



Rastreia Saúde: A Spatiotemporal Disease Tracking System through Open Unstructured Data and GIS

Rastreia Saúde: Um Sistema de Rastreamento Espaço-temporal de Doenças através de Dados Abertos Não Estruturados e SIG

Luiz Henrique Anjos Cardim¹ and Nádia Puchalski Kozievitch²

¹ Universidade Tecnológica Federal do Paraná, Departamento de Informática, Curitiba, Brasil. luizcardim@alunos.utfpr.edu.br.

ORCID: <https://orcid.org/0000-0002-2255-3680>

² Universidade Tecnológica Federal do Paraná, Departamento de Informática, Curitiba, Brasil. nadiap@utfpr.edu.br.

ORCID: <https://orcid.org/0000-0003-2286-9623>

Recebido: 03.2021 | Aceito: 05.2021

Abstract: Automated disease tracking has become an increasingly important tool today. This article describes the prototype of a disease tracking system for the Brazilian territory, preliminarily tested at the state level, in Paraná, and at the municipal level, in Curitiba. This study aims to extract and present relevant information in the health segment from unstructured data, extracted from news portals. The system generates data that allows analysis at different levels of granularity, from small municipalities to the national level. The results of the study shows the viability of the system and allows the authors to identify some patterns in the processed data.

Keywords: GIS. Spatial Database. Unstructured Data. Disease Tracking.

Resumo: O rastreamento automatizado de doenças tem se tornado uma ferramenta cada vez mais importante nos dias atuais. Esse artigo descreve o protótipo de um sistema de rastreamento de doenças para o território brasileiro, testado preliminarmente em nível estadual, no estado do Paraná e em nível municipal, na cidade de Curitiba. O objetivo do estudo é extrair e apresentar informações relevantes através de dados não estruturados, obtidos de portais de notícias. O sistema gera dados que permitem a análise em diferentes níveis de granularidade, desde municípios pequenos até a escala nacional. Os resultados do estudo demonstram a viabilidade do sistema e permitiram aos autores identificarem alguns padrões nos dados processados.

Palavras-chave: SIG. Banco de Dados Espacial. Dados não Estruturados. Rastreamento de Doenças.

1 INTRODUCTION

According to the World Health Organization (WHO)¹, today's cities are facing a triple health burden: infectious diseases (such as pneumonia, dengue, HIV/AIDS, tuberculosis, pneumonia); non-communicable diseases (such as heart disease, stroke, asthma) and other respiratory illnesses, cancers, diabetes and depression; and violence and injuries, including road traffic injuries. Better integration of health information systems is widely believed to be a pre-condition for more effective systems and information sharing. And better integration includes better opportunities for linking data between different registries, integration of relevant data sources in a central database or platform, integration of information on health and social care, and opportunities to integrate data at the personal level. Within the challenges and barriers, we can mention the data availability, data standards, linking data from different databases, legislation and ICT infrastructure, e-health applications, and governance and cooperation².

¹ <https://www.who.int/health-topics/urban-health>

² <https://www.euro.who.int/en/publications/abstracts/promoting-better-integration-of-health-information-systems-best->

If crises such as the COVID-19 pandemic are considered, data combined with context and meaning turns into knowledge for informing public health response. The impact of this pandemic has even changed the way data is shared (RDA COVID-19 Working Groups, 2020), along with guidelines to urban planners toward topics such as open data, open research, standards, and trustworthy data repositories (Fonseca et al., 2020).

In parallel, several initiatives provide open health data to support planning and policy-making (such as Open Cities Project³ And European Data Portal⁴) to support primary health care and local services⁵, open data from public-private partnerships⁶, open data from researchers⁷, open unstructured data (like news portals). Since the approval of the Brazilian Law in Information Access in 2011⁸, public agencies are making available their data through transparency portals (e.g., Curitiba with Instituto de Planejamento de Curitiba (IPPUC)⁹ and Curitiba Open Data Portal¹⁰).

However, despite all efforts in the segment, the Brazilian territory still lacks adequate tools for such finality, since the diseases relevant to Brazil can be different from those of other countries (KINDHAUSER; MARY KAY; WORLD HEALTH ORGANIZATION, 2003) and, in addition, most disease tracking platforms are based primarily on the English language. We also emphasize that, by limiting the tracking of diseases to the territory of only one country, it is also possible to achieve a greater degree of data granularity, also covering small and medium-sized municipalities.

The study presents an alternative to fill this gap, with a prototype for a disease tracking system. The system performs the collection and processing of data in an automated way through news portals, linking the processed information with spatial data from Brazilian municipalities. The first version of this paper (CARDIM; KOZIEVITCH, 2020) (responsible for presenting the processing of unstructured data), was now extended to include an automated web prototype for visualization (Rastreia Saúde), along with the findings/challenges of processing and integrating unstructured data. The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the prototype. Section 4 presents the results. And finally, Section 5 contains the conclusions of the study.

2 RELATED WORK

Among the works in this research segment, one of the pioneers is the alert system through e-mails from ProMed-Mail¹¹ (MADOFF, 2004). This system, which continues to be widely used today, has become the source of data for several other disease tracking platforms developed later. In its flow, before the news received on the portal is published, it is checked by specialists, which makes the platform a reliable and recognized source of data in the disease tracking segment. The WHO also maintains an alert portal for the emergence of new communicable diseases¹² with a similar model to ProMed-Mail, however with a much lower update frequency.

Another reference study in the segment is the work of Freifeld et al. (2008) that presents the initial architecture of the HealthMap¹³ platform, one of the first disease tracking platforms based on unstructured data. HealthMap extracts alerts on communicable diseases on a global basis by extracting data from several sources, including news sites and also ProMed reports.

[practices-and-challenges](#)

³ <https://www.worldbank.org/en/region/sar/publication/planning-open-cities-mapping-project>

⁴ <https://www.europeandataportal.eu/en/highlights/open-health-data-european-data-portal>

⁵ https://www.euro.who.int/_data/assets/pdf_file/0003/376833/almaty-acclamation-mayors-eng.pdf

⁶ <https://repositoriodatasharingfapesp.uspdigital.usp.br/>

⁷ <https://dataverse.harvard.edu/>

⁸ http://www.planalto.gov.br/ccivil_03_ato2011\2014/2011/lei/l12527.htm

⁹ <http://ippuc.org.br/>

¹⁰ <https://www.curitiba.pr.gov.br/dadosabertos/>

¹¹ <https://promedmail.org/>

¹² <https://www.who.int/csr/don/en/>

¹³ <https://www.healthmap.org/>

A similar approach is used in the study by Lan et al. (2012), which presents the STEWARD¹⁴ platform. However, in this system, only ProMED-Mail records are used, organizing them in the dimensions of space and time. According to the authors, using only the ProMed database can limit the dynamics of identifying new outbreaks, but it also reduces noise in the data presented because it is a more reliable source of information. In a subsequent study, Lan et al. (2014) presented the Newsstand¹⁵ platform, a system that can track different types of subjects, including health-related news in the dimensions of space and time.

Some studies tried to track the location and timing of diseases using even more dynamic methods for data extraction, like Social Media data. Among them, the study of Jayawardhana and Gorsevski (2019), which tries to track the location and timing of disease occurrence (flu) using data from Twitter.

Another example is the study of Sankaranarayanan et al. (2009) which processes Tweets by identifying whether they are news and also which news segment they belong to. The platform also offers a web interface for consulting processed data.

Another approach, used by Chunara et al. (2013), was to extract data through crowdsourcing, in a platform called flu near you¹⁶, that provides a form for users to self-report symptoms of respiratory diseases, such as fever, cough, shortness of breath, among others. The platform also allows data visualization through a web view of the maps.

We also cite the Brazilian studies InfoDengue (CODECO et al., 2018) and MonitoraCovid-19 (FIOCRUZ, 2021). InfoDengue is the result of a joint effort by several segments and several research institutions. The system uses multiple data sources and a probabilistic model to estimate, in a semi-automated way, dengue cases in more than two thousand Brazilian cities. MonitoraCovid-19 is a Fiocruz project that uses various data sources to generate spatiotemporal visualizations on various aspects of Sars-CoV-2, such as the number of deaths and vaccination data, and others.

Among the review studies in the segment, Choi et al. (2016) carried out a systematic review of the main disease tracking systems and studies related to them. The study presents the differences between the main platforms and their strengths and weaknesses. The authors also highlight the importance of these systems and the need for countries with a shortage of them to seek to implement them.

Mohanty et al. (2019) present a review of the disease tracking applications available for Android and IOS platforms. The study concluded that there is great potential in this segment, especially for solutions that serve health professionals and public health authorities.

We also cite studies in related areas or support of disease tracking, such as the study of Castro and JR. (2018), which describes the prototype of a tool to index textual and geographical information in a combined way. For textual indexing, the study used Natural Language Processing (NLP) techniques such as removing stop words and ranking through the Inverse Document Frequency (IDF).

Considering the public health data from Curitiba, several studies can be mentioned (DE OLIVEIRA et al., 2018; CAVALCANTE et al., 2018; LIMA et al., 2019). The study of de Oliveira et al. (2018) presented a characterization of Paraguay's public health data and using the information about the city of Asuncion a comparison was made with Curitiba's public health data. Cavalcante et al. (2018) surveyed the citizens of Curitiba to list the most important features in a health app. The study also presented a prototype of the application's screens for the features most required in the survey. In the study of Lima et al. (2019), open data of Curitiba public health is aggregated with transportation data, analyzing the accessibility to Curitiba public health units via public transportation.

3 THE PROTOTYPE RASTREIA SAÚDE

The prototype architecture (Figure 1) shows four main parts of the system: the front-end, the back-end API, the database, and the data extractor. In the "Data extractor architecture" section we present the relationship between the "data extractor" and "database" modules. In the "Visualization" section we present the relationship

¹⁴ <http://steward.umiacs.umd.edu>

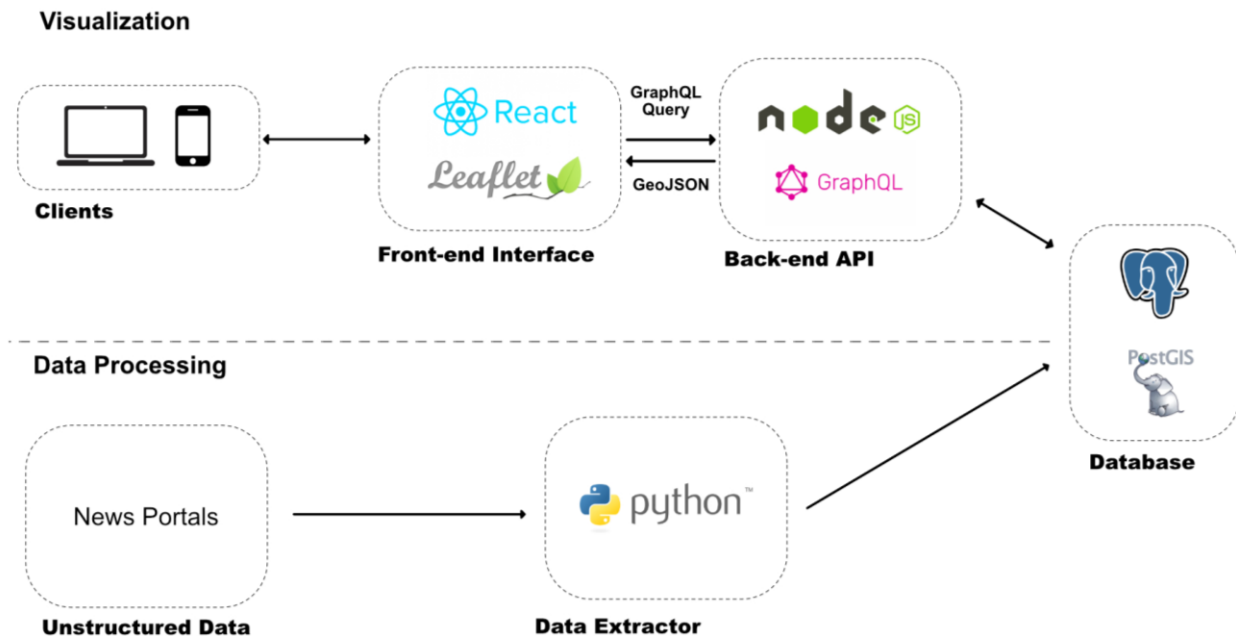
¹⁵ <http://newsstand.umiacs.umd.edu/web/>

¹⁶ <https://flunearyou.org/>

between the “front-end”, “back-end API” and “database” modules. The modules were grouped in this way to separate the two major contributions of this study, which are the processing of unstructured data and visualization.

Among the software used in the project, in the front-end we use ReactJS¹⁷ (version 17) and Leaflet¹⁸ (version 1.7), two Javascript libraries, to create the web interface. In the back-end API, we use Node¹⁹ (version 14.15) with Apollo Server²⁰ (version 2.20) to create a GraphQL²¹ endpoint. In the data extractor, we use the Python²² language (version 3.8.5) with the Scrapy framework²³ (version 2.3.0) to perform the data crawling. For NLP we use the Spacy²⁴ library (version 2.3.2) configured for the Portuguese language. The database used was PostgreSQL²⁵ (version 12.3) with the Postgis²⁶ extension (version 2.5).

Figure 1 – The system architecture of Rastrea Saúde.



Source: The authors (2021).

3.1 Data Extractor Architecture

As described in the study of Lan et al. (2012), the extraction of correct places from unstructured documents is a challenging task, which evolves processing data through a pipeline, that breaks the data cleaning and formatting into several subsequent steps. The general architecture of the data extractor developed in our study is shown in Figure 2.

The system has three sources of data:

- a) HTML from the news portals extracted through web crawling;
- b) the shapefile of all the Brazilian municipalities and other information like the population of

¹⁷ <https://reactjs.org/>

¹⁸ <https://leafletjs.com/>

¹⁹ <https://nodejs.org/>

²⁰ <https://www.apollographql.com/docs/apollo-server/>

²¹ <https://graphql.org/>

²² <https://www.python.org/>

²³ <https://scrapy.org/>

²⁴ <https://spacy.io/>

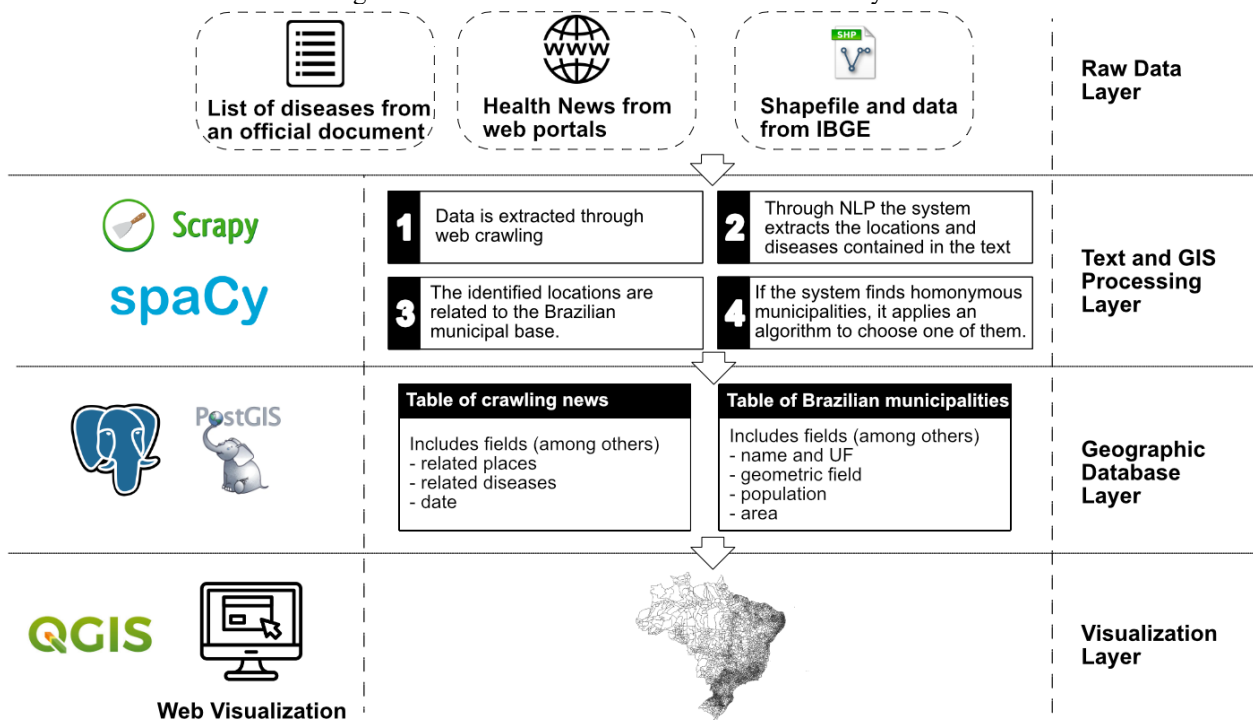
²⁵ <https://www.postgresql.org/>

²⁶ <https://postgis.net/>

each city, extracted from the IBGE²⁷;

c) the list of infectious and parasitic diseases (Ministério da Saúde, 2010), obtained from the SUS²⁸ and manually inputted into the system.

Figure 2 – The architecture of the data extraction system.



Source: The authors (2021).

To our analysis, we use the state of Paraná as a reference. The state is located in the south of Brazil and has the 5th largest population in the country, estimated in 2020 at 11,516,840 inhabitants according to the IBGE²⁹. Paraná has 399 municipalities, and its three most populous cities are Curitiba (the capital), Londrina, and Maringá.

We used data from five news portals of the state: The SESA (Health secretary of the government of Paraná) Portal³⁰, Bandab³¹, Bem Paraná³², AEN³³ (the official news agency of Paraná government) and Tribuna do Paraná³⁴, with the data range from January 1, 2019, to August 20, 2020.

The choice of what diseases to track was based on data published by the Ministry of Health of Brazil (MINISTÉRIO DA SAÚDE, 2010). All the diseases listed on this document were searched, except rabies. The reason for this exception is that rabies in Portuguese is called raiva and the word raiva also means angry in Portuguese, a very common word that could generate a lot of noise in the extracted data.

The complete list of tracked diseases is: aids, amebíase, ancilostomíase, ascaridíase, botulismo, brucelose, cancro mole, candidíase, coccidiodomicose, cólera, coqueluche, criptococose, criptosporidíase, dengue, difteria, doença de chagas, doença de lyme, diarréia, doença meningocócica, donovanose, enterobíase, escabiose, esquistossomose mansônica, estrogiloidíase, febre amarela, febre maculosa brasileira, febre purpúrica brasileira, febre tifóide, filariase por wuchereria bancrofti, giardiase, gonorreia, hanseníase, hantavírose, hepatite a, hepatite b, hepatite c, hepatite d, hepatite e, herpes, histoplasmose, HPV, influenza,

²⁷ Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics)

²⁸ Sistema Único de Saúde (The Brazilian Universal Health Care System)

²⁹ <https://cidades.ibge.gov.br/brasil/pr/panorama>

³⁰ <http://www.saude.pr.gov.br/>

³¹ <https://www.bandab.com.br/>

³² <https://www.bemparana.com.br/>

³³ <http://www.aen.pr.gov.br/>

³⁴ <https://www.tribunapr.com.br/>

leishmaniose, leptospirose, linfogranuloma venéreo, malária, meningite, mononucleose, oncocercose, paracoccidiodomicose, parotidite, peste, poliomelite, psitacose, rubéola, sarampo, shigelose, sífilis, cisticercose, tétano, toxoplasmose, tracoma, tuberculose and varicela.

3.1.1 PROCESSING THE UNSTRUCTURED DATA

The processing of unstructured data (HTML obtained from news portals) had the following steps:

- a) first, the system searches for the term bairro (neighborhood) or bairros (neighborhoods). If it finds, the processing of the current page is interrupted, as the system may confuse neighborhood names with city names;
- b) the system extracts the date of publication (or last update) of the news;
- c) the system searches in the text content (body of news and title) all occurrences of the diseases being tracked, using regular expressions. This comparison is not case-sensitive, but it does respect the accentuation of terms. We also emphasize that this list includes the formal names of disease and the popular names for which it is known. For example, for Influenza, we also look for the term gripe (flu);
- d) diseases found with different terms are grouped into only one key term. For example, coronavirus, SARS-CoV2, COVID, and COVID-19 are grouped into COVID-19;
- e) if the system finds any diseases, then it does an NLP to identify the locations contained in the text;
- f) the system removes the names of cities that can generate too much noise in the data, such as the municipality of Saúde (health) in the state of Bahia;
- h) the locations found in the previous steps are searched at the base of Brazilian municipalities, previously extracted from IBGE;
- i) if there are homonymous municipalities, the system selects only one of them, following two rules: first, it checks whether the publication portal is in the same state as any of the identified municipalities (proximity rule, like in the study of Lan et al. (2014)). If the first strategy is not met, it then selects the municipality with the largest population;
- j) finally, if the system has identified at least one disease and one location in the news, it adds the record to the database.

Despite reaching a level of municipal granularity for the Brazilian territory, our approach has some limitations: one is that the name of some municipalities is easily confused with other localities or terms in the Portuguese language (step f of data processing). Thus, the only data available for these cases is the “smoothed rate” (Eq. 2) view, which considers the numbers obtained in all neighboring municipalities.









Regarding granularity, the system only reaches the municipal level. It would be possible to reach a level of neighborhoods, but the process would have to be adjusted, using neighborhood tables instead of the cities table, other sources of news, etc.









Another limitation is about the types of news processed by the system. To measure the relevance of a disease to a given region, not all health news is of equal importance. As an example, we put a news item reporting the inauguration of a health unit that will treat multiple diseases. In this case, the system will extract each of the diseases mentioned in that news and relate to the municipality where the health unit will be inaugurated, generating noise in the data.

3.1.2 TABLES OF THE SYSTEM

As shown in Figure 2, this prototype uses two tables (Figure 3), one containing the Brazilian municipalities and the other with the data tracked through the news. The table of Brazilian municipalities contains 5572 records, which includes all the Brazilian municipalities. The table of tracked news has a total of 2263 rows. The description of each of the fields in these two tables is presented below:

Figure 3 - The tables used by this prototype.

crawling.municipios		
 id	<i>integer</i>	« pk »
 geom	<i>public.geometry</i>	
 cd_mun	<i>character varying(7)</i>	
 nm_mun	<i>character varying(60)</i>	
 sigla_uf	<i>character varying(2)</i>	
 area_km2	<i>double precision</i>	
 populacao	<i>integer</i>	
 municipios_pkey	<i>constraint</i>	« pk »

crawling.crawling_news		
 url	<i>character varying(500)</i>	« pk »
 idades	<i>jsonb</i>	« nn »
 titulo	<i>character varying(500)</i>	« nn »
 doencas	<i>character varying(50)[]</i>	
 data	<i>date</i>	
 tipo	<i>character varying(20)</i>	
 tipo_predicted	<i>character varying(20)</i>	
 crawling_news_pkey1	<i>constraint</i>	« pk »

Source: The authors (2021).

The fields of the table municipios (municipalities):

- a) Id – The identification of a municipality (primary key).
- b) geom – The geometry representation of the borders of the municipality.
- c) cdmun – A numerical code for the municipality, used by IBGE.
- d) nmmun – The name of the municipality.
- e) siglauf – The abbreviation of the state to which the municipality belongs.
- f) areakm2 – The total area of the municipality in square kilometers.
- g) populacao – The total population of the municipality.

The fields of the table crawlingnews:

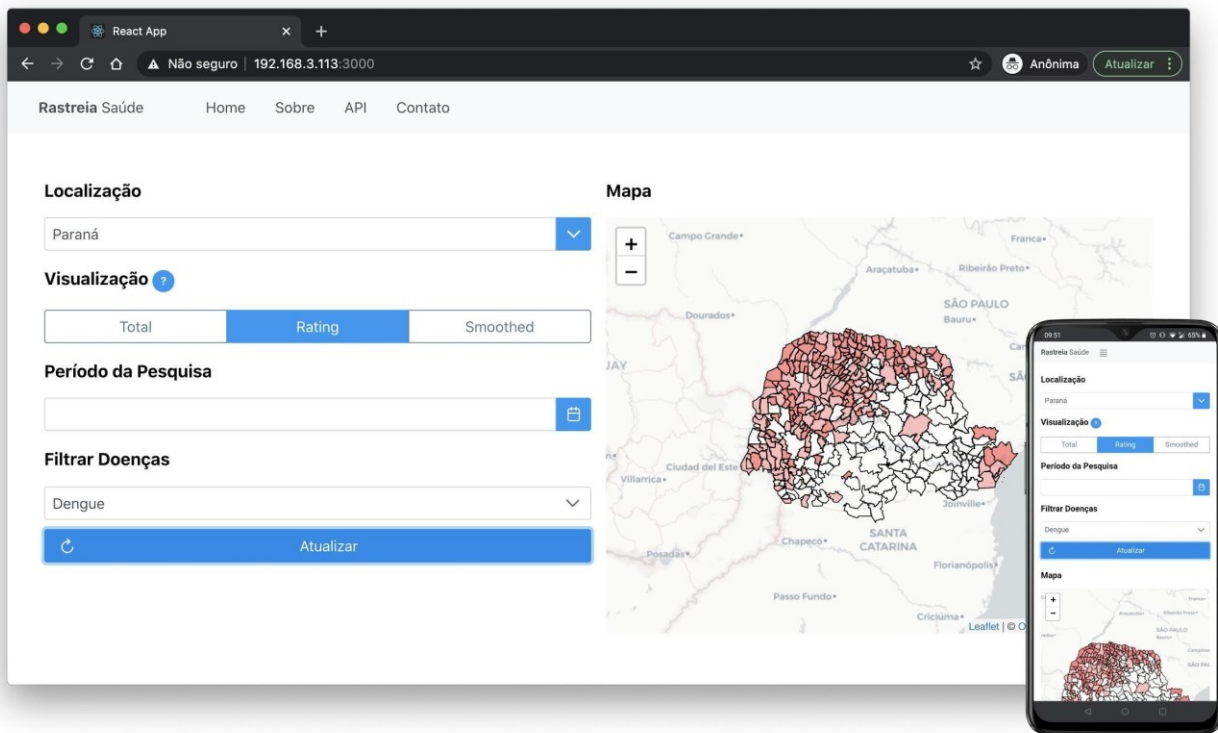
- a) url – The url of the processed news.
- b) cidades – List of cities identified in the news. In addition to the name of the municipality, it also records the name of the state.
- c) titulo – The title of the processed news.
- d) doencas – The list of diseases identified in the news.
- e) data – The date of the news.
- f) tipo – It allows the system administrator to label the type of news, such as notification of a disease case, vaccination campaign, etc. This field will allow the creation of a data set for machine learning in future studies.
- g) tipopredicted – In this field, the news type predictions made by the ML algorithm will be recorded.

3.2 Visualization

In this section, we describe the web visualization created for the prototype presented in this study. Figure 4 presents the main screen of the prototype, with a horizontal menu at the top, a set of filters on the left, and the map on the right. JavaScript language was used for the development of the system visualization due to the rich ecosystem of this language for this purpose. NPM, one of its package managers, is the world’s largest software registry (NPM, 2021) and most of these packages are published as open-source software.

ReactJS was used to create the user interfaces, along with Leaflet (to create the maps) and PrimeReact (UI components and to make the interface responsive). We developed a responsive interface because it fits both mobile devices and desktops. According to Statcounter (2021), smartphones represented 51.81% of internet access in Brazil in February 2021, against 47.34% of desktop devices.

Figure 4 - The main screen of Rastreia Saúde in desktop and mobile devices.



Source: The authors (2021).

The filters in the main screen have the following functionalities:

- a) **Localização** – It allows you to limit your search to just one state or to search the entire Brazilian territory;
- b) **Visualização** – It allows to classify the map based on the total numbers, classification, or smoothed classification. The difference between these classifications is presented later in this article;
- c) **Período da Pesquisa** – It allows to limit the data range of the displayed data;
- d) **Filtrar Doenças** – It allows limiting the data presented to just one disease. If the user performs a search without filtering a particular disease, the system will also return a list with the total news identified for each disease (Figure 5), based on the other selected filters.

Figure 5 - The list of identified diseases for a place.

Filtrar Doenças

Nenhuma doença selecionada ▼

↻ **Atualizar**

Doenças Identificadas

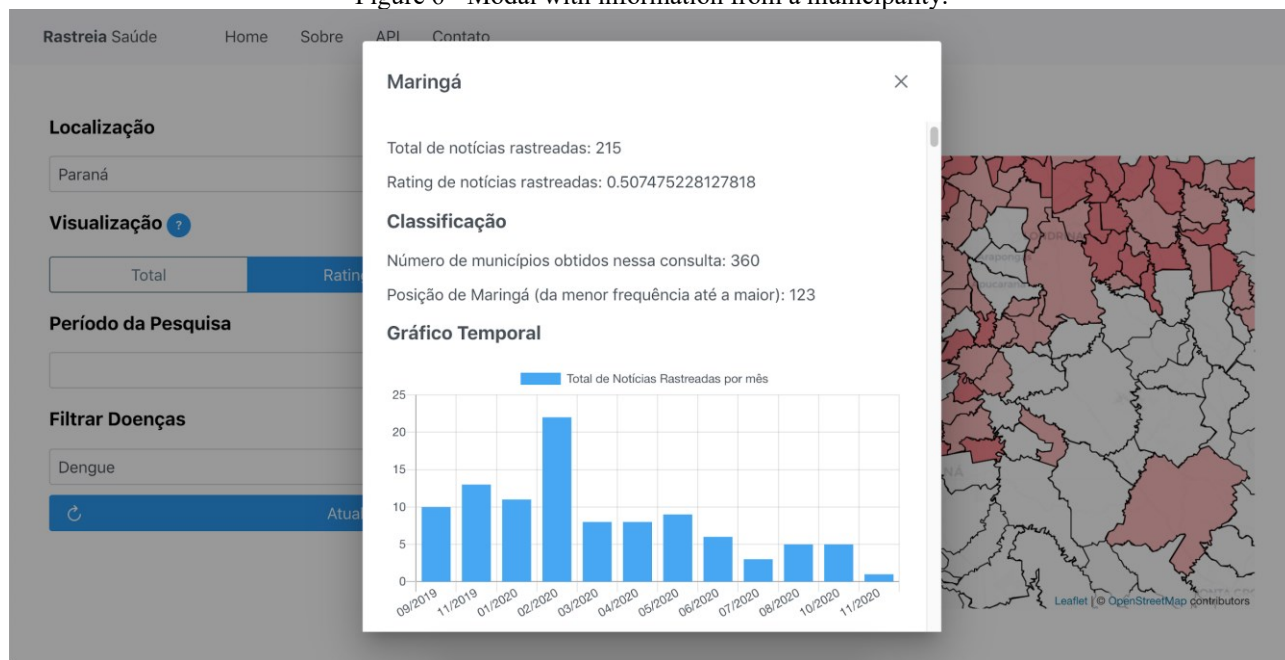
COVID-19 1072	Dengue 271	Sarampo 177	Febre Amarela 161	Influenza 104
Rubeola 54	HIV 46	Meningite 36	Vericela 34	Tuberculose 33

Source: The authors (2021).

It is also possible to perform a temporal analysis of the data. Clicking on one of the municipalities opens a modal with complementary information (Figure 6). This modal allows the user to check, among

others, the total news tracked for the selected municipality, based on the selected filters. It also features a graph with the last 12 months tracked, as well as the complete list of tracked news.

Figure 6 - Modal with information from a municipality.

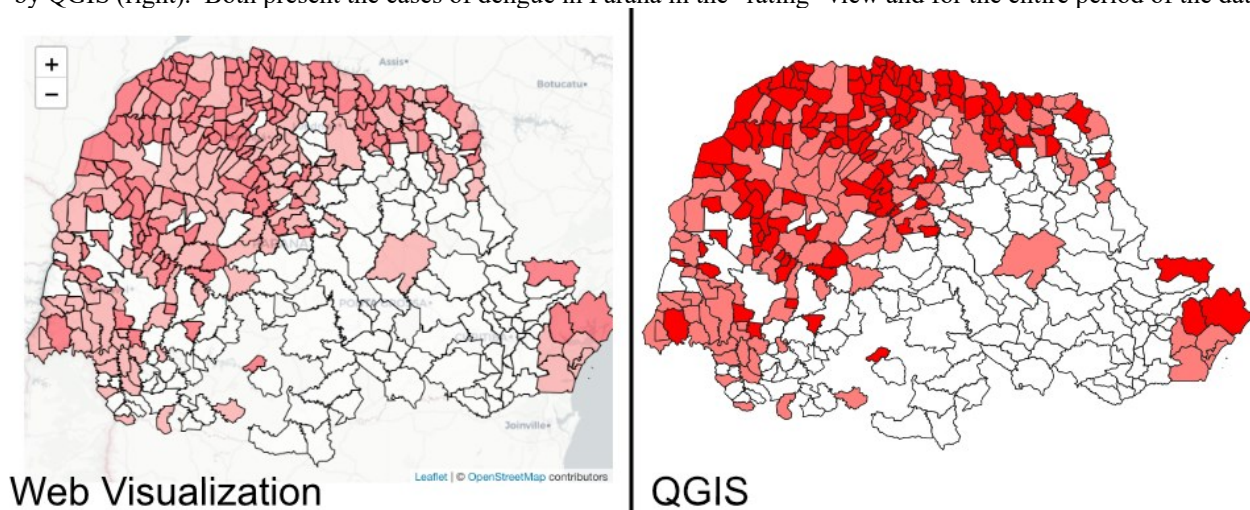


Source: The authors (2021).

4 RESULTS AND DISCUSSION

In this section, graphs and maps are presented to check the data obtained through the data extractor. Here QGIS was used to generate the maps in order to validate the map generated by the prototype interface, as shown in Figure 7. Note that the figures are similar, and the user does not need to know the advanced features of any application to visualize the data.

Figure 7 - Comparison between the map generated by the prototype interface of this study (left) and the map generated by QGIS (right). Both present the cases of dengue in Paraná in the “rating” view and for the entire period of the data.



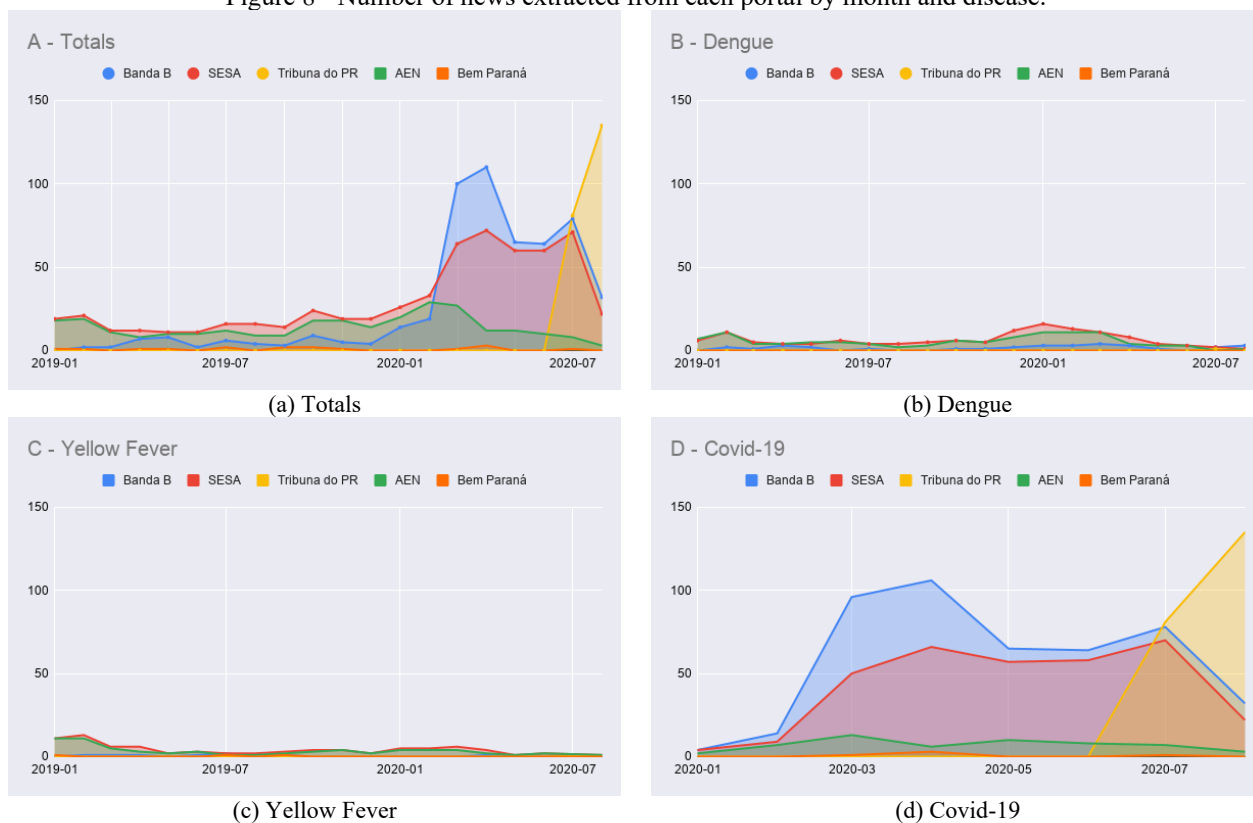
Source: The authors (2021).

4.1 Analyzing data on a state scale

The distribution in time of the extracted news is presented in Figure 8 for 3 diseases (Dengue, Yellow Fever, and COVID-19), and also the totals. Note that the Tribuna do Paraná portal does not keep a long historical period of its news, so in the last extraction of our algorithm, it obtained data only from the last 2

months of this portal. Another important point is the correlation between the number of news extracted on the SESA and the AEN portal. The reason is that both portals are managed by the same organization, the Paraná government.

Figure 8 - Number of news extracted from each portal by month and disease.



Source: The authors (2021).

Figure 8 presents an increase in the total number of health-news since about February, although news related to Dengue and Yellow Fever, two highly relevant diseases in the state of Paraná in the last few months, has been declining. The reason is the high frequency of news related to COVID-19, which, as we can see in Table 1, has almost a greater number of news published than all other diseases combined.

Table 1 - The top 10 diseases by the number of news.

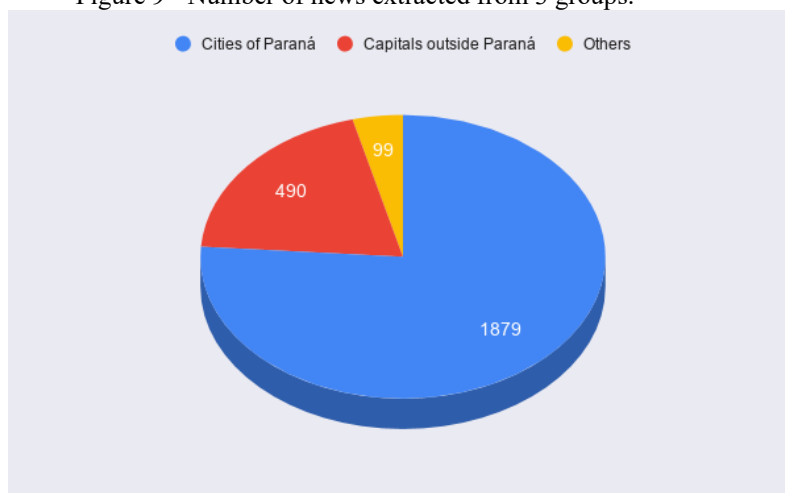
Disease	Related news	Percent
COVID-19	1072	49.69
Dengue	271	12.56
Measles	177	8.20
Yellow Fever	161	7.46
Influenza	104	4.82
Rubella	54	2.50
HIV	46	2.13
Meningitis	36	1.66
Varicella	34	1.62
Tuberculosis	33	1.57
Others	134	6.21

Source: The authors (2021).

The other tracked diseases for Paraná are: syphilis, leptospirosis, hepatitis b, whooping cough, leprosy, polio, malaria, diphtheria, tetanus, leishmaniasis, brucellosis, hepatitis c, Chagas disease, herpes, HPV, cysticercosis, hepatitis e, hepatitis a, giardiasis, typhoid and cholera.

To check the extracted data, we also selected the cities identified in states other than the news source (in this case we extracted the news identified in municipalities outside Paraná), and, from this selection, we separated the news related to capitals. Thus, the news identified was segmented into three groups: news within the state of Paraná, news from capitals outside Paraná, and others. The totals for each of these three groups are shown in Figure 9. We separated the data in this way because the five portals used in this research are located in the state of Paraná, and, therefore, most of the news published by them is related to cities in that state. As this study aims to present a prototype to track diseases at the national level, we added this step to identify outliers and possible noise generators in data from other states. However, the results presented in this study are not affected by this checking step.

Figure 9 - Number of news extracted from 3 groups.



Source: The authors (2021).

On the group of “others”, a manual check on the data was carried out. In this group, there are cities with great potential to generate false alerts due to having names with famous international places (like the city of Colômbia in the state of São Paulo or the city of Tailândia in the state of Pará) or with very common words in the Portuguese language (like the city of Central (Central) in the state of Bahia or the city of Campanha(Campaign) in the state of Minas Gerais. In these cases, the name of these cities was added as an exception to be ignored by the system pipeline (like described in step 5 of processing).

We then generated choropleth maps (Figure 10) for the two diseases with the highest number of news in the observed period, trying to identify the regions with higher relevance for each one of them. For each map, we present three versions:

- a) A version based on the total number of news of a given disease-related to each city.
- b) A version based on a raw ratio per thousand (rrpt), obtained using the Eq. (1).
- c) A version based on a smoothed rate per thousand (srpt), obtained using Eq. (2).

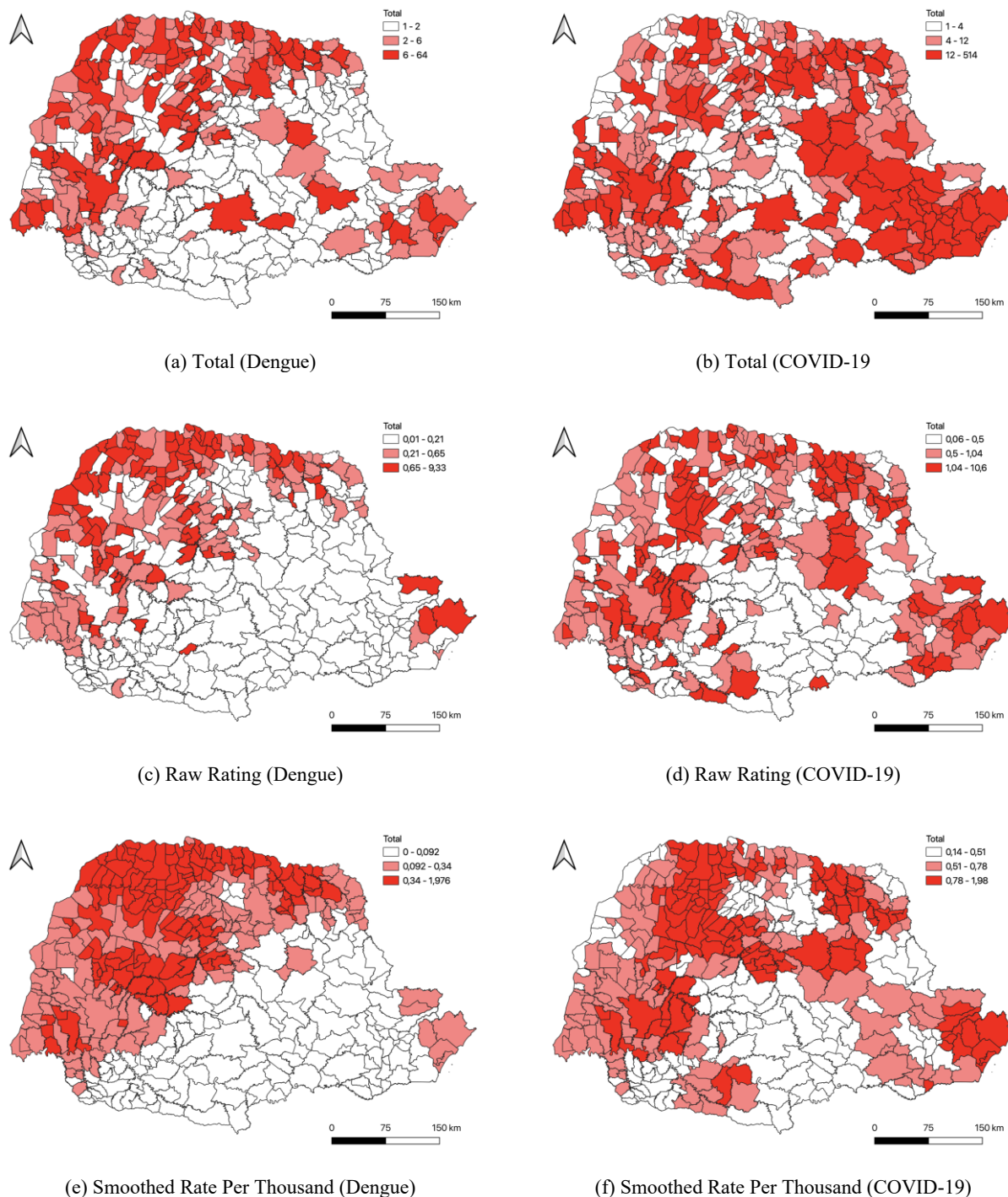
$$rrpt = \frac{T}{P} * 1000 \tag{1}$$

$$srpt = \sum_{i=1}^n \frac{T(i)}{P(i)} * 1000 \tag{2}$$

Where:

- n – is the number of bordering neighbors (including the city itself).
- T – is the total of news related to some disease for the city.
- P – is the population of the city.

Figure 10 - The distribution of the identified diseases on the map.



Source: The authors (2021).

Through the raw rate per thousand (Eq. 1), we seek to reduce the bias of the system alerts for larger cities and highlight the alerts of small and medium cities, considering population density. According to (Anselin et al., 2006) raw rate serves as an estimate for an underlying risk, i.e., the probability for a particular event to occur. The smoothed rate per thousand (Eq. 2) tries to balance the data and thus allows the identification of patterns by regions. According to Anselin et al. (2006) smoothed rates tend to empathize broad trends.

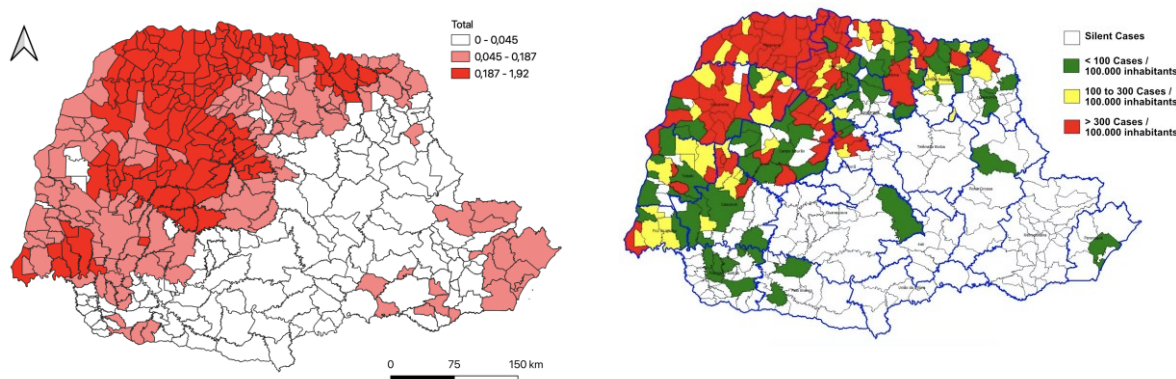
Lastly, we chose to classify the map with a quantile scale because, as Brewer and Pickle (2002) demonstrated in their study, this is one of the most efficient scales in the presentation of choropleth maps, being suitable for several different types of epidemiological data.

Figure 10 shows complementary visualizations of the diseases. For example, to identify the regions

most affected by dengue, Figure 10-E showed a better-defined pattern, while for COVID-19 the map in Figure 10-B was closer to reality. The reason for this is because COVID-19 has an equivalent publication frequency in almost all regions of the state, making it difficult to identify patterns by region.

For validation purposes, we also compared a map generated from the data extracted by our system with a map for the dissemination of official Dengue data (SESA, 2021) in the state of Paraná (Figure 12). Both reflect the period from July 28, 2019, until February 22, 2020.

Figure 11 - Comparison between the map generated through the data obtained in this study and the official data from the government of Paraná.



(a) Smoothed Rating Map of Dengue cases, generated through news data.

(b) Official Report of dengue cases, from Paraná government.

Source: (a) The authors (2021) and (b) Adapted from SESA (2021).

Although some municipalities present divergent data between the two images, it is possible to verify that the northwest region presented a high index in both maps.

4.2 Analyzing data on a city scale

We also analyzed the data on a municipal scale using the municipality of Curitiba as a reference. Curitiba is the capital and largest city in the state of Paraná, with an estimated population in 2020 of 1,948,626 million inhabitants according to IBGE³⁵. An important point is that the city government of Curitiba makes available various types of data through an open data portal and, one of them, is the data of health appointments of its health units. Among the fields contained in the health appointments data is the ICD³⁶ of the diagnosed disease. Thus, we filter the cases of measles, influenza, and dengue using their respective ICDs described in Table 2:

Table 2 - The ICD of searched diseases.

Disease	ICD
Measles	B05
Influenza	J11
Dengue	A90

Source: The authors (2021).

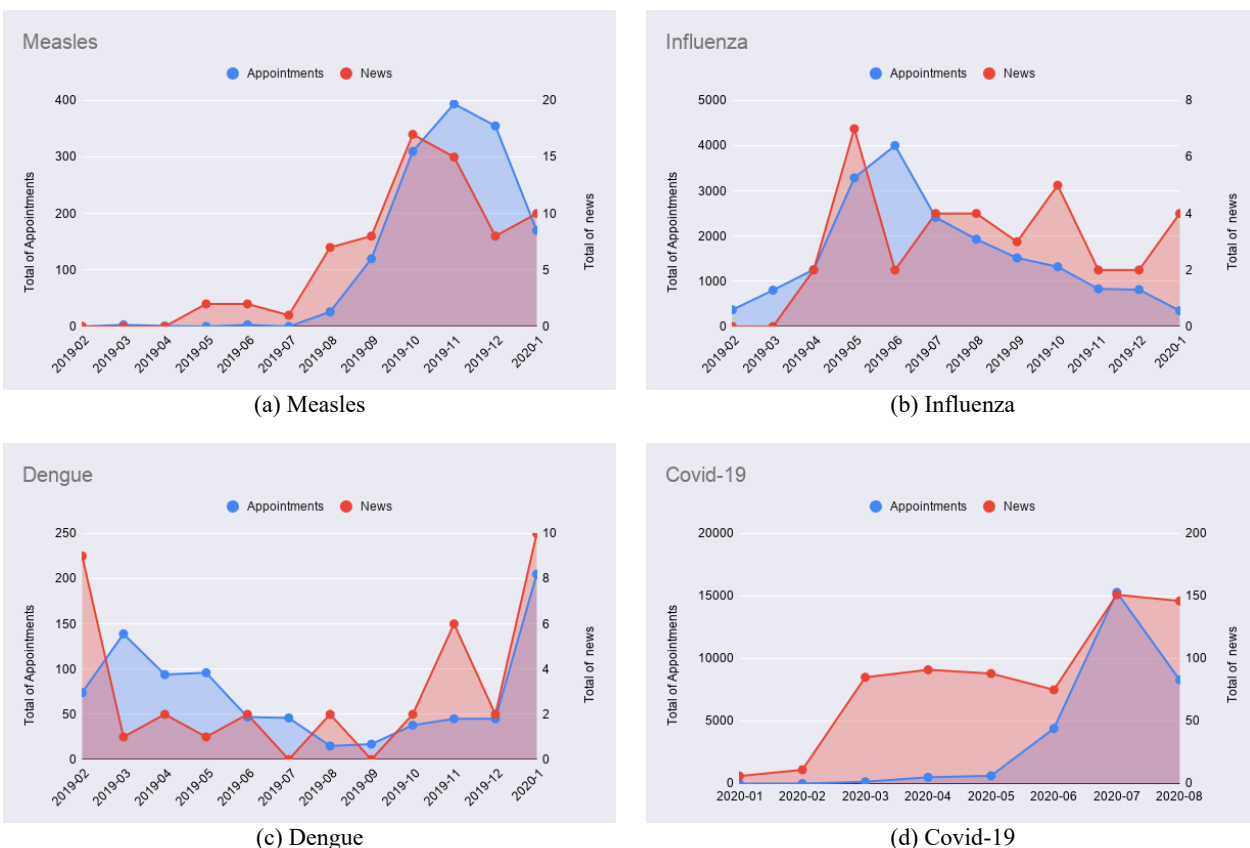
To obtain the number of COVID-19 cases, we used another dataset³⁷, also from the open data portal of the municipality of Curitiba. This dataset was more up-to-date and contained data from March until August 2020.

³⁵ <https://cidades.ibge.gov.br/brasil/pr/curitiba/panorama>

³⁶ International Classification of Diseases

³⁷ <http://dadosabertos.c3sl.ufpr.br/curitiba/CasosCovid19/>

Figure 12 - Comparison between the number of diagnoses and the number of news per month in the city of Curitiba.



Source: The authors (2021).

The initial idea was to compare the number of cases of each disease from the beginning of 2019 to the current date (August 2020) with the number of related news processed by our system. However, we encountered two obstacles: The first was that Curitiba data from health appointments has a gap in January 2019; the second was that the health appointments dataset was limited until mid-February 2020. In this way, we removed January 2019 and included January 2020 in our analysis, thus comprising one year. The only exception was COVID-19, which was analyzed between January and August 2020.

4.3 Findings

In the temporal analysis of the data, we identified an abnormal increase in the volume of news related to infectious diseases between February and August (Figures 8B and 8C), generated mainly by the great attention that COVID-19 has received. This increase generated a distortion in the visualization of other diseases when analyzed on the same scale as the COVID-19 (Figure 8). However, when the data were analyzed in isolation, it was possible to perceive a relationship between its volume of occurrence and the volume of news related to it (Figure 12).

In Figure 10, the diseases strongly influenced by environmental factors (such as dengue), were highlighted in more well-defined regions. Another observed fact was that the relevance of diseases could be identified for a large number of cities, even those of small and medium-size.

Among the challenges faced in the development of this study, we cite the difference in the nomenclature of some Brazilian municipalities, between data from the polygons base and data from the population base, both obtained from the IBGE web-site. In these situations, we did a manual check on the municipality's website to verify the correct spelling. Another challenge was to define a period for data analysis, as some databases had a very short data history (or a gap in the data for a given period), while others comprised a very long uninterrupted period. We try to cover as much data as possible within the periods where most data sources were available.

There are also challenges inherent to projects involving geographic information retrieval (GIR), such as place name disambiguation (Machado et al., 2010), where the algorithm needs to identify the location of an ambiguously named place. To solve this problem, our study used an approach based on filter layers and a final resolution based on the distance from the source of the news and the population of the city.

Another challenge is related to the particularities of the crawling process for each news portal. The AEN website³⁸ paginator, for example, loads the news list via AJAX³⁹. Thus, for each portal, a specific data search algorithm was parameterized.

The data processed and used in this study is available in the Harvard Dataverse repository⁴⁰. All the code developed in this project was also made available on the GitHub platform. The code was split into two parts: The data extractor⁴¹ and the web platform for visualization⁴².

5 CONCLUSIONS

The analysis of unstructured data and patterns has been already mentioned as a potential benefit to healthcare organizations seeking to formulate more effective strategies (WANG et al., 2015).

The first version of this paper (CARDIM; KOZIEVITCH., 2020) (responsible for presenting the processing of unstructured data), was now extended to include an automated web prototype for visualization (Rastreia Saúde), along with the findings/challenges of processing and integrating unstructured data. As a preliminary test, for analysis on a state scale, data from Paraná state were used with 5 different sources. For the analysis on a municipal scale, open data from the municipality of Curitiba were also used. The graphs and maps generated on the data extracted by the platform showed some patterns and confirmed that the system can extract relevant information even in small municipalities.

The analysis showed an abnormal increase in the volume of news related to infectious diseases, generated mainly by the great attention that COVID-19 has received. Another observed fact was that the relevance of diseases could be identified for a large number of cities, even those of small and medium-size. Different city names and data range for different data sources are among the challenges which can be cited.

As future work, we can mention further testings, the expansion of the base of news portals extracted and processed by the system (covering a larger area), along with machine learning (to segment and filter the types of news processed by the system) and other processing steps (such as the differentiation between large cities and states with the same name, such as the state and cities of São Paulo and Rio de Janeiro).

Acknowledgments

The authors thank the support of the UTFPR (edital DIREC 01/2020) for this research. They also thank the Brazilian Institute of Geography and Statistics (IBGE), Curitiba Urban Research and Planning Institute (IPPUC), and the city government of Curitiba for sharing part of the data used in this study.

Authors Contribution

The authors Luiz Henrique Anjos Cardim and Nádia Puchalski Kozievitch declare that they are responsible for preparing the manuscript entitled “Rastreia Saúde: A Spatiotemporal Disease Tracking System through Open Unstructured Data and GIS”. Cardim, as the main author, was responsible for the conceptualization, curation, formal analysis of the data, for the development of the investigative process, and for the design of the study methodology. Kozievitch contributed through the provision of resources and acted in the supervision, validation, and review of the work as a whole.

³⁸ <http://www.aen.pr.gov.br/>

³⁹ Asynchronous JavaScript e XML

⁴⁰ <https://doi.org/10.7910/DVN/1YY646>

⁴¹ <https://github.com/luizhcardim/rastreiasaude-extractor>

⁴² <https://github.com/luizhcardim/rastreiasaude>

Interest conflicts

The authors declare that there are no conflicts of interest.

References

- ANSELIN, L.; LOZANO, N.; KOSCHINSKY, J. Rate Transformations and Smoothing. **Spatial Analysis Laboratory Department of Geography**, 2006.
- BREWER, C. A.; PICKLE, L. Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. **Annals of the Association of American Geographers**, v. 92, n. 4, p. 662–681, 2002. DOI. [10.1111/1467-8306.00310](https://doi.org/10.1111/1467-8306.00310).
- CARDIM, L. H. A.; KOZIEVITCH, N. P. Spatiotemporal disease tracking through open unstructured data and GIS. In: XXI Brazilian Symposium on Geoinformatics – GeoInfo, 2020, On-line. **Proceedings...** São José dos Campos: INPE, 2020. p.130–141.
- CASTRO, M. Q.; JR, C. A. D. Ferramenta Para Recuperação de Informação Utilizando Indexação Espacial e Textual. In: XIX Brazilian Symposium on Geoinformatics – GeoInfo, 2018, Campina Grande, PB, Brazil. **Proceedings...** MCTIC/INPE. p.158–163, 2018.
- CAVALCANTE, J. L. S. B.; NETO, M. S.; KOZIEVITCH, N. P. Utilização e Estudo de Dados de Saúde Georreferenciados Para Desenvolvimento de Aplicação Móvel. In: XIX Brazilian Symposium on Geoinformatics – GeoInfo, 2018, Campina Grande, PB, Brazil. **Proceedings...** MCTIC/INPE. p.170–175, 2018.
- CHOI, J.; SHIM, E.; WOO, H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. **BMC public health**, 16(1), 1238. DOI. <https://doi.org/10.1186/s12889-016-3893-0>.
- CHUNARA, R.; AMAN, S.; SMOLINSKI, M.; BROWNSTEIN, J. Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. **Online Journal of Public Health Informatics**, v. 5, n. 1, 2013.
- CODECO, C.; COELHO, F.; CRUZ, O.; OLIVEIRA, S.; CASTRO, T.; BASTOS, L. Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. **Revue d'Épidémiologie et de Santé Publique**, V. 66, n. 5, p.386, 2018. DOI. [10.1016/j.respe.2018.05.408](https://doi.org/10.1016/j.respe.2018.05.408).
- FIOCRUZ. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT). **MonitoraCovid-19**. Rio de Janeiro, 2020. Available at: <https://bigdata-covid19.iciet.fiocruz.br/>.
- FONSECA, K. V. O.; KOZIEVITCH, N. P.; BERARDI, R. C. G.; SCHMEISKE, O. R. M. Information Technology Macro Trends Impacts on Cities: Guidelines for Urban Planners. In: J. C. Augusto (Org.); **Handbook of Smart Cities**. p.1–24, 2020. Cham: Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-15145-4_58-1.
- FREIFELD, C.; MANDL, K.; BROWNSTEIN, J. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. **Journal of the American Medical Informatics Association : JAMIA**, v. 15, p. 150–7, 2008.
- JAYAWARDHANA, U. K.; GORSEVSKI, P. V. An ontology-based framework for extracting spatio-temporal influenza data using Twitter. **International Journal of Digital Earth**, v. 12, n. 1, p. 2–24, 2019. Taylor & Francis.
- KINDHAUSER, MARY KAY AND WORLD HEALTH ORGANIZATION. (2003). **Communicable diseases 2002: global defence against the infectious disease threat / edited by Mary Kay Kindhauser**.

- World Health Organization. <https://apps.who.int/iris/handle/10665/42572>.
- LAN, R.; ADELFO, M. D.; SAMET, H. Spatio-Temporal Disease Tracking Using News Articles. In: Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health, HealthGIS, 2014, New York, NY, USA. **Proceedings...** New York: ACM, 2014. p.31–38.
- LAN, R.; LIEBERMAN, M. D.; SAMET, H. The Picture of Health: Map-Based, Collaborative Spatio-Temporal Disease Tracking. SIGSPATIAL International Workshop on Use of GIS in Public Health, HealthGIS, New York, NY, USA. **Proceedings...** New York: ACM, 2012. p.27–35.
- LIMA, C. D.; PEIXOTO, A. M.; GOMES-JR, L. C.; LUDERS, R.; FONSECA, K. V. O. Avaliação da Qualidade do Transporte Público no Acesso a Unidades de Saúde de Curitiba. In: Workshop de Computação Urbana (COURB), 3, 2019. Gramado, RS, Brasil. **Proceedings...** Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 71-82.
- MACHADO, I.; ALENCAR, R.; JUNIOR, R.; DAVIS JR, C. An Ontological Gazetteer for Geographic Information Retrieval. In: XI Brazilian Symposium on Geoinformatics – GeoInfo, 2010, Campos do Jordão, SP, Brazil. **Proceedings...** São José dos Campos: INPE, 2010. p.21–32, 2010.
- MADOFF, L. ProMED-mail: An Early Warning System for Emerging Diseases. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 39, p. 227–32, 2004.
- MINISTÉRIO DA SAÚDE. **Doenças Infecciosas e Parasitárias: guia de bolso. 8 ed.** Brasília: Ministério da Saúde, 2010. Available at: https://bvsmis.saude.gov.br/bvs/publicacoes/doencas_infecciosas_parasitaria_guia_bolso.pdf.
- MOHANTY, B.; CHUGHTAI, A.; RABHI, F. Use of Mobile Apps for epidemic surveillance and response – availability and gaps. **Global Biosecurity**, v. 1, p. 37, 2019.
- NPM. **About npm | npm Docs**. 2021. Available at: <https://docs.npmjs.com/about-npm>. Accessed on: 13 feb. 2021.
- OLIVEIRA, M. F. A. DE; KOZIEVITCH, N. P.; BIM, S. A.; LEGAL-AYALA, H. Caracterização dos Dados Públicos de Saúde do Paraguai. In: Escola Regional de Banco de Dados (ERBD), 2018, Rio Grande, RS, Brasil. **Proceedings...** Porto Alegre, RS, Brasil: SBC. p.12 – 21, 2018.
- RDA COVID-19 WORKING GROUPS. **RDA COVID-19 Working Group Recommendations and Guidelines, 1st release**. Research Data Alliance, 2020. Available at: <https://www.rd-alliance.org/system/files/RDA%20COVID-19%3B%20recommendations%20and%20guidelines%2C%201st%20release%2024%20April%202020.pdf>.
- SANKARANARAYANAN, J.; SAMET, H.; TEITLER, B. E.; LIEBERMAN, M. D.; SPERLING, J. TwitterStand: news in tweets. In: 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS, November 4-6, 2009, Seattle, Washington, USA. **Proceedings...** New York: ACM, 2009. p.42–51, 2009.
- SESA. Situação da Dengue, Chikungunya e Zika vírus no Paraná. Paraná, 2021. Available at: http://www.dengue.pr.gov.br/sites/dengue/arquivos_restritos/files/documento/2020-11/boletimdengue27_2020_1.pdf. Accessed on: 05 apr. 2021.
- STATCOUNTER. **Desktop vs Mobile vs Tablet Market Share Brazil | StatCounter Global Stats**. 2021. Available at: <https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet/brazil>. Accessed on: 12 feb. 2021.
- WANG, Y.; KUNG, L.; TING, C.; BYRD, T. A. Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care. In: International Conference on System Sciences, 5-8 jan, 2015, Kauai, HI,

USA. **Proceedings...** New York: IEEE, 2015. p.3044–3053.

Main author biography



Luiz Henrique Anjos Cardim was born in the city of São Paulo. He has a degree in Distributed Systems Development from the Federal Technological University of Paraná (UTFPR), post-graduated in Database Design and Management by PUC-PR, and is a master's degree student in Applied Computing at UTFPR. His interests involve geographic information systems, extraction of spatiotemporal information through unstructured data, and open data visualizations.



Esta obra está licenciada com uma Licença [Creative Commons Atribuição 4.0 Internacional](https://creativecommons.org/licenses/by/4.0/) – CC BY. Esta licença permite que outros distribuam, remixem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuem o devido crédito pela criação original.