Charles McGrath

Spatial Epidemiology: Final Paper

12/18/2024

Analyzing the Effect of Covid-19 Vaccination Rates on Mortality at a County Level

Introduction

After the first recorded cases of Covid-19 in the US in January 2020, the country saw drastic change during the next three years. In the US alone, there were a hundred million confirmed cases and more than one million deaths. Although cases continued to persist, mortality rates declined significantly after the winter of 2021. Numerous studies have shown that the Covid-19 vaccine, introduced by Pfizer in December 2020, reduced the rate of severe sickness and death. In experiments conducted during early 2021, Seyed Moghadas and others found that vaccination reduced death rates by 63.5% (Moghadas, 2021). Research has also shown that the impact of Covid-19 is unevenly distributed across the geography and population of the US. While rural areas have seen far less cases and deaths than urban areas, the lack of access to healthcare facilities caused significant vulnerabilities (Cuadros, 2021). Covid-19 was felt hardest among minorities, such as Hispanics and Blacks living in urban areas, and disproportionately affected older people. However, there are very few studies that have analyzed the effects of vaccination rates and other factors on Covid-19 mortality rates on a county-level basis. Focusing on county-level data can reveal geographic variation across broader regions of the country and localized rural and urban areas. The purpose of my project was to analyze the effect of vaccination rates and socioeconomic indicators on Covid-19 mortality using geospatial analyses with Python. My hypothesis was that counties with low vaccination rates, high poverty rates, and more urbanization would correspond with higher Covid-19 mortality rates.

Data Sources

This study used four primary data sources. Firstly, a table of Covid-19 death counts per week was downloaded from the CDC's website. The data, which was collected from February 2021 to May 2023, was measured across the US and contained county-wide information. From the CDC's Covid-19 database, a CSV of Covid-19 vaccination rates per county was used. In addition, the USDA's rural-urban commuting area codes (RUCA), which classify census tracts using population density, commuting times, and urbanization (USDA, 2020). A CSV containing the rate of families living below poverty, organized by county-level across the US, was also used. For purposes of reference, a CSV containing weekly Covid-19 cases per county was included. Finally, county boundary shapefiles of Missouri and the US were used in the analysis.

Study Design

This project used a cross-sectional study to measure the rate of vaccination and other factors across the population of US counties, and determine if these factors influenced the Covid-19 mortality rate. A cross-sectional study is a retrospective study in which descriptive information is collected at a specific time point to understand the distribution and factors of diseases. In this study, the data was aggregated and analyzed at two time points, the year 2021 and the year 2022. The Covid-19 mortality data was aggregated over 2021 and 2022 separately, and the cumulative vaccination rate was measured on the first of January of 2022 and 2023. Data for these time points were not available for poverty and urbanization data, so the most recent available data was used.

By using a cross-sectional study, the researcher can calculate occurrence rate over an entire population of interest. Using Covid death counts, it is possible to calculate the mortality rate for each county. Sampling is required for calculating vaccination and poverty rates, because the US population is much too large to survey. However, the sampling is based on the population, rather than individuals in a group, so it is possible to calculate the rate of causal inference. Also, no sampling is required for mortality rates, because deaths are reported to the government. Overall, a cross sectional study such as this one is highly efficient and cost-effective, as it only requires the use of pre-existing government data.

Methods

The methods used in this study can be broken into four steps: data cleaning, visualization, modeling, and analysis.The first step in this project was to clean and process the data for visualization and analysis. All data layers were joined with the US county boundary shapefile using the county FIPS code, and counties in the lower 48 states were selected. Then, Covid death counts were aggregated per county for the year 2021 and 2022. This was done to provide a specific time point of a year when analyzing Covid-19 cases. This aggregated death counts layer was then joined to the covid-19 case layer, which included county population data. A feature was created for the mortality rate per county, which was calculated by dividing the number of Covid-induced deaths by the county population. In addition, the percentage of population with one vaccination dose for each county on or nearest to January 1 of 2021 was selected. The same process was repeated for January 1, 2022. As RUCA values were organized per census tract, the most common RUCA value for each county was selected. To clean the data, I used the Numpy, Pandas, Geopandas, and Matplotlib Python packages within Jupyter Notebook. Next, I visualized the data by plotting a series of maps. Figure 1 shows the percentage of population per US county with one dose as of January 2022, and Figure 2 shows the percentage with one dose as of January 2023. Figures 3 and 4 show aggregated Covid mortality rates during 2021 and 2022, respectively. Figure 5 is a map of the most common rural/urban code by county in 2010. Figure 6 is a map of the poverty rates by county, as of 2022.

The next step in the project was to determine the effect of the three explanatory variables on Covid-19 mortality rates using linear regression. First, the data for all three variables was merged together, separately for the 2022 and 2023 data. Then, the most important features were selected and converted into a new data table. Since I predicted a negative association between vaccination rates and Covid-related deaths, I created a feature containing the percentage of county population without the vaccine by subtracting 1 from the vaccination rate. Next, a linear regression was performed on each individual feature from the 2023 data. To perform this regression, the data was split into X and y train and test sets using the Scikit Learn package. The x values included the explanatory variable: percent without vaccine per county, the most common RUCA, or the poverty rate. The y values, as the dependent variable, were set as the mortality rate per county. The test size was to be 20% of the entire dataset. Next, the linear regression model was trained on the training sets, and predictions were made using the test set. The R-squared value was computed from the predictions and the ground truth values. After testing each variable, a linear regression was performed on all three explanatory variables at once using Scikit Learn. This entire workflow was repeated for the 2022 data. Scatter plots were created to visualize the relationship between the explanatory and dependent variables.

Results

The percentage of population with the first vaccination dose was similar between 2022 and 2023. An average of 58.9% of people across US counties had received their first dose in January 2023. 109 counties reported a vaccination dose of above 95%. North Carolina and Texas had the highest number of counties (12) with a 95% vaccination rate. In January 2022, an average of 52.9% of people across US counties had received their first dose. Only 31 counties reported a vaccination dose of above 95% in January 2022. Texas also had the highest number of counties with a 95% vaccination rate, but only had 5 of them. Figures 1 and 2 show some notable changes. Both Georgia and Nebraska had a majority of counties with under 40% vaccination rates in 2022. While Nebraska largely stayed that way in 2023, Georgia's vaccination rates increased significantly. In general, counties in the northeast, Florida, and the Southwest had high vaccination rates.

Figure 1

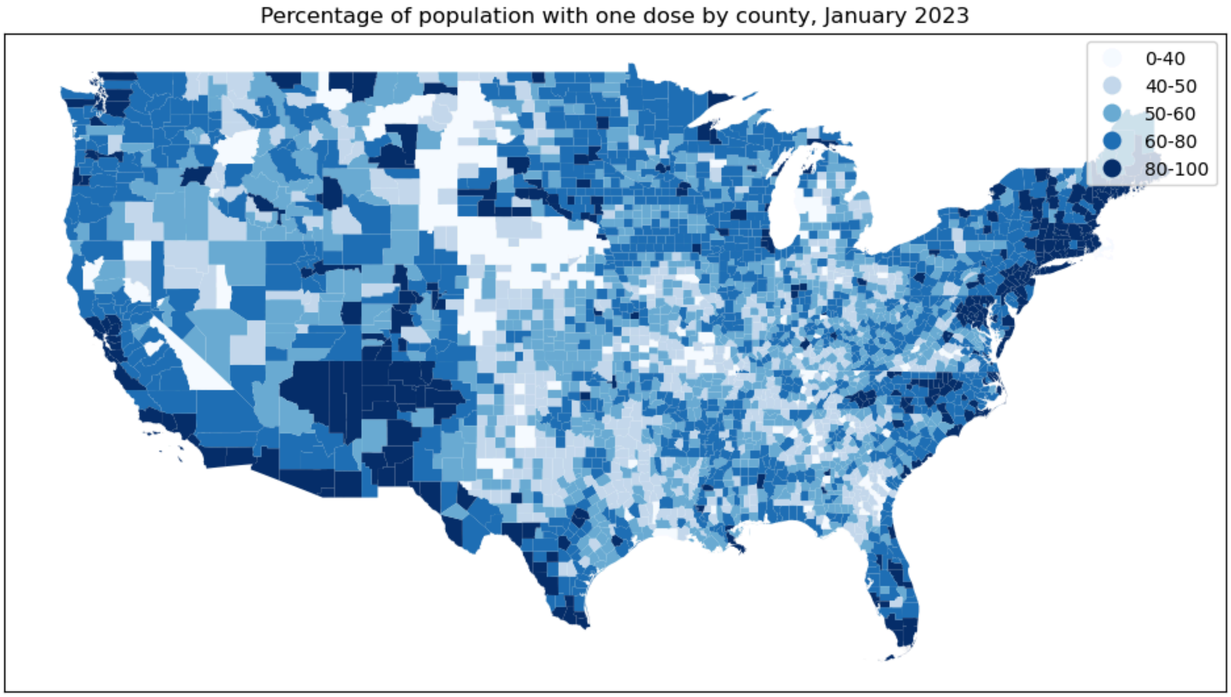
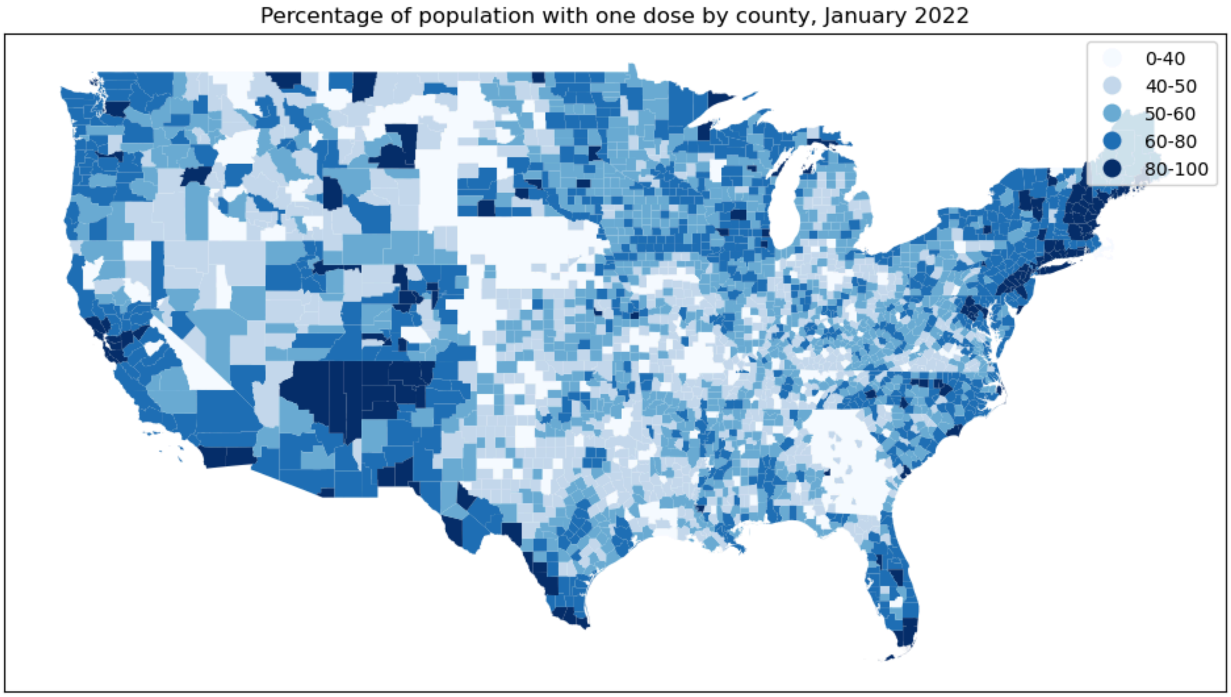

Percentage of population with one dose by county, January 2023
0-40
40-50
50-60
60-80
80-100

Figure 2


Percentage of population with one dose by county, January 2022

Covid mortality rates, on the other hand, changed drastically from 2022 to 2023. While the average mortality rate across US counties was 0.028% in 2022, it was 0.01% in 2023. Five counties contained a mortality rate of above 0.2% in 2023, but in 2022, 108 counties did. The highest mortality rates were concentrated in the Midwest, Northeast, Florida, and the Southwest in both 2022 and 2023.

Figure 3

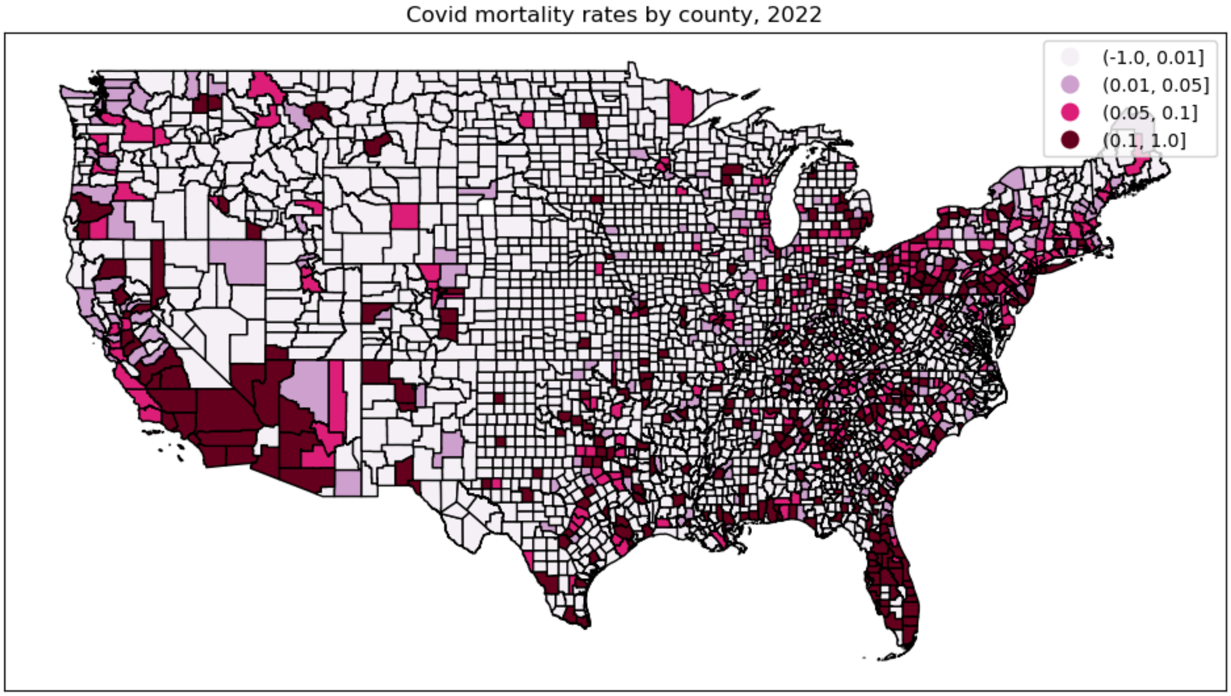
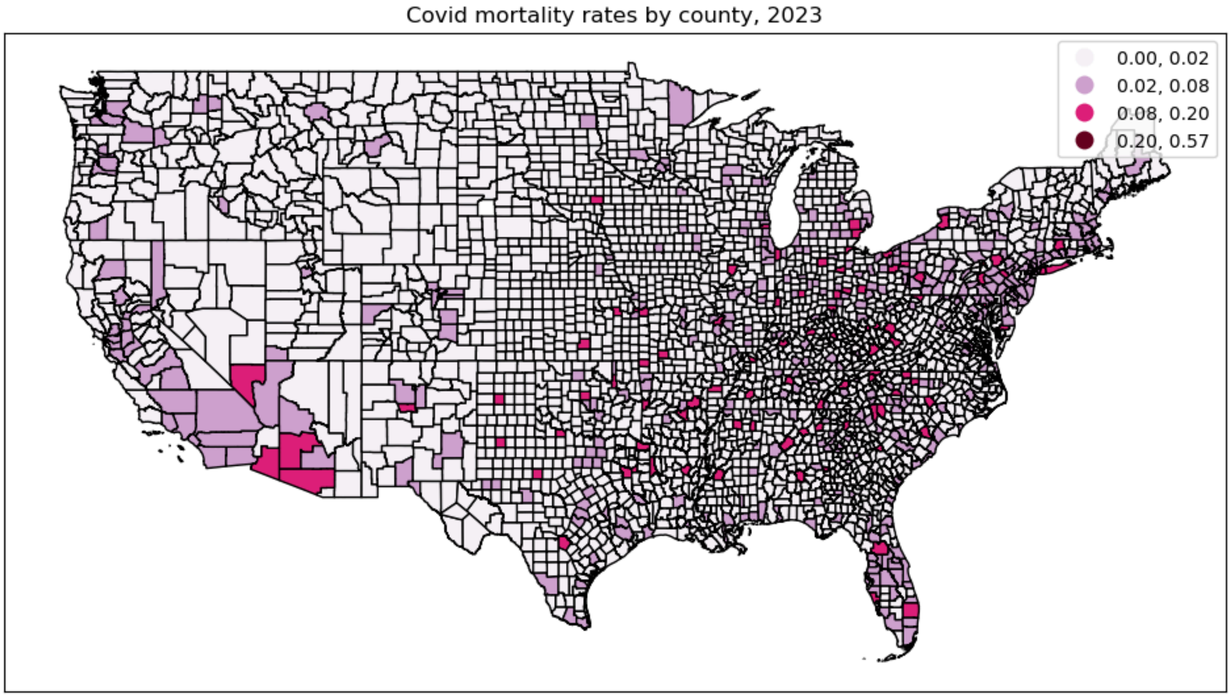
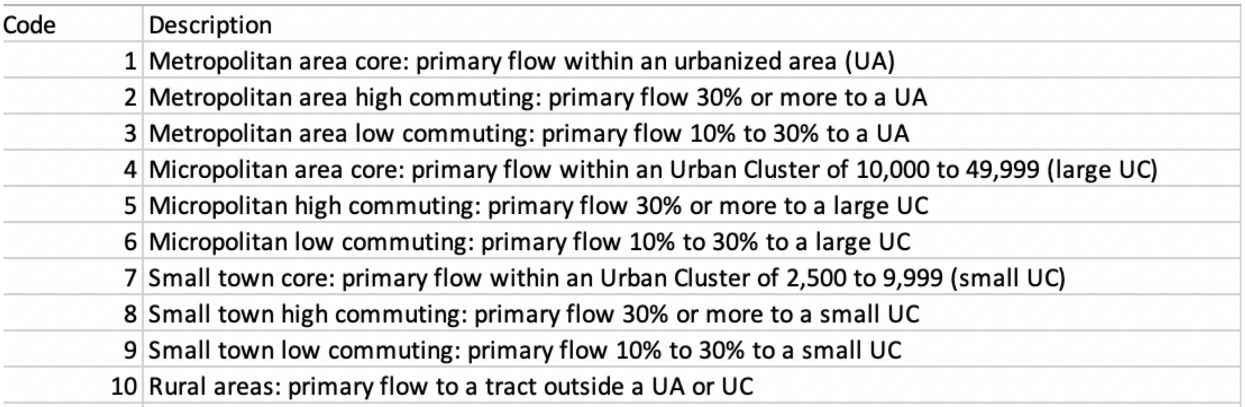


Figure 4



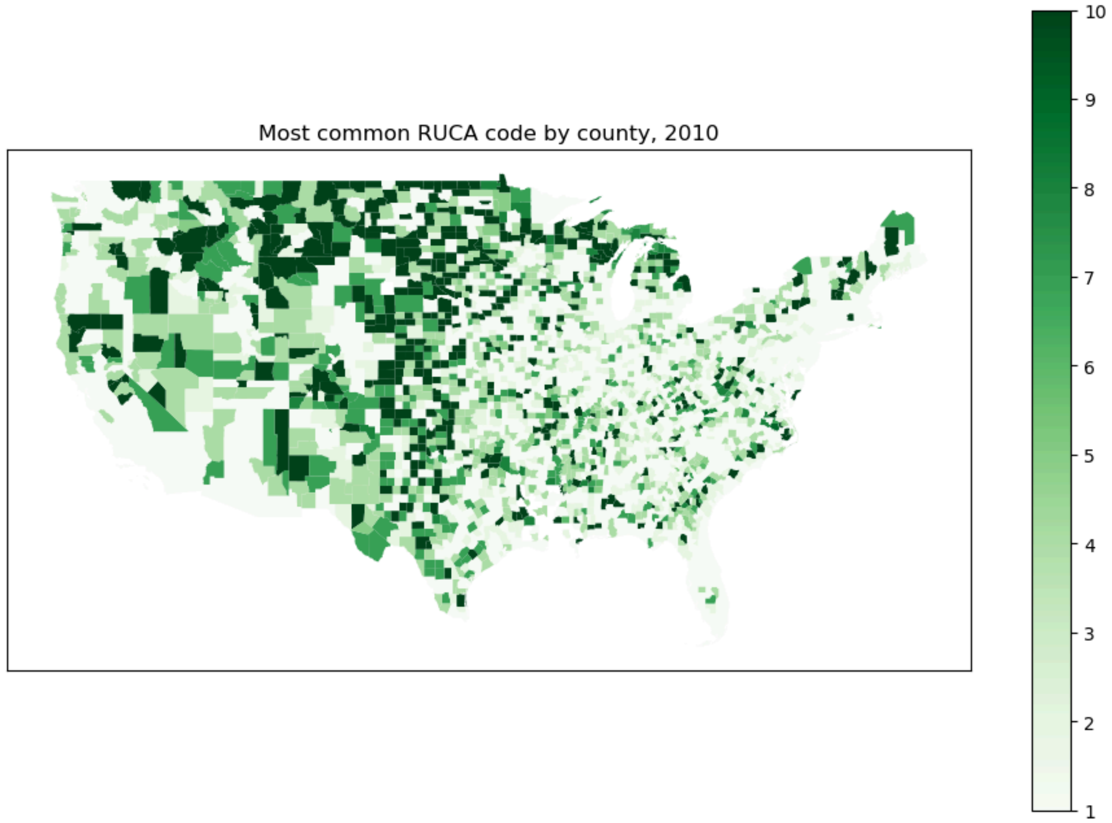
The map of rural and urban areas shows that the areas with the highest concentration of urban areas are the northeast and the Midwest, with some pockets in California and the Southeast also urbanized. The highest concentration of rural areas is located in the Great Plains and the northern Rocky Mountains.

Table 1: Description of Rural Urban Area Codes



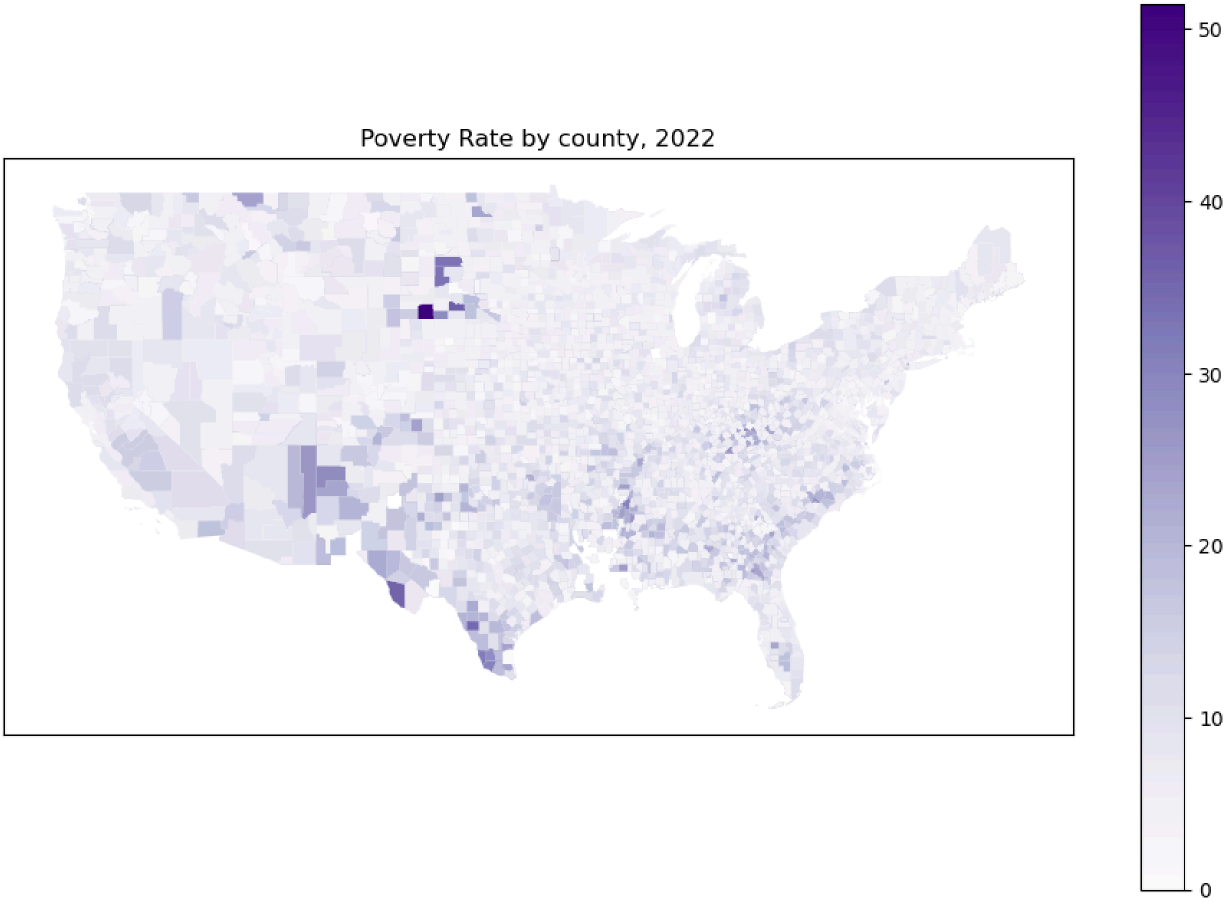
Code	Description
1	Metropolitan area core: primary flow within an urbanized area (UA)
2	Metropolitan area high commuting: primary flow 30% or more to a UA
3	Metropolitan area low commuting: primary flow 10% to 30% to a UA
4	Micropolitan area core: primary flow within an Urban Cluster of 10,000 to 49,999 (large UC)
5	Micropolitan high commuting: primary flow 30% or more to a large UC
6	Micropolitan low commuting: primary flow 10% to 30% to a large UC
7	Small town core: primary flow within an Urban Cluster of 2,500 to 9,999 (small UC)
8	Small town high commuting: primary flow 30% or more to a small UC
9	Small town low commuting: primary flow 10% to 30% to a small UC
10	Rural areas: primary flow to a tract outside a UA or UC

Figure 5



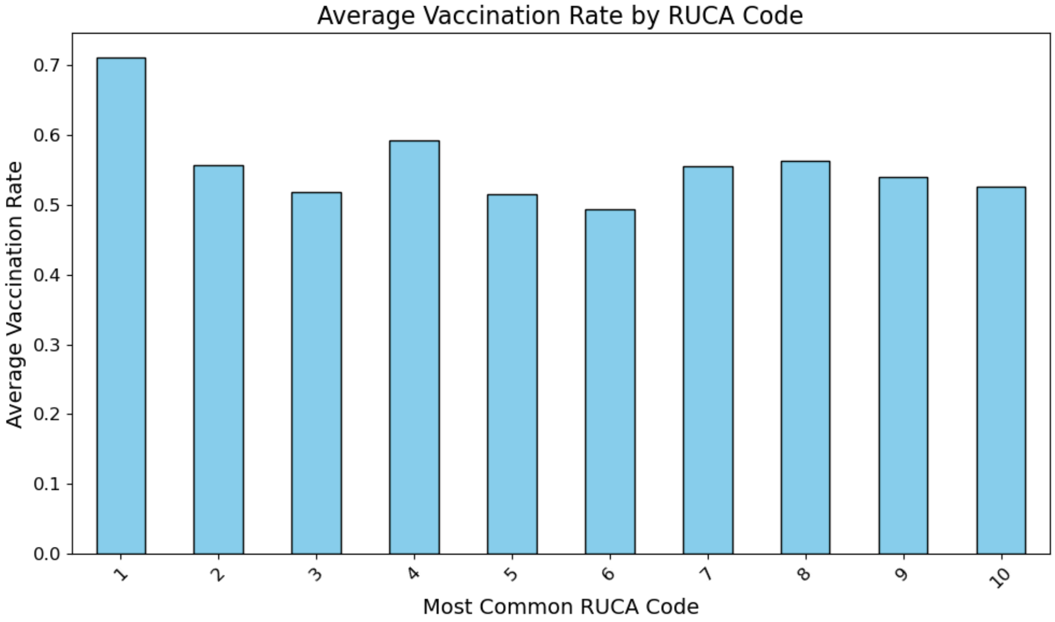
The map of poverty rates across the US reveal higher rates in the Southeast, Southwest, and the Great Plains. The lowest poverty rates are in the Pacific Coast and the northeast.

Figure 6



Without performing further analyses, these visualizations reveal a few interesting trends. The regions where vaccination rates are lowest, namely the Great Plains, are also predominantly rural areas and have moderate to high poverty rates. However, the highest Covid-19 mortality rates were in the northeast and the southwest, which are predominantly urban and have lower poverty rates in general. Based on these visualizations, a county-wide visualization reveals that lower vaccination rates are not the primary cause of lower mortality rates. Another apparent trend is that vaccination rates and urbanization rates have a slight negative relationship, shown in Figure 7. Counties with a RUCA code of 1, meaning they are heavily urbanized, have an average vaccination rate of more than 10% more than any other code. However, counties coded as rural areas do not have a significantly lower vaccination rate than those coded as small towns.

Figure 7



The results of the linear regression are shown in Table 2. The R-squared values were very low across each variable, and none of the explanatory variables were found to have significant predictive power with respect to mortality rates. The rural-urban score had an R-squared value of 0.26 for the 2022 data, and a 0.185 for the 2023 data. In general, rural counties were slightly more likely to have higher mortality rates. The percent without vaccine and the poverty rate both had an r-squared value of approximately zero. The linear regression with all features had an R-squared value of 0.249 and 0.217, for 2022 and 2023, respectively.

Table 2

Linear Regression	2022	2023
Feature	R squared	R squared
Rural-urban score	0.26	0.185
Percent without vaccine	0.057	0.095
Poverty Rate	0.0014	-0.002
All features	0.249	0.217

Figure 8

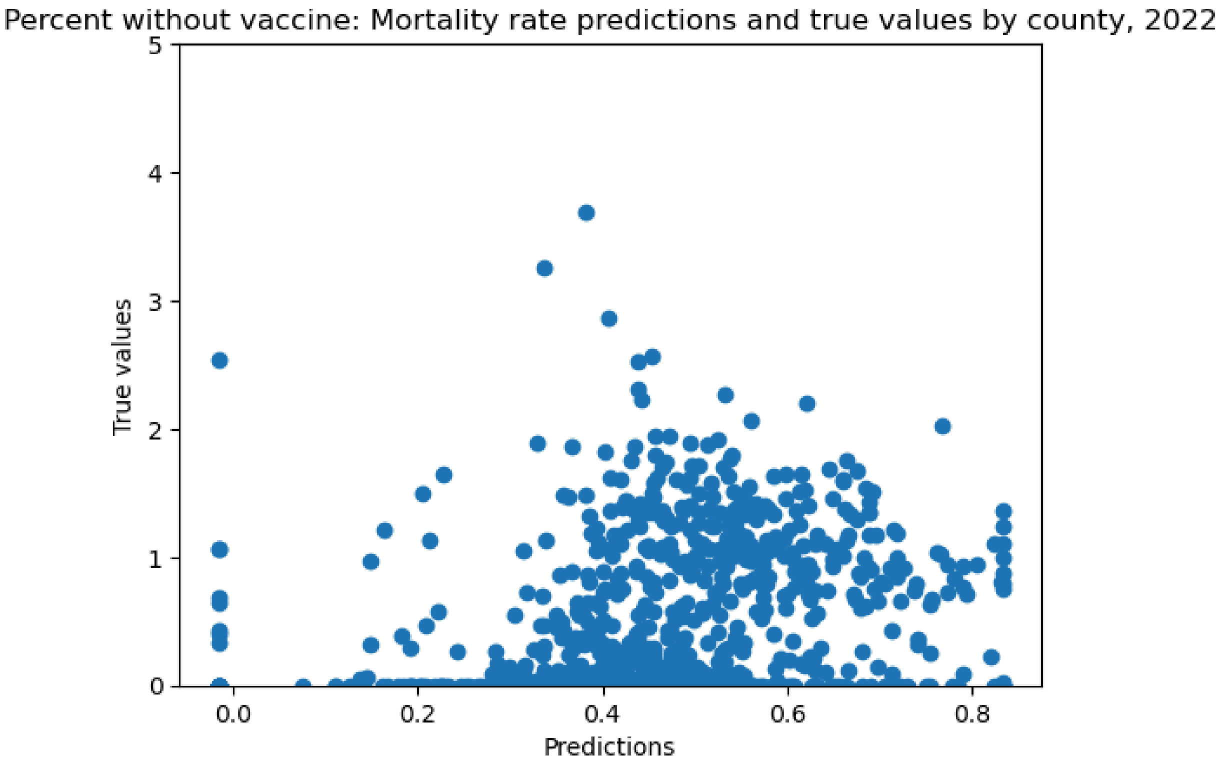


Figure 8 shows the relationship between the predicted and true values with the percentage of county populations without the vaccine in 2022. The scatter plot reveals little relationship between the percentage without vaccine and Covid-19 mortality rates. The predicted mortality rates are less variable than the true values.

Discussion

While the maps revealed interesting spatial patterns in the data, the results of this project were inconclusive and demonstrated a need for more advanced data modeling. The main goal of the project was to find out if vaccination rates had a strong effect on Covid-19 mortality. The weak relationship between the two variables and the low R-squared value from performing linear regression on the 2022 and 2023 datasets revealed that vaccination rates did not have a strong impact on Covid-19 mortality on a county level. The same goes for poverty rates and rural and urban areas. The addition of all three factors as explanatory variables into the linear regression increased the R squared value to about 0.2 for both 2022 and 2023, meaning that they explained about 20% of the variation in mortality rates.

The magnitude and distribution of Covid-19 mortality is dependent on a host of factors that are not included, and likely served as confounding variables. Age is a major confounder, because the elderly are much more likely to die of Covid. I was not able to standardize the dataset, because age was not included as a variable in the county-level mortality data. People with pre-existing conditions were also significantly more likely to die of Covid than those without, and this data was not included. While I did incorporate poverty rates to account for socioeconomic status, other variables such as race, ethnicity, and access to healthcare could also confound the data. If more relevant variables are included and the effects of confounding variables are controlled, the model's predictions will likely be more accurate.

In addition to controlling for confounding variables, more advanced data modeling should be performed. While a linear regression model is a simple and efficient way of measuring the strength of relationship between two variables, it cannot capture complex relationships between the data. A decision tree or random forest model might capture the complexity of the relationship between vaccination rates and mortality rates, for example. This study could also be expanded to account for differences in geographic variation by using a multilevel framework, which includes nested geographic relationships such as states, counties, and census tracts. Finally, spatial clustering could be computed using software such as SATScan, which can detect spatial clusters or counties with high or low Covid-19 mortality rates.

There are also characteristics of the Covid-19 vaccination and mortality rates that may have impacted the results. Covid-19 fluctuated dramatically due the spread of variants such as Delta and Omicron, which have different transmissibility and mortality rates, and are affected by vaccines in different ways. Vaccination and poverty rates are determined from sampling, which may suffer from various biases. Finally, Covid-19 spreads faster in cities than rural areas, and cities also happen to have higher vaccination rates on average. This pairing makes it difficult to study the effect of vaccination rates on mortality rates alone.

Conclusion

This project sought to measure the county-level effect of vaccination rates on mortality using a cross-sectional study from 2021 to 2022. In particular, it sought to explain this relationship using geographic factors. Despite shortcomings of the results, there are a few important takeaways from this project. The rate of first-dose vaccination increased slightly from 2022 to 2023, but mortality rates decreased drastically throughout the country. However, vaccination rates across the US cannot fully explain the decline in mortality rates on a county-level basis. Areas of the country with higher vaccination rates, such as cities in the northeast and the southwest, often experienced higher deaths than rural areas. Future studies should control for biological, socioeconomic, and environmental factors and use advanced modeling to create better predictions of Covid-19 mortality.

Works Cited

"AH Covid-19 Death Counts by County and Week, 2020-present", *National Center for Health Statistics (NCHS),* Centers for Disease Control and Prevention (CDC). 5 April 2023

"Cartographic Boundary Files- Shapefile", *United States Census Bureau,* 2018

"Covid-19 Vaccinations in the United States, County", *National Center for Health Statistics (NCHS),* Centers for Disease Control and Prevention (CDC), 12 May 2023,

Cuadros, Diego F et al. "Dynamics of the Covid-19 epidemic in urban and rural areas in the United States", *Annals of Epidemiology,* Volume 59, 2021, p.16-20

"Families below poverty by State", *HDPulse:* National Institute of Minority Health and Health Disparities (NIH), 2018-2022

Moghadas, Sayed et al. "The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States", *Clinical Infectious Diseases*, Volume 73, Issue 12, 15 December 2021, p. 2257–2264

"Rural-Urban Commuting Area Codes", *United States Department of Agriculture (USDA),* 2020

"United States Covid-19 Community Levels by County" *CDC Covid Response,* Centers for Disease Control and Prevention (CDC), 28 June 2024