https://doi.org/10.48009/4_iis_2020_187-192

ACCURACY OF SIMPLE FORECASTING METHODS IN PREDICTING COVID-19 INCIDENCE IN NEW YORK CITY

John C. Stewart, Robert Morris University, stewartj@rmu.edu
Diane A. Igoche, Robert Morris University, igoche@rmu.edu
Gary Alan Davis, Robert Morris University, davis@rmu.edu

ABSTRACT

Traditional forecasting methods have been used for several years and, and therefore, have become pervasive in a variety of industries and fields of study. With the recent COVID-19 pandemic, as with any area of uncertainty, it has been difficult to predict the incidence rates for future periods. The importance of forecasting cannot be overstated, when planning for the level of medical and first responder resources necessary to adequately treat afflicted patients. This study compares the accuracies of some common, traditional forecasting methods to the actual incidence rates, in order to determine which methodology contributes the most useful level of incidence predictability.

Keywords: Forecasting, Prediction, COVID-19, Analytics

INTRODUCTION

Simple forecasting methods have been taught in introductory statistics courses (Lind et al., 2008), and have been used to assist in the prediction of demand and other components for future time periods (Brown, 1956). Various methods of forecasting have proven to be more effective than others for different patterns, trends, cycles, and extent of noise reduction in data (Chambers, et al., 1971). Recently, as with most uncertain patterns in data, there has been an urgent need to determine the trend in virus incidence, in order to be prepared for the number of patients needing critical treatment. However, as with any dataset and forecasting method, noise is a pervasive element that can lead to inherent error in the outcome or results. The objective of any forecasting application is to reduce the amount of error, and thereby, increase the accuracy in predicting the next time period. This study evaluated the accuracy of some of the traditional forecasting methods to determine their ability to predict the incidence of COVID-19 in New York City, during the months of the initial rise of incidence rates.

Theory and Application of Forecasting

Forecasting uses historical data to make predictive estimates that have been used in different sectors, notably supply chain management (Brown, 1959; Box, et al., 1976), healthcare, and retail. As a result, forecasting in supply chain management has seen very innovative applications. This increased use of forecasting within supply chain management can be attributed to the increase in computing power and the advent of software programs that allow for rapid calculation of forecasting results. Forecasting use cases in business also include the use of historical data to predict customer behavior in response to the release of new products and services.

Forecasting has been applied to scenarios outside of those previously mentioned. Since forecasting can be both statistical and based on human judgment (Farrow et al., 2017), there have been positive outcomes when forecasting methods are used in decision making for human services. Decision makers can readily translate the results of forecasting models into disaster-recovery techniques. For example, weather forecasting can be used to predict and track storms. In a similar manner, forecasting can also be used in healthcare to mitigate, manage, and even prevent the spread of disease. As seen during the recent COVID-19 outbreak, nations were able to conduct capacity planning for health facility and personnel needs and use forecast methods for measured re-openings.

Health forecasting dates to the time of Hippocrates of Kos (460 BC), when forecasting methods were used to determine the occurrence of certain diseases of that period (Soyiri & Reidpath, 2013). Hippocrates employed both statistical and human judgement from his expertise as a Physician. Considerations for forecasting during an epidemic or pandemic include the utilization of real-time data, which can present unique challenges. Up until 2014, there were challenges surrounding disease forecasting because the forecast results were difficult to implement. The Centers for Disease Control and Prevention, and the Division of Vector-Borne Diseases launched the Epidemic Prediction Initiative (EPI) to address the challenges that arose from forecasting diseases. The EPI effort made real-time forecasting or "nowcasting" (i.e., predicting the near future, recent past, and present) possible, which identifies patterns during the yearly influenza seasons. The EPI initiative also helped develop forecasting models for the COVID-19 pandemic (Giannone, Reichlin, & Small, 2008).

Time-series data are used in health forecasting, and specifically, within disease forecasting. *Trend*, *Seasonality* and *Cyclicality* are all components of time-series analysis that are considered when developing forecasting models. *Trend* is considered to be the long-term change in a time-series that is not affected by components in the data. *Seasonality* is a repetitive occurrence within the data. Finally, *Cyclicality* is the extent to which incident data points are influenced by overall patterns (Soyiri & Reidpath, 2012). There are also unexplained movements of time-series that must be factored into forecasting models. These unexplained components also affect the type of forecasting method that is used in a study.

Forecasting methods have been used to study patterns and trends during notable epidemics, including the Avian Influenza epidemic, and the Ebola epidemic. The models developed for these diseases were used to predict the trajectory of the epidemics, and to predict the benefits of various intervention strategies to slow the spread of the diseases. Performance measures are used for model selection, including the *Mean Square Error* (MSE), *Mean Absolute Deviation* (MAD), and *Mean Absolute Percentage Error* (MAPE). The "best" model will have the lowest MSE, MAD, and/or MAPE values. The basic forecasting methodologies include the following: *Moving Average*, *Exponential Smoothing*, *Trend-Adjusted Exponential Smoothing*, and *Time-Series* (Yule, 1909; Brown, 1956; Brockwell, 2016). Each forecasting methodology is further defined below:

> *Moving Average Forecasting Technique*: This method is widely used for forecasting long-term trends. The Moving Average Technique calculates the average of the most recent observations in a dataset.
>
> *Exponential Smoothing Technique*: This method is adequate for series that have no trend or seasonality. This method forecasts future values by using a weighted average of previous values within data or a series. The exponential smoother in this method assumes that the level will be constant.
>
> *Trend Adjusted Exponential Smoothing Technique:* This method is also known as *Holt's Double Exponential Smoothing method*. Unlike simple exponential smoothing, this method works with series that have trend and level without seasonality. There are other methods that can handle series with trend and seasonality, however this method is popular because it is computationally inexpensive.

RESEARCH METHODOLOGY

Our research focused on the following research area: How accurate are traditional, quantitative forecasting methods at predicting the incidence of COVID-19 in a major metropolitan location, like New York City (NYC)? Data from this location was chosen for a few reasons. First, availability. NYC was one of the first areas to have complete data due to the incidence prior to other areas of the country. Second this data showed a complete curve: a rise in incidence, a peak, and a downward trend. The objective was to determine if a curve could be fit to the entire dataset.

The specific research questions to be answered are as follows:

> Q_1: Are traditional forecasting methods capable of offering any level of accurate insight into the incidence of COVID-19 in New York City?
>
> Q_2: Which traditional quantitative forecasting method was most accurate in fitting a curve to the incidence of COVID-19 in New York City?

To answer these research questions, the study evaluated several different forecasting methods: Moving Average, Exponential Smoothing, Trend-Adjusted Exponential Smoothing, and Time-Series. All methods were applied against the COVID-19 daily incidence data from February 29 to May 18, 2020 in New York City. These data were gathered from the city of New York website ("Coronavirus Disease 2019," 2020).

The forecasting methods were performed in *Microsoft Excel*. *Frontline Solver* was used to optimize the smoothing constant(s) in exponential smoothing. In addition, higher order polynomials were evaluated in attempts to fit a curve to the data. For each forecasting method the differences between the actual and the forecast values were determined and the lowest MAD was calculated.

RESULTS

Using the moving average technique, a 7-day moving average yielded the lowest Mean Absolute Deviation (MAD) of 289.001. The exponential smoothing method resulted in a MAD of 467.15. Using exponential smoothing with a trend, the optimal MAD (using alpha = 1 and beta = 0.0891) was 463.96. The study evaluated second through fifth order polynomial time series methods, with the 3rd order polynomial yielding the lowest MAD at 729.39. **Table 1:** Mean Absolute Deviation (MAD) for Each Forecasting Method shows the comparison of MAD results for each forecasting method tested during the study.

Table 1: Mean Absolute Deviation (MAD) for Each Forecasting Method.

Method	Specifics	MAD
Moving Average	7-day	209.001
Exponential Smoothing	α= 1.0	467.15

Exponential Smoothing w/ Trend	α= 1.0
	β= 0.0891
		463.59

Time Series (polynomial)	5 4 3 2	8393.35 766.74 729.39 830.59

In order to answer the first research question, "Q$_1$: Are traditional forecasting methods capable of offering any level of accurate insight into the incidence of COVID-19 in New York City?" The data was evaluated against various standard forecasting methods. While there is an inherent limit to forecasting beyond the next time period, the curves outlined by some of the methods fit the data fairly well. Applying these methods to similar data and similar defensive mechanisms (i.e., ubiquitous social distancing) may give insight into the potential outline of the incidence curve and thus the timeline for needed resources to meet the strain on the medical infrastructure.

To address the second research question, "Q$_2$: Which traditional quantitative forecasting method was most accurate in fitting a curve to the incidence of COVID-19 in New York City?," MAD was used as the criteria to compare the accuracies of each method, in order to determine the most accurate method. According to our results, the 7-day moving average showed the lowest MAD and, therefore, proved to be the most accurate of the methods tested.

Figure 1: MAD of 7-Day Moving Average presents the 7-day moving average forecast compared to the actual daily incidence of COVID-19. As depicted in **Figure 1**, and discussed previously, the 7-day moving average forecast fits the actual COVID-19 incidence data with the least amount of error.

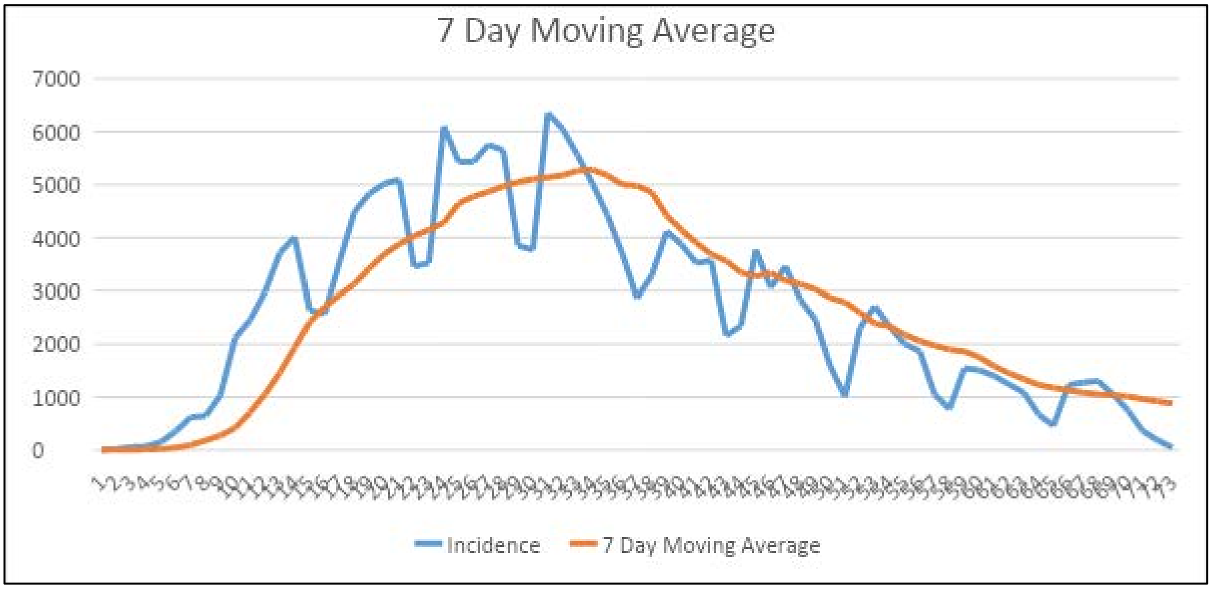

Figure 1: MAD of 7-Day Moving Average

This study evaluates higher order time series methods, and evaluates the data against 2nd, 3rd, 4th, and 5th order polynomials. As discussed previously, the 3rd order polynomial was the best fit of the four higher order methods (MAD = 729.39). **Figure 2: 3rd Order Polynomial Forecast** shows the fit of the 3rd order polynomial curve compared to the actual daily incidence of COVID-19 cases. However, the accuracy of the 3rd order polynomial was substantially *less* than the simple moving average, or exponential smoothing methods tested in this study.

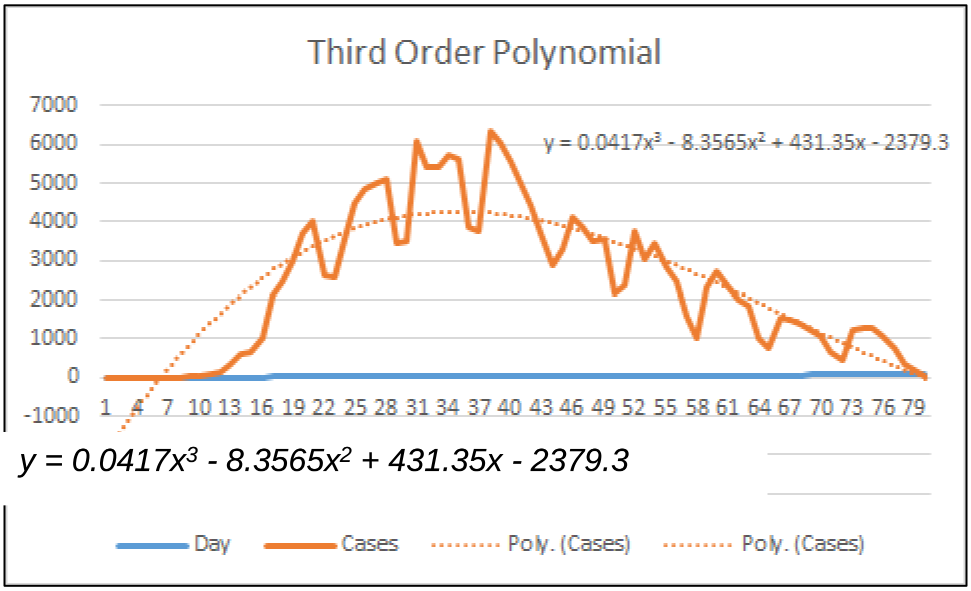

Figure 2: 3rd Order Polynomial Forecast

DISCUSSION AND CONCLUSION

This study explored the use of traditional forecasting methods applied to the incidence of the COVID-19 virus in New York City. Based on the criteria of comparing the forecasting outcomes using MAD, it was found that the lowest MAD, and, therefore the "best" fit of the predicted to actual incidence is the *7-day moving average*. Given that forecasting has value in determining only a few time periods in the future, the 7-day moving average may have some limited value in determining the slowing of the rate of incidence, and may give some insight into how soon an apex might be reached in the incidence rate. The forecasting in this study assumes that all of the factors that existed previously (i.e., similar human behavior) will be present in any future attempts to apply these forecasts to the incidence of COVID-19.

The findings of the current study are consistent with other studies that employed the 7-day moving average for effective forecast modeling. For example, researchers Yu, Hsu, and Yang (2019) were able to create a forecasting model to predict daily electricity usage in Taiwan. The researchers selected the 7-day moving average for their forecasting model because it yielded more accurate forecasts than previous models. In their 2014 study, Wong and Lai employed a 7-day moving average of daily weather reporting data to predict emergency ambulance usage in Hong Kong. The researchers' model provided reasonably accurate forecasts of daily ambulance calls that were one to seven days in advance. As in the current study, many past researchers used the 7-day moving average to forecast the progression of disease. An Iowa State University (2012) study constructed a model to forecast the progression of avian-based diseases. The ISU model used the 7-day moving average to forecast the transmission of disease, as well as identify and predict known disease triggers. Finally, Soyiri and Reidpath (2013) used data from over 70,000 respiratory deaths that occurred over a 13-year period in New York City to develop a forecasting model. As in the past-cited examples, the authors were able to use the 7-day moving average to construct an accurate predictive model of respiratory deaths that result from weather and air quality factors. As in the current study, the 7-day moving average has demonstrated efficacy in the construction of models across a diverse array of forecasting applications.

There are some limitations in this study that should be noted. The reliance on simple, traditional methods could be a possible limitation of the usefulness of the results from this study. To minimize this limitation, future studies involving COVID-19 incidence data could utilize more advanced (or sophisticated) forecasting methods to determine if more accurate results could be obtained. Finally, this analysis was limited to the relatively short timeframe from February 29 to May 18, 2020. The analysis in this study could be repeated at some point in the future, when additional data are available, and the COVID-19 incidence rates are more definitive.

REFERENCES

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. New York, McGraw Hill.

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Chambers, J.C., K.M. Satinder, and D.D. Smith. How to choose the right forecasting techniques. *Harvard Business Review* (July - August 1971), 45-74.

Coronavirus Disease 2019 (Covid 19). *City of New York*. Retrieved May 18, 2020, Retrieved from https://www1.nyc.gov/site/doh/covid/covid-19-main.page.

Farrow, D. C., Brooks, L. C., Hyun, S., Tibshirani, R. J., Burke, D. S., & Rosenfeld, R. (2017). A human judgment approach to epidemiological forecasting. *PLoS Computational Biology, 13*(3).

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: the real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665-676.

Iowa State University, Avian research: data on avian research described by researchers at Iowa State University. (2012, Feb 03). *Health & Medicine Week.*

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2012). *Statistical techniques in business & economics*. New York, NY: McGraw-Hill/Irwin.

Soyiri, I. N., & Reidpath, D. D. (2013). An overview of health forecasting. *Environmental Health and Preventive Medicine*, *18*(1), 1.

Soyiri, I. N., & Reidpath, D. D. (2013). The use of quantile regression to forecast higher than expected respiratory deaths in a daily time series: A study of New York City data 1987-2000. *PLoS One*, *8*(10).

Wong, H., & Lai, P. (2014). Weather factors in the short-term forecasting of daily ambulance calls. *International Journal of Biometeorology, 58*(5), 669-78.

Yu, K. W., Hsu, C. H., & Yang, S. M. (2019). *A model integrating ARIMA and ANN with seasonal and periodic characteristics for forecasting electricity load dynamics in a state*. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE).

Yule, G. U. (1909). The applications of the method of correlation to social and economic statistics. *Journal of the Royal Statistical Society*, *72*(4), 721-730.