Towards Measuring Air Quality Index with Country and States Granularity from Sentinel-5 Data

Jordan Salas Cuno
Computer Institute
Universidade Federal Fluminense
Niterói, Brazil
jordansalas@id.uff.br

Aura Conci
Computer Institute
Universidade Federal Fluminense
Niterói, Brazil
Email: aconci@ic.uff.br

Andouglas G. da Silva-Júnior
Grad. Prog. in Elec. and Comp. Eng.
Instituto Federal do Rio Grande do Norte
Natal, Brazil
Email: andouglas.silva@ifrn.edu.br

Luiz M. G. Gonçalves
Grad. Prog. in Elec. and Comp. Eng.
Universidade Federal do Rio Grande do Norte
Natal, Brazil
E-mail: lmarcos@dca.ufrn.br

***Abstract*—One of the topics that have been recently discussed with greater emphasis is air quality, not only because it is a topic directly related to climate change and the greenhouse effect, but also because it has a strong link in the transmission of respiratory diseases. Low-quality indices of air can worsen the symptoms of patients with the COVID-19 pandemic, being one of the most recent examples. Being able to identify and monitor air quality in different geographic regions of Brazil will provide information on possible measures to be taken by authorities and citizens in general. Air quality indices are currently monitored by some entities using technology of sensors strategically installed at fixed points in cities. Another type of technology available is the high-resolution spectrometer system installed on satellites that constantly monitor the surface of the planet. Sentinel-5 is one of these satellites that generates a large amount of data for climate monitoring and that is used in this research. Basically, in this work, we propose a system that takes data from the Sentinel-5 in order to provide a measure of air quality to be used in the prediction of the pandemic dynamics.**

***Index Terms*—Air Quality Sensor, Sentinel-5 Data, Pandemic Dynamics, Data-Driven Modeling**

I. Introduction

In the current pandemic context, studying and developing applications with AI to predict the dynamics of pandemics has immediate use in public management. The techniques based on LSTM and MAE have applicability in time series prediction [1], but it remains to be determined how accurately they can determine data from a spread. Analyzing the actions to combat the virus, both in the social and economic spheres, there is a need for more realistic epidemiological data, with local models that allow the authorities to make decisions in a coherent way, given the geographic and demographic reality of each region, state or municipality. Thus, we believe that the development of AI-based methods to study the dynamics of pandemics can better predict parameters such as the predicted number of cases, deaths, and contamination rate, among others, compared to traditional techniques. Among these data, we find some that are related to the pandemic itself (number of deaths, number of sick people in the ICU, number of daily cases) and others that are not directly related to it, but indirectly, such as measures of agglutination of people, the existence of comorbidities, population, isolation level, as well as climate and atmospheric parameters such as humidity, temperature, and air quality. In this work, we study the last one (air quality), which the literature indicates has a great influence on the development of diseases that cause cardio-respiratory failure [2]. These diseases can be associated with poor air quality, and are, in a first analysis, one of the main factors of aggravation of the Sars-Cov-2 virus, causing mortality in most patients who are affected by the virus and are carriers of the same.

Over time the air quality has been affected by pollution and different kinds of factors or events. Among them, we can cite the constant consumption of fossil fuels, a huge amount of fires, and others. The degradation of air quality has increased in the last decades, but, at the beginning of 2020^{ies} the COVID-19 pandemic has been declared by World Health Organization (WHO). This disease caused by the Sars-Cov-2 event has changed a lot the air quality index because one of the actions most used to stop the spread of the disease was social isolation or, in more rigid situations, a complete lockdown. These kinds of measures were used mainly in the big cities, keeping thousands of people at home, and decreasing significantly the gas emission generated by vehicles, industries, and people agglomeration. Determining the air quality in different periods of time and different geographic areas in Brazil gives an overview of the impacts climatic global.

A set of measured atmospheric parameters is usually used as input to determine the quality of the air. In this research, the data are generated by the Sentinel-5 satellite. One of its main characteristics is the constant monitoring of the earth's surface, and thanks to this resource the process can be done in different regions. To make the experiment feasible, our geographic area of study is defined as for being in Brazil,

This work is sponsored by CAPES under grant 88887.506889/2020-00.

which means that the process to determine the quality of the air involves a cartographic analysis of the cities and states of Brazil. Questions as the granularity of data should also be thought about during our research.

Hence, our motivation is that our studies will contribute to the larger project, called *Methods for predicting the dynamics of viral epidemics and pandemics with clustered data analysis from the perspective of artificial intelligence.*, developed within the scope of CAPES-EPIDEMIAS Program, by the Natalnet Associated Laboratories (UFRN) in partnership with the Media-Lab (UFF). The main objective of this project is to predict the dynamics of advancing the COVID-19 pandemic using Artificial Intelligence methods, that is, with data-driven models. The approach used, called data-driven, uses the most modern Machine Learning technologies to make an (informed) prediction of the future behavior of the pandemic. The main advantage of this approach is the incorporation of other aspects for the prediction that influence the behavior of the pandemic, but which cannot be used in traditional epidemiological models (SIR and SEIRD, among others [1]), such as mobility indexes, climatic factors, and air pollution, the latter being the topic of study of the present work. The larger project proposes the creation of software tools for the analysis of pandemics, with the availability of curve predictions to health management authorities through a virtual page. Furthermore, the technological tools developed during this project can be reused in the mitigation of future viral pandemics and endemics.

As part of the project, the team designed and implemented the n-Covid system, which presents itself as an easy-to-reuse modular tool for rapid prediction of future epidemics. At the moment, the project has its active website (http://ncovid.natalnet.br/), already with some basic features such as presenting the curve of deaths by COVID-19 for the states of Brazil and its short-term prediction (7 days) given by an LSTM type prediction model. The predictions are presented for each state in Brazil and also for the country as a whole.

Thus, the main idea of this article is to contribute with the description of a method that can be used to determine air quality data within a period of time, focused on a geographic area corresponding to Brazil, initially, detailing the characteristics of the Sentinel-5 satellite and the data set that are used. in this research. As highlighted above, the estimates computed in the larger project use, among other variables, the air quality indicator (AQI), for which we contribute with the use of Sentinel-5 data that can be processed at a level of granularity desired by the tools at the end of our research.

II. Air Quality Theory

The AQI is one of the most used indicators to evaluate the air quality of an area. Based on several parameters analysis it is possible to verify whether the air in a certain place is dangerous to humans and animals, mainly when related to respiratory diseases. In the literature on air quality and its relationship with Covid-19 [2], [3], one of the main topics discussed and suggested to the end is its improvement. Improving air quality means minimizing respiratory problems, as well as other chronic diseases that are responsible for the majority of deaths from Covid-19. Thus, in this section, we discuss the air quality index and the most common methods used to acquire the parameter values and compute this index. We present the main theoretical topics related to its determination, including physical means for capturing atmospheric parameters, which can either be using sensors (in-place) or using data from the earth's surface captured by satellites.

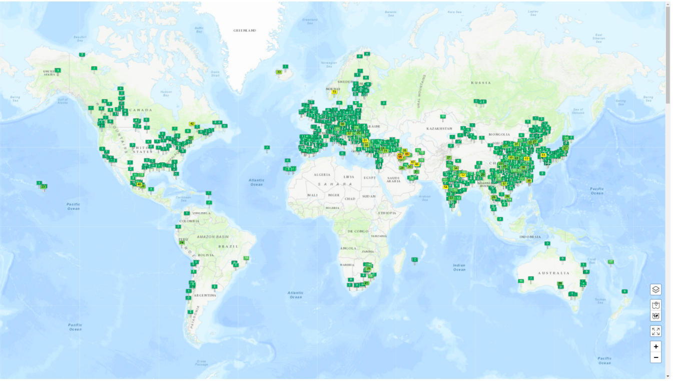

Fig. 1. World map of existing AQI stations

A. AQI Data from Sensors and Satellite

The Air Quality Index (AQI) can be calculated based on measurements of atmospheric parameters, the main of them is particulate matter (PM2.5 and PM10). Other pollutants are Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), and Carbon Monoxide (CO) emissions. Most ground stations have been built nowadays for monitoring both PM2.5 and PM10 data, but there are few exceptions where only PM10 is available [4]. It is important to notice that it is not necessary to get all the parameter values to estimate the AQI. As the final value is given by the worst parameter value, if only one parameter is out of the interval indicating the good range, the index results will be out as well. All measurements are based on hourly readings. For instance, an AQI reported at 8 am means that the measurement was done from 7 am to 8 am.

Basically, there are two ways to acquire air quality data. The first one is in-loco, where different kinds of sensors, all together in a station, can be used to get their specific information, on the variables above (Ozone, PM, and so on). The Figure 1 shows several, but not all of them, existing air quality stations around the world [4]. It can be seen that there are basically a few of those located in Brazil, mainly in the southern regions or cities of the country.

On the other hand, it is possible to acquire this information using data from satellites, which uses information computed from the radiation of light from regions on the earth's surface to compute the value of the parameters. Air quality scientists discovered a long time that each pollutant has specific radiation. Hence, there exist several systems nowadays in order to measure pollution from satellite data, such as the *World's Air Pollution: Real-time Air Quality Index* [4], the *Air quality* *index (AQI) and PM2.5 air pollution* [5], and the *Air Pollution: Real-time Air Quality Index (AQI)* [6]. Knowing the air quality index for some particular city is possible, as for example the Hanoi Air Quality Index (AQI) and Vietnam Air Pollution [7].

Other sites such as the *Health and Air Quality Data Pathfinder* from NASA [8] provide access open for scientists. They say that *Air pollution is one of the largest global environmental and health threats*. Hence, they have put instruments aboard NASA satellites and airborne platforms that continually acquire data about these pollutants.

In our work, we use data provided by the Sentinel-5 [9], which is used in our computations in order to give an AQI with better quality. So, in the current work, we focus only on the Sentinel-5 data. Besides, we notice that the use of in-loco (PM, humidity, and temperature) sensors, discussed next, is also an option in our big project, and another researcher in parallel work to ours is developing these sensors for AQI determination [10].

B. Particulate Matter Sensors

One parameter that is very used to estimate the air quality is the particulate matter (PM), where the most common sizes are 2.5 and 1 micrometer. Several entities and scientific works showed these particulates can facilitate the spread of diseases, such as Covid-19. For example, the United States Environmental Protection Agency (EPA) affirms that small particles less than 10 micrometers can get deep into the lungs, and in some cases can affect even the bloodstream [11]. In this sense, it is suggested that the AQI, PM2.5, NO2, and temperature parameters are the four variables that could promote the sustained transmission of Covid-19 [12]. Other interesting works have shown that the lockdown, a countermeasure broadly used to try to stop the spread of Covid-19 worldwide, changed significantly the levels of particulate matter in the air [10], [13]–[15].

To acquire this data, in most cases, PM sensors are used, which are able to detect the number of particles that passes between an emitter and a light receiver in a time interval. One quite used is the DSM501A, which generally, is one of the sensors embedded in the station [10].

C. In-loco Sensors for PM, Umidity, and Temperature

The work done by Bezerra [10] proposes a system to collect and provide air quality data to processes that can help real-time monitoring of indoor environments such as classrooms, offices, laboratories, and other public spaces. The authors built a system for future application in order to suggest actions for improving indoor air quality or alerting to preventive measures as necessary. Their system is composed of a sensors module with low-cost sensors to monitor PM2.5, temperature, and humidity. A basic software infrastructure to store and process the sensor's data is also provided in the work. The system also encompasses the development of an IoT web platform capable of providing data sets in the form of Air Quality Indicator (AQI) parameters.

This platform allows to access the raw data with temporal and geographically annotated parameters, that are made available to the academy and the population, by way of a web server. These data are input to the same Data Analysis module as ours will go through, which uses Machine Learning (ML) and Artificial Intelligence (AI) techniques in order to indicate if the air quality is adequate or not. The system also suggests actions that can be taken to improve the quality of the air [2].

In the following Section, we briefly discuss the analytics data tools used in the project in order to process data from Covid-19 together with air quality data and others. Our data is thought to be input to that developed tool.

D. Covid-19 Dynamics Prediction

Estimating the dynamics of epidemiological data for Covid-19 is the goal of the larger project to which we intend to contribute with some part of it in this article. The use of a methodology based on a data-driven modeling, which allows predicting parameters closer to reality from training with data from cases that have already taken place in other countries or regions [1], is one of the options already implemented. Our preliminary works [1], [2] indicate that this is a possible way forward. Improvements in the theoretical-methodological part, such as the use of data obtained with more quality, guarantee better results with better maps of the distribution of the pandemic.

Actually, the pandemic numbers can be understood as time series data. Hence, the basic flow for time series prediction using data-driven techniques can be summarized in collecting the data, processing the data, training the prediction model, evaluating the prediction model, optimizing the model (if necessary), and using it. In order to make predictions, and achieve the project objectives, this flow unfolds in several lines of research, being our work one of these.

Indeed, we intend to show here how data from the Sentinel-5, which will be detailed next in Section III can help to increase the data quality. Actually, several parameters can be used for dynamics prediction of Covid-19, besides AQI [2].

III. Reading and Processing Sentinel-5 AQI Data

Sentinel-5 is a low-orbit earth satellite system designed to provide information on air quality and climate composition, in addition to monitoring the ozone layer [9]. Sentinel-5 is part of the European Earth Observation Program (Copernicus) which is directed by the European Commission (EC), The main objective is to carry out atmospheric measurements with high Spatio-temporal resolution, related to air quality, components of the climate system, ozone and UV radiation [16].

A. Features

The Sentinel-5 mission consists of a high-resolution spectrometer system operating in the ultraviolet to shortwave infrared range, consisting of 7 different spectral bands UV-1 (270-300nm), UV-2 (300-370nm), VIS (370-500nm), NIR-1 (685-710nm), NIR-2 (745-773nm), SWIR-1 (1590-1675nm) and SWIR-3 (2305-2385nm). The spectral resolution varies from 1nm in UV1 to 0.25nm in SWIR channels, with the main climatic components being O_3, NO_2, SO_2, HCHO, CO, CH_4 . The sweep angle of the sensor is 108°, considering the height of the satellite orbit (817km) the corresponding distance of the track on the ground is 2670km. Figure 2 illustrates how this is performed, based on the TROPOMI measurement principle [16].

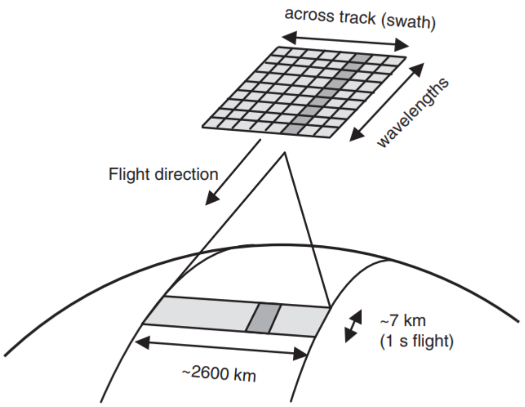

Fig. 2. TROPOMI measurement principle. The dark-gray ground pixel is imaged on the two-dimensional detector as a spectrum. All ground pixels in the 2600km wide swath are simultaneously measured (Veefkind et al., 2012)

B. Data Structure

Information provided by the Sentinel-5 mission is organized into a three-tiered structure.

- Level-0: Contains information about the satellite's orientation, this information is saved but not available to users.
- Level-1B: Contains geolocated and corrected terrestrial radiation information.
- Level-2: Contains geophysical information derived from the processing of measured data provided by Level-1B.

The data used in this research are obtained from Level-2. At this level, there are three types of flows to work with this data. The first is NRT, which is the *near real-time* stream available 3 hours after scanning. The other is the offline stream, where information is available after a few days. And, the last is the reprocessing stream, where possible missing information is corrected. Table I shows an example of the Level 2 products, with identifier and institution.

C. Dataset

The information collected by Sentinel-5 is available on the Copernicus Open Access Hub website, using APIs it is possible to download the information to be processed, in this work a script was developed with the purpose of performing searches automatically within the normal of the website, they are defined some mandatory parameters in the API request, a period of time must be defined in days, weeks or months, define which Level-2 product will be downloaded, in addition to a georeferenced polygon to define on which area of the earth's surface where the experiments will be done.

TABLE I
LEVEL 2 PRODUCTS

Product	**Identifier**	**Institution**
Cloud	L2__CLOUD_	DLR
NPP-VIIRS Clouds	L2__NP_BDx	RAL
HCHO	L2__HCHO__	BIRA/DLR
SO_2	L2__SO2___	BIRA/DLR
O_3 Total Column	L2__O3 ____	BIRA/DLR
O_3 Tropospheric Column	L2__O3_TCL	IUP/DLR
Aerosol layer height	L2__AER_LH	KNMI
Ultra violet aerosol index	L2__AER_AI	KNMI
O_3 Full Profile	L2__O3__PR	KNMI
O_3 Tropospheric Profile	L2__O3_TPR	KNMI
NO_2	L2__NO2___	KNMI
CO	L2__CO____	SRON/KNMI
CH_4	L2__CH4___	SRON/KNMI

In this research the period of time defined is from January 2020 to June 2021, the selected product of Level-2 is Nitrogen Dioxide (L2__NO2___), and the territory of Brazil is defined as an area geographic area of study. The result is a set of files (.nc), each of which represents a satellite scan range. The information within the files has a structure of groups and layers to structure the data needed in the experiment. Figure 3 shows a ".nc" generic file description for a Level-2 file.

IV. IMPLEMENTATIONS

As described in the previous section, the data provided by Sentinel-5 has a structure that needs to be correctly consumed. In addition to this particularity, the time period needs to be considered when evaluating the NO_2 rates in the air, Also, considering the (huge) territory of Brazil as a geographic area of study, this implies adding a cartographic process. Thus, our basic architecture consists of a workflow with three steps, Dataset Processing, Polygonal Mask Generation, and Region Segmentation, which has been defined as illustrated in Figure 4.

A. Dataset Processing

Each file downloaded from Sentinel-5 has an approximate size of 320MB. Considering the huge extension of Brazil, as a country, each day of scanning can contain up to 3 files. As a result of working with the information that Sentinel-5 generates every day, it is expected that the data size reaches the terabyte (TB). As the purpose of the experiment is to consider only the NO_2 indices, an information extraction process is carried out together with the temporal processing. The result of this first step is a set of smaller files, which are easy to use and with data from NO_2 in chronological order.

B. Polygonal Mask Generation

When defining a geographic area in the API requests, the information returned by the Copernicus Open Access Hub website refers to the area that is contained in the complete reading range that Sentinel-5 forms around the earth. Based on Sentinel-5's processing pipeline, it is necessary to establish a cartographic analysis of Brazil, in order to be able to delimit and segment polygonal regions independent of each

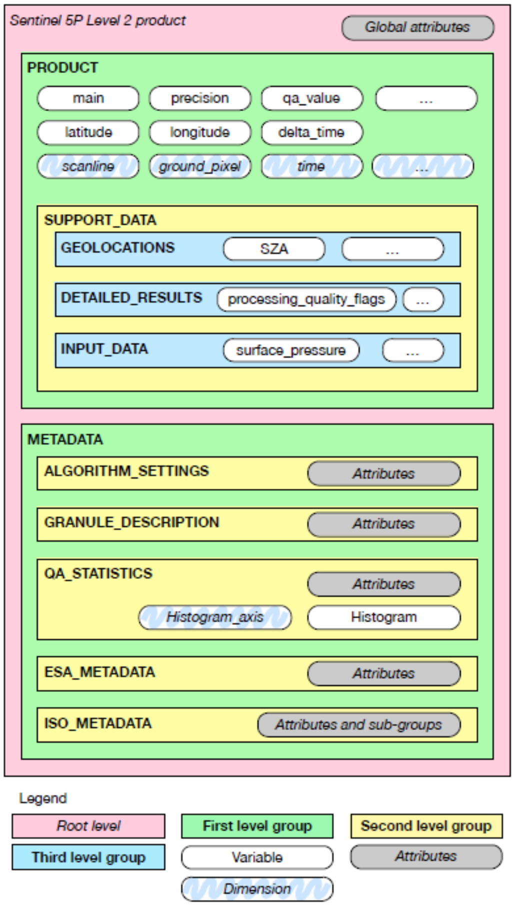

Fig. 3. Description of the generic structure of Level-2 files (.nc) [9]

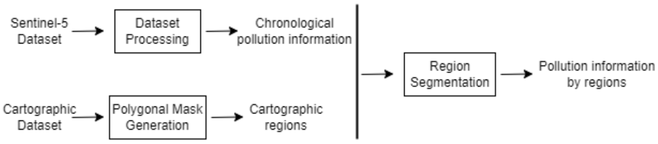

Fig. 4. Workflow of our proposed methodology.

other. In this research, 27 polygonal masks are segmented, corresponding to the 27 states of Brazil.

C. Region Segmentation

The last step is to use the results of the two previous steps as input. The NO_2 indexes are chronologically structured with punctual information, and the masks are delimited as geographic polygons. Then, the Sentinel-5 data are segmented and, thus, it is possible to get the NO_2 indexes for each Brazil's region.

V. Experiments and Results

In order to show a visual result based on the devised workflow described above, we first restrict the day's interval to 6 days, i. e., it was considered a data range between the 1^{st} and 6^{th} of January 2019. As the area of interest, we used the polygons masks inside the cartographic data of the Brazilian states Maranhão and Ceará. At the end of the processing flow, wee obtained a mask of NO_2 indexes with unity in a number of molecules per cm^2, as shown in Figures 6 and 7.

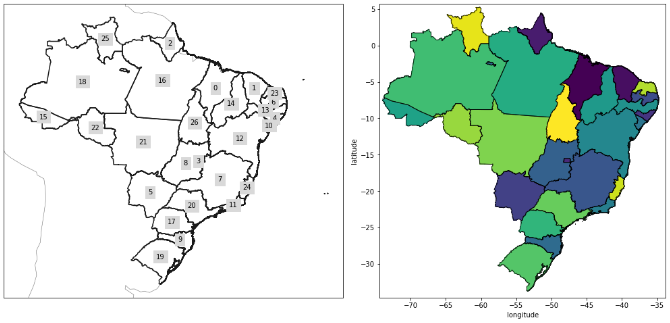

Fig. 5. Polygonal masks

Statistical operations of mean, median, and standard deviation could be performed to evaluate the results obtained from our first implementation, based on the time interval used to acquire data. Table II shows the media of NO_2, within a 6 days interval, for the same two states presented before.

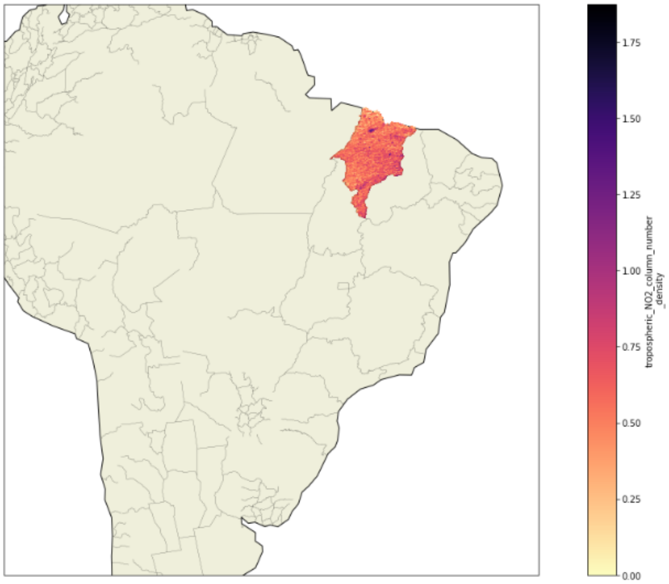

Fig. 6. NO_2 pollution in the Maranhão region

TABLE II
MEAN OF INDICES NO_2

Region	Identifier	Time (days)	Molecules/cm^2
Maranhão	L2__NO2___	6	0.4395276184343702
Ceará	L2__NO2___	6	0.6643574976584711

VI. Conclusion

As seen, air quality indices are currently measured by some organisms using technology based on devices composed of

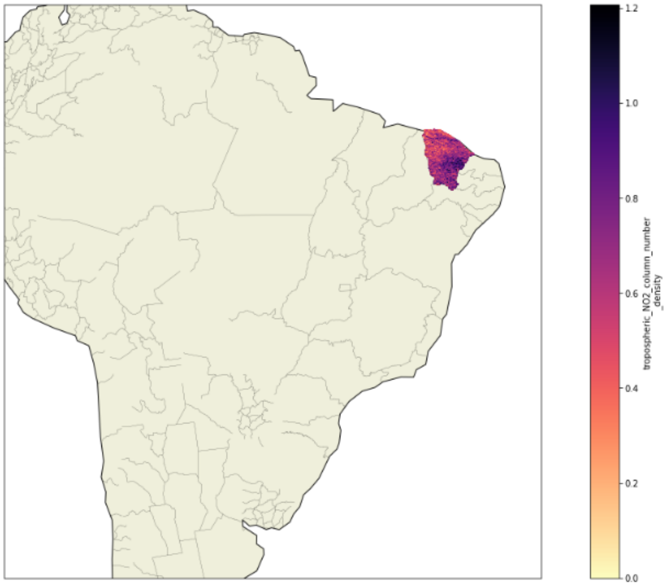

Fig. 7. NO_2 pollution in the Ceará region

sensors at fixed points in cities. These measures are accurate, however, most cities in Brazil, for example, do not have this kind of technology available. Towards solving this problem of data completion, in this work, we study the possibility of using another type of methodology that is available, which is the high-resolution spectrometer system installed on satellites that constantly monitor the surface of the planet. In this case, we use data provided by the Sentinel-5 satellite, which is an orbital system that generates a large amount of data for climate and environmental monitoring.

Hence, to this end, we have shown how data from the Sentinel-5 can help to increase reliability in the data quality, completing it where there is a lack of more accurate sensor devices. Actually, this parameter will be used by our data scientists for predicting the dynamics prediction of Covid-19, increasing confidence in their works [2].

In future work, we intend to study how interpolation and extrapolation techniques can be used in order to devise a more reliable value when a finer granularity is necessary for the data. We believe that using splines or other mathematical surfaces could play an important role in this process. Indeed, the use of deep learning is another option that can be studied for this.

References

[1] I. G. Pereira, J. M. Guerin, A. G. S. Júnior, G. S. Garcia, P. Piscitelli, A. Miani, C. Distante, and L. M. G. Gonçalves, "Forecasting covid-19 dynamics in brazil: A data driven approach," *International Journal of Environmental Research and Public Health*, vol. 17, pp. 1–26, 7 2020.

[2] D. P. Aragão, E. V. Oliveira, A. A. Bezerra, D. H. dos Santos, A. G. da Silva Junior, I. G. Pereira, P. Piscitelli, A. Miani, C. Distante, J. S. Cuno, A. Conci, and L. M. Gonçalves, "Multivariate data driven prediction of covid-19 dynamics: Towards new results with temperature, humidity and air quality data," *Environmental Research*, vol. 204, p. 112348, 2022.

[3] P. Fermo, B. Artíñano, G. De Gennaro, A. M. Pantaleo, A. Parente, F. Battaglia, E. Colicino, G. Di Tanna, A. Goncalves da Silva Junior, I. G. Pereira, G. S. Garcia, L. M. Garcia Goncalves, V. Comite, and A. Miani, "Improving indoor air quality through an air purifier able to reduce aerosol particulate matter (pm) and volatile organic compounds (vocs): Experimental results," *Environmental Research*, vol. 197, p. 111131, 2021.

[4] A. Q. Scale, "World's air pollution: Real-time air quality index," 2022. Online resource, available at https://waqi.info/, accessed on 08.10.2022.

[5] W.-A.-P. Project, "Air quality index (aqi) and pm2.5 air pollution," 2022. Online resource, available at https://www.iqair.com/, accessed on 08.10.2022.

[6] R.-T.-A.-Q.-I. Project, "Air pollution: Real-time air quality index (aqi)," 2022. Online resource, available at https://aqicn.org/city/vietnam/hanoi/, accessed on 08.10.2022.

[7] W.-A.-Q.-I. Hanoi, "Hanoi air quality index (aqi) and vietnam air pollution," 2022. Online resource, available at https://www.iqair.com/, accessed on 08.10.2022.

[8] NASA, "Health and air quality data pathfinder - open access for open science earth data," 2022. Online resource, available at https://www.earthdata.nasa.gov/learn/pathfinders/health-and-air-quality-data-pathfinder, accessed on 08.10.2022.

[9] The-European-Space-Agency, "Sentinel-5," 2022. Online resource, available at https://sentinel.esa.int/web/sentinel/missions/sentinel-5, accessed on 08.10.2022.

[10] A. A. Bezerra, D. H. dos Santos, and L. M. G. Gonçalves, "Real-time air quality monitoring using optical sensors to prevent severe acute respiratory syndromes," in *Latin America Optics And Photonics Conference*, vol. 1, (Recife), OSA, Optica Publishing Group - Formerly OSAA, August 2022.

[11] U.S.-Environmental-Protection-Agency, "Air quality index (aqi)," 2022. Online resource, available at https://www.airnow.gov/aqi-and-health/, accessed on 08.10.2022.

[12] H. Li, X.-L. Xu, D.-W. Dai, Z.-Y. Huang, Z. Ma, and Y.-J. Guan, "Air pollution and temperature are associated with increased covid-19 incidence: A time series study," *International Journal of Infectious Diseases*, vol. 97, pp. 278–282, 2020.

[13] S. Lokhandwala and P. Gautam, "Indirect impact of covid-19 on environment: A brief study in indian context," *Environmental Research*, vol. 188, p. 109807, 2020.

[14] B. M. Hashim, S. K. Al-Naseri, A. Al-Maliki, and N. Al-Ansari, "Impact of covid-19 lockdown on no2, o3, pm2. 5 and pm10 concentrations and assessing air quality changes in baghdad, iraq," *Science of the Total Environment*, vol. 754, p. 141978, 2021.

[15] R. P. Singh and A. Chauhan, "Impact of lockdown on air quality in india during covid-19 pandemic," *Air Quality, Atmosphere & Health*, vol. 13, no. 8, pp. 921–928, 2020.

[16] J. Veefkind, I. Aben, K. McMullan, H. Förster, J. de Vries, G. Otter, J. Claas, H. Eskes, J. de Haan, Q. Kleipool, M. van Weele, O. Hasekamp, R. Hoogeveen, J. Landgraf, R. Snel, P. Tol, P. Ingmann, R. Voors, B. Kruizinga, R. Vink, H. Visser, and P. Levelt, "Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications," *Remote Sensing of Environment*, vol. 120, pp. 70–83, 2012. The Sentinel Missions - New Opportunities for Science.