A Spatial Exploration of COVID-19 Vaccine Uptake in the United States

Maeve Hutchinson
ID: 200051807
Department of Computer Science
City, University of London
January 2022

***Abstract*—This report investigates COVID-19 vaccination in the United States, focussing spatial variation in vaccine uptake based upon demographic and political factors. It was found that there is a significant negative relationship between the proportion of people who voted republican in a given county, and the vaccine uptake in that county. It was also found using a Geographically Weighted Regression model that vaccine uptake varies spatially across the US, and the demographic factors contributing to that uptake also vary. The US was clustered using K-means clustering, and the factors contributing to uptake in each cluster was examined.**

I. Problem Statement

The COVID-19 virus has dominated every aspect of our lives since its outbreak in December 2019. In the United States alone it has caused an estimated 836,000 deaths thus far [1]. The most effective method of mitigating further impact is vaccination, which has been proven to dramatically reduce the chances of becoming seriously ill with the virus [2]. Despite this, many people globally who are able to get vaccinated have chosen not to. In particular, in the US there has been a strong anti-vaccine sentiment, with protests all across the country [3]. Thus, this report explores the spatial dynamics of vaccine uptake in the US, with the aim of answering the following research questions:

- Is there a relationship between demographic factors and vaccine uptake?
- Does vaccine uptake vary spatially in the US?
- Can these spatial variations be explained by demographic factors? Do these variations also vary spatially?

The data used is county-level first vaccine uptake in the US [4]. Alongside this, the 2020 Presidential Election data will be used [5], as much of the vaccine skepticism is politically driven. The Atlas of Rural and Small-Town America, [6], provides county-level demographic data for the whole country under the categories of people, jobs, and income.

These datasets are suitable for answering the research questions, as they are at a sufficiently small geographical division to capture demographic differences spatially, and cover many different demographic factors. All the data are from reliable governmental sources and are from 2015 to 2021, so provide adequately recent data to describe an event as current as COVID-19.

II. State of the Art

Despite the specific domain of COVID-19 vaccination being distinctly new, there are already several papers relating to the spatial dynamics of COVID-19 vaccine uptake [7]. The first paper of interest, [8], aimed to produce novel techniques for visually analysing vaccine uptake through the use of rotavirus vaccine data. There was a particular focus on future uses of the visualisation techniques developed in exploring COVID-19 vaccine uptake in the United States, making it very relevant. More levels of resolution are explored, including state and ZIP code alongside county, and temporal dynamics of vaccine uptake are explored alongside spatial.

Of particular relevance to this project are the methods developed involving hotspot maps to identify areas requiring further interventions, and visualisation maps of vaccination uptake. These techniques reveal the importance of interactivity when visualising data at a county-level resolution — interaction over space allows outliers and hotspots to be readily identified compared to static maps. Although this cannot be displayed in this report, interactive maps will be used heavily throughout the analysis process of this project to explore the data in detail.

The second paper of interest, [9], explores the exact same vaccination dataset as this report, but earlier in time, ending in July 2021. This paper also aims to explore demographic and spatial dynamics of vaccine uptake in the US and uses Social Vulnerability Index (SVI) to provide demographic data in many categories similar to those explored in this report. The key difference is that this paper does not include any data about political views, which is an important aspect of this report, and the demographic areas explored are slightly different.

The techniques used are extremely relevant — first, Ordinary Least Squares (OLS) regression is presented as a baseline to model vaccine uptake as a function of demographic factors, and then Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR) are implemented. GWR models, as opposed to OLS, allow each explanatory covariate, in this case, demographic and political factors, to have a different coefficient in different spatial regions, which makes it an ideal model for exploring my final research question of whether reasons for low vaccine uptake vary spatially.

III. Properties of the Data

The vaccination dataset is provided by the CDC and collected nationally at vaccination centres. The data comes with various columns detailing the uptake of the vaccine, however, the only column to be used is the percentage of the population in a given county that has received at least one dose, as this will best capture vaccine hesitancy. The data is updated regularly, so temporal data is provided, however, this will be disregarded and only the most recent entry for each county will be used, which is from December 31st 2021.

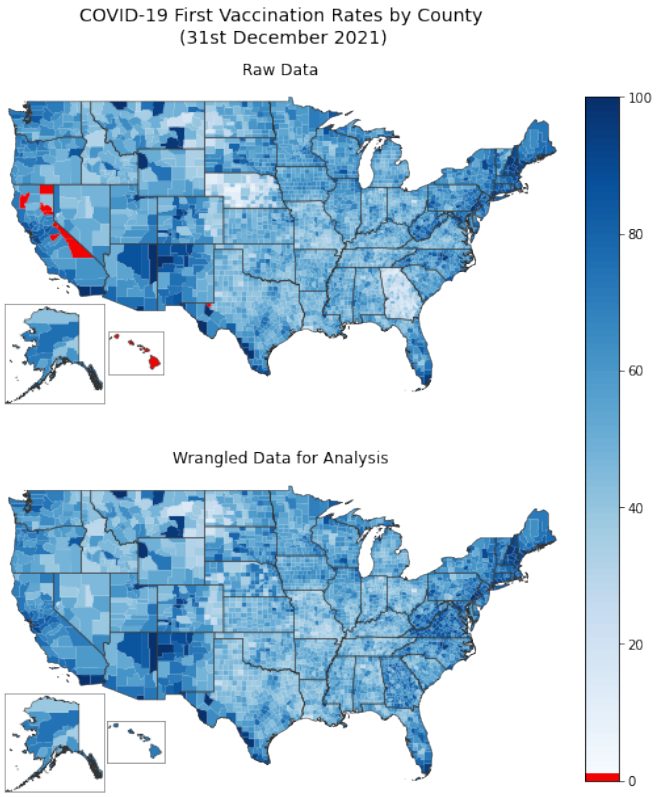

Fig. 1. Map of COVID-19 first vaccination rates by county before and after cleaning the data.

The data is provided at a county level. When vaccination rates are plotted on a map (Fig. 1) it can be seen that there are several counties with missing values, as indicated in red. This is likely due to the varying infrastructure in place for vaccination in different states and counties. This issue was dealt with by imputing the mean of other counties within the same state with available data. This method was chosen as there were very few missing values, so it will likely not impact the findings considerably, especially as visually it seems that counties within the same state often have a similar uptake level. For Hawaii, data was found separately directly from its official website [10].

From figure 1 it can also be observed that certain states, such as Georgia and Nebraska are on average considerably lighter than surrounding states. So, the percentage of vaccinations with unknown county were calculated for each state. It was found that a few states had a very high number of records missing a county, which would cause the uptake to appear lower when viewed at county level. Notably, Georgia had over 80% of records missing a county. For the 4 states with greater than 25% unknown county records, the counties were redistributed using the total state mean uptake. After wrangling (fig. 1) it can be seen that Georgia and Nebraska are now a more similar colour to surrounding states, as expected.

The election data was also provided in a time series format, so only the data from 2020 was selected. The only feature attained from this data was the percentage of people in a given county that voted for the Republican Party. This feature was chosen because in the news it appears that most vaccine skepticism is coming from supporters of the Republican Party, so this relationship is worth investigating.

From the Atlas of Rural and Small-Town America, the columns chosen were county population, percentage with a college degree or above, per capita income, unemployment rate, percentage living in poverty, percentage white non-Hispanic people, percentage foreign-born, percentage of non-English speaking households, and percentage over 65. When plotted, both the election data and demographic data did not appear to have any missing or erroneous values (fig. 2).

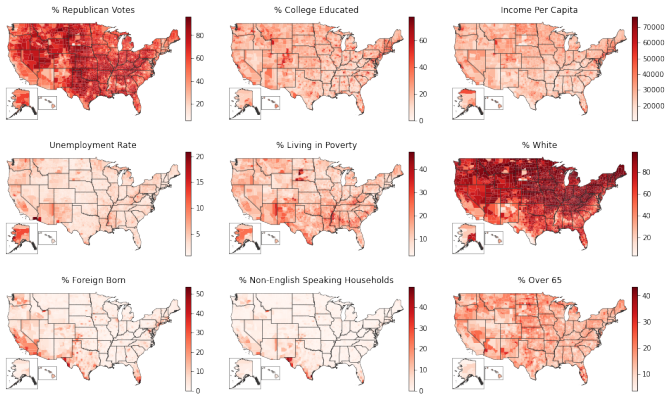

Fig. 2. Maps of selected demographic features by county.

All three datasets were joined based upon FIPS, a unique geographical code assigned to counties. For plotting purposes, a shapefile of the US counties was also used [11].

IV. Analysis

A. Approach

Task 0 - Data Preparation

Initial visualisation and data cleaning, as described in the previous section.

Task 1 - Exploration

Initial exploration of the selected features from task 0 will be undertaken both visually and quantitatively. Scatter plots of each feature against first vaccine uptake will be visualised and a pearson correlation matrix of all the features will be calculated. From this, if there are any features that do not

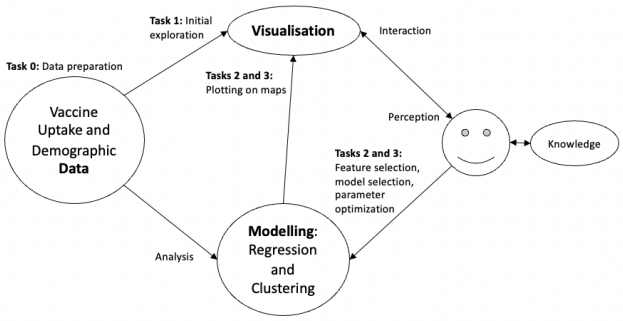

Fig. 3. Analysis workflow diagram.

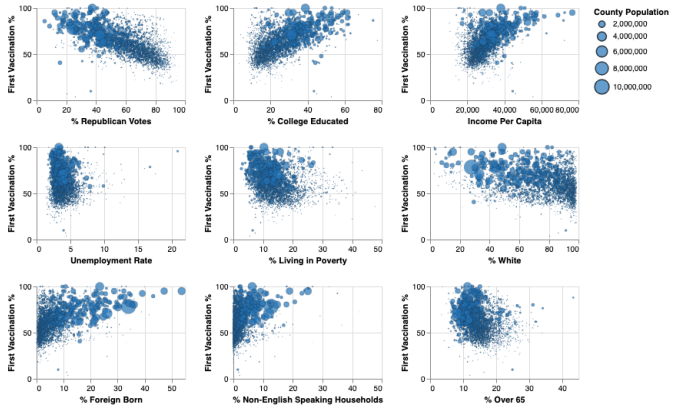

Fig. 4. Scatter graphs of each feature against first vaccine uptake. Each point represents an individual county.

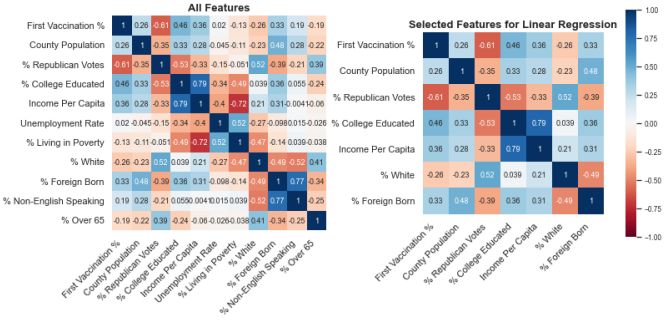

Fig. 5. Correlation matrices of all features and vaccine uptake before and after feature selection.

appear to be linearly related to vaccine uptake, they will be disregarded from further analysis. The variance inflation factor will be calculated to check for collinearity, and if necessary more features will be disregarded.

Task 2 - Regression Modelling

As in [9], first, an ordinary least squares (OLS) regression model will be fitted to the data, both for all the features and each individual feature. The model residuals for each feature will be visualised on a map to check for spatial heterogeneity, looking towards answering the research questions about spatial variation. It is expected from [9] that there will be spatial heterogeneity, which justifies fitting a geographically weighted regression (GWR) model, as it takes into account differences over space. The correlation coefficients of the GWR for each feature over space will be plotted on a map to visualise the heterogeneity.

Task 3 - Clustering

Finally, a k-means clustering model will be fitted to the data based upon the GWR. This will allow the final research question to be investigated. The results of the clustering will be visualised on a map, and the centroids of each cluster will be analysed.

B. Process

Task 1 - Exploration

Figure 4 shows each feature, besides population, plotted against vaccine uptake. Population is shown as the size of the data point. There are several features that do not appear to be linearly related to the vaccine uptake, such as unemployment rate, percent living in poverty, and percent over 65. To confirm whether these should be disregarded, a Pearson correlation matrix was plotted (fig. 5). All features with a value of less than 0.25 were disregarded, reducing the number of features to the 6 shown in figure 5. In figure 5 it appears that there may be some problems with collinearity, but VIF scores were calculated and they were all below 5 so no more features were removed.

The feature with the strongest linear relationship to vaccine uptake is, in fact, the percentage voting republican with a Pearson value of -0.61. This suggests that the stereotype that Republican supporters are more vaccine-hesitant could be true.

Task 2 - Regression Modelling

Using the selected features, an OLS model was fitted to the data. It had an r^2 value of 0.403, which suggests that this model does not best capture the relationship between the features and vaccine uptake.

So, OLS models were fitted for each feature individually, and the residuals for each of those models was visualised on a map (fig. 6). It can be seen that there is spatial variation in the residuals, in particular with the East Coast having high residuals for most of the categories, which displays that the model is predicting uptakes that are too high for that area. This suggests that vaccine uptake does vary spatially in the US.

Due to the spatial variation, a GWR model was fitted next. This takes into account the spatial variation because the coefficient for each covariate can vary across counties. Each feature was standardised, the GWR model was fitted to the standardised features and each feature coefficient for each county was calculated. These are visualised in figure 7. The r^2 value for the GWR model was much better than for the OLS model at 0.8, which suggests that the reason for vaccine uptake also varies spatially. Figure 7 clearly shows the geographical variation of each feature contribution to vaccine uptake, however, it is difficult to interpret, so counties will

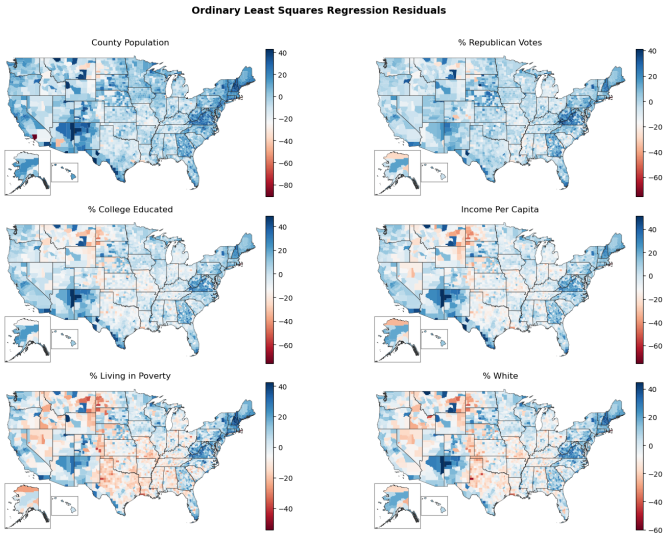

Fig. 6. OLS residuals for each feature in each county.

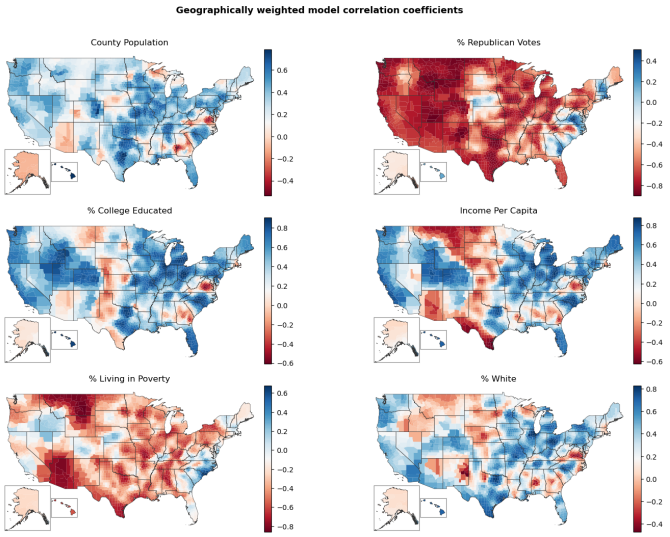

Fig. 7. GWR model coefficients for each feature in each county.

now be clustered.

Task 3 - Clustering

For the final task, k-means clustering was undertaken based upon the coefficients from the GWR model. First, the distortion score for various values of k was calculated, and plotted in figure 8. From this visualisation, we can see that the elbow of the graph is at $k = 5$, which suggests that is the optimal value. Models with values from 2 to 7 were fitted and plotted, to visually confirm whether $k = 5$ is the optimal value (fig. 8). It can be seen that 2 is clearly too few clusters, and beyond 6 it is very hard to interpret. Thus, it was decided that visually $k = 4$ was the clearest model.

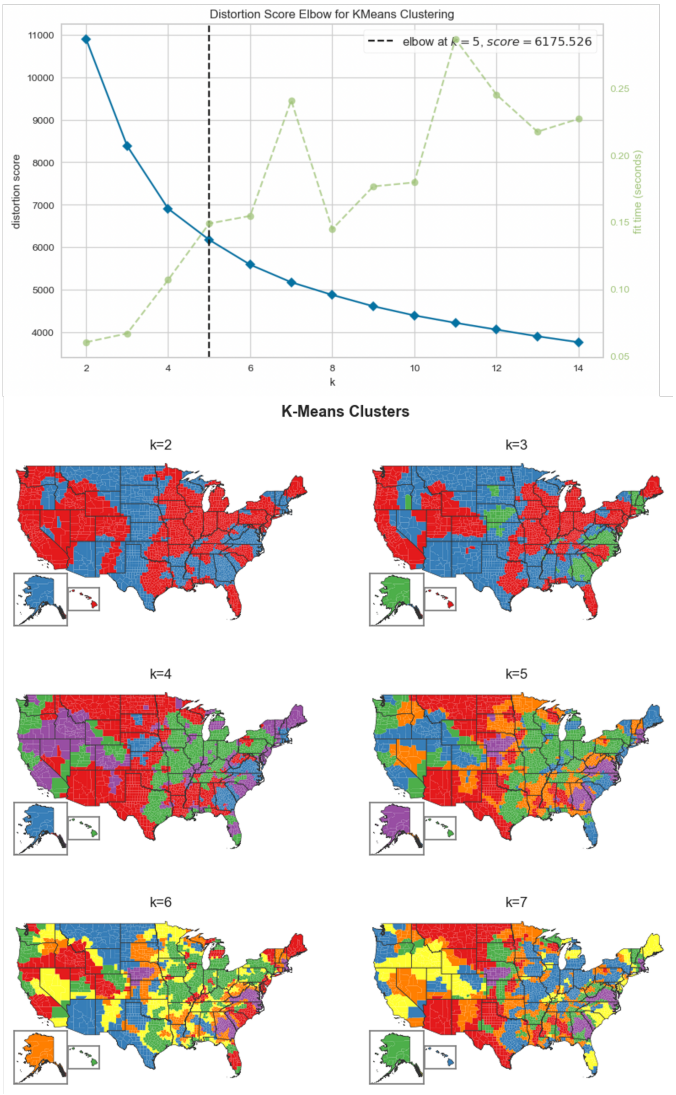

Fig. 8. Elbow plot and K-Means clusters from k=2 to k=7 of each county based upon the GWR model coefficients.

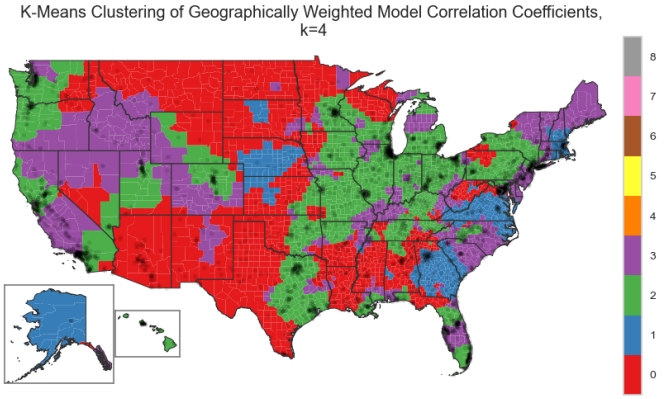

Fig. 9. K-means clustering of the US counties based upon the GWR model coefficients, k=4, with major cities shown.

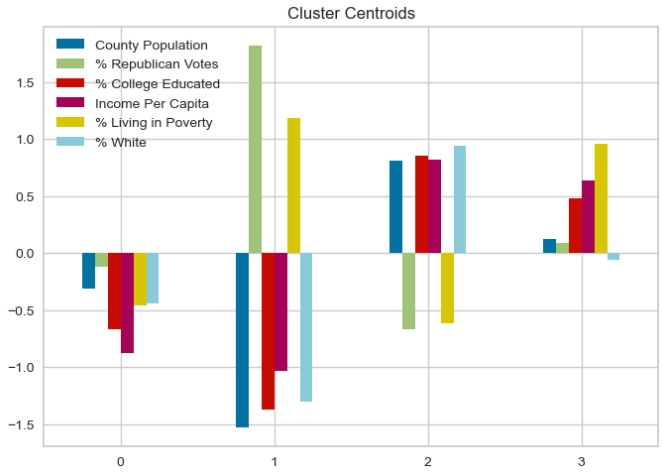

Fig. 10. Cluster centroid values for each cluster.

C. Results

It has been found that vaccine uptake does vary spatially in the US, and the demographic and political reasons for vaccine uptake levels also vary spatially. Figure 9 visualises the final k-means clustering model based upon the GWR coefficients. Also plotted are the major cities in the US, to see which clusters they fall into. It can be seen that cluster 2 contains most of the cities, so that is likely the most metropolitan, whereas cluster 1 contains the whole of Alaska and areas such as Nebraska, so is likely to be more rural.

Figure 10 visualises the cluster centroids as a bar chart. Because each feature was standardised before the GWR was fitted, the features are all on the same scale. It can be seen that the most important feature contributing to vaccine uptake levels in rural areas (cluster 1) is political outlook, and the second is percent living in poverty. This suggests that low vaccine uptake is due to a mixture of both political factors and lack of access to resources.

V. Critical Reflection

Broadly, the methods used were able to answer the research questions posed. The choice of a GWR model was based upon precedent of the same method being used on a very similar dataset, so that part of the analysis was relatively successful in capturing the spatial variation.

However, the extension to clustering was less successful - whilst some useful insights were gained, the clusters do not really follow any recognisable pattern and are not particularly clear cut when visualised. Perhaps there are better methods of capturing the variation described by the GWR other than traditional clustering techniques.

The data was also limiting in that there were only certain demographic factors available at a county level, and so that regression could be used, the categories needed to be quite distinct from each other. So, perhaps a better model could be built using different categories.

The key lesson learnt is that when dealing with similar data to this is that GWR models perform much better than OLS, and are very good at capturing spatial variation.

REFERENCES

[1] H. Ritchie et al., "Coronavirus Pandemic (COVID-19)," Our World in Data, Mar. 2020, Accessed: Jan. 09, 2022. [Online]. Available: https://ourworldindata.org/coronavirus

[2] A. Pormohammad et al., "Efficacy and Safety of COVID-19 Vaccines: A Systematic Review and Meta-Analysis of Randomized Clinical Trials," Vaccines, vol. 9, no. 5, p. 467, May 2021, doi: 10.3390/vaccines9050467.

[3] B. Haring, "Anti-Vaccination Protests Go Worldwide As Backlash To Mandates Grows," Deadline, Sep. 19, 2021. https://deadline.com/2021/09/anti-vaccination-protests-worldwide-backlash-to-mandate-grows-1234839338/ (accessed Jan. 09, 2022).

[4] "COVID-19 Vaccinations in the United States,County — Data — Centers for Disease Control and Prevention." https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh (accessed Jan. 09, 2022).

[5] MIT Election Data And Science Lab, "County Presidential Election Returns 2000-2020." Harvard Dataverse, 2018. doi: 10.7910/DVN/VOQCHQ.

[6] "USDA ERS - Atlas of Rural and Small-Town America." https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/ (accessed Jan. 09, 2022).

[7] A. Mollalo, A. Mohammadi, S. Mavaddati, and B. Kiani, "Spatial Analysis of COVID-19 Vaccination: A Scoping Review," IJERPH, vol. 18, no. 22, p. 12024, Nov. 2021, doi: 10.3390/ijerph182212024.

[8] T. C. Mast, D. Heyman, E. Dasbach, C. Roberts, M. G. Goveia, and L. Finelli, "Planning for monitoring the introduction and effectiveness of new vaccines using real-word data and geospatial visualization: An example using rotavirus vaccines with potential application to SARS-CoV-2," Vaccine: X, vol. 7, p. 100084, Apr. 2021, doi: 10.1016/j.jvacx.2021.100084.

[9] A. Mollalo and M. Tatar, "Spatial Modeling of COVID-19 Vaccine Hesitancy in the United States," IJERPH, vol. 18, no. 18, p. 9488, Sep. 2021, doi: 10.3390/ijerph18189488.

[10] "Hawaii COVID-19 Data." https://health.hawaii.gov/coronavirusdisease2019/current-situation-in-hawaii/ (accessed Jan. 09, 2022).

[11] U. C. Bureau, "Cartographic Boundary Files," Census.gov. https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html (accessed Jan. 09, 2022).

Section	**Words**
Abstract	96
Problem Statement	248
State of the Art	356
Analysis: Approach	491
Analysis: Process	526
Analysis: Results	173
Critical Reflection	184