A Few Strategies for the Statistical Modeling of the COVID-19 Pandemic Data

Seth Arreola, Gwendolyn Lind, Caleb Peña
Research Advisor: Dr. Sam Behseta

12/11/2020

Department of Mathematics

Main Research Objectives

In this study we utilize machine learning, statistical models, including time-series models to address two main goals for this project:

- Predict and track daily COVID cases by county
- Classify the counties based on their COVID data

Data Resources

Before we begin our research we need to find some data that best fit with our objectives.

- Mobility data
  - Google https://www.google.com/covid19/mobility/
  - Apple https://covid19.apple.com/mobility
- COVID infection data
  - USA.Facts.org https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
- From these sources we created and updated two primary data sets
  - Two most populated counties in each state
  - Seven southern California counties

Mobility Data: Google

Google data is available by Community Mobility Reports

- The reports chart show movement trends over time by geography, across different categories of places.
- Collects aggregated data and focuses more on where people spend their time.
- Interpret the mobility data by creating "baseline days" from a 5-week period through Jan 3 – Feb 6, 2020

Mobility Data: Google

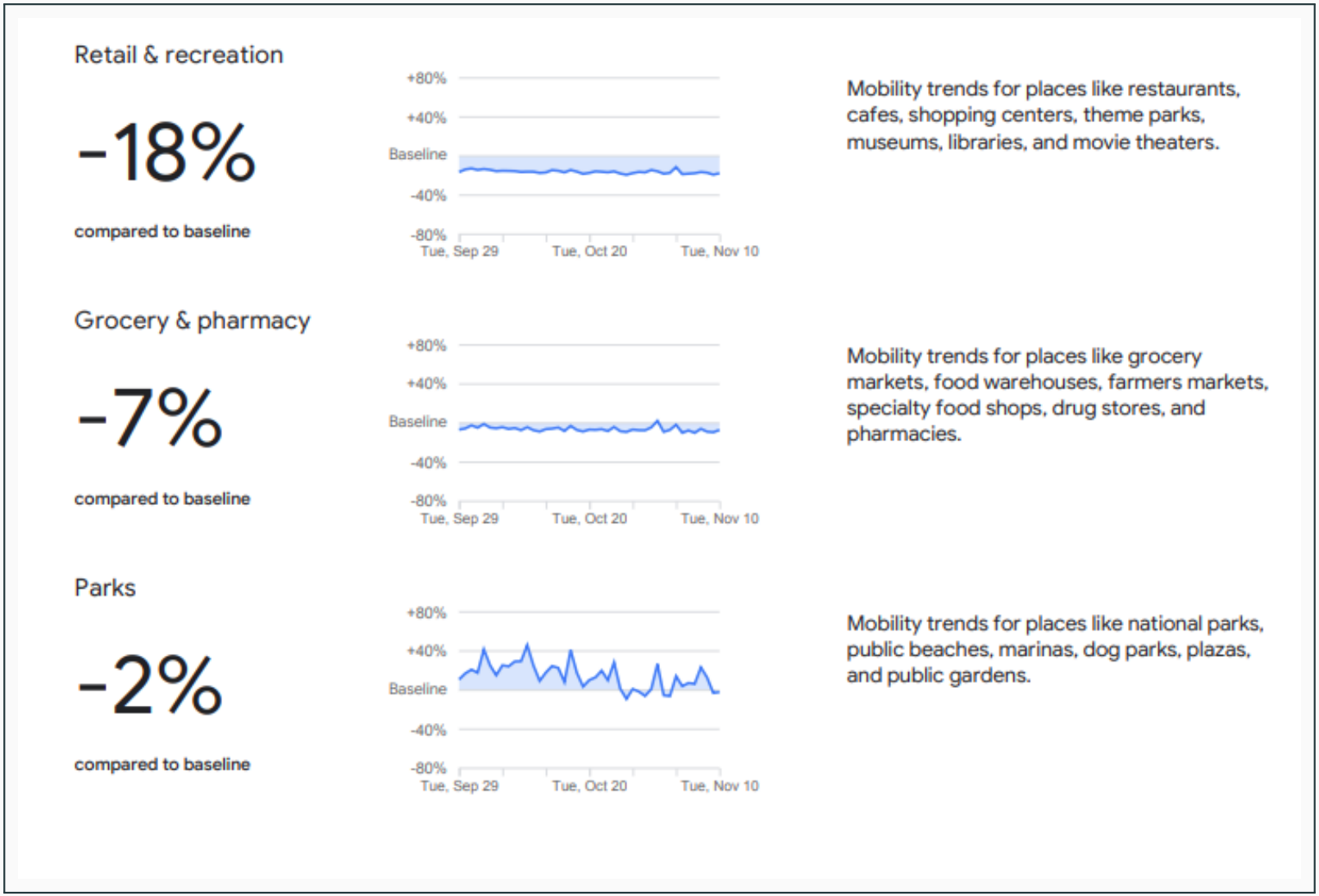


Figure 1: How retail, grocery, and park trends on Nov 10th compare to baseline (U.S.)

Mobility Data: Apple

Apple's Data set is very similar to Googles

- Apple tracks mobility in three categories driving, walking, and transit.
- Collects the data from requested directions
- Interpret the mobility data by creating "baseline days" from the day of January 13th, 2020.

Mobility Data: Apple

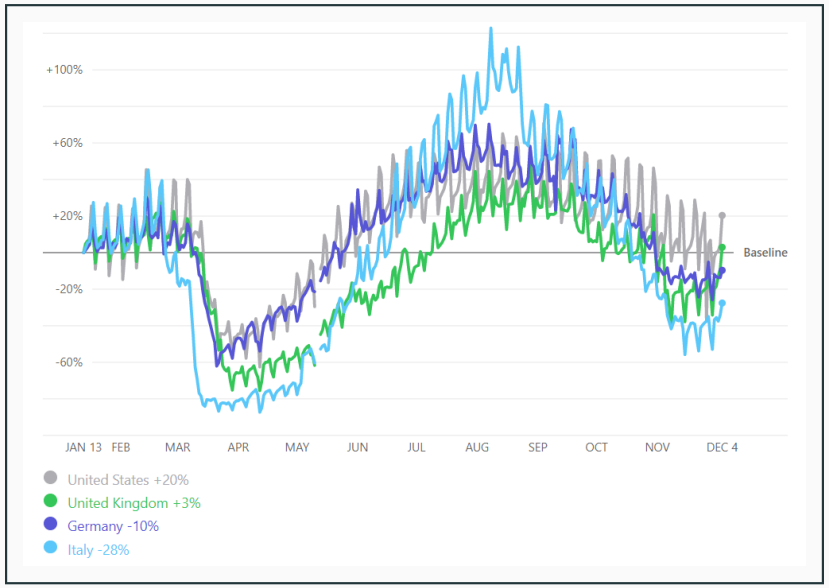
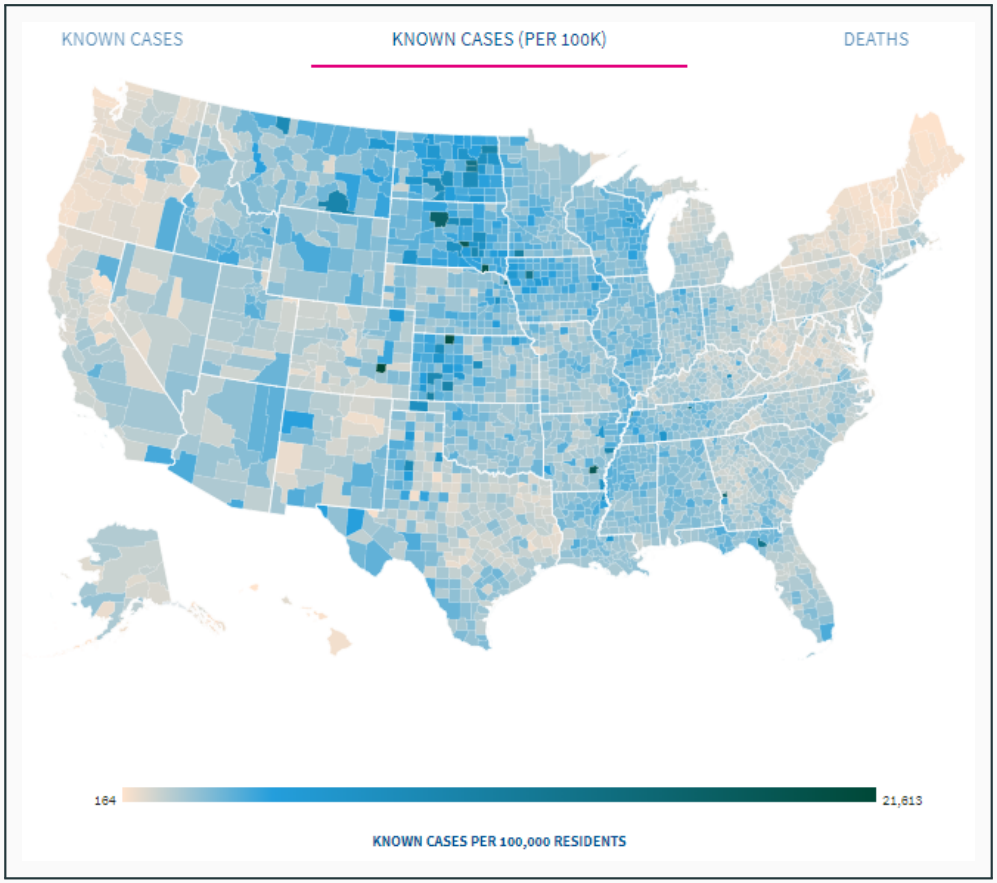


Figure 2: Example of Apple Mobility trends in the U.S. compared to baseline

3860	geo_type	region	transporta	alternative_	sub-region	country	1/13/2020	1/14/2020	1/15/2020	1/16/2020	129.25	107.79
3861	county	Orange County	walking		California	United Sta	100	104.74	115.1	122.04	144.22	168.84
3862	county	Orange County	walking		New York	United Sta	100	102.78	109.61	95.27	127.25	98.68
3863	county	Orange County	walking		Florida	United Sta	100	97.45	101.7	104.94	134.51	162.93

Figure 3: Example of Apple's raw data.

Mobility Data: USA FACTS



- Tracks COVID-19 data daily by state and county
  - Number of cases
  - Number of Deaths

Data Wrangling: The Next Step

In order to predict covid cases for a specific day we have to not only compile the data into one data set, but also consider the following:

- The time range from exposure to development of symptoms of COVID-19 is 2 to 14 days.
- The mean incubation period from COVID-19 is 4-5 days.
- Most contagious in the 48 hours before symptoms develop and they remain contagious for up to 10 days.

Data Wrangling

In order to capture this affect, a naive approach is taken. We have n data points which tell us how many cases were reported each day.

$$X_{cases,1} \ldots X_{cases,n}$$

We know cases reported today are based on past interactions. Thus, we create $Y_{cases,i}$ where

$$Y_{cases,i} = \sum_{j=1}^{7} X_{cases,i-j-4}$$

Thus, $Y_{cases,i}$ is our attempt to account for the number of people contagious in the previous week plus four additional days

Data Wrangling

Similarly, for mobility, we need to account for the mobility in regions of the previous week (plus four days), in order to predict cases for any given day. We have n mobility points for n days,

$$X_{mobility,1} \dots X_{mobility,n}$$

And we create $Y_{mobility,i}$, where,

$$Y_{mobility,i} = \frac{1}{7} \sum_{j=1}^{7} X_{mobility,i-j-4}$$

This was done for all Google and Apple parameters.

Methods and Results

In this study, our Methods and Results are split into two sections:

- Prediction Methods and Results
- Classification Methods and Results

We examine Prediction Methods and Results first.

Prediction Methods Applied

Methods we applied for Prediction:

- Ridge Regression
- Lasso Regression
- Poisson Regression
- GLARMA: Generalized Linear Auto-Regressive Moving-Average Modeling

Shrinkage and Regularization

- When the dimensionality of the parameter space or the number of predictors is too large or when some of the predictors are correlated, the usual estimation of the parameters will be very biased.
- To address these problems statisticians created a series of methods known as regularization techniques.
- The main idea is to add some bias to the estimates while reducing their uncertainty.
- This also makes it possible to remove some of the variables by shrinking their estimates to 0.

Shrinkage Methods: Ridge and Lasso Regression

For the usual regression or linear modeling, regularization is achieved by adding a penalty term to the criterion for finding the estimates.

$$\text{Ridge} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

$$\text{Lasso} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

The latter term in each expression serve as penalty terms - i.e. the larger the value of λ, the higher the penalty is for a model having larger coefficients.

Note that when $\lambda = 0$ the model becomes the standard least squares regression. This means the space of the constraint was large enough to encompass the true least square estimates.

Poisson Regression for Estimating Infected Cases

- We let the response or number of cases in each data block to follow a Poisson distribution, namely $y_i \sim Pois(\lambda)$
- We them model $\lambda = \exp(\alpha + \beta x^T)$, where α establishes a baseline for the disease rate, β is the vector of all coefficients, and x is the vector of all predictors.
- We estimate the components of β with the method of maximum likelihood estimation.

Introduction to the Time-Series Method Applied: GLARMA

- We model the number of cases at time t with a Poisson distribution whose intensity rate is λ_t
- We let λ_t be an exponential model of predictors
- The model then can be written as following:

$$y_t = Pois(\lambda_t)$$

$$\lambda_t = \exp(\alpha + \beta x_t^T + z_t).$$

$$z_t = \sum_{i=1}^{p} \phi_i (z_{t-i} + e_{t-i}) + \sum_{j=1}^{q} \theta_i e_{t-i},$$

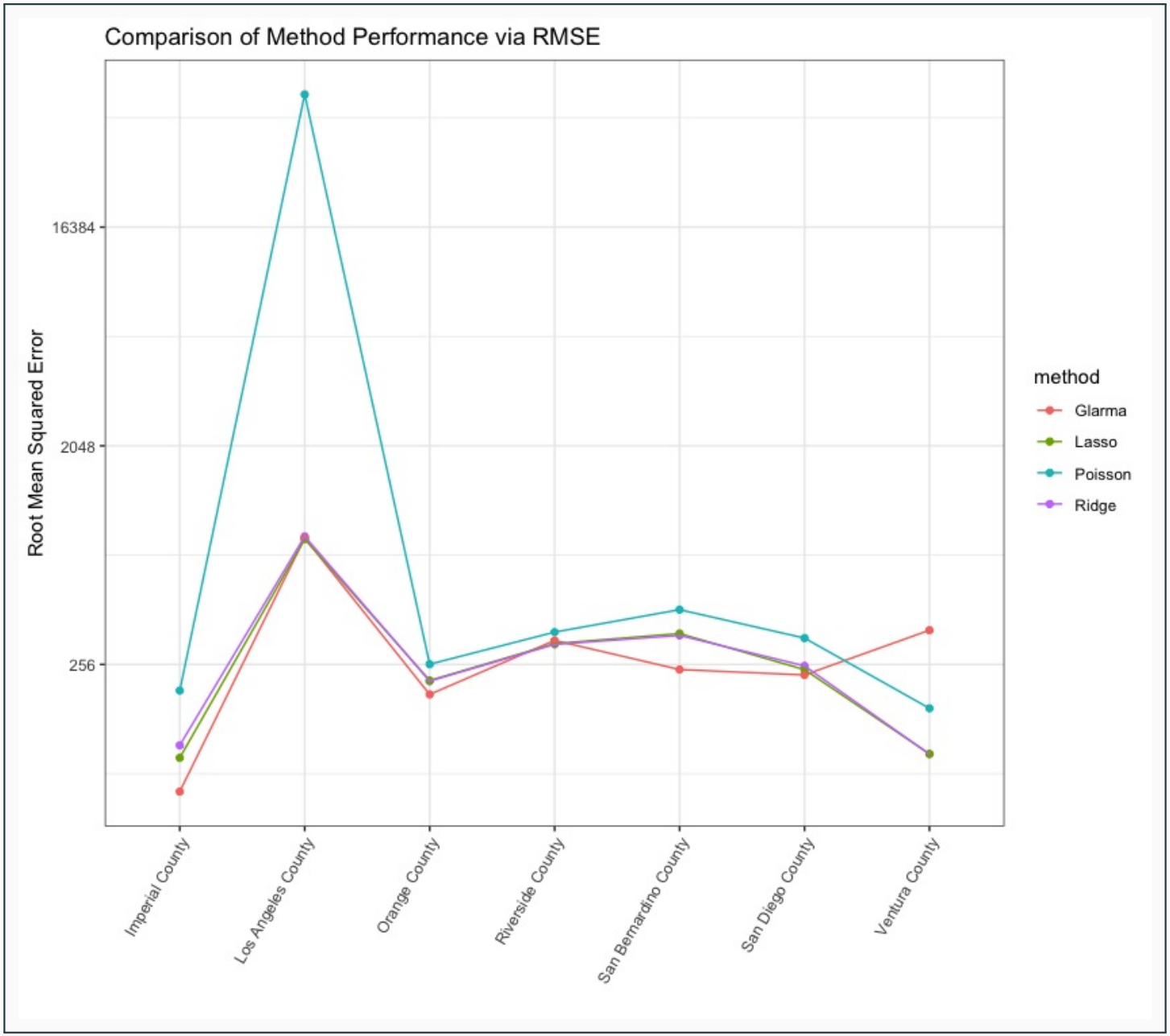
where α indicates the baseline, x_t is the vector of the predictors, z_t is comprised of auto-regressive terms of lag p and moving-average terms of lag q, and e_t represents the error terms.

- All parameters in the model are estimated by a maximum likelihood estimation procedure.

Prediction Results

Now lets discuss the accuracy of our methods on the seven SoCal counties.

Prediction Results



Prediction Results

Comparing the MSE between Lasso and Glarma Directly

County	Lasso	Glarma
Imperial	12631.81	5405.74
Los Angeles	738376.93	718519.75
Orange	46334.03	36757.22
Riverside	95997.32	103606.51
S.B.	115787.20	58113.45
San Diego	57312.96	55318.32
Ventura	21726.24	125700.84

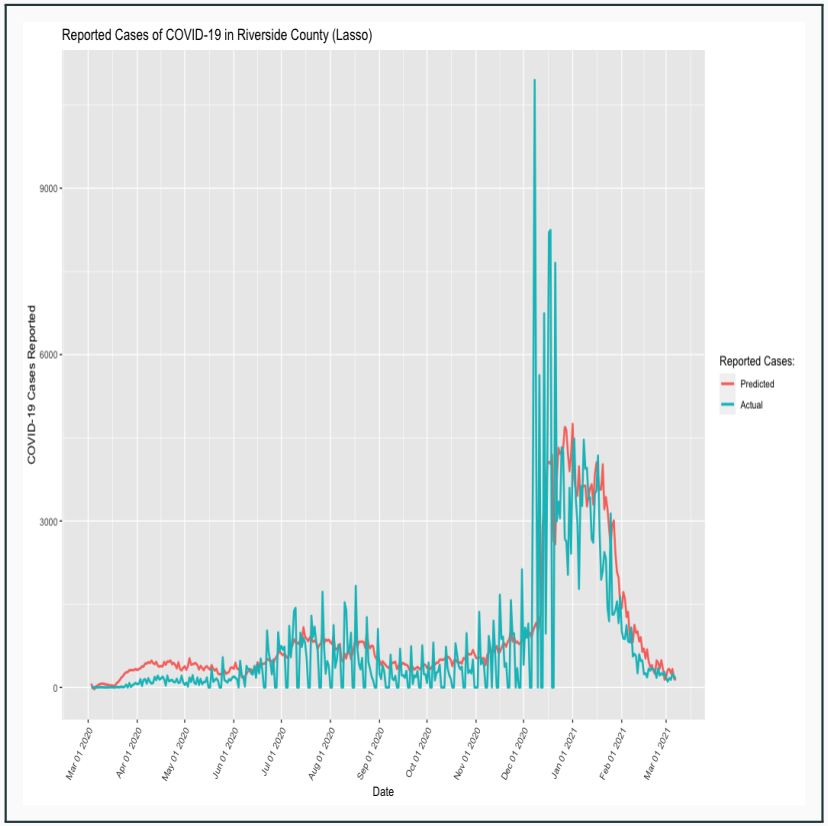
Table 1: Note: S.B. stands for San Bernardino

Lasso seemingly outperforms Glarma in Riverside and Ventura county

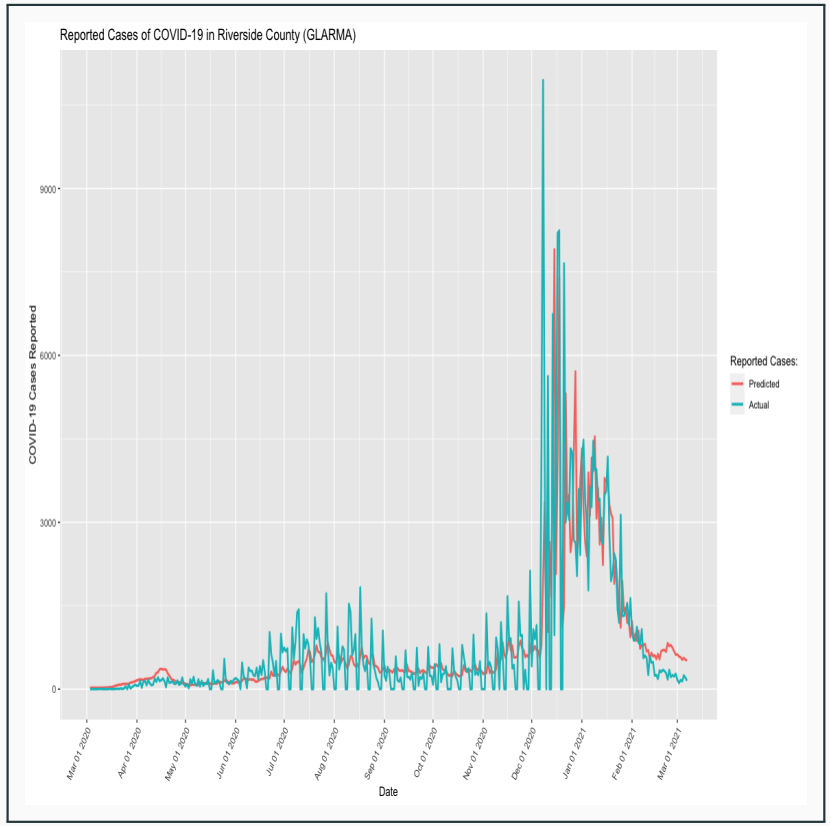
Prediction Results

A comparison of the Lasso and Glarma models on Riverside County.

Lasso



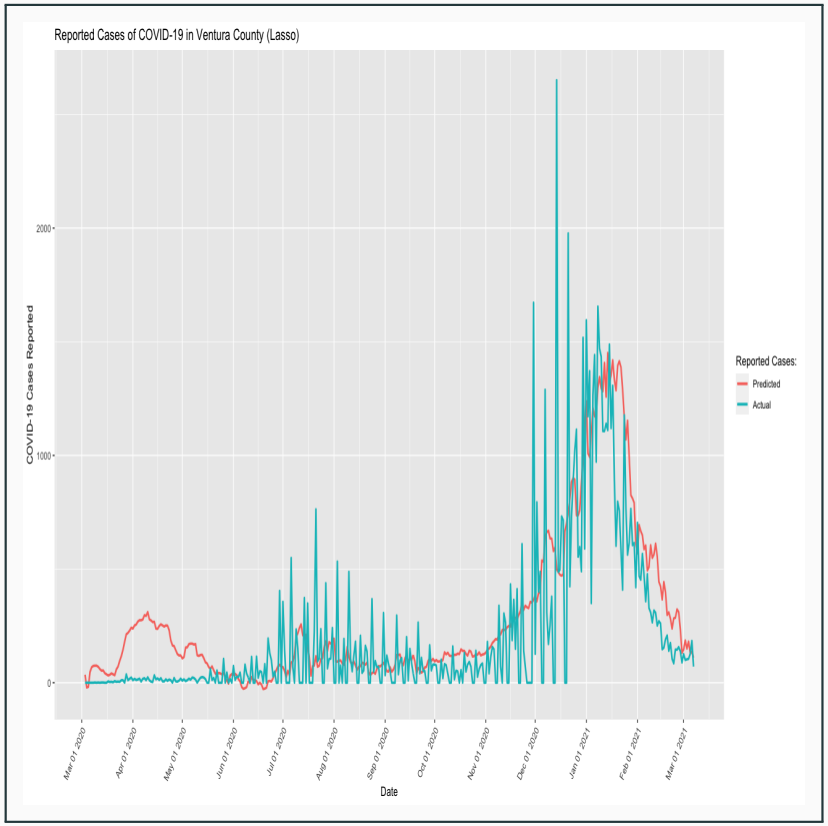
Glarma



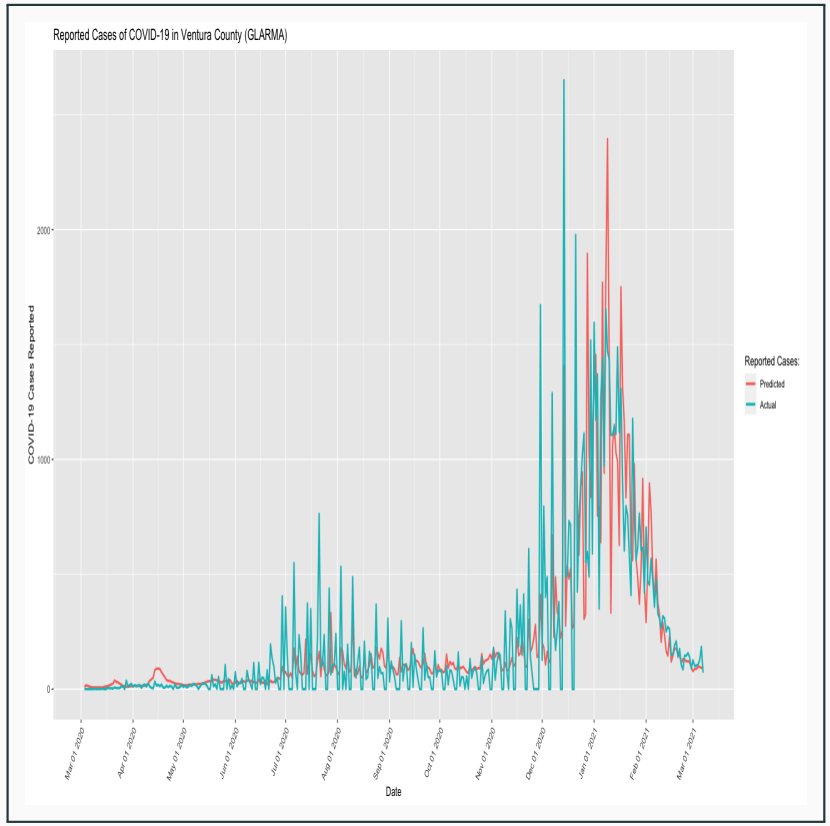
Prediction Results

A comparison of the Lasso and Glarma models on Ventura County.

Lasso



Glarma



Prediction Results

In Both cases, Riverside and Ventura, we can see that the Glarma model captures the trend of the data more accurately, while lasso only captures the average.

Prediction Results

Now lets compare all of the methods on one of the other SoCal counties which do not contain a high number of "0" observations. Specifically, Orange County where CSUF resides. We begin with the least accurate method and cycle through to the most accurate.

Prediction Results

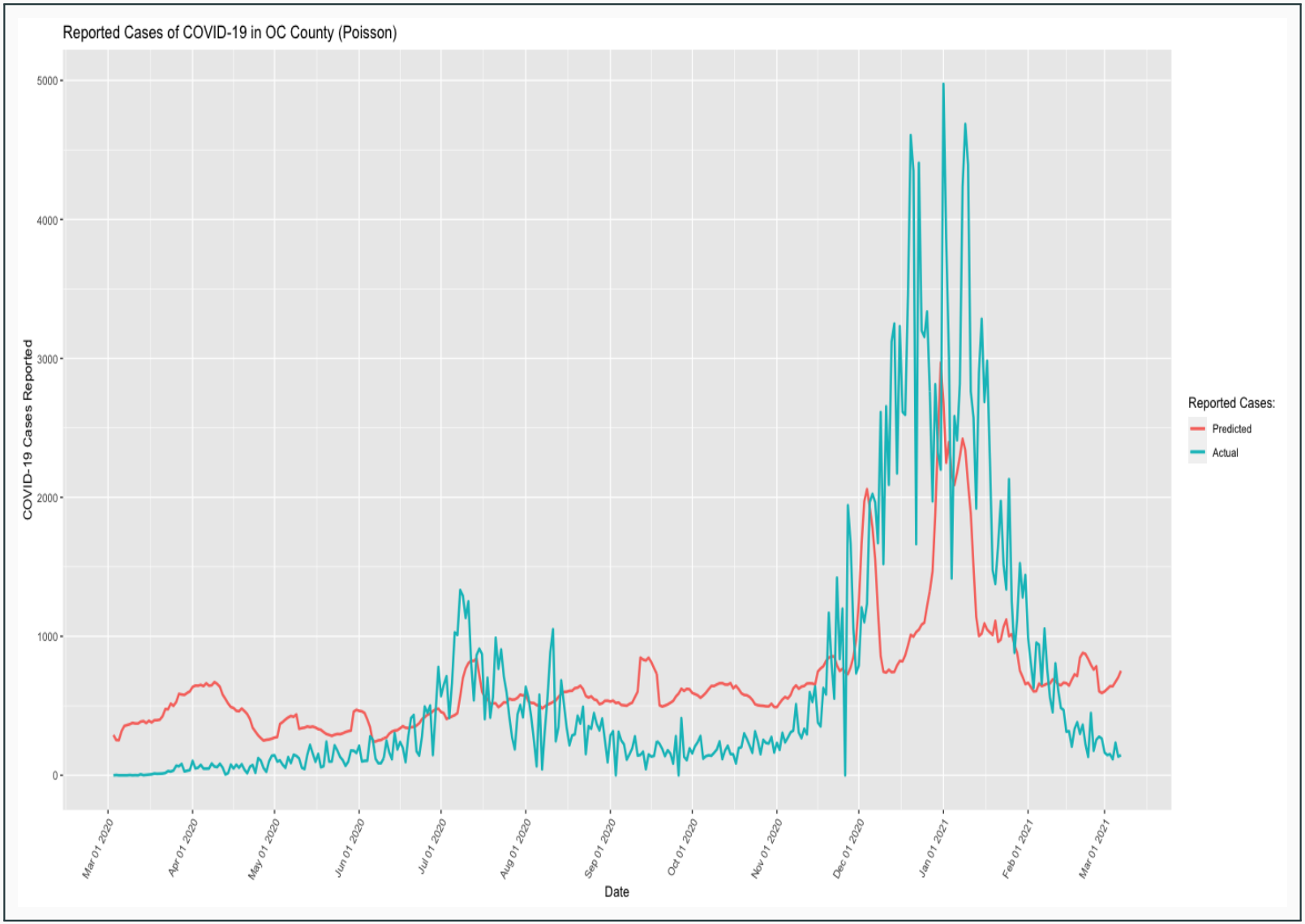


Figure 4: Poisson model on OC

Prediction Results

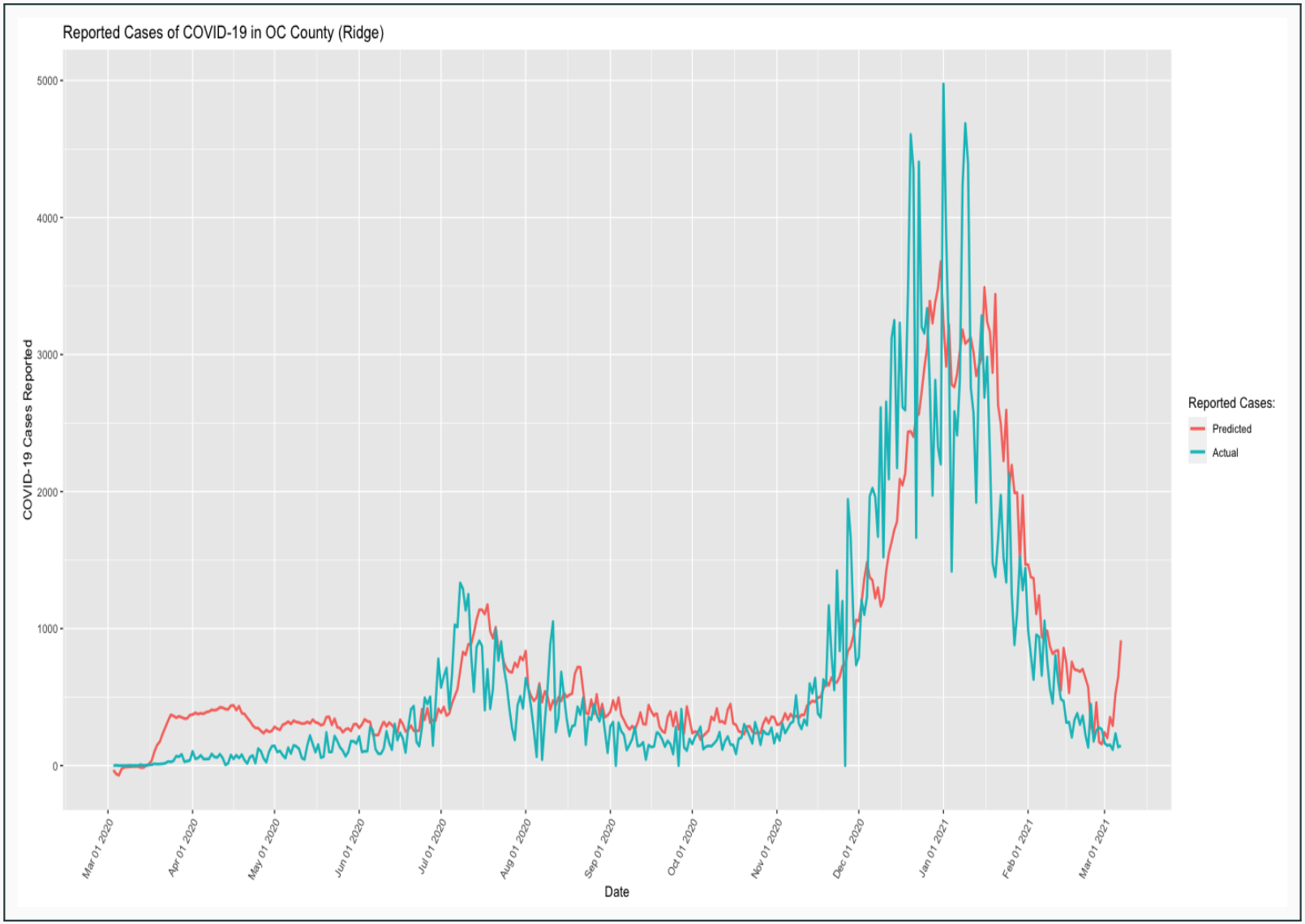


Figure 5: Ridge model on OC

Prediction Results

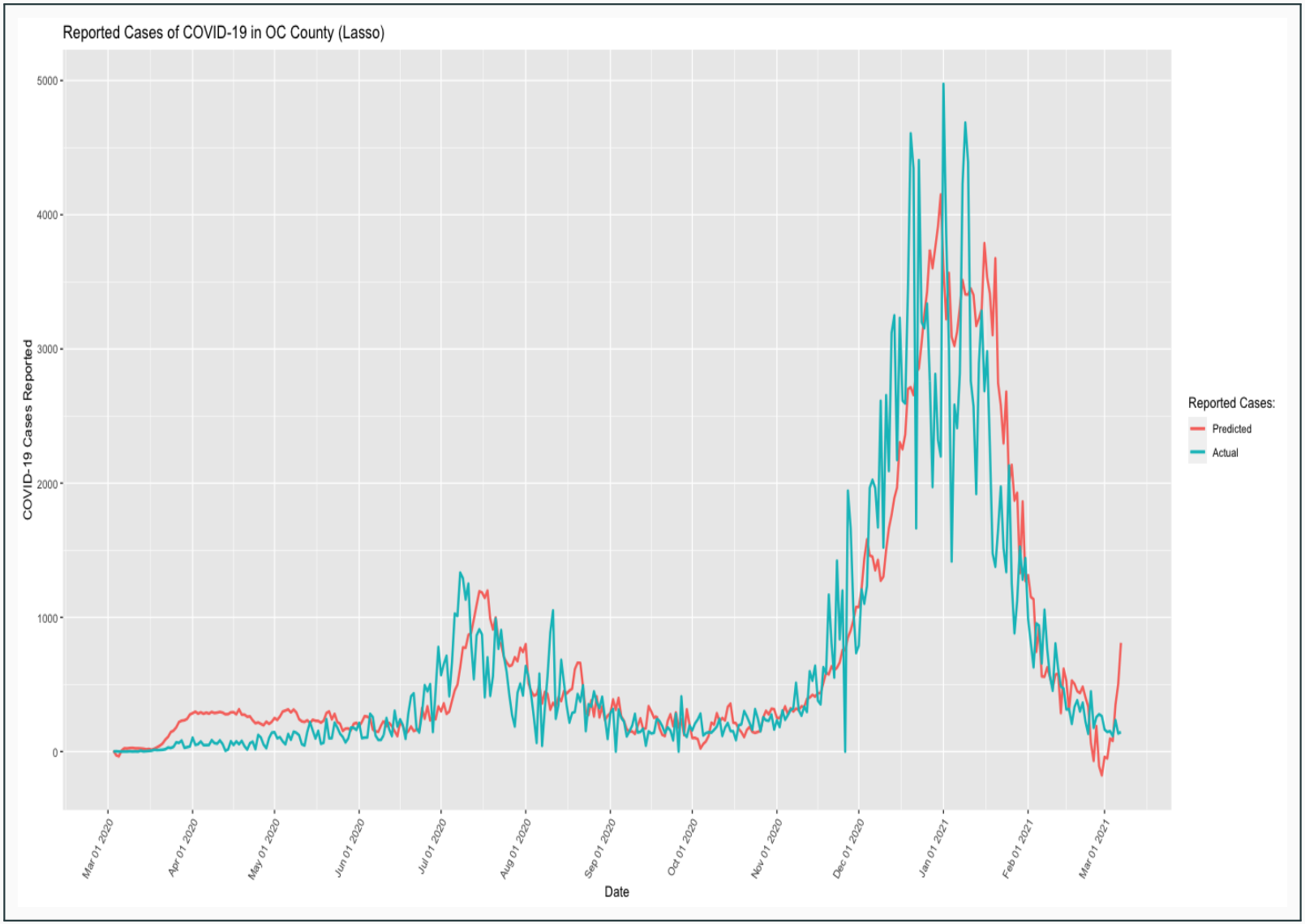


Figure 6: Lasso model on OC

Prediction Results

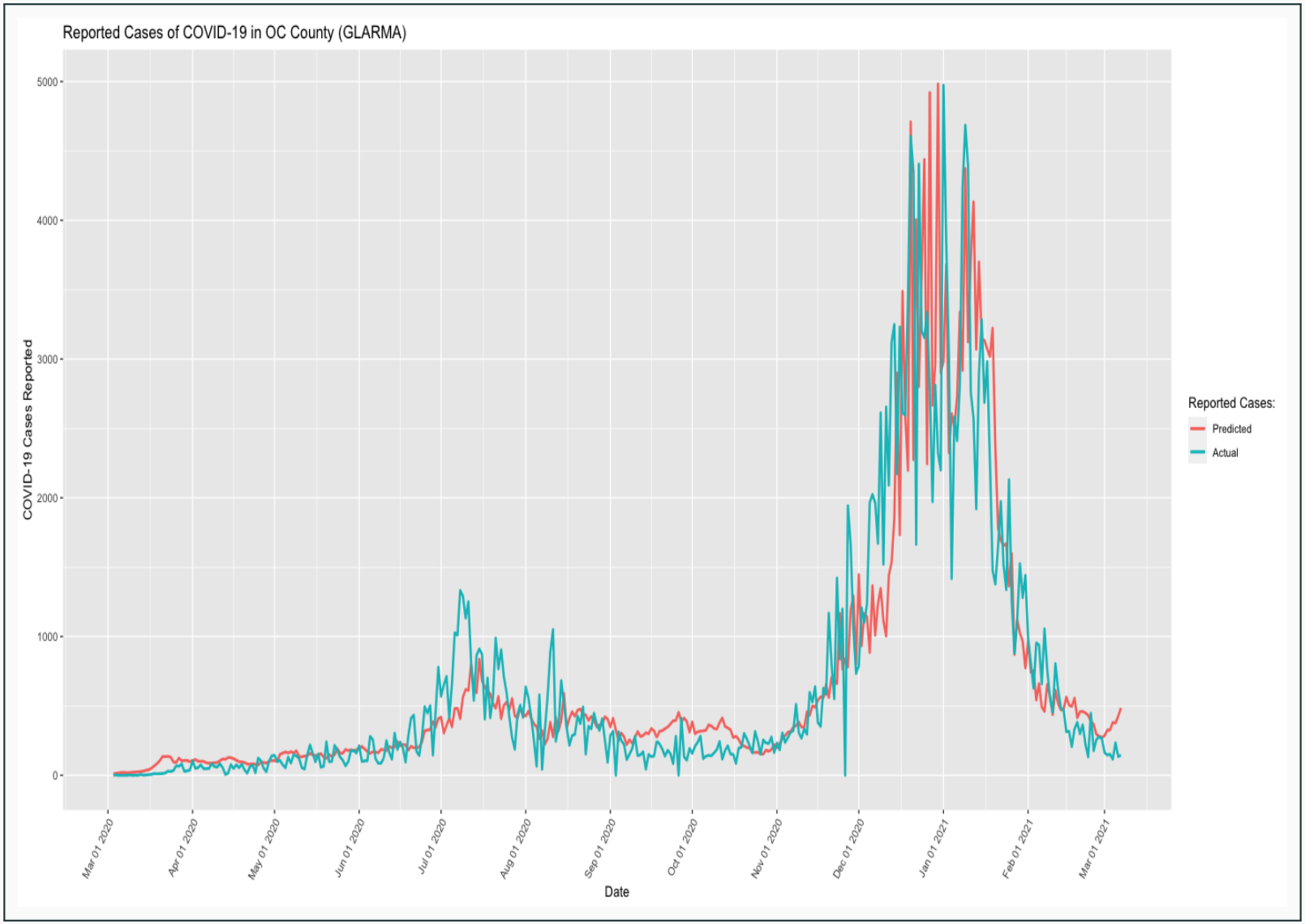


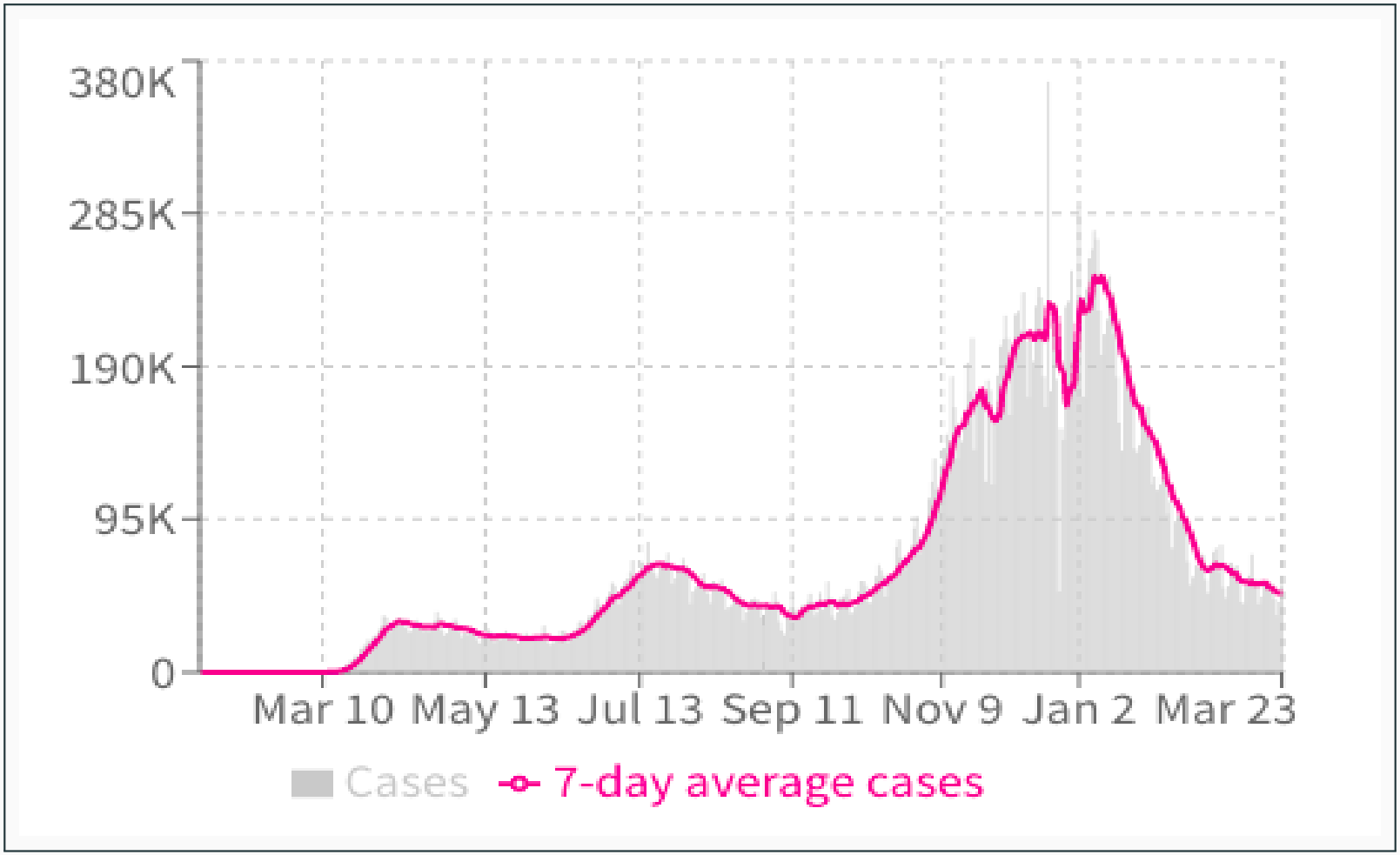
Figure 7: Glarma model on OC

Prediction Results

Now lets take a look at the trend of new COVID cases, and examine how we capture this trend with the lasso model.

Prediction Results

First consider the trend of new COVID cases via USAFacts.org



Prediction Results

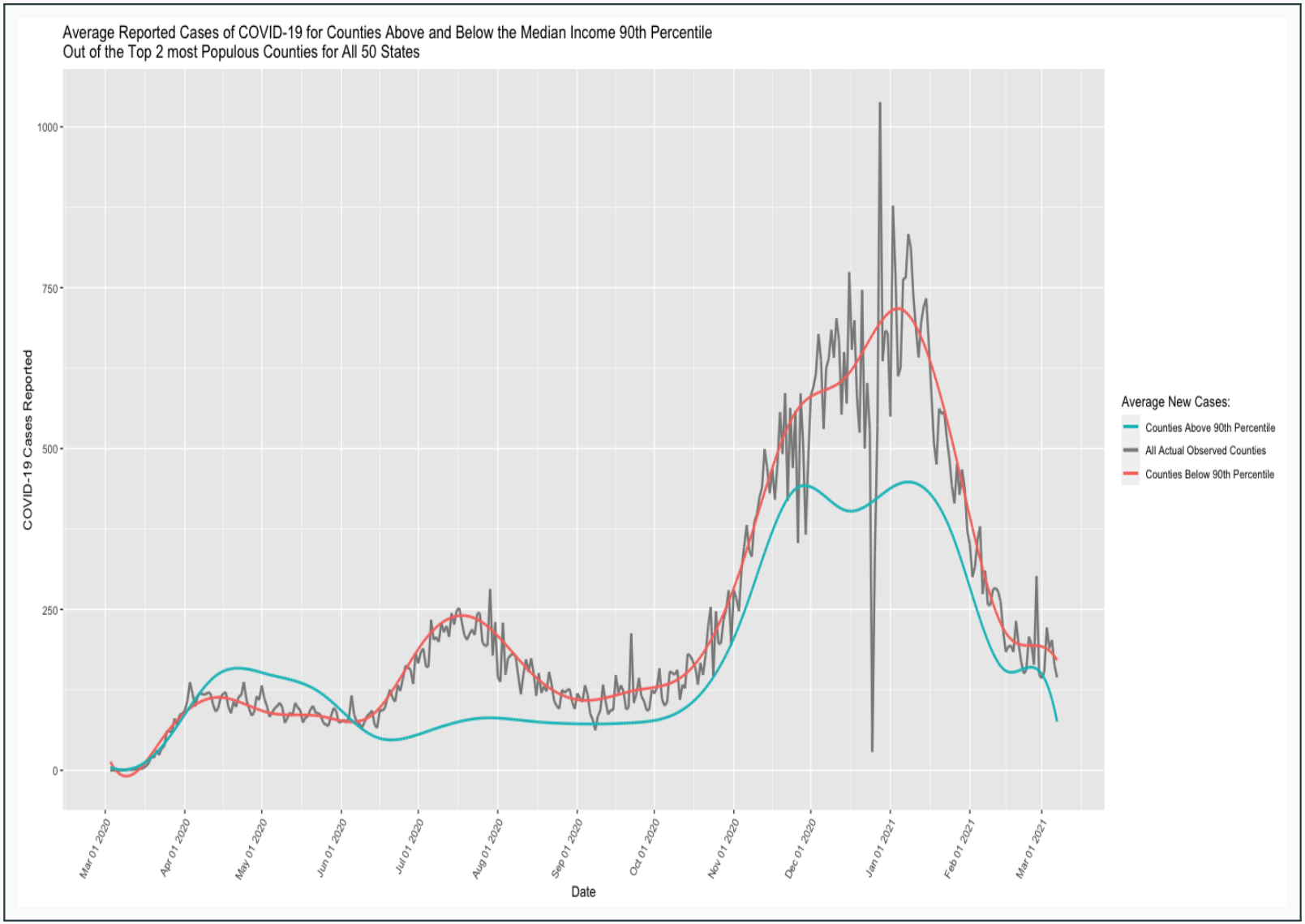


Figure 9:

Prediction Results

From examining the Average COVID in counties Below vs Above the 90th median income percentile plot from the last slide we can see the disparity in cases between counties of different wealth.

Classification: Overview

- Prediction: Mobility Data (by County) $\rightarrow$ Covid Data
- Classification: Mobility Data + Covid Data $\rightarrow$ County

- Why classification? We can learn more about...
  - Variable importance
  - How Covid impacts different types of populations

Classification: Overview

Methods that we applied for Classification:

- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- K-Nearest Neighbors
- Neural Networks

Classification: Logistic Regression

- Observations are classified based on their predicted probabilities.
- Logistic regression is easy to generalize...
  - to more than one predictor
    $log(\frac{E[Y|X]}{1-E[Y|X]}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$
  - to more than two categories - by looking at multiple pairwise regressions.

Classification: Linear Discriminant Analysis (LDA)

- Uses Bayes Theorem
- For K categories, we wish to maximize:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

where π_k is the prior probability that $K = k$ and $f_k(x)$ is the likelihood function.

  - We need to estimate π_k and $f_k(x)$
  - $\hat{\pi}_k = \frac{n_k}{n}$, assume $f_k(x)$ follows a Gaussian distribution
  - Gaussians have two parameters
    - $\hat{\mu}_k = \frac{\sum_{i:y_i=k} x_i}{n_k}$
    - *Assume the covariance matrix $\sum$ is common to all classes*

Adapted from *An Introduction to Statistical Learning*, by James et.al (Springer, 2013)

Classification: Quadratic Discriminant Analysis (QDA)

- QDA is almost identical to LDA
- We relax the assumption that all classes share a single covariance matrix
- The decision boundary is no longer a line, but a curve

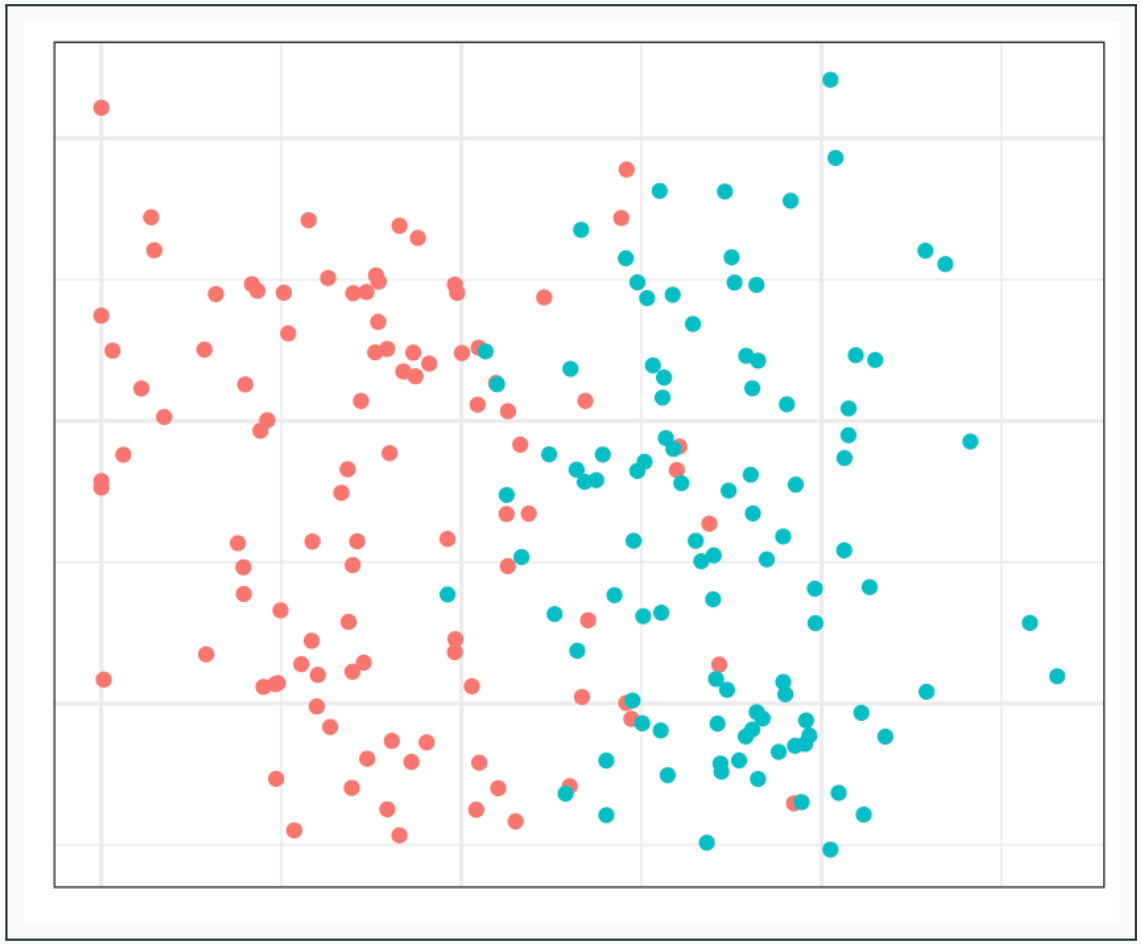
Adapted from *An Introduction to Statistical Learning*, by James et.al (Springer, 2013)

Classification: K-Nearest Neighbors (KNN)

- The previous approaches used **parametric** methods.
  $E[Y|X] = f(X)$
- **Non-parametric** models do not assume $f(X)$ follows some known functional form.
  - Harder to interpret
  - More powerful predictions
- Algorithm:
  - Choose a value of k
  - Choose a measure of similarity (e.g. Euclidean Distance)
  - Identify the k-most-similar observations
  - Classify by plurality vote

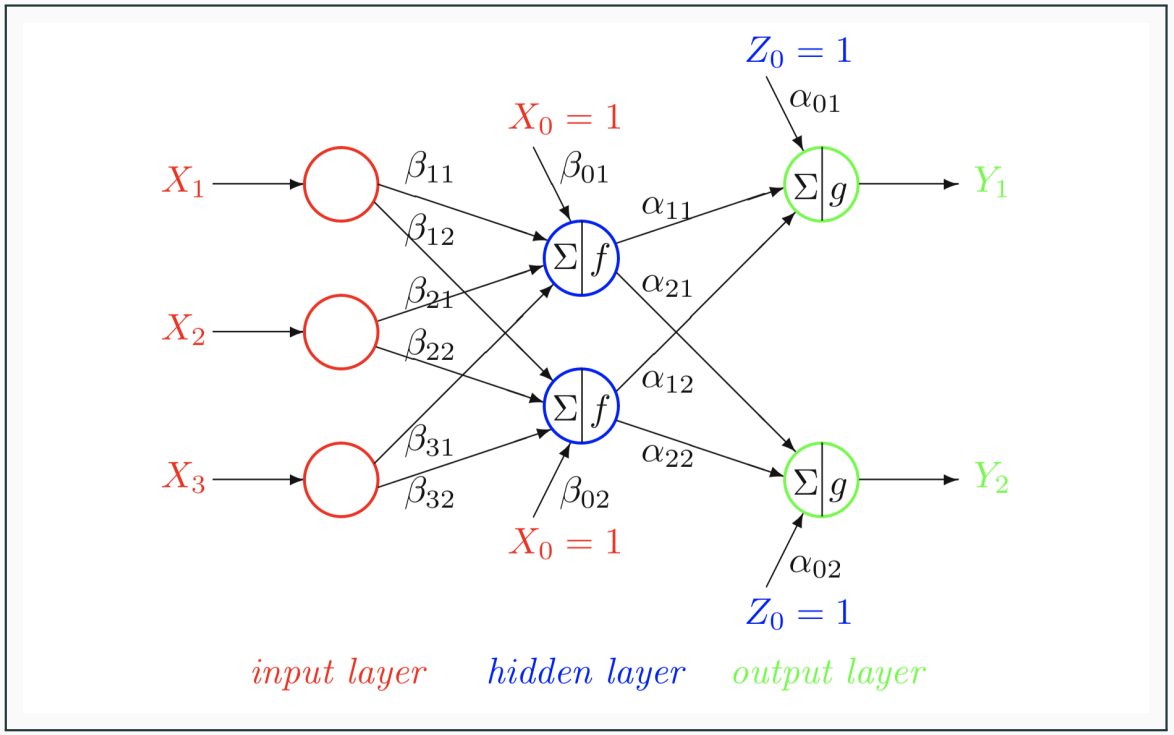
Adapted from *An Introduction to Statistical Learning*, by James et.al (Springer, 2013)

Classification: K-Nearest Neighbors (KNN)



Adapted from *An Introduction to Statistical Learning*, by James et.al (Springer, 2013)

Classification: Neural Networks

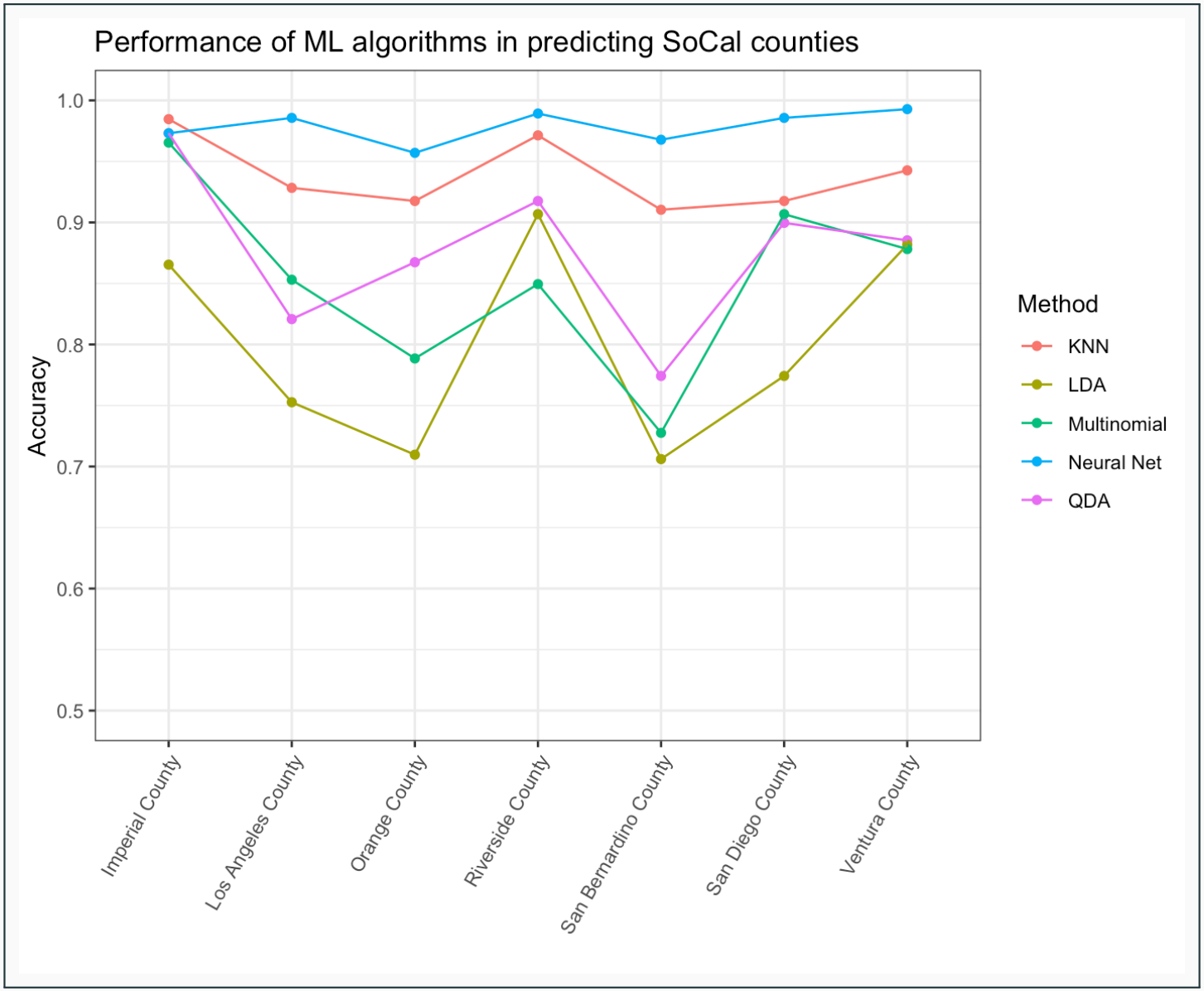


A neural network with a single hidden layer

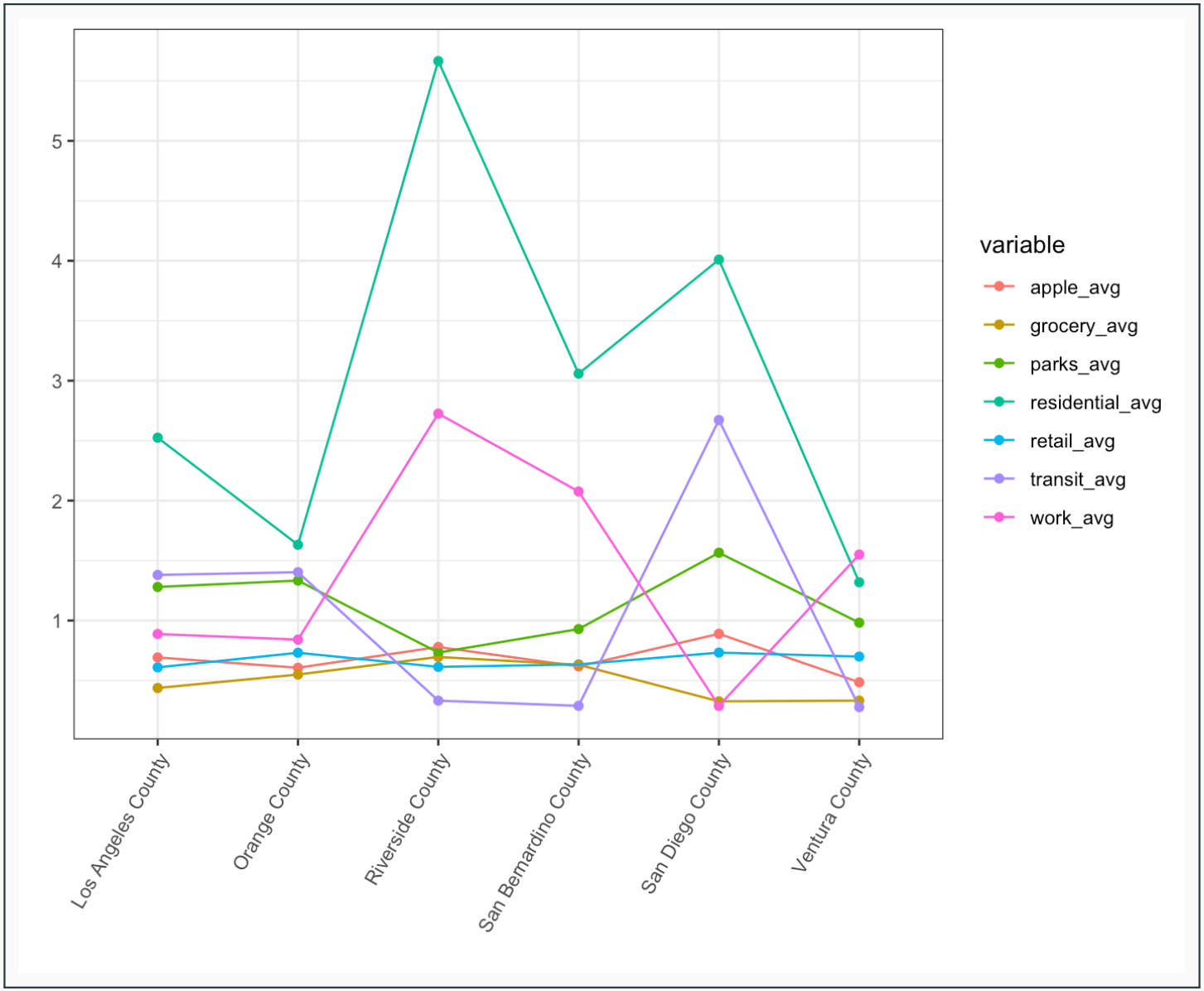
Results: Comparison of Classification techniques

- Accuracy of the classification models decreases drastically as the number of classes (counties) increases.
- We focus our attention on seven Southern California counties.

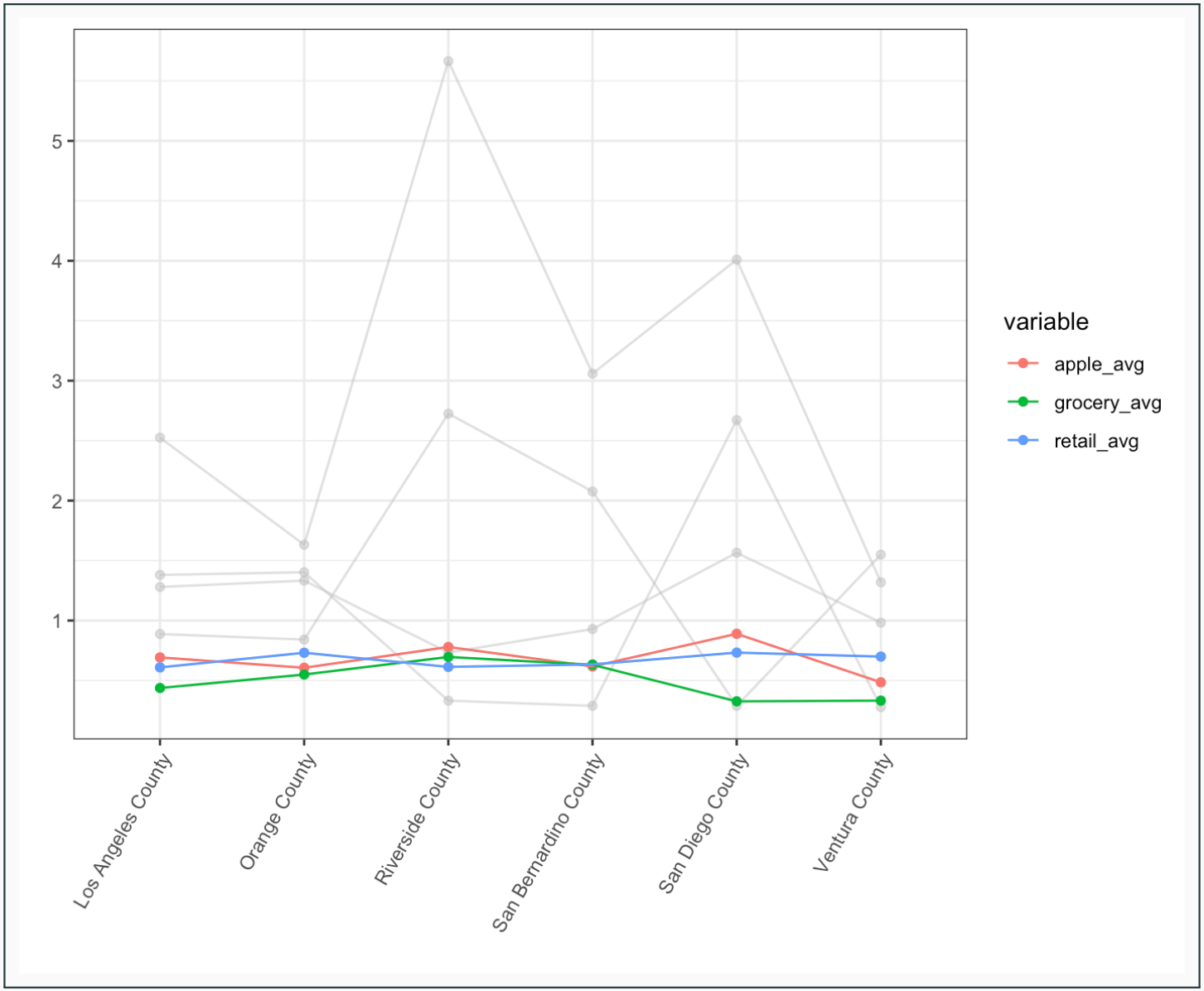
Results: Comparison of SoCal Counties



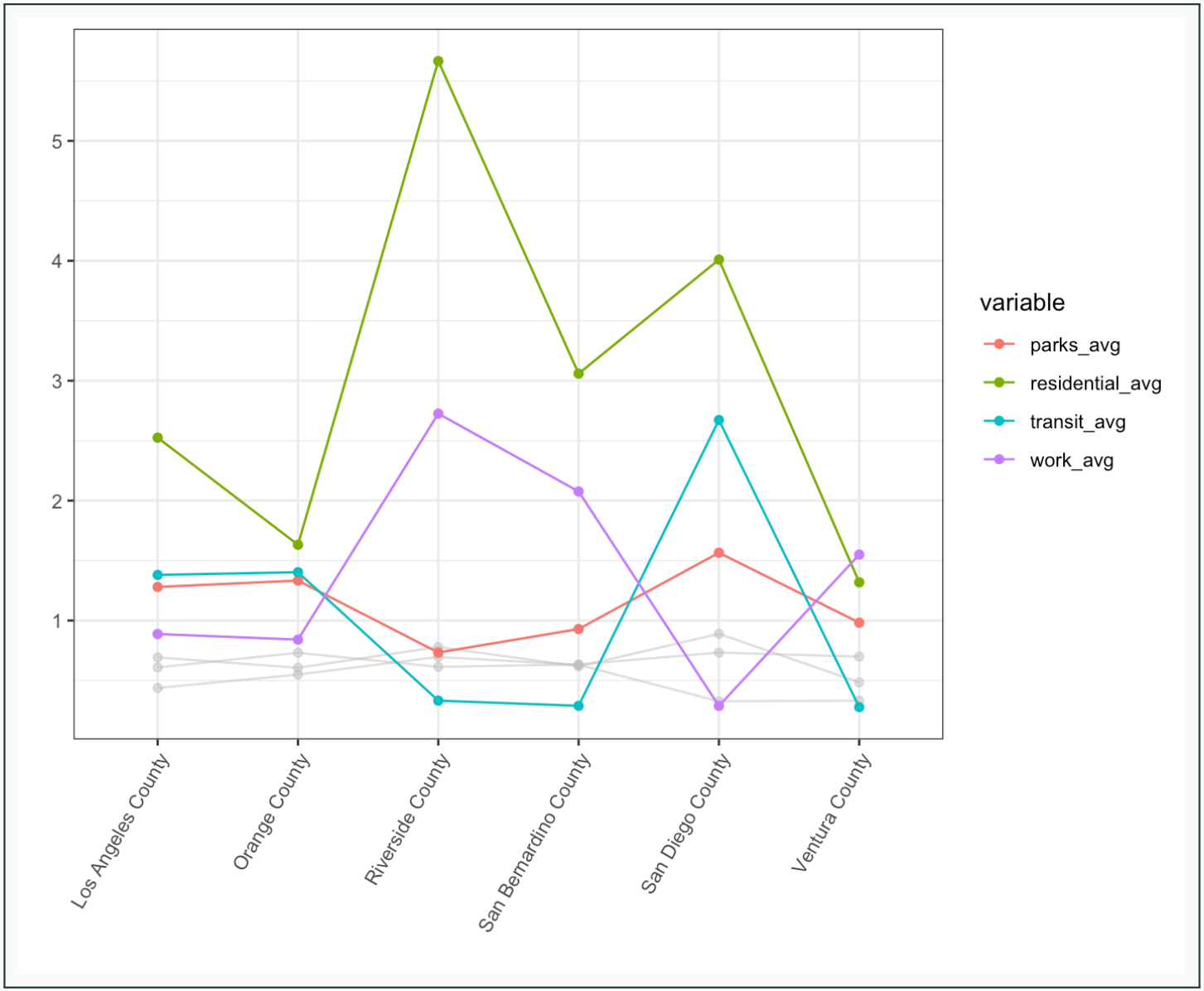
Results: Variable Importance



Results: Variable Importance



Results: Variable Importance



Conclusion

- Goal: Predict and Track daily new COVID-19 cases by county AND classify the counties based on their COVID-19 data
- Prediction:
  - Using simple statistical methodologies and machine learning yielded accurate models
    - (1) GLARMA
    - (2) Lasso Regression
    - (3) Ridge Regression
    - (4) Poisson
  - The two most populated counties in each state data set showed that the "most rich" (top 10%) counties were affected by COVID-19 differently

Conclusion

- Classification:
  - Each county has either the same impact or a distinct impact based on their variables from the COVID-19 data
    - Transit in LA vs Transit in Riverside

References

- Google mobility: https://www.google.com/covid19/mobility/
- Apple Mobility: https://covid19.apple.com/mobility
- USA Facts: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
- COVID-19 History: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7332915/
- COVID-19 Biology:https://nyti.ms/2GO6PDV

Thank you!

- Seth Arreola: setharreola8888@gmail.com
- Gwendolyn Lind: gwendolynlind@csu.fullerton.edu
- Caleb Peña: wilberforce116@csu.fullerton.edu
- Cameron Abrams: cwjabrams@csu.fullerton.edu