Spatiotemporal disease tracking through open unstructured data and GIS

Luiz H. A. Cardim[1], **Nádia P. Kozievitch**[1]

[1]Departamento de Informática - Universidade Tecnológica Federal do Paraná (UTFPR)

luizcardim@alunos.utfpr.edu.br, nadiap@utfpr.edu.br

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. This study aims to extract relevant information in the health segment from unstructured data, extracted from news portals. The system should generate data that allows analysis at different levels of granularity, from small municipalities to the national level. The results of the study demonstrated the viability of the system and allowed the authors to identify some patterns in the processed data.*

1. Introduction

The dynamics of the current world, where millions (or even billions) of people move every day between different neighborhoods, cities, countries, or even continents, created the need to think about equally dynamic ways of monitoring some types of information, such as the spread of communicable diseases. COVID-19 showed us that a virus can, in a matter of months, and starting from some local cases, quickly turn into a global pandemic [World Health Organization 2020]. The impact of this pandemic has even changed the way data is shared [RDA COVID-19 Working Groups 2020], trying to make data sharing simpler. We add to this highly dynamic scenario of human mobility the growing urban cluster that has occurred in recent years, creating cities that are increasingly larger and with greater population density [United Nations 2019]. In this scenario, the delay in identifying the outbreak of a new disease can generate disastrous conditions, putting human lives at risk.

The internet offers us a very rich source of information about the occurrence of diseases in certain regions, like open data from cities[1], open data from public-private partnerships[2], open data from researchers[3], open unstructured data (like news portals), and others, however, its immense volume and heterogeneity of data makes the task of synthesizing all this information manually very costly or even impractical in the case of large geographic spaces or very long periods. In this way, several types of research have been carried out to automate this collection and synthesis of data into relevant information, making it possible to geographically monitor outbreaks at a global level and for extended periods.

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

[1]As example we cite the Curitiba Open Data Portal (`https://www.curitiba.pr.gov.br/dadosabertos/`) used in this study

[2]`https://repositoriodatasharingfapesp.uspdigital.usp.br/`

[3]`https://dataverse.harvard.edu/`

those of other countries [Kindhauser, Mary Kay and 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.

This 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 rest of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the project architecture. 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-Mail[4] [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 are published, they are checked by specialists, which makes the platform a reliable and recognized source of data in the disease tracking segment. The World Health Organization (WHO) also maintains an alert portal for the emergence of new communicable diseases[5] 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 HealthMap[6] platform, one of the first disease tracking platforms based on unstructured data. HealthMap extracts alert on communicable diseases on a global basis by extracting data from several sources, including news sites and also ProMed reports.

A similar approach is used in the study by [Lan et al. 2012], which presents the STEWARD[7] 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 Newsstand[8] 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], that tries to track the location and timing of a 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

[4] https://promedmail.org/
[5] https://www.who.int/csr/don/en/
[6] https://www.healthmap.org/
[7] http://steward.umiacs.umd.edu
[8] http://newsstand.umiacs.umd.edu/web/

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 you[9], 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.

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 it.

[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 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] carried out a survey with 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, analysing the accessibility to Curitiba public health units via public transportation.

3. System 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 prototype developed in our study is shown in Figure 1. The system has three sources of data:

1. HTML from the news portals extracted through web crawling.
2. The shapefile of all the Brazilian municipalities and other information like the population of each city, extracted from the IBGE[10].
3. The list of infectious and parasitic diseases, obtained from the SUS[11] and manually inputted in the system.

[9] `https://flunearyou.org/`

[10] Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics)

[11] Sistema Único de Saúde (The Brazilian Universal Health Care System)

The choice of what diseases to track was based on data published by the Ministry of Health of Brazil [Silva and Ferreira 2006]. 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, coccidioidomicose, cólera, coqueluche, criptococose, criptosporidíase, dengue, difteria, doença de chagas, doença de lyme, diarréia, doença meningocócica, donovanose, enterobíase, escabiose, esquistossomose mansônica, estrongiloidíase, febre amarela, febre maculosa brasileira, febre purpúrica brasileira, febre tifóide, filaríase por wuchereria bancrofti, giardíase, gonorreia, hanseníase, hantavirose, hepatite a, hepatite b, hepatite c, hepatite d, hepatite, e, herpes, histoplasmose, HPV, influenza, leishmaniose, leptospirose, linfogranuloma venéreo, malária, meningite, mononucleose, oncocercose, paracoccidioidomicose, parotidite, peste, poliomelite, psitacose, rubéola, sarampo. shigelose, sífilis, cisticercose, tétano, toxoplasmose, tracoma, tuberculose and varicela.

Among the software used in the project, in the data extractor we use the python[12] language (version 3.8.5) with the scrapy framework[13] (version 2.3.0) to perform the data crawling. For NLP we use the spacy[14] library (version 2.3.2) configured for the Portuguese language. The database used was PostgreSQL[15] (version 12.3) with the Postgis[16] extension (version 2.5).

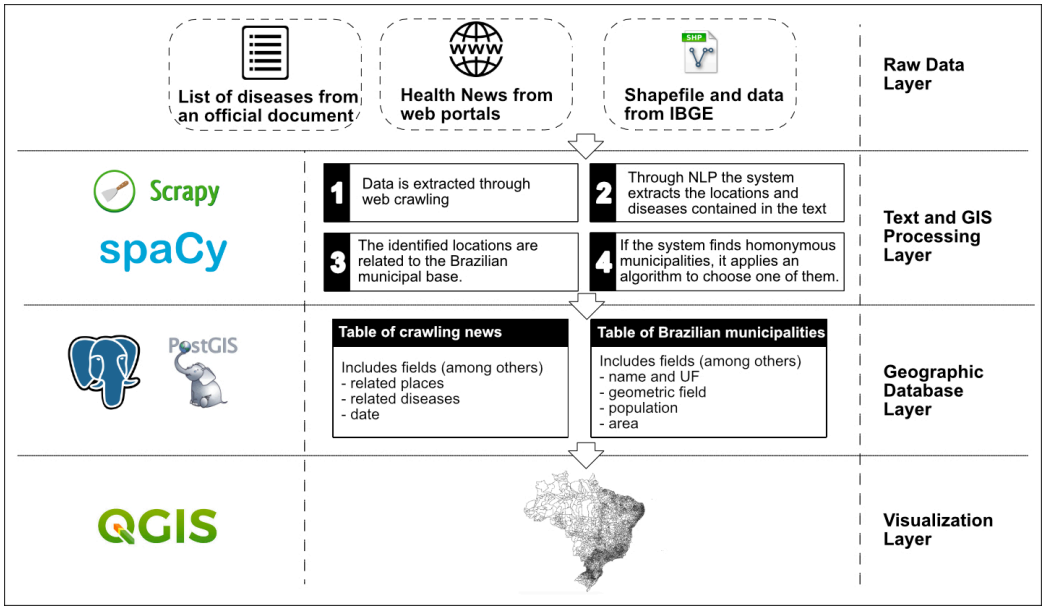


Figure 1: The architecture of the data extraction system.

[12] https://www.python.org/
[13] https://scrapy.org/
[14] https://spacy.io/
[15] https://www.postgresql.org/
[16] https://postgis.net/

3.1. Processing the Unstructured Data

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

1. 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.
2. The system extracts the date of publication (or last update) of the news.
3. The system searches in the text content (body of news and title) all occurrences of the diseases being tracked, using regular expressions.
4. Diseases found with different terms are grouped into only one key term. For example, coronavirus, COVID, and COVID-19 are grouped into COVID-19.
5. If the system finds any diseases, then it does an NLP to identify the locations contained in the text.
6. 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.
7. The locations found in the previous steps are searched at the base of Brazilian municipalities, previously extracted from IBGE.
8. 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.
9. Finally, if the system has identified at least one disease and one location in the news, it adds the record to the database.

4. Results and Discussion

4.1. Analyzing data on a state scale

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 IBGE. 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) Portal[17], Bandab[18], Bem Paraná[19], AEN[20] (the official news agency of Parana government) and Tribuna do Paraná[21], with the data range from January 1, 2019 to August 20, 2020. The distribution in time of the extracted news is presented in Figure 2 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

[17] http://www.saude.pr.gov.br/
[18] https://www.bandab.com.br/
[19] https://www.bemparana.com.br/
[20] http://www.aen.pr.gov.br/
[21] https://www.tribunapr.com.br/

on the SESA and the AEN portal. The reason is that both portals are managed by the same organization, the Paraná government.

The data processed and used in this study were made available in the Harvard Dataverse repository[22].

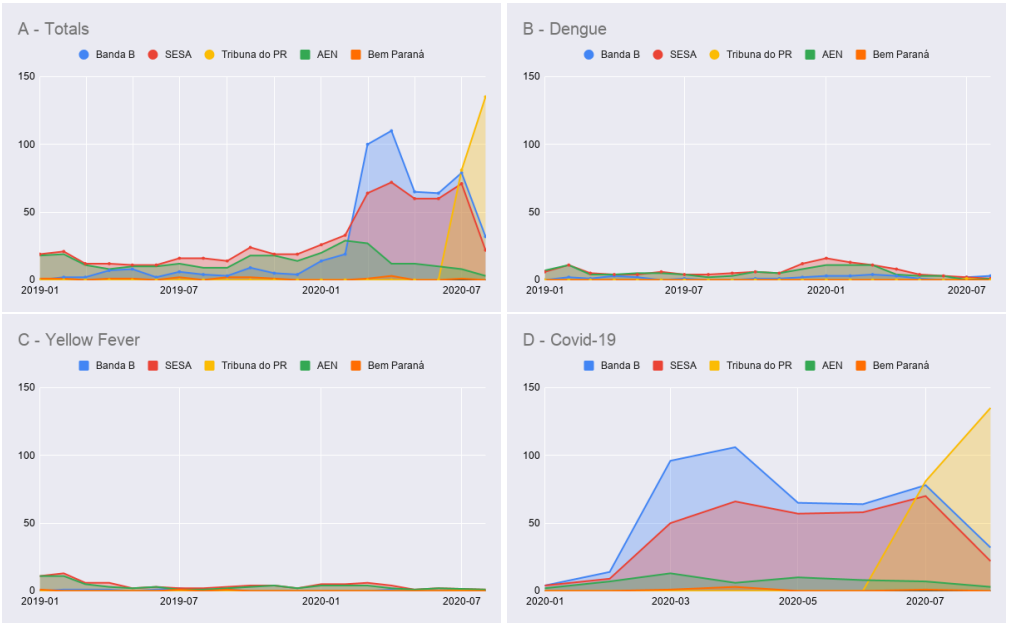


Figure 2: Number of news extracted from each portal by month and disease.

Figure 2 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 a greater number of news published than all other diseases combined.

Disease	**Related news**
COVID-19	1072
Dengue	271
Measles	177
Yellow Fever	161
Influenza	104
Rubella	54
HIV	46
Meningitis	36
Varicella	34
Tuberculosis	33

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

[22] https://doi.org/10.7910/DVN/1YY646

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 3.

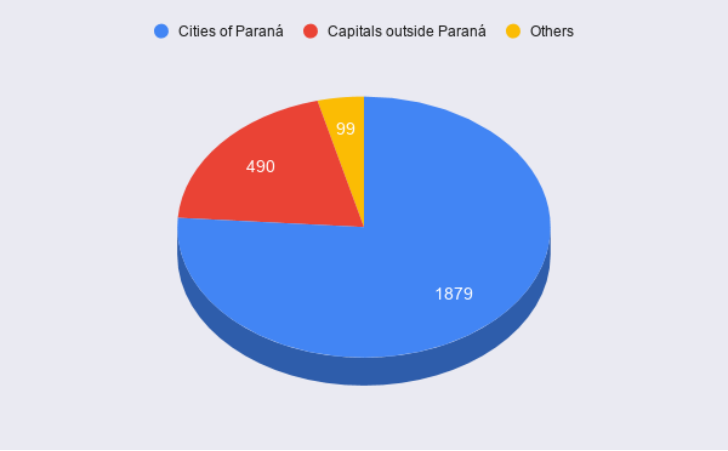


Figure 3: Number of news extracted from 3 groups.

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 4) 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:

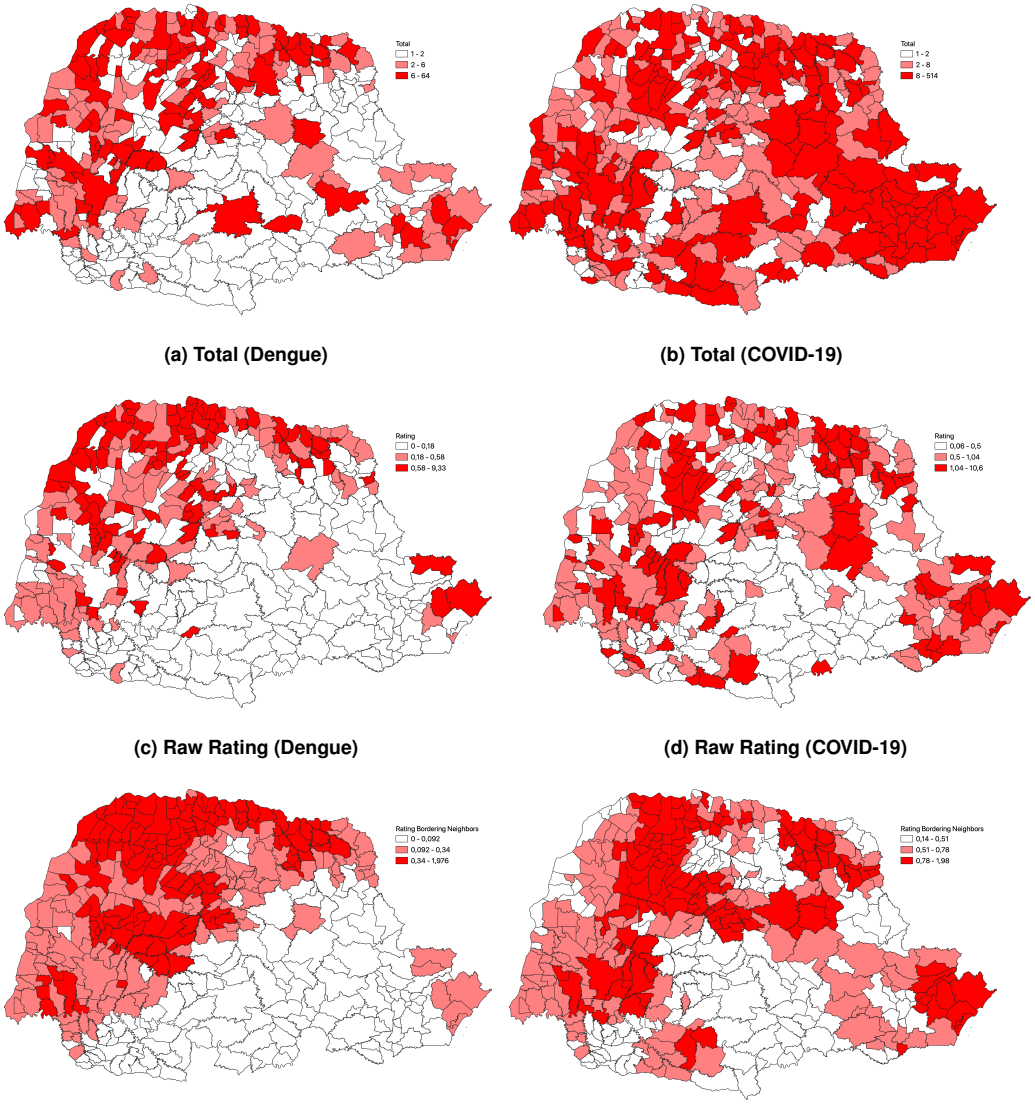
1. A version based on the total number of news of a given disease-related to each city.
2. A version based on a raw ratio per thousand (rrpt), obtained using the equation 1.
3. A version based on a smoothed rate per thousand (srpr), obtained using the equation 2.

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

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

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.



(a) Total (Dengue) **(b) Total (COVID-19)**

(c) Raw Rating (Dengue) **(d) Raw Rating (COVID-19)**

(e) Smoothed Rate Per Thousand (Dengue) **(f) Smoothed Rate Per Thousand (COVID-19)**

Figure 4: The distribution of the identified diseases on the map.

Through the raw rate per thousand (equation 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 estimates for an underlying risk, i.e., the probability for a particular event to occur. The smoothed rate per thousand (equation 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 4 shows complementary visualizations of the diseases. For example, to identify the regions most affected by dengue, Figure 4-E showed a better-defined pattern, while for COVID-19 the map in Figure 4-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.

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 IBGE. An important point is that the city government of Curitiba makes available various types of data through an open data portal[23] 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 ICD[24] of the diagnosed disease. Thus, we filter the cases of measles, influenza, and dengue using their respective ICDs described in Table 2:

Disease	**ICD**
Measles	B05
Influenza	J11
Dengue	A90

Table 2: The ICD of searched diseases.

To obtain the number of COVID-19 cases, we used another dataset, 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.

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 of 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.

[23] `https://www.curitiba.pr.gov.br/dadosabertos/`
[24] International Classification of Diseases

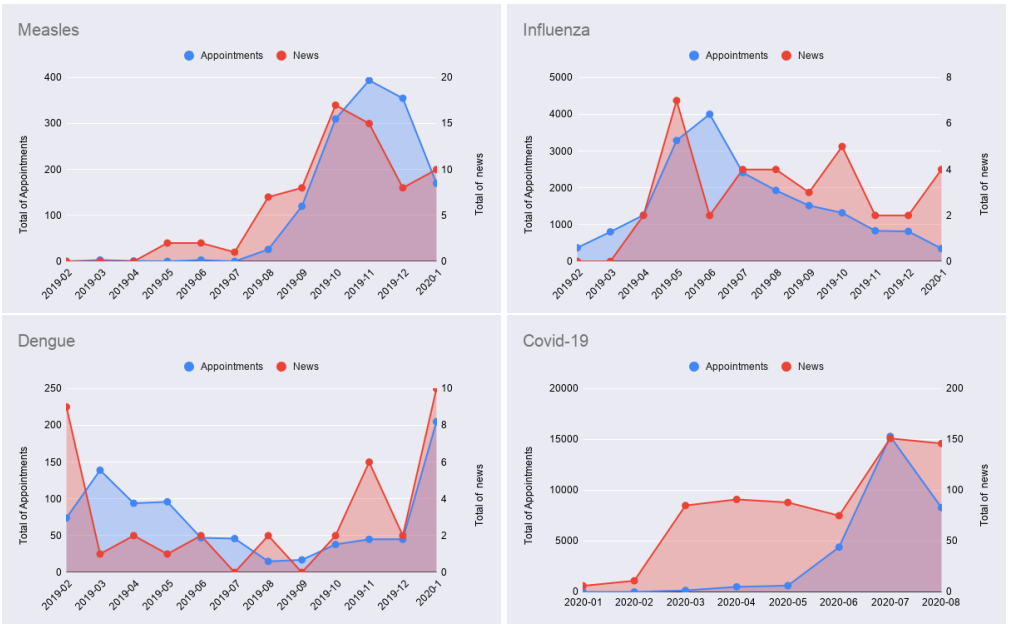


Figure 5: Comparison between the number of diagnoses and the number of news per month in the city of Curitiba.

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 (Figure 2A), 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 2). 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 5).

Analyzing the spatiality of the data (Figure 4), we realized that 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 we 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 website. 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.

5. Conclusions

This study presented the architecture and preliminary results of a prototype system for disease tracking system on a national scale. As a preliminary test, for analysis on a state

scale, data from Paraná 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 on small municipalities. However, the system needs further testing before it can be made available for public access. In future studies, we intend to expand the base of news portals extracted and processed by the system, covering the entire national territory and develop a mobile interface for users to consult information. Future studies can also apply machine learning to segment and filter the types of news processed by the system. Other processing steps should also be added to the system, such as the differentiation between large cities and states with the same name, such as São Paulo and Rio de Janeiro.

6. Acknowledgments

The authors 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.

References

Anselin, L., Lozano, N., and Koschinsky, J. (2006). Rate transformations and smoothing. *Spatial Analysis Laboratory Department of Geography*.

Brewer, C. A. and Pickle, L. (2002). Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4):662–681.

Castro, M. Q. and Jr., C. A. D. (2018). Ferramenta para recuperação de informação utilizando indexação espacial e textual. In Vinhas, L. and Campelo, C. E. C., editors, *XIX Brazilian Symposium on Geoinformatics - GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, 2018*, pages 158–163. MCTIC/INPE.

Cavalcante, J. L. S. B., Neto, M. S., and Kozievitch, N. P. (2018). Utilização e estudo de dados de saúde georreferenciados para desenvolvimento de aplicação móvel. In Vinhas, L. and Campelo, C. E. C., editors, *XIX Brazilian Symposium on Geoinformatics - GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, 2018*, pages 170–175. MCTIC/INPE.

Choi, J., Shim, E., and Woo, H. (2016). Web-based infectious disease surveillance systems and public health perspectives: A systematic review. *BMC Public Health*, 16.

Chunara, R., Aman, S., Smolinski, M., and Brownstein, J. (2013). Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. *Online Journal of Public Health Informatics*, 5(1).

de Oliveira, M. F. A., Kozievitch, N. P., Bim, S. A., and Legal-Ayala, H. (2018). Caracterização dos dados públicos de saúde do paraguai. In *ERBD 2018*, page 12, Porto Alegre, RS, Brasil. SBC.

Freifeld, C., Mandl, K., and Brownstein, J. (2008). Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association : JAMIA*, 15:150–7.

Jayawardhana, U. K. and Gorsevski, P. V. (2019). An ontology-based framework for extracting spatio-temporal influenza data using twitter. *International Journal of Digital Earth*, 12(1):2–24.

Kindhauser, Mary Kay and World Health Organization (2003). Communicable diseases 2002 : global defence against the infectious disease threat / edited by mary kay kindhauser. https://apps.who.int/iris/handle/10665/42572.

Lan, R., Adelfio, M. D., and Samet, H. (2014). Spatio-temporal disease tracking using news articles. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health*, HealthGIS '14, page 31–38, New York, NY, USA. Association for Computing Machinery.

Lan, R., Lieberman, M. D., and Samet, H. (2012). The picture of health: Map-based, collaborative spatio-temporal disease tracking. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health*, HealthGIS '12, page 27–35, New York, NY, USA. Association for Computing Machinery.

Lima, C. D., Peixoto, A. M., Gomes-JR, L. C., Luders, R., and Fonseca, K. V. O. (2019). Avaliação da qualidade do transporte público no acesso a unidades de saúde de curitiba. In *Anais do III Workshop de Computação Urbana (COURB 2019)*, volume 1, Gramado, RS, Brasil. SBC.

Madoff, L. (2004). Promed-mail: An early warning system for emerging diseases. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 39:227–32.

Mohanty, B., Chughtai, A., and Rabhi, F. (2019). Use of mobile apps for epidemic surveillance and response – availability and gaps. *Global Biosecurity*, 1:37.

RDA COVID-19 Working Groups (2020). *RDA COVID-19 Working Group Recommendations and Guidelines, 1st release*. Research Data Alliance.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In Agrawal, D., Aref, W. G., Lu, C., Mokbel, M. F., Scheuermann, P., Shahabi, C., and Wolfson, O., editors, *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009, November 4-6, 2009, Seattle, Washington, USA, Proceedings*, pages 42–51. ACM.

Silva, T. P. T. e. and Ferreira, I. d. L. M. (2006). Doenças infecciosas e parasitárias: guia de bolso. *Cadernos de Saúde*, 22:2498 – 2498.

United Nations (2019). World urbanization prospects: The 2018 revision. https://www.un-ilibrary.org/content/publication/b9e995fe-en.

World Health Organization (2020). Timeline of who's response to covid-19. https://www.who.int/news/item/29-06-2020-covidtimeline. accessed June 3, 2020.