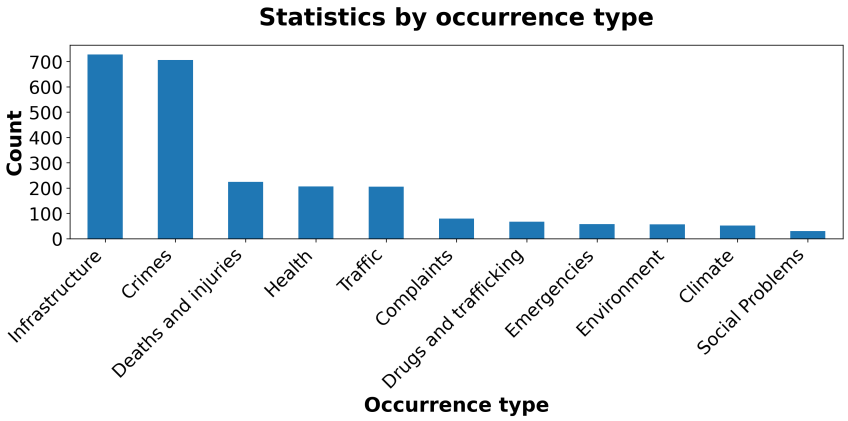

Figure 4. Number of occurrences per category.

The map in Figure 6 shows the spatial distribution of occurrences in Bauru. We can visualize these occurrences by categories. For example, there is a predominance of occurrences related to infrastructure and crime, spread throughout the city's urban area. Occurrences related to environmental issues are more frequent in the western and southern regions. Also, there is a concentration of traffic-related incidents in the downtown region.

We can refine the search by category and time period of interest. Clicking on a marker displays an informative pop-up with details about the occurrence, including title, category, date, exact address, source, and a link to the corresponding news article or post. Furthermore, it is possible to perform a joint analysis of specific occurrence types. In addition to the map with the markers, we can switch to a heatmap view, as shown in Figure 7, which shows the areas with the highest concentration of occurrences in hot red.

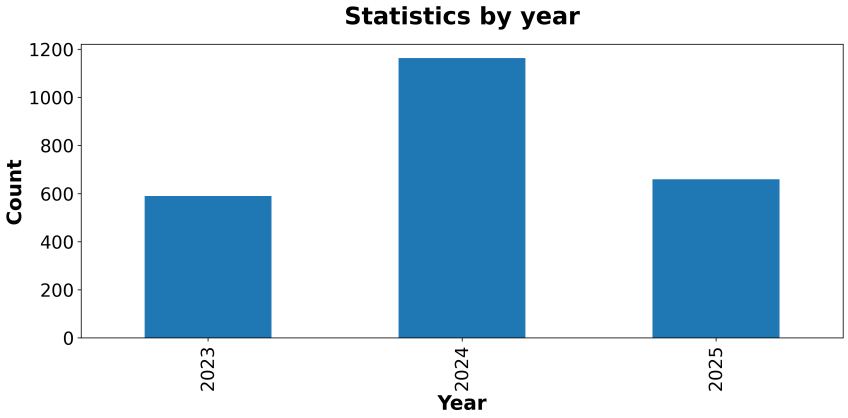

Figure 5. Number of occurrences per year.

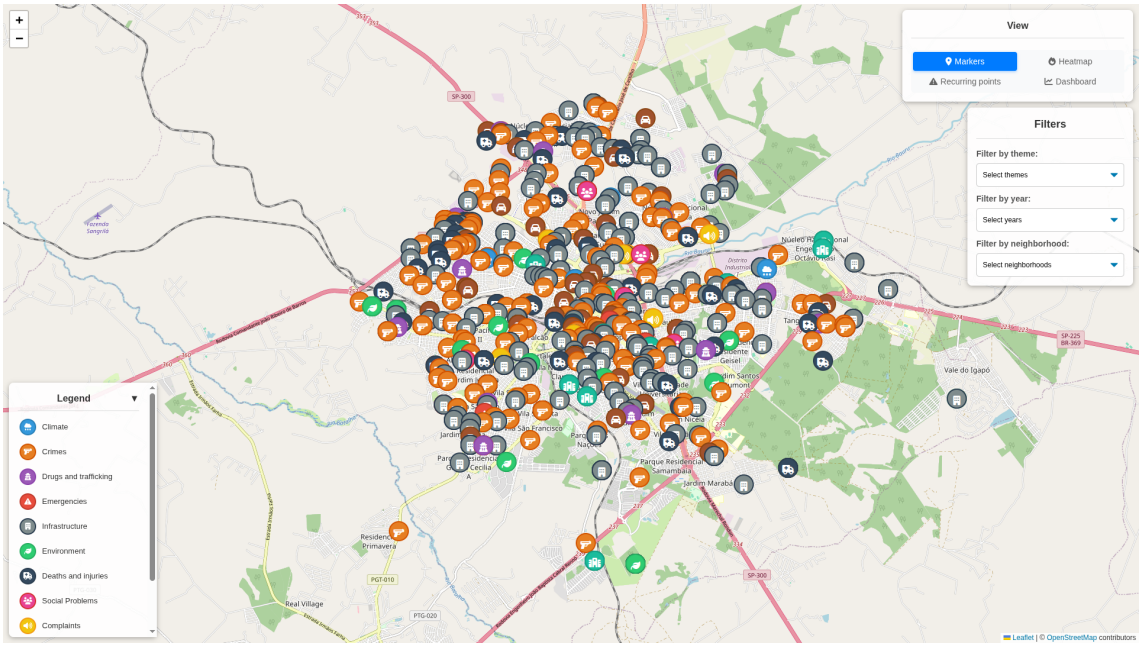

Figure 6. Map of occurrences in Bauru.

Another implemented analysis is the identification of recurring problems. Using a spatial clustering technique based on coordinate rounding, it is possible to identify locations where the same type of occurrence has been reported multiple times. Figure 8 presents the result of this analysis, in which markers highlight locations with a high recurrence of a specific type of occurrence. This map shows that infrastructure-related occurrences are the most frequent in various locations throughout the city. Clicking on a marker displays the most common address associated with the location, the type of occurrence, and the number of times it has been reported.

Figure 9 presents a comparison between some of the main neighborhoods of Bauru and the most frequent categories of occurrences among them. This analysis allows us to identify areas of the city that concentrate a greater number of records, revealing spatial patterns associated with certain types of urban problems.

Several other analyses are possible through the maps and charts implemented by

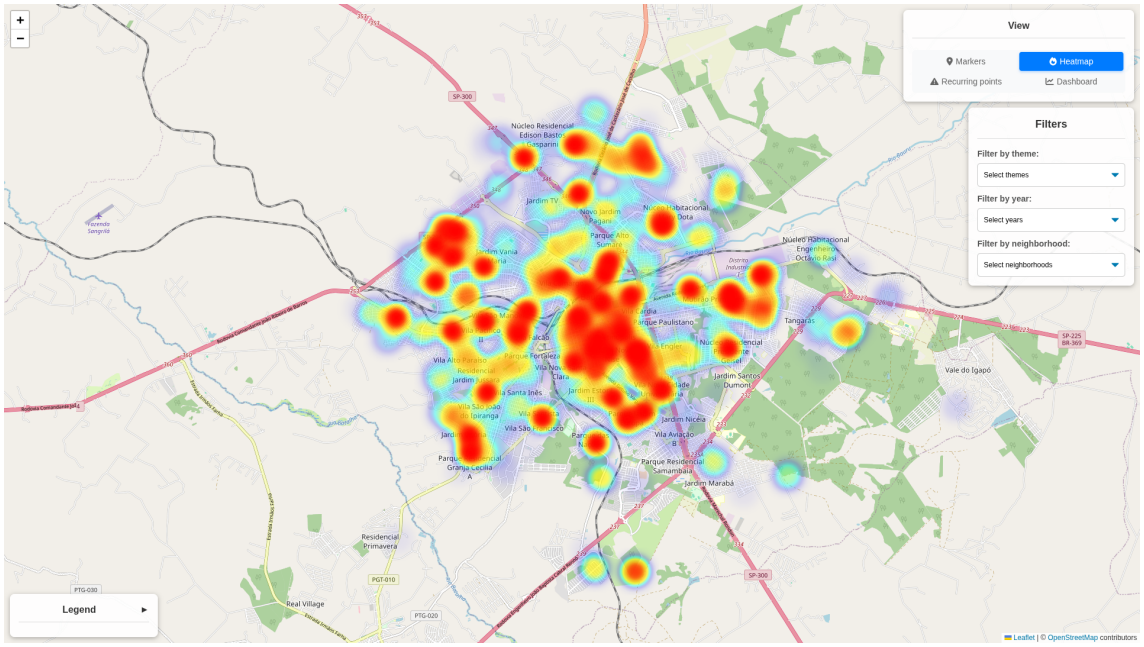

Figure 7. Heatmap of occurrences in Bauru.

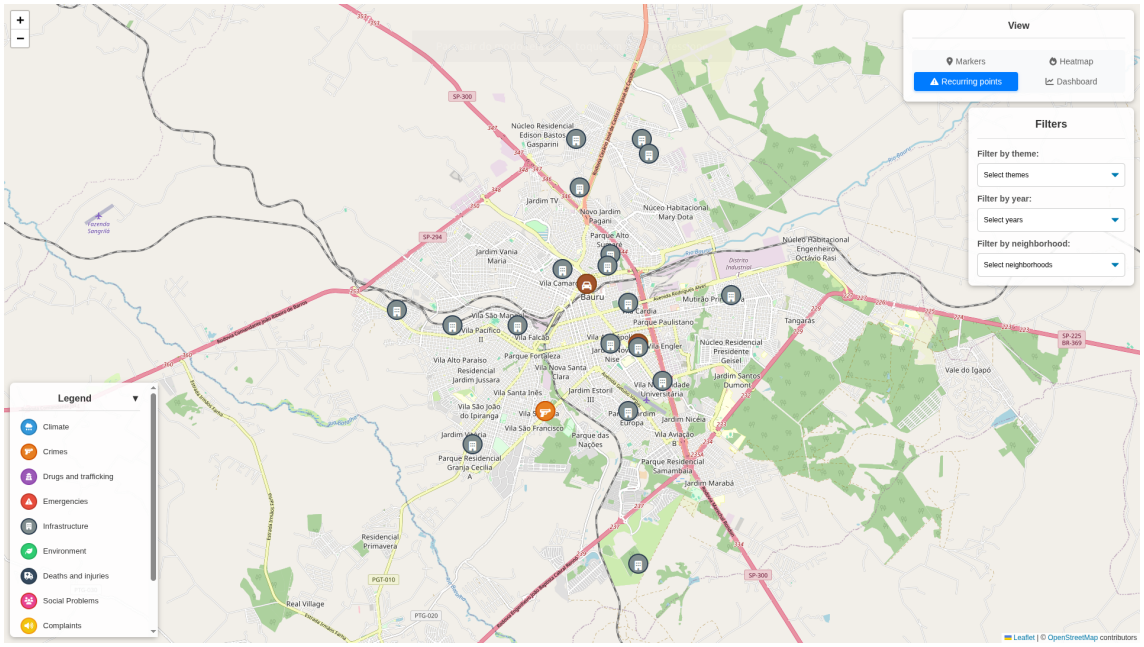

Figure 8. Map of recurrences in Bauru.

our methodology, reinforcing the tool's role as an instrument to support evidence-based decision-making. It may be replicated for other cities and may be helpful to understand and plan solutions for the reported problem, especially for those cities without data.

6. Conclusions

This study proposed a methodology for identifying urban occurrences using web scraping by collecting information from local news and social media posts. This methodology was implemented as a tool for mapping and analyzing urban occurrences in the municipality

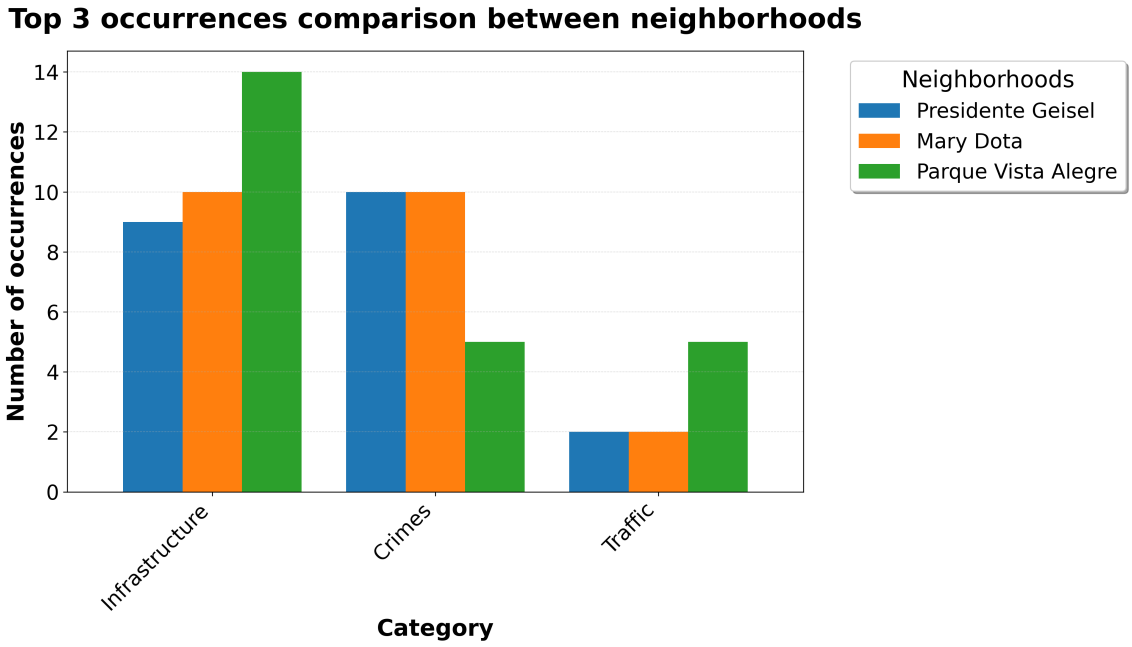

Figure 9. Comparison of occurrences between neighborhoods in Bauru.

of Bauru, allowing for an interactive visualization of spatial data related to city problems.

The methodology is adaptable and can be applied to any city or region that has local information available online, whether from news websites, social networks, or other internet sources. This type of solution is especially useful for cities that do not have or do not make available data on their urban problems. It is important to emphasize that data collection must respect copyright and the right to privacy. Furthermore, care must be taken to use reliable sources to avoid spreading misinformation. The expected gains from implementing the tool include increased transparency and efficiency in urban management, support for smarter allocation of public resources, and, consequently, an improvement in the quality of life of the population affected by the reported occurrences.

As future work, we intend to extend the application to receive continuous data updates and incorporate new sources of information. Another future work is to collect and analyze data in other cities and regions to enable broader comparative studies on urban problems, making this data openly available to any interested person, thus consolidating the methodology as a tool to support data-driven urban management.

References

Abernathy, P. M. (2020). Will local news survive? News deserts and ghost newspapers. Technical report, University of North Carolina.

Agonafir, C., Pabon, A. R., Lakhankar, T., Khanbilvardi, R., and Devineni, N. (2022). Understanding New York City street flooding through 311 complaints. *Journal of Hydrology*, 605:127300.

Anantharam, P., Barnaghi, P., Thirunarayan, K., and Sheth, A. (2015). Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology*, 6(4).

Boeing, G. and Waddell, P. (2017). New insights into rental housing markets across the united states: Web scraping and analyzing craigslist rental listings. *Journal of Planning Education and Research*, 37(4):457–476.

Bolta, V. and Hassani, M. (2023). Using human mobility patterns to forecast outliers in citizen complaints data. In *Proceedings of the 2023 IEEE International Conference on Big Data*, BigData 2023, pages 5166–5175.

Bondielli, A., Ducange, P., and Marcelloni, F. (2020). Exploiting categorization of online news for profiling city areas. In *Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems*, EAIS 2020, pages 1–8.

Brasil (2011). Lei n^o 12.527, de 18 de novembro de 2011 – institui a Lei de Acesso à Informação.

D'Andrea, E., Ducange, P., Loffreno, D., Marcelloni, F., and Zaccone, T. (2018). Smart profiling of city areas based on web data. In *Proceedings of the 2018 IEEE International Conference on Smart Computing*, SMARTCOMP 2018, pages 226–233.

Dias, R. S., Cioni, J. C., Kaiser, I. M., Peixoto, A. S. P., and Manzato, G. G. (2015). Cadastramento de informações urbanas do município de Bauru-SP utilizando sistemas de informação geográfica. In *Anais do 8^o Congresso de Extensão Universitária da UNESP*. Universidade Estadual Paulista.

Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y., and Barrios, S. (2021). Web scraping versus twitter api: A comparison for a credibility analysis. In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*, iiWAS '20, page 263–273, New York, NY, USA.

Eisenstein, J. (2018). *Introduction to Natural Language Processing (NLP)*. MIT Press.

Eshleman, R. and Yang, H. (2014). "hey 311, come clean my street!": A spatio-temporal sentiment analysis of twitter data and 311 civil complaints. In *Proceedings of the 2014 IEEE 14th International Conference on Big Data and Cloud Computing*, BDCloud 2014, pages 477–484.

Gao, S., Janowicz, K., and Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467.

Gulinelli, É. L. (2016). O saneamento e as águas de Bauru: uma perspectiva histórica (1896-1940). Master's thesis, Universidade Estadual Paulista (Unesp).

Harrison, C., Eckman, B., Hamilton, R., Hartswick, P., Kalagnanam, J., Paraszczak, J., and Williams, P. (2010). Foundations for smarter cities. *IBM Journal of Research and Development*, 54(4):1–16.

He, J., Zhang, W., and Yang, M. (2024). The spatial and temporal characteristics of urban public safety under the residents' complaints: Evidence from 12345 data in beijing, china. *Journal of Urban Management*, 13(2):217–231.

Hong, Z., Wang, H., Lyu, W., Wang, H., Liu, Y., Wang, G., He, T., and Zhang, D. (2023). Urban-scale poi updating with crowd intelligence. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4631–4638.

Jiao, Y., Li, C., Yao, Z., Weng, C., Lian, A., and Dong, R. (2024). How can online citizen complaints provide solutions to refine environmental management: A spatio-temporal perspective. *Information Processing & Management*, 61(2):103611.

Krause, A. B. P. (2020). Intervenções públicas em ocupações irregulares: um estudo de caso sobre a Favela Ferradura na cidade de Bauru. *Projectare: Revista de Arquitetura e Urbanismo*, 1(10).

Laudon, K. C. and Laudon, J. P. (2022). *Management information systems: Managing the digital firm – 17th ed.* Pearson Education.

Lee, W., Gross, K. J., Yong, C., Chelmis, C., and Zois, D.-S. (2025). Who reaps the benefits of smart management of neighborhood complaints? Impact of online participatory forums on neighborhood equity. *Cities*, 158:105716.

Macedo, E. T. d., Salles, M. C. T., Nunes, E. R., Martins, M. d. F., and Ribeiro, R. O. (2018). Problemas urbanos que interferem na (in) sustentabilidade de cidades: um estudo no município de Serra Redonda - PB. *Revista Brasileira de Planejamento e Desenvolvimento*, 7(3):1–24.

Magagnin, R. C. and da Silva, A. N. R. (2008). Reflexos da dependência do transporte motorizado individual em cidades brasileiras de médio porte: a questão da mobilidade no município de Bauru. *Olhares sobre Bauru*, 1:159–170.

Mitchell, R. (2018). *Web Scraping with Python*. O'Reilly Media, Sebastopol, CA.

Osorio-Arjona, J., Horak, J., Svoboda, R., and García-Ruíz, Y. (2021). Social media semantic perceptions on madrid metro system: Using twitter data to link complaints to space. *Sustainable Cities and Society*, 64:102530.

Páez, A. and Boisjoly, G. (2022). *Exploratory Data Analysis*, pages 25–64. Springer International Publishing, Cham.

Ramos, F. J. d. C. (2020). Indicadores socioeconômicos locais para a cidade de Bauru: um diagnóstico sob a ótica da competência em informação e midiática. Master's thesis, Universidade Estadual Paulista (Unesp).

Sta, H. B. (2017). Quality and the efficiency of data in "smart-cities". *Future Generation Computer Systems*, 74:409–416.

Taveira, M. F., Mariano, R. d. A., Trinta, P. Q., Bento, M. B. C., and Rocha, C. (2026). Resiliência urbana em grandes eventos: a atuação do COR-Rio no G20. *Revista De Administração Pública*, 60:e2025–0579.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison Wesley, 1 edition.

United Nations (2025). World urbanization prospects 2025: Summary of results. UN DESA/POP/2025/TR/ NO. 12, Department of Economic and Social Affairs – Population Division, New York.

Vincenzi, A. M. R., Delamaro, M. E., Dias Neto, A. C., Fabbri, S. C. P. F., Jino, M., and Maldonado, J. C. (2018). *Automatização de teste de software com ferramentas de software livre*. Elsevier.