Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

COVID-19 spatiotemporal research with workflow-based data analysis

Srikar Chintala [a,*,1], Ritvik Dutta [b,1], Doron Tadmor [c,1]

[a] University Preparatory Academy, San Jose, CA 95125, USA
[b] Monta Vista High School, Cupertino, CA 95014, USA
[c] Imperial College Business School, Imperial College London, London SW7 2BU, UK

ARTICLE INFO

Keywords:
COVID-19
Coronavirus
Dynamic maps
KNIME
Spatial visualization
Visual analysis

ABSTRACT

Given the pertinence and acceleration of the spread of COVID-19, there is an increased need for the replicability of data models to verify the veracity of models and visualize important data. Most of these visualizations lack reproducibility, credibility, or accuracy, and are static, which makes it difficult to analyze the spread over time. Furthermore, most current visualizations depicting the spread of COVID-19 are at a global or country level, meaning there is a dearth of regional analysis within a country. Keeping these issues in mind, a replicable, efficient, and simple method to generate regional COVID-19 visualizations mapped with time was created by using the KNIME software, an open-source data analytics platform that can create user-friendly applications or workflows. For this analysis, Albania, Sweden, Ukraine, Denmark, Russia, India, and Australia were closely observed. Among the maps generated for the aforementioned countries, it was noticed that regions with a high population or high population density were often the epicenters within their respective country. The regions caused the virus to spread to their neighboring regions: kickstarting the "domino effect", leading to the infection of another region until the country is overwhelmed with cases—what we call a proximity trend. These dynamic maps are crucial to fighting the pandemic because they can provide insight as to how COVID-19 spreads by providing researchers or officials with an accurate and insightful tool to aid their analysis. By being able to visualize the spread, health and government officials can dive deeper to identify the sources of transmission and attempt to stop or reverse them accordingly.

1. Background

Infecting seventy million people and causing over a million deaths (WHO Coronavirus Disease (COVID-19) Dashboard, 2020), COVID-19 has proven to be a serious threat to human health, production, life, social functioning, and international relations. Given the virus's profound negative effects on the world through the many forced lockdowns, fighting back against it is of the most critical importance. In the fight against COVID-19, big data technologies have played an important role in many aspects, including the rapid aggregation of multi-source big data, rapid visualization of epidemic information, and spatial tracking of confirmed cases, among others (Zhou et al., 2020). However, the main challenge is finding strategies to adjust traditional data analysis methods and improve the speed and accuracy of the information provided.

Concerning the ongoing pandemic, we are facing an "infodemic": a mass release of COVID-19 related information making it difficult to navigate through and understand. Because of this "infodemic", researchers are presented with 3 major issues. First, the data's availability and presentation is often very inaccurate. Second, data is often fragmented into static snapshots, preventing the viewer from seeing the bigger picture of cases over time. Many COVID-19 tracking maps currently exist; unfortunately, none of them can present meaningful, time-frame independent data. Consequently, researchers cannot perform an accurate and insightful analysis of the spread of the virus over time. Third, most data visualizations are very inefficient, making it difficult to create multiple visualizations with one template and impractical to replicate by other researchers. Keeping each of the aforementioned issues in mind, a method to successfully create accurate and insightful visualizations was developed. To ensure accuracy within this method, data from a reputable source was used. To make this method as efficient as possible, KNIME, a workflow analysis tool, was used to create a template file that could be edited with user parameters to easily replicate the process for multiple different countries. Furthermore, most visualizations regarding country cases of COVID-19 are at a

[*] Corresponding author at: University Preparatory Academy, San Jose, CA 95125, USA.
E-mail addresses: schintala@upatoday.com (S. Chintala), rdutta596@student.fuhsd.org (R. Dutta), doron.tadmor20@imperial.ac.uk (D. Tadmor).
[1] Contributed equally.

https://doi.org/10.1016/j.meegid.2020.104701
Received 31 August 2020; Received in revised form 24 December 2020; Accepted 28 December 2020
Available online 31 December 2020
1567-1348/© 2020 Elsevier B.V. All rights reserved.

national level. In other words, there is a dearth of visualizations available for regional and provincial impacts of COVID-19 within any given country, except the United States. Currently, there is plenty of evidence suggesting that effective maps can aid the prevention of viral spread.

Given the pertinence and acceleration of COVID-19, this study focused on creating replicable and efficient workflows that create visualizations to understand the spread of COVID-19, verify the accuracy of other models, promote knowledge and information sharing, create research collaborations with other researchers, develop pilot studies on COVID-19 for future research, cultivate professional data analysis.

2. Material and methods

2.1. Data

The data source used within this study includes the National Science Foundation's and the Spatiotemporal Innovation Center's (STC) data on Github, as it is reliable and constantly updated. STC conducts world-class research through partnerships among academic institutions, national laboratories, industrial organizations, and other public/private entities, and via international collaborations. On Github, STC provides data on twenty-four countries, which includes confirmed, death, and recovered cases within each country. This research focused on seven of the twenty-four – Albania, Sweden, Ukraine, Russia, India, Denmark, and Australia. With further investigation of the data, there is no need to modify it in any way, attesting to the STC's reliability in producing such data.

2.2. Software tools

First, a platform was needed to be chosen for the base of the method created within this study. KNIME, R, and Rapidminer were all initially chosen as possible candidates due to their widespread usage in data analytics. Rapid Miner is an integrated environment for machine learning and predictive analysis based on the Java programming language that provides a powerful graphical user interface. However, it is most suited for those needing to work with database files because the software requires the user to manipulate SQL statements and files. R is an open-source programming language that can provide quality representation of data in the form of charts, plots, and mathematical equations. However, the language does not provide a user-friendly environment as it requires the user to iterate through every line of code and make edits to the code itself to make it work for different inputs, thus demonstrating a steep learning curve. The last two mentioned coding platforms are very complex to learn and difficult to utilize if the user presents a lack of knowledge of those topics. Among all, KNIME was chosen due to its various built-in components for machine learning and data mining. KNIME was initially developed for pharmaceutical research, but now it is used in many different or diverse fields like business intelligence, business forecasting, financial analysis (Chahal and Gulia, 2016). KNIME also allows Python, R, and WEKA to be integrated within its workflows. These built-in components made the workflow more efficient by replacing entire blocks of code with KNIME commands. These commands are optimized to work as fast as possible and by labeling the built-in command with its purpose, the workflow can provide those with limited coding knowledge a basic understanding of how the workflow works. The works of Pynam et al. (2018) and Ranjan and Agarwal (2017) corroborate this conclusion.

2.3. Workflow

This workflow requires two inputs: provincial COVID-19 data and a base map file of the country the user wishes to study, as shown in Fig. 1. This workflow can create multiple outputs: the map in a .html file format or a .gif file format. The workflow automatically generates a chart that showcases the changes in the number of cases over time. Each of the options within the workflow is shown in Fig. 2.

Each map's creation requires a separate workflow within the KNIME workspace as well as its data which were obtained from the Spatiotemporal Innovation Center's Github. Next, mapshaper.com was used to simplify the country .geojson file. There, the .geojson file was loaded onto the editor by unchecking "detect line intersections". After loading the file, the "Visvalingam/weighted area" feature was used to simplify the map to 5% of the original format. Mapshaper.com's simplification process shrinks .geojson files by unnecessary map detail and exports it as a .json file. The .json file must be opened from the system file manager and saved as a .geojson file. After the import of the workflows, the downloaded data must now be loaded on to the workflow. To do this, the "File Reader" node must be clicked and the path of desired .csv, which contains the COVID-19 data to be analyzed, must be entered. Afterward, more data must be loaded into the "Table Creator" node which demands the path of the simplified .geojson file, two integer inputs for the longitude and latitude of the centroid of the country to be analyzed, a string that to title the map with, the file path where the workflow outputs the .html map, the default zoom value of the map, two boolean values: one that enables the output of a .html map and one that enables the output of a .gif map, and the integer input for the number of days to increment by within the data. Setting an increment of one would take every day in the dataset and so on. Subsequently, the "Unpivoting" node was used to reformat the data frame so that each date, which was originally its column in the .csv file, is now its data point under a new column called "Dates". This was done by ensuring that the "Include" value columns only had dates, and "Retailed" columns only had "hasc" and "county". Thereafter, the "Column Rename" and "Column Filter" nodes must be checked to make sure that there are no errors with the new data frame's columns. From there, the entire workflow can be run

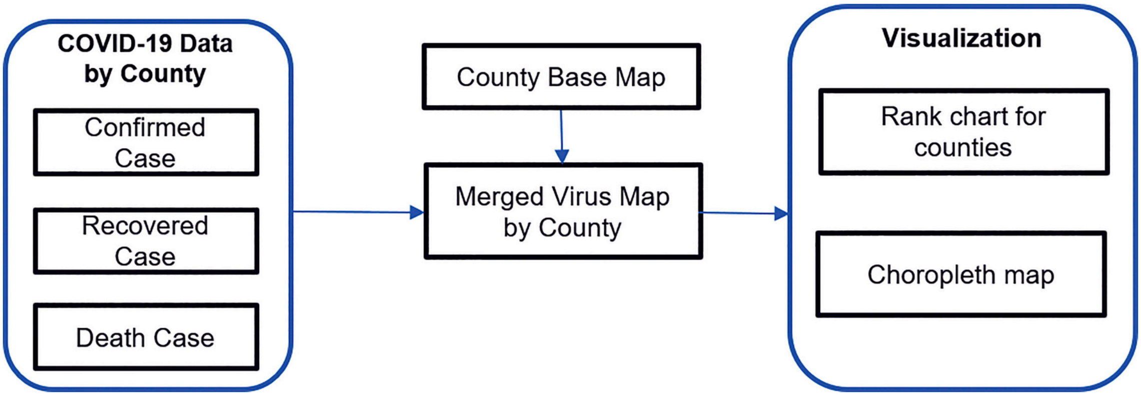

Fig. 1. This represents the general flowchart we used, or in other words, the general methodology we used. It required using the COVID-19 case data combined with a base map of a country. Then, you would use a data processing and visualization application to create a dynamic map of the country.

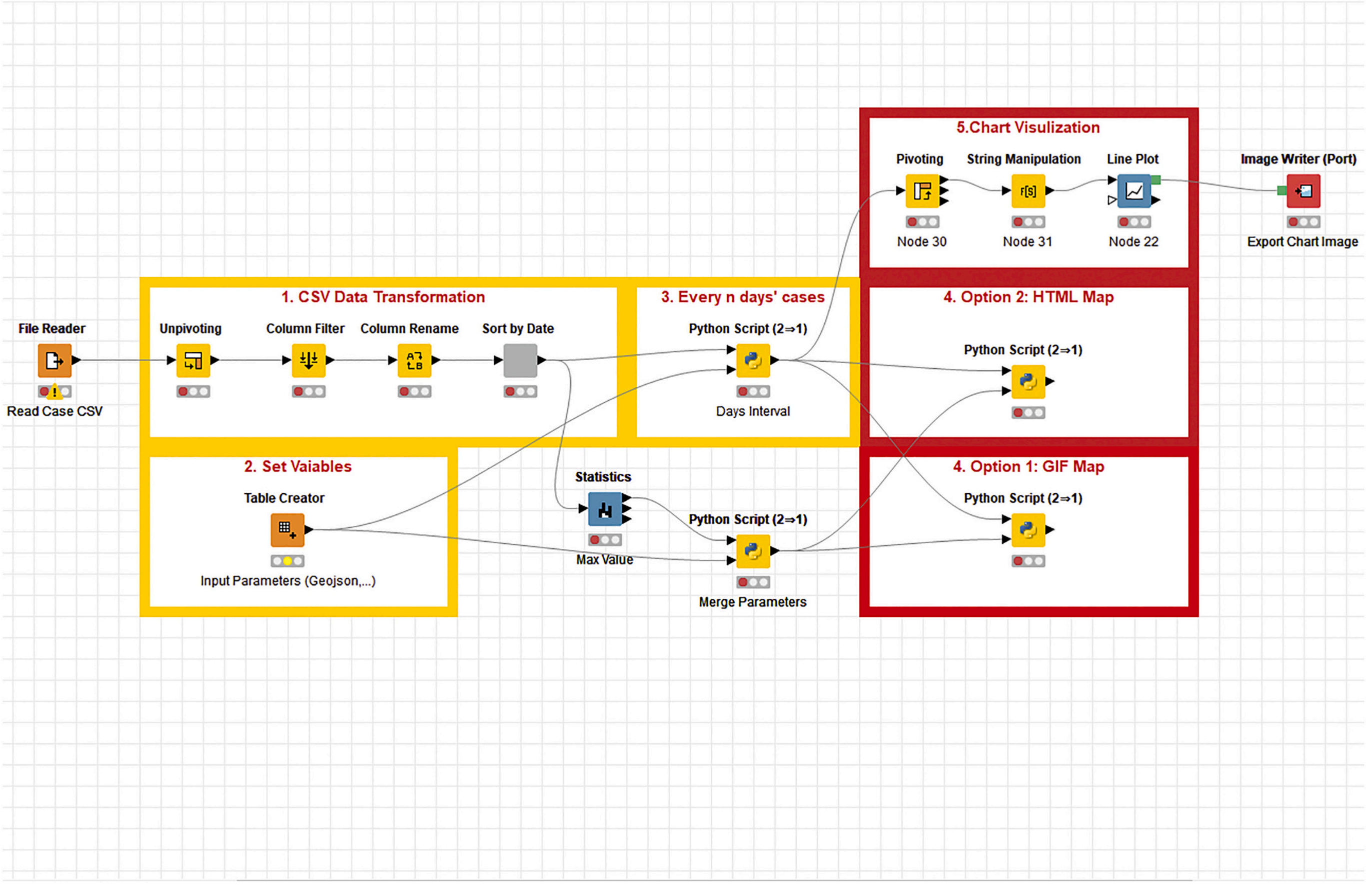

Fig. 2. This is the workflow that we created in KNIME to create our choropleth maps for the various countries while utilizing Python. We added the capability to produce .gif files and chart visualizations.

by clicking the "Run all Nodes" button, which runs the existing code within the workflow to compile the map. Visual instructions for this procedure will be provided in the supplementary material. The result while inputting the base map for Albania and Albania's daily confirmed cases data is displayed in Fig. 3. This procedure within the workflow tool was then repeated for each country and case that we wanted to observe

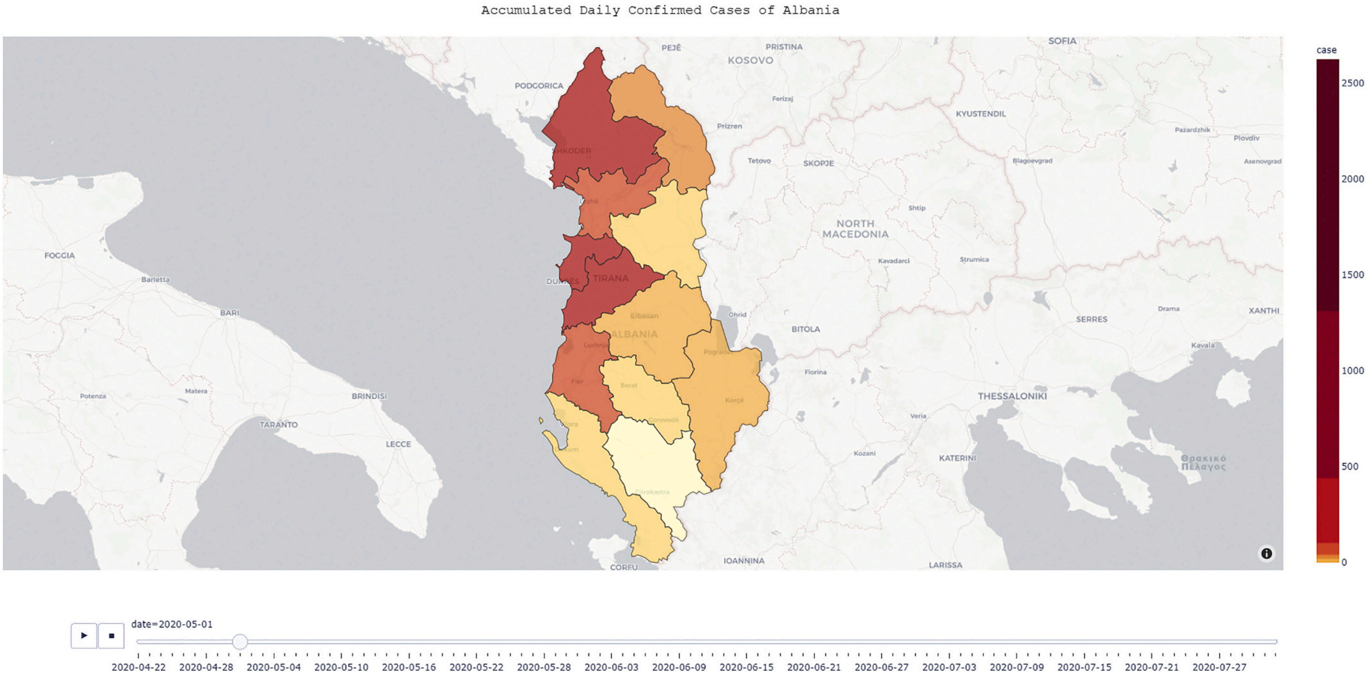

Fig. 3. This represents the dynamic map created for the country of Albania on the cumulative daily confirmed COVID-19 cases.

by creating a new workflow and editing the parameters. Figs. 4 and 5 showcase the compilation of choropleth maps depicting the number of cases in Sweden and Ukraine.

Different data subjects such as confirmed, death, and recovered were considered. By changing some parameters around along with the source files, a new choropleth map showcasing each of the different data subjects could be created in a short period, proving the replicability of the workflow for different parameters. As a tool, KNIME was very simple even for the untrained eye; the only changes needed to be made within a country were changing the titles (of graph and output) and source files.

3. Results

The choropleth maps allow us to visualize the relationships between each county, which is crucial in understanding the spread of COVID-19. This workflow fulfills this demand by providing a graphical display as well as a map over a time series. For example, the line graph in Fig. 6 shows that Albania's Vlorë region has relatively stable cases up until 6/11/2020, where it then surpasses seven regions in cumulative cases. While this trend can easily be identified from a line graph, visualizing it on a map allows us to draw different types of conclusions. Figs. 7 and 8 can give insight into how proximity to counties with higher cases can influence other provinces' risk of contracting more cases. On 6/11, Vlorë only had 6 cases, while Fier, the county north of Vlorë, had 64. Just 13 days later on 6/24, Vlorë's cases rose to 75. This trend can be seen in the following months as well. While no definite conclusion as to whether or not proximity was the sole factor that caused the increase in cases can be made, choropleth maps allow the viewer to better understand the relationships between regions than if cases were displayed on a line graph.

For instance, India is a very heavily populated country as it is home to about 1.35 billion people. India comprises twenty-nine states, with three of the top ten most populated states exhibiting the most amount of spread during the 111 days tracked: Maharashtra, Tamil Nadu, and Delhi. Regions farthest away from the major Indian cities within those states did not exhibit much change in the number of cases. On the contrary, cities like Mumbai, Chennai, and Delhi had a significant impact on the spread within their respective state. Each of the aforementioned cities has large hubs of transport and rank in the top four for the most amount of passenger traffic within all Indian airports (Asher, 2020) and also rank in the top five for the most densely populated Indian cities (Top 10 Most Densely Populated Districts of India, 2020) "Top 10 Most Densely Populated Districts of India", 2011). These cities seem to influence the states that are close in proximity. In particular, Gujarat, a relatively smaller state with a smaller population density as compared to Maharashtra, is seen to exhibit an alarmingly large number of cases for its size: the fourth-most in the country. Gujarat borders Maharashtra and is close to Mumbai, where many Gujaratis live. West Bengal, the most densely populated Indian state, and Uttar Pradesh, the most populated Indian state, have maintained a relatively low number of cases. Uttar Pradesh can do this even with the proximity to New Delhi. The number of cases on June 29, 2020, in India is shown in Fig. 9.

A similar phenomenon where the areas with denser population have more cases can be observed in Russia as well. Home to around 145 million residents, Russia's population is spread out over 84 subdivisions. Cases in the two most populated and densely populated cities in Russia, Moscow, and St. Petersburg (City Population, 2020), are seen to rapidly increase within their respective subdivisions. These two cities are also home to the country's largest airports for commercial passengers (Air-Mundo, 2020). Subdivisions surrounding these cities were shown to have had a large increase in the number of cases after the initial onset being in those cities, much like the instance with Gujarat and Mumbai. The Kaliningrad Oblast, geographically located between Poland, Belarus, and Lithuania, shows a rather modest number of cases even though many countries with open borders are close to it. However, Russia wholly is unable to contain the spread, as evidence from the map suggests that the areas with large numbers of cases tend to spread to the nearby subdivisions, increasing the number of cases on a widespread level. The number of cases on June 29, 2020, in Russia is shown in Fig. 10. This phenomenon was present in all of the other countries we analyzed as well.

4. Discussion

Since this approach is generalizable to any dataset that captures coronavirus data in regions, there is a huge potential to expand the workflow and the research scope. Three main directions for this

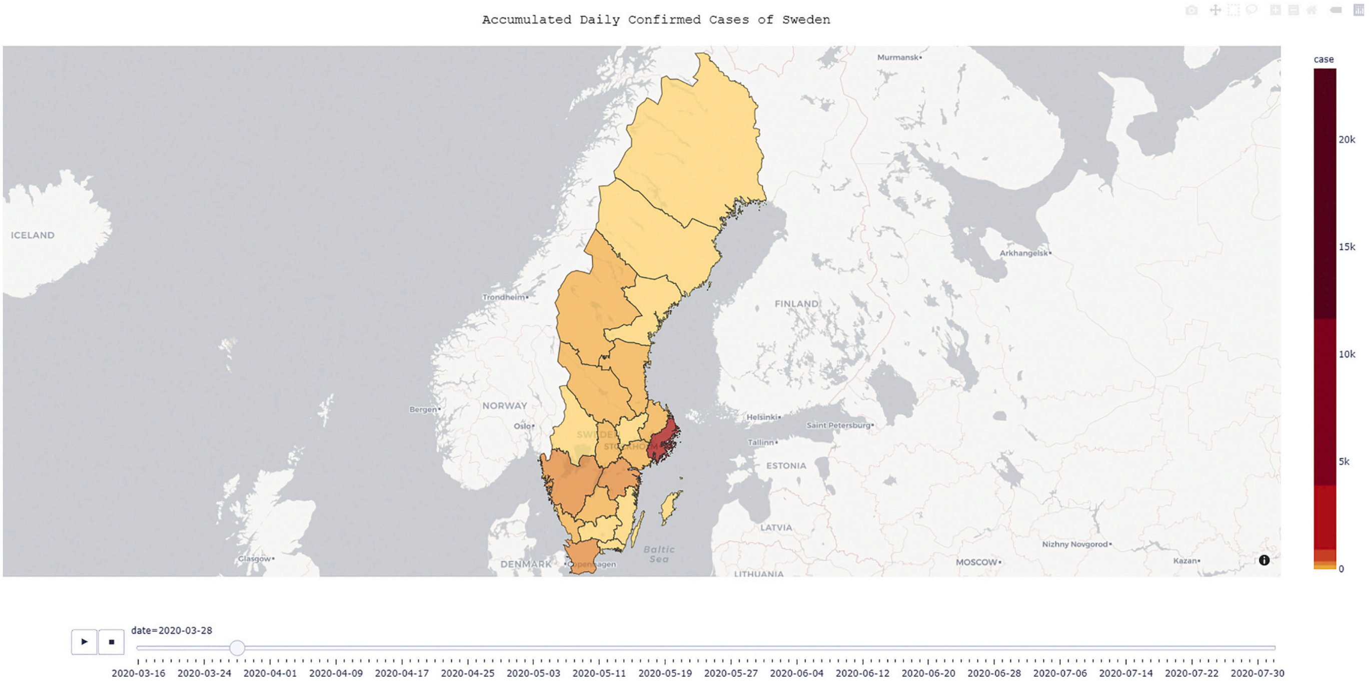

Fig. 4. This represents the dynamic map created for the country of Sweden on the cumulative daily confirmed COVID-19 cases.

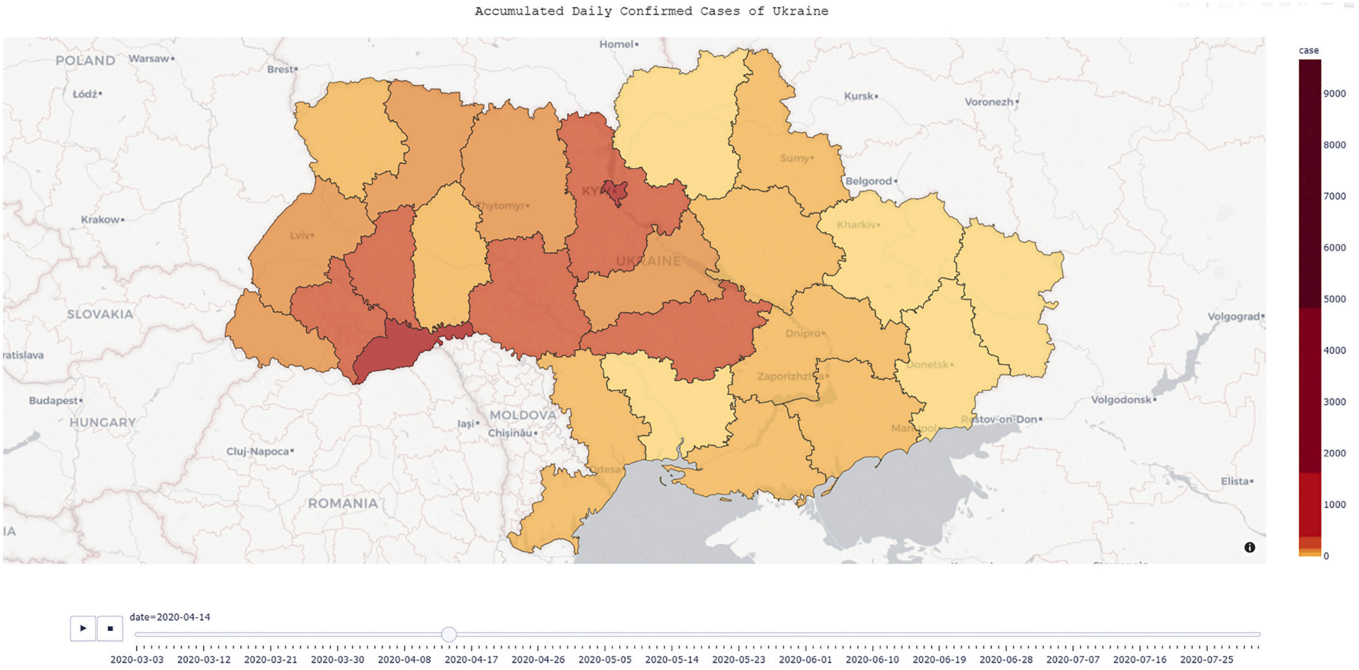

Fig. 5. This represents the dynamic map created for the country of Ukraine on the cumulative daily confirmed COVID-19 cases.

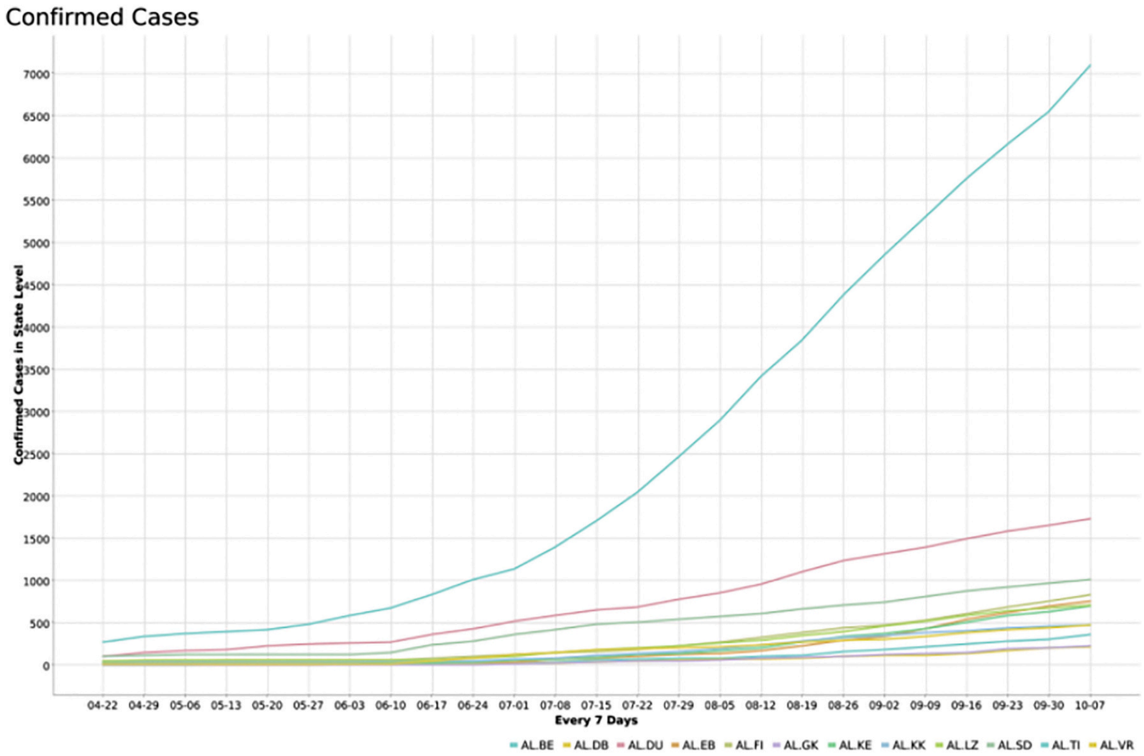

Fig. 6. This represents a line graph of the cumulative daily confirmed COVID-19 cases in different regions of Albania.

workflow-based model include statistically analyzing additional countries, using additional advanced displays with workflow extensions (mp4, reports, charts) to possibly create small displays, incorporating additional factors, such as demographics, socioeconomics, and policies as map layers, and creating large and user-friendly interactive maps that allow the user to zoom around a global map and look at different layers. A feature to allow users to use "day averages" to change the time series to display the map at the frequency of choice is currently being developed. Furthermore, this workflow-based model has strong capabilities towards statistically analyzing continents wholly as well. Currently, we are planning on creating a base map for each continent and combining the raw data for each country in the continents into one data file. In doing so, it would allow us to create a dynamic map for each continent, allowing researchers to track the spread of COVID-19 at a larger standpoint rather than analyzing each country individually. This can be very beneficial to containing the spread of COVID-19 as it would

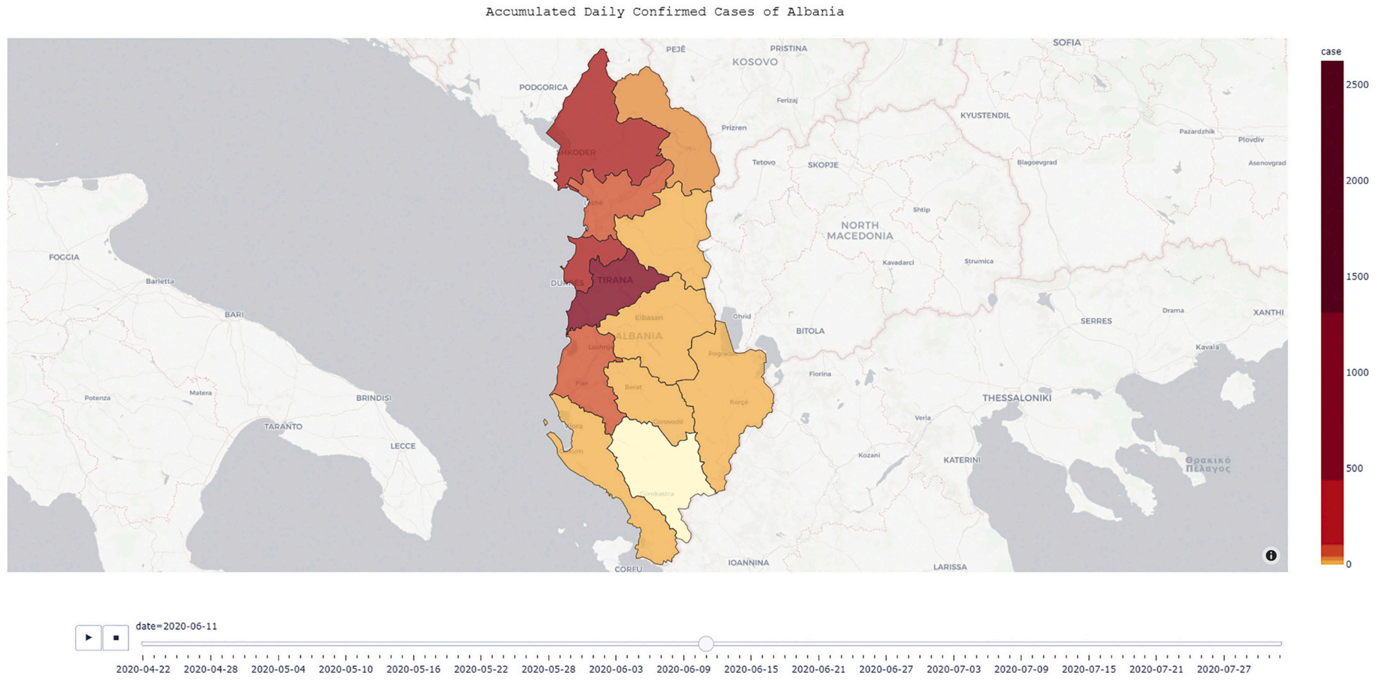

Fig. 7. This dynamic map illustrates the amount of cumulative daily confirmed COVID-19 cases for Albania on 6/11/20.

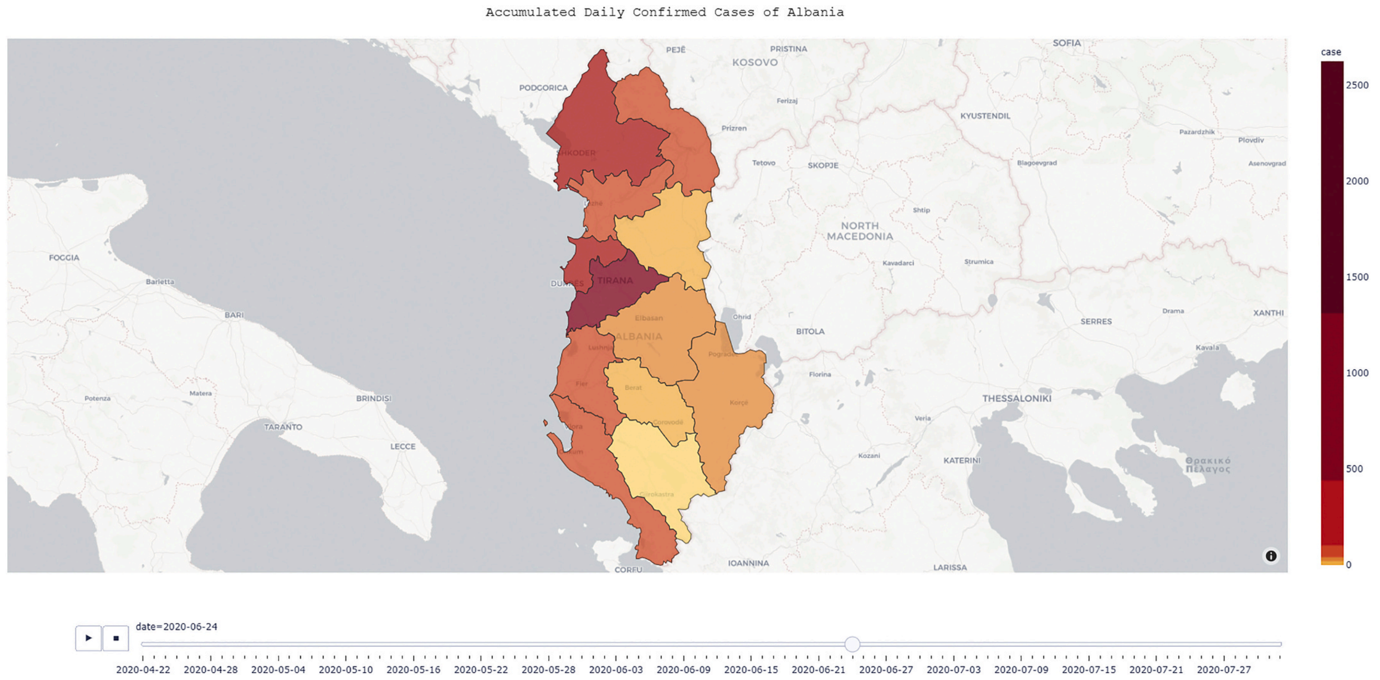

Fig. 8. This dynamic map illustrates the amount of cumulative daily confirmed COVID-19 cases for Albania on 6/24/20.

provide information on how to prevent the spread throughout an entire continent. From there, by preventing spread from that continent, it would immobilize the spread to other continents. Hence, this type of analysis would be extremely beneficial.

5. Conclusion

Using the maps generated, further analysis of the relationship between provinces and their neighboring provinces in another country should be conducted. This study may be able to determine the relationship between province cases and province responses to COVID-19. As stated previously, there have been a plethora of studies on how the United States' state policies are affecting each state differently – it will be beneficial to conduct a similar analysis of other countries. We saw clearly with the analyzed countries the impact of proximity to increased COVID-19 cases; yet, given the lack of city level data for each country, it is also difficult to tell entirely whether this is true or false. However, the identification of the trend does provide the opportunity that governments within a country will start to collect city level data in order to look into this theory much closer. The proximity trend opens up various causes that can be analyzed and deduced from the dynamic maps to fight against the spread of COVID-19 to show which regions need to tighten

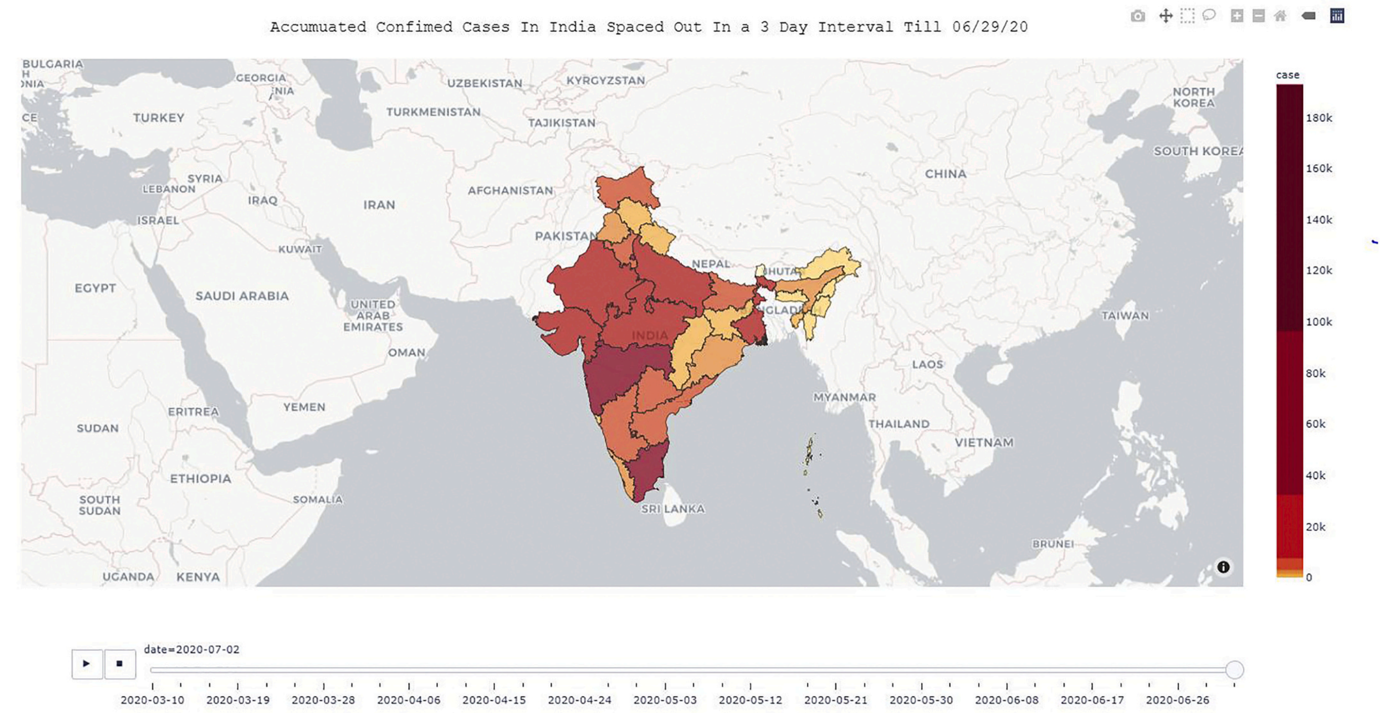

Fig. 9. This dynamic map illustrates the amount of cumulative daily confirmed COVID-19 cases for India on 6/29/20.

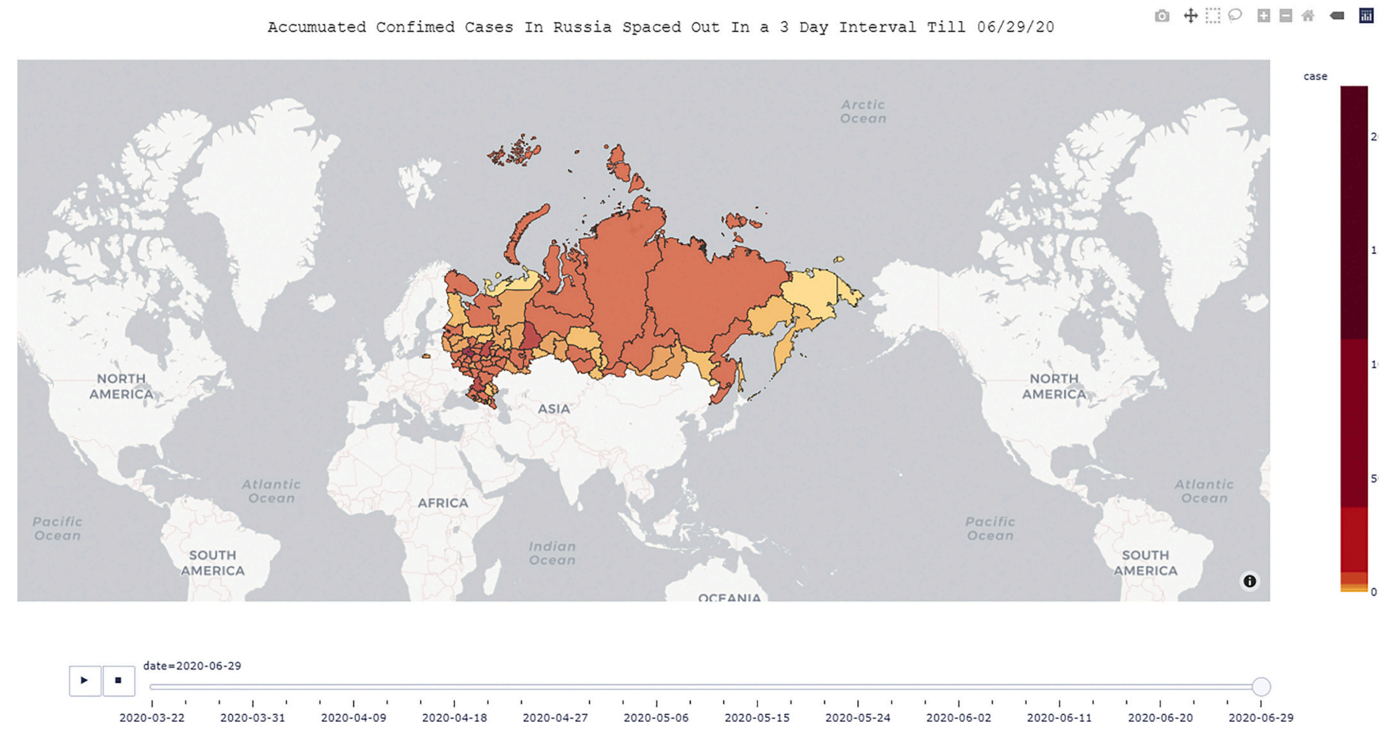

Fig. 10. This dynamic map illustrates the amount of cumulative daily confirmed COVID-19 cases for Russia on 6/29/20.

on their border restrictions. Further applications and iterations of this study may be able to suggest changes within the foreign policy to possibly immobilize the spread of the virus within the countries in question.

Declaration of Competing Interest

None

Acknowledgments

We would like to express our appreciation for Dr. Shuming Bao and

Dr. Tao Hu for their valuable and constructive feedback during the planning and development of this research work. We would also like to thank Harvard's Center for Geographic Analysis, Wuhan University's Geocomputation Center for Social Science, and the China Data Lab for their support and guidance in conducting this research. Data was provided by the Spatiotemporal Innovation Center.

Please contact Dr. Shuming Bao (sbao@umich.edu) or Dr. Tao Hu (taohu@fas.harvard.edu) for access to the workflows created by this study. To view all the maps created within the study already, visit https://covid-19.stcenter.net/index.php/animated-maps/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104701.

References

AirMundo, 2020. List of major airports in Russia. Retrieved August 26, 2020, from. https://airmundo.com/en/countries/russia/.

Asher, P.B., 2, J, 2020. India - leading airports by number of passengers handled. June 022020. Retrieved from. https://www.statista.com/statistics/589115/indian-airports-passenger-traffic/.

Chahal, H., Gulia, P., 2016. Comprehensive study of open-source big data mining tools. Int. J. Artificial Intelligence Knowledge Discovery 6 (1).

City Population, 2020. Russia: Federal Districts and Major Cities. Retrieved August 26, 2020, from. https://www.citypopulation.de/en/russia/cities/.

Pynam, V., Spanadna, R., Srikanth, K., 2018. An extensive study of data analysis tools (rapid miner, Weka, R Tool, Knime, Orange). Int. J. Comput. Sci. Eng. 5 (9).

Ranjan, R., Agarwal, S., 2017. Detailed analysis of data mining tools. Int. J. Eng. Res. Technol. 6 (5).

Top 10 Most Densely Populated Districts of India. (2020). Retrieved August 26, 2020, from https://www.census2011.co.in/facts/topdistrictdensity.html.

WHO Coronavirus Disease (COVID-19) Dashboard, 2020. https://covid19.who.int/.

Zhou, C., Su, F., Pei, T., 2020. COVID-19: challenges to GIS with big data. Geogr. Sustainab. 1 (1).