United States COVID-19 Hospital Resource Forecasting

By
Jerad Williams

Agenda

- Problem statement
- Hypothesis
- Proposed approach
- Data wrangling steps
- Statistical data analysis
- Forecasting time-series with ARIMA
- Model diagnosis
- Predicting values
- Forecasting time-series with SARIMA
- Model diagnosis
- Predicting values

Problem

- During COVID-19 times, experts need to know the gap between expected and the existing resources.
- By knowing this beforehand, they can work on their preparation strategies
- By using time-series forecasting for predicting hospital bed use, need for intensive bed use and invasive ventilator use, we can achieve this goal.
- Data used is from IHME (Institute for Health Metrics and Evaluation.

Hypothesis

- NULL Hypothesis (H0): Input time series is stationary
- Alternate Hypothesis (H1): Input time series is non-stationary
- The confidence level will be at 95%.

Proposed approach

- IHME’s database has three columns for each dimension distributed by date
  - One represents mean
  - Second represents lower uncertainty bound
  - Third represents upper uncertainty bound
- Techniques used:
  - ARIMA
  - SARIMAX
- Evaluation strategies:
  - Comparison of observed data points and forecasts.
  - Usage of Mean Squared Error and Root Mean Squared Error
  - Usage of Q-Q plots
  - Checking the Akaike Information Criterion (AIC) value.

Data Wrangling Steps

- Mean aggregation of mean columns by date
- Summation of lower and upper uncertainty bound columns by date
- Missing values imputation using mean if the distribution of corresponding column is normal, median if distribution is skewed.
- Alternatively, MICE algo was also used for missing value imputation
- For outlier manipulation, z-scores of all the values of a column are found and any z-scores above 3 or -3 are replaced with 3 and -3 respectively.
- Alternatively, usage of box plot analysis can be done for outlier manipulation. All values above 96th percentile or below 4th percentile were replaced with 96th percentile value and 4th percentile value respectively.

Statistical Data Analysis

- Extraction of data frame into time-series
  - Converted all dimensions into time-series
  - Obtained plots of all time-series and saw their patterns.
  - Based on plots, we could infer that need of resources was 0 till Feb 2020 and began increasing after that.
- Decomposing the time-series
  - The four components obtained after time-series decomposition were:
    - Observed
    - Trend
    - Seasonal
    - Residual
  - High amount of seasonality could be observed from the data

Statistical Data Analysis

- Checking Stationarity
  - If a series satisfies following conditions, it is said to be stationary
    - Constant mean
    - Constant variance
    - An autocovariance that doesn’t depend on time
  - Procedures used:
    - Plotting rolling statistics
    - Dickey-fuller statistics
  - To plot rolling averages, simple moving averages and exponential moving averages were used.
  - These plots were used to see which plot gave constant mean and variance and better Dickey-fuller statistics.
  - Subtracted the obtained series from the original series to obtain the stationary series

Statistical Data Analysis

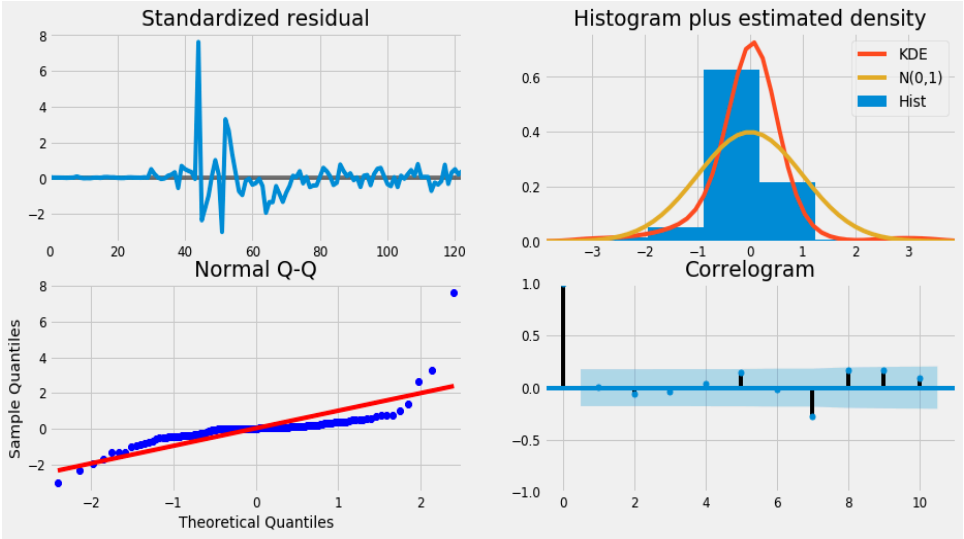
- Checking stationarity
  - Checked for trend and seasonality components and if present, subtracted them from the original series since presence of both of them makes the series non-stationary
  - ARIMA converts non-stationary series to stationary, hence, no transformations are needed
- Plotting ACF and PACF
  - Used ACF to find autocorrelation function with q lag
  - Used PACF to find partial autocorrelation function with p lag

Forecasting time-series with ARIMA

- ARIMA consist of 2 terms
  - AR
  - MA
- AR corresponds to the difference value. This is today's value minus yesterday's value or value-on-value change.
- MA corresponds to moving average terms.
- Three integers (p, d, q) are typically used to parametrize ARIMA models.
  - p: number of autoregressive terms (AR order)
  - d: number of non-seasonal differences (differencing order)
  - q: number of moving-average terms (MA order)
- Best (p,d,q) are (2,2,2) with AIC = 1684.39

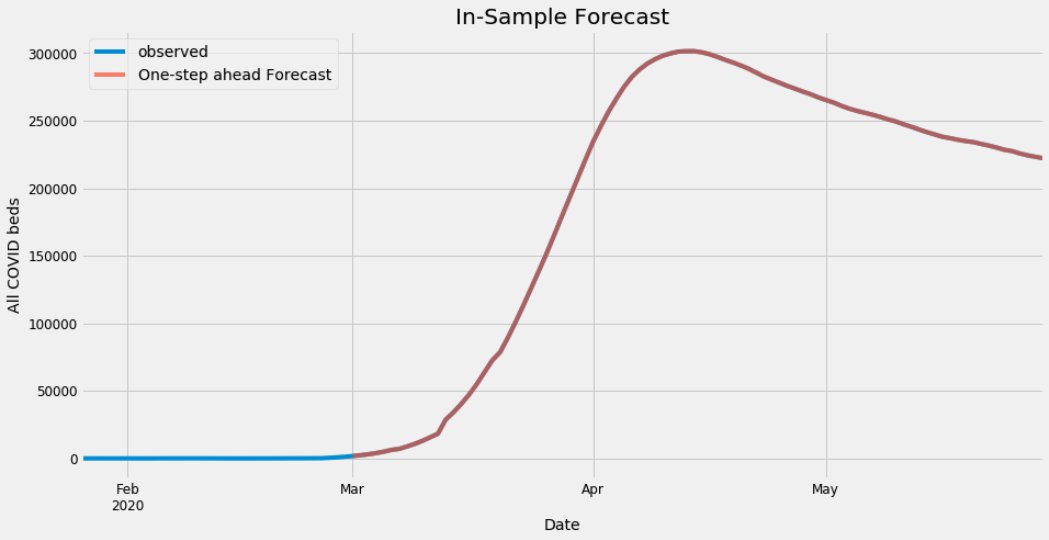
Model diagnosis

- The Standardized Residual corresponds to random noise.
- The Histogram suggests that the residuals have normal distribution.
- The Normal Q-Q plot suggests that the theoretical and sample quantiles are very close to each other. The more close the sample quantiles to the line, the more normal their distribution will be.
- The Correlogram has all autocorrelations of different lags in between the blue shaded area.



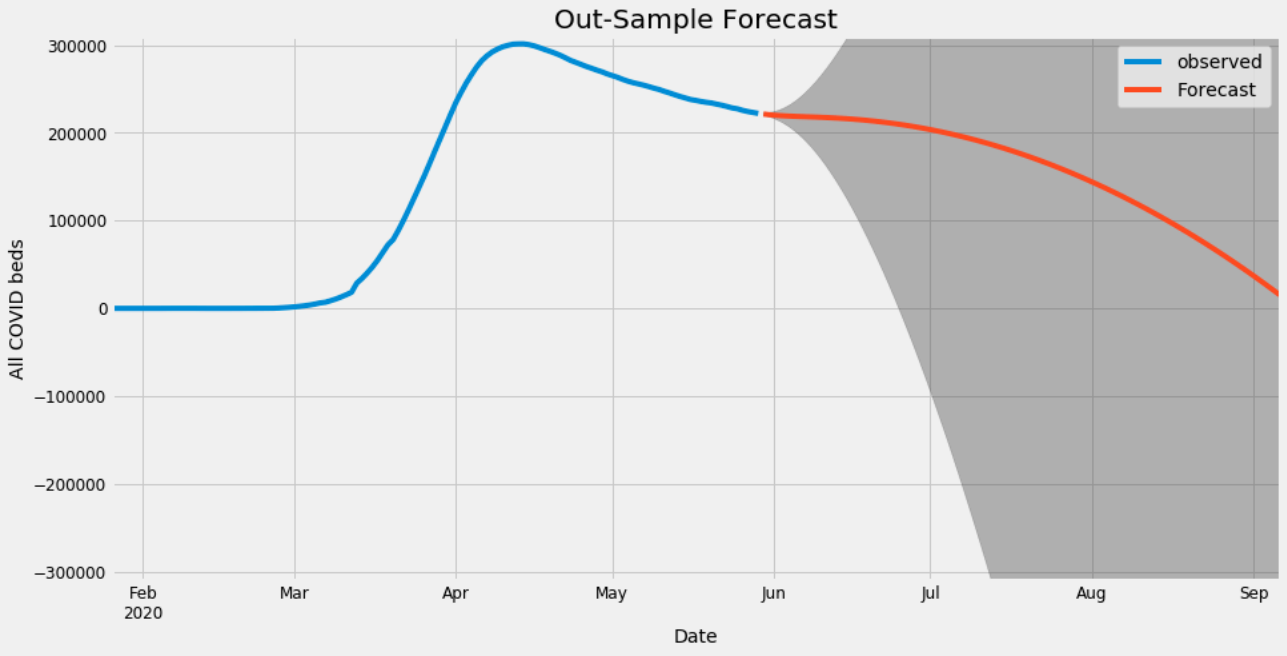
Predicting values

- Model validation after 1st March 2020 was carried out
- The predicted values align really well with the observed values.
- Mean squared error was also very close to zero.



Predicting values

We predicted the requirements of the number of COVID-19 beds, ICU beds, and ventilators after 29th May 2020 with 95% confidence interval.

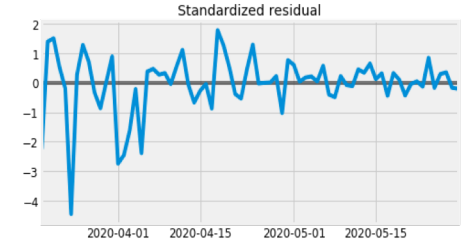


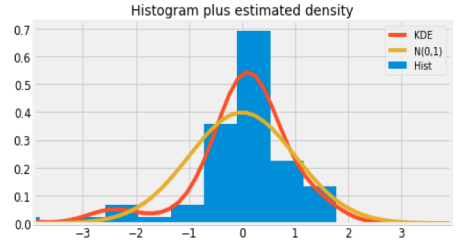
Forecasting time-series using SARIMA

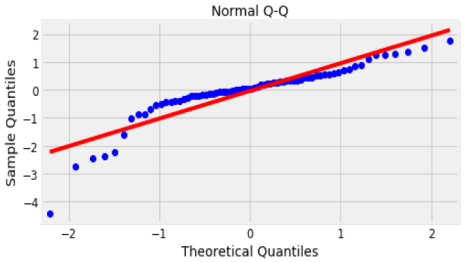
- SARIMA stands for Seasonal Autoregressive Integrated Moving Average
- Seasonal Elements
  - There are four seasonal elements that are not part of ARIMA that must be configured; they are:
    - P: Seasonal autoregressive order.
    - D: Seasonal difference order.
    - Q: Seasonal moving average order.
    - m: The number of time steps for a single seasonal period
- Best (p,d,q) and (P,D,Q,m) found are SARIMA (1, 2, 2)x(0, 2, 2, 12)12 - AIC:1045.545

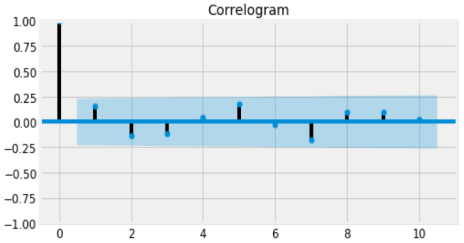
Model diagnosis

- The standardized residual corresponding to random noise.
- The histogram suggests that the residuals have normal distribution.
- The Normal Q-Q plot suggests that the theoretical and sample quantiles are very close to each other. The more close the sample quantiles to the line, the more normal their distribution will be.
- The correlogram has all autocorrelation of different lags in between the blue shaded area.



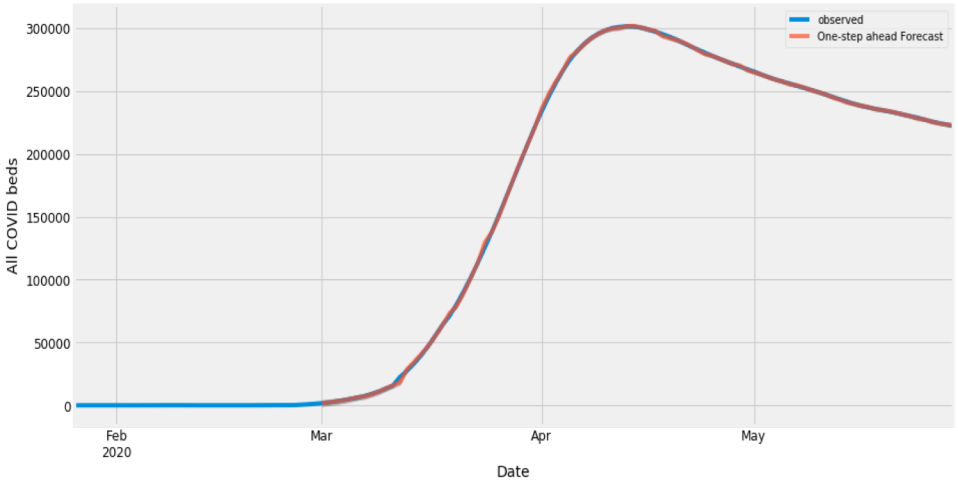






Predicting values

- Validated the model by plotting the predicted values on the data after the 1st of March 2020.
- Overall, our forecasts align with the true values very well.
- However, we calculated the Mean-Squared Error 916.37 which is bigger than the ARIMA model, which implies that the ARIMA model is better than SARIMA.



Predicting values

We predicted the requirements of the number of the bed after 29th May 2020 with 95% confidence interval.