The Impact of Contingency Measures on the COVID-19 Reproduction Rate

Regina Sousa, Daniela Oliveira, Francini Hak, and José Machado(✉)

ALGORITMI/LASI, University of Minho, Braga, Portugal
{regina.sousa,daniela.oliveira,francini.hak}@algoritmi.uminho.pt,
jmac@di.uminho.pt

Abstract. The SARS-CoV-2 virus had a major impact on the health of the world's population, causing governments to take progressively more cautious measures. All of these measures took into account the pandemic situation in the region in real time, with the aim of slowing down the spread of the infection as much as possible and reducing the associated mortality. This article aims to study the impact of preventive measures on the spread of COVID-19 and the consequent impact on excess deaths. In order to obtain the results presented, Big Data techniques were used for data storage and processing. As a result it can be concluded that Gross Domestic Product (GDP) is directly proportional to the Human Development Index (HDI), Higher GDP per capita are associated with a higher number of new cases of COVID-19 and R-index is inversely proportional to the severity of the contingency measures.

Keywords: Covid-19 · Contigency Measures · Proliferation rate · Correlation · Big Data Analysis · Spark · PowerBI

1 Introduction

In March 2020, the World Health Organization (WHO) declared a worldwide pandemic derived from the new class of Coranavirus. Contingency measures were adopted, changing face-to-face activities to a virtual format, depriving some habits, closing local businesses and forcing people to confine themselves at home for months [4]. Being a novelty, no country knew how to properly deal with the pandemic and which measures would be most effective, always taking into account other concerns, such as the country's economy and people's mental health [7].

A consequence of the pandemic was the daily production of data, that grew exponentially. Big Data techniques were applied to treat and analyze the data in the proper way, in order to generate knowledge to support decision-making processes [11]. Thus, it was possible to monitor in real time the evolution of cases of infection and deaths caused in each country [5]. In this sense, this article aims to apply Big Data techniques to verify whether or not the contingency measures contributed to reducing the spread of the Covid-19 virus.

ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023

J. M. Machado and H. Peixoto (Eds.): AISCOVID 2022, LNICST 485, pp. 28–37, 2023.
https://doi.org/10.1007/978-3-031-38204-8_3

The paper is divided into five sections. First, an introduction is made in which the reader is informed about the theme and purpose of the document. The second section provides context for the topic discussed. Following that, the study's materials and methods are described. The results are presented in Sect. 4. Finally, conclusions are reached, as well as next steps for future work.

2 Background

With the accelerating adoption of Information Technologies (IT), the amount of data produced daily has been increasing exponentially [12]. Traditional data storage and management processes are no longer enough. In this sense, the term Big Data emerged, capable of extracting, processing and analyzing large amounts of data in real time, dealing with complex and different data structures [10].

Big Data is characterized by 5 Vs, such as: (i) volume in relation to the amount of data collected; (ii) velocity defined by the time of processing and manipulation of the data; (iii) variety of acquired data types; (iv) veracity to the quality, reliability and accuracy of the data provided; (v) value obtained with the collected data [10].

Data analysis is recognized by the ability to transform raw data into knowledge, in order to support the decision-making. This process is usually divided into three phases: Storage, Processing and Visualization [12]. As such, there are several tools destined to each phase of the process, however, some can run more than one phase. Some commonly used tools in the storage phase are MongoDB, Apache Cassandra and CouchDB. For the processing phase there are tools like Apache Spark, Apache Hadoop and Samza. For the last phase we can find tools such as PowerBI, Tableau and Stat iQ.

Big data has proven to be essential in the healthcare sector, being able to handle a huge amount of data to provide real-time monitoring of epidemic outbreaks [2]. In relation to previous outbreaks of epidemics and pandemics, COVID-19 is unprecedented, so it was necessary that datasets be created and made available as open access, for possible analysis of the daily numbers of new infections broken down by country or cities [4]. Thus, computational techniques allowed us to visualize the spread of the virus in real time, have an accurate detection of infected patients, and obtain an optimized contact tracing. However, there are still points that need to be analyzed and this article aims to analyze whether the application of contingency measures in fact contributed to reducing the spread of Covid-19.

3 Materials and Methods

The main goal of this research work is to study the impact that imposed containment measures have had on the proliferation rate of COVID-19 infection, worldwide. To do this, public datasets were used with data on the rates of measures applied in different geographic areas over time, as well as data on COVID-19 indicators in each region and their reproduction rate. According to the defined

case study, the following research question emerged, which this work aims to answer:

How did government measures during COVID-19 affect the reproduction rate?

3.1 Big Data Architecture

Given the high amount of data for processing, it was necessary to develop an architecture subdivided by four distinct phases, the first being the choice and fusion of the various data sources, whose output will be the input of the next phase, called Storage. The processed data will then go through the central phase of the research, which is the data processing and its respective treatment using the Apache Spark tool. In order to be able to draw conclusions and construct some Business Intelligence indicators, the Visualization phase was planned, using the PowerBI tool. The architecture present in the Fig. 1 meticulously demonstrates all the processes.

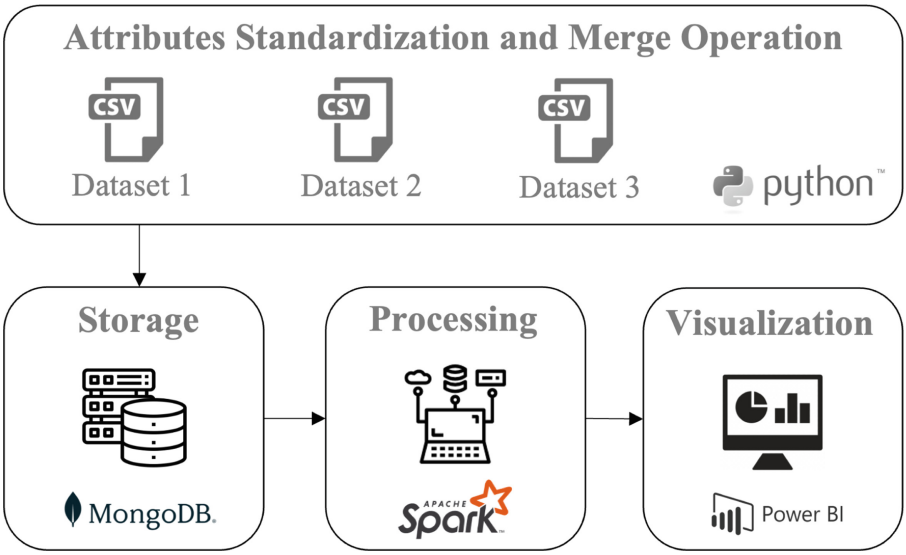


Fig. 1. Big Data pipeline designed and its phases implemented.

Data Collection and Storage: To answer the case study correctly, 3 public datasets were chosen and will be presented below. All datasets are in CSV format although with different time horizons.

– **First Dataset - WHO COVID-19 Global Data:** The WHO COVID-19 Global Data dataset contains information associated with the COVID-19 pandemic in each of 236 countries from January 3, 2020, to March 1, 2022. It consists of 8 attributes that serve as the basis for subsequent datasets, with information such as the number of new COVID-19 cases per day and cumulative,

and the number of COVID-19 deaths per day, for each country, its code, and its respective region for the World Health Organization (WHO) [8].

- **Second Dataset - COVID-19 Government Response Tracker:** This dataset consists of 6 attributes that are intended to represent the degree of severity associated with the government measures imposed in each country. Each observed country is associated with its daily percentage index of health and confinement for exposure, called ***Containment Health Index For Display***, and its health and confinement index designated on a scale of 0 to 3. The time horizon for this dataset starts on January 1, 2020 and ends on February 3, 2021 [3].
- **Third Dataset - COVID-19 by *Our World in Data*:** This information source is composed of 67 attributes referring to different indicators evaluated against the impact of COVID-19, such as confirmed cases and deaths, excess mortality, hospital and ICU occupation, policy responses, tests and positivity, vaccinations, and reproduction rate. In this case study, special attention was given to the ***Reproduction Rate*** (**R** indicator) indicator, which estimates in real time the effective reproduction rate of COVID-19. The ***Location*** and ***Date*** attributes were also necessary to be able to join with the other datasets chosen [1].
- **Final Dataset Acquisition:** To get to the final dataset, the three data sets described above were joined using a script developed in Python using the Pandas library. This library is one of the most used today for the treatment of data with different types of data, allowing simple and effective manipulation [9]. The merge operation of the different datasets was done through the attributes ***Location*** and ***Date***, which are present in both. To do this, it was necessary to first rename these two attributes in all the datasets so that they had the same designation. Additionally, it was necessary to format all dates in the *yyyy-mm-dd* format. Lastly, the final dataset obtained is uploaded into a Mongo database.

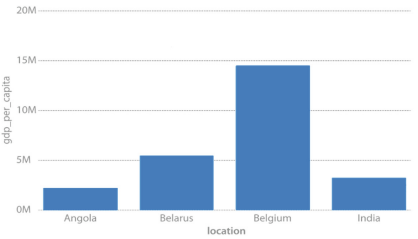
Data Processing: An Extraction, Transformation, and Loading (ETL) process was developed, where data stored in the Mongo database was obtained (*Extract*) to be properly treated (*Transformation*). Using the PySpark library, the transformation of this data was performed. Finally, the resulting data was saved in a Databricks table to be able to be queried by a data visualization platform (*Loading*). The data processing phase was subdivided into the following tasks:

- **Redundant columns removal:** After an analysis of the dataset, the removal of some redundant and duplicate columns that had no relevance to the case study was performed;
- **Removal of new negative COVID-19 case records:** Removal of records with negative new COVID-19 cases: Records whose number of new daily COVID-19 cases had a negative value were removed;
- **Removal of R indicator null values:** For the primary purpose of analyzing the R indicator, records with a reproduction rate of null have been removed.

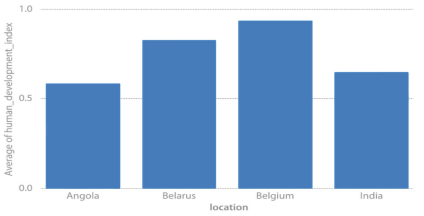
Data Visualization: Data visualization is the last stage of the proposed architecture, ending with the graphical representation of information and the creation of dashboards to analyze the impact of government measures on the COVID-19 reproduction rate. The tool chosen for information visualization was PowerBI and allows connection to other data sources such as Spark and Azure Databricks [6]. A connection was made to the cluster created on the Databricks platform to query the data resulting from the processing step described above.

4 Results

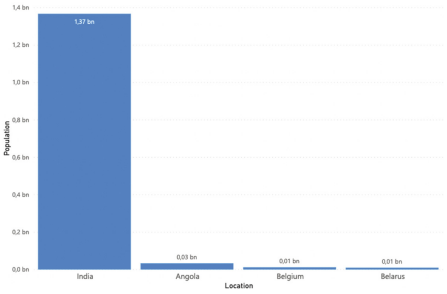
Taking into account the previously described architecture, all the preprocessing and data manipulation culminated in dashboards, built in PowerBI. Therefore, some of the components of the final dashboard will be analyzed and discussed in the following section. Initially, bar charts were generated (one to verify the gross domestic product (GDP) per capita by location and another for the human development index also by location).



(a) Gross domestic product per capita per location.



(b) Human development index per location.



(c) Number of Population per location.

Fig. 2. Initial feature analysis.

Analyzing the results, the Fig. 2a shows the 4 selected countries. This selection was made taking into account its representativeness, presenting two countries with a low per capita GDP (Angola and India), a country with a medium GDP value (Belarus) and a country with a high GDP value (Belgium). It should also be noted that two countries with low GDP per capita were selected so that it was possible to relate the amount of population to the remaining parameters, considering for Angola a population of 32 million [13] and for India of 1.3 billion [14]. The Fig. 2b shows the relations between the human development index by country, confirming that although all of them have relatively high rates, Angola and India are a little below average.

On the other hand, the Fig. 3 represents the evolution of the virus proliferation rate (R) for each country, in the time interval of the collected data. Through its analysis, it can be concluded that countries with higher values for the R indicator also have a higher Gross Domestic Product (GDP) per capita and a higher human development index, which may also be related to the fact that their populations are older. However, the R-indicator of India, which has a significantly lower GDP per capita, is also high due to the fact that this country maintains a high containment health index when compared to the number of new COVID-19 cases, this correlation can be analyzed in Fig. 3 and Fig. 4.

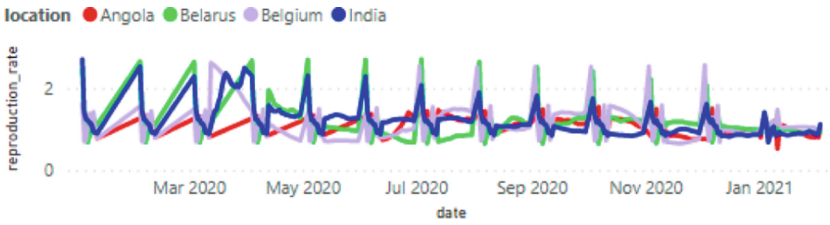


Fig. 3. Virus proliferation rate, over time.

Now presenting the correlations between various parameters, the relationship between Containment Health For Display (severity of containment measures) and the percentage of new cases per country is illustrated in Fig. 4. In this figure it can be concluded that the rate of contingency measures in most countries is directly proportional to the average number of cases. Only in India, the rate of contingency measures is higher than the number of cases, which is justified by the Fig. 3, where it is shown that the number of cases decreased probably due to the level of restrictions imposed over a long period of time.

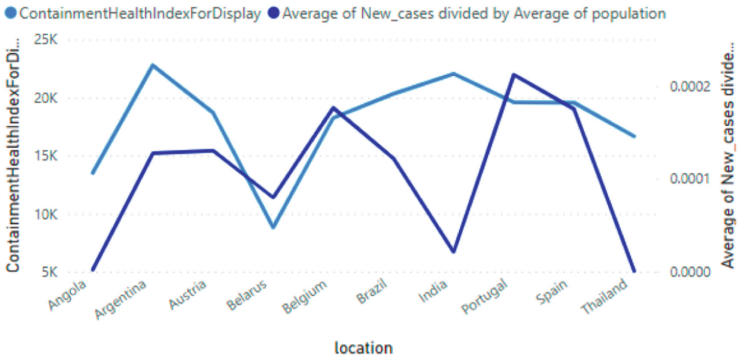
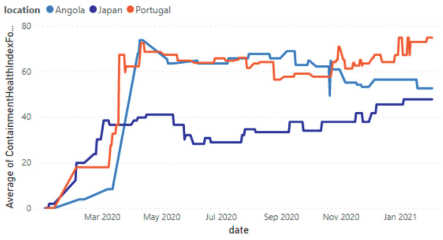
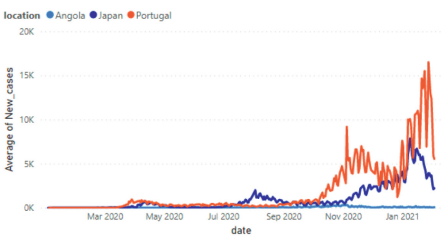


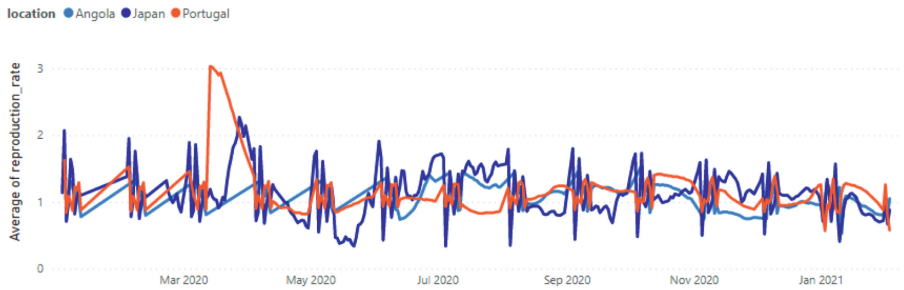
Fig. 4. Evolution of severity of contingency measures with the number of new cases, per location.



(a) Average Contingency Measure Index over time per location.



(b) Average number of new cases over time per location.



(c) Variance of the average virus proliferation rate over time per location.

Fig. 5. Comparative analysis between the means of the contingency measures, number of new infection cases COVID-19 and R-indicator.

A detailed analysis of the graphs presented in Fig. 5a, Fig. 5b and Fig. 5c shows that the containment measures emerged in response to the high probability of infection (proliferation rate).

An example is Portugal where, when the reproduction rate and new cases rose sharply, the severity of the containment measures increased considerably. Here, too, we can see that the containment measures affect the reproduction

rate, because when the measures were kept at a high level, the proliferation rate was somewhat contained.

Moreover, comparing Japan and Portugal, the higher value of the reproduction rate of the former is about half of the latter, this is due to the fact that Japan implemented containment measures before Portugal, and thus the probability of infection was much lower.

Overall, analyzing the Fig. 6, it can be seen that the containment measures emerged in response to a high proliferation rate. It is further concluded that stabilization of the containment measures at high levels provided stabilization of the proliferation rate at relatively low levels.

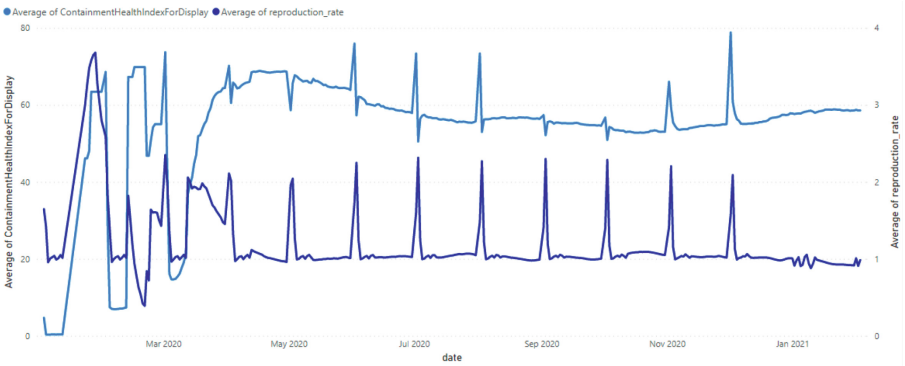


Fig. 6. Overall assessment of the average of the containment measures against the average of the R-indicator.

5 Conclusions and Future Work

A Big Data architecture was developed with emerging technologies in the market, such as Apache Spark, Azure Databricks, NoSQL databases such as MongoDB, and PowerBI, with the purpose of analyzing the impact of government measures against the proliferation rate of COVID-19.

The GDP was analyzed in relation to the population of each country and its Human Development Index (HDI). It was found that the GDP is directly proportional to the HDI, as opposed to the value of each country's population, as is the case of Angola and India. It was also concluded that countries with a higher GDP per capita are associated with a higher number of new cases of COVID-19, which is related to the existence of better access to health care and, consequently, a higher number of tests performed. Another premise is that air transport is also more abundant in these countries, so the COVID-19 contamination of travelers is directly proportional.

The severity index of contingency measures applied in each country was found to increase in the face of high R-indicator values and high numbers of new cases

of COVID-19 infection. The R-indicator is therefore inversely proportional to the severity of the contingency measures.

Therefore, it can be concluded that the application of contingency measures influenced the pandemic situation in each country differently. There are countries that used contingency measures as prevention, and in these cases a decrease in the number of cases is more effective.

On the other side, there are countries that used contingency measures as a response to the increase in the number of cases. In this last scenario the control of the pandemic was not as effective. The graphs that best illustrate this phenomenon are shown in Figs. 5a, 5b and 5c.

Acknowledgements. This work has been supported by FCT-Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. The grant of Regina Sousa is supported by the project "Integrated and Innovative Solutions for the well-being of people in complex urban centers" within the Project Scope NORTE-01-0145-FEDER-000086. Francini Hak thanks the Fundação para a Tecnologia (FCT) for the grant 2021.06230.BD.

References

1. Appel, C., et al.: Data on covid-19 (coronavirus) by our world in data. https://github.com/owid/covid-19-data/tree/master/public/data. Accessed 01 May 05 2022
2. Bragazzi, N.L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., Wu, J.: How big data and artificial intelligence can help better manage the covid-19 pandemic. Int. J. Environ. Res. Public Health **17** (2020). https://doi.org/10.3390/ijerph17093176
3. of Government, B.S.: Oxford covid-19 government response tracker. https://static-content.springer.com/esmart%3A10.1038%2Fs41562-021-01079-8/MediaObjects/41562_2021_1079_MOESM4_ESM.xlsx. Accessed 01 May 05 2022
4. Hak, F., Abelha, A., Santos, M.: Open science in pandemic times: a literature review. vol. 177, pp. 552–555. Elsevier B.V. (2020). https://doi.org/10.1016/j.procs.2020.10.077
5. Leung, C.K., Chen, Y., Shang, S., Deng, D.: Big data science on covid-19 data, pp. 14–21. Institute of Electrical and Electronics Engineers Inc., December 2020. https://doi.org/10.1109/BigDataSE50710.2020.00010
6. Microsoft: Tutorial: Introdução ao serviço de criação no power bi. microsoft documentation. https://docs.microsoft.com/pt-pt/power-bi/fundamentals/service-get-started, July 2020, Accessed 16 May 2022
7. Oliveira, D., et al.: Management of a pandemic based on an openehr approach. Procedia Comput. Sci. **177**, 522–527 (2020)
8. Organization, W.H.: Who coronavirus (covid-19) dashboard — who coronavirus (covid-19) dashboard with vaccination data. https://covid19.who.int/data, Accessed 01 May 2022
9. Pandas: Pandas - python data analysis library. https://pandas.pydata.org/, Accessed 01 May 2022
10. Sagiroglu, S., Sinanc, D.: Big data: a review, pp. 42–47 (2013). https://doi.org/10.1109/CTS.2013.6567202

11. Sousa, R., Lima, T., Abelha, A., Machado, J.: Hierarchical temporal memory theory approach to stock market time series forecasting. Electronics **10**(14), 1630 (2021)
12. Tsai, C.-W., Lai, C.-F., Chao, H.-C., Vasilakos, A.V.: Big data analytics: a survey. J. Big Data **2**(1), 1–32 (2015). https://doi.org/10.1186/s40537-015-0030-3
13. Worldometer: angola population live (2022). https://www.worldometers.info/world-population/angola-population/. Accessed 16 May 2022
14. Worldometer: India population live (2022). https://www.worldometers.info/world-population/india-population/. Accessed 16 May 2022