Relationships Between Nighttime Light, Covid-19, and Economic Variables in the United States

Katherine Marquis

Covid-19 has caused over 1 million deaths in the United States by May 2022 (WHO Coronavirus). Nighttime light (NTL), as used as an indicator of human activity, may reflect Covid-19 transmission and severity. This study attempts to assess the relationship between NTL and Covid-19 by assuming that high NTL radiance levels indicate a large amount of human contact to spread Covid-19 and that significant drops in NTL radiance levels correspond to quarantine or deaths due to Covid-19. In order to examine this relationship, NTL monthly composite satellite images from 2018-2023 were collected from NASA, county information was collected from the U.S. Census Bureau, Covid-19 case and death statistics at the county level were provided by Johns Hopkins, Covid-19 vaccination information was collected by the CDC, and economic variables were collected by Opportunity Insights. After preprocessing this data, various regression models were applied to the data to analyze relationships.

1. Introduction

This project aims to use machine learning models to obtain economic indicators from satellite images in order to study the spatial economic impacts of Covid-19 policies. The Covid-19 pandemic caused massive disruptions, as there are a cumulative total of over 750 million confirmed cases and over 6.5 million deaths globally as of February 2023. In the United States alone, there is a cumulative total of over 100 million confirmed cases and over 1 million deaths (WHO Coronavirus). This global emergency has significant impacts on people's health, the economy, the job market, and daily life across the country (Rubinyi et al., 2020). Unlike local governments and health departments, which may be biased or inaccurate in their reporting of statistics, satellite images should accurately reveal changes in economic activity due to Covid-19 because they are a reliable source of information on population dynamics, energy consumption, social interaction, urbanization, and infrastructure (Rubinyi et al., 2020).

Light usage can be used as a proxy for economic activity, so nighttime light (NTL) satellite images can be used to accurately examine population dynamics and economic infrastructures during Covid-19. The plausibility that light data corresponds to economic activity follows from the assumption that light is a normal good, and thus follows the same logic of earlier work that uses consumption decisions as a proxy for income (Donaldson & Storeygard, 2016). Light usage can be measured through satellite images taken at night, which capture the visible and infrared electromagnetic radiation from the Earth's surface. These NTL images depict urban areas and transportation networks as bright spots against a dark background, and thus we can measure the intensity of the light in the images to assess economic activity in the area (Stathakis et al., 2021). NTL satellite imagery has already been used to successfully estimate population density,

measure GDP, and model urbanization (Zhang et al., 2022; Henderson et al., 2009). Due to the correlation between human activity and light, there should be a change in the NTL satellite image data during the Covid-19 pandemic lockdown. This project aims to investigate, visualize, and quantify changes in economic activity during Covid-19 by analyzing NTL satellite images using statistical methods and machine learning.

2. Literature Review

Previous work uses NTL images to model urbanization, calculate economic variables at the national and subnational level, model resilience during natural disasters, and find correlation between measles transmission and population density. From this, NTL has been applied to the Covid-19 pandemic in limited capacity. Various papers that use NTL images prior to Covid-19 are summarized below.

NTL images from 1992 to 2008 of India, China, Japan, and the United States captured by the Operational Linescan System from the Defense Meteorological Satellite Program (DMSP/OLS) have been used to model urbanization by Zhang and Seto. After testing the ability of NTL images to map urbanization dynamics by finding correlations, unsupervised machine learning clusterings, such as the Isodata clustering algorithm, were applied to model urbanization dynamics (Zhang & Seto, 2011).

Henderson, Storeygard, and Weil use NTL images from DMSP/OLS processed by NOAA to model income growth measures. They develop a statistical framework to estimate GDP growth from NTL data, which is based on the assumption that there is a simple constant elasticity relationship between total observable lights and total income. To predict income growth, they develop best fit elasticity of measured GDP growth with respect to NTL growth and produce estimates. Then, this is applied to predict GDP growth on subnational and supranational areas (Henderson et al., 2009).

To get a deeper understanding of how NTL images can be used as a proxy for economic activity and development, Bruederle and Hodler studied the correlation of multiple specific indicators of human development and NTL intensity. The NTL data from the National Oceanic and Atmospheric Administration (NOAA) was combined with geo-coded Demographic and Health Surveys for 29 African countries from 2991 to 2013 in order to be analyzed using linear regression and various numerical transformations. They concluded that NTL intensity is an indicator of a wealthy, educated, and healthy population as NTL intensity is correlated with household income, education, and health (Bruederle & Hodler, 2018).

In addition to finding that NTL intensity correlates with socioeconomic characteristics and human dynamics, they can be used to model and predict economic resilience. NTL images collected by DMSP/OLS from 1992 to 2013 have been used to model economic resilience during

Hurricane Katrina. A statistical framework was introduced that was able to model the economic recovery from the hurricane, and statistical analyses were carried out to determine which socioeconomic factors affect recovery the most (Qiang et al., 2020).

Further, DMSP/OLS NTL images from 2000-2004 were used to show that measles seasonality can be explained by spatiotemporal changes in population density, as modeled by satellite images of light usage. It was found that measles transmission and population density are highly correlated. The study demonstrated that population density is highly important for predicting epidemic progression at the city level, and the fine measurement of population density has implications for public health, crisis management, and economic development (Bharti et al., 2011).

Since NTL has been shown to correlate with various socioeconomic factors at the national and subnational level and it has successfully been utilized in research related to natural disasters and illness infection and transmission, some previous work has been done in applying NTL images to Covid-19 research.

Most notably, a correlation between NTL intensity and Covid-19 incidence and mortality rate was found. The paper used monthly cloud-free NTL images for the contiguous United States in 2019 and 2020 from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day Night Band (DNB) on the Joint Polar-orbiting Satellite System, as provided by the Earth Observation Group of the Payne Institute for Public Policy. For 2019 and 2020, the average NTL intensity over the year was calculated for each county. Since NTL represents the intensity of human activity, NTL should change from 2019 to 2020 due to most counties limiting social interaction for the majority of 2020, and NTL should be correlated with Covid-19 cases and incidence. To assess this, a negative binomial mixed model was used, and it was found that NTL intensity is correlated with both Covid-19 incidence and mortality (Zhang et al., 2022).

For each municipality in Greece, VIIRS DNB monthly images were compared before Covid-19 from 2017-2019 and during Covid-19 in 2020. The sum of light intensity of these NTL images was compared across cities and it was found that dense urban areas exhibit less change in NTL, likely because of emissions from a large established infrastructure, and areas that rely on tourism exhibited a large decrease in NTL during the pandemic (Stathakis et al., 2021).

Additionally, the Black Marble VIIRS NTL images from 2012 through 2020 were combined with economic and electricity data to predict India's national GDP during Covid-19 using various machine learning methods, in which Lasso Regression was most effective. Thus, NTL images may be used to calculate GDP in the short term or in areas where there is not updated or accurate GDP information (Dasgupta, 2022).

Overall, past research using NTL images is applicable to expected and observed changes in human dynamics during Covid-19. Past research focuses largely on finding correlations between NTL and other types of data, mostly economic variables or human dynamics. Based on past research, some application of NTL images in Covid-19 research has been done, but this project aims to expand the scope of it.

3. Data

3.1 Data Sources

3.1.1 Counties

The list of counties was taken from the dataset "fips2county," which contains information on all counties in the United States. Each row in the table contains a state FIPS, 3-digit county FIPS, county FIPS, county name, state name, state abbreviation, and a state-county combined name (Connell, 2022).

3.1.2 NTL Images

The NTL images used in this project were from the "VIIRS Nighttime Day/Night Band Composites Version 1" dataset, which was accessed using Google's Earth Engine API and was provided with the processed images by the Earth Observation Group from the Payne Institute for Public Policy. The images used were altered by the provider using the VIIRS Cloud Mask (VCM) to adjust for low quality images due to cloud cover. Two bands accompany each image, that is each pixel contains two types of information. One is the average radiance band, which gives the radiance value for each pixel as calculated by averaging the usable observed radiance values for said pixel throughout the month. Most calculations are done using this band. The second is the number of cloud-free coverages, which is the number of observations for each pixel used to calculate the average radiance, as some are filtered out by the VCM. This band can be used to determine which areas of the image are more reliable and which areas have fewer observations (Earth Observation Group).

3.1.3 County Shapefiles

The geospatial for each United States county was retrieved using the "TIGER: United States Census Counties 2018" dataset through access with Google's Earth Engine API. Each row in the dataset represents a county and contains the land area, water area, latitude, longitude, FIPS code, and various legal identifiers for the county (U.S. Census Bureau).

3.1.4 Covid-19 Cases and Deaths

The Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University was used to find the number of Covid-19 cases and deaths by county each month. The data was accessed through their GitHub repository, where the confirmed cases, confirmed deaths, recovered cases, active cases, incidence rate, and case fatality ratio are provided daily from January 22, 2020 through March 9, 2023 at the county level (or equivalent).

The dataset also includes the location (county, state, and country) as well as geographical location (latitude and longitude) for each entry (Dong, Du, & Gardner, 2020).

3.1.5 Vaccinations

The Center for Disease Control (CDC) dataset titled "COVID-19 Vaccinations in the United States,County" was used for vaccination data at the county level. It provides administrative data such as the date of reporting to the CDC Covid-19 Data Tracker, the County Federal Information Processing Standard State Code (FIPS), the Morbidity and Mortality Weekly Report (MMWR) week, the recipient county, the recipient state, the 2019 Census county population, the Social Vulnerability Index (SVI) of the county, and the completeness percentage (the proportion of vaccinated individuals who have a valid and reported FIPS code to all vaccinated individuals). Each data entry has the previously mentioned administrative information along with Covid-19 vaccination information such as the number of people who have completed a vaccination series (defined by those having the second dose in a two-dose series or having the first dose in a one-dose series), the percentage of the county population who has completed a vaccination series, the number of people who have completed their primary series and received a booster dose by county, the percentage of the county population who have completed their primary series and received a booster dose, the percentage of vaccinated individuals with respect to SVI, and the percentage of vaccinated individuals with a booster dose with respect to SVI. Additionally, the dataset has entries of the above categories separated by age. The age breakdowns are 5 years or older, 12 years or older, ages 5-17, 18 years or older, and 65 years or older (COVID-19 Vaccinations).

3.1.6 Policies

The CDC "COVID-19 State and County Policy Orders" dataset was used for tracking changes in major policies at the state and county level. Each row in the dataset signifies the start or stop of a policy, and the attributes given are state abbreviation, county (if applicable), FIPS (if applicable), policy level (state or county), date, policy type, start or stop, comments, source, total phases, and geocoded state (COVID-19 State and County Policy).

3.1.7 Economic Variables

The economic data used in this paper was provided by the Opportunity Insights Economic Tracker. The data was accessed through their GitHub repository, which has multiple datasets. One such dataset, titled "Affinity - County - Daily," is from Affinity Solutions and provides credit and debit card spending at the county level daily by listing the year, month, day, FIPS code, frequency of updating the data, spending relative to January 2020, and an indicator to mark that the number provided for that entry is up to date. Another dataset, which is provided by Lightcast, is titled "Job Postings - County - Weekly" and lists by county, state, FIPS code, and week the number of job postings, the number of job postings that fall into ONET job zone levels 1 and 2, and the number of job postings that fall into ONET job zone levels 3, 4, and 5. Third,

the data frame titled “Employment - County - Weekly” was created using Paychex and Intuit data and for each entry containing a year, week, county, and FIPS code the row includes employment rates relative to those of January 4-31, 2020 for all workers, the workers in each of the four quartiles of wage distribution, and the upper, lower, and middle two quartiles of wage distribution. Finally, the data frame titled “UI Claims - County - Weekly” gives the count of unemployment insurance claims and the rate of unemployment insurance claims per 100 people in the 2019 workforce for each entry of county, FIPS, state, and week in the dataset (Chetty et al., 2023).

3.2 Data Collection and Preprocessing

3.2.1 County-Level Covid-19 and Economic Data Collection and Preprocessing

The first step in data collection and preprocessing was to analyze the “fips2county” data frame. After confirming that it contains all counties in the United States and has no missing information, any rows corresponding to Hawaii or Alaska were dropped because this project is only considering the contiguous United States. Only the county name, state abbreviation, and FIPS code columns were kept for each entry.

With this list of locations (counties, states, and FIPS codes) to use, the CSSE Covid-19 Data Repository was processed. Each dataset represents a day’s worth of information, so the datasets corresponding to the last day of each month from March 2020 through February 2023 were considered. When looping through these datasets, the first step is to determine if the dataset covers all counties from the list previously assembled. When checking if the CSSE Covid-19 Data Repository covered all counties, it should be noted that there are multiple counties for which no data was recorded near the beginning of Covid-19 (March 2020 and/or a few months following). Many counties in Utah and two in Massachusetts had no data recorded at all. Given that the counties with no information at all and the missing few rows for other counties is a small amount of data compared to entries for all counties and for all months from March 2020 through February 2023, the rows with missing data were dropped. So the data frame now contained a row for each date and county combination that had information on cases and deaths in the CSSE Repository, where the dates were the months from March 2020 to February 2023 and the counties were all counties in the contiguous United States. Each row had the date, county, state, FIPS, total cases, new cases that month, total deaths, and new deaths that month.

The vaccinations data needed to then be collected. Since the CDC provides a large dataframe, removing unnecessary rows and columns makes it more manageable to then analyze and combine with the existing data frame. To first reduce the “COVID-19 Vaccinations in the United States,County” data frame, the columns of unnecessary attributes were dropped. The columns kept were date, FIPS, county, state, population, number of people with a completed primary series, population percentage with a completed primary series, number of people with a booster dose, population percentage with a booster dose, SVI category, population percentage with a

completed primary series with respect to SVI, and population percentage with a booster dose with respect to SVI. Then, the FIPS codes were converted to integers, as they were strings in the original data, and all rows with unknown FIPS codes or counties in Hawaii or Alaska were dropped. The dates were converted to integer days, months, and years because the original dates came as strings. For every county, the last data entry for each month of said county was kept and the rest were dropped, so the most updated data for the county was kept to represent each month.

Now that the data frame was significantly smaller and more manageable, it could be combined with the previous working dataset. The first step was to check that all counties in the working dataset had data in the reduced vaccinations dataset, which they did. Then, the information for each county and month combination from the vaccinations data was added to the working dataset. The vaccination dataset was not complete, so any missing information from that was filled in with NaN.

Next the policy data was preprocessed. Because of the variety of policy type labels and syntax, only stay at home orders were analyzed. To do this, the data frame was reduced to only the rows that contain “Shelter in Place” as the policy type. This information was added to the working dataset as a column of 0s and 1s, where 0 meant there was no stay at home order for the county and month that the row of the dataset represents and 1 meant there was a stay at home order at that time and place. If no stay at home order was issued or lifted at the county level, the state’s stay at home order start and stop dates were used.

Then, the various economic variables were preprocessed and added to the working data frame. From “Affinity - County - Daily,” the amount of credit and debit card consumer spending daily relative to January 2020 was added to the data frame. It covered 1704 of the counties of interest. From “Job Postings - County - Weekly,” the number of job postings relative to January 2020 was added to the data frame. It covered 3142 of the counties of interest. From “Employment - County - Weekly,” the relative employment rate as compared with January 2020 was added. It covered 1999 counties of interest. From “UI Claims - County - Weekly,” the number and rate of unemployment insurance claims were added to the data frame. It covered 1425 counties of interest. For each dataset, if a county was included in the dataset then the latest information from each month was added to the data frame. If a county was not included in a dataset, then NaN values were added to the data frame for that county.

This finalized the working data frame with monthly county-level Covid-19 and economic statistics.

3.2.2 County-Level NTL Data Collection and Preprocessing

The next step was to collect the monthly NTL Images and their numerical attributes for all the dates and counties of interest, as used in the construction of the previous data frame. Using

Google Earth Engine's NTL dataset titled "VIIRS Nighttime Day/Night Band Composites Version 1," all NTL images from January 2018 through March 2023 were retrieved. Each NTL image is a monthly composite, so this was a collection of 63 monthly composite NTL images. Using the list of counties generated earlier and the "TIGER: US Census Counties 2018" dataset, the county shapefiles were retrieved and each NTL image was clipped to the boundary of each county. Thus, for each county of interest there existed a collection of 63 monthly composite NTL images dated from January 2018 through March 2023 in which each NTL image was clipped to only contain data from within the county boundary. For each image, the average radiance band was used to calculate mean radiance, standard deviation of radiance, sum of lights, and standardized sum of lights all at the county level. This information was saved in the format of a data frame, where each row was an entry for a date (by month and year) and a county and the other attributes of the row were the mean, standard deviation, sum of lights, and standardized sum of lights.

3.2.3 Geospatial Data Collection and Preprocessing

The "TIGER: US Census Counties 2018" dataset contains a column specifying the land area for each county, so the data frame was reduced to only have two columns, which gave the FIPS code and land area in square meters. Since county land area in square meters is a large number and thus would be inconvenient for calculations, another column was added to the data frame for land area in square miles. This was saved as a data frame for future use in preprocessing and feature engineering.

3.3.3 Final Preprocessing

The final step in preprocessing was to combine all relevant data frames. This meant combining Covid-19 statistics, economic statistics, NTL data, and county data all into one data frame. In doing this, it left a final data frame where each row represented a specific date and county with various county-level statistics. The columns of the combined data frame are listed below.

```
['Date', 'County', 'State', 'FIPS', 'Total_Cases', 'Total_Deaths',
'New_Cases', 'New_Deaths', 'Month', 'Year', 'SVI', 'Series_Complete_Num',
'Series_Complete_Pct', 'Series_Complete_SVI', 'Booster_Complete_Num',
'Booster_Complete_Pct', 'Booster_Complete_SVI', 'Population',
'Stay_at_home_order', 'Spending', 'Job_Postings', 'Employment_Rate',
'UI_Count', 'UI_Rate', 'NTL_Mean', 'NTL_Std_Dev', 'NTL_SOL',
'NTL_SOL_Standardized', 'County_Land_Area_mi2']
```

This was saved as one large data frame for all entries from January 2028 through March 2023. All Covid-19 statistics were filled in as NaN for all dates before March of 2020.

3.3 Feature Engineering

In order to maximize the effectiveness of predictive models, finding the optimal relationship(s) between variables is crucial. Combining features in various ways to create more allows for more relationships between variables and can lead to more effective models. Thus, various features were created using the existing ones collected from the data sources.

3.3.1 Feature Engineering with Covid-19 Statistics

First, Covid-19 incidence rate, death rate, and case fatality ratio were calculated and added to the data frame. Their formulas and definitions are described below.

Incidence Rate (*new_cases/population*): the number of new Covid-19 cases in a month per person in a county

Cumulative Incidence Rate (*total_cases/population*): the number of cumulative Covid-19 cases through a specific date per person in a county

Death Rate (*new_deaths/population*): the number of new Covid-19 deaths in a month per person in a county

Cumulative Death Rate (*total_deaths*): the number of cumulative Covid-19 deaths through a specific date per person in a county

Case Fatality Ratio (*new_deaths/new_cases*): the proportion of Covid-19 cases that are fatal in a specific month and county

Cumulative Case Fatality Ratio (*total_deaths/total_cases*): the proportion of Covid-19 cases that are fatal in a county through a specified date

3.3.2 Feature Engineering with Geospatial County Data

Since Covid-19 spreads through close contact, areas with many people living in small areas likely have relatively high transmission of Covid-19. Thus, NTL, Covid-19 cases, and other statistics of interest may correlate with population density. It was calculated according to the description below and added to the data frame.

Population Density (*population/county_land_area*): the number of people living per square mile in a county according to the 2018 United States Census

3.3.3 Feature Engineering with NTL

Various NTL statistics can portray different aspects of human dynamics. So, manipulating NTL data in various ways could lead to better models. The engineered features involving NTL mean radiance, standard deviation, sum of lights, and standardized sum of lights are below.

Normality Divergence (*(sol - $sol_{2018\text{-}2019\ avg}$)/$sol_{2018\text{-}2019\ avg}$*): the change between the sum of lights for a specific month in 2020-2023 and the average sum of lights value for that same month in 2018-2019, and then it normalizes it relative to that month's average sum of lights value in 2018-2019

Mean with Respect to Population (*mean/population*): the average monthly NTL radiance normalized by population

Standard Deviation with Respect to Population (*std_dev/population*): the monthly standard deviation of NTL radiance emitted normalized by population

Sum of Lights with Respect to Population (*sol/population*): the amount of light emitted per person in a county

Standardized Sum of Lights with Respect to Population (*stand_sol/population*): the standardized amount of light emitted per person in a county

Normality Divergence with Respect to Population (*norm/population*): the normalized change in light divided by the population

Mean with Respect to Population Density (*mean/pop_density*): the average monthly NTL radiance normalized by population density

Standard Deviation with Respect to Population Density (*std_dev/pop_density*): the monthly standard deviation of NTL radiance emitted normalized by population density

Sum of Lights with Respect to Population Density (*sol/pop_density*): the amount of light emitted per person in a county density

Standardized Sum of Lights with Respect to Population Density (*stand_sol/pop_density*): the standardized amount of light emitted per person in a county density

Normality Divergence with Respect to Population Density (*norm/pop_density*): the normalized change in light divided by the population density

3.4 Standardization

The final step in processing the data was to standardize it. The data frame with all collected, preprocessed, and engineered statistics was standardized to contain all values between -1 and 1

for interpretability of model output. A data frame with the same information from between March 2020 through March 2022 was also standardized.

3.5 Data Validation and Visualizations

Before any models can be used, it is important to verify that the aggregate data is correct. The economic data is difficult to validate with another source, but the total Covid-19 cases and aggregate change in NTL can be visualized and verified.

3.5.1 Covid-19 Cases Visualization and Validation

Comparing the aggregate Covid-19 case data to other sources can be used to verify that the collected data is correct.

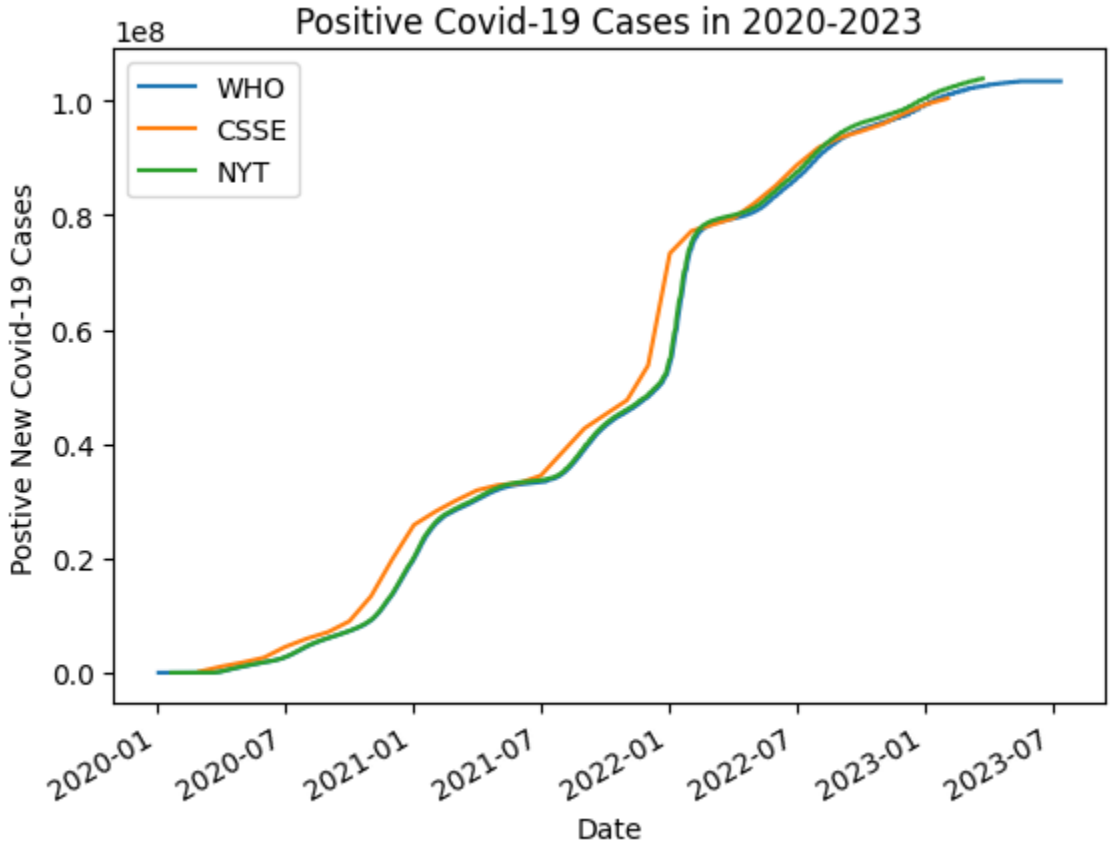


Fig. 1: Total positive Covid-19 cases in the United States in 2020-2023 as reported by the World Health Organization (WHO), Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (CSSE), and the New York Times (NYT). Since all accounts are similar, it can be assumed that the data is accurate.

3.5.2 NTL Data Visualization and Validation

The basis of this project assumes that NTL is correlated with human activity. Confirming this association is present in the data is therefore important. Since there are no easily accessible

sources to compare NTL data, examining the images and observing an aggregate change in NTL data during quarantine will show the correlation between human activity and NTL radiance.

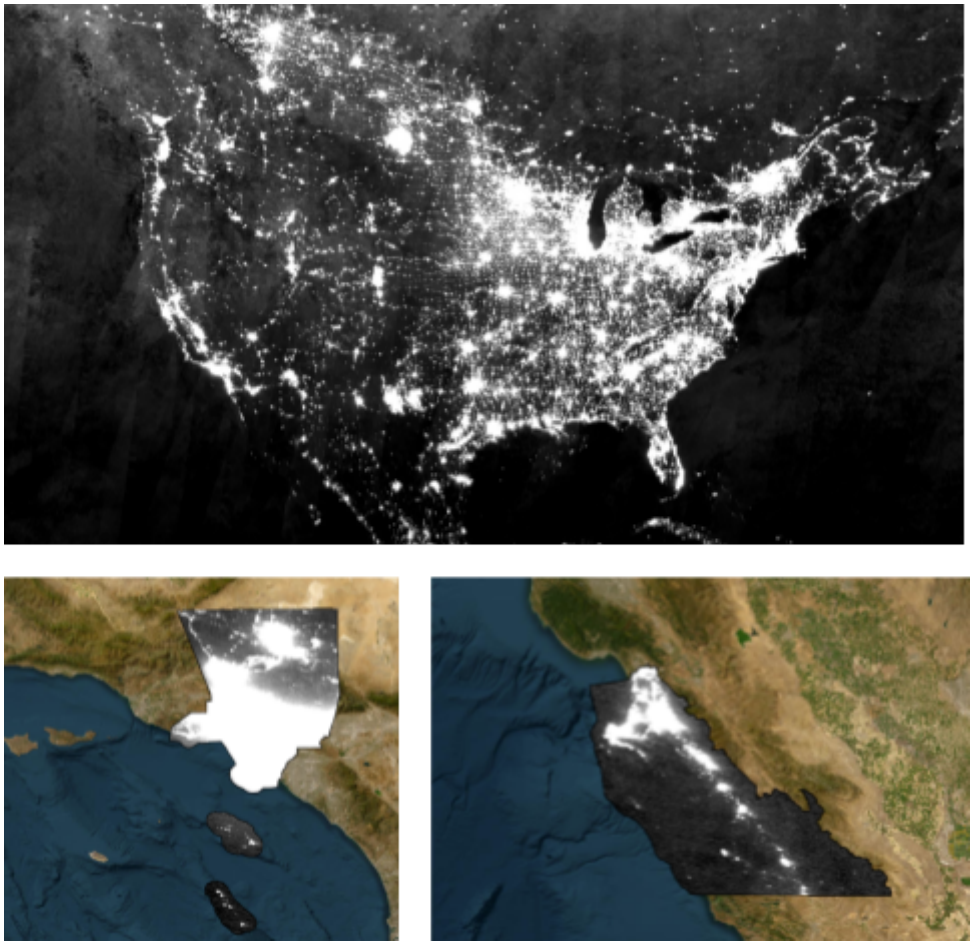


Fig. 2: Various NTL monthly composite images for January 2020. The locations shown are the contiguous United States (top), Los Angeles County, California (bottom left), and Monterey County, California (bottom right). The difference in amount and intensity of light between Los Angeles County and Monterey County is likely due to population. Los Angeles emits much more light because there are more people living there.

Now that a plausible correlation can be seen between NTL and population, the trend should be confirmed numerically.

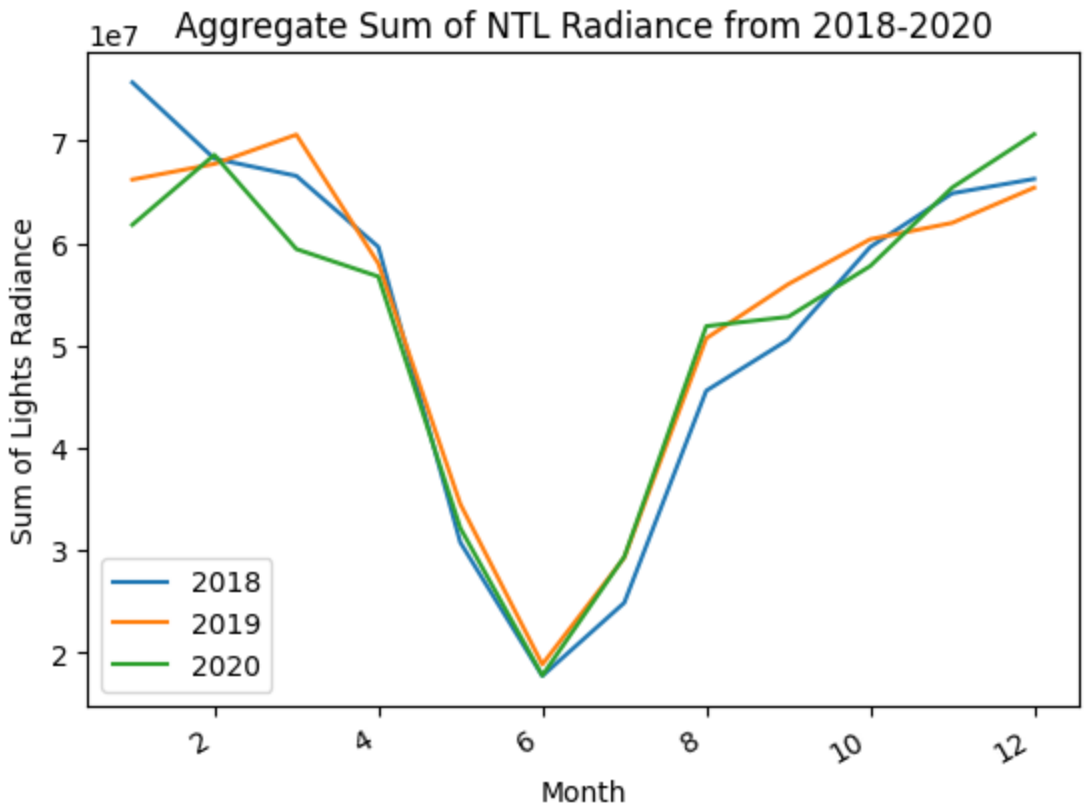


Fig. 3: Monthly sum of lights for the contiguous United States in 2018-2020. A significant drop in NTL radiance can be seen during quarantine in March 2020. The lack of change in other months can be explained by urban areas dominating an aggregate sum of lights calculation. Since urban areas have large populations, lack of light due to Covid-19 cases and deaths is not prominent. This trend can be observed and confirmed by Fig. 3. The consistent change in seasonal NTL sum of lights is likely due to human activity and environmental conditions, such as vegetation cover in the summer and reflective snow in the winter (Wang et al., 2021).

The aggregate change in sum of lights has been observed, but a better trend can be visualized. NTL has previously been used to show human activity, so in areas of lower population there should be a decrease in light if there is a decrease in activity due to high levels of Covid-19. Splitting the counties into three categories by population should show a difference in NTL response to Covid-19 based on the population. If these categories are labeled urban, semi-urban, and rural, the aggregate response to low, medium, and high levels of Covid-19 can be observed.

Fig. 4: Change in sum of lights between 2018-2019 and 2020 for rural, semi-urban, and urban counties. Counties that are considered rural are those that fall in the lowest third by population, counties that are considered semi-urban are those that fall in the middle third by population, and counties that are considered urban are those that fall in the highest third by population. Each month, the counties in each urbanization category are split into thirds based on monthly Covid-19 cases and these are deemed the low, medium, and high Covid-19 levels. The aggregate change in sum of lights from the average 2018-2019 value and 2020 value is calculated and plotted by Covid-19 severity for rural areas (top left), semi-urban areas (middle left), and urban areas (bottom left). It can be seen that in rural counties, high numbers of Covid-19 cases correlate with drops in NTL, and in urban counties, high numbers of Covid-19 correlate with increases of NTL. The difference between the change in NTL for high levels of Covid-19 and low levels of Covid-19 is plotted for each county type. This is shown monthly for rural counties (top right), semi-urban counties (middle right), and urban counties (bottom right) by the black line, and the purple line is the average value of said line. This again confirms that the change in NTL during the Covid-19 pandemic is negative in rural areas and positive in urban areas.

From these graphs and images, we can confirm that the data shows a correlation between NTL and human activity, but the change of NTL during Covid-19 is reliant on the population of the county. When looking to quantify relationships further, the urbanization of the area must be taken into account.

3.5.3 Data Visualizations

After validating the accuracy of the aggregate data and confirming a general trend, other visualizations can be done with the data to observe relationships between statistics and ensure the data makes sense.

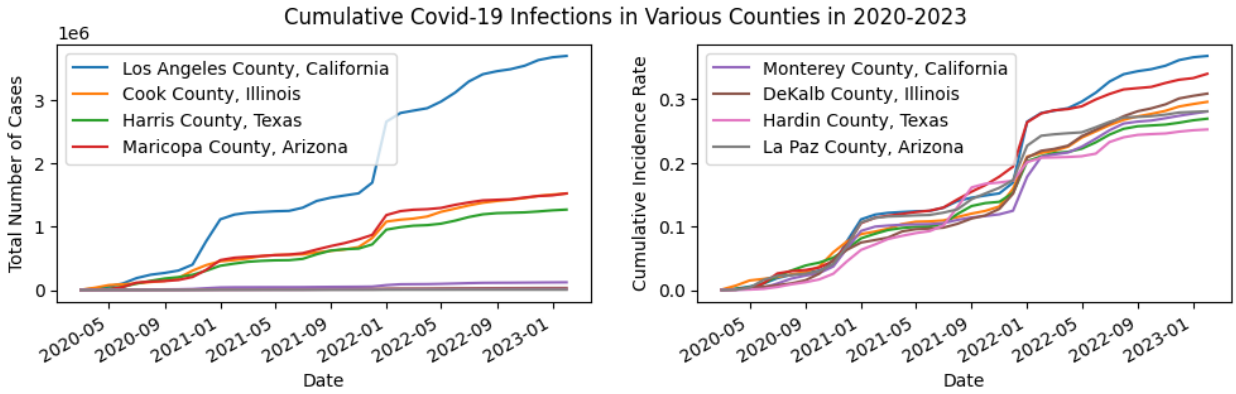


Fig. 5: Total number of Covid-19 cases and incidence rate in the four largest counties in the United States and smaller counties in their respective states. The left graph shows the total number of cases over time, and the right graph shows the cumulative incidence rate over time. An extreme difference in the total number of Covid-19 cases is consistent with common assumptions because it is largely dependent and limited by county population. When accounting for population size and graphing incidence, which is total cases per person in the county, the rates are similar. Many of the higher incidence rates still correspond to the counties with the largest populations, but this is acceptable because there will be more opportunity to transmit Covid-19 in densely populated urban areas.

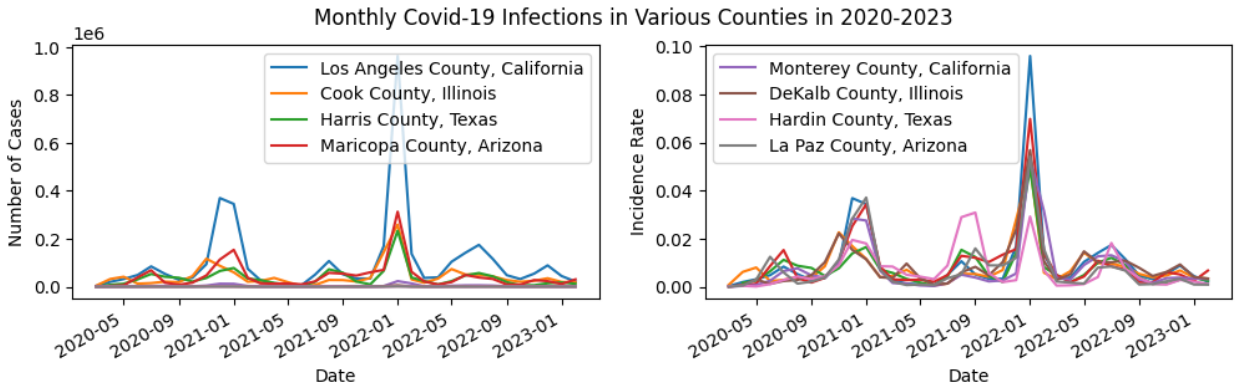


Fig. 6: The monthly number and incidence rate of Covid-19 in various counties in the United States. These counties are the same as shown in Fig. 4, and this graph shows that the monthly

and cumulative cases data is consistent because the large spikes of monthly cases correspond to the points with steeper increases in total cases.

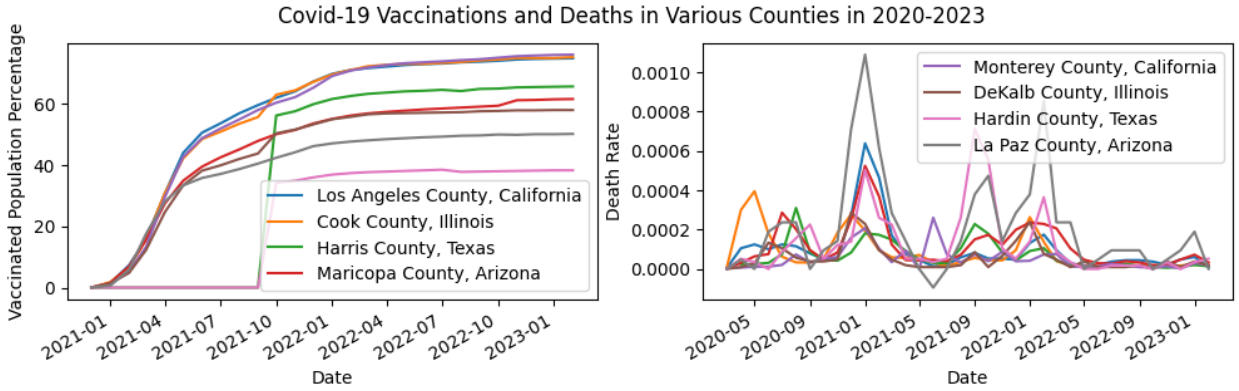


Fig. 7: Covid-19 vaccinations and monthly deaths in various counties. From this, it can be seen that Covid-19 vaccinations have an impact on deaths due to Covid-19. Notice that the two counties with the lowest vaccination percentage have larger spikes in deaths than other counties with higher vaccination rates.

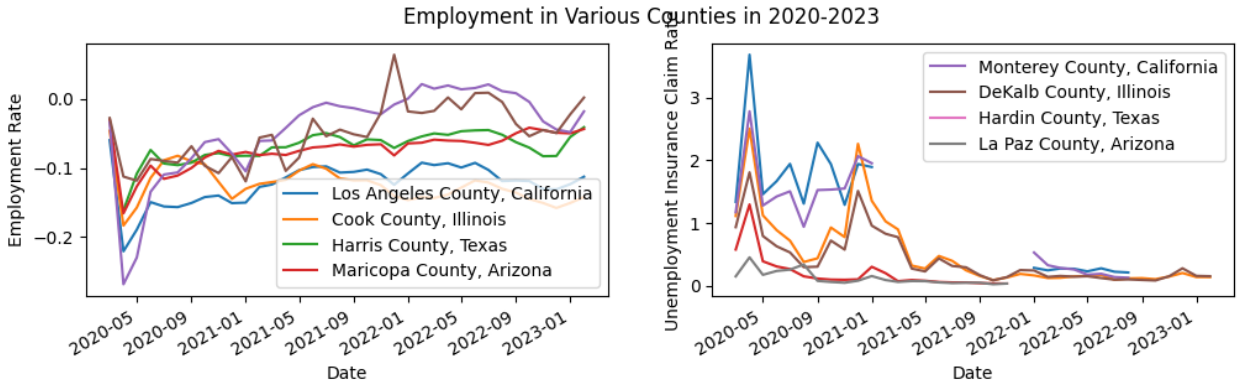


Fig. 8: Employment rate and unemployment insurance claims rate in 2020-2023. The rates are calculated relative to the statistics of January 2020. Large drops during lockdown in March 2020 and slow recovery can be seen.

From these figures, the expected trends in the data can be confirmed. Changes in cases, deaths, vaccinations, and employment align with expected drops and spikes.

4. Models

4.1 Background

The model used was an Ordinary Least Squares (OLS) regression on panel data using entity and time fixed effects. Entity effects control for variables that change across entities but are constant over time, and time fixed effects control for variables that are constant over entities but change over time. The formula for a dataset with n entities and T time periods is:

$$Y_{it} = \alpha_i + \beta X_{it} + \delta_t + u_i + e_{it}$$
$$i = 1,\ldots, n\,;\; t = 1,\ldots, T$$

Here, Y_{it} is the outcome variable for entity i at time t, α_i is the intercept for each entity i, X_{it} is a vector of predictors for entity i at time t, δ_t is the unknown coefficient for the time regressor for time t, u_i is the within-entity error term for entity i, and e_{it} is the overall error term. The interpretation of β states that for a given entity when a predictor changes one unit over time, the outcome will change by β units. This applies to whatever scale the variables are in (Torres-Reyna, 21).

In this model, the entity fixed effects control for differences across counties. In order to find specific relationships, the counties were classified as either rural or urban based on population, as a county was considered rural if it fell in the lower half of county population size or urban if it fell in the upper half of county population size. Further, the stay at home order was included in all input variables of the models since it affects human dynamics significantly, the time period for variables was limited to March 2020 - December 2021, and a heteroskedastic covariance estimator was used because uniform variance cannot be assumed. Finally, all data that was used in the models was standardized as previously mentioned.

4.2 Results

The OLS regression model was used to find relationships between Covid-19 statistics, NTL radiance, and economic variables for all counties, rural counties, and urban counties. The tables and descriptions below summarize the results.

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.1289
Estimator:	PanelOLS	R-squared (Between):	0.5160
No. Observations:	32970	R-squared (Within):	0.2280
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.2742
Time:	17:10:42	Log-likelihood	7.439e+04
Cov. Estimator:	Robust		
		F-statistic:	2159.0
Entities:	1561	P-value	0.0000
Avg Obs:	21.121	Distribution:	F(2,31386)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	488.27
		P-value	0.0000
Time periods:	22	Distribution:	F(2,31386)
Avg Obs:	1498.6		
Min Obs:	703.00		
Max Obs:	1559.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.0029	0.0025	-1.1369	0.2556	-0.0078	0.0021
New_Cases	0.5298	0.0170	31.025	0.0000	0.4955	0.5624
Stay_at_home_order	0.0015	0.0006	2.2943	0.0218	0.0002	0.0027

F-test for Poolability: 4.1188
P-value: 0.0000
Distribution: F(1561,31386)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.2170
Estimator:	PanelOLS	R-squared (Between):	0.8301
No. Observations:	33414	R-squared (Within):	0.2117
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.2695
Time:	17:10:47	Log-likelihood	-6.031e+04
Cov. Estimator:	Robust		
		F-statistic:	4415.8
Entities:	1530	P-value	0.0000
Avg Obs:	21.839	Distribution:	F(2,31861)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	7.8618
		P-value	0.0004
Time periods:	22	Distribution:	F(2,31861)
Avg Obs:	1518.8		
Min Obs:	1409.0		
Max Obs:	1525.0		

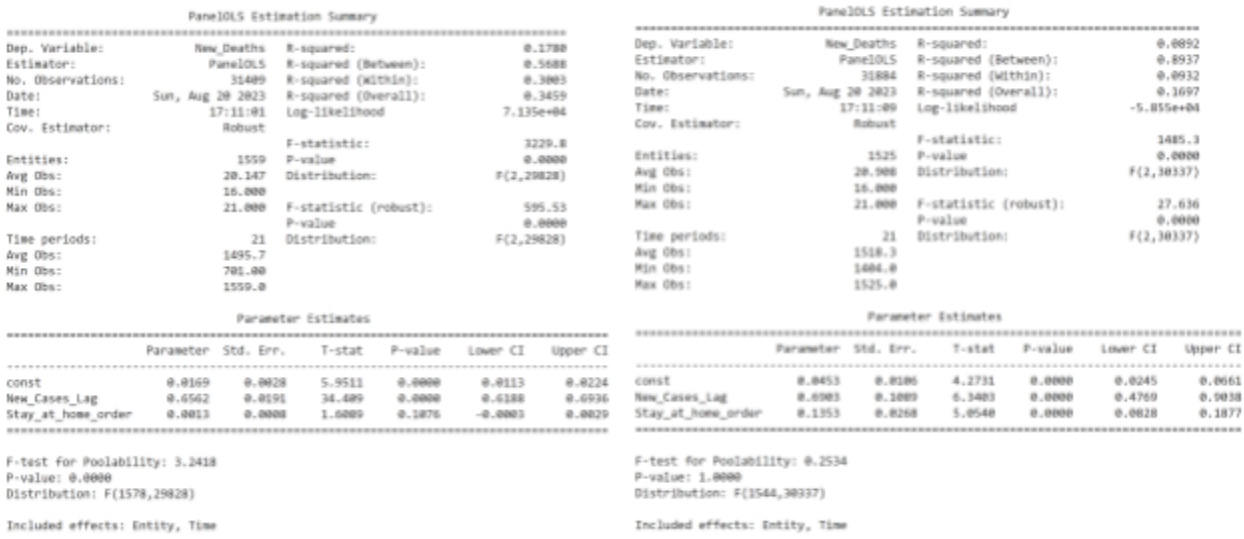
Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0050	0.0396	0.1256	0.9001	-0.0726	0.0825
New_Cases	1.0527	0.4052	2.5978	0.0094	0.2585	1.8470
Stay_at_home_order	-0.0166	0.0570	-0.2913	0.7708	-0.1282	0.0950

F-test for Poolability: 0.5479
P-value: 1.0000
Distribution: F(1550,31861)

Included effects: Entity, Time

Table 1: OLS results using Covid-19 monthly cases to model Covid-19 monthly deaths for rural counties (left) and urban counties (right). The model uses Covid-19 cases and deaths from the same month.



PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.1780
Estimator:	PanelOLS	R-squared (Between):	0.5688
No. Observations:	31489	R-squared (Within):	0.3003
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.3459
Time:	17:11:01	Log-likelihood	7.135e+04
Cov. Estimator:	Robust		
		F-statistic:	3229.8
Entities:	1559	P-value	0.0000
Avg Obs:	20.147	Distribution:	F(2,29828)
Min Obs:	16.000		
Max Obs:	21.000	F-statistic (robust):	595.53
		P-value	0.0000
Time periods:	21	Distribution:	F(2,29828)
Avg Obs:	1495.7		
Min Obs:	701.00		
Max Obs:	1559.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0169	0.0028	5.9511	0.0000	0.0113	0.0224
New_Cases_Lag	0.6562	0.0191	34.409	0.0000	0.6188	0.6936
Stay_at_home_order	0.0013	0.0008	1.6089	0.1076	-0.0003	0.0029

F-test for Poolability: 3.2418
P-value: 0.0000
Distribution: F(1578,29828)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.0892
Estimator:	PanelOLS	R-squared (Between):	0.8937
No. Observations:	31884	R-squared (Within):	0.0932
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.1697
Time:	17:11:09	Log-likelihood	-5.855e+04
Cov. Estimator:	Robust		
		F-statistic:	1485.3
Entities:	1525	P-value	0.0000
Avg Obs:	20.900	Distribution:	F(2,30337)
Min Obs:	16.000		
Max Obs:	21.000	F-statistic (robust):	27.636
		P-value	0.0000
Time periods:	21	Distribution:	F(2,30337)
Avg Obs:	1518.3		
Min Obs:	1484.0		
Max Obs:	1525.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0453	0.0106	4.2731	0.0000	0.0245	0.0661
New_Cases_Lag	0.6903	0.1089	6.3403	0.0000	0.4769	0.9038
Stay_at_home_order	0.1353	0.0268	5.0540	0.0000	0.0828	0.1877

F-test for Poolability: 0.2534
P-value: 1.0000
Distribution: F(1544,30337)

Included effects: Entity, Time

Table 2: OLS results using Covid-19 lagged monthly cases to predict Covid-19 monthly deaths for rural counties (left) and urban counties (right). The lagged monthly cases use the count from the previous month to predict the number of deaths in the next month.

PanelOLS Estimation Summary

Dep. Variable:	New_Cases	R-squared:	0.0334
Estimator:	PanelOLS	R-squared (Between):	0.6460
No. Observations:	66384	R-squared (Within):	0.0383
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.3015
Time:	17:27:00	Log-likelihood	-5.102e+04
Cov. Estimator:	Robust		
		F-statistic:	1892.4
Entities:	3091	P-value	0.0000
Avg Obs:	21.477	Distribution:	F(2,63270)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	13.066
		P-value	0.0000
Time periods:	22	Distribution:	F(2,63270)
Avg Obs:	3017.5		
Min Obs:	2112.0		
Max Obs:	3084.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.0173	0.0028	-6.2395	0.0000	-0.0227	-0.0119
NTL_SOL	0.2747	0.0551	4.9810	0.0000	0.1666	0.3827
Stay_at_home_order	0.0562	0.0214	2.6229	0.0087	0.0142	0.0981

F-test for Poolability: 4.2268
P-value: 0.0000
Distribution: F(3111,63270)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Cases	R-squared:	0.0003
Estimator:	PanelOLS	R-squared (Between):	0.0022
No. Observations:	32970	R-squared (Within):	0.0032
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0032
Time:	17:27:09	Log-likelihood	8.611e+04
Cov. Estimator:	Robust		
		F-statistic:	5.2299
Entities:	1561	P-value	0.0054
Avg Obs:	21.121	Distribution:	F(2,31386)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	8.6303
		P-value	0.0002
Time periods:	22	Distribution:	F(2,31386)
Avg Obs:	1498.6		
Min Obs:	703.00		
Max Obs:	1559.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1445	0.0002	-768.75	0.0000	-0.1449	-0.1442
NTL_SOL	0.0023	0.0006	4.0085	0.0000	0.0012	0.0034
Stay_at_home_order	0.0002	0.0004	0.6177	0.5368	-0.0005	0.0010

F-test for Poolability: 19.425
P-value: 0.0000
Distribution: F(1581,31386)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Cases	R-squared:	0.0313
Estimator:	PanelOLS	R-squared (Between):	0.6449
No. Observations:	33414	R-squared (Within):	0.0325
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.2869
Time:	17:27:14	Log-likelihood	-3.663e+04
Cov. Estimator:	Robust		
		F-statistic:	514.88
Entities:	1530	P-value	0.0000
Avg Obs:	21.839	Distribution:	F(2,31861)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	12.177
		P-value	0.0000
Time periods:	22	Distribution:	F(2,31861)
Avg Obs:	1518.8		
Min Obs:	1409.0		
Max Obs:	1525.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0280	0.0143	1.9488	0.0513	-0.0002	0.0561
NTL_SOL	0.2758	0.0625	4.4107	0.0000	0.1532	0.3983
Stay_at_home_order	0.1593	0.0457	3.4883	0.0005	0.0698	0.2489

F-test for Poolability: 4.5247
P-value: 0.0000
Distribution: F(1550,31861)

Included effects: Entity, Time

Table 3: OLS results using NTL sum of lights to model Covid-19 monthly cases for all counties (top), rural counties (bottom left), and urban counties (bottom right).

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.0012
Estimator:	PanelOLS	R-squared (Between):	0.3257
No. Observations:	66384	R-squared (Within):	0.0018
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0354
Time:	17:42:38	Log-likelihood	-1.052e+05
Cov. Estimator:	Robust		
		F-statistic:	16.757
Entities:	3091	P-value	0.0000
Avg Obs:	21.477	Distribution:	F(2,63270)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	8.0381
		P-value	0.0003
Time periods:	22	Distribution:	F(2,63270)
Avg Obs:	3017.5		
Min Obs:	2112.0		
Max Obs:	3084.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0130	0.0052	2.5032	0.0123	0.0028	0.0232
NTL_SOL	0.1070	0.0417	2.5650	0.0103	0.0252	0.1888
Stay_at_home_order	0.0751	0.0214	3.5078	0.0005	0.0331	0.1170

F-test for Poolability: 0.0760
P-value: 1.0000
Distribution: F(3111,63270)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.0003
Estimator:	PanelOLS	R-squared (Between):	-0.0005
No. Observations:	32970	R-squared (Within):	-0.0064
Date:	Sun, Aug 20 2023	R-squared (Overall):	-0.0054
Time:	17:42:43	Log-likelihood	7.227e+04
Cov. Estimator:	Robust		
		F-statistic:	4.8727
Entities:	1561	P-value	0.0170
Avg Obs:	21.121	Distribution:	F(2,31386)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	7.5379
		P-value	0.0005
Time periods:	22	Distribution:	F(2,31386)
Avg Obs:	1498.6		
Min Obs:	703.00		
Max Obs:	1559.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.0882	0.0002	-326.12	0.0000	-0.0887	-0.0798
NTL_SOL	-0.0021	0.0007	-3.1565	0.0016	-0.0035	-0.0008
Stay_at_home_order	0.0017	0.0007	2.3913	0.0168	0.0003	0.0031

F-test for Poolability: 9.5889
P-value: 0.0000
Distribution: F(1581,31386)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	New_Deaths	R-squared:	0.0010
Estimator:	PanelOLS	R-squared (Between):	0.2886
No. Observations:	33414	R-squared (Within):	0.0018
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0301
Time:	17:42:46	Log-likelihood	-6.438e+04
Cov. Estimator:	Robust		
		F-statistic:	15.567
Entities:	1530	P-value	0.0000
Avg Obs:	21.839	Distribution:	F(2,31861)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	7.5138
		P-value	0.0005
Time periods:	22	Distribution:	F(2,31861)
Avg Obs:	1518.8		
Min Obs:	1489.0		
Max Obs:	1525.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0754	0.0140	5.3902	0.0000	0.0480	0.1029
NTL_SOL	0.0951	0.0465	2.0451	0.0409	0.0040	0.1863
Stay_at_home_order	0.1519	0.0413	3.6807	0.0002	0.0710	0.2328

F-test for Poolability: 0.8806
P-value: 0.9996
Distribution: F(1550,31861)

Included effects: Entity, Time

Table 4: OLS results using NTL sum of lights to model Covid-19 monthly deaths for all counties (top), rural counties (bottom left), and urban counties (bottom right).

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0006
Estimator:	PanelOLS	R-squared (Between):	0.0004
No. Observations:	28865	R-squared (Within):	0.0143
Date:	Tue, Aug 22 2023	R-squared (Overall):	0.0049
Time:	00:29:19	Log-likelihood	-1.456e+04
Cov. Estimator:	Robust		
		F-statistic:	7.9380
Entities:	1309	P-value	0.0004
Avg Obs:	21.815	Distribution:	F(2,28473)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	5.8398
		P-value	0.0029
Time periods:	22	Distribution:	F(2,28473)
Avg Obs:	1357.5		
Min Obs:	1226.0		
Max Obs:	1365.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1301	0.0031	-41.475	0.0000	-0.1362	-0.1239
Incidence_Rate	-0.0102	0.0050	-2.0456	0.0408	-0.0199	-0.0004
Stay_at_home_order	-0.0418	0.0160	-2.6055	0.0092	-0.0732	-0.0103

F-test for Poolability: 69.864
P-value: 0.0000
Distribution: F(1389,28473)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0013
Estimator:	PanelOLS	R-squared (Between):	0.0014
No. Observations:	3855	R-squared (Within):	0.0222
Date:	Tue, Aug 22 2023	R-squared (Overall):	0.0070
Time:	00:35:33	Log-likelihood	-3182.8
Cov. Estimator:	Robust		
		F-statistic:	2.3119
Entities:	179	P-value	0.0992
Avg Obs:	21.536	Distribution:	F(2,3653)
Min Obs:	17.000		
Max Obs:	22.000	F-statistic (robust):	1.6995
		P-value	0.1829
Time periods:	22	Distribution:	F(2,3653)
Avg Obs:	175.23		
Min Obs:	106.00		
Max Obs:	179.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1483	0.0112	-13.239	0.0000	-0.1703	-0.1263
Incidence_Rate	0.0044	0.0142	0.3066	0.7592	-0.0235	0.0322
Stay_at_home_order	-0.0931	0.0510	-1.8274	0.0677	-0.1931	0.0068

F-test for Poolability: 67.174
P-value: 0.0000
Distribution: F(199,3653)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0004
Estimator:	PanelOLS	R-squared (Between):	0.0002
No. Observations:	26010	R-squared (Within):	0.0074
Date:	Tue, Aug 22 2023	R-squared (Overall):	0.0026
Time:	00:33:54	Log-likelihood	-1.058e+04
Cov. Estimator:	Robust		
		F-statistic:	5.2358
Entities:	1190	P-value	0.0053
Avg Obs:	21.857	Distribution:	F(2,24797)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	4.4636
		P-value	0.0115
Time periods:	22	Distribution:	F(2,24797)
Avg Obs:	1182.3		
Min Obs:	1121.0		
Max Obs:	1186.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1285	0.0031	-41.292	0.0000	-0.1346	-0.1224
Incidence_Rate	-0.0123	0.0051	-2.3915	0.0168	-0.0223	-0.0022
Stay_at_home_order	-0.0260	0.0162	-1.6110	0.1072	-0.0577	0.0056

F-test for Poolability: 68.710
P-value: 0.0000
Distribution: F(1190,24797)

Included effects: Entity, Time

Table 5: OLS results using Covid-19 monthly incidence rate to model employment rate for all counties (top), rural counties (bottom left), and urban counties (bottom right).

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0007
Estimator:	PanelOLS	R-squared (Between):	0.0028
No. Observations:	29865	R-squared (Within):	0.0185
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0079
Time:	18:41:26	Log-likelihood	-1.456e+04
Cov. Estimator:	Robust		
		F-statistic:	9.9215
Entities:	1369	P-value	0.0000
Avg Obs:	21.815	Distribution:	F(2,28473)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	12.566
		P-value	0.0000
Time periods:	22	Distribution:	F(2,28473)
Avg Obs:	1357.5		
Min Obs:	1226.0		
Max Obs:	1365.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1259	0.0032	-39.105	0.0000	-0.1323	-0.1196
NTL_SOL	-0.0137	0.0032	-4.2583	0.0000	-0.0200	-0.0074
Stay_at_home_order	-0.0426	0.0160	-2.6647	0.0077	-0.0740	-0.0113

F-test for Poolability: 69.870
P-value: 0.0000
Distribution: F(1389,28473)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0012
Estimator:	PanelOLS	R-squared (Between):	0.0014
No. Observations:	3855	R-squared (Within):	0.0210
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0067
Time:	18:41:32	Log-likelihood	-3182.8
Cov. Estimator:	Robust		
		F-statistic:	2.2745
Entities:	179	P-value	0.1030
Avg Obs:	21.536	Distribution:	F(2,3653)
Min Obs:	17.000		
Max Obs:	22.000	F-statistic (robust):	1.6656
		P-value	0.1892
Time periods:	22	Distribution:	F(2,3653)
Avg Obs:	175.23		
Min Obs:	105.00		
Max Obs:	179.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1488	0.0130	-11.474	0.0000	-0.1743	-0.1234
NTL_SOL	-0.0014	0.0248	-0.0570	0.9545	-0.0501	0.0473
Stay_at_home_order	-0.0928	0.0510	-1.8198	0.0689	-0.1928	0.0072

F-test for Poolability: 67.158
P-value: 0.0000
Distribution: F(199,3653)

Included effects: Entity, Time

PanelOLS Estimation Summary

Dep. Variable:	Employment_Rate	R-squared:	0.0005
Estimator:	PanelOLS	R-squared (Between):	0.0034
No. Observations:	26010	R-squared (Within):	0.0134
Date:	Sun, Aug 20 2023	R-squared (Overall):	0.0069
Time:	18:41:36	Log-likelihood	-1.058e+04
Cov. Estimator:	Robust		
		F-statistic:	6.7516
Entities:	1190	P-value	0.0012
Avg Obs:	21.857	Distribution:	F(2,24797)
Min Obs:	1.0000		
Max Obs:	22.000	F-statistic (robust):	11.001
		P-value	0.0000
Time periods:	22	Distribution:	F(2,24797)
Avg Obs:	1182.3		
Min Obs:	1121.0		
Max Obs:	1186.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1232	0.0033	-37.836	0.0000	-0.1296	-0.1168
NTL_SOL	-0.0135	0.0031	-4.3894	0.0000	-0.0195	-0.0075
Stay_at_home_order	-0.0274	0.0161	-1.7004	0.0891	-0.0590	0.0042

F-test for Poolability: 68.673
P-value: 0.0000
Distribution: F(1210,24797)

Included effects: Entity, Time

Table 6: OLS results using NTL sum of lights to model Covid-19 monthly cases for all counties (top), rural counties (bottom left), and urban counties (bottom right).

5. Discussion

From the OLS regression results, connections can be found within the data and observations can be made about these relationships.

5.1 Covid-19 Monthly Cases and Covid-19 Monthly Deaths

Multiple linear regression was used to test if Covid-19 monthly cases and the stay at home order significantly model Covid-19 monthly deaths in rural and urban counties. The statistical results of this regression can be found in Table 1 and are interpreted below.

For rural counties, the fitted model was New_Deaths = 0.5290 * New_Cases + 0.0015 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.2742, F(2, 31386) = 488.27, p = 0.0000), and it was found that monthly cases (β = 0.5290, p = 0.0000) and the stay at home order (β = 0.0015, p = 0.0218) significantly modeled monthly deaths.

For urban counties, the fitted model was New_Deaths = 1.0527 * New_Cases. The overall regression was statistically significant (R^2 = 0.2695, F(2, 31861) = 7.8618, p = 0.0004), and it was found that monthly cases (β = 1.0527, p = 0.0094) significantly modeled monthly deaths but the stay at home order (β = -0.0166, p = 0.7708) did not.

As is intuitively expected, this model confirms that Covid-19 cases and deaths are significantly correlated.

Multiple linear regression was then used to test if lagged Covid-19 monthly cases and the stay at home order significantly predict Covid-19 monthly deaths in rural and urban counties. The difference between this model and the last is that the monthly cases used as an input are from the previous month rather than in the same month as the new deaths used as an output variable. The statistical results of this regression can be found in Table 2 and are interpreted below.

For rural counties, the fitted model was New_Deaths = 0.0169 + 0.6562 * New_Cases. The overall regression was statistically significant (R^2 = 0.3459, F(2, 29828) = 595.53, p = 0.0000), and it was found that monthly cases (β = 0.6562, p = 0.0000) significantly predicted monthly deaths but the stay at home order (β = 0.0013, p = 0.1076) did not.

For urban counties, the fitted model was New_Deaths = 0.0453 + 0.6903 * New_Cases + 0.1353 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.1697, F(2, 30337) = 27.636, p = 0.0000), and it was found that monthly cases (β = 0.6903, p = 0.0000) and the stay at home order (β = 0.1353, p = 0.0000) significantly predicted monthly deaths.

As is intuitively expected, this model confirms that Covid-19 cases are predictive of Covid-19 deaths.

From these models, it can be seen that Covid-19 monthly cases are correlated to both Covid-19 deaths in the same and subsequent month. Since not all Covid-19 cases that result in death have the date of infection and the date of death in the same month and not all of them have the date of infection and the date of death in separate months, the most accurate model of cases and deaths likely lies between the two above. That is, the most accurate model uses a shorter time period than months to predict deaths from cases.

5.2 NTL Sum of Lights and Covid-19 Monthly Cases

Multiple linear regression was used to test if NTL sum of lights and the stay at home order significantly model Covid-19 monthly cases in rural and urban counties. The statistical results of this regression can be found in Table 3 and are interpreted below.

For all counties, the fitted model was New_Cases = -0.0173 + 0.2747 * NTL_SOL + 0.0562 * Stay_at_home_order. The overall regression was statistically significant ($R^2 = 0.3015$, $F(2, 63270) = 13.066$, $p = 0.0000$), and it was found that NTL sum of lights ($\beta = 0.2747$, $p = 0.0000$) and the stay at home order ($\beta = 0.0562$, $p = 0.0087$) significantly modeled monthly cases.

For rural counties, the fitted model was New_Cases = -0.1445 + 0.0023 * NTL_SOL. The overall regression was statistically significant ($R^2 = 0.0032$, $F(2, 31386) = 8.6303$, $p = 0.0002$), and it was found that NTL sum of lights ($\beta = 0.0023$, $p = 0.0000$) significantly modeled monthly cases but the stay at home order ($\beta = 0.0002$, $p = 0.5368$) did not.

For urban counties, the fitted model was New_Cases = 0.0280 + 0.2758 * NTL_SOL + 0.1593 * Stay_at_home_order. The overall regression was statistically significant ($R^2 = 0.2869$, $F(2, 31861) = 12.177$, $p = 0.0000$), and it was found that NTL sum of lights ($\beta = 0.2758$, $p = 0.0000$) and the stay at home order ($\beta = 0.1593$, $p = 0.0005$) significantly modeled monthly cases.

From these models, positive correlations between NTL and cases and NTL and the stay at home order are observed. As is consistent with past literature, the positive correlation between NTL and Covid-19 cases is due to the connection between NTL and human activity. With more people moving and interacting, there is more NTL being observed. This interaction then provides an opportunity for the transmission of Covid-19. Thus, the positive correlation between NTL and Covid-19 cases is significant and has meaning. Further, a positive correlation between the stay at home order and Covid-19 cases is observed. Although the stay at home order helped reduce the spread of Covid-19, the policy was implemented in times of large spikes of cases. So even though the stay at home order helped slow the spread of Covid-19, it has a positive correlation with new cases because it was only active in periods with high levels of Covid-19 transmission.

5.3 NTL Sum of Lights and Covid-19 Monthly Deaths

Multiple linear regression was used to test if NTL sum of lights and the stay at home order significantly model Covid-19 monthly deaths in rural and urban counties. The statistical results of this regression can be found in Table 4 and are interpreted below.

For all counties, the fitted model was New_Deaths = 0.0130 + 0.1070 * NTL_SOL + 0.0751 * Stay_at_home_order. The overall regression was statistically significant ($R^2 = 0.0354$, $F(2, 63270) = 8.0381$, $p = 0.0003$), and it was found that NTL sum of lights ($\beta = 0.1070$, $p = 0.0103$) and the stay at home order ($\beta = 0.0751$, $p = 0.0005$) significantly modeled monthly deaths.

For rural counties, the fitted model was New_Deaths = -0.0802 + -0.0021 * NTL_SOL + 0.0017 * Stay_at_home_order. The overall regression was statistically significant (R^2 = -0.0054, F(2, 31386) = 7.5379, p = 0.0005), and it was found that NTL sum of lights (β = -0.0021, p = 0.0016) and the stay at home order (β = 0.0017, p = 0.0168) significantly modeled monthly deaths.

For urban counties, the fitted model was New_Deaths = 0.0754 + 0.0951 * NTL_SOL + 0.1519 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.0301, F(2, 31861) = 7.5318, p = 0.0005), and it was found that NTL sum of lights (β = 0.0951, p = 0.0409) and the stay at home order (β = 0.1519, p = 0.0002) significantly modeled monthly deaths.

It is expected that as more people die due to Covid-19, fewer households will be lit up and NTL will decrease. This can be seen in rural areas because NTL and deaths have a negative correlation, but in urban areas the correlation is positive. This is because in rural areas with smaller populations, it is noticeable when fewer households are lit up from residents dying due to Covid-19. In urban areas where the population is larger and there is more constant lighting from street lights, industrial buildings, and other sources, the change in total light from the loss of a household's light is negligible. Individual deaths have little impact on the sum of lights in urban areas, but spikes of Covid-19 deaths and cases occur when there is human activity. So, as human activity increases and NTL increases from this, Covid-19 cases will increase, as examined in Section 5.2. With an increase in Covid-19 cases, there is an increase in Covid-19 deaths. The loss of light from these deaths is not noticeable in urban areas due to the high levels of light already emitted. Overall, the negative correlation between NTL and Covid-19 deaths in rural areas occurs because noticeably fewer households are being lit up as people die, but the correlation between NTL and Covid-19 deaths is positive in urban areas because individual deaths have less of an impact on areas that are heavily lit and spikes in Covid-19 cases and deaths happen when there are high levels of human activity, shown by an increase in NTL.

5.4 Covid-19 Monthly Incidence Rate and Employment Rate

Multiple linear regression was used to test if Covid-19 monthly incidence rate and the stay at home order significantly model employment rate in rural and urban counties. The statistical results of this regression can be found in Table 5 and are interpreted below.

For all counties, the fitted model was Employment_Rate = -0.1301 + -0.0102 * Incidence_rate + -0.0418 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.0049, F(2, 28473) = 5.8398, p = 0.0029), and it was found that incidence rate (β = -0.0102, p = 0.0408) and the stay at home order (β = -0.0418, p = 0.0092) significantly modeled employment rate.

For rural counties, the fitted model was Employment_Rate = -0.1483 + -0.0931 * Stay_at_home_order. The overall regression was somewhat statistically significant (R^2 = 0.0070, F(2, 31386) = 1.6995, p = 0.1892), and it was found that the stay at home order (β = -0.0931, p =

0.0677) significantly modeled employment rate but incidence rate (β = 0.0044, p = 0.7592) did not.

For urban counties, the fitted model was Employment_Rate = -0.1285 + -0.0123 * Incidence_rate. The overall regression was statistically significant (R^2 = 0.0026, F(2, 24797) = 4.4636, p = 0.0115), and it was found that incidence rate (β = -0.0123, p = 0.0168) significantly modeled employment rate but the stay at home order (β = -0.0260, p = 0.1072) did not.

From these models, it can be seen that when Covid-19 incidence increases and people stay at home, employment rate decreases. The stay at home order may have more of an impact on employment rate in rural areas because of the type of work being done, and this may suggest there are fewer opportunities for remorse work in rural areas.

5.5 NTL Sum of Lights and Employment Rate

Multiple linear regression was used to test if NTL sum of lights and the stay at home order significantly model employment rate in rural and urban counties. The statistical results of this regression can be found in Table 6 and are interpreted below.

For all counties, the fitted model was Employment_Rate = -0.1259 + -0.0137 * NTL_SOL + -0.0426 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.0079, F(2, 28473) = 12.566, p = 0.0000), and it was found that NTL sum of lights (β = -0.0137, p = 0.0000) and the stay at home order (β = -0.0426, p = 0.0077) significantly modeled employment rate.

For rural counties, the fitted model was Employment_Rate = -0.1488 + -0.0928 * Stay_at_home_order. The overall regression was somewhat statistically significant (R^2 = 0.0067, F(2, 3653) = 1.6656, p = 0.1892), and it was found that the stay at home order (β = -0.0014, p = 0.9545) significantly modeled employment rate but NTL (β = 0.0044, p = 0.7592) did not.

For urban counties, the fitted model was Employment_Rate = -0.1232 + -0.0135 * NTL_SOL + -0.0274 * Stay_at_home_order. The overall regression was statistically significant (R^2 = 0.0069, F(2, 24797) = 11.001, p = 0.0000), and it was found that NTL (β = -0.0135, p = 0.0000) and the stay at home order (β = -0.0274, p = 0.0891) significantly modeled employment rate.

These models match those relating incidence rate and employment rate. NTL is positively correlated with cases. If cases increase the incidence rate increases, so NTL and incidence rate are positively correlated. Thus, the negative correlation between NTL and employment rate agrees with the negative correlation between incidence rate and employment rate.

6. Conclusion

Overall, there are strong and significant relationships that can be modeled by Covid-19, NTL, and economic variables. These relationships are dependent on the population of the area in question, as rural and urban counties react differently to Covid-19. Specifically, high levels of Covid-19 are associated with large losses of NTL in rural areas, but high levels of Covid-19 are associated with increases of NTL in urban areas. All of the relationships discussed here are dependent on the assumption that NTL is correlated with human activity, as used in previous literature. The models here confirm it, as NTL is positively correlated with Covid-19 cases, which increase when there is more contact between people. Although, the relationship between NTL and Covid-19 deaths vary by county. In rural counties, NTL and deaths have a negative correlation, but in urban counties, NTL and deaths have a positive correlation due to the constant excess of light in urban areas. Finally, it was observed that NTL and Covid-19 incidence rate are significant indicators of employment rate during the pandemic, as both NTL and Covid-19 incidence rate have a negative correlation with employment rate.

7. Acknowledgements

I would like to thank the Caltech SURF program for this opportunity, Prof. Pawel Janas for being my mentor, and John Niccolai from Global Fixed Income at Citadel Investment Group for funding the project.

8. References

Bharti, N., Tatem, A.J., Ferrari, M.J., Grais, R.F., Djibo, A., & Grenfell, B.T. (2011). Explaining Seasonal Fluctuations of Measles in Niger Using Nighttime Lights Imagery. *Science, 334*: 1424-1427. https://doi.org/10.1126/science.1210554

Bruederle A, Hodler R (2018) Nighttime lights as a proxy for human development at the local level. PLOS ONE 13(9): e0202231. https://doi.org/10.1371/journal.pone.0202231

Chetty, R., Friedman, J.N., & Stepner, M. (2023). The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. *Opportunity Insights*. https://www.tracktherecovery.org/

Connell, C. (2022). *The Ultimate State-County-FIPS Tool.* Towards Data Science. Retrieved July 5, 2023 from https://towardsdatascience.com/the-ultimate-state-county-fips-tool-1e4c54dc9dff

COVID-19 State and County Policy Orders. (2023). Centers for Disease Control and Prevention. Retrieved July 5, 2023, from https://healthdata.gov/dataset/COVID-19-State-and-County-Policy-Orders/gyqz-9u7n

COVID-19 Vaccinations in the United States,County. (2023). Centers for Disease Control and Prevention. Retrieved July 3, 2023, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

Dasgupta, N. (2022). Using satellite images of nighttime lights to predict the economic impact of COVID-19 in India. *Advances in Space Research, 70*(4): 863-879. https://doi.org/10.1016/j.asr.2022.05.039

Donaldson, D. & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives, 30*(4), 171–198. https://doi.org/10.1257/jep.30.4.171

Dong, E., Du, H., & Gardener, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet, 20*(5): 533-534. https://doi.org/10.1016/S1473-3099(20)30120-1

Earth Observation Group. (n.d.). *VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1.* Earth Engine Data Catalog. Retrieved July 3, 2023, from https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMSLCFG

Henderson, J.V., Storeygard, A., & Weil, D.N.. 2012. Measuring Economic Growth from Outer Space. *American Economic Review*, *102*(2): 994-1028.

Qiang, Y., Huang, Q., & Xu, J. (2020). Observing community resilience from space: Using nighttime lights to model economic disturbance and recovery pattern in natural disaster. *Sustainable Cities and Society, 57.* https://doi.org/10.1016/j.scs.2020.102115

Rubinyi, S., Goldblatt, R., & Park, H. (2020, May 4). *Nighttime lights are revolutionizing the way we understand COVID-19 and our world*. World Bank Blogs. https://blogs.worldbank.org/sustainablecities/nighttime-lights-are-revolutionizing-way-we-understand-covid-19-and-our-world

Stathakis, D., Liakos, L., & Baltas, P. (2021). COVID-19 Pandemic Assessment by Night-Lights. *IEEE International Geoscience and Remote Sensing Symposium IGARSS*. https://doi.org/10.1109/IGARSS47720.2021.9553441

Torres-Reyna, O. (2007). *Panel Data Analysis Fixed and Random Effects using Stata*. Princeton University. https://www.princeton.edu/~otorres/Panel101.pdf

U.S. Census Bureau. (2018). *TIGER: US Census Counties 2018*. Retrieved July 3, 2023, from https://developers.google.com/earth-engine/datasets/catalog/TIGER_2018_Counties

Wang, Z., Roman, M., Kalb, V., Miller, S., Zhang, J., & Shrestha, R. (2021). Quantifying uncertainties in nighttime light retrievals from Suomi-NPP and NOAA-20 VIIRS Day/Night Band data. *Remote Sensing of Environment, 263*. https://doi.org/10.1016/j.rse.2021.112557

WHO Coronavirus (COVID-19) Dashboard. (n.d.). World Health Organization. Retrieved February 21, 2023, from https://covid19.who.int/

Zhang, Q. & Seto, K. (2011). Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sensing of Environment, 115*(9): 2320-2329. https://doi.org/10.1016/j.rse.2011.04.032

Zhang, Y., Peng, N., Yang, S., & Jia, P. (2022). Associations between nighttime light and COVID-19 incidence and mortality in the United States. *International Journal of Applied Earth Observation and Geoinformation*, *112*. https://doi.org/10.1016/j.jag.2022.102855