THE DEVELOPMENT OF **A PREDICTIVE MODEL OF COVID-19 MORTALITY** AND ITS RELATIONSHIP WITH ENVIRONMENTAL POLLUTANTS AND OUTDOOR MOBILITY


(Wally Skalij / Los Angeles Times)

Danielle Heymann, Graduate Student

Research Advisor: Dr. GuanNan Wang

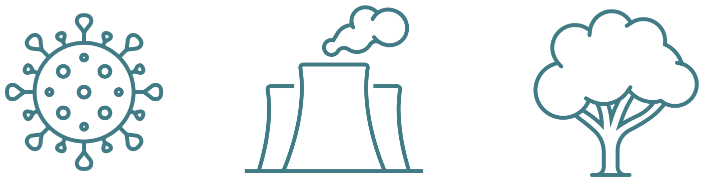
Department of Computer Science

Computational Operations Research (COR)

College of William and Mary

OBJECTIVE

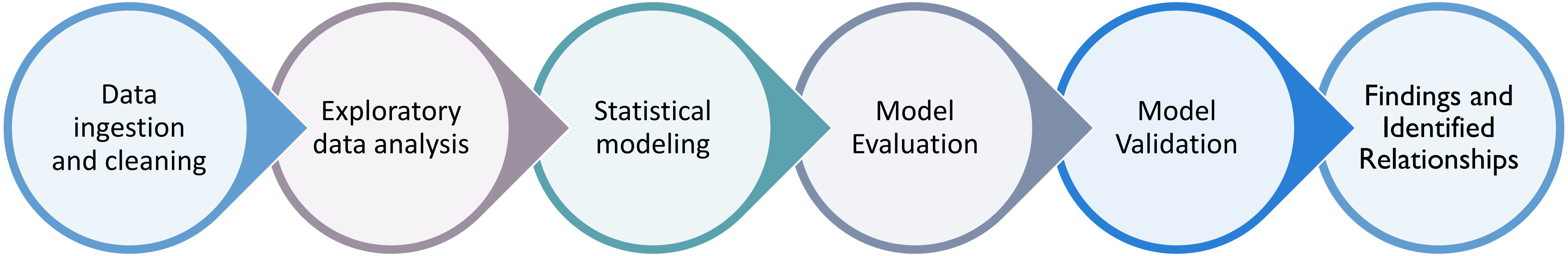
- Model the death and infection counts during the COVID-19 pandemic
  - Data sources
    - Air quality index (AQI), CO, SO2, NO2, PM10, PM2.5, Ozone **(EPA)**
    - Outdoor Mobility Data **(Google)**
    - COVID-19 Data **(JHU)**



- Area of Interest for Research
  - California counties with early surges of COVID-19 cases:
    - Santa Barbara, Sacramento, Ventura
  - Matrix of models considered and created:
    - GLM: General Linear Model
    - TSGLM: Time Series following Generalized Linear Model

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

METHODOLOGY



DATA INGESTION, CLEANING, AND EDA

DATA GATHERING

- JHU COVID-19 Infection and Death Counts
  - Used dataset maintained by Dr. GuanNan Wang
  - Additional remedial measure:
    - Resolved Santa Barbara false jump in death count
- Google Mobility Data
  - Extracted outdoor park visits mobility data
- EPA Pollutants data
  - Sources:
    - Pre-extracted aggregated data for each pollutant by quarter
    - Real time raw data from queries to Air Quality System (AQS) API
      - Json files imported to python

DATA CLEANING + REMEDIAL MEASURES

Key Reasons for Data Cleaning:

- Formatting and Naming Discrepancies
  - Merging API and pre-extracted EPA data
- COVID counts represented as difference (delta) in cumulative counts per day
- Date Conventions
  - Time Series based on **days**
- Missing Observations
  - Kalman Smoothing
- Outlier Detection and Removal
  - Conventional method: more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3

Techniques and Remedial Measures:

- Multiple Daily Observations
  - Grouped by specific county and pollutant combination
  - Averaged to one mean observation
- Missing Observations
  - Get/fit a State Space Model via ARMIA model
  - Estimate missing values by Kalman smoothing
  - R package: imputeTS (Moritz, Steffen, and Bartz-Beielstein, Thomas)

MODELING: INPUT

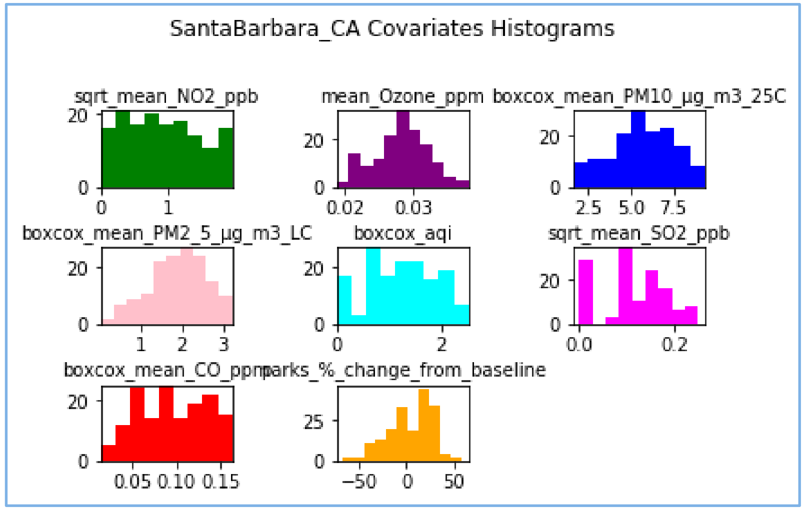
Models created	Response, Y	Autoregressive component
GLM with lag, TSGLM	Death_diff	AR(1), AR(2)
GLM with lag, TSGLM	Infections_diff	AR(1), AR(2)
TSGLM	Death_diff	AR(1), delayed infections

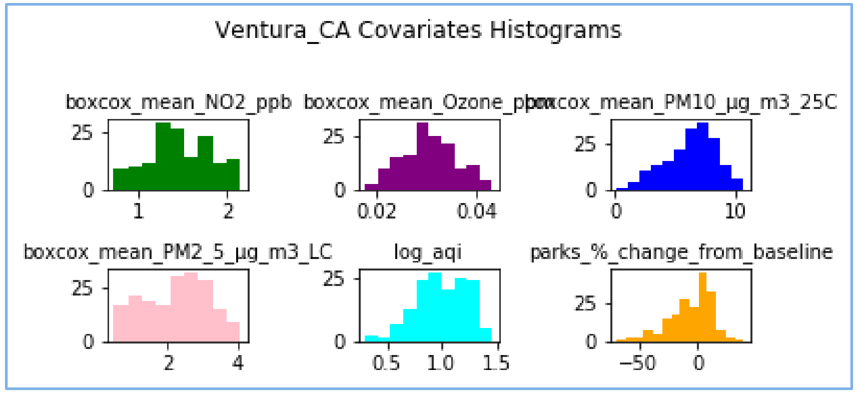
County	Source	Covariates, Xi
Santa Barbara	EPA pollutants	NO2_ppb AQI Ozone PM10 PM2.5 SO2 CO
	Google	Parks Percent change
	JHU *varies for model	Death_diff lagged Infection_diff lagged Infection_diff lagged optimally
Sacramento	EPA pollutants	NO2_ppb AQI Ozone PM10 PM2.5
	Google	Parks Percent change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

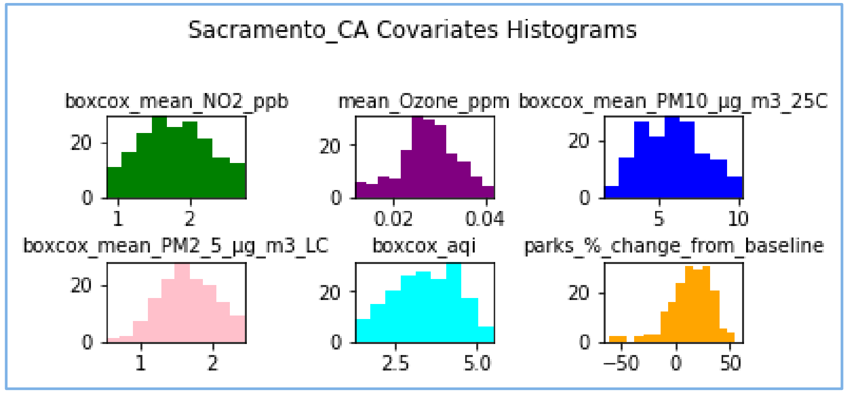
R Libraries Utilized:

- mgcv (for glm, general linear model)
- tscount (for tsglm, time series general linear model)

TRANSFORMATIONS OF COVARIATES + HISTOGRAMS







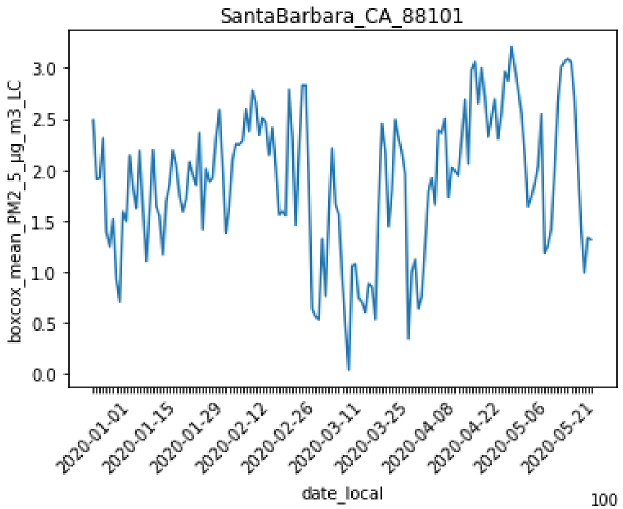
FINAL DATASET CREATION CONSIDERATIONS

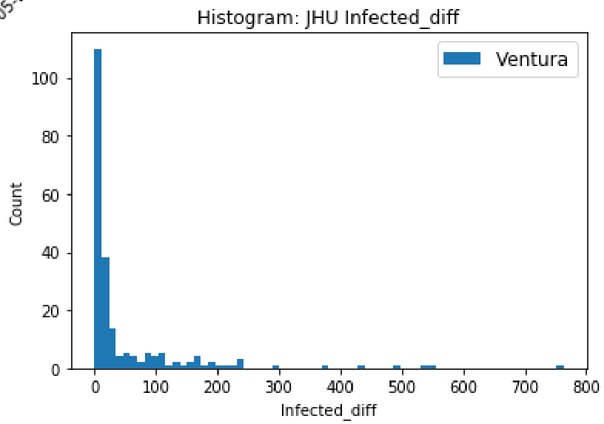
- Time series plots generated for each covariate and response variable for anomaly detection

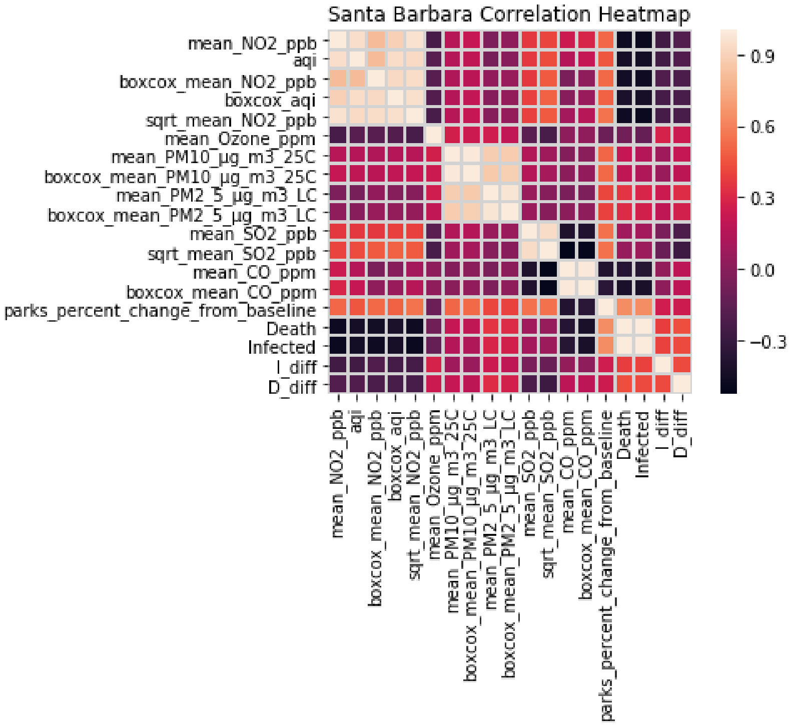
- Histograms generated for all variables as additional quality diagnostic

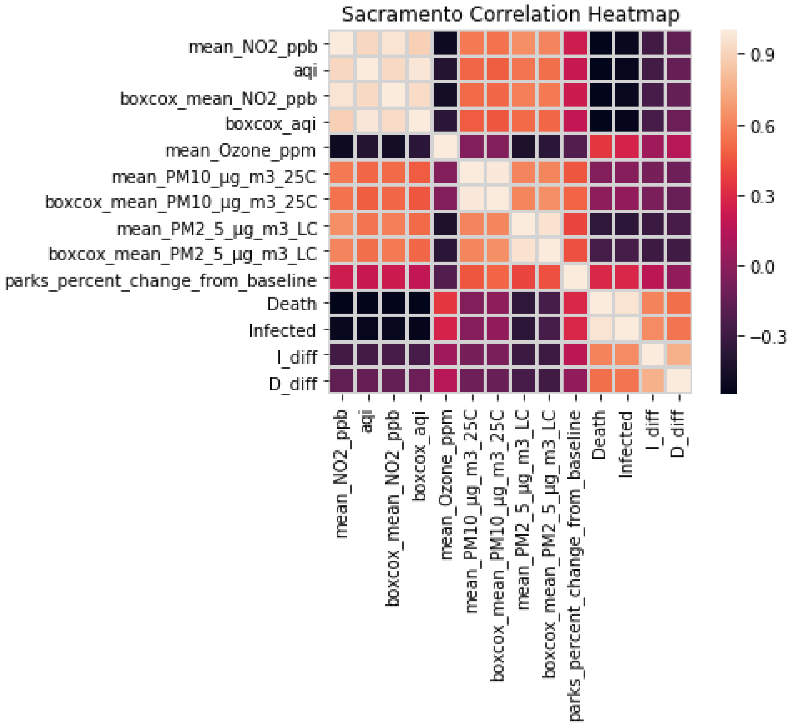
- Correlation matrix and heatmap generated for each county

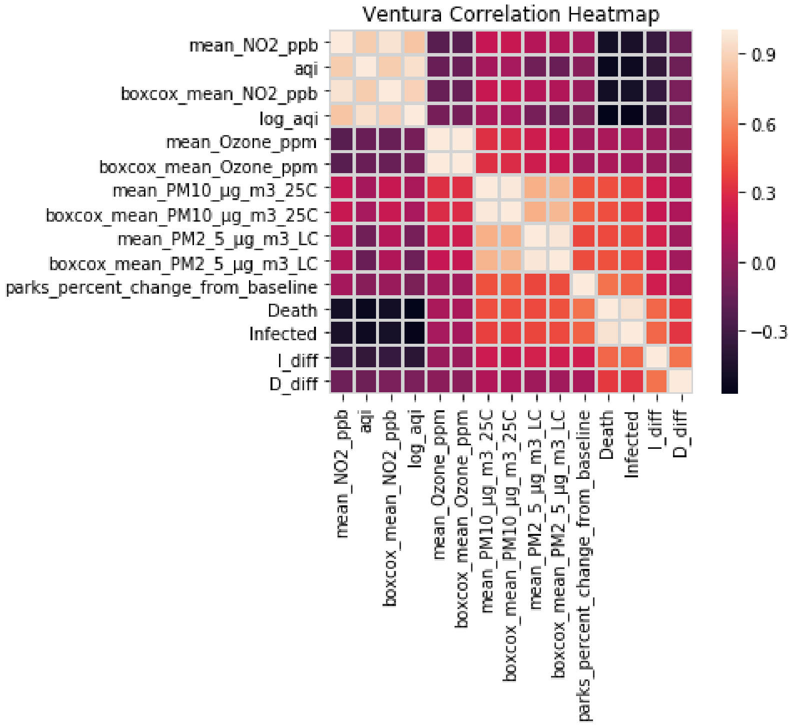
- Some covariates left out to avoid multicollinearity:
  - kept AQI; removed NO_2
  - kept PM2.5; removed PM10





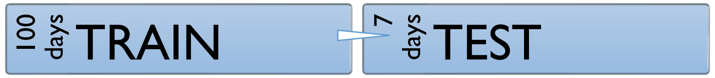






FINAL DATASET CREATION CONSIDERATIONS: CORRELATION HEATMAPS

MODELING, EVALUATION, VALIDATION



BASELINE GLM MODELS

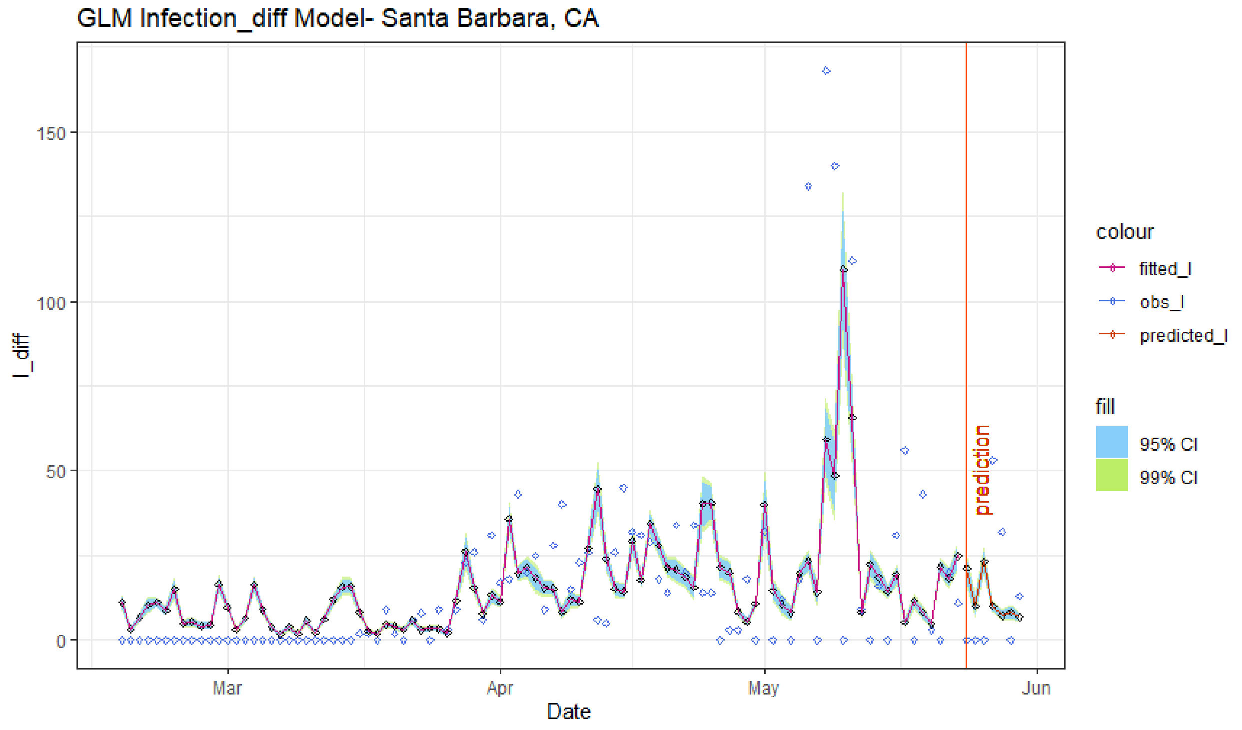
- General Linear Model
- mgcv R package (Wood)
- Poisson Family
  - Response variable in terms of "counts"
- Autoregressive component
  - Infection_diff with AR(2)
  - Death_diff with AR(2)

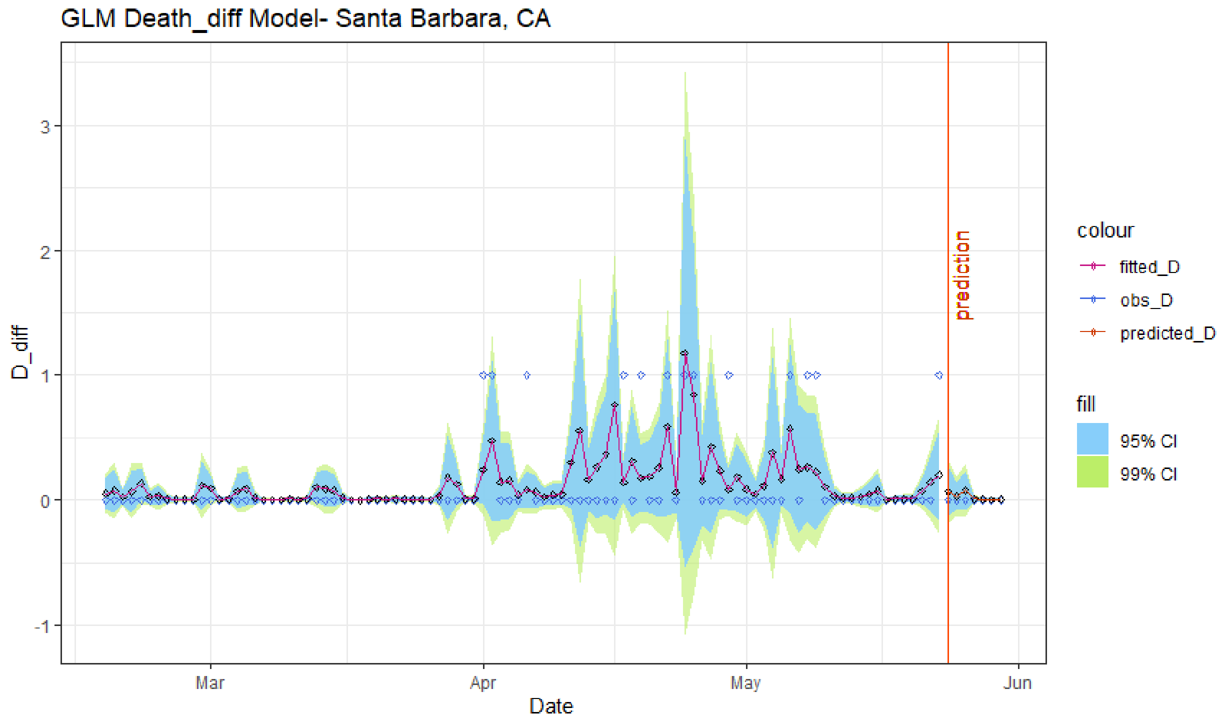
Poisson AR(2):
$E[y_t|X_t, y_{t-1}, y_{t-2}] = \exp(\delta_0 + X_t\delta_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t)$

SANTA BARBARA COUNTY

Baseline GLM Models

MSPE	
I_diff	D_diff
520.2754	0.0017114

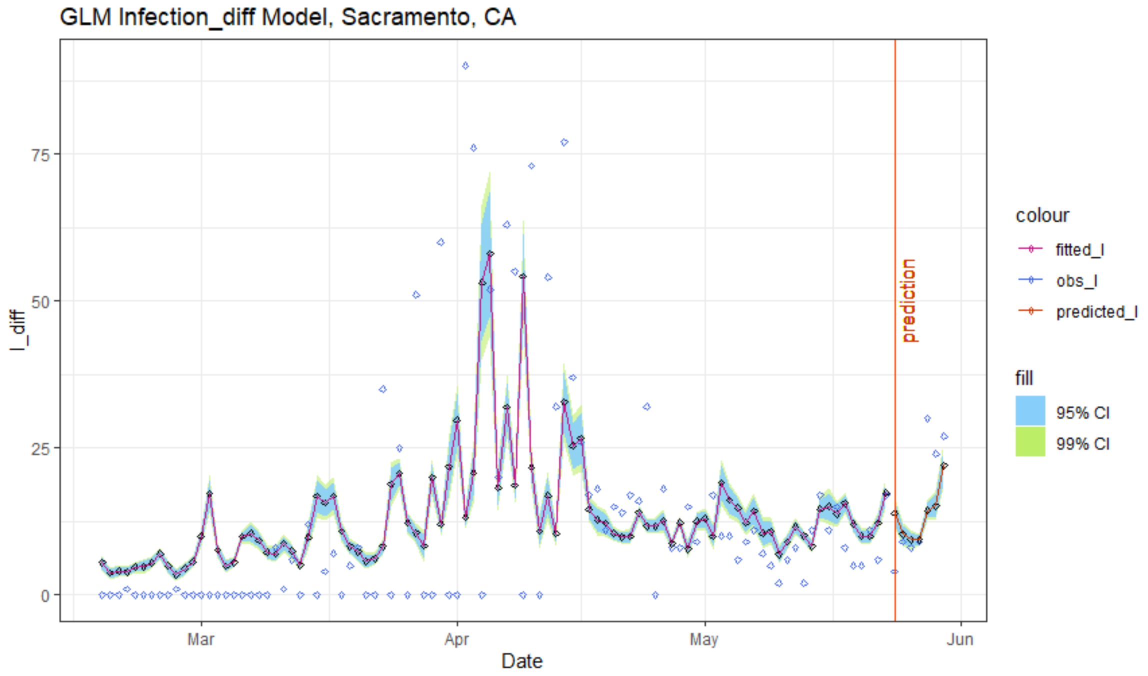


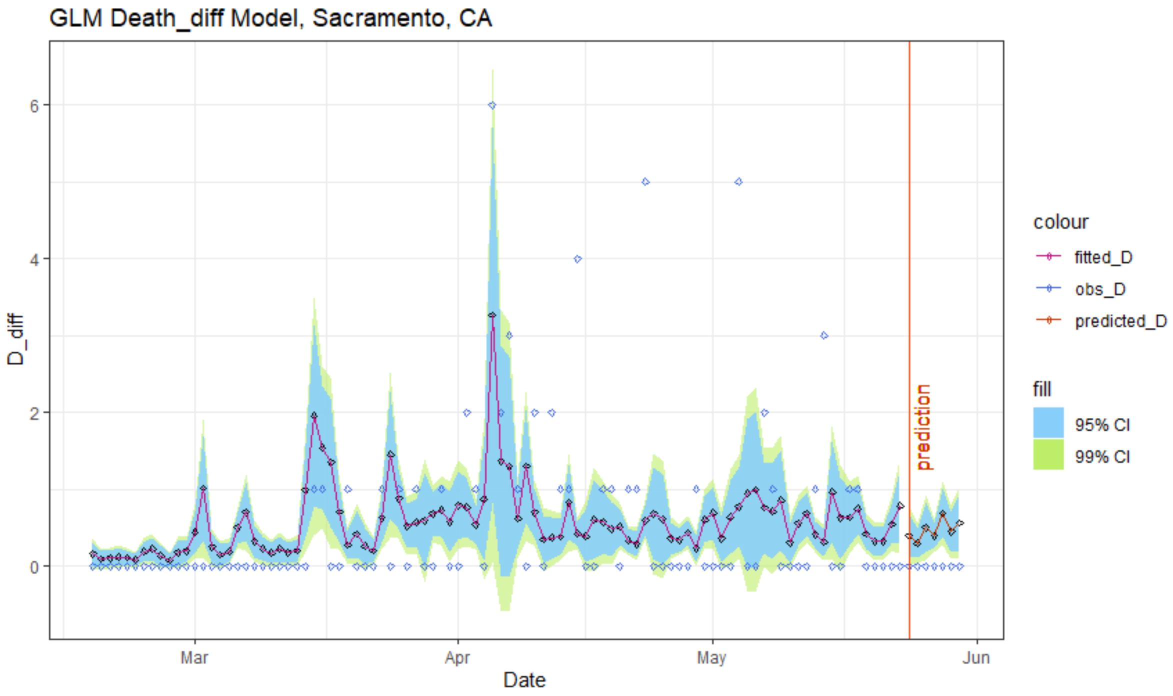


SACRAMENTO COUNTY

Baseline GLM Models

MSPE	
I_diff	D_diff
64.32438	0.2364802

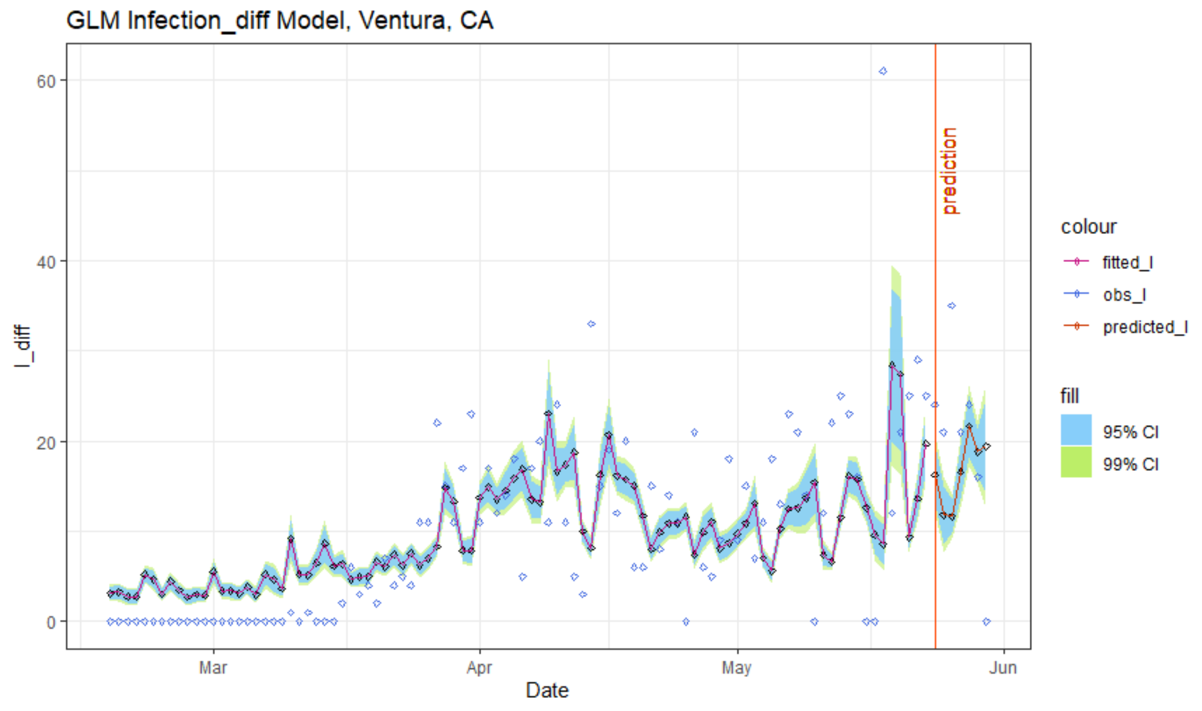


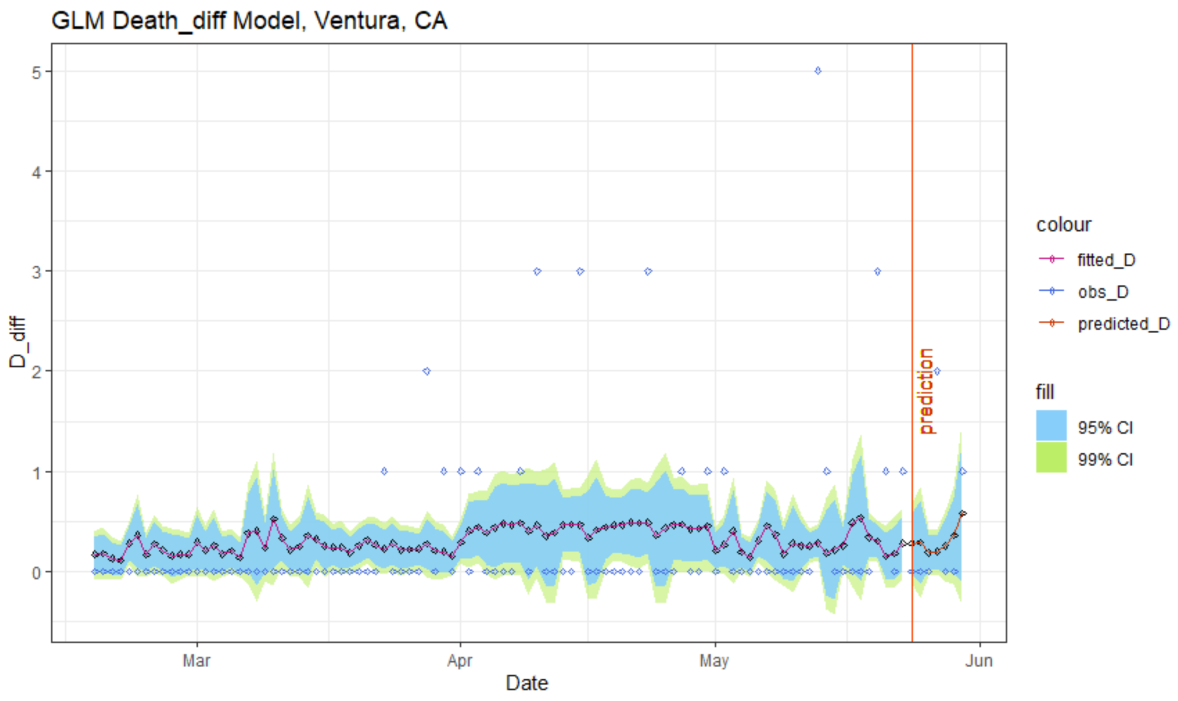


VENTURA COUNTY

Baseline GLM Models

MSPE	
I_diff	D_diff
156.8412	0.5457606





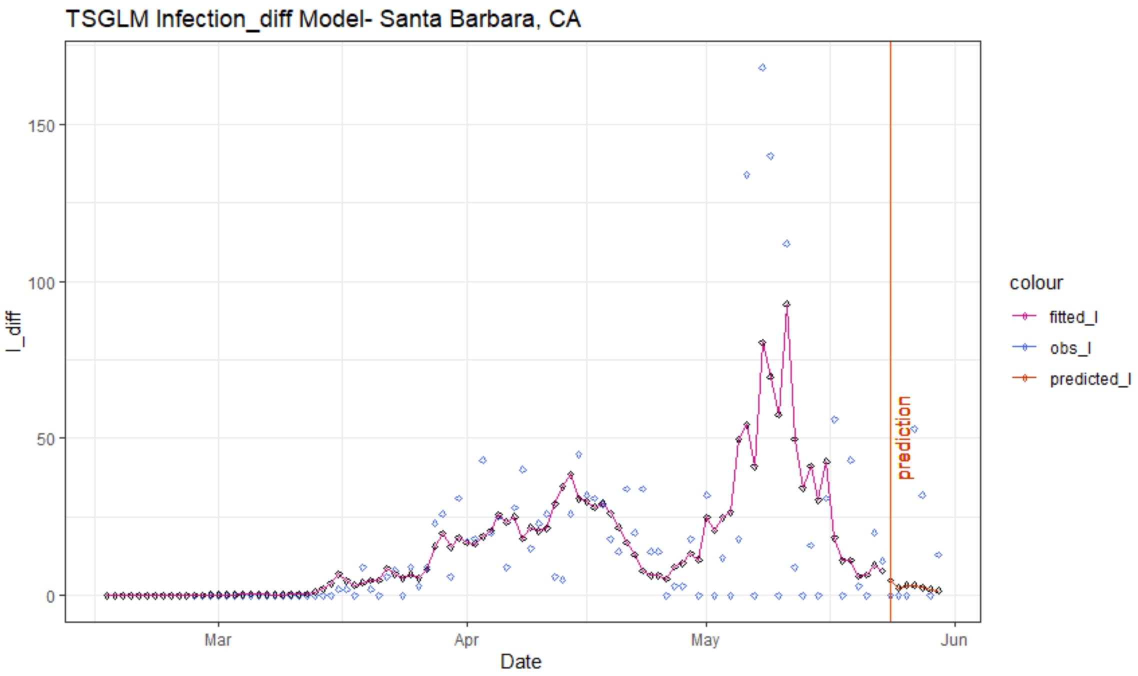
TSGLM MODELS

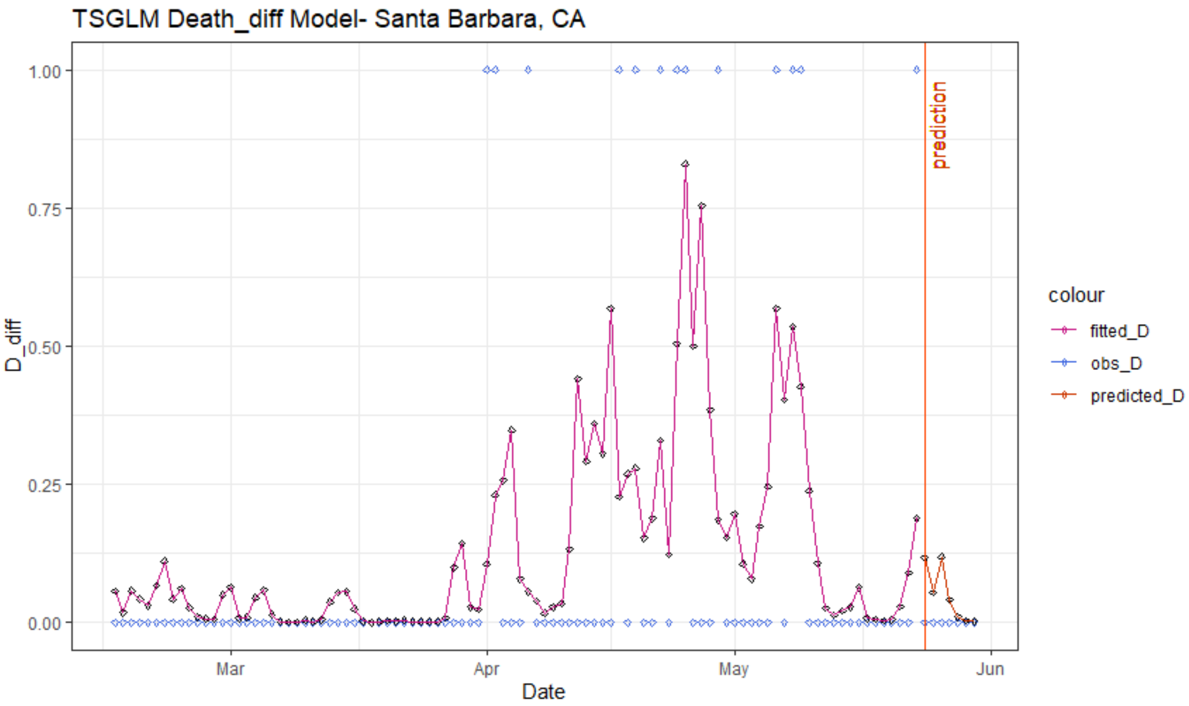
- Time Series following Generalized Linear Model
- tscount R package (Liboschik, Fokianos, Fried)
- Poisson Family with log link
  - Response variable in terms of "counts"
- Autoregressive (AR) and Moving Average components (MA)
  - AR(1), MA(1)

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

SANTA BARBARA COUNTY

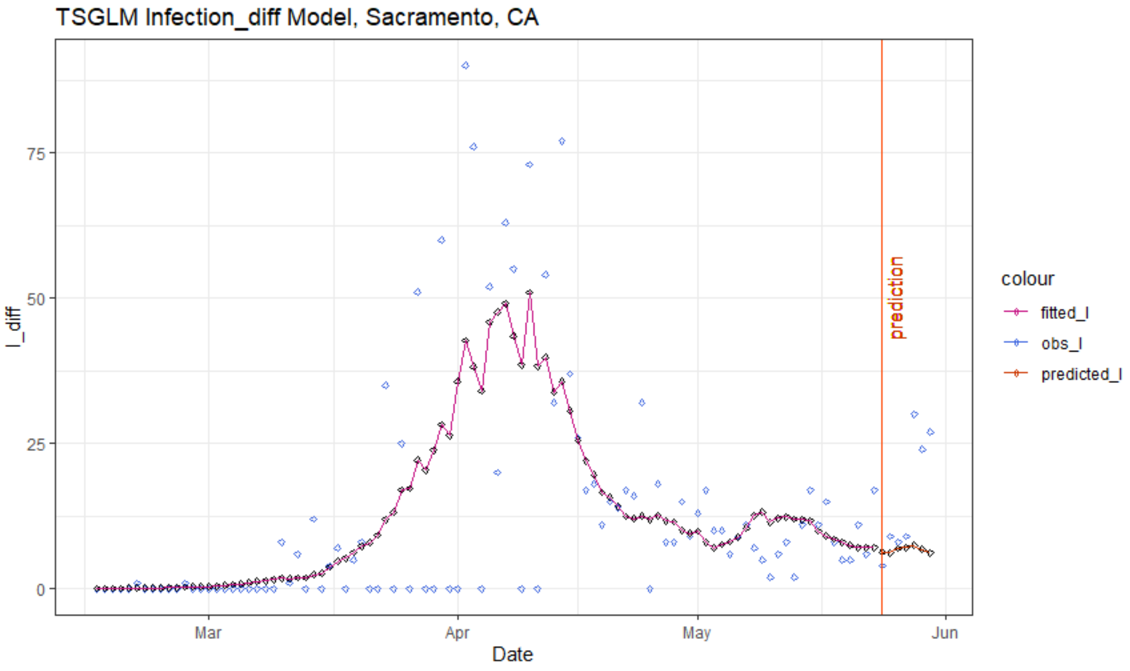
TSGLM Models

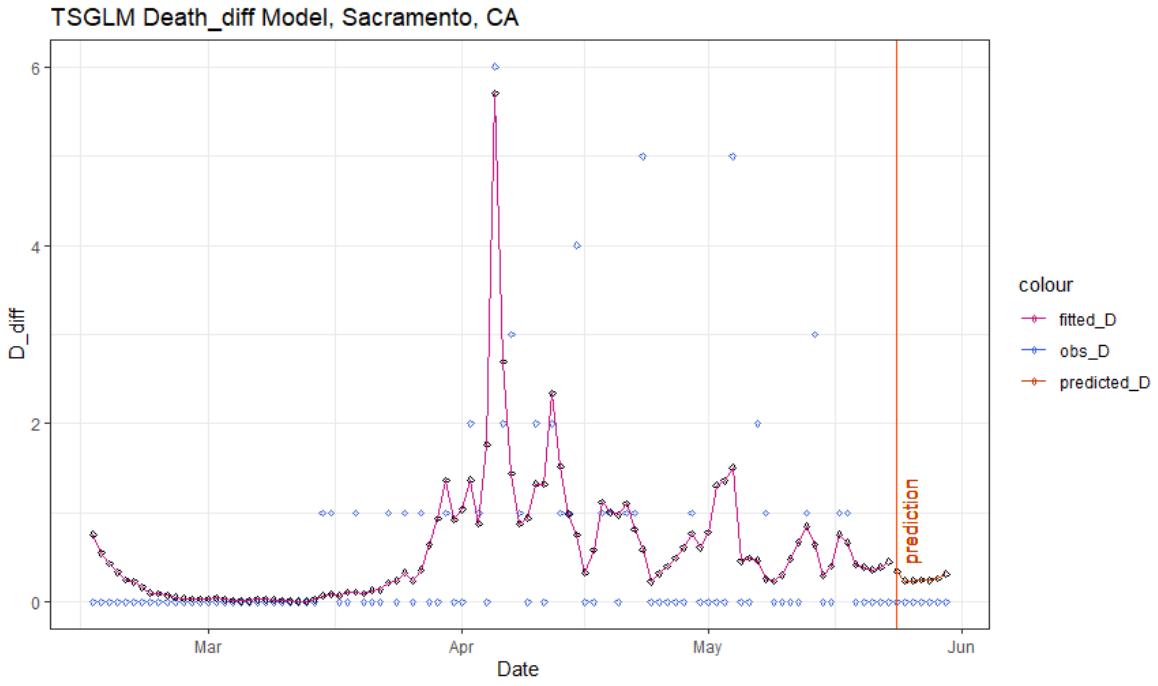




SACRAMENTO COUNTY

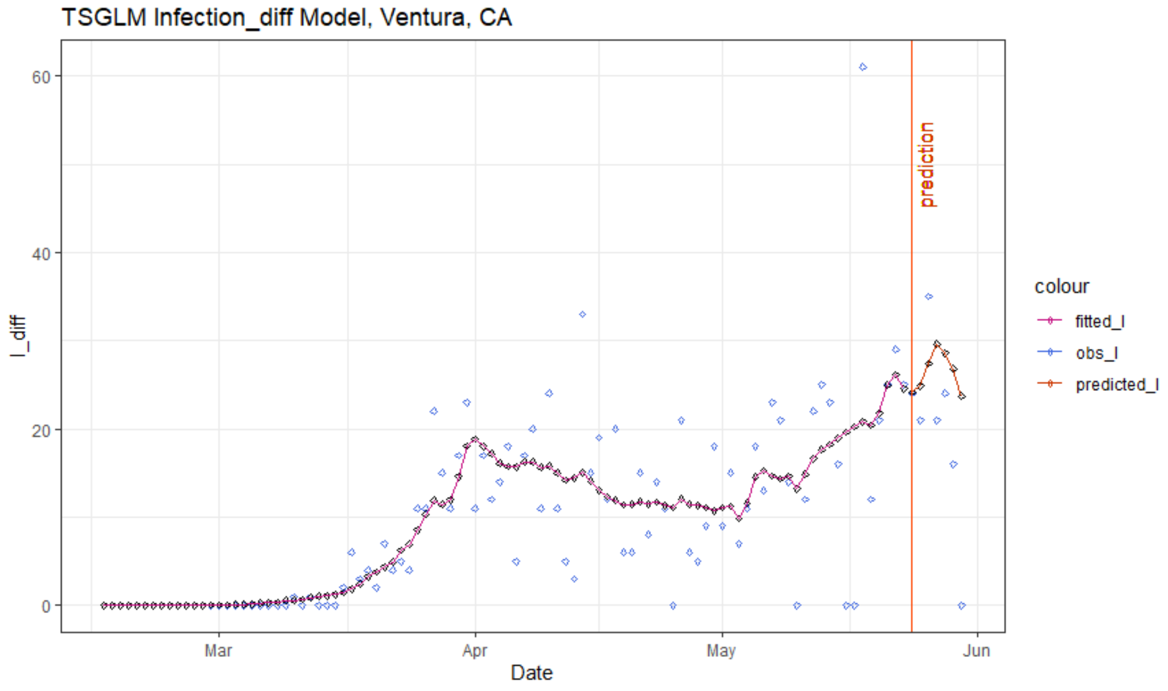
TSGLM Models

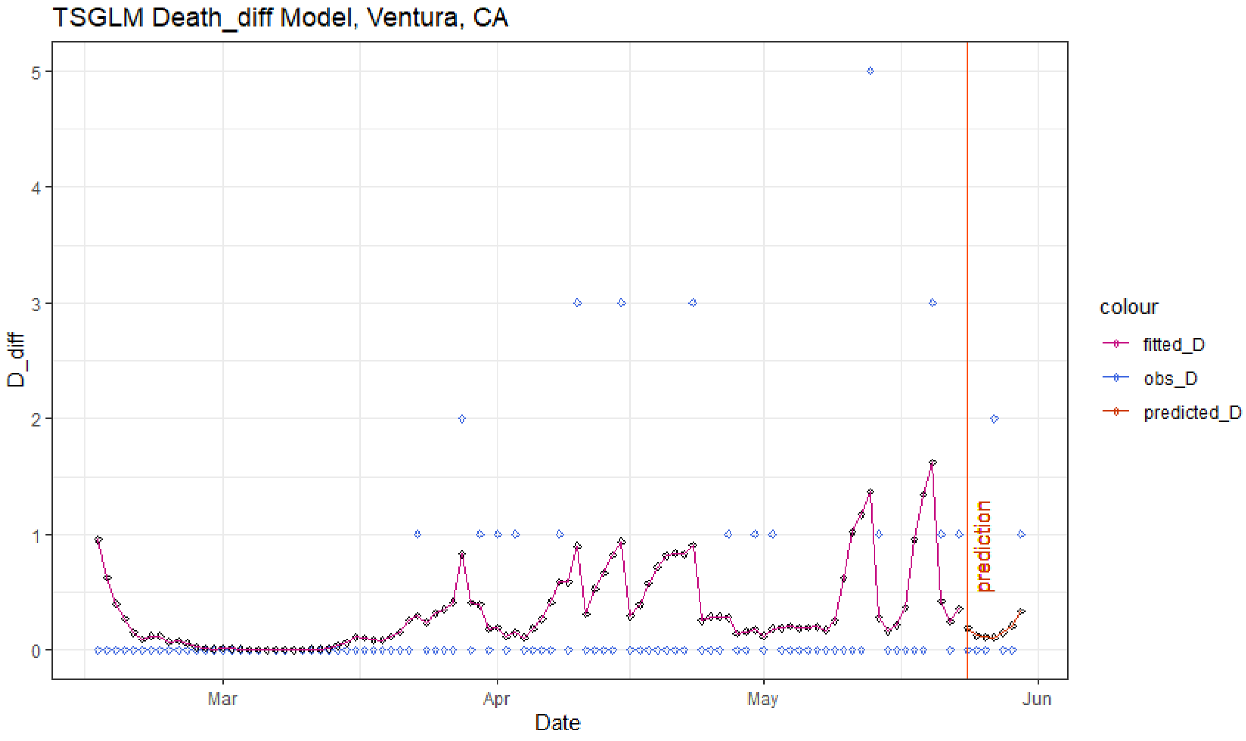




VENTURA COUNTY

TSGLM Models





UPDATED TSGLM MODEL

- Adapted from the previous Death_diff response model
- Incorporate Infection_diff as a covariate with delay
  - 15 models generated for each county with Infection_diff covariate taking delay value, d, in [1,15]
    - $x_{delay\_infection} = X_{t-d}$
  - Optimal delay time determined via selecting model with minimum MSPE

Poisson AR(1), MA(1):

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

SANTA BARBARA COUNTY

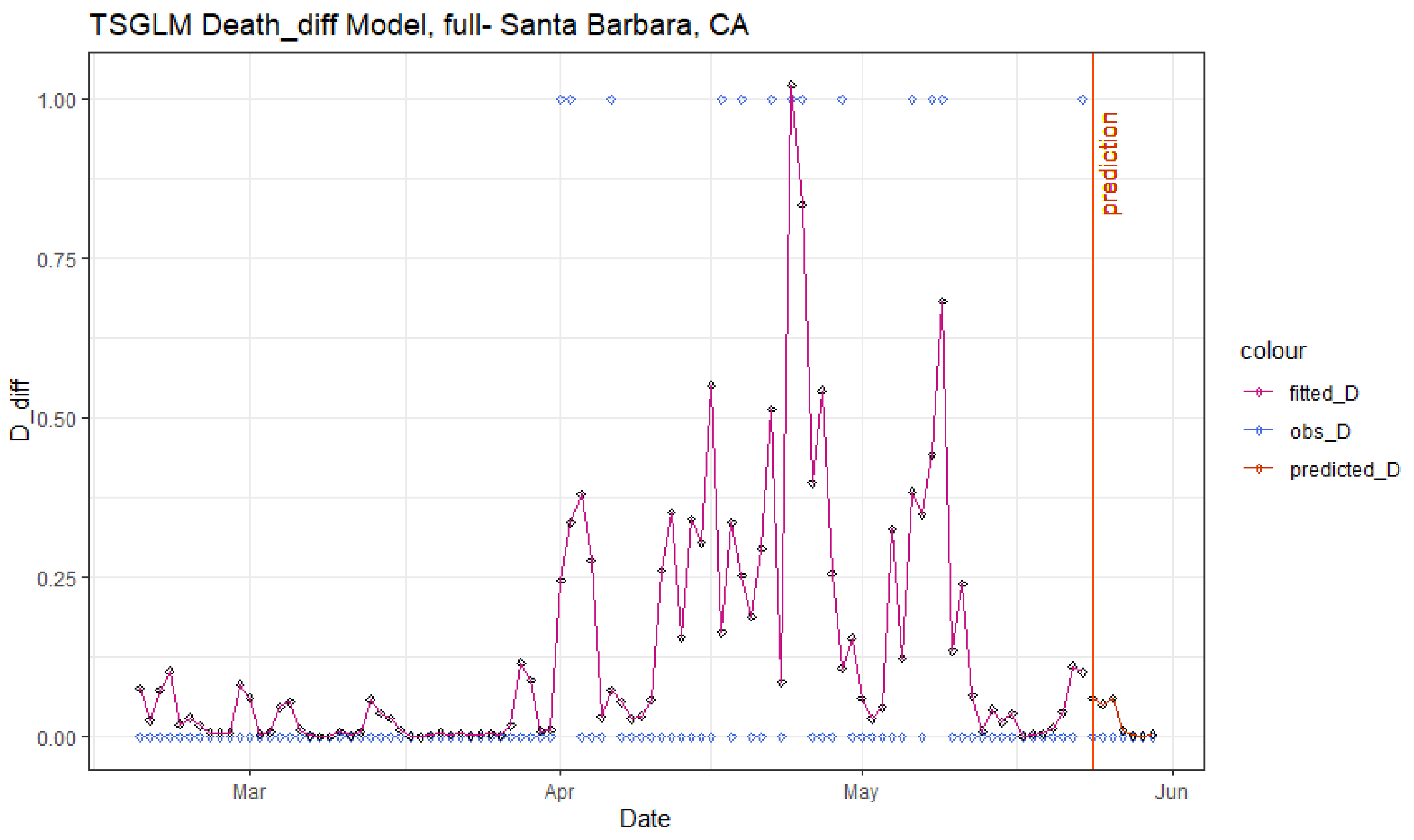
TSGLM Models (updated)

```
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
```

* No significant coefficients



SACRAMENTO COUNTY

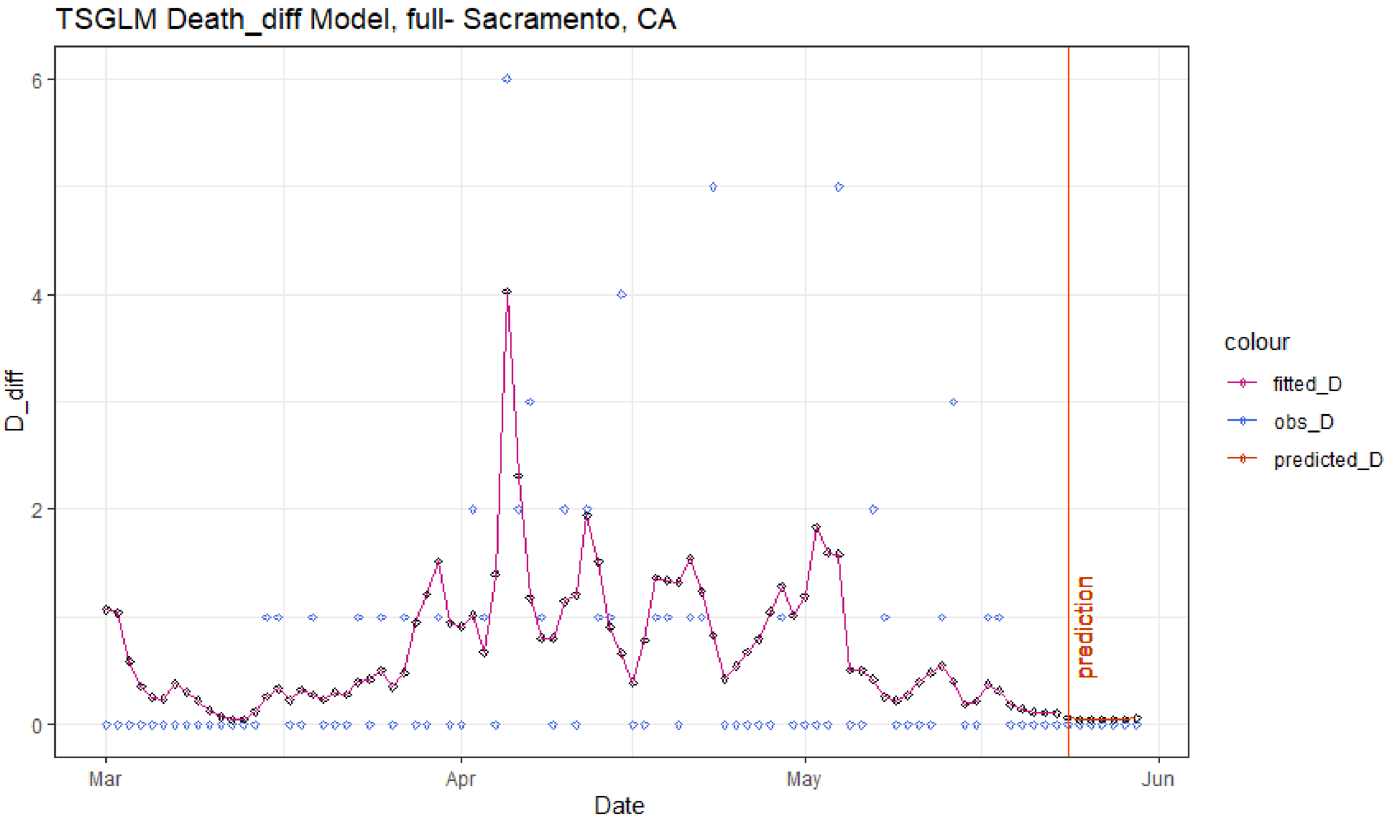
TSGLM Models (updated)

```
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
```

* Significant coefficients: parks % change from baseline



COMPUTER SCIENCE
COLLEGE OF WILLIAM AND MARY

VENTURA COUNTY

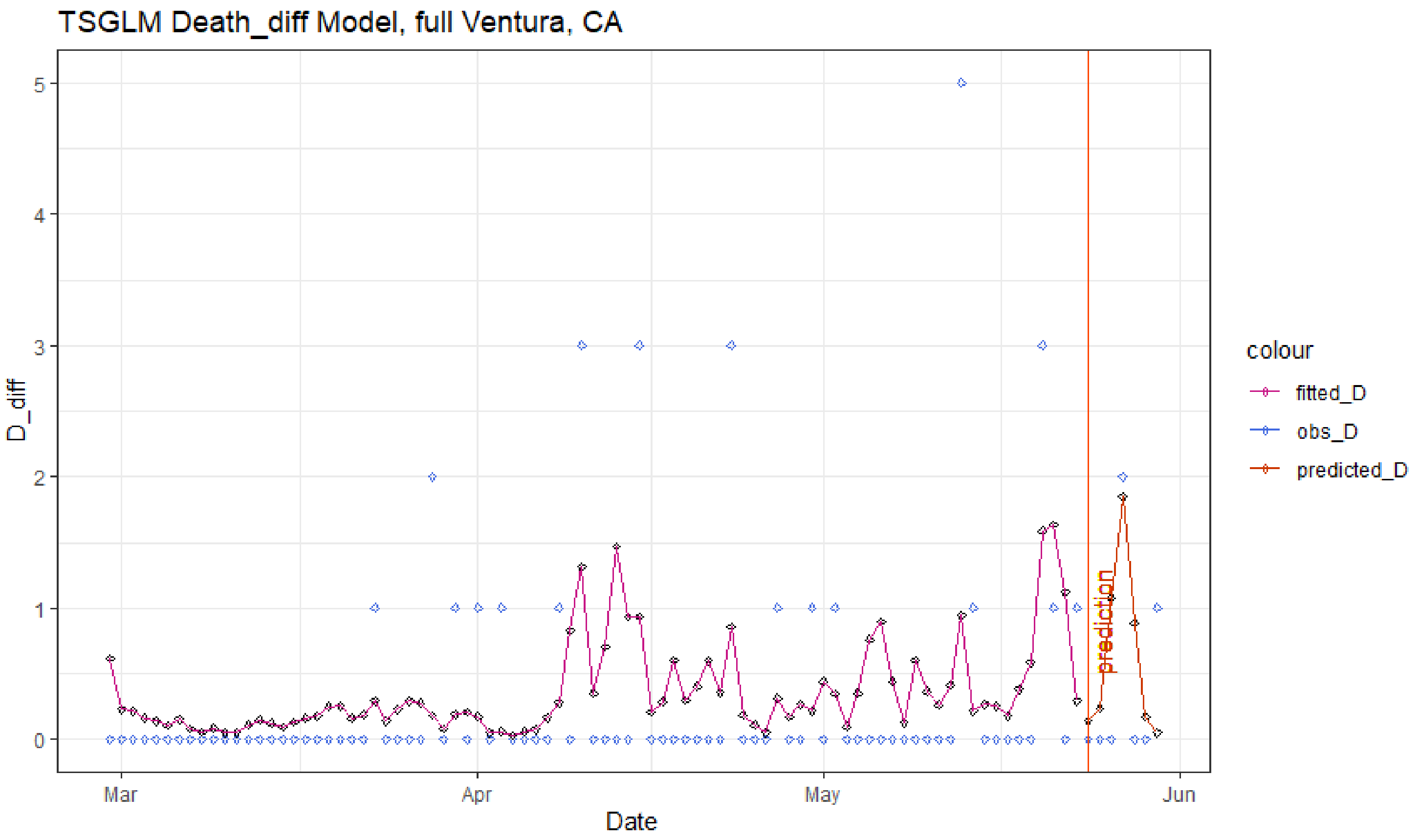
TSGLM Models (updated)

```
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: Ozone, Infections (t -13 days)



MSPE COMPARISON RESULTS

	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

REFERENCES AND SOURCES

- https://personal.utdallas.edu/~pbrandt/pests/parp.pdf
- https://www.cdc.gov/mmwr/volumes/69/wr/mm6922e1.htm
- https://www.google.com/covid19/mobility/
- https://aqs.epa.gov/aqsweb/airdata/download_files.html
- https://www.airnow.gov/?city=Washington&state=DC&country=USA
- https://aqs.epa.gov/aqsweb/documents/data_api.html
- https://github.com/CSSEGISandData/COVID-19
- 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.

THANK YOU