**Predicting Economic Recovery During a Pandemic**

**Adam Curry | DSC630 | Bellevue University**

**Executive Summary**

The economic impacts of the COVID-19 respiratory illness have been devastating for countries across the world. While deaths and infections should be the chief focus of the illness, the livelihoods and economic impacts from the virus have also been overwhelming. Countries across the world have done everything from widespread economic shutdowns to laissez-faire government response. The United States’ approach has been unique to the rest of the world, as federalism allowed each state to take a different approach in dealing with the pandemic. The government response of certain states, Google mobility, and the state’s economic impacts were the focus of this analysis.

This analysis chose the following states/territories as a result of their disparate geographies and their relatively wide range of government intervention: California, New York, Florida, Washington D.C., Hawaii, and Nebraska [4]. Using predictive analytic techniques and data from Google, Affinity, Womply, Paychex (for middle class - $27,000 - $60,000), and a government stringency index derived by Oxford university, this analysis found that the stringent government response to the pandemic and Google movement influenced the state’s economic outcomes [4] [5] [6]. Understanding that the government response often suppressed mobility, it makes intuitive sense that the stringency levels would correlate with the mobility data. When combined, these data elements predicted the economic impacts with relatively high confidence. In addition, the economic data elements for the various states allowed this analysis to forecast 60 days into the future, to see how well a state’s economy will recover following the month of August 2020. The findings yielded: the less stringent the response, the better the economic performance.

**Introduction/background of the problem**

The SARS-CoV-2 virus that causes the COVID-19 respiratory illness has plagued the world in 2020 and has had massive impacts to the lives and livelihoods of people across the world [1]. As of today, the total death toll in the United States is 240,265 and as of September, the country lost 10.7 million jobs [2] [3]. Federalism allowed each state to take different measures in their approach to the pandemic, which seems to impact their overall economic performance. For example, New York’s GDP is down 12.1% from quarter four 2019 to quarter two 2020 whereas South Dakota is down 8.7%. This paper quantified how a state’s pandemic responses impacted their economy and made predictions on their economic future using features that theoretically contribute to a growing economy. Using predictive analytics (regression, PCA, K-Means, autoregression), the final analysis predicted an economic index as well as the future economic wellbeing of six unique states, based on the state’s pandemic response and Google mobility.

The metrics used in this analysis include a Google mobility, Affinity middle class debit/credit card spending, Womply merchant behavior, Paychex, Intuit, Earnin, and Kronos employment data of middleclass families ($27,000 - $60,000), and a government stringency index derived by Oxford university [4] [5] [6]. These relatively diverse sources of data provide insight into the movement of the state’s citizens, their spending habits, how well small business are performing, and how stringent the state’s government response was to the pandemic. The middle class was chosen for the analysis because, intuitively, they would need to travel to work more often than a professional who can work from home. For example, a bus driver or store clerk would need to go to their place of employment to work whereas a marketing professional or data scientist could work from their home office. The baseline for all these metrics (minus the Stringency Index) was January 4th – 31st. This means that a number above or below the baseline of zero is an indicator of how well or poor the specific metric performed, or put another way anything below a zero is performing worse than January.

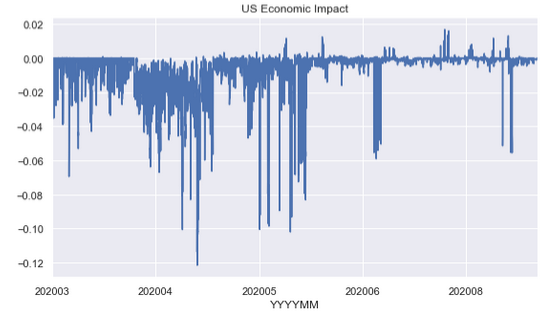
In order to quantify the economic impacts, the card spending, merchant behavior, and employment data were combined to derive an economic index. This was obtained by simply multiplying the three variables together to form the derived economic metric, econ\_derived. This economic index was used as a barometer of how a state performed between the months of March and August 2020. The data were limited to these months as March was when the pandemic “start date” occurred in the United States and August would contain “seasoned” data, meaning the data shouldn’t change since it is now three months old.

**Methods**

The project followed as closely as possible to the CRISP-DM process, except for a true “technical deployment” stage. This project won’t see a traditional production deployment, as it is limited to the confines of this course. However, the project followed the other stage. In fact, it jumped back and forth between data understanding, data preparation, modeling, and evaluation, and went through several iterations. This paper along with the presentation of the code and analysis will be considered the deployment stage for the CRISP-DM model. All analysis was done using a combination of Python and R programming languages.

To prepare the data, every data source needed to be cleansed and combined to form a final matrix style dataset. Cleaning the data included updating state abbreviations and adding datetime intervals as these would be considered primary keys for the joins made on the datasets. In order to do this with relative ease and to ensure the steps didn’t have to be repeated, the data was loaded to an SQLite database. From here, the data could be easily loaded back to the notebook without having to rerun the prior steps.

To create the derived metric, three features that appeared to mirror economic outcome were simply multiplied together. The general idea behind the derived economic metric was to include middle income spending and middle-income employment and overall small business revenue, because if the baseline of these three variables is greater than zero, then it could be a good economic sign. This would seem to indicate that when middle class employment is up along with spending then small business revenue is also performing well. The image below shows this metric’s performance throughout the year. The graph makes intuitive sense, given what we know about the pandemic timeline (i.e. April is when most of the government response to the pandemic occurred). Its important to note, some economies tend to not perform as well as others. This should be considered when interpreting the final analysis.



The z-score of all the metrics were defined to visually see how far certain observations fell outside of the mean. Using anything that was less than or equal to two standard deviations, and didn’t include the month of April, showed nine states that had poor performing records. April was omitted as nearly all states were performing poorly during this time. Visually seeing that states underperformed outside of April showed me that the analysis could proceed, as states continued to underperform compared to other states.

The next step was to visualize the data. A correlation matrix showed that some items were highly correlated, but nothing reached above 90%, so I was safe to assume that multicollinearity would not be an issue with this analysis. Boxplots showed that there were several observations that are outside of the interquartile for every month observed. However, April stood out as the lowest performing metric on all four graphs, which makes intuitive sense, as this is when the initial government response occurred.

The data were then moved to R where a PCA and k-Means analysis could be conducted on the dataset. Using the inputs 'econ\_derived', 'StringencyIndexForDisplay', and 'gps\_retail\_and\_recreation', PCA was conducted to find the most impactful variables for the analysis. As a side note, gps\_retail\_and\_recreation was used by itself to as the sole Google mobility metric for this clustering algorithm, because the others could be influenced by the state’s geography. For example, transit will look entirely different in New York than in Nebraska. The purpose of this analysis is to find correlations in the government response in accordance with the economic outcome, so the other mobility variables were omitted from the k-means and PCA. The analysis used the average silhouette method to find the optimal number of clusters for the k-Means algorithm. In this scenario, there were a total of three clusters. These clusters were then transferred back to the original dataset and moved to the SQL database for additional analysis in Python. The purpose of these models was to develop a method of visualizing the disparity in government response and Google mobility. In other words, they were supplementary to the overall problem statement: states acted differently. The results are described further downstream and will offer clarity on the overall stringency index.

The next model used was a multi-linear regression analysis. The regression analysis’ objective was to see if predictions could be made on the target variable econ\_derived using Google mobility and the stringency index. The analysis removed the month of April, as nearly all states were underperforming during this time, which means the month itself was an outlier. Backward elimination was utilized on each of the six states to ensure values that were not statistically significant (p<0.05) were removed from the predictions.

The final model was an autoregression model. This type of model forecasts using a linear combination of past values of the target variable. In other words, the regression runs against itself with a set of specified lagged values. This model splits the data at a given point-in-time instead of the traditional 80/20 (70/30) rule. This is because the more recent records should influence the forecasting model more than the older records. Meaning, July should be more impactful than April. I put the model inside a function in order to easily run each state through the analysis with a different number of lags on three separate autoregression models. This also made it easier to run each of the six states through their own forecasts. I chose 14, 30, and 60-day lag methods to see how much the lag influenced the 60 day forecast.

**Results**

The results of the PCA revealed the two most impactful variables that would apply to the k-means analysis. One of the variables accounted for 73.7% of the variance and the other accounted for 19.4% of the variance. The k-means algorithm took the PCA results and revealed that three optimal clusters defined the data. These clusters spread out across the research period revealed that cluster 1 had the least stringent government response with the best performing economy, cluster 2 had the second most stringent government response but with a mixed performing economy, and cluster 3 had the most stringent response with the worst performing economy.

The multilinear regression analysis found that certain variables were good predictors for the derived economic target variable, while others were less impactful. For example, New York’s google transit stations variable was statistically significant where this variable was completely omitted from Nebraska’s population. This makes intuitive sense, as Nebraska is much more rural and less densely populated than New York. However, all six states showed that the stringency index was statistically significant (even with April removed). The regression models were evaluated and showed an R2 for New York was .8, California .698, Nebraska .758, DC .773, Hawaii .655, and Florida .645. The root mean square error was very low on all findings, ranging from .088 to .143, showing relatively strong predictions overall.

This autoregression model forecast looked 60 days into the future. Each of the six states was run through three separate forecasting models, one lagging at 14 days, one at 30 days, and one at 60 days. Nebraska’s mean square error was very low on a 60-day lag and showed a steady economic prediction at or above the zero baseline. New York’s forecast prediction performed best on a 30-day lag. The 60-day lag produced very erratic predictions, so much so, the prediction was omitted from the final graph as to visualize the other predictions easier. New York remained below the zero baseline. California’s forecasted prediction performed the best on a 60-day lag and remained below the zero economic baseline. Washington D.C. was like New York, so the 60-day lag was omitted. Washington D.C. had the worst economic forecast of all six states. Hawaii was very odd, in that it had very volatile economic performance throughout the year. The best forecasting model was the 14-day lag and it was below the zero baseline. Florida performed the best on a 60-day lag and remained at or just below the zero baseline. However, Florida’s predictions were volatile, while their economic performance was stable outside of April and May.

**Conclusion**

The conclusion drawn from the PCA and k-means analysis was that the government stringency index and Google retail mobility index influenced the economic index. A scatter plot comparing the derived economic index with a log scaled government stringency index showed cluster 1 didn’t have very many observations above or below the zero baseline, cluster 2 had several observations above and below the baseline and cluster 3 was well below the baseline. This seems to indicate that the stricter the government response and the less retail mobility, the more volatile and ultimately poorer the economy performs.

The regression analysis reinforced the k-means analysis, as the predictors applied in the multilinear regression models were statistically significant and were overall, good predictors of the economic index. Some states didn’t require all variables to make their predictions, but this also reinforces the reason why these six disparate geographical states were selected for these analyses. The conclusion that can be drawn from this analysis seems to offer additional insight to the hypothesis that the higher government stringency and decrease in Google mobility negatively influence the economic performance.

The autoregression models predicted that certain states will see an economic index that is at or above the zero baseline, while others will remain well below. Hawaii’s volatile economic performance could be a result of the high tourism and it is a destination location (even during the pandemic). This could explain why a 14-day lag performed the best, as any dates that were lagged beyond 14 days made erratic predictions. States that had less stringent or shorter lasting government stringency responses appear to have economies that were forecasted to stay at or above the zero baseline for the next 60 days. This conclusion can be drawn from states like Nebraska and Florida, both of which had mean stringency indexes below the other four states. However, California also had a lower stringency index compared to the other states but saw lower economic production. This could be a result of other Google mobility metrics having a heavier influence on the target variable. It could also be an indication that California contains a very diverse type of behaviors. Perhaps certain parts of the state tend were more mobile than others.

There are a couple of things to keep in mind when reviewing these findings. There could be other outside variables that influence the final result. For example, a Covid-19 vaccine could drastically alter the way citizens behave. Also, some economies tend to traditionally outperform others. This is evidenced in Hawaii’s results where the derived economic index was erratic, often above the zero baseline and below the baseline, meaning the seasonality of the data influences the results. However, it can be concluded that the predictors used in the k-means, multilinear regression, and the auto regression were good at predicting the target derived economic variable. The findings seem to indicate that the more stringent the government and the less mobile a state’s citizens, the more volatile and ultimately poorer the economy performs.

**Acknowledgments**

I’d like to acknowledge my classmates and professor’s candor on my analysis up to this point. Their feedback helped me deliver a more polished product. I’d also like to acknowledge my family for being patient with me during the development of this project. I spent several weekends isolated in my office and several late nights on this analysis, and they remained patient with me as I continued my analysis.

**References**

1 Sauer, L. (). Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

2 https://coronavirus.jhu.edu/map.html

3 Ettlinger, M. (20, October 2020). COVID-19 Economic Crisis: By State. Retrieved from https://carsey.unh.edu/COVID-19-Economic-Impact-By-State

4 Chetty, R., Friedman, J., Hendren N., Stepner M. (01, November 2020). The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. Retrieved from https://opportunityinsights.org/wp-content/uploads/2020/05/tracker\_paper.pdf

5 https://tracktherecovery.org

6 Hale, T. Angrist, N. Cameron-Blake, E. Hallas, L. Kira, B. Majumdar, S. Petherick, A. Phillips, T. Tatlow, H. Webste, S. (02, September 2020). Variation in government responses to COVID-19. Retrieved from https://www.bsg.ox.ac.uk/sites/default/files/2020-09/BSG-WP-2020-032-v7.0.pdf