# An Analysis of the COVID-19 Pandemic in Ohio Using Twitter Data

Tarun Paravasthu and Madeleine Johnson

***Abstract*— In this report, we discuss the results of descriptive analysis and gradient boosting regression on early Covid-19 cases in Ohio. This analysis can help identify which areas were and continue to be most at risk to new diseases.**

## I. INTRODUCTION

### A. Background

Ohio's experience with COVID-19 was unique in some ways, but also reflective of broader trends created by the pandemic across the United States. Some of the similar challenges faced by Ohio to the rest of America were significant cases in prisons, a need to expand hospital capabilities, quarantining and stay at home requirements, and an increase in unemployment claims. One of the main differences between Ohio and the rest of the United States was how swiftly governor Mike DeWine acted at the start of the pandemic. He was one of the first governors to declare a state of emergency for the state; there were only three known cases on March 9th when the state of emergency was declared. DeWine also took aggressive measures such as canceling festivals in early March which was initially met with backlash but later praised as a good public health move. Ohio was also the first state in the country to announce statewide school closings, on March 12th, 2020. The state was also a forerunner in testing sterilization techniques to combat the mask shortage and shifting manufacturing capabilities to supplies to help keep their citizens safe. There was some controversy surrounding abortions being grouped in with nonessential medical procedures initially. Overall Wikipedia has more positive than negative content of Ohio's handling of the pandemic. **(a)**

### B. Summary

We performed descriptive and predictive analysis to examine the correlation between these tweets. This involved analyzing specific similarity scores, calculating and visualizing county averages, and analyzing awareness over time. For the model we used several machine learning and optimization techniques to predict cases based on other variables.

### C. Results

Some of our descriptive findings were that illness and core (covid) had the 3rd and 4th highest average normalized jaccard awareness values, behind sports and entertainment, and that Hamilton county had the highest cases and deaths. We found that we were able to build a fairly accurate model to predict cases using the data, with deaths, age brackets, vaccination rates, county, and race being important factors to the model.

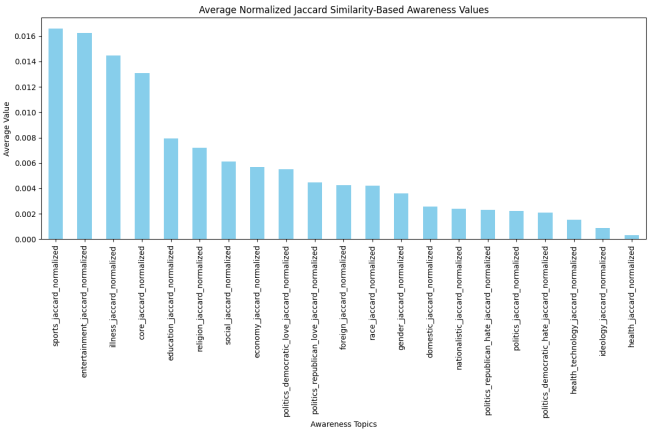


Fig. 1.

## II. DATA

### A. Datasets

Each observation in the dataset contains 180 variables. There are a total of 10,472 observations in the dataset, with 30% of the being used for training and the other 70% being used to test the model. The earliest observations in the dataset were collected in November of 2019 and the latest a few months after. The dataset is the culmination of over 46 million tweets from 91 thousand unique users.

### B.

There were a total of 21 variables for normalized jaccard topics, with averages shown in Fig. 1. The top 5 average awareness values were for sports, entertainment, illness, core, and education. There was a pretty large drop in awareness between core and education. The rest of the variables descend pretty linearly after that. The top awareness value being sports indicates a high level of awareness regarding athletics across Ohioans tweets during the beginning of the COVID-19 pandemic. Although this chart has no element of time to it one would expect some of these numbers to evolve throughout the pandemic. Tweets regarding education may have peaked the week of March 12th when schools were ordered to switch to remote learning. **(b)**

### C.

The county with the highest mean awareness value for the core_jaccard_normalized variable is Delaware county, closely followed by Richland county. This means that tweets created in these counties on average show a higher degree of awareness and engagement surrounding topics that are

central to the COVID-19 pandemic. Other counties that exhibited high awareness values are Clermont county, Medina county, and Morrow county. There were a few counties who's awareness values were so low that they do not show up on the graph. These counties include Hocking county, Holmes county, Champaign county, Highland county, and Paulding county. All core values are shown in Fig. 2. **(c)**

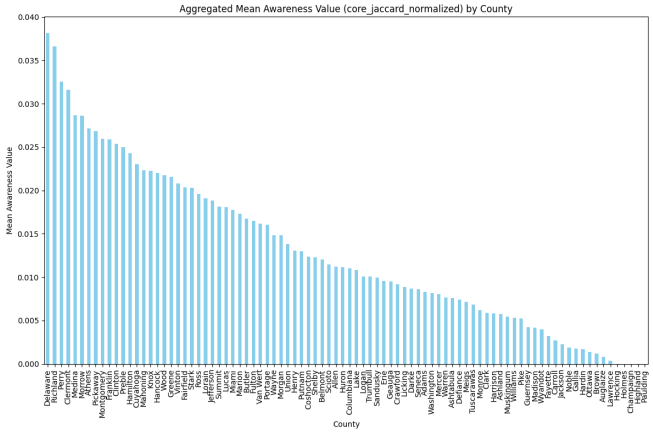


Fig. 2.

*D.*

We created two county-level maps (Fig. 3 and Fig. 4) to visualize the average number of cases and the average number of deaths by county. The five counties with the highest number of cases per capita are Hamilton county, Cuyahoga county, Pickaway county, Miami county, and Franklin county. The five counties with the highest deaths per capita are Hamilton county, Miami County, Cuyahoga county, Portage county, Mahoning county. More data analysis would be needed to help explain why the counties with the most cases did not necessarily have the highest number of deaths. Potential reasons include a higher number of seniors living there, residents being less able to isolate, and willingness or ability to seek medical care when sick. **(d)**

*E.*

We also created a line chart for each normalized jaccard awarness score over time (Fig. 5).

The single biggest spike in awareness level was in the race topic, 20 days from the start of when tweets started being collected. There was a spike that was double the size of the next largest spike, which was gender around day 70 of data collection. Also starting around day 70 illness and education started trending upwards and remained elevated for the rest of the collection period. The sports topic had multiple small spikes throughout the entire collection period which most likely are correlated to sporting events. The largest sports spike happened around day 115 of data collection and was also the third largest spike during the entire collection period. **(e)**

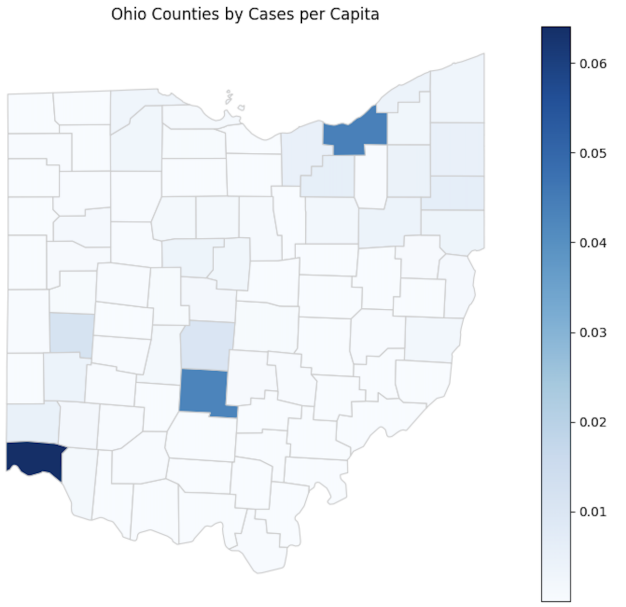


Fig. 3.

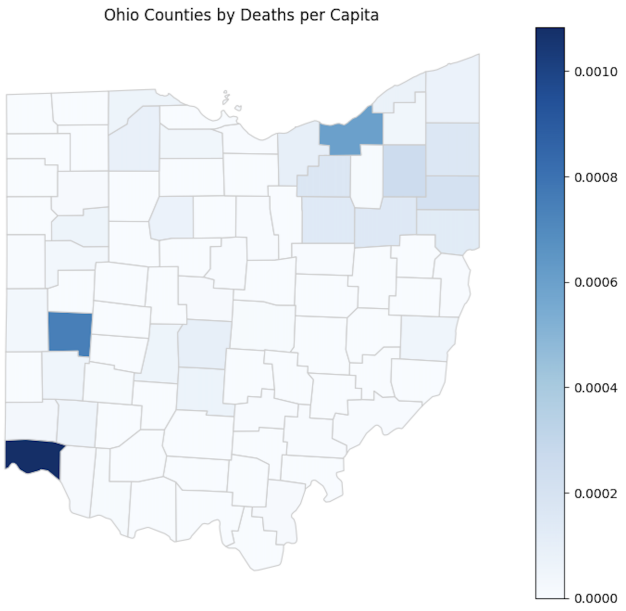


Fig. 4.

## III. METHODS

We experimented with several different kinds of models, including Linear Regression, Support Vector Machines, Decision Trees, Neural Networks, Gradient Boosting and Ensembling. Ultimately the best result was acheived with sci-kit learn's GradientBoostingRegressor and Bayesian Parameter Optimization.

### *A. Preprocessing*

Due to most of the variables being similarity or awareness scores, there wasn't a lot of preproccessing work; especially since the Gradient Boosted Models we ended up using do not

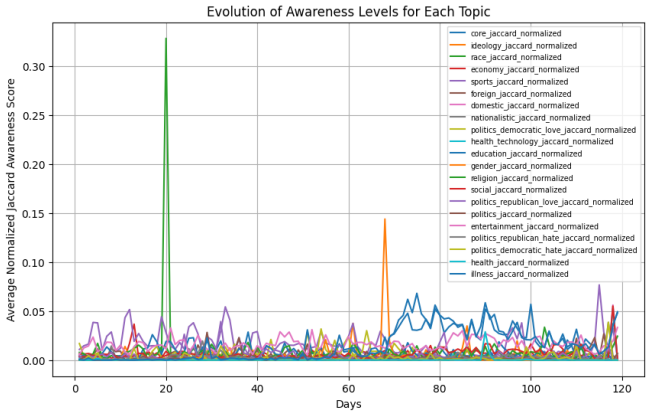


Fig. 5.

require standardization. The only base columns we changed were making "date_index_converted" a numeric variable, and One Hot Encoding counties. We also merged any outside data by county. We then isolated the "cases" column as the y variable and the rest as the X.

### B. Early Attempts

We first fitted a basic Linear Regression as a benchmark, which scored an $R^2$ of around 0.51. We then moved on to decision trees, using sklearn's RandomForestRegressor and ExtraTreesRegressor. These preformed better, ranging from 0.70 to 0.80, indicating that decision tree based models were well suited to the data. In particular, ExtraTreesRegressor was able to acheive scores of 0.95 and over on certain splits of the training data, although it seemed to overfit and did far worse on the actual test. This overfitting would become a recurring theme due to the 30-70 train-test split. We also tried Neural Networks and Stacked Generalization (creating a meta-model from other models), but these proved unfruitful. We also tried a Support Vector Machine, but quickly gave up when it performed worse than Linear Regression.

### C. Gradient Boosting

We then decided to try Gradient Boosting models as then next step from decision trees. These are ensembling models, which means they combine other models. Specifically, they combine several "weak learner" decision trees into a final prediction. Some of the ones we tried were XGBoost, LightGBM, CatGBM, and Sci-kit's Gradient Boosting Regressor. Initially we settled on XGBoost as the best and spent a long time trying to optimize it, plateuing at an $R^2$ of about 0.875. However, we later found out that the Sci-kit model, despite being slower and somewhat less complex, performed better with a final $R^2$ of 0.90152.

### D. Outside Datasets

We attempted to use several outside sources of data, including information about flu cases, age, education, population density, and covid vaccination. However, the only one that improved the model was county vaccination rates[1], despite being from 2023. These improved the XGBoost model score by about 0.01, and the Sci-kit model score by about 0.03.

### E. Parameter Optimization

We found the XGBoostRegressor model to be difficult to tune using automation, so manual tuning acheived the best results. However, when moving on the the GradientBoostingRegressor, we found a Bayesian Parameter Optimization to be very effective. Using Scikit-Optimize's BayesSearchCV, we where able to iteratively sample parameters more purposefully than a random or grid search. In addition, it is useful for continuous parameters like learning rate or n-estimators. We first optimized the parameters setting n_estimators to a low value, then n_estimators with the other parameters it generated.

The optimization provided us with an optimal number of estimators and learning rate. However, with the prior knowledge that models were prone to overfitting, we decided to manually decrease subsample from 1 to 0.9, and max depth from 4 to 3. These changes dramatically improved the model by increasing the randomization of trees created and decreasing tree depth, or how much the model molds itself to the training data. These were the final parameters: learning_rate=0.252991836, max_depth=3, n_estimators=373, random_state=0, subsample=0.9.

## IV. RESULTS

### A. Results and Difficulties

As previously stated, the final $R^2$ score of the Gradient Boosted Regressor on the Kaggle site was 0.90152.

When examining the feature importances of the model, deaths, age brackets, vaccination rates, county, and race all were high. Deaths in particular was by far the best predictor of cases. Certain counties such as Pickaway and Marion were also easily distinguished. Out of the similarity values, core and illness were the highest, with economy and politics after.

The biggest difficulty we faced was overfitting. With a 30-70 split, even cross validation was sometimes not perfectly representative of the test data and much trial and error was needed to improve accuracy.

### B. Next Steps

At this point, we believe the model has been optimized fairly well; a further increase would most likely come from a better model, feature engineering, or an outside variable that is highly predictive of cases.

One area to explore might be a better optimization of XGBoost. According to most literature it tends to outperform GradientBoostingRegressor (and is much faster), so it is possible that we were just unable to optimize it properly. In this regard, the sci-kit model has fewer paramters and is thus easier to get a better result out of.

[1]1. Source: https://data.cincinnati.com/covid-19-vaccine-tracker/ohio/39/