Roland DePratti
CIC Working Group
Research Report
04/03/2024

Overview

It is important to analyze the impact of the public policy decisions that were implemented to manage the COVID-19 epidemic. In our current heated political environment, health professionals and data scientists must provide factual data to illuminate the discussion. Both the CDC and Oxford University have published a great amount of data on COVID-19. Making sense of that data will allow more informed decisions for the next pandemic. This work aims to use machine learning models to identify the impact of the various public policy decisions on the health consequences of COVID-19 in Virginia.

The data used to develop and test the models are comprised of 155 weeks of COVID-19 data, including policies, cases, hospitalizations, and deaths for the state of Virginia from 01/01/2020 until 12/26/2022. The policy, cases, and death data were obtained from the Oxford Covid Research Group [1]. The weekly hospitalization data were captured from CDC hospitalization data [2]. The work focused on the impact of school closing, face covering, and vaccination policies on weekly new cases, new hospitalizations, and new deaths. The data was analyzed using two machine learning algorithms: SAMIRAX Time Series Analysis [3] and ElasticNet Regression [4].

The paper includes the following sections: I) Executive Summary; II) Preprocessing; III) Virginia Covid Cycles And Policy Decisions; IV) SAMIRAX Analysis; V) ElasticNet Analysis, and VI) Performance Comparison. References are listed at the end of the paper. My GitHub site which contains the Jupyter Notebook, project paper, and presentation slides is listed in the references.

I) Executive Summary

- Time series analysis highlighted a seasonal cycle of 18 weeks. This cycle appeared to coincide with the peaks between COVID variants.

- In Virginia, the Omicron strain of COVID-19 was responsible for many more cases than both the alpha and delta strains (24% more cases than alpha and 84% more cases than delta).

- Although the Omicron variant was viewed as milder and less serious, it accounted for only 6.4% fewer hospitalizations than alpha and 71% more than delta. The largest week of COVID-related hospitalizations occurred during the height of the Omicron strain.

- During the Omicron phase of the disease, total deaths were fewer than alpha (41% fewer), however, it accounted for 88.6% more deaths than the delta strain. The week with the second highest reported deaths due to COVID occurred during the Omicron phase.

- Two algorithms, Samirax and ElasticNet, were applied to the data to model the relationship between the three policy decisions (Face Coverings, Vaccination, and School Closings) and the pattern of COVID-19 weekly cases, hospitalizations, and deaths. The Samirax algorithm did not produce a very accurate model and was discarded. The ElasticNet algorithm developed an accurate model. The ElasticNet model also provided information about the relationship between health outcomes and policy decisions. The later results are listed in the summary.

- School closings had a decreasing impact on new COVID cases, especially when that was the policy two or three weeks prior. Stricter vaccination policies also decreased cases, but only if that was the week's policy. Older facial covering policies appeared to increase new cases.

- School closings also decreased new hospitalizations, especially when that was the policy two or three weeks prior. Stricter vaccination policies also decreased cases, but only if that was the policy two weeks prior. For hospitalizations. facial covering policies appeared to increase new cases.

- For COVID-19 deaths, the number of deaths 2 weeks prior was a small predictor of a decrease in deaths this week. However, current and prior face-covering policy again was a predictor in an increase in deaths this week.

- These counterintuitive results for face-covering policy might be the result of the political backlash in some quarters. In many states, there was a big gap between the policy and whether citizens followed the policy. In the case of school closings, public schools, and many other schools followed suit, so the gap between policy and the follow-up was smaller.

II) <u>Preprocessing</u>

The Oxford policy file was a daily file that was summarized to the weekly level to align with the CDC hospitalization data. COVID health data (daily cases, hospitalizations, and deaths) were summarized by adding daily numbers to the weekly level and policy data (i.e., school closing, face covering, and vaccination indicators) were summarized to the highest policy level for that week. The original Oxford health data was cumulative. Weekly new cases and deaths were generated by subtracting the current week from the prior week. The date on the new weekly rows was represented as the Monday of the week. The CDC hospitalization data was already represented as weekly data but was represented as Sunday each week. The weekly date had to be converted to the Monday of the week to align with the Oxford weekly data. All health data was defined as integers and the policy indicators were defined as ordinal categories. After preprocessing, the file contained all the policy decisions to be used as predictors, as well as the weekly health counts to be used as responses or outcomes.

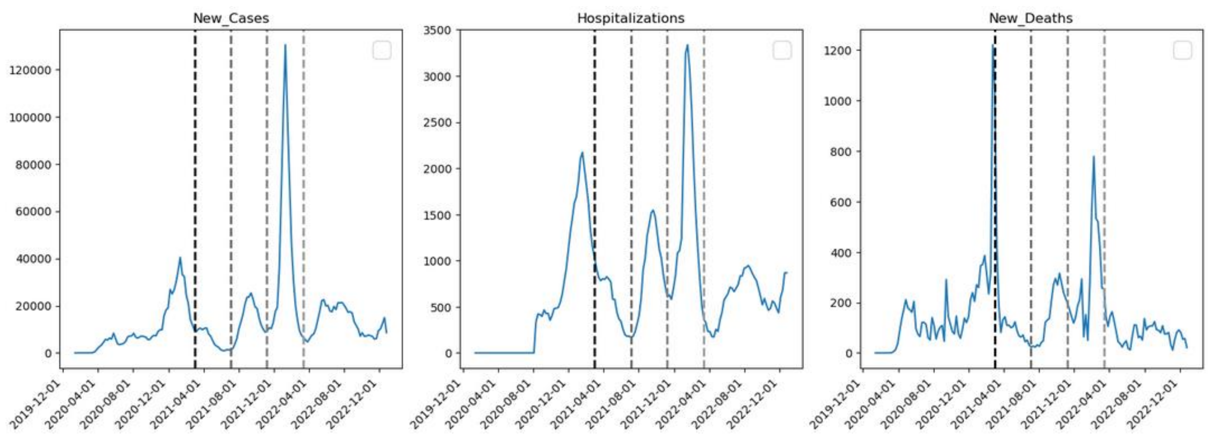
When examining time series data, it is important to determine that the data is stationary, otherwise the results will be unpredictable. Two tests were applied to determine stationarity, the Augmented Dickey-Fuller Test (ADF) and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. Both tests indicated stationarity at the .05 level.

Policy changes do not have immediate results. It may take several weeks before policy changes have an impact on health outcomes. This is especially true for something like hospitalizations. For that reason, we added with each week's data the policy indicators in place one (P1W), two (P2W), or three weeks prior (P3W). That way we could measure the relationship between earlier policy changes and the week most likely to affect it.

III) Virginia Covid Cycles And Policy Decisions

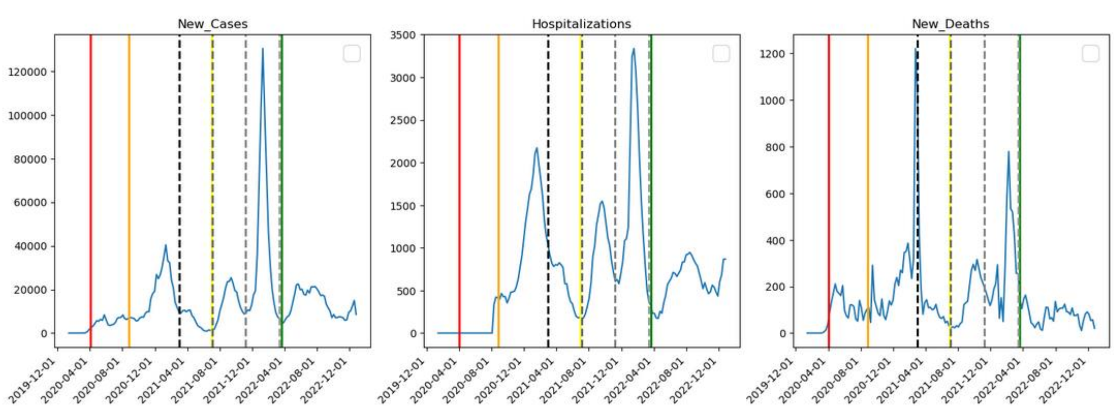
a) COVID Cycles

The gray lines represent the estimated onset of the different strains of COVID-19 in Virginia: alpha, beta, delta, and OMICRON. Incidentally, these 'cycles, echo the 18-week seasonality found in the data. In the middle of each cycle is a peak new cases day. Hospitalizations peak about x weeks after new cases, while deaths peak about x weeks after hospitalizations.



b) School Policy Changes

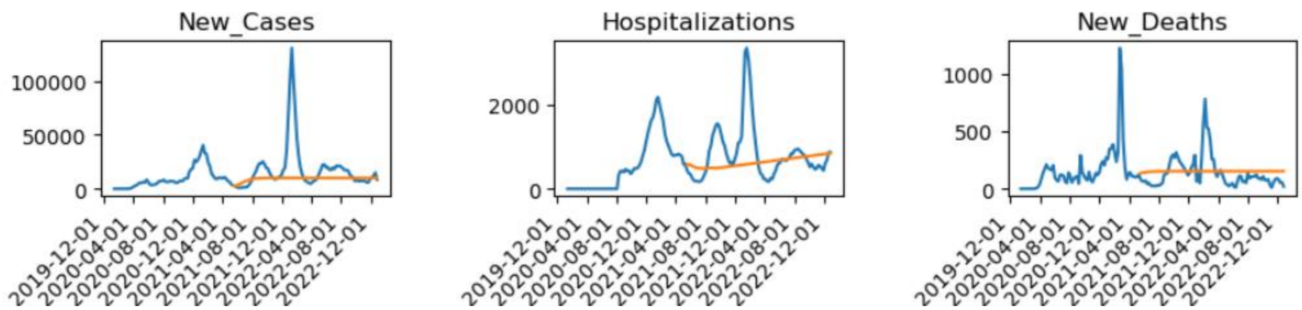
These charts add lines representing decreasing school closing policies (red = all closed, orange = some closed, yellow = open with accommodations, i.e. masks, green = open without restrictions). The highest peak in new cases and hospitalizations, and the second highest in deaths happened during OMICRON, after schools were opened.



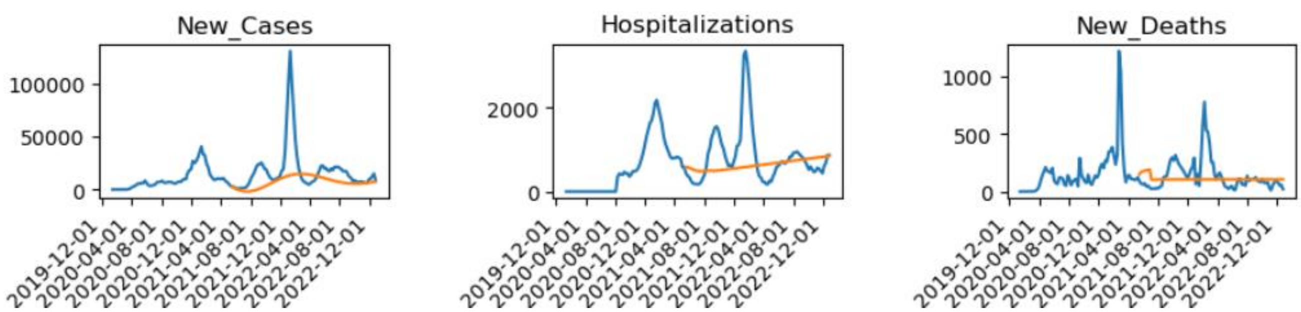
IV) <u>SAMIRAX Models</u>

SAMIRAX is a library of functions to complete time series analysis. It considers seasonality and allows the inclusion of additional predictors besides time. Many models were built using different combinations of current and prior policy indicators. However, the predictive ability of the SAMIRAX-developed models was poor. Although the predictions followed the general direction of the curve. The root mean square errors were very large. Below are some charts for some of the models, comparing predictions to actual values. The blue line represents the actual values, and the red lines represent the predictions.

a) SAMIRAX Models with No Additional Predictors



b) SAMIRAX Models with One Additional Predictor

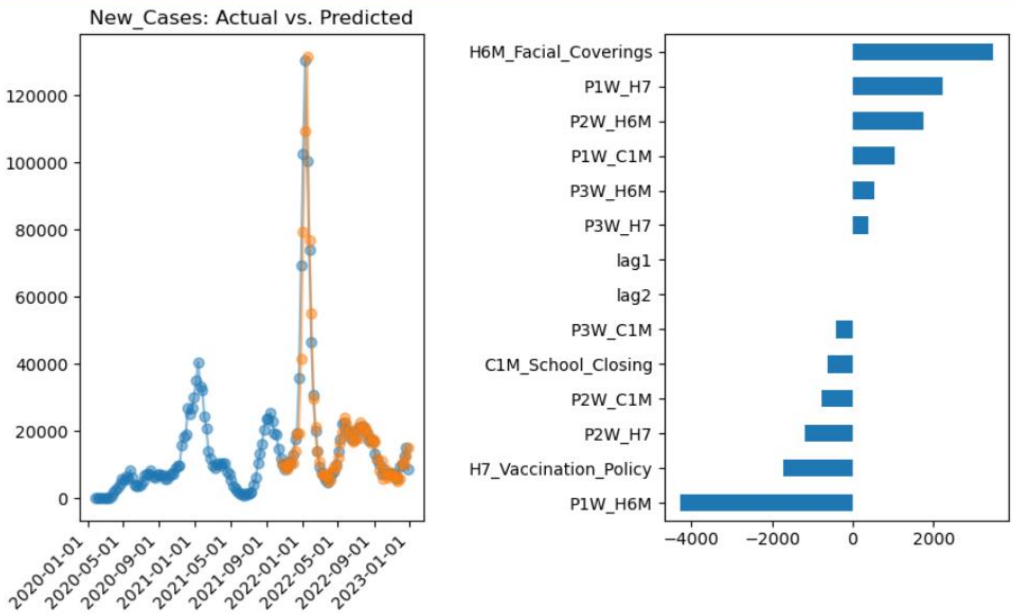


V) ElasticNet Models

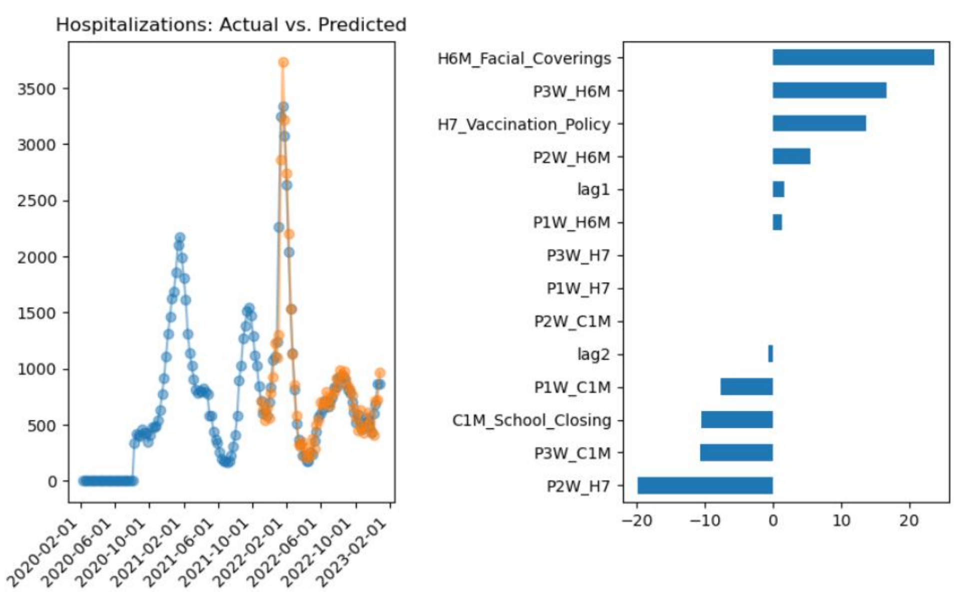
ElasticNet is a linear regression algorithm. It limits the correlation between predictors (high coefficients). It can eliminate predictors already represented in other predictors (feature selection). I supplied the algorithm with all the predictors and let it select the important ones. I then asked it to weigh predictors on importance (Very Important Features).
The ElasticNet models performed much better. Predictions closely aligned with the actuals, reporting much smaller root mean square errors (see performance comparison section). The model also provided some interesting insights into how policy indicators affected the health counts. The following charts demonstrate the actual results listed in the summary. The left diagram shows the actual new cases (blue) and the predictions for the later weeks of the pandemic. The predictions closely overlap the actuals. The right diagram shows the association of the policy indicators to the

new cases. Bars to the left of the center line represent a negative association with the response variable, and bars to the right associate a positive association with the response variable.

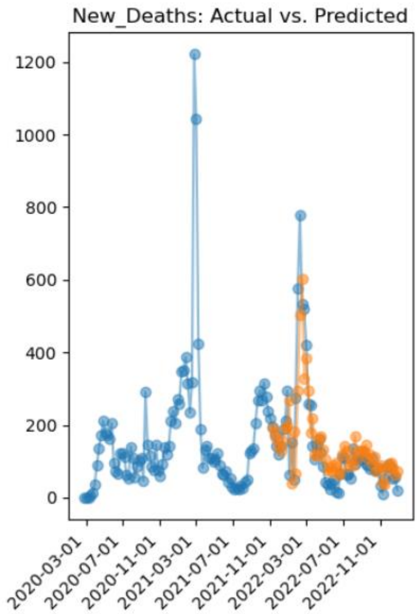
a) New Cases

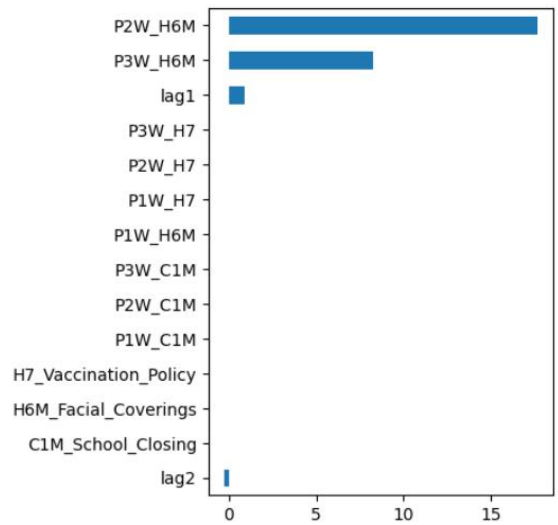


b) New Hospitalizations



c) New Deaths





VI) Result Comparison

The chart below compares the results for the various models on performance measures. The chart demonstrates that the ElasticNet models have much smaller errors. The performance RMSE improvement between ElasticNet and SAMIRAX for new cases was 68% (7,7260 vs 2,2865). For hospitalizations, the performance RMSE improvement between ElasticNet and SAMIRAX was 76% (164 vs 684). A 38% improvement occurred in deaths (81 vs 132).

Response/Model	Algorithm	RMSE	Improvement
New Cases	Sarimax	2,2865.27	
	ElasticNet	7,296.88	68%
New Hospitalizations	Sarimax	683.89	
	ElasticNet	164.02	76%
New Deaths	Sarimax	131.78	
	ElasticNet	81.17	38%

Table 1 Algorithm Performance Comparison Results

VII) Future Work:

Shikhar Johri recommended that the AutoTS library is a good method to experiment with a set of time series algorithms. The library easily builds pipelines, runs various algorithms, and completes performance metrics on each algorithm. I have started some experiments incorporating that pipeline.

References

[1] https://github.com/OxCGRT/covid-policy-dataset

[2] https://data.cdc.gov/Public-Health-Surveillance/Weekly-United-States-COVID-19-Hospitalization-Metr/akn2-qxic/about_data

[3] https://www.statsmodels.org/stable/examples/notebooks/generated/statespace_sarimax_stata.html

[4] https://machinelearningmastery.com/elastic-net-regression-in-python/

[5] https://github.com/rdpratti/Analyzing-Pandemic-Response