# Web Traffic Forecasting

Time Series Analysis


**Shree Karimkolikuzhiyil**; Programming; Statistics, M.S. (Distance); shreejesh@tamu.edu
**Jingcheng Xia**; Computations; Computer Science, B.S. (On-Campus); sixtyfour64@tamu.edu
**Jackson Smith**; Analysis; Statistics, M.S. (Distance); jackson.t.smith@tamu.edu
**Samuel Burge**; Writing; Statistics, M.S. (Distance); samuelburge@tamu.edu
**Max Kutschinski**; Theory; Statistics, M.S. (Distance); mwk556@tamu.edu


## Introduction and Motivation

The objective of this analysis is to forecast daily unique visitors to an academic website over a 30-day horizon. Predicting website traffic allows IT departments to manage project throughput and prioritize maintenance and enhancements to website functionality and effectively allocate web server resources. Web traffic is also a key indicator of customer growth and expansion, as well as sustaining recurring customers and ingrained growth. The details provided by web traffic throughput reports contain many metrics, including page loads, returning visitors, and unique visits, each of which conveys a different picture and set of information for an organization. As well, having a picture of expected throughput and confirming (or denying) expectations with reality allows a business to understand unexpected growth and/or unexpected decay in business development.

The data contains five years of daily time series data of user visits. There are four features in the data set, which include daily counts for the number of page loads, first-time visitors, returning visitors, and unique visitors.[1] An initial plot of the data shows strong seasonality and volatility, but doesn't appear to have any discernible trend or cyclical behavior. An explanation for this could be due to the nature of the website. Students would likely be the largest share of users for a website of this nature, and the seasonality seems associated with the academic calendar typically seen at academic institutions.

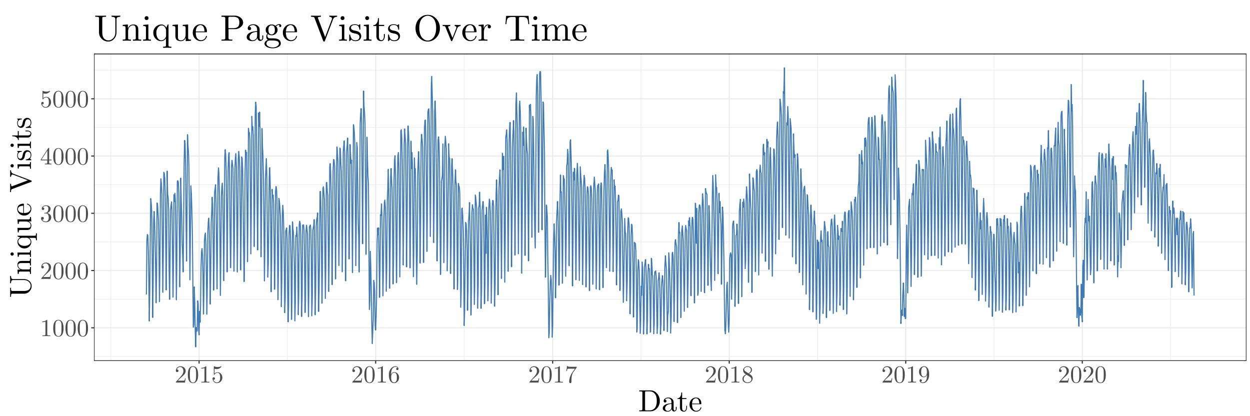



Figure 1

[1] A visit is defined as a stream of hits on one or more pages on the site on a given day by the same user within a 6-hour window, identified by the IP address of the specific device. Returning visitors are identified through allowed cookies on a user's device, and the total number of returning and first-time visitors is, by definition, the number of unique visitors.

## Modeling

### SARIMA

Stationarity is a common assumption underlying many time series procedures. As such, it is important to assess the level of stationarity prior to modeling and make the appropriate adjustments if necessary.

Shumway and Stoffer [2019] describe a stationary time series as one whose properties do not depend on the time at which the series is observed. More specifically,

(i) *the mean value function* $\mu_t = E(x_t)$ *is constant and does not depend on time* $t$

(ii) *the autocovariance function* $\gamma(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$ *depends on times* $s$ *and* $t$ *only though their lagged difference.*

The strong seasonality that is apparent in Figure 1 is indicative of non-stationarity, since seasonality will affect the value of the time series at different times. Seasonality is defined as a recurring pattern at a fixed and known frequency based on the time of the year, week, or day. Figures 2 and 3 aim to identify the type of seasonality present in the data. Figure 2 plots a subset of the first several weeks and indicates that there exists weekly seasonality, whereas Figure 3 uses locally weighted scatterplot smoothing (Lowess) to emphasize the inherent annual seasonal behavior.

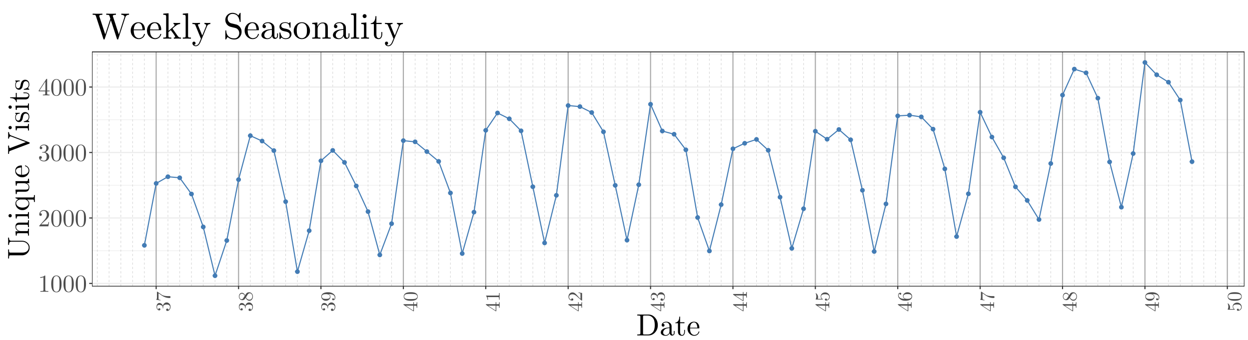



Figure 2: Sample of weekly page visits

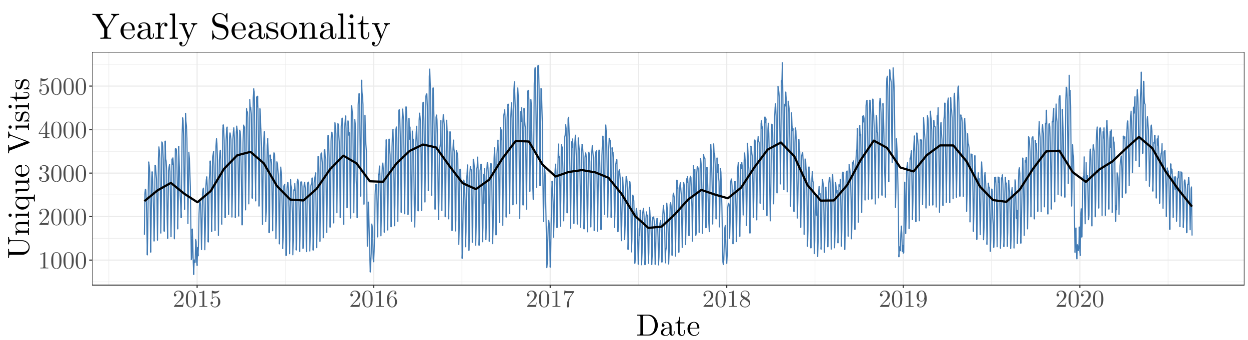



Figure 3: Smoothing via Lowess

A popular approach in addressing non-stationarity due to seasonality is to eliminate these effects via seasonal differencing. The seasonal difference of a time series is the series of changes from one season to the next, which is defined as follows:

$$\nabla_s x_t = x_t - x_{t-s} \tag{1}$$

One challenge with the unique visits, however, is the complex seasonality. Multiple seasonal patterns exist within the time series, and the family of SARIMA(p,d,q)(P,D,Q)[s] models only allow for a single seasonal difference. In an attempt to handle the complex seasonality, we performed a two-step seasonal differencing approach by taking the annual difference of the time series, then taking the weekly difference of the transformed time series from the previous step. This can be more clearly shown as

$$x_t^* = \nabla_7 \nabla_{365} x_t = (1 - B^7)(1 - B^{365})x_t = (x_t - x_{t-7}) - (x_{t-365} - x_{t-372}) \tag{2}$$

where B is the backshift operator. Time plots of the aforementioned transformation steps series at different stages are displayed in Figure 4. The differenced series appears to be stationary with a constant mean and variance.

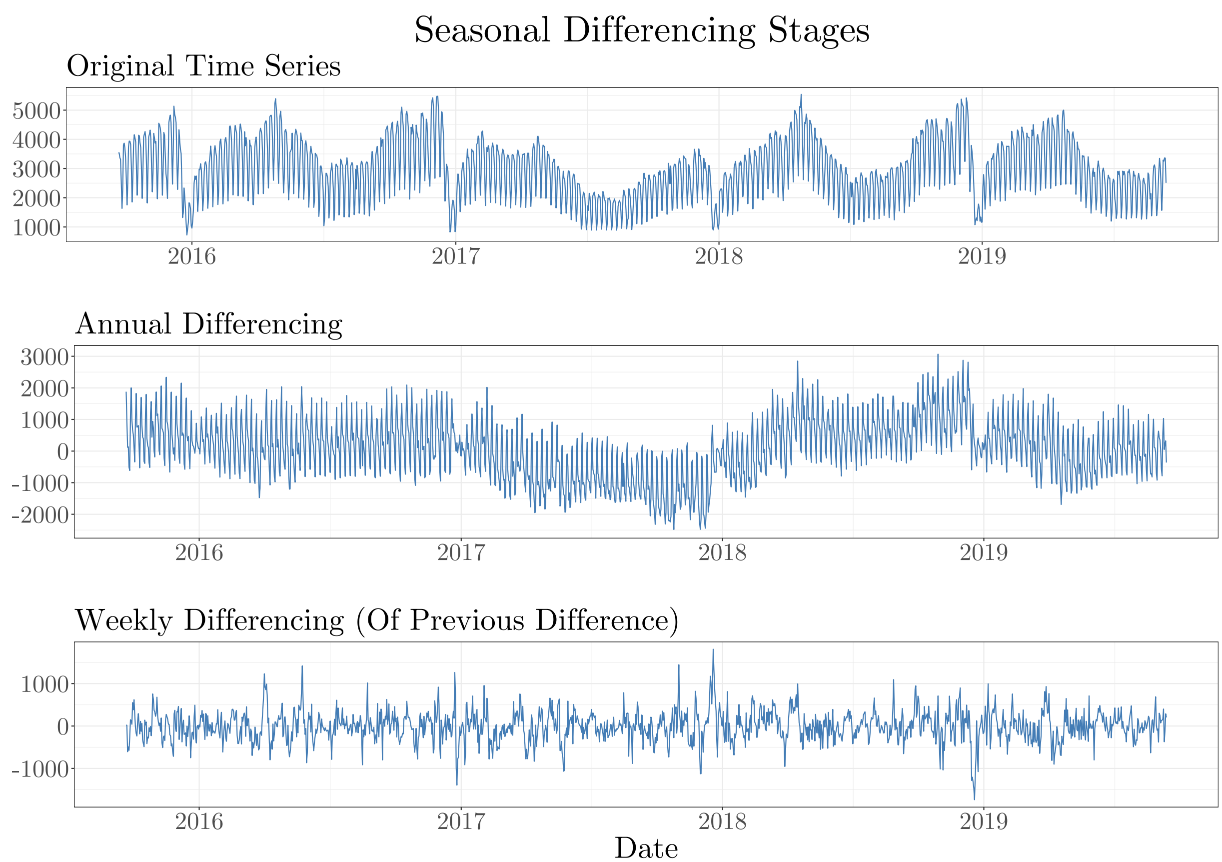



Figure 4: Time plots of differenced series

The ACF and PACF of the differenced series $x_t^*$ are displayed in Figure 5. Neither the ACF nor the PACF seems to cut off after a certain lag, which would be indicative of an AR or MA process. Rather, both of them appear to tail off over time, making it difficult to determine specific orders for the family of ARMA(p,q) models defined in Eq(3) as:

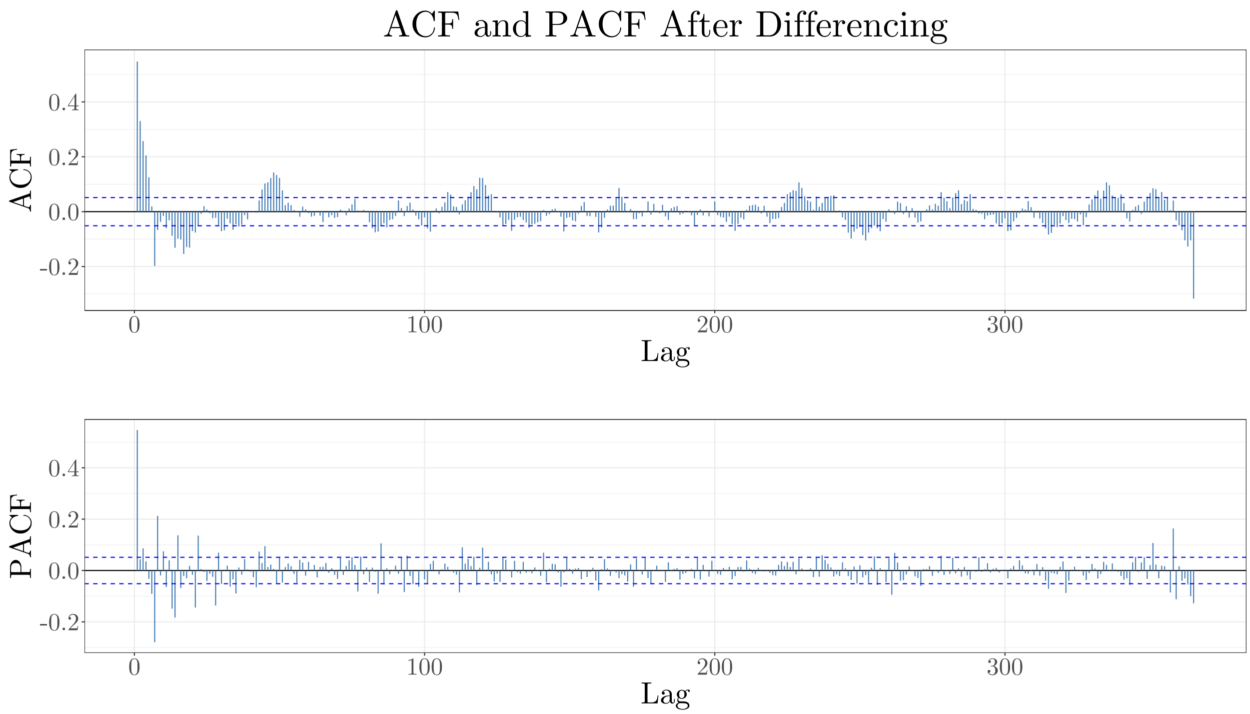



Figure 5: ACF and PACF

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 \omega_{t-1} + \cdots + \theta_p \omega_{t-q} \tag{3}$$

where $\phi_p \neq 0, \theta_p \neq 0, \sigma_w^2 > 0$, and the model is causal and invertible.

We opted to fit a range of ARMA(p,q) models with small orders, with the final model selected based the forecasting accuracy of the models on a test or hold-out set since our primary goal for this analysis is forecasting. The model selection criterion for the fitted ARMA models is shown in the table below.

```
## Series: diff
## Model: ARIMA(1,0,2)(0,0,2)[7]
##
## Coefficients:
##          ar1      ma1      ma2     sma1     sma2
##       0.9429  -0.3374  -0.2588  -0.6548  -0.1751
## s.e.  0.0162   0.0300   0.0299   0.0276   0.0268
##
## sigma^2 estimated as 66504:  log likelihood=-10137.51
## AIC=20287.03   AICc=20287.09   BIC=20318.72
```

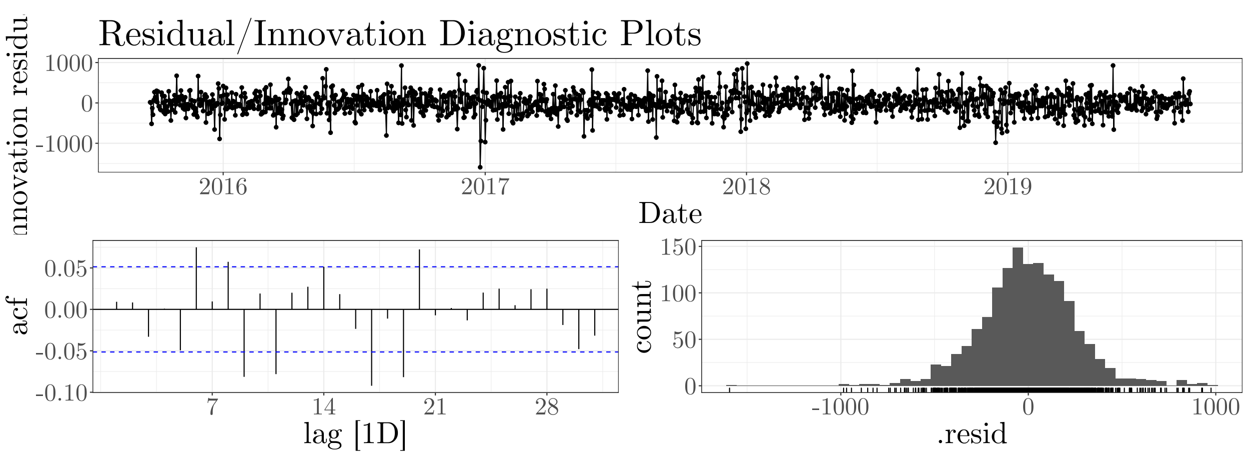
```
## # A tibble: 3 x 9
##   .model  sigma2 log_lik    AIC   AICc    BIC ar_roots  ma_roots    sigma
##   <chr>    <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>    <list>      <dbl>
## 1 test   202693. -10947. 21906. 21906. 21938. <cpl [1]> <cpl [9]>   450.
## 2 test2   66747. -10091. 20196. 20196. 20233. <cpl [1]> <cpl [16]>  258.
## 3 test3   66293. -10086. 20188. 20188. 20230. <cpl [8]> <cpl [16]>  257.
```

| Model | Innov. Std. Error | Log-Likelihood | AIC | AICc | BIC |
|---|---|---|---|---|---|
| SARMA(1,2)(2,2)[7] | 257.88 | -10137.51 | 20287.03 | 20287.09 | 20318.72 |
| ARMA(1,2) | 300.81 | -10359.03 | 20726.06 | 20726.09 | 20747.19 |
| ARMA(2,2) | 300.85 | -10358.72 | 20727.44 | 20727.48 | 20753.85 |
| ARMA(2,1) | 301.20 | -10360.94 | 20729.88 | 20729.91 | 20751.01 |
| ARMA(1,1) | 301.44 | -10362.60 | 20731.20 | 20731.21 | 20747.04 |
| ARMA(1,0) | 301.77 | -10364.66 | 20733.32 | 20733.32 | 20743.88 |
| ARMA(0,1) | 314.88 | -10426.44 | 20856.88 | 20856.89 | 20867.44 |

The parameters of the ARMA(1,2) model are estimated via conditional least squares and are displayed in Eq(4).

$$\hat{x}_t = 0.7164_{(0.0414)} x_{t-1} + \omega_t - 0.1953_{(0.0489)} \omega_{t-1} - 0.089_{(0.0325)} \omega_{t-2} \tag{4}$$

Figure 6 displays of plot of the residuals of the fitted ARMA(1,2) model. Initially, the residuals seem to behave like white noise, being centered around zero with a constant variance. However, further analysis of autocorrelation plot and formal testing using the Box-Ljung test indicate that the innovations are correlated (i.e., they are not white noise). The autocorrelation in the innovations does appear small for most lags, given how similar the various models are this is likely the best fit we can obtain from the ARMA(p,q) modelling procedure.



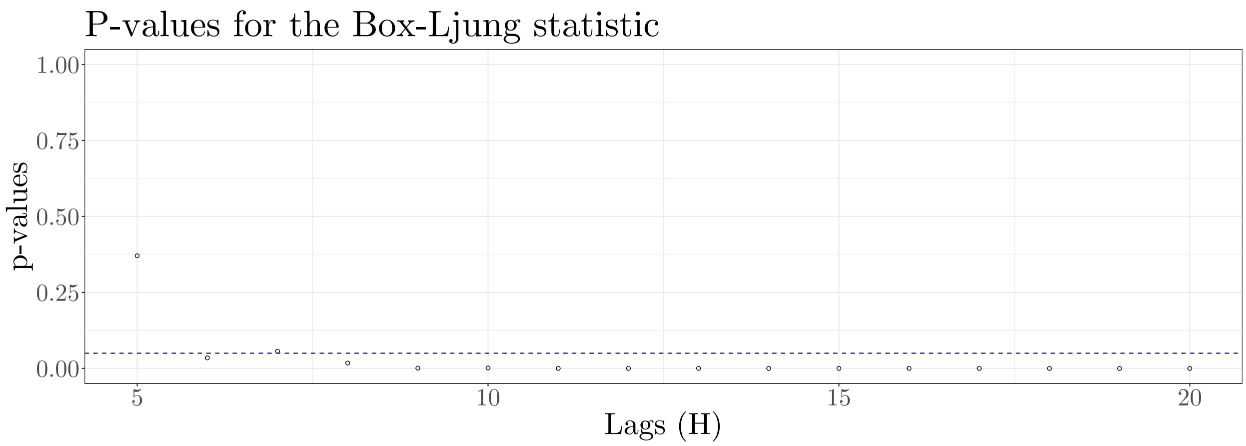
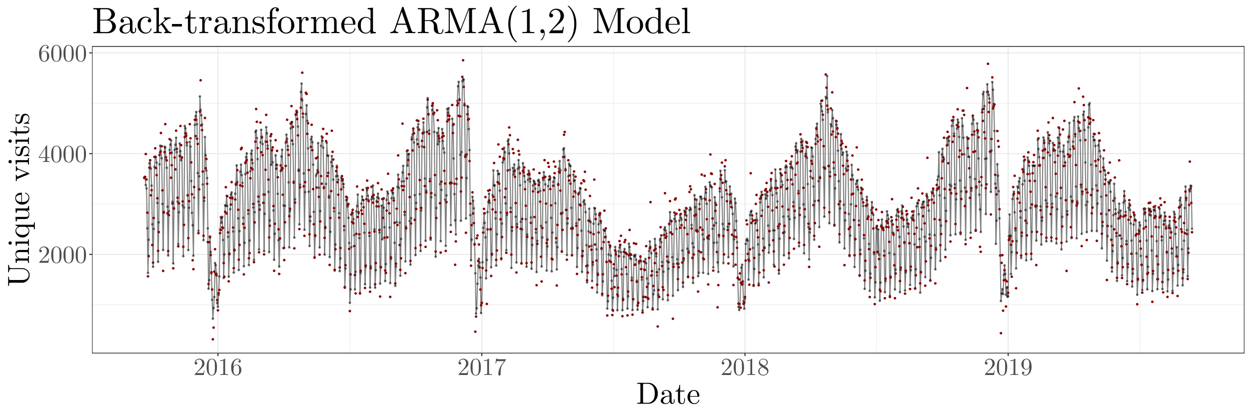



Table 1 contains the model's polynomial roots. Since they appear to be different from each other by a reasonable margin, we can conclude that there is no parameter redundancy in the model.

Table 2: Roots of polynomials

| AR Roots | MA Roots |
|---|---|
| 1.06+0i | 0.64+0.80i |
| NA | -0.92+0.44i |
| NA | -0.23-1.00i |
| NA | 0.64-0.80i |
| NA | -0.23+1.00i |
| NA | -0.92-0.44i |
| NA | -0.78-0.98i |
| NA | 1.02+0.00i |
| NA | 0.28+1.22i |
| NA | -1.25+0.00i |
| NA | 0.28-1.22i |
| NA | 1.13+0.54i |
| NA | -0.78+0.98i |
| NA | 1.13-0.54i |
| NA | 1.42+0.00i |
| NA | -2.72+0.00i |

The fitted values of the ARMA(1,2) model are plotted against the actual values of the seasonally differenced series $x_t^*$ (Figure 7) and the.



Fitted model shown by black line and the actual values of the training set are shown in red.

### Prophet

Taylor and Letham [2018]

O'Hara-Wild [2020]

## Results

All the above models were trained on a training set, and the predictive accuracy was then evaluated on a test set. The training set consists of page visits starting from 2015-09-21 until 2019-09-13, and test set contains the data from 2019-09-14 to 2020-08-19, making up the last 341 observations of the data. Our primary evaluation metrics for model comparison are the root mean squared error (RMSE) and mean absolute error (MAE), which both have the advantage of being measured on the same scale as the data (i.e., the number of website visits). Using these are our primary accuracy measures gives us more interpretable results. For clarification, the RMSE and MAE are defined as

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

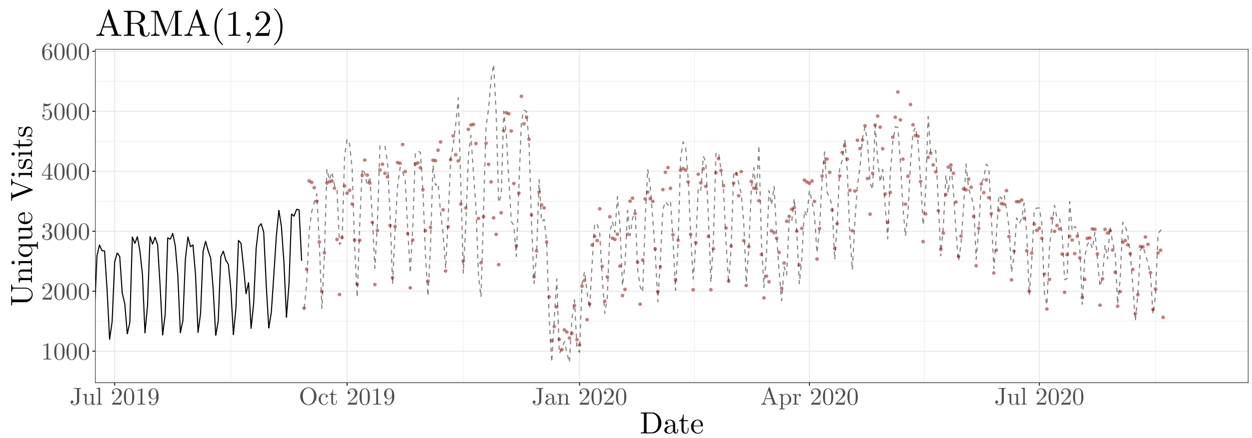
In addition to these accuracy measures, a common-sense baseline serves as a sanity check, and is often used as a benchmark for more advanced time series models. Given daily data with yearly seasonality, a common-sense baseline is to predict the number of unique visits at time t to be equal to the number of unique visits at t-365. In other words, a random walk model making a constant prediction with yearly seasonality, which is known as a seasonal naive model (Eq. X)

$$\hat{x}_t = x_{t-365}$$

The forecast accuracy for the ARMA(1,2) model shows the best performance on the test set,

Table 3: Accuracy measures from the test set performance of fitted models.

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | RMSSE |
|---|---|---|---|---|---|---|---|
| SARMA(1,2)(2,2)[7] | -1.25 | 430.61 | 302.12 | 99.33 | 100.46 | 0.71 | 0.77 |
| ARMA(2,2) | -3.93 | 432.13 | 302.79 | 99.45 | 99.45 | 0.71 | 0.77 |
| ARMA(1,2) | -3.91 | 432.15 | 302.81 | 99.46 | 99.46 | 0.71 | 0.77 |
| ARMA(2,1) | -3.92 | 432.14 | 302.83 | 99.43 | 99.50 | 0.71 | 0.77 |
| ARMA(1,1) | -3.85 | 432.16 | 302.87 | 99.45 | 99.55 | 0.71 | 0.77 |
| ARMA(1,0) | -3.63 | 432.31 | 302.99 | 99.56 | 99.61 | 0.71 | 0.77 |
| ARMA(0,1) | -3.03 | 432.71 | 303.49 | 99.87 | 99.87 | 0.71 | 0.78 |
| Prophet | 278.74 | 554.16 | 431.50 | 8.07 | 15.64 | 1.52 | 1.27 |
| Seasonal Naive | 9.59 | 676.70 | 498.06 | 53.16 | 536.76 | 1.17 | 1.21 |



## Conclusion

The d

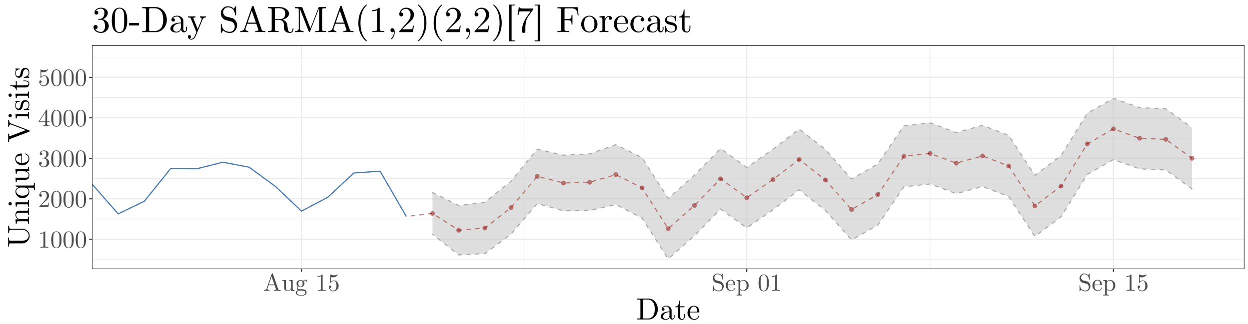



Table 4: 30-Day ARMA(1,2) Forecast.

| Date | Lower 95% CI | Forecast | Upper 95% CI |
|---|---|---|---|
| 2020-08-20 | 1123 | 1639 | 2155 |
| 2020-08-21 | 615 | 1224 | 1834 |
| 2020-08-22 | 646 | 1279 | 1912 |
| 2020-08-23 | 1129 | 1783 | 2437 |
| 2020-08-24 | 1884 | 2556 | 3227 |
| 2020-08-25 | 1704 | 2391 | 3077 |
| 2020-08-26 | 1709 | 2409 | 3109 |
| 2020-08-27 | 1859 | 2597 | 3336 |
| 2020-08-28 | 1525 | 2271 | 3017 |
| 2020-08-29 | 516 | 1262 | 2009 |
| 2020-08-30 | 1089 | 1835 | 2582 |
| 2020-08-31 | 1748 | 2495 | 3241 |
| 2020-09-01 | 1281 | 2027 | 2773 |
| 2020-09-02 | 1730 | 2476 | 3223 |
| 2020-09-03 | 2222 | 2972 | 3722 |
| 2020-09-04 | 1714 | 2466 | 3218 |
| 2020-09-05 | 985 | 1737 | 2490 |
| 2020-09-06 | 1351 | 2104 | 2857 |
| 2020-09-07 | 2298 | 3051 | 3805 |
| 2020-09-08 | 2367 | 3121 | 3875 |
| 2020-09-09 | 2127 | 2881 | 3635 |
| 2020-09-10 | 2305 | 3060 | 3814 |
| 2020-09-11 | 2056 | 2811 | 3566 |
| 2020-09-12 | 1070 | 1825 | 2580 |
| 2020-09-13 | 1558 | 2313 | 3069 |
| 2020-09-14 | 2601 | 3356 | 4112 |
| 2020-09-15 | 2970 | 3726 | 4481 |
| 2020-09-16 | 2739 | 3495 | 4251 |
| 2020-09-17 | 2712 | 3468 | 4224 |
| 2020-09-18 | 2244 | 3000 | 3755 |

# References


Mitchell O'Hara-Wild. *fable.prophet: Prophet Modelling Interface for 'fable'*, 2020. URL https://fable.tidyverts.org. R package version 0.1.0.

R.H. Shumway and D.S. Stoffer. *Time Series: A Data Analysis Approach Using R*. A Chapman & Hall book. CRC Press, Taylor & Francis Group, 2019. ISBN 9780367221096.

Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080. URL https://doi.org/10.1080/00031305.2017.1380080.