A Comprehensive Study of Covid-19 in Florida

Julian Bennett* and Lauren Eriksen†
Project Advisor: Xingjie Helen Li ‡

Abstract

In this paper, we develop a series of modified Susceptible-Infected-Removed (SIR) models to study the COVID-19 pandemic, assess the effectiveness of government protocols, and investigate the rationale behind them in Florida from March to June 2020.

Our research involves the application of multiple regressions to pinpoint and identify Covid-19 lockdown periods, followed by a series of modifications to the SIR model and parameters fitting from the data. This involved creating new model approximations, such as the time-delayed SIR model and the reinfected SIR model, in order to take into account factors such as incubation and reinfection, and get the lowest error discrepancy possible for our infection rate. We were able to conclude that as we included more factors, our error rate became lower and our simulation results became more accurate compared to the actual data. We could also identify outlier metropolitan areas and draw certain conclusions on the level of government performance. We then moved on to find the correlations, if any, between infection rates and external factors. We looked at demographic and weather data to demonstrate whether correlations existed. We found that there are a few factors with high correlations, including graduate education and low temperatures.

Keywords: Susceptible-Infected-Removed model (SIR), time-delayed SIR, reinfected SIR, parameter fitting, error discrepancy, multivariate regression, statistical correlation

1 Introduction

The rise of dangerous pandemic and epidemic viruses has left scientists and researchers with the need to counteract them. Countermeasures could include

*North Carolina State University, jabenne7@ncsu.edu
†Purdue University, leriksen@purdue.edu
‡Department of Mathematics and Statistics, University of North Carolina at Charlotte, xli47@uncc.edu

finding a way to measure them, how quickly they spread within a population, how long symptoms appear, the potential for reinfection, and most importantly, a way to predict the spread of a particular infectious disease. The SIR (susceptible-infected-removed) model was designed to be an epidemiological model for the infection of a disease that measures its particular growth rate [5, 3].

Although the original SIR models are used very frequently to model disease infection rates, we wanted to see how accurate it would be for Covid-19. However, achieving the most accurate prediction model is more complicated than it may seem because these SIR equations do not account for many factors. Although the original SIR model turns out to be too simplistic for states like Florida, it still remains a great foundational model. Throughout our research, we modified the original SIR equations to fit the real data more closely.

Our motivation for the research came from the use of SIR models for the prediction of pandemics. We wanted to see if it was possible to fit the parameters from the data and accurately model infection rates for the disease.

We can even consider a previous study on the spread of Covid-19 in North Carolina [1]. This study was split into different metros that depended on the size of the population and took into account the incubation rate and the removal rate for the Covid-19 victims. Even this modified SIR model for North Carolina does not fit our data for Florida. This then brings up other topics of interest, such as splitting up Florida metropolitan areas, finding the growth rate, and finding the incubation and removal rate of Covid-19. Furthermore, if these changes don't fit our data, we want to study the possible directions of modifying the original equations to get better predictive models of the spread of Covid-19 in Florida.

We chose to study Covid-19 cases in Florida because this state is known for its tourism, education, politics, and large population. These factors lead to large, diverse populations and include visitors from many different places. We wanted to see whether these factors had any effect on Covid-19 infection rates.

In Section 2, we explore the metropolitan areas included in our study and explain how we processed real-world data to prepare for modeling. Next, in Section 3, we introduce the various modified SIR equations, the numerical approximation scheme, and the error assessment methods used. Section 4 presents an additional outcome of our study: the identification of protocol adherence dates and the evaluation of their effectiveness in each region. Section 5 demonstrates the results of our modified SIR models. In Section 6, we focus on the statistical correlation analysis to identify why certain metro areas experienced higher growth rates and took longer to follow protocols. Section 7 discusses the limitations of our current methods and suggests directions for future improvement. Finally, in Section 8, we conclude by summarizing the key outcomes of this project.

2 Data description and processing

In this section, we describe the data and talk about how we process the infected cases in Florida from March to June 2020 before we implement various SIR models to model the growth rate of infection.

2.1 Florida's Metro Areas

In order to study different parts of Florida, we needed to find a way to combine counties because it was too hard to compare 68 different counties. In total, there are 22 metropolitan (metro) areas, but we decided to focus on the 11 major metro areas, based on population and popularity. These 11 metro areas each consisted of populations exceeding 500,000, while also being very attractive for tourism or politics [6]. Because our research focuses on the growth rate of Covid-19, these particular characteristics play a crucial role in our findings. Our 11 metro areas consist of Cape Coral, Deltona, Jacksonville, Lakeland, Miami, North Port, Orlando, Palm Bay, Pensacola, Tallahassee, and Tampa Bay, which are illustrated in Figure 1.

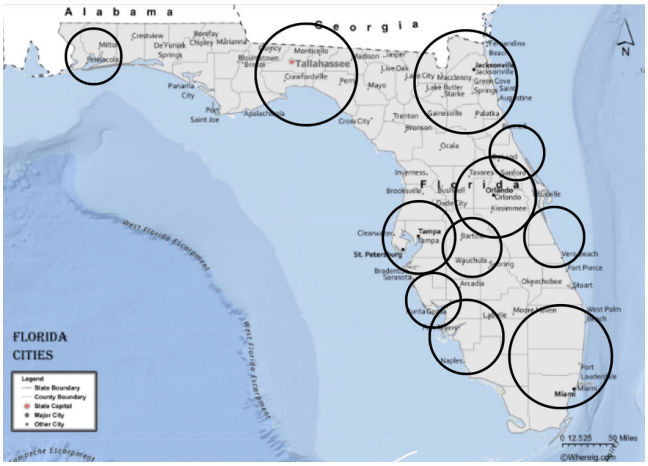


Figure 1: The 11 metro areas of state of Florida.

2.2 Processing of Covid-19 Data in Florida

Covid-19 was a worldwide pandemic that began in March 2020. We used the dataset of Covid-19 cases in the state of Florida from March 2020 to June 2020 from the Johns Hopkins Coronavirus Resource Center [7]. This dataset provides the daily number of cases in each county. When there are dates without reported infection numbers, we estimated and then filled in the missing values using data from nearby days. Specifically, we first take the logarithm of the infection numbers from the nearest dates with available case values, and then

apply linear interpolation to estimate the logarithm of numbers for the missing dates. There were multiple government regulations to keep Covid-19 cases low in order for hospitals to stay afloat [8]. We used these regulations from March 2020 to June 2020 to split up our data into several time periods. Although government protocols suggest a method for defining time intervals, each metro region did not necessarily adhere to these guidelines. To determine the actual time periods for each metro region, we adjusted the intervals by maximizing the R^2 scores of linear regressions on the logarithm of the observed data over the corresponding periods. Once the optimal time intervals are identified, we then fine-tuned the parameters of various modified SIR models.

For example, in March, there is an increase in Covid-19 cases because very few Covid-19 regulations had been enforced at that point. However, multiple regulations were enforced in April, so there is a decrease in the rate of infection. Consequently, there is a need to split the actual data into several time intervals before we fit parameters of SIR models. These were the types of trends we were looking for in order to apply them to our model. Specifically speaking, when we studied each metro, we adjusted the dates of each time interval to maximize the R^2 score of linear regression of the raw data. Once we found these dates over the entire time window, these became our lockdown periods for splitting and analyzing the data for each metro.

Meanwhile, these dates become more important when we compared different metros because we were able to find when each metro actually followed protocols. This enabled us to identify the metros that were quick to follow these laws and those that took a little longer. We then further investigated the reasons why particular metros were taking a longer time to follow the protocols. In particular, we studied the factors that might have played a role in this, such as gender, education, location, and weather. These were the types of questions we pursued while conducting an in-depth analysis of the results that came from our research.

Remark 1. In summary, we follow these steps to process the Covid-19 data in Florida and model the infection growth rate.

1. For each metro area, we use linear regression of the logarithm of the actual data to determine the splitting of time periods. The government protocol dates provide us with an initial guess of where to split the intervals, and then we adjust the interval by maximizing the R^2 scores over all periods.

2. Then, within each time period, we adjust the parameters of various SIR models to visually match the slopes of the real data.

   We quantify the intrinsic modeling error across different SIR models by calculating the error term (6).

3. Meanwhile, we further investigated the reasons why particular metros took much longer to follow the protocols. In particular, we studied the factors that might have played a role in this, such as gender, education, location, and weather.

3 Methods

In this section, we first review the mathematical expression of the original SIR model, and propose various modifications on SIR to account for some unique features of Florida.

3.1 Review of the original SIR Model

When modeling and predicting the evolution of a pandemic, the simplest model to use is a susceptible-infected-removed (SIR) model. It helps us find the growth rate of any particular infection, in our case Covid-19 [5, 3]. This model aims to predict and estimate how people change from susceptible to infected to removed over time. Mathematically, the SIR model is set up through these important variables:

- t: number of passing days since first case of Covid-19;
- $S(t)$: number of susceptible people at time t;
- $I(t)$: number of infected people at time t;
- $R(t)$: number of removed people at time t;
- $\beta(t)$: average infection rate at time t;
- $\gamma(t)$: average removal rate at time t.

We can assume constant parameters $\beta(t) \equiv \beta$ and $\gamma(t) \equiv \gamma$ to simplify the modeling in the first place. These variables can be put together into three important equations that will approximate the rate of change for each population in the original SIR model.

$$\frac{dS(t)}{dt} = -\beta I(t)S(t), \tag{1a}$$

$$\frac{dI(t)}{dt} = \beta I(t)S(t) - \gamma I(t), \tag{1b}$$

$$\frac{dR(t)}{dt} = \gamma I(t). \tag{1c}$$

We are also given the initial points for each equation: $S(0) = S_0$, $I(0) = I_0$, and $R(0) = R_0$. We aim to find the rate of change for the susceptible, infected, and removed populations from the data. Note that equation (1a) is always negative. This is because everyone starts in the "susceptible population" and as people contract the infection, they will move to the "infected population". Equation (1b) is the most important equation, though, because we want to fit the parameters from our data [4] and predict the growth rate of the Covid-19, which is $\frac{dI(t)}{dt}$.

Looking at equation (1b) more closely, we can approximate this around $t = 0$ (when the pandemic began). This will give us $\frac{dI(t)}{dt} \approx I(t)[\beta S_0 - \gamma]$. This

equation has a solution of $I(t) = I(0)e^{kt}$ where $k = \beta S(0) - \gamma$. This equation will be very important throughout our project because we can log-transform it to get $\log I(t) = kt + \log I_0$ and thus find the growth rate k through linear regression over each lockdown period.

On the other hand, if we sum all three equations, we find that the right side adds up to zero, whereas the left side adds up to $\frac{d}{dt}(S(t) + I(t) + R(t))$. This means that $S(t)+I(t)+R(t)$ must equal a constant because the derivative of any constant equals zero. This tells us that the original SIR model always assumes a constant population, which is not the case in reality. In addition, the original SIR model does not take into account the incubation period of Covid-19, as each equation in (1) describes the same point in time, t. In later sections, we proposed modified SIR models to better model the reality of Covid-19 in Florida metros.

3.2 Adjustments to the original SIR Model

In this section, we considered several approaches to modify the original SIR equations (1), taking into account the specific characteristics of Covid-19 and the relative levels of tourism in Florida over time.

Time-delayed SIR Model. The first thing that needed to be done was adding a time-delay factor due to the incubation period of Covid-19. This means that there is a certain amount of time between contracting the infection and becoming infectious. In order to show this, we introduced τ_1 into the original SIR model. τ_1 gives us the average number of days for the incubation period. In addition, there is also a period of time for Covid-19 to be completely removed from the infected population, so we also introduced τ_2 to take this into account. It gives us the average number of days for removal. With these variables, we created a new set of equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta I(t-\tau_1)S(t), \\ \frac{dI(t)}{dt} &= \beta I(t-\tau_1)S(t) - \gamma I(t-\tau_2), \\ \frac{dR(t)}{dt} &= \gamma I(t-\tau_2). \end{aligned} \tag{2}$$

We will use these equations as a basis to fit the parameters β and γ by comparing the growth rate k from the data.

Since k is the slope of the original data, we can adjust the parameters β and γ to match the slope of k for each period. We decided to split the entire dataset into 5 lockdown periods because this is where the data naturally separated and followed certain protocols that were enforced. It was observed from (1b) that the approximation around $(t = 0)$, which corresponds to the beginning of each period, yields $\frac{dI(t)}{dt} \approx I(t)[\beta S_0 - \gamma]$. The solution to this approximated equation is $I(t) = I(0)e^{kt}$, where $k = \beta S(0) - \gamma$. Therefore, adjusting γ is mathematically

equivalent to adjusting β. However, for numerical simulations, it is more robust to focus on tuning γ. So, we tuned the parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, and γ_5 for each period based on a visual comparison between the predicted solutions of the SIR model (2) and the actual data on the logarithmic scale. As a result, we will have a new β which is the infection rate for each period. Although we tuned the parameters using a visual comparison, in Section 3, we introduced an error quantity (6) to measure the discrepancy of the slope k between the approximated and actual data.

Repeated Infection SIR Model. Another factor to consider is the fact that people can get Covid-19 multiple times. This means that after people enter the recovered population, they can later be re-entered into the susceptible population. This can be described by a new parameter μ, which represents the fraction of people that survived their earlier Covid-19 infection and have become susceptible again. Hence, the new set of equations read:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta I(t-\tau_1)S(t) + \mu R(t), \\ \frac{dI(t)}{dt} &= \beta I(t-\tau_1)S(t) - \gamma I(t-\tau_2), \\ \frac{dR(t)}{dt} &= \gamma I(t-\tau_2) - \mu R(t) \end{aligned} \tag{3}$$

This version of the model is a more realistic representation of Covid-19 because many people ended up getting Covid-19 multiple times.

New SIR Model with Incoming Tourism Flow. Another way to improve the original SIR model (1), is by accounting for the people that visited Florida for spring break, family, or work. Since Covid-19 started in March and many college students went to Florida for spring break, the state had many visitors that could have led to a significant increase in the spread of Covid-19. To take this factor into account, we introduce the parameter ϵ to account for the increase or decrease in the total population. This can be shown in this new set of equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta I(t-\tau_1)S(t) + \epsilon V(t), \\ \frac{dI(t)}{dt} &= \beta I(t-\tau_1)S(t) - \gamma I(t-\tau_2), \\ \frac{dR(t)}{dt} &= \gamma I(t-\tau_2), \end{aligned} \tag{4}$$

$V(t)$ denotes the average incoming tourism population during the underlying period, which can be treated as an external source term.

Note that the only equation that changed was the first one, as the new people traveling into Florida are now considered part of the susceptible population. The obstacle with this addition was finding the tourism data to account for people

flying into the state. There was no comprehensive data available for each metro area currently under consideration. In addition, we were unsure how to account for people who drove into the state. Therefore, we will study and evaluate the performance of (4) in our future work.

3.3 Numerical Approximations and Error Assessments

We used the Euler method to numerically approximate the solutions of our proposed models. Because our time factor is counted on a daily basis, the time step size is fixed to 1. For instance, the numerical approximation of (2) becomes:

$$\begin{aligned}\tilde{S}(t+1) &= \tilde{S}(t) - \beta\tilde{I}(t-\tau_1)\tilde{S}(t),\\ \tilde{I}(t+1) &= \tilde{I}(t) + \beta\tilde{I}(t-\tau_1)\tilde{S}(t) - \gamma\tilde{I}(t-\tau_2),\\ \tilde{R}(t+1) &= \tilde{R}(t) + \gamma\tilde{I}(t-\tau_2).\end{aligned} \tag{5}$$

With these equations, we can input them into MATLAB and use the MATLAB function dde23 to find values for $\tilde{S}(t)$, $\tilde{I}(t)$, and $\tilde{R}(t)$ at each discrete time grid t. The most important quantity of this output is the values of $\tilde{I}(t)$ because they represent the infected population. We can use these values to fit the growth rate over period i of the SIR model by $\log \tilde{I}(t) = \tilde{k}_i t + \text{intercept}$. With this equation, we can find the slope $\tilde{k}_i$ of each period $i = 1, \ldots, 5$. This whole procedure was applied to the original SIR (1), time-delayed SIR (2), as well as the repeated infection SIR (3).

To mathematically compare the performance of various SIR models, we defined $|\tilde{k}_i - k_i|$ to calculate the discrepancy between $\tilde{k}_i$ obtained from the SIR model and k_i obtained from the data. In other words, it would show mathematically how similar our slopes were to each other.

However, because each lockdown period had a different length of time, we needed to take into account how long each period was when we computed $|\tilde{k}_i - k_i|$ for $i = 1, \ldots, 5$. We therefore proposed a weighted error formula to compare the growth rate of infection from the modeled SIR versus the actual dataset. The new discrepancy formula that we came up with is:

$$\frac{\sum_{i=1}^{5} |\tilde{k_i} - k_i| \times \text{length of } i\text{th period}}{\text{total length of all periods}}. \tag{6}$$

This definition of error quantity improved our comparison between the updated and original SIR models. With this formula, we were able to model the evolution of the pandemic within each metro over the five lockdown periods.

4 Study of Protocol Effectiveness

In addition to studying the Covid-19 infection growth rate in Florida using various SIR models, we also analyzed the effectiveness of protocols across different metro areas using linear regression. This was necessary because, although the

government protocol release dates are available [8], the actual dates when the protocols were followed varied significantly across different metro regions. The results of our project can also help identify the actual dates when the protocols were followed in each metro area.

Linear regression is often shown in the form $y = \beta_0 + \beta_1 x$, where β_1 represents the slope and β_0 represents the y-intercept. Connecting this concept with our project, we used the data that we found for the number of infected people over time [4], and plotted it in log scales. Recall that we divided the data over time where the initial division was based on the protocol release date. We then adjusted the splitting dates per metro by checking the quality of the regression lines and R^2 scores. Once our regression lines are plotted and adjusted, we were able to pinpoint the actual date when each metro area started following protocol, which can be used as an assessment to study the effectiveness of protocols.

Results on actual protocol followed dates. For each metro area, we were able to create graphs using linear regression. We adjusted the five periods to represent our data in each graph, and found the best regression for each of these periods. One thing we wanted to find was when each metro started following protocols. This was found by adjusting the temporal length of each period and optimizing the regression quality over that given period. We considered it to be a quality regression when our error numbers were below 3%. We summarized all our findings in Table 1 so it was easier to compare each metro within Florida.

Metro	First Case	Protocol Followed
Miami	03/07	04/05
Orlando	03/12	04/05
Lakeland	03/17	04/07
Jacksonville	03/11	04/08
Tampa Bay	03/02	04/08
Tallahassee	03/19	04/19
North Port	03/08	04/23
Deltona	03/08	04/25
Cape Coral	03/07	04/26
Pensacola	03/05	04/26
Palm Bay	03/17	04/29

Table 1: Study of protocol effectiveness in Florida. The actual protocol announcement date is March 29, 2020 [8].

Looking at this table, we can easily see the divide in metros that followed protocols early and those that took a lot longer. As shown in Table 1, the first five metros obeyed the protocols much faster than the other six. In addition, these first five metros have some of the highest tourism in Florida. We would like to point out that according to Florida's tourism statistics from [10], there was still a considerable amount of incoming tourism in March and April 2020. One would think that with more people, more cases would appear, but it actually

turns out to be the other way according to the data. This could be due to the fact that these metros with a lot of tourism tend to follow protocols more strictly.

5 Results on various SIR models

In this section, we will compare the performance of various SIR models given the actual infection cases data of Florida from March to June 2020.

5.1 Results for the Pensacola metro area

We first use the Pensacola metro area as a demonstration example.

After identifying the optimal time period divisions, we apply various SIR models to infer the model parameters. It is important to note that the choice of SIR model significantly impacts the quality of the fit. In Figure 2, we present graphical examples for the Pensacola metro area, comparing the original SIR fit with an improved SIR fit. These examples illustrate the importance of first determining the time periods by maximizing the R^2 values in linear regression over each time interval, and subsequently modifying the SIR model for better accuracy.

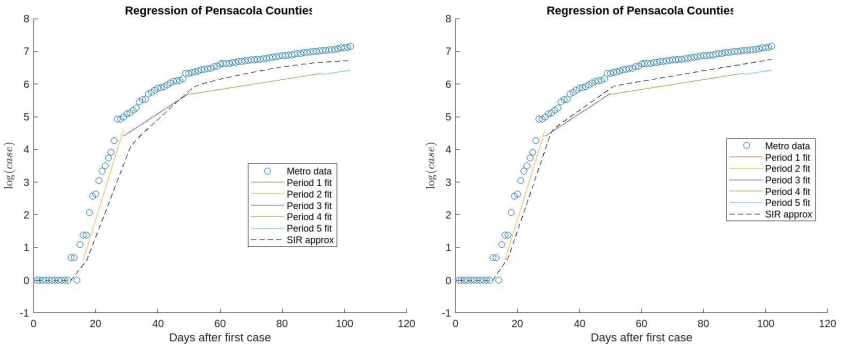
The graph on the left represents the results from the original SIR (1) fitting. The fitted line of original SIR and the linear regressed dashed line of raw data over each period have a relative error value of 1.71%. The graph on the right represents the results from the improved time-delay SIR model (2). Note that the slopes of the SIR fitted results and the linear regressed dashed line are very similar to the slope values of the actual data scatter circles, with a relative error value of 0.68%. Hence, Figure 2 explains why we initially determine time periods by maximizing the R^2 scores over each period, and then improve the original SIR by modifying it to better match the slope values (infection growth rates) of the actual datasets.

5.2 Results for the Jacksonville metro area

Next, we use the Jacksonville metro as another demonstration to show that the time-delayed SIR can better capture the transition of growth rates over different time periods compared to the original SIR model. More precisely speaking, we compare the simulation results for the Jacksonville region via the original SIR (1) with those via the time-delayed SIR (2) in Figure 3. Notice that the original SIR model does not take the incubation time into account, so it is less capable of capturing the transition behaviors of different Covid-19 infection periods.

5.3 Comparison of Various SIR models

In this subsection, we compare the performance of our time-delayed and repeated infection SIR models. We initially adjusted the parameters and numerically computed the solutions to try to match our original data. We then



(a) Poor parameter fitting (b) Optimal parameter fitting

Figure 2: Comparison of the original SIR (1) fit with a reinfection SIR fit (2) for the Pensacola Counties. In both figures, the scatter circles denote actual data points; the black dashed lines denote the results from various SIR approximations; and the colored solid line segments denote the linear regression of actual data over each time period.

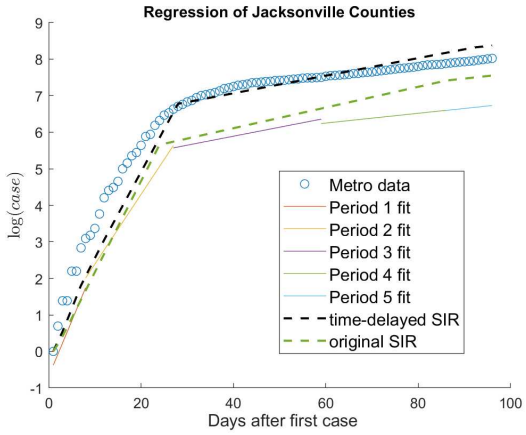


Figure 3: Comparison of the simulation results via the original SIR (1) (dashed-green) with those via the time-delayed SIR (2) (dashed-black) in the Jacksonville metro. The scatter circles denote actual data points.

calculated the error discrepancies of the growth rates using our discrepancy formula (6) in each metro. We summarized the error discrepancies in Table 2 to compare the performance among various models.

Metro	Only Delayed	Reinfected
Cape Coral	0.95%	0.66%
Deltona	0.85%	0.36%
Jacksonville	1.87%	1.53%
Lakeland	0.67%	0.64%
Miami	1.50%	0.48%
North Port	1.83%	1.61%
Orlando	1.86%	1.80%
Palm Bay	0.90%	0.81%
Pensacola	0.98%	0.91%
Tallahassee	0.40%	0.27%
Tampa	1.95%	1.66%

Table 2: Error of growth rate of various modified SIR models. The error is defined in (6) and the reference growth rates are obtained from the raw data.

All of our errors were below 2%, which means that our model was quite accurate for the growth rate of infection, when compared to real world data. One thing to point out, though, is that once we introduced this reinfected factor, the error rates for the repeated infection SIR model (3) decreased significantly for all metros. This means that the modified SIR became more accurate to model the growth rate of Covid-19 infection when we included a time delay and the possibility of reinfection.

6 Statistical Correlation Study

In order to find correlations between outside factors and Covid-19 spread, we first have to introduce multivariate regression. This is just a slightly more complicated version of linear regression. Instead of having one dependent and one independent variable, multivariate regression contains one dependent and multiple independent variables. Each of these independent variables must also be independent of each other in order for the model to meet the underlying assumptions of the analysis. When using multivariate regression, the main things to look for are the p-value and the R^2 values. The p-value tells you whether the model results are statistically significant or not, and the R^2 value tells you about the level of correlation. When observing a p-value, we use 0.05 as our significance threshold. If the number is above 0.05, we consider the data to be not significant and therefore the R^2 value is irrelevant. If the p-value is below 0.05, we consider our data to be significant and we can now look at the R^2 value. The R^2 value is a number between 0 and 1. The closer the number is to 1, the more correlated the model's results are.

We wanted to set up two different multivariate regression equations. The first would include data from different age ranges, genders, income, and education. The second would include data from high and low temperatures and various types of weather. All factors were tested for correlations with COVID-19 infection rates.

6.1 Study of Demographic Data

The first of the external categories of factors we researched was demographic data. We chose this as a correlation factor because things like age, gender, and education can play a significant role in the effects of certain situations. Note that demographic data is accessible to the public, so it was easy to find. In this study, we modeled a multivariate regression of four independent variables: age, gender, income, and education [2]. Each of these categorical variables was divided into multiple subcategories, as shown in the table below. All these variables depended on our chosen y: the weighted average of Covid-19 growth rate k_i. We chose this as our y because the demographic data is provided by the year, not by the day. We then normalized all the factors to a single measurement to fit them into one model. We opted to represent the factors as percentages, as many were already presented this way in the data.

Demographic	P-value
0-20 Age	0.57568
20-50 Age	0.98726
50-70 Age	0.91047
70-80+ Age	0.74878
Male	0.31745
Female	0.22341
0-50k Income	0.37335
50-100k Income	0.36109
100-150k Income	0.37373
150-200k+ Income	0.33624
High School Grad	0.22308
Some College	0.096412
College Grad	0.26149
Post-Grad	0.042838

Demographic	R^2
Age	0.165
Gender	0.162
Income	0.516
Education	0.833

Table 3: Multivariate regression study of Covid-19 growth rate and demographic data.

From Table 3, the most observable trend is that the majority of the p-values seem to have no correlation to our Florida Covid-19 data, since almost all p-values are exceptionally above 0.05. This result is not surprising when considering some of these factors. However, considering the factor of income, this Florida study provides valuable information. In the previous study of Covid-19 in North Carolina [1], there was a correlation with education and income,

however, for Florida we only have a marginal correlation with education. Therefore, achieving this result offers new conclusions that are neither repetitive nor redundant compared to previous studies.

The one factor that is significantly correlated with Covid-19 cases in Florida is post-graduates. The 'Some College' factor is highlighted as well just for the sake of it being a significantly lower p-value than the other factors, even though it is not below 0.05. Looking at the post-graduate demographics, we can interpret the statistical meaning behind its correlation, but we need to look more closely at the R^2 scores before any conclusions are made.

When looking at R^2 values, we can only take the numbers into account whether the p-values are below 0.05, otherwise they are not significant. Since education had the only relevant p-value, we can see that its R^2 is relatively close to 1, showing that post-grad is an important predictor variable for the multivariate regression modeling the Covid-19 spread.

Remark 2. Note that the variables listed in Table 3 contribute to the intercept in this multivariate regression model. For example, the intercept might represent a male aged from 0 to 20 with an income of $0-50$k and a high school diploma. This causes a limitation in the modeling approach used in this study, as the predictor variables of multivariate regression are supposed to be independent.

6.2 Study of Weather Data

Another one of the external factors we wanted to study was the different types of weather in each metro area. For each day, we found the highest temperature, the lowest temperature, and the type of weather (sunny, rainy, cloudy, foggy, snowy) in that specific metro region [9]. These represent our independent variables for the multivariate regression. Since these predictor variables are all dependent on each other, we had to consider them separately. The first one we looked at was the type of weather. In order to handle this data, we had to use dummy variables. In the simulation, we use MATLAB commands: we first convert data to categorical types via command 'categorical'; then we convert categorical data to dummy variables via command 'dummyvar'; lastly, we perform multivariate regression via 'mvregres'.

We collected the results from each metro and put them in Table 4.

From this table, we noted that the only p-values that were less than 0.05 were from the Deltona metro with period 1 and the Orlando metro with period 4. After looking at the corresponding R^2 values, we found that Deltona had a high value, but Orlando had a much lower value. This means that the results of our correlation study are more interpretable for the Deltona metro, but less so for the Orlando metro.

After looking at our results for different types of weather, we wanted to expand and look at the highest and lowest temperature for each day. We wanted to use the multivariate expression for these variables and compare it to the data above to see if any new results appeared. For the highest values, we categorized temperatures above 80 as H (high), above 60 as M (medium), and anything

Metro	P1	P2	P3	P4	P5
Cape Coral	0.661	0.892	0.567	0.732	0.727
Deltona	0.00112	0.678	0.914	0.636	0.524
Jacksonville	0.94	0.89	0.471	0.419	0.812
Lakeland	NA	0.763	0.58	0.401	0.0919
Miami	0.111	0.712	0.321	0.833	0.712
North Port	0.522	0.89	0.286	0.539	0.58
Orlando	0.468	0.966	0.669	0.0338	0.208
Palm Bay	NA	NA	0.362	0.683	0.371
Pensacola	0.363	0.974	0.551	0.526	0.393
Tallahassee	NA	0.165	0.392	0.112	0.593
Tampa Bay	0.11	0.772	0.529	0.347	0.573

Table 4: The p-values of multivariate regression of Covid-19 growth rates over each lockdown period verse weather types for each metro.

below that as L (low). We then found the p-value for each metro and each of their periods in Table 5.

Metro	P1	P2	P3	P4	P5
Cape Coral	0.348	NA	NA	0.773	NA
Deltona	0.12	0.2	0.728	0.363	0.711
Jacksonville	0.422	0.0526	0.402	0.938	NA
Lakeland	NA	0.158	0.00389	0.153	NA
Miami	0.111	NA	0.0165	0.937	NA
North Port	0.315	0.00471	0.997	0.864	NA
Orlando	NA	0.72	0.0792	0.0248	0.505
Palm Bay	NA	NA	0.925	0.752	0.371
Pensacola	1	0.341	0.28	0.733	NA
Tallahassee	NA	0.338	0.218	NA	0.781
Tampa Bay	0.309	0.0924	0.836	0.972	NA

Table 5: The p-values of multivariate regression of Covid-19 growth rates over each lockdown period verse daily highest temperatures for each metro.

From Table 5, the most significant observation comes from Lakeland period 3, Miami period 3, North Port period 2, and Orlando period 4. All of these data points have a p-value of less than 0.05. From above, we found that Orlando with period 4 was also significant in the types of weather data. This means that we can look into this data and investigate causal factors of the highest temperatures for further studies.

When we looked at the lowest values, we categorized temperatures above 60 as H (high), above 50 as M (medium), and equal or below 50 as L (low). We changed the cutoffs because it is a different data set and our cutoffs need to reflect the numbers we have. Thus, the p-values are shown in Table 6.

Metro	P1	P2	P3	P4	P5
Cape Coral	0.786	NA	0.0673	0.228	NA
Deltona	0.332	0.0281	0.65	0.846	NA
Jacksonville	NA	0.0447	0.115	0.203	NA
Lakeland	NA	0.00056	0.307	0.174	NA
Miami	NA	NA	NA	0.927	NA
North Port	0.522	0.00019	NA	0.272	NA
Orlando	NA	NA	0.00000319	0.0104	NA
Palm Bay	NA	NA	NA	NA	NA
Pensacola	1	0.000000428	0.564	0.842	NA
Tallahassee	NA	0.329	0.758	0.397	NA
Tampa Bay	0.606	0.0000885	NA	0.094	NA

Table 6: The p-values of multivariate regression of Covid-19 growth rates over each lockdown period verse daily lowest temperatures for each metro.

We can see that this table for lowest temperatures looks a little different than the previous study based on highest temperatures. Here, we see a lot of zeros and quite a few extremely small values. The most important ones to look at are from Deltona period 2, Jacksonville period 2, Lakeland period 2, North Port period 2, Orlando period 3 and 4, Pensacola period 2, and Tampa Bay period 2.

Evidently, the high and low temperatures have a better statistical correlation with Covid-19 cases than the type of weather. Most of our data from the types of weather was inconclusive, but a decent amount of data from the high and low temperatures showed some correlation.

7 Discussion

In this section, we discuss the future directions of our various modified SIR models and the statistical correlation study.

Through Figure 2, Figure 3 and Table 2, it is clear that the time-delayed refection SIR model can better model the growth rate of infection within Florida. However, several areas for improvement remain: First, we did not account for the number of death cases, which makes the current reinfection model incomplete. Second, we manually adjusted the parameters to visually match the simulation slope to that of the actual data, but we plan to use deep learning techniques to infer the parameters directly from the data.

Regarding the statistical correlation study, our first limitation is the use of MATLAB's black-box command as our multivariate regression tool. As a result, we are uncertain whether the extremely low p-value and 'NA' outcomes in Table 6 are truly due to strong evidence against the null hypothesis or if they stem from numerical issues. Additionally, we assumed that the categorical variables for weather are uncorrelated in our multivariate regression, which may

be an invalid assumption. We plan to use partial least squares regression to account for potential correlations. Alternatively, we could perform a Principal Component Analysis (PCA) first and use the uncorrelated PCA scores as input variables for the regression.

8 Conclusion

Throughout our research, we have observed, compared, and analyzed 11 different metros in Florida. We initially found that about half of the metros followed the protocols early, while the other half took about a month longer. Meanwhile, we studied the original and modified SIR models to model the spread of Covid-19 in Florida. After we adjusted our SIR model, we found that when we included the aspect of reinfection, our reinfected time-delay SIR model became much more accurate for predicting the growth rate of infection. Lastly, we decided to find whether any outside factors were correlated with the Covid-19 infection rates in Florida. Through our research, we found that education and low temperatures had the most correlation.

Looking into the future, we would like to expand the current study to several directions. Since we only focused on the first six months of 2020, we want to now look at the next six months to see if our model is still accurate in terms of prediction of the growth rate of infection. In addition, we would like to look more deeply into certain correlations and what they really mean. We want to know why daily low temperatures and high education had a correlation with infection rates and what this could mean for future pandemics.

Acknowledgements

This work was supported in part by the NSF REU grant DMS-2150179.

References

[1] C. Alzate and X. H. Li. Measuring the effectiveness of covid-19 preventative methods in north carolina with a time-delayed sir model, 2021.

[2] U. C. Bureau. Deltona, fl. *Census Reporter*, 2022.

[3] I. Cooper, A. Mondal, and C. G. Antonopoulos. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, 2020.

[4] E. Dong, J. Ratcliff, T. D. Goyea, A. Katz, R. Lau, T. K. Ng, B. Garcia, E. Bolt, S. Prata, D. Zhang, et al. The johns hopkins university center for systems science and engineering covid-19 dashboard: data collection process, challenges faced, and lessons learned. *The lancet infectious diseases*, 22(12):e370–e376, 2022.

[5] D. Smith, L. Moore, et al. The sir model for spread of disease-the differential equation model. *Convergence*, 2004.

[6] M. to Florida Blog. What are florida's 10 largest metropolitan areas?, 2024.

[7] The confirmed covid-19 cases in florida. access in June 2022.

[8] Florida department of health covid-19 announcements. access in June 2022.

[9] Past weather in orlando, florida, usa, 2020.

[10] Visit florida: Covid-19 fl tourism impacts. access in June 2022.