The development of predictive model of COVID-19 mortality and its relationship with environmental pollutants and outdoor mobility

Danielle Heymann
Advisor: Dr. GuanNan Wang
Department of Computer Science
The College of William and Mary
November 10, 2020

ABSTRACT

The novel coronavirus, COVID-19 has rapidly altered daily lives around the globe. Many questions emerged as the disease spread at alarming rates with a wide scale of severity. COVID-19 had its first U.S. case on January 21, 2020 [6]. California was one of the first states to have a significant surge in cases, and therefore it was used to conduct this research. The specific counties used in this research include the following: Santa Barbara, Sacramento, and Ventura. Each county was treated independently throughout the research process to avoid policy influences on a more generalized model. Therefore, this is a case study on 3 counties, each with a unique model. The objective of this research was to model the daily death count during the COVID-19 pandemic through utilizing data from pollutant indicators (Air quality index, CO, SO2, NO2, PM10, PM2.5, Ozone) and outdoor mobility data. The timeframe considered was from February through May 2020. The exploratory data analysis and preliminary modeling process revealed a relationship between the increase of pollutant indicators and the increase of infection counts. General Linear Models (GLM) and Time Series following Generalized Linear Models (TSGLM) [8] were implemented. The models developed for each county with response of daily death count did reveal some relationships. The Ventura County model revealed a positive relationship between predictor variables, ozone and daily infection count, with respect to response variable, daily death count. The Sacramento County model revealed a negative relationship between predictor variable outdoor mobility to parks and death count. The Santa Barbara County model did not reveal any significant regression coefficients, and therefore, no relationship can be inferred. Additionally, for two of the three counties studied, the model suggests that death observations are related to the infection observations from roughly two weeks prior to a specified observation date.

INTRODUCTION

With the uncertainty around COVID-19 and the factors that could contribute to a rise in cases and deaths, I decided to see if these counts could be explained by environmental markers. Additionally, I wanted to explore the relationship, if it existed, between changes in outdoor mobility to outdoor parks and the COVID-19 daily infection and death counts. Outdoor parks have become an outlet for exercise, socially distanced events, and time away from virtual meetings during the COVID-19 pandemic. In counties where air quality is problematic, I wanted to understand if visiting outdoor parks had a weak, strong, or nonexistent relationship to the COVID-19 daily infection and death counts. It is known that ozone in the atmosphere can exacerbate lung diseases including asthma, emphysema, and bronchitis [7]. Moreover, ozone in the atmosphere can make lungs more susceptible to infection [7]. These established connections have contributed to the designation of individuals with such conditions to high-risk groups during this pandemic. Given that this is a novel coronavirus, there is not a tremendous amount of research to identify possible connections between the air pollutant indicators and COVID-19 counts. Moreover, as social distancing is a relatively new concept, we do not have much information on visiting outdoor parks while social distancing yet additionally having exposure to air pollutants. These connections have not yet been thoroughly explored, which makes it an appealing topic.

DATA INGESTION AND ASSUMPTIONS

The models developed depend on data which can be broken down into the following groups and their respective source: Air quality index (AQI), CO, SO2, NO2, PM10, PM2.5, Ozone (EPA), Outdoor Mobility Data (Google), and COVID-19 Data (John's Hopkins University). For the EPA data, the data comes from both pre-extracted reports produced by the EPA and queries to the EPA's AirNow Developer Tools API. More detail on the data cleaning and merging will be addressed in this section. From the Outdoor Mobility Data report that Google maintains, the "percent parks change from baseline" data was extracted. The COVID-19 Data was pulled from the JHU data repository and has been further cleaned and maintained by Dr. GuanNan Wang.

The data was obtained specifically for the subset of California counties with early surges of COVID-19 cases, which includes, Santa Barbara County, Sacramento County, and Ventura County. While maximum complete for a specific variable covers range of January 1, 2020 through August 17, 2020, the data of all merged variables only overlaps over the range February 16, 2020 through May 30, 2020. Therefore, the timeframe of range February 16, 2020 through May 30, 2020 was used in the modeling methods to consider the complete selection of covariates.

The following table shows the models created and with respective response variables used throughout the analysis process:

Models created	**Response, Y**	**Autoregressive component**
GLM with lag, TSGLM	Death_diff	AR(1), AR(2)
GLM with lag, TSGLM	Infection_diff	AR(1), AR(2)
TSGLM	Death_diff	AR(1), delayed infections

The response variables, Death_diff and Infection_diff, were derived from the cumulative death and infection counts. These response variables represent in the difference of death or infection count, respectively, per day. Assuming that the data has no significant errors, this number should always be greater than or equal to zero, as if think of the cumulative density of these counts, the shape is only steady or increasing. The assumptions of the model is that we are specifically looking at death or infection counts, and do not consider a "recovered" or "removed" compartment. In other words, the construction of these models assumes that an individual who was infected with COVID-19 can recover, but they still contribute an observation to the "infected" count that cannot be removed from the set.

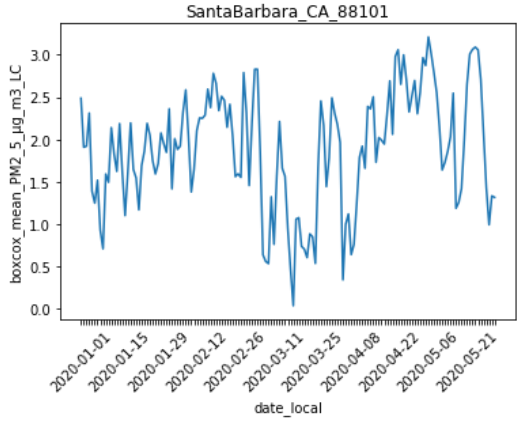
The following table shows the covariates used throughout the analysis process:

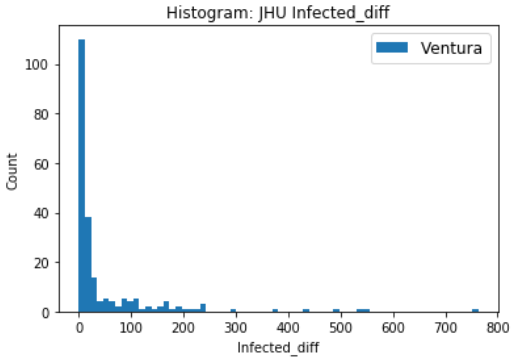
County	Source	Covariates, Xi
Santa Barbara	EPA pollutants	NO2_ppb
		AQI
		Ozone
		PM10
		PM2.5
		SO2
		CO
	Google	Parks Percent change
	JHU *varies for model	Death_diff lagged
		Infection_diff lagged
		Infection_diff lagged optimally
Sacramento	EPA pollutants	NO2_ppb
		AQI
		Ozone
		PM10
		PM2.5
	Google	Parks Percent change
	JHU *varies for model	Death_diff lagged
		Infection_diff lagged
		Infection_diff lagged optimally
Ventura	EPA pollutants	NO2_ppb
		AQI
		Ozone
		PM10
		PM2.5
	Google	Parks Percent change
	JHU *varies for model	Death_diff lagged
		Infection_diff lagged
		Infection_diff lagged optimally

There were a variety of data cleaning techniques implemented to address problems with the data. Modifications include: addressing formatting and naming discrepancies, representing COVID-19 counts as difference (delta) in cumulative counts per day, establishing date conventions, accounting for missing observations, and detecting and removing outliers. The Google "parks percent change from baseline" data was complete and did not present major issues aside from outlier detection and removal as well as repairing missing observations. The EPA pollutants data was the most complicated data to clean. The EPA published pre-extracted datasets for specific environmental indicators on a quarterly basis. Accordingly, I used 7 datasets from the pre-extracted EPA pollutant indicator data which I filtered specifically for the three counties of interest. I noted that it would be advantageous to have a longer span of data, including the months after the initial surge in cases. Therefore, in order to get data through the month of July, I used the EPA's AirNow Developer Tools API to submit my custom queries and obtain more data. In the python file that I created for data cleaning and Exploratory Data Analysis (EDA), I created a function to query the API and return the json file to be processed into a dataframe. From this point, I could neatly clean and re-subset the data, but I noticed some discrepancies in naming conventions used in the pre-extracted data and the API generated data. Therefore, I rigorously cleaned and reorganized the data in a uniform way for ease of merging later. There were different available reading types for each measure, but I decided to use the daily mean of the AQI, CO, SO2, NO2, PM10, PM2.5, and Ozone because it presented a simple interpretation with lower expectation of highly influential outliers.

Histograms and time series plots aided in anomaly detection for each of these pollutant covariates. There was evidence that Outliers were detected and removed through conventional methods; an observation less than *Q1* - 1.5*IQR* or greater than 1.5*IQR* - Q3 was removed from the dataset. Following outlier detection, I used the R package, 'imputeTS', to fill in missing observations through an iterative process of fitting a space model via Auto Regressive Integrated Moving Average modeling (ARIMA) and estimating missing values by Kalman smoothing [9].

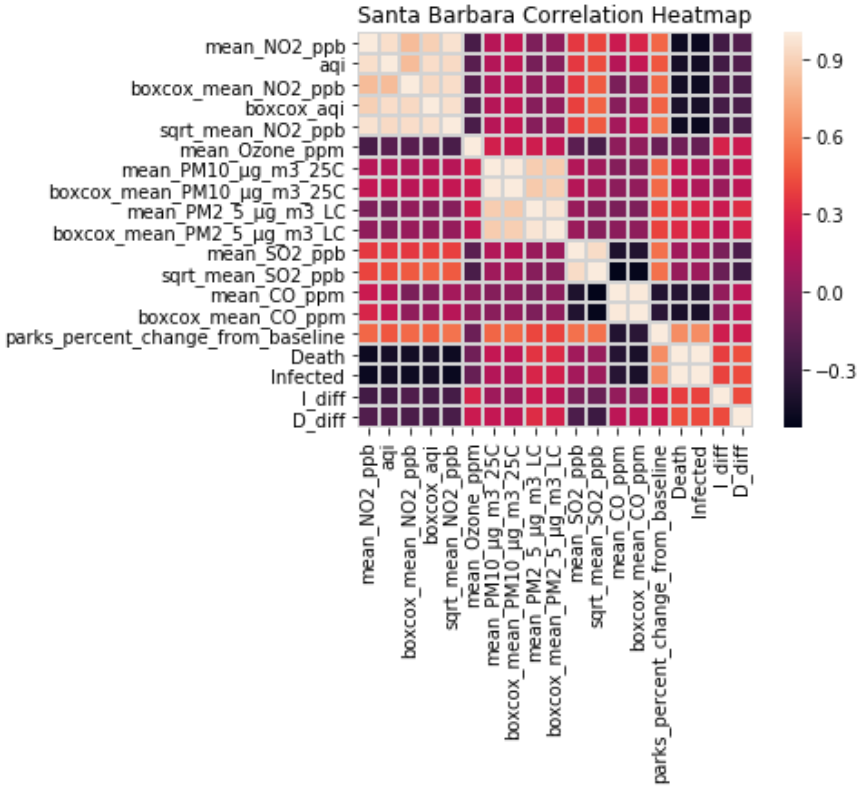
The example graphs below, which display complete and cleaned data, show the importance of plotting to ensure that anomalies have been detected and adequately resolved.





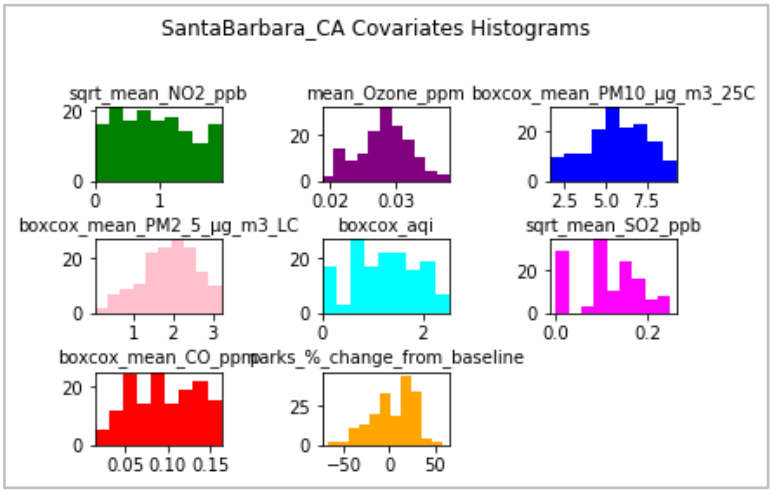
One instance in which the histogram was especially useful was with the JHU daily death count data. The histogram showed relatively low and sensibly counts, while the time series plot revealed a sensible progression, with exception of one extreme observation, in which 28 deaths occurred in a day. As this seemed unreasonable, some news searches revealed that this was in fact an error. A report from the Santa Barbara Edhat explained that the Santa Barbara County Public Health Department had a data error in which there was underreporting, which was accounted for by a jump in 28 deaths on one specific day [10]. The report also had a table which traced back each of the 28 deaths to the specific date on which it occurred, which made it easy for me to edit the data. Accordingly, I updated the data, and the histogram and time series plots looked more sensible.

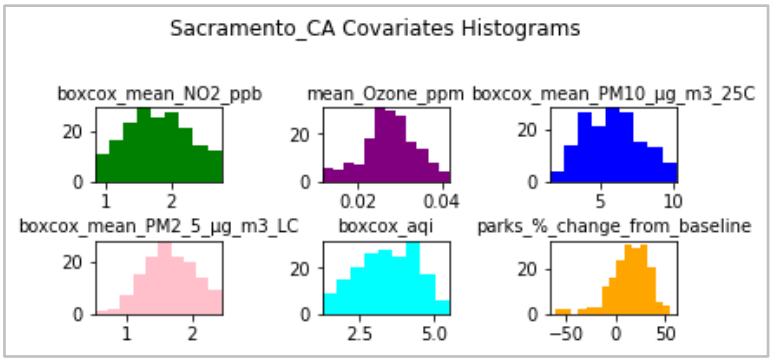
The next step in cleaning the data was to address multicollinearity through the correlation matrix and visual heatmaps. Most notably, some covariates have strong correlation with each other, such as AQI and NO_2, and PM2.5 and PM10. The correlation heatmap of Santa Barbara County reveals this information. As a result, I decided to keep only AQI and PM2.5 as covariates for the modeling. This decision came from the fact that AQI had a stronger correlation with the daily infection count and daily death count than that of NO_2. Likewise, PM2.5 had a stronger correlation with the daily infection count and daily death count than that of PM10. The correlation heatmaps of Ventura County and Sacramento County can be found in the Appendix.

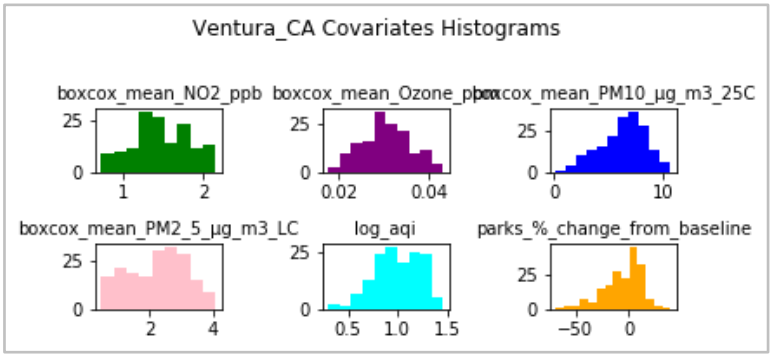


Next, following the cleaning of the data, I applied transformations to the covariates to yield a more normalized dataset. This was a necessary remedy for highly skewed data. Transformations used include square root transformations, log transformations, and Box-Cox transformations. Different transformations were tested out, and I kept the transformation which achieved the most normal looking plot for each covariate. Additionally, zeroes in the data were taken into consideration while deciding how to transform each covariate. These techniques succeeded to yield somewhat normal distributions for each covariate.

The figures below show the histograms of the transformed covariates.







MODEL DEVELOPMENT AND VALIDATION

This section addresses the modeling process and validation considerations. Following the data cleaning procedure, I generated an organized dataset for each county, which included two types of response variables (daily death count and daily infection count) as well as covariates from the EPA data and Google mobility data. Each response variable is represented in its own model, as we only want one response. Restating earlier modifications, the response variables, daily infection count and daily death count, are counts per day, and not cumulative. Thus, we can use the Poisson log link regression to model the data. The dataset was split into two subsets, train and test. The train data contained roughly 100 days while the test data contained 7 days. It would be reasonable to expect that one week could be predicted with the given training data. The models were developed strictly with the train data and validated strictly with the preserved test data. To measure predictive accuracy and validate the model, I focused on the mean square prediction error, for which the formula is provided below. The summation is over *n* observations in the test set.

$$\text{MSPE} = \frac{1}{n}\sum_{i=0}^{n}(\hat{y} - y)^2$$

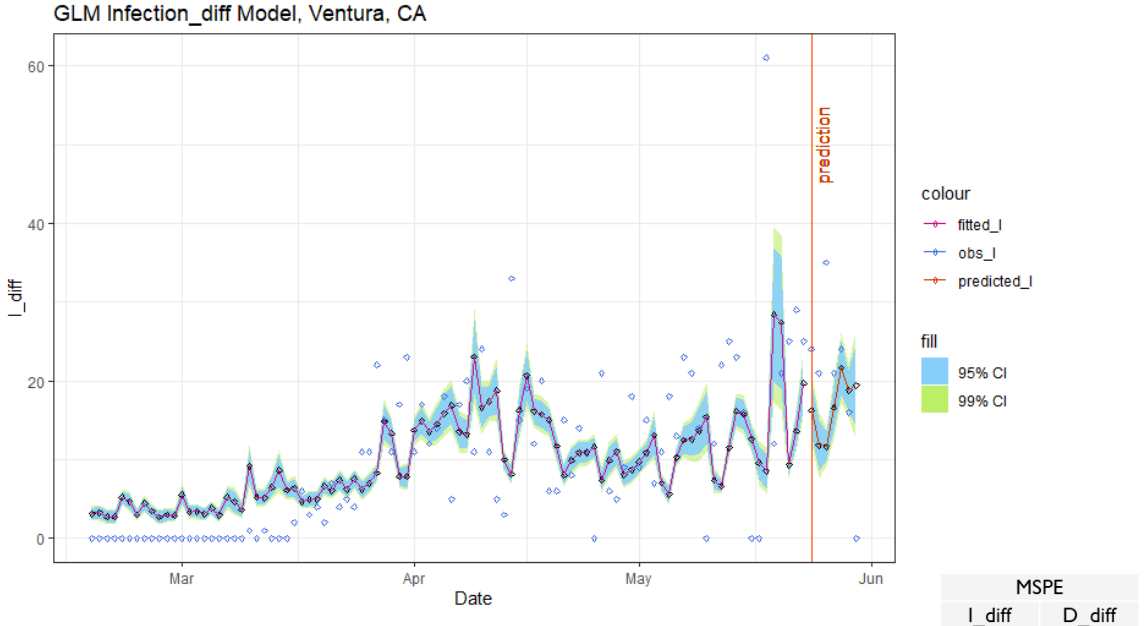
Three phases were implemented during the model development. Ultimately the goal was to create a model specifically for daily death count, as death is a more absolute reported value. The reporting for infections is not as accurate as the reporting for deaths due to the nature of severity between the two scenarios. An individual who is asymptomatic may not take a test during the period that he or she was infected, and so that observation is lost. Therefore, death is a more absolute and definite measure. Nevertheless, building models with daily infection count as the response variable is a valid exploratory measure and can be an additional source of insight for the death model. For example, if there were no significant coefficients for the pollutant covariates in the infection model, it would be hard to believe that there could be a significant coefficient for the pollutant covariates in the death model. On the other hand, if multiple covariates had significant coefficients in the infection response model, and one of those covariates also had a significant coefficient in the death response model, that could be explainable and sensible. This is because if we expect that the significant covariates can explain the response in the death model, it is likely that at least those covariates would explain the response in the infection model. Moreover, I thought that it would be interesting to see which covariates, if any, were significant and could explain the response of the infection model.

To begin the modeling efforts, a baseline Generalized Linear Model (GLM) was created for each response variable and for each county. I used the R package, 'mgcv.' I manually created an autoregressive component to the model by generating two columns in the covariate matrix to correspond to y_{t-1} and y_{t-2} for both the death response model and the infection response model.

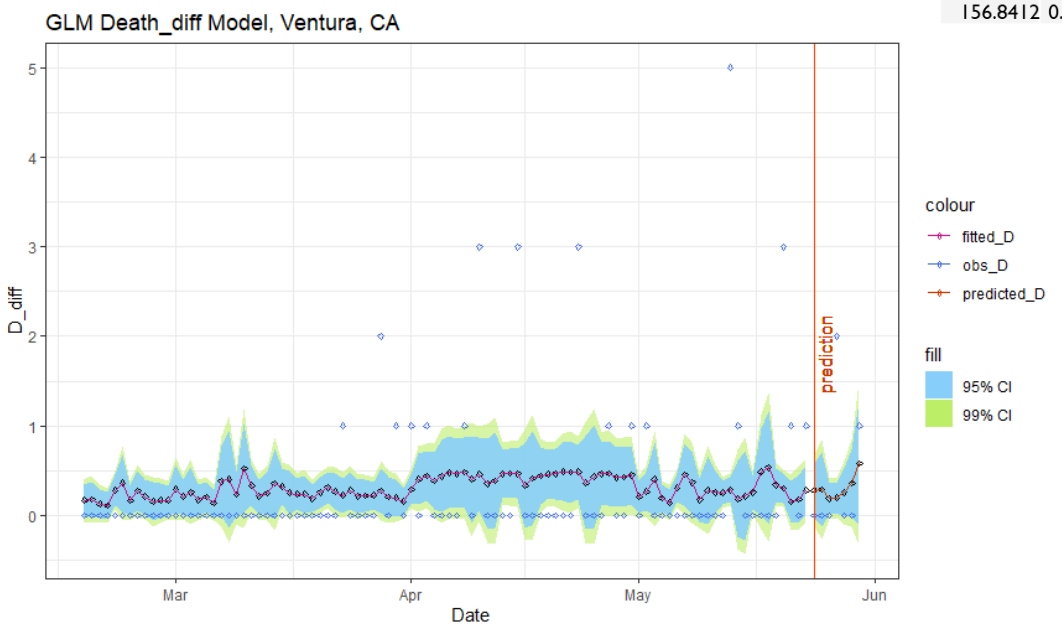
This resulted in 6 models. The general equation for this baseline Poisson AR(2) model is stated below.

$$E[y_t|X_t,y_{t-1},y_{t-2}] = \exp(\delta_0 + X_t\delta_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t)$$

The baseline model was generated to see and understand how each model fit on a time series plot, evaluate significant coefficients, and determine the mean squared prediction error of the model. I will use Ventura County for the example plots and output throughout the analysis, and the plots and output of the other two counties can be found in the Appendix. Example plots for the death response model and infection response model are displayed below.



MSPE	
I_diff	D_diff
156.8412	0.5457606



The model output for death response model for Ventura County is displayed below. We can see that at this point, with the MSPE quite high for the infection response model (156.84), there is no use in interpreting the hypothesis test applied to the coefficients at this point. For the death response model, we can explore the hypothesis test results and note that there are many significant coefficients at this phase in the model development. The Appendix contains additional model output for the other counties.

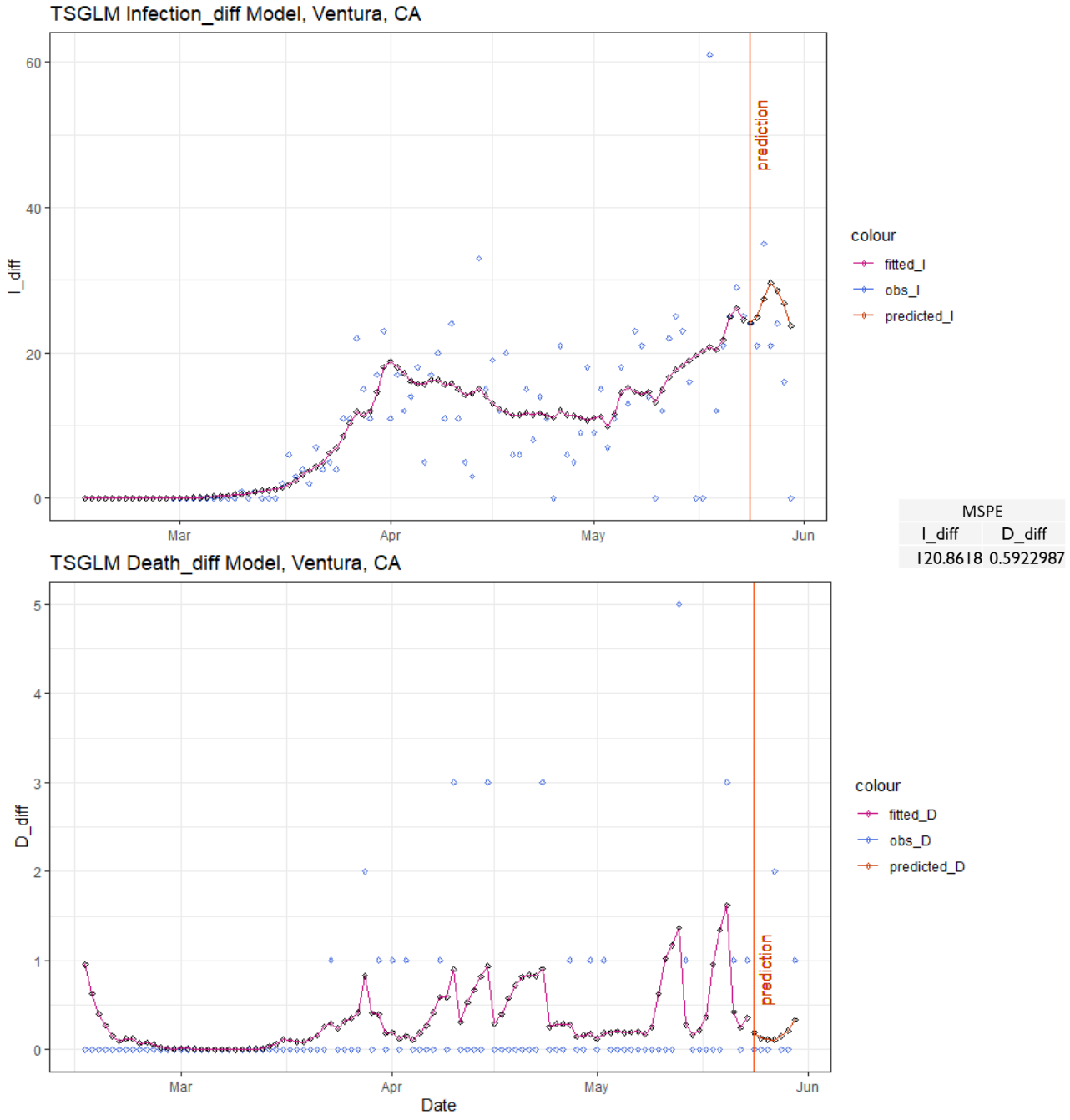
```
Call:
glm(formula = (D_diff ~ D_diff_t_1 + D_diff_t_2 + log_aqi + boxcox_mean_Ozone_ppm +
    boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = V_glm_train_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0142  -0.8767  -0.7059  -0.5736   4.4272

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       0.874798   1.622329   0.539   0.5897
D_diff_t_1                       -0.056498   0.246875  -0.229   0.8190
D_diff_t_2                       -0.022571   0.248902  -0.091   0.9277
log_aqi                          -1.122400   1.214258  -0.924   0.3553
boxcox_mean_Ozone_ppm           -29.798062  17.403336  -1.712   0.0869 .
boxcox_mean_PM2_5_Âµg_m3_LC      -0.116773   0.239717  -0.487   0.6262
parks_percent_change_from_baseline  0.004154   0.014372   0.289   0.7725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 113.44  on 93  degrees of freedom
Residual deviance: 109.32  on 87  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 161.38

Number of Fisher Scoring iterations: 7
```

```
Call:
glm(formula = (I_diff ~ I_diff_t_1 + I_diff_t_2 + log_aqi + boxcox_mean_Ozone_ppm +
    boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = V_glm_train_dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.549  -2.607  -1.009   1.056  11.597

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       2.944529   0.332587   8.853  < 2e-16 ***
I_diff_t_1                        0.026222   0.002802   9.360  < 2e-16 ***
I_diff_t_2                        0.022744   0.002947   7.718 1.18e-14 ***
log_aqi                          -1.071681   0.244750  -4.379 1.19e-05 ***
boxcox_mean_Ozone_ppm           -12.607467   3.228812  -3.905 9.44e-05 ***
boxcox_mean_PM2_5_Âµg_m3_LC      -0.014566   0.043672  -0.334  0.73873
parks_percent_change_from_baseline  -0.007489   0.002710  -2.763  0.00572 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1066.27  on 95  degrees of freedom
Residual deviance:  782.91  on 89  degrees of freedom
AIC: 1081.1

Number of Fisher Scoring iterations: 6
```

In the second phase of the modeling, the Time Series following Generalized Linear Model (TSGLM) was implemented as a Poisson AR(1) MA(1) model. Once again, the model was built separately for each of the response variables, daily death count and daily infection count. The general model equation is displayed below.

$$E[y_t|\mathbf{X}_t, y_{t-1}, y_{t-2}] = \exp(\delta_0 + \mathbf{X}_t\delta_1 + \alpha_1 y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t)$$

Below shows the plot for the Ventura County models and the associated MSPE. Both models show improved fit and reduced MSPE. The other counties' plots can be found in the Appendix.



MSPE	
I_diff	D_diff
120.8618	0.5922987

The output below reveals that in the infection model parks percent change from baseline was a significant coefficient with a positive coefficient and the Box-Cox transformed mean of PM2.5 was a significant coefficient with a negative coefficient for the infection response model. Once again, because the MSPE is so high (although reduced from the baseline model), it is difficult to infer anything definitive. In the death response model, it is evident that the Box-Cox transformed the mean Ozone and the Box-Cox mean PM2.5 are both significant coefficients, negative and positive, respectively. The Ozone coefficient, which is negative, has a much stronger value. This finding is interesting, and it is something to further explore in the next phase of the model development. The model outputs from the other counties can be found in the Appendix.

```
Call:
tsglm(ts = V_tsglm_train_dat$I_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = V_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -0.3682    0.13987   -0.64237   -0.09410
beta_1                             -0.0451    0.04611   -0.13552    0.04525
alpha_1                             1.0000    0.02748    0.94613    1.05387
log_aqi                             0.5079    0.13568    0.24199    0.77384
boxcox_mean_Ozone_ppm               6.1450    1.22297    3.74799    8.54193
boxcox_mean_PM2_5_Âµg_m3_LC         -0.0375    0.02034   -0.07736    0.00238
parks_percent_change_from_baseline  0.0014    0.00155   -0.00163    0.00443
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -305.873
AIC: 625.7459
BIC: 643.8407
QIC: 538.7396

Call:
tsglm(ts = V_tsglm_train_dat$D_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = V_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        1.6913    0.74155     0.2379    3.14476
beta_1                            -1.0000    0.26744    -1.5242   -0.47583
alpha_1                            1.0000    0.05091     0.9002    1.09978
log_aqi                           -1.6362    0.56197    -2.7376   -0.53475
boxcox_mean_Ozone_ppm             -3.3954    5.80151   -14.7661    7.97538
boxcox_mean_PM2_5_Âµg_m3_LC         0.0142    0.06293    -0.1092    0.13751
parks_percent_change_from_baseline -0.0125   0.00375    -0.0199   -0.00519
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -58.04896
AIC: 130.0979
BIC: 148.1927
QIC: 85.04033
```

In the third phase of modeling, once again a TSGLM model was used, with the Poisson AR(1) MA(1), but this time, the model was only generated for daily death count as the response variable. To make the model more thorough with data that is already available, a column for delayed infection is introduced to the covariate regression matrix. As previously, the general equation for the model is displayed below.

$$E[y_t|\mathbf{X}_t, y_{t-1}, y_{t-2}] = \exp(\delta_0 + \mathbf{X}_t\delta_1 + \alpha_1 y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t)$$

However, we have a new variable in the covariate matrix, $x_{delay\_infection} = X_{t-d}$, where d is an amount of days in which the infection count is delayed. To determine the optimal amount of delay applied to this measure, I created a function to iterate through the through model generation, altering the column associated with delayed daily infection count, I_diff. This way, for each iteration, the delay value increases by 1 day (starting from -15). The MSPE is saved to a table, and then it will be clear which delay time produces the most effective model by inspecting which delay amount the smallest MSPE corresponds to. Pseudo code for this method, which I name 'model_compare,' is written below. The input, dat_df is a dataframe for the complete dataset for a particular county, containing both response and predictor variables as well as train and test time spans of the data.

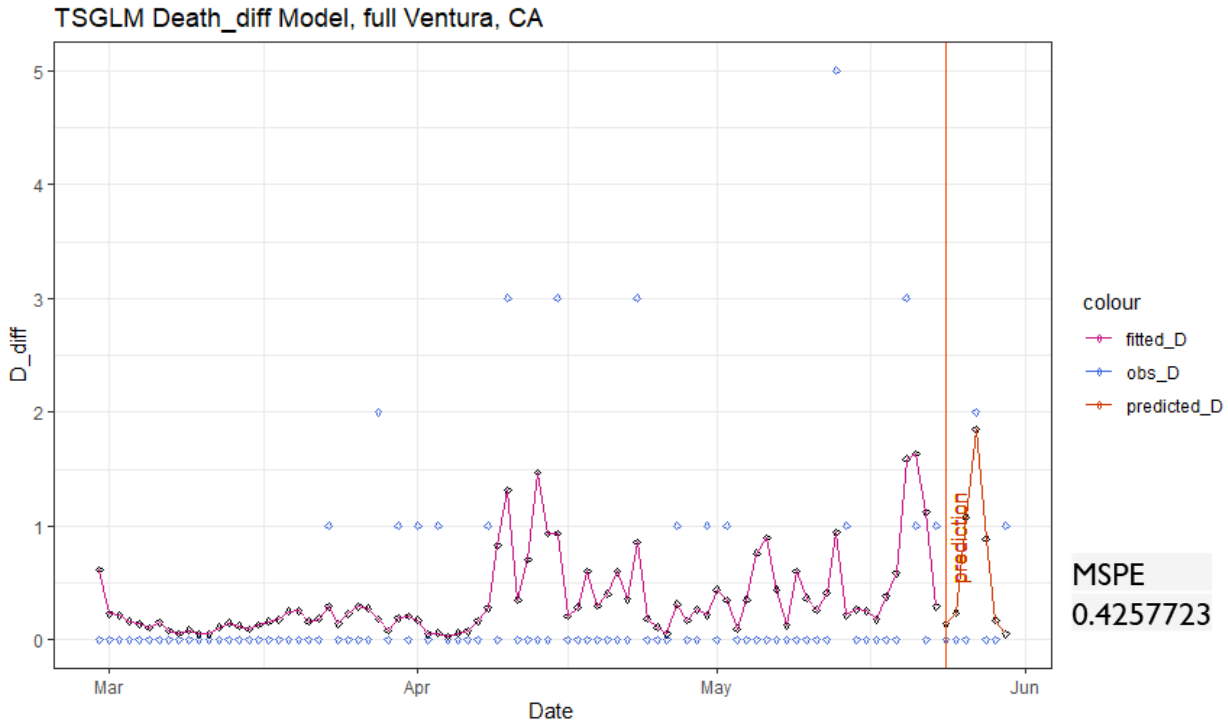
```
model_compare <- function(dat_df)
    For delay_amt in 1...15:
        Create vector where delayed_I_diff = I_diff_{t-delay_amt}
        Add vector delayed_I_diff as a column to existing covariate matrix
        Split train and test data
        Generate TSGLM Poisson AR(1) MA(1) model with selected input
        Generate predicted values
        Send model through MSPE calculating function
        Append MSPE to results_table
    Return results_table
```

The complete code can be found in the associated R scripts written to develop the models and conduct this study. Below is an example of the output:

```
model_compare(V_tot_dat)
 X1 X0.529175594025805
  1           0.5291756
  2           0.5681251
  3           0.5376217
  4           0.5181157
  5           0.5493178
  6           0.6312362
  7           0.5309314
  8           0.5264215
  9           0.5310032
 10           0.6324898
 11           0.5756301
 12           0.5339643
 13           0.4257723
 14           0.5394722
 15           0.5921500
```

Therefore, 13 days was used for the delay period in the delayed daily infection counts column of the covariate matrix. This version of the model performed best when the MSPE is compared to the previous models, and it is in part because of the introduced covariate.

Below is the plot for Ventura County associated with this updated model.



The model output is displayed below.

```
Call:
tsglm(ts = V_tot_train$D_diff, model = list(past_obs = 1, past_mean = 1),
    xreg = tmp_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -3.0622    1.58721    -6.1730    0.04869
beta_1                             -0.7398    0.26066    -1.2507   -0.22892
alpha_1                             0.7187    0.13759     0.4491    0.98842
I_diff_t_n                          0.0797    0.02266     0.0353    0.12412
log_aqi                             1.5669    1.16548    -0.7174    3.85121
boxcox_mean_Ozone_ppm              34.1514   14.21933     6.2820   62.02078
boxcox_mean_PM2_5_Âµg_m3_LC         -0.1627    0.12138    -0.4006    0.07521
parks_percent_change_from_baseline -0.0121    0.00863    -0.0290    0.00481
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 8
Log-likelihood: -60.16595
AIC: 136.3319
BIC: 155.8731
QIC: 137.4416
```

Significant coefficients for this model include Box-Cox transformation of mean Ozone as well as the daily infection count with a 13-day delay. The Ozone covariate has a strong positive

coefficient, and the delayed infection coefficient has a weak positive coefficient. The parks percent change from baseline contributes a weak negative coefficient.

The table below shows the MSPE values from all of the constructed models.

	MODEL	GLM		TSGLM		TSGLM (updated)	
	Response, Y	I_diff	D_diff	I_diff	D_diff	D_diff	*optimal delay days for I_diff:
County	Santa Barbara	520.2754	0.0017114	501.3653	0.0046582	0.0014522	3
	Sacramento	64.32438	0.2364802	178.8306	0.0746677	0.0028939	14
	Ventura	156.8412	0.5457606	120.8618	0.5922987	0.4257723	13

It is evident that the most complex updated TSGLM model yields quite low MSPE values, which validates its effectiveness in modeling the scenario.

CONCLUSION AND FUTURE WORK

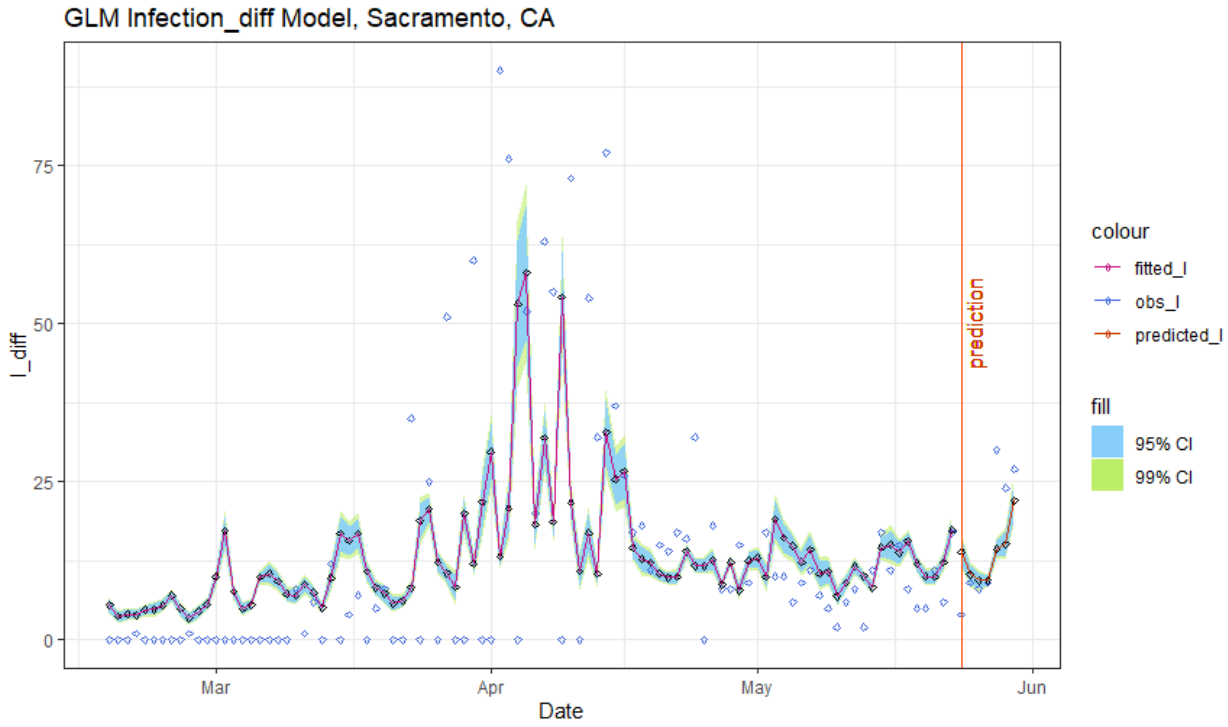
The main objective of the research study, to develop a predictive model of COVID-19 mortality and uncover its relationship with environmental pollutants and outdoor mobility, brought forward interesting findings. To begin with, the Poisson TSGLM AR(1) MA(1) models with response variable daily death count (and with delayed infections introduced to the covariate matrix) yielded significant regression coefficients. For Ventura County, the covariates with significant regression coefficients include the mean Ozone (with Box-Cox transformation), park visits as a percent change from baseline, and the daily infection count with a 13-day delay. The Ozone covariate has a strong positive coefficient, and the delayed infection coefficient has a weak positive coefficient. The park visits as a percent change from baseline contributes a weak negative coefficient. As the expected value of the dependent response variable, daily death count, is related to the covariates through the log-link, and so the Ozone and delayed infection count contribute to greater daily death counts while the park visits as a percent change from baseline contributes to a reduced daily death count expectation. For Sacramento County, the covariates with significant regression coefficients include park visits as a percent change from baseline, and the daily infection count with a 14-day delay. Most notably, the park visits as a percent change from baseline contributes a weak negative coefficient. Finally, for Santa Barbara County, there were no covariates with significant regression coefficients, which could be due to a large amount of zeroes in the daily death count. The full model outputs for each phase of modeling can be found in the Appendix.

This research revealed that two out of the three California counties studied revealed a relationship between the environmental pollutants, outdoor mobility, and the response of daily death count from COVID-19. While the scope of the research focused on February until June of 2020, it would be interesting to develop the model through the late summer months and early fall, during which California was struck by a wave of wildfires. The pollutant measurements during that period could potentially reveal a stronger relationship between the more extreme values of pollutant observations and the daily death count as a result of COVID-19. I am interested in gathering data from that time and using it as input to develop a new predictive model. The three California counties initially used would be appropriate for the study. If this research was done with a larger team of people and multiple computers, I would suggest developing the model on a larger scale, looking at perhaps a complete state. If the entire country was to be used for the model scope, then it would be necessary to add in covariates to represent policy markers implemented by each state to reduce the spread of COVID-19. I believe that expanding the model scope could lead to more interesting findings.

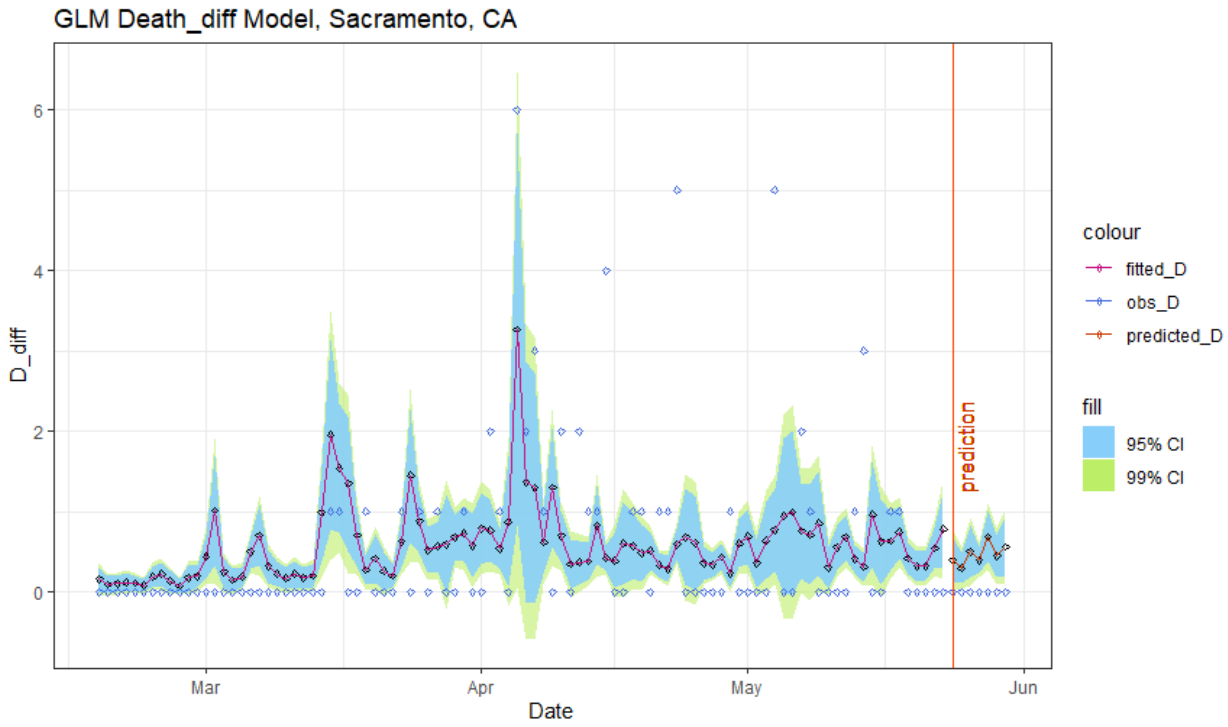
APPENDIX

The python data ingestion and EDA notebook, R modeling scripts, and source data files, as well as the exported datasets from the python data ingestion and EDA can be found at: https://github.com/4dh/COVID19_Enviromental_project

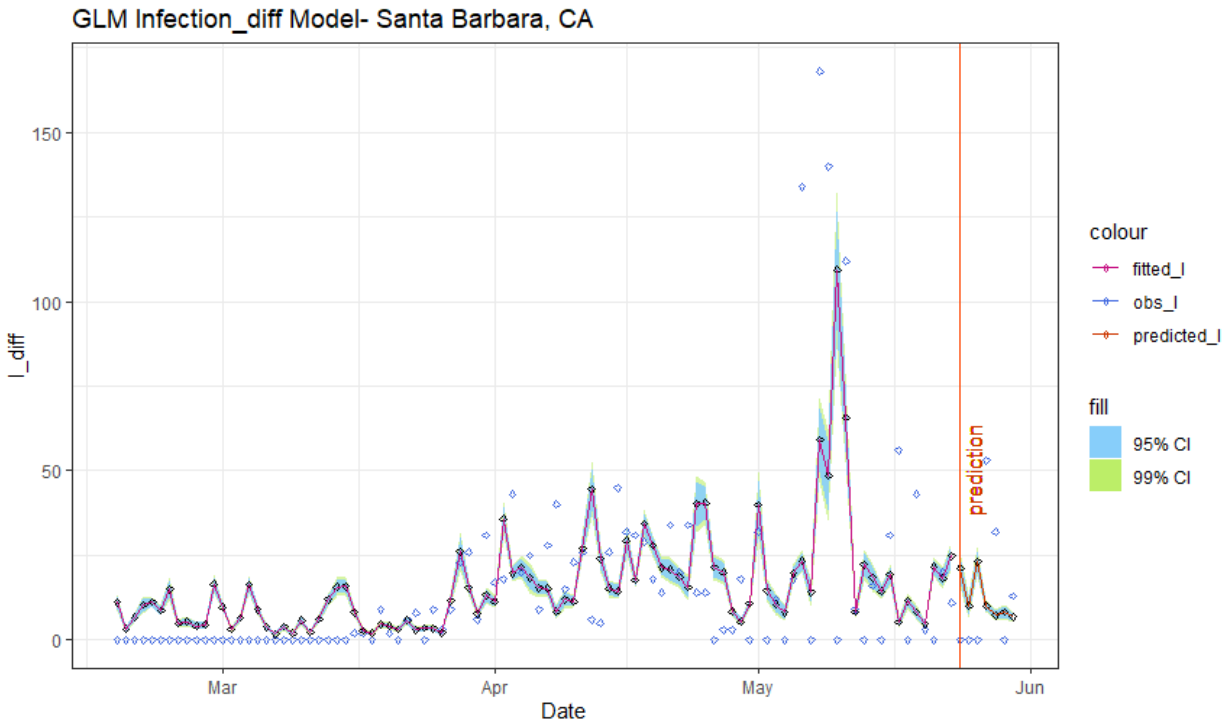
Plots for all phases of the model and their respective model fit output:



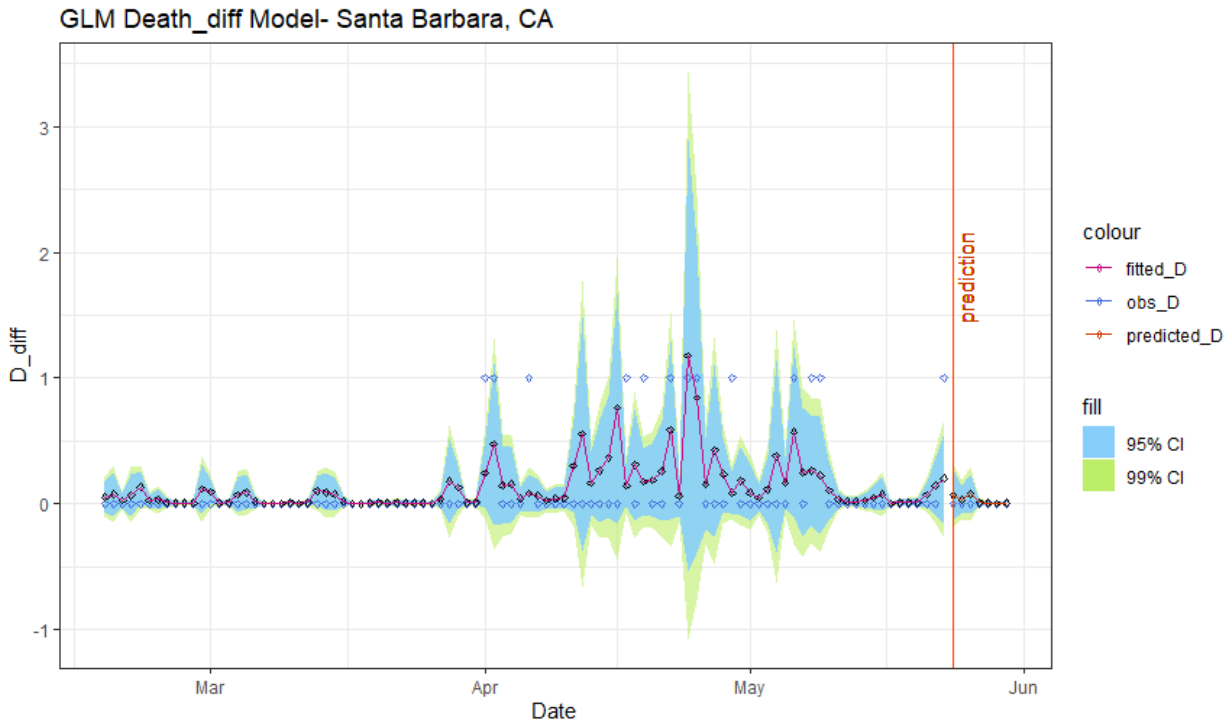

```
Call:
glm(formula = (I_diff ~ I_diff_t_1 + I_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = S_glm_train_dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.4097   -3.3565   -1.5944    0.6099   13.8349

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       2.914297   0.283594  10.276  < 2e-16 ***
I_diff_t_1                        0.010361   0.001267   8.180 2.83e-16 ***
I_diff_t_2                        0.014759   0.001188  12.422  < 2e-16 ***
boxcox_aqi                        0.060283   0.037391   1.612  0.10691
mean_Ozone_ppm                   20.623554   6.746463   3.057  0.00224 **
boxcox_mean_PM2_5_Âµg_m3_LC      -1.040553   0.092220 -11.283  < 2e-16 ***
parks_percent_change_from_baseline  0.002391   0.001693   1.413  0.15780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2253.2  on 95  degrees of freedom
Residual deviance: 1757.4  on 89  degrees of freedom
AIC: 2044.2

Number of Fisher Scoring iterations: 6
```



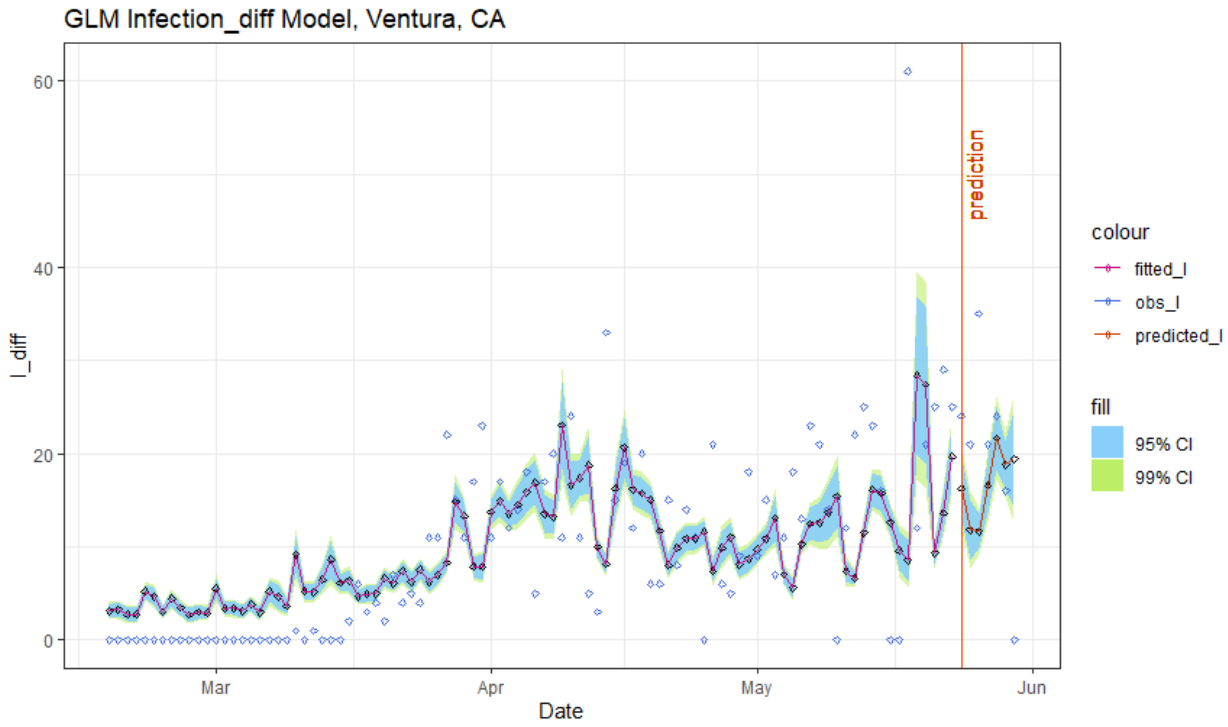

```
Call:
glm(formula = (D_diff ~ D_diff_t_1 + D_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = S_glm_train_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7027  -1.0802  -0.6714   0.3440   3.5298

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.673396   1.287573  -0.523  0.60098
D_diff_t_1                       0.068793   0.102627   0.670  0.50265
D_diff_t_2                       0.093486   0.097649   0.957  0.33839
boxcox_aqi                       0.100560   0.188277   0.534  0.59327
mean_Ozone_ppm                  56.410405  31.236183   1.806  0.07093 .
boxcox_mean_PM2_5_Âµg_m3_LC     -1.368833   0.433598  -3.157  0.00159 **
parks_percent_change_from_baseline -0.011354   0.007201  -1.577  0.11485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.19  on 95  degrees of freedom
Residual deviance: 125.59  on 89  degrees of freedom
AIC: 212.52

Number of Fisher Scoring iterations: 6
```



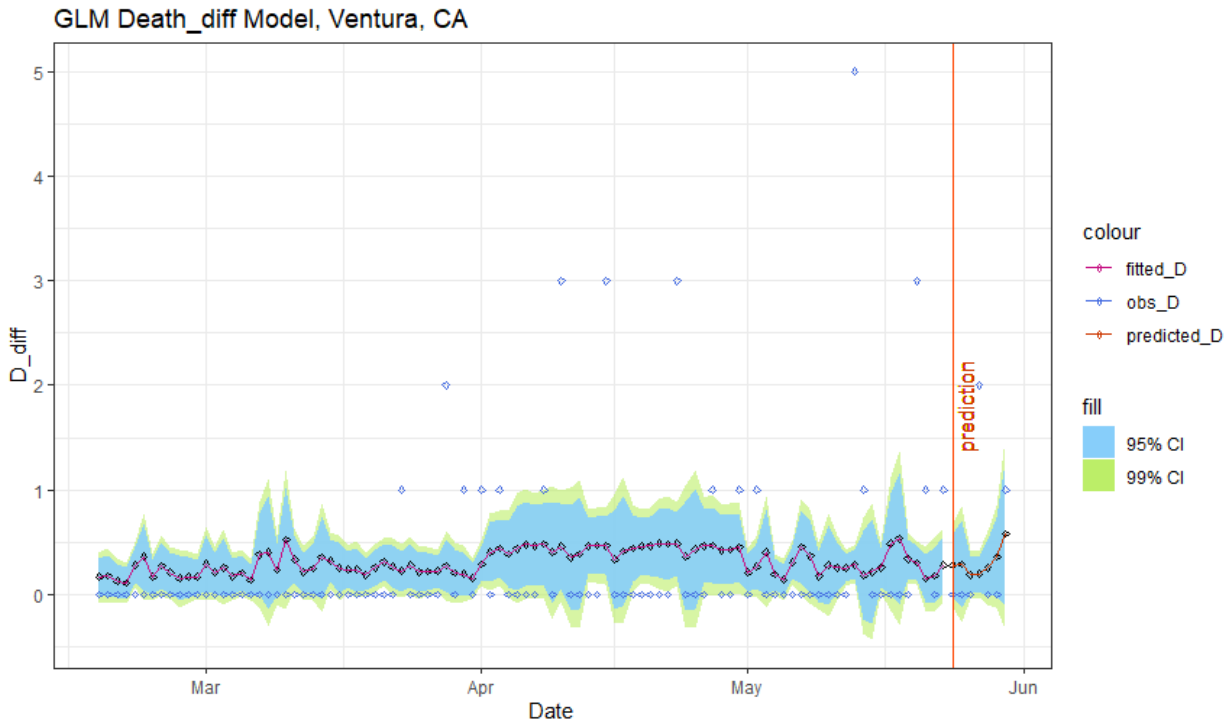

```
Call:
glm(formula = (I_diff ~ I_diff_t_1 + I_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + sqrt_mean_SO2_ppb +
    boxcox_mean_CO_ppm + parks_percent_change_from_baseline),
    family = poisson(), data = SB_glm_train_dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.7875   -4.2843   -1.9781    0.9918   15.7164

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -0.8200422  0.2564743  -3.197  0.00139 **
I_diff_t_1                         0.0017599  0.0006763   2.602  0.00926 **
I_diff_t_2                         0.0102974  0.0005715  18.017  < 2e-16 ***
boxcox_aqi                        -0.6143757  0.0626466  -9.807  < 2e-16 ***
mean_Ozone_ppm                    14.1247870  8.4418358   1.673  0.09429 .
boxcox_mean_PM2_5_Âµg_m3_LC         0.8477516  0.0548520  15.455  < 2e-16 ***
sqrt_mean_SO2_ppb                  6.7308520  0.6348776  10.602  < 2e-16 ***
boxcox_mean_CO_ppm                 8.2568887  1.1525782   7.164 7.84e-13 ***
parks_percent_change_from_baseline -0.0193159  0.0022913  -8.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3203.5  on 95  degrees of freedom
Residual deviance: 2143.2  on 87  degrees of freedom
AIC: 2422.2

Number of Fisher Scoring iterations: 6
```



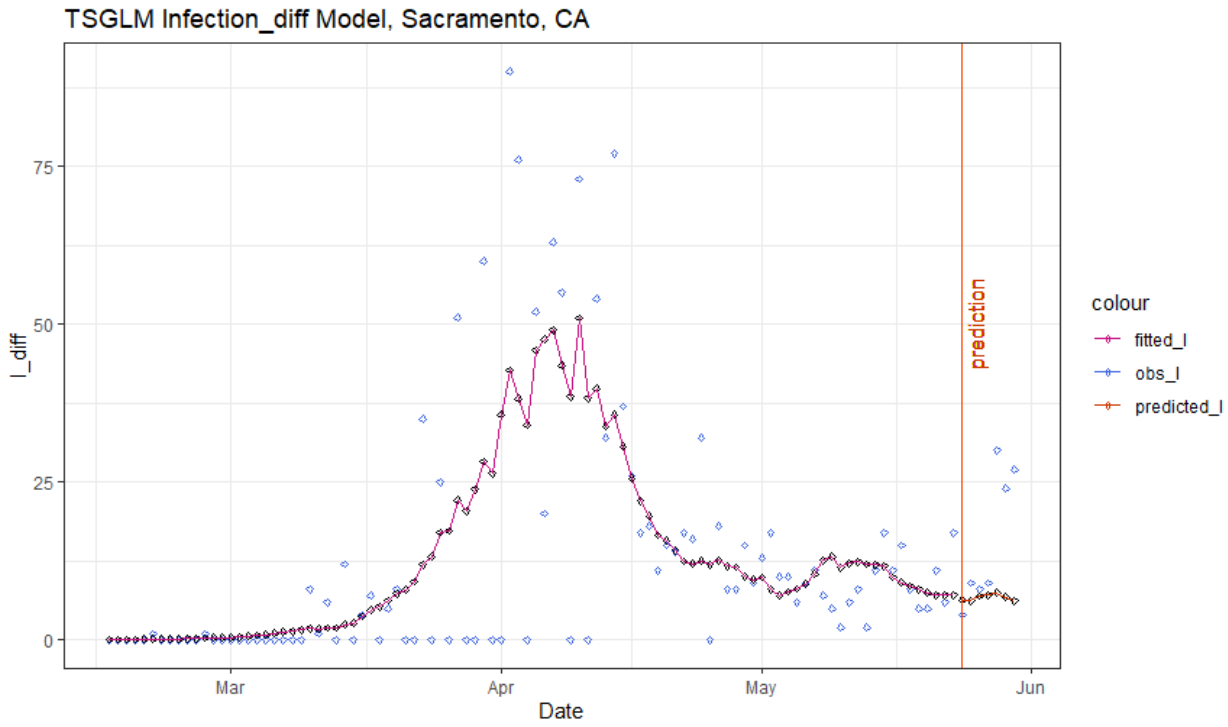

```
glm(formula = (D_diff ~ D_diff_t_1 + D_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + sqrt_mean_SO2_ppb +
    boxcox_mean_CO_ppm + parks_percent_change_from_baseline),
    family = poisson(), data = SB_glm_train_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2344  -0.4651  -0.2238  -0.1064   1.7673

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -8.90637    3.48472  -2.556   0.0106 *
D_diff_t_1                        -0.35145    0.71600  -0.491   0.6235
D_diff_t_2                        -0.55895    0.78061  -0.716   0.4740
boxcox_aqi                        -0.67230    0.70861  -0.949   0.3427
mean_Ozone_ppm                    77.50299   98.01493   0.791   0.4291
boxcox_mean_PM2_5_Âµg_m3_LC         1.87653    0.80985   2.317   0.0205 *
sqrt_mean_SO2_ppb                 -2.81847    6.93670  -0.406   0.6845
boxcox_mean_CO_ppm                 7.88604   12.58451   0.627   0.5309
parks_percent_change_from_baseline -0.02335    0.02751  -0.849   0.3961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 51.984  on 95  degrees of freedom
Residual deviance: 32.526  on 87  degrees of freedom
AIC: 76.526

Number of Fisher Scoring iterations: 6
```



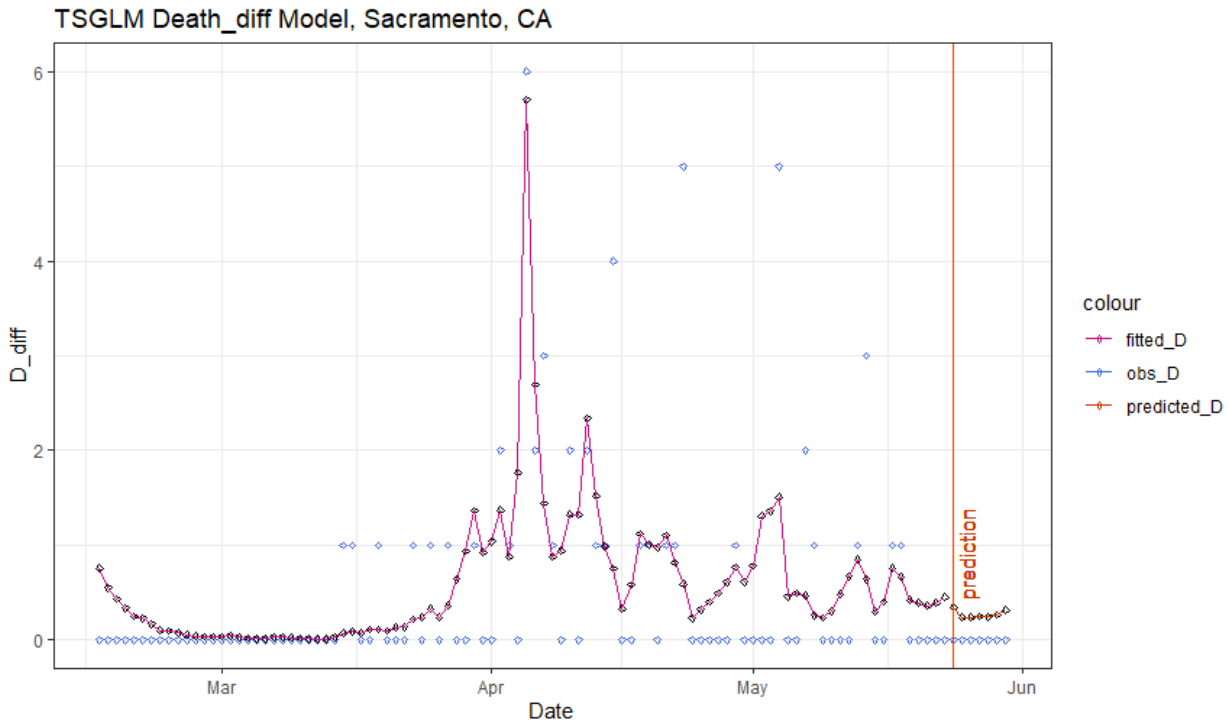

```
Call:
glm(formula = (I_diff ~ I_diff_t_1 + I_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = S_glm_train_dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.4097   -3.3565   -1.5944    0.6099   13.8349

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       2.914297   0.283594  10.276  < 2e-16 ***
I_diff_t_1                        0.010361   0.001267   8.180 2.83e-16 ***
I_diff_t_2                        0.014759   0.001188  12.422  < 2e-16 ***
boxcox_aqi                        0.060283   0.037391   1.612  0.10691
mean_Ozone_ppm                   20.623554   6.746463   3.057  0.00224 **
boxcox_mean_PM2_5_Âµg_m3_LC      -1.040553   0.092220 -11.283  < 2e-16 ***
parks_percent_change_from_baseline  0.002391   0.001693   1.413  0.15780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2253.2  on 95  degrees of freedom
Residual deviance: 1757.4  on 89  degrees of freedom
AIC: 2044.2

Number of Fisher Scoring iterations: 6
```



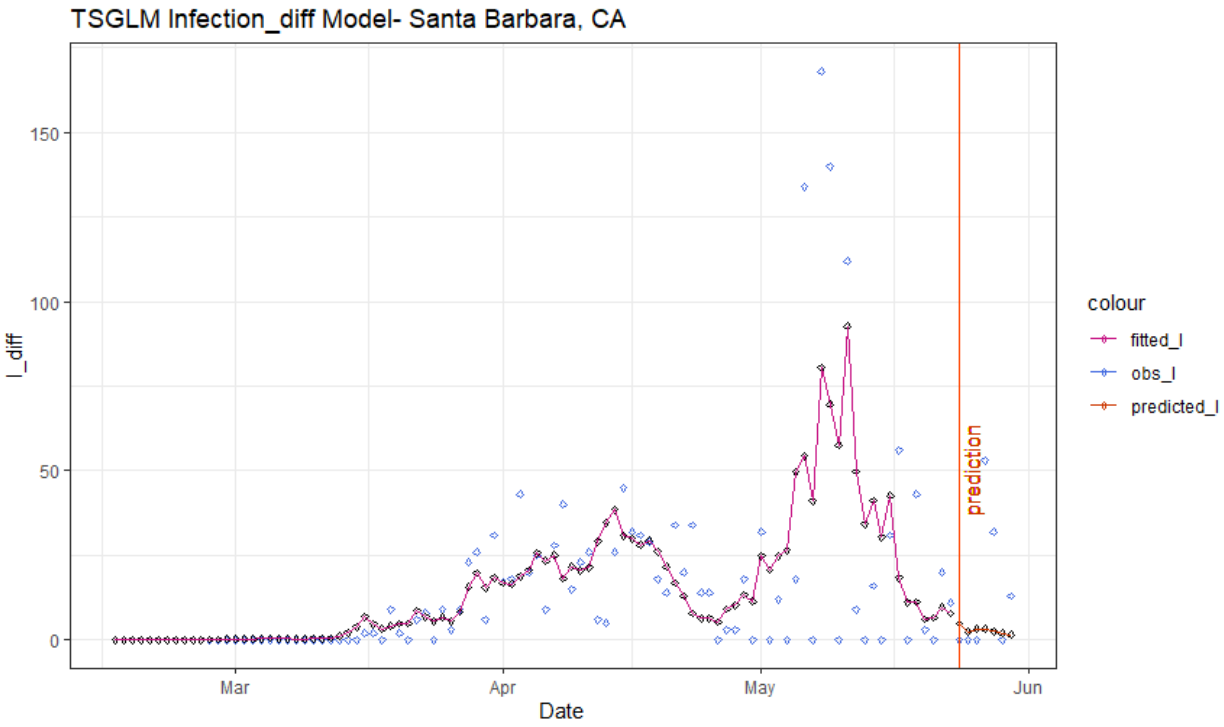

```
Call:
glm(formula = (D_diff ~ D_diff_t_1 + D_diff_t_2 + boxcox_aqi +
    mean_Ozone_ppm + boxcox_mean_PM2_5_Âµg_m3_LC + parks_percent_change_from_baseline),
    family = poisson(), data = S_glm_train_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7027  -1.0802  -0.6714   0.3440   3.5298

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.673396   1.287573  -0.523  0.60098
D_diff_t_1                       0.068793   0.102627   0.670  0.50265
D_diff_t_2                       0.093486   0.097649   0.957  0.33839
boxcox_aqi                       0.100560   0.188277   0.534  0.59327
mean_Ozone_ppm                  56.410405  31.236183   1.806  0.07093 .
boxcox_mean_PM2_5_Âµg_m3_LC     -1.368833   0.433598  -3.157  0.00159 **
parks_percent_change_from_baseline -0.011354   0.007201  -1.577  0.11485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.19  on 95  degrees of freedom
Residual deviance: 125.59  on 89  degrees of freedom
AIC: 212.52

Number of Fisher Scoring iterations: 6
```



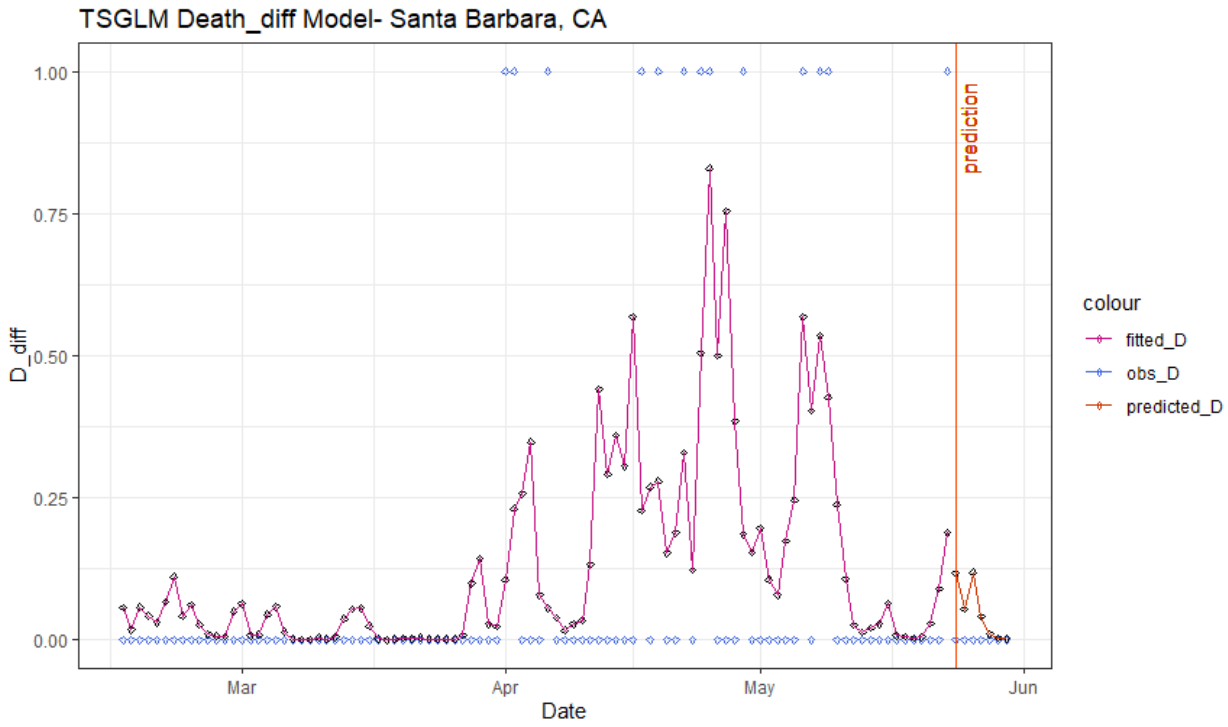

```
Call:
tsglm(ts = S_tsglm_train_dat$I_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = S_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -0.2289   0.053596   -0.33393  -0.123838
beta_1                             -0.0755   0.014437   -0.10384  -0.047249
alpha_1                             1.0000   0.013781    0.97299   1.027011
boxcox_aqi                          0.1020   0.015647    0.07136   0.132690
mean_Ozone_ppm                      8.2946   2.188566    4.00506  12.584084
boxcox_mean_PM2_5_Âµg_m3_LC         -0.1086   0.040430   -0.18780  -0.029320
parks_percent_change_from_baseline -0.0027   0.000952   -0.00457  -0.000834
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -623.0784
AIC: 1260.157
BIC: 1278.252
QIC: 1264.262
```



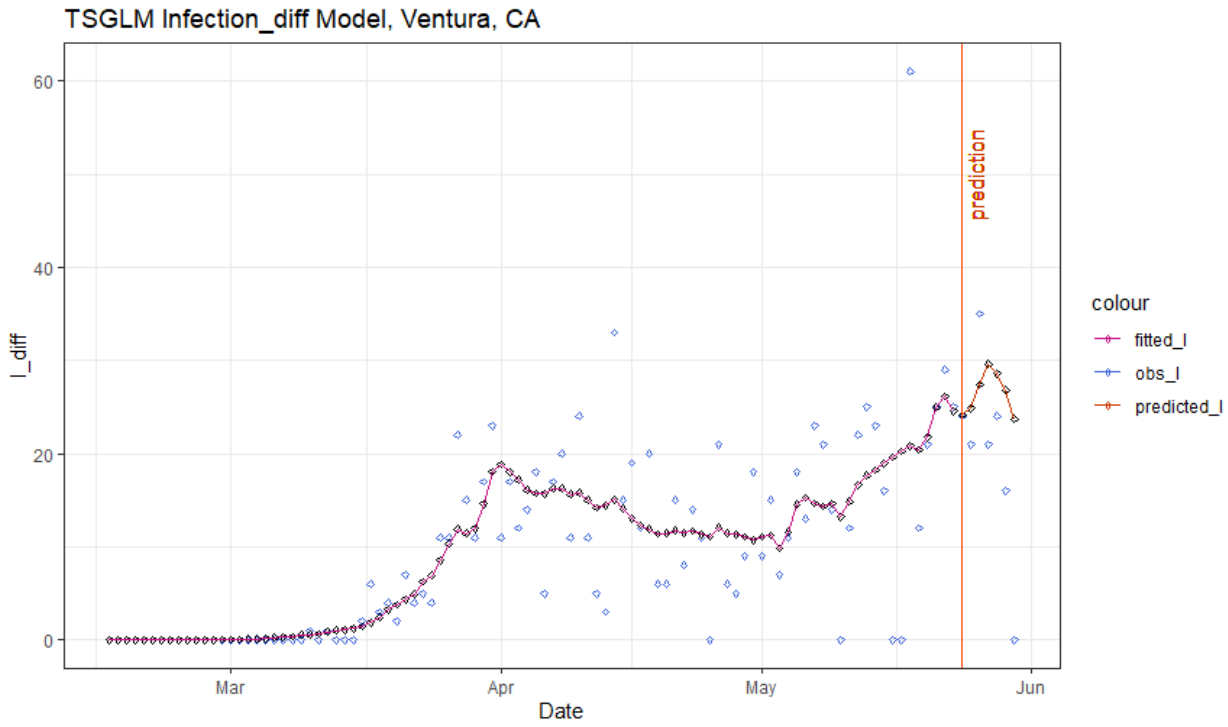

```
Call:
tsglm(ts = S_tsglm_train_dat$D_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = S_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                         0.5615    0.42915    -0.2796    1.40262
beta_1                             -0.6740    0.17408    -1.0152   -0.33280
alpha_1                             1.0000    0.04411     0.9135    1.08646
boxcox_aqi                         -0.1113    0.08046    -0.2690    0.04641
mean_Ozone_ppm                      1.6976    6.65351   -11.3431   14.73819
boxcox_mean_PM2_5_Âµg_m3_LC          0.0158    0.16075    -0.2993    0.33087
parks_percent_change_from_baseline -0.0130    0.00237    -0.0177   -0.00838
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -85.83533
AIC: 185.6707
BIC: 203.7654
QIC: 192.1582
```



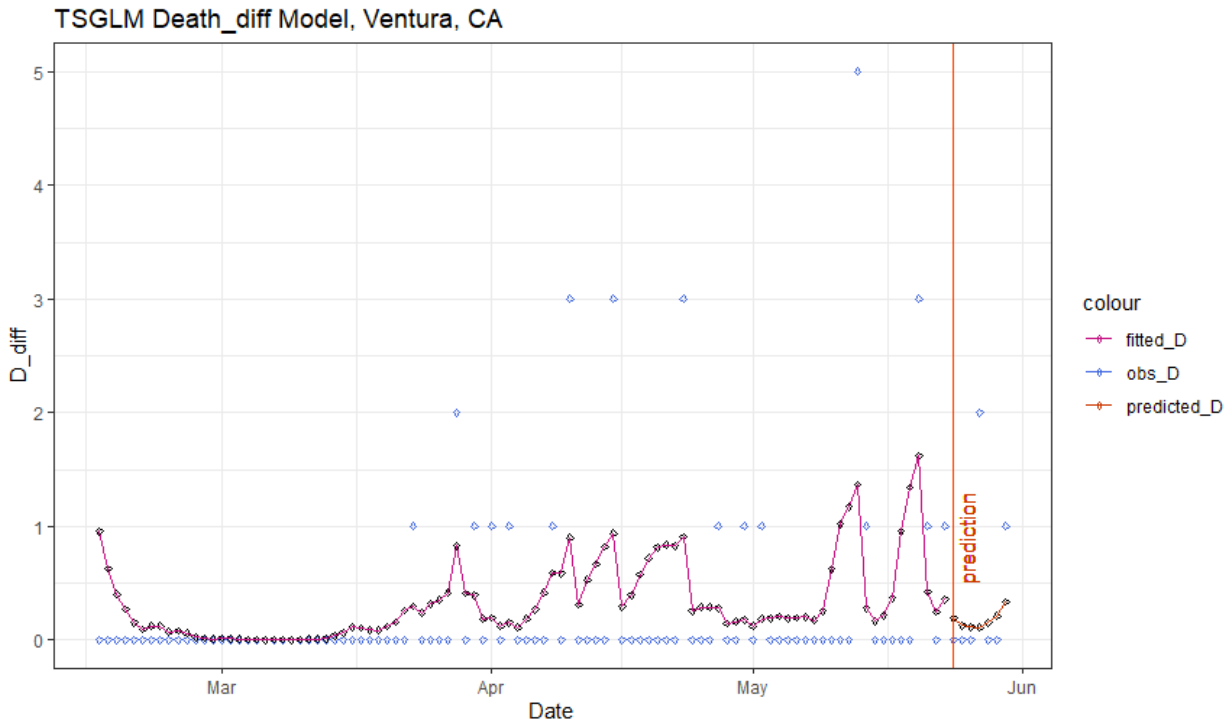

```
Call:
tsglm(ts = SB_tsglm_train_dat$I_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = SB_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -1.0413    0.15836    -1.3517    -0.7309
beta_1                             -0.1763    0.01873    -0.2130    -0.1396
alpha_1                             1.0000    0.01253     0.9754     1.0246
boxcox_aqi                          0.1486    0.04486     0.0607     0.2366
mean_Ozone_ppm                      7.0574    5.64978    -4.0160    18.1307
boxcox_mean_PM2_5_Âµg_m3_LC          0.2579    0.02984     0.1994     0.3163
sqrt_mean_SO2_ppb                   1.5214    0.39342     0.7503     2.2925
boxcox_mean_CO_ppm                  2.6153    0.52460     1.5871     3.6435
parks_percent_change_from_baseline -0.0165    0.00158    -0.0195    -0.0134
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 9
Log-likelihood: -755.7447
AIC: 1529.489
BIC: 1552.754
QIC: 1539.55
```



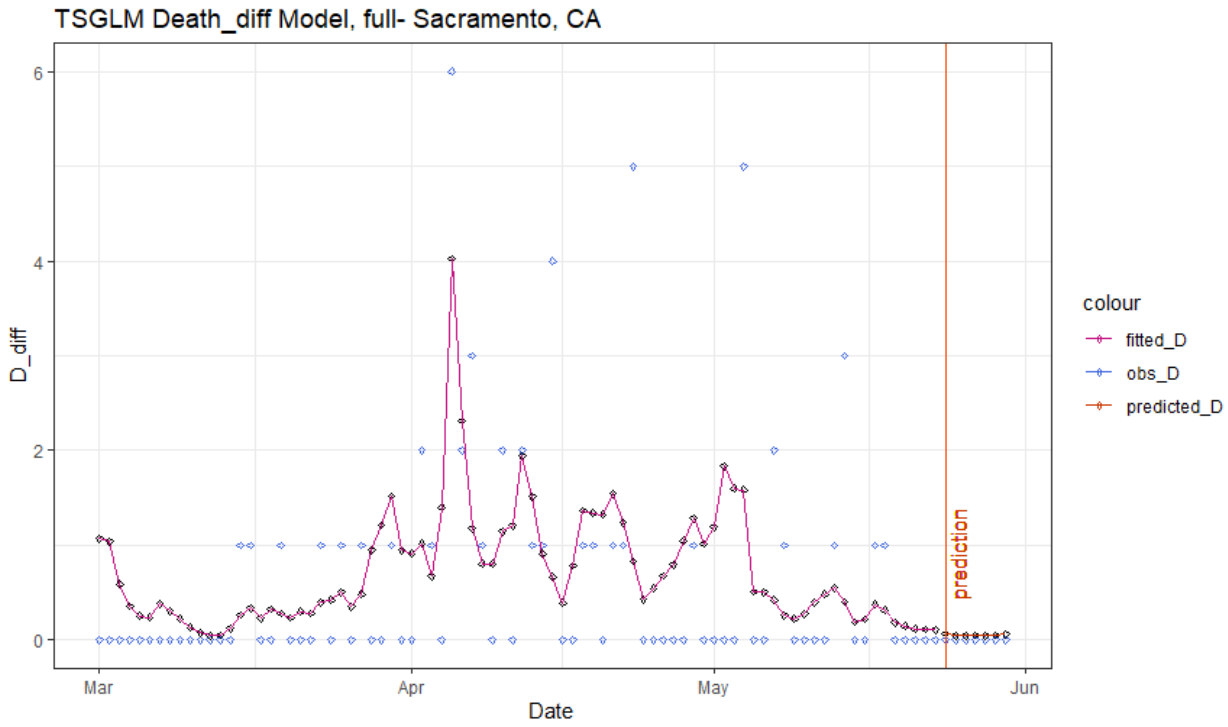

```
Call:
tsglm(ts = SB_tsglm_train_dat$D_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = SB_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                         -6.298     3.2094   -12.5882   -0.00752
beta_1                              -0.349     0.4462    -1.2238    0.52520
alpha_1                              0.416     0.2468    -0.0673    0.90007
boxcox_aqi                          -0.178     0.5321    -1.2207    0.86486
mean_Ozone_ppm                       3.074    80.4861  -154.6764  160.82338
boxcox_mean_PM2_5_Âµg_m3_LC          1.595     0.7649     0.0959    3.09421
sqrt_mean_SO2_ppb                    1.963     6.7657   -11.2977   15.22324
boxcox_mean_CO_ppm                  11.037    10.5273    -9.5959   31.67035
parks_percent_change_from_baseline  -0.028     0.0224    -0.0718    0.01581
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 9
Log-likelihood: -29.75441
AIC: 77.50882
BIC: 100.7735
QIC: 77.1148
```




```
Call:
tsglm(ts = S_tsglm_train_dat$I_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = S_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -0.2289   0.053596   -0.33393  -0.123838
beta_1                             -0.0755   0.014437   -0.10384  -0.047249
alpha_1                             1.0000   0.013781    0.97299   1.027011
boxcox_aqi                          0.1020   0.015647    0.07136   0.132690
mean_Ozone_ppm                      8.2946   2.188566    4.00506  12.584084
boxcox_mean_PM2_5_Âµg_m3_LC         -0.1086   0.040430   -0.18780  -0.029320
parks_percent_change_from_baseline -0.0027   0.000952   -0.00457  -0.000834
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -623.0784
AIC: 1260.157
BIC: 1278.252
QIC: 1264.262
```

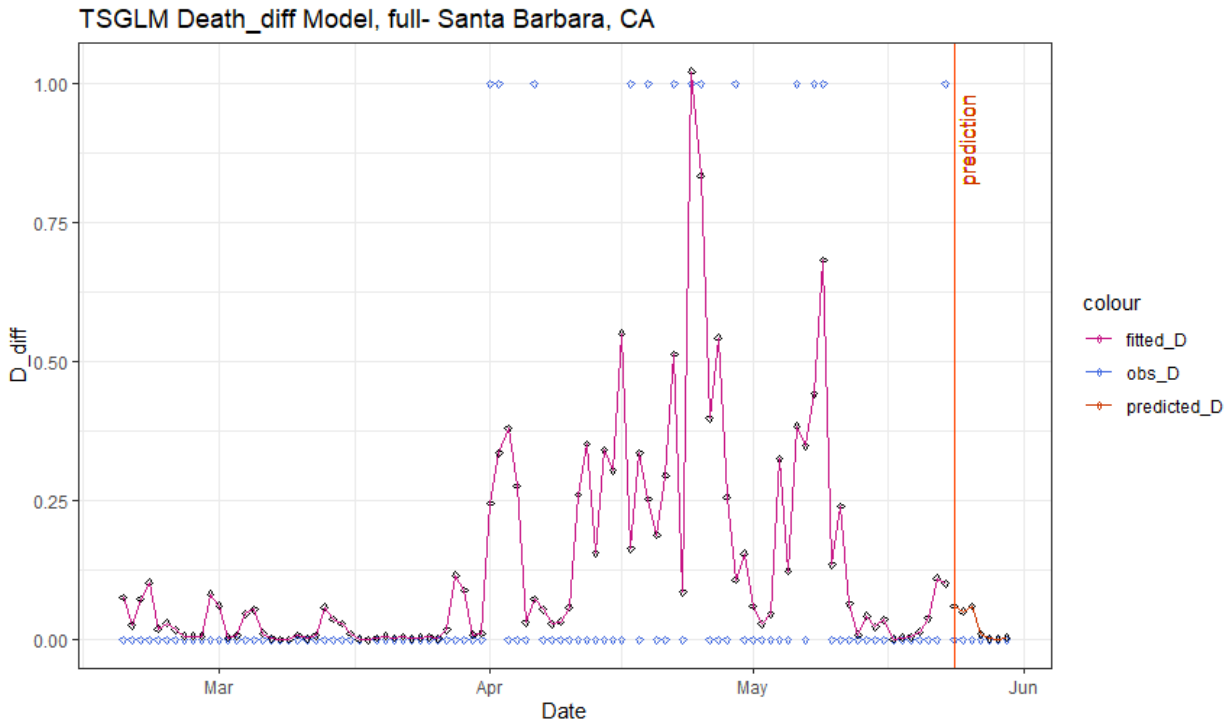



```
Call:
tsglm(ts = S_tsglm_train_dat$D_diff, model = list(past_obs = 1,
    past_mean = 1), xreg = S_tsglm_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                         0.5615    0.42915    -0.2796    1.40262
beta_1                             -0.6740    0.17408    -1.0152   -0.33280
alpha_1                             1.0000    0.04411     0.9135    1.08646
boxcox_aqi                         -0.1113    0.08046    -0.2690    0.04641
mean_Ozone_ppm                      1.6976    6.65351   -11.3431   14.73819
boxcox_mean_PM2_5_Âµg_m3_LC         0.0158    0.16075    -0.2993    0.33087
parks_percent_change_from_baseline -0.0130    0.00237    -0.0177   -0.00838
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 7
Log-likelihood: -85.83533
AIC: 185.6707
BIC: 203.7654
QIC: 192.1582
```

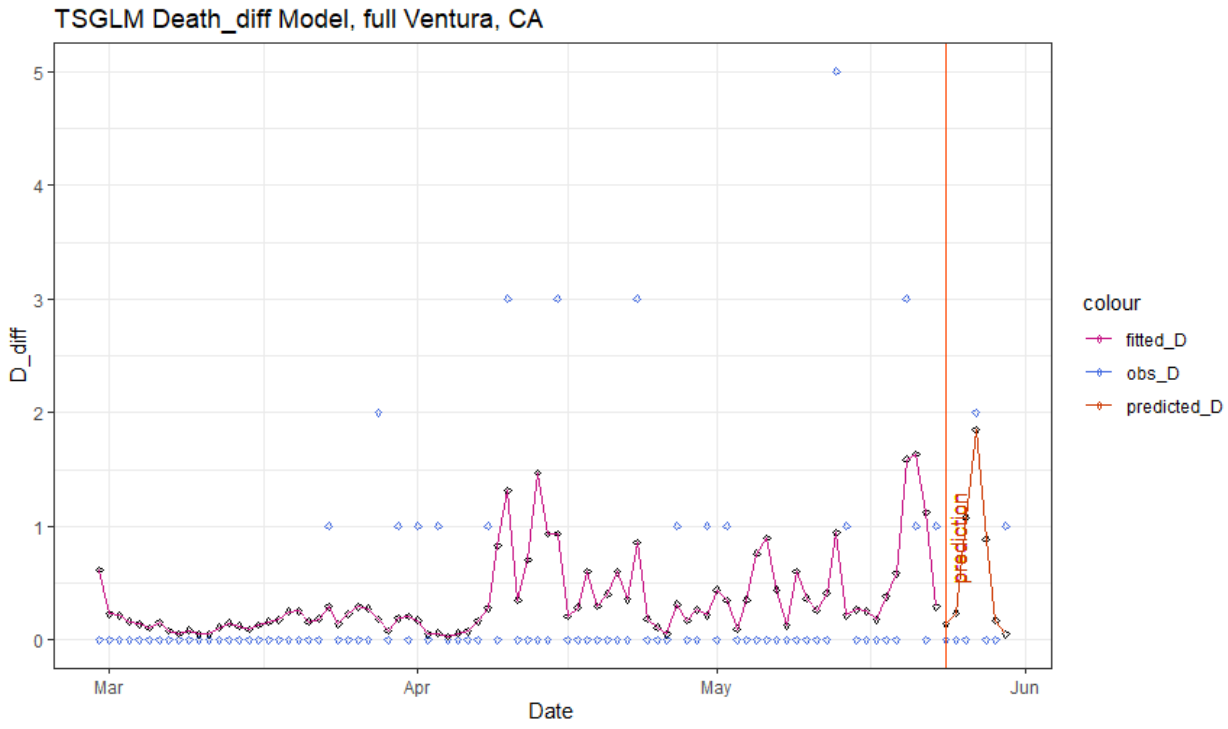



```
Call:
tsglm(ts = S_tot_train$D_diff, model = list(past_obs = 1, past_mean = 1),
    xreg = tmp_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                    Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                           0.2988    0.31629   -0.32108     0.9187
beta_1                               -0.5675    0.15099   -0.86344    -0.2716
alpha_1                               1.0000    0.07400    0.85497     1.1450
I_diff_t_n                            0.0027    0.00347   -0.00411     0.0095
boxcox_aqi                           -0.1102    0.08688   -0.28046     0.0601
mean_Ozone_ppm                        1.8029    7.00902  -11.93456    15.5403
boxcox_mean_PM2_5_Âµg_m3_LC           0.0984    0.27536   -0.44125     0.6381
parks_percent_change_from_baseline   -0.0147    0.00260   -0.01981    -0.0096
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 8
Log-likelihood: -88.46104
AIC: 192.9221
BIC: 212.3686
QIC: 53.90557
```

```
model_compare(S_tot_dat)
X1 X0.188793918768143
 1         0.188793919
 2         0.252250067
 3         0.255106368
 4         0.187548276
 5         0.214285637
 6         0.187446873
 7         0.149605261
 8         0.101025910
 9         0.068406329
10         0.065441505
11         0.059078270
12         0.026473376
13         0.027380330
14         0.002893892
15         0.004658448
```




```
Call:
tsglm(ts = SB_tot_train$D_diff, model = list(past_obs = 1, past_mean = 1),
    xreg = tmp_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                    Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                          -8.4229     3.8380   -15.9453    -0.9005
beta_1                               -0.0642     0.4556    -0.9572     0.8287
alpha_1                               0.0750     0.3092    -0.5310     0.6810
I_diff_t_n                            0.0125     0.0104    -0.0079     0.0329
boxcox_aqi                           -0.0784     0.6957    -1.4419     1.2851
mean_Ozone_ppm                       41.2354    89.9593  -135.0816   217.5525
boxcox_mean_PM2_5_Âµg_m3_LC           1.6729     0.8870    -0.0656     3.4115
sqrt_mean_SO2_ppb                    -3.6334     7.6248   -18.5777    11.3109
boxcox_mean_CO_ppm                   12.0704    14.0057   -15.3802    39.5210
parks_percent_change_from_baseline   -0.0184     0.0291    -0.0755     0.0387
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 10
Log-likelihood: -28.7795
AIC: 77.559
BIC: 103.0978
QIC: 77.52773

model_compare(SB_tot_dat)
X1 X0.0016862988573222
 1          0.001686299
 2          0.004152524
 3          0.001452201
 4          0.007945295
 5          0.003927659
 6          0.008922809
 7          0.009187940
 8          0.012137358
 9          0.009449313
10          0.002448130
11          0.007570677
12          6.957146744
13          0.028792194
14          0.394972391
15          0.640243373
```




```
Call:
tsglm(ts = V_tot_train$D_diff, model = list(past_obs = 1, past_mean = 1),
    xreg = tmp_reg_mat, link = "log", distr = "poisson")

Coefficients:
                                  Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)                        -3.0622    1.58721    -6.1730    0.04869
beta_1                             -0.7398    0.26066    -1.2507   -0.22892
alpha_1                             0.7187    0.13759     0.4491    0.98842
I_diff_t_n                          0.0797    0.02266     0.0353    0.12412
log_aqi                             1.5669    1.16548    -0.7174    3.85121
boxcox_mean_Ozone_ppm              34.1514   14.21933     6.2820   62.02078
boxcox_mean_PM2_5_Âµg_m3_LC        -0.1627    0.12138    -0.4006    0.07521
parks_percent_change_from_baseline -0.0121    0.00863    -0.0290    0.00481
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: poisson
Number of coefficients: 8
Log-likelihood: -60.16595
AIC: 136.3319
BIC: 155.8731
QIC: 137.4416
```

```
model_compare(V_tot_dat)
X1 X0.529175594025805
 1           0.5291756
 2           0.5681251
 3           0.5376217
 4           0.5181157
 5           0.5493178
 6           0.6312362
 7           0.5309314
 8           0.5264215
 9           0.5310032
10           0.6324898
11           0.5756301
12           0.5339643
13           0.4257723
14           0.5394722
15           0.5921500
```

REFERENCES

[1] "AirData Website File Download Page." *EPA*, Environmental Protection Agency, aqs.epa.gov/aqsweb/airdata/download_files.html.

[2] "AirData Website File Download Page." *EPA*, Environmental Protection Agency, aqs.epa.gov/aqsweb/documents/data_api.html.

[3] Brandt, Patrick T., and John T. Williams. 2001. A Linear Poisson Autoregressive Model: The Poisson AR(p) Model. Political Analysis 9(2): 164-84.

[4] "COVID-19 Community Mobility Reports." *Google*, https://www.google.com/covid19/mobility/

[5] CSSEGISandData. "CSSEGISandData/COVID-19." *GitHub*, github.com/CSSEGISandData/COVID-19.

[6] "Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 4 June 2020, www.cdc.gov/mmwr/volumes/69/wr/mm6922e1.htm.

[7] "Health Effects of Ozone Pollution." *EPA*, Environmental Protection Agency, 10 Sept. 2020, www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution.

[8] Liboschik T, Fokianos K, Fried R (2017). "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models." *Journal of Statistical Software*, doi.org/10.18637/jss.v082.i05.

[9] Moritz, Steffen, and Thomas Bartz-Beielstein. "ImputeTS: Time Series Missing Value Imputation in R." *The R Journal*, doi.org/10.32614/RJ-2017-009.

[10] Staff, byEdhat. "Public Health Adds 28 Missed COVID-19 Deaths Due to Data Error." *Edhat*, 1 Aug. 2020, www.edhat.com/news/public-health-adds-28-missed-covid-19-deaths-due-to-data-error.

[11] Wood SN (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society (B)*, **73**(1), 3-36.