Web Traffic Forecasting

Time Series Analysis

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

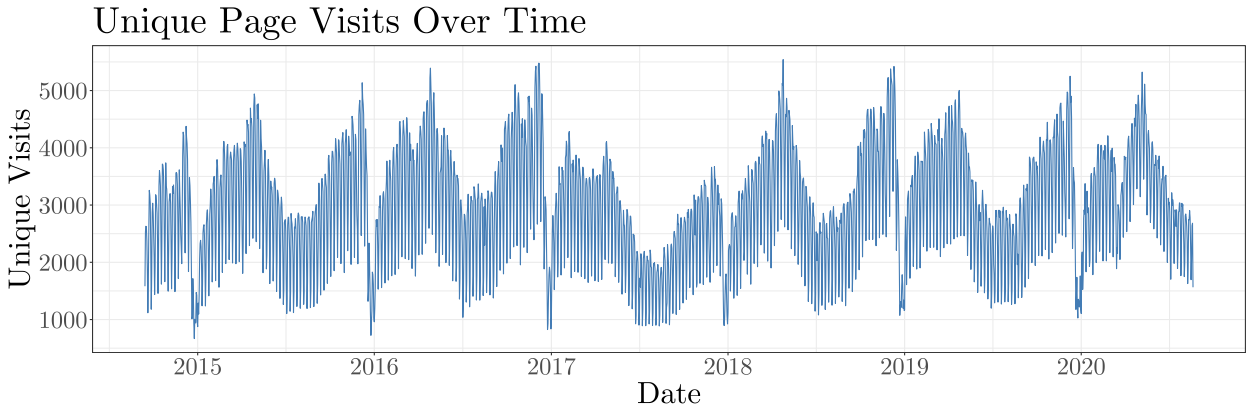


Figure 1

[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.

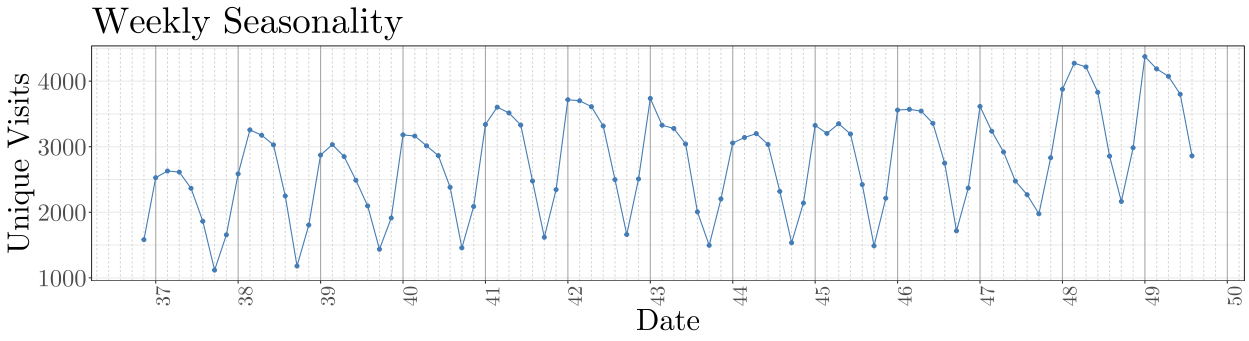


Figure 2: Sample of weekly page visits

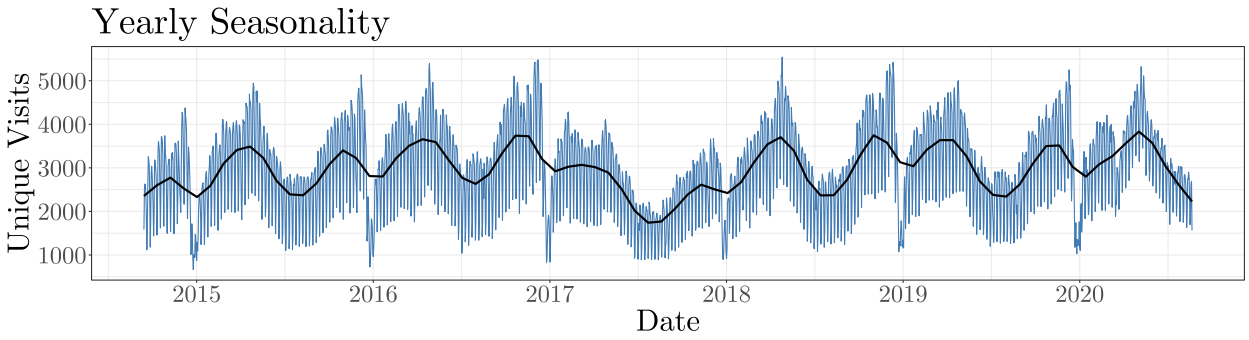


Figure 3: Smoothing via Lowess

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 (1).

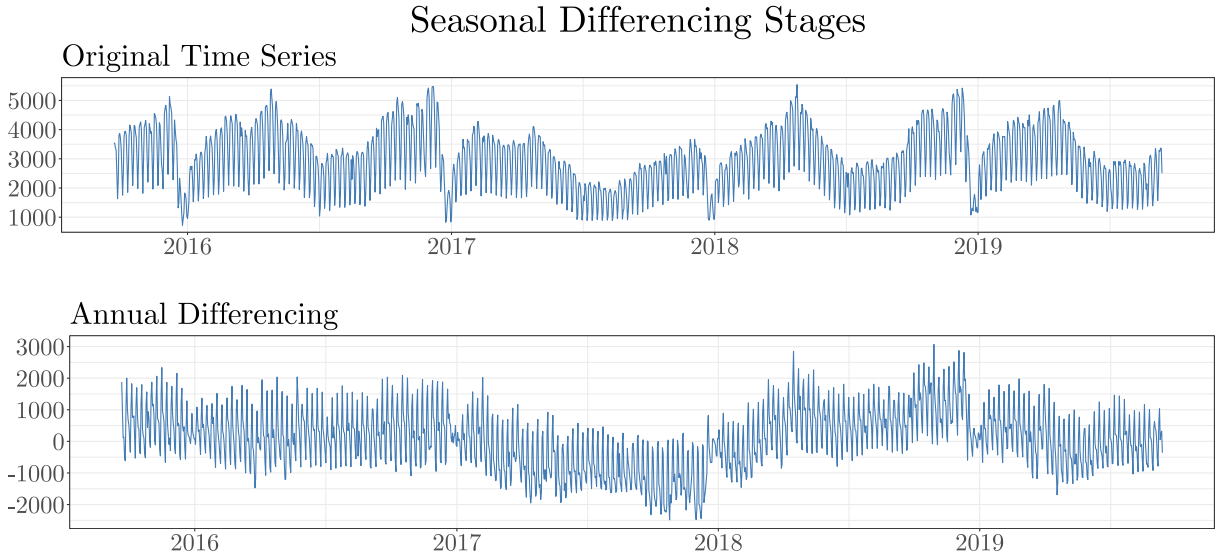
$$\nabla_s x_t = x_t - x_{t-s} \tag{1}$$

One challenge with the unique visits, however, is the complex seasonality. Multiple seasonal patterns exist within the time series, and the family of seasonal ARIMA(p,d,q)(P,D,Q)[s] models only allow for a single seasonal difference. In an attempt to handle the complex seasonality, and to coerce the data into a form we could use with time series packages for estimating the models, we performed a two-step seasonal differencing approach as displayed in (2) and (3) by first taking the annual difference of the time series, and then taking the weekly difference of the transformed time series from the previous step.

$$\tilde{x}_t = \nabla_{365} x_t = (1 - B^{365}) x_t = (x_t - x_{t-365}) \tag{2}$$

$$x_t^* = \nabla_7 \nabla_{365} x_t = (1 - B^7)(1 - B^{365}) x_t = (x_t - x_{t-7}) - (x_{t-365} - x_{t-372}) \tag{3}$$

where B is the backshift operator. Time plots of the aforementioned transformation steps are displayed in Figure 4. The series $\{\tilde{x}_t\}$ does not appear stationary, but the series $\{x_t^*\}$ appears to be stationary with a constant mean and variance. Our approach was to treat the annually differenced time series as the input time series, and fitting various $SARIMA(p,d,q)(P,1,Q)_s$ models where the seasonal lag is the weekly difference to in effect fit models to $\{x_t^*\}$.



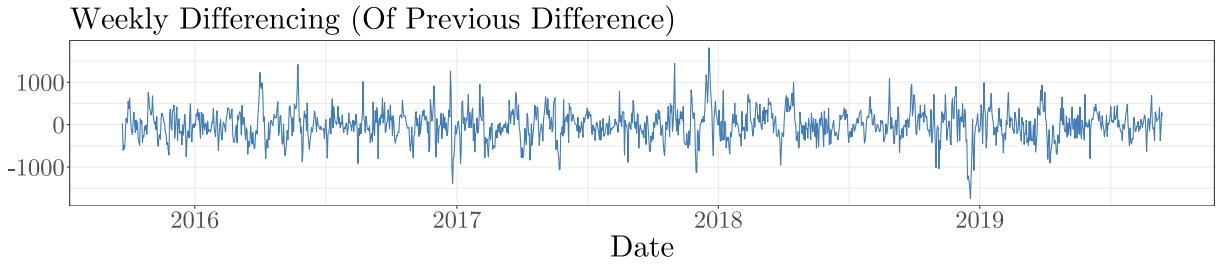


Figure 4: Time plots of differenced series

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 SARIMA$(p, d, q)(P, 1, Q)_7$ models defined in (4).

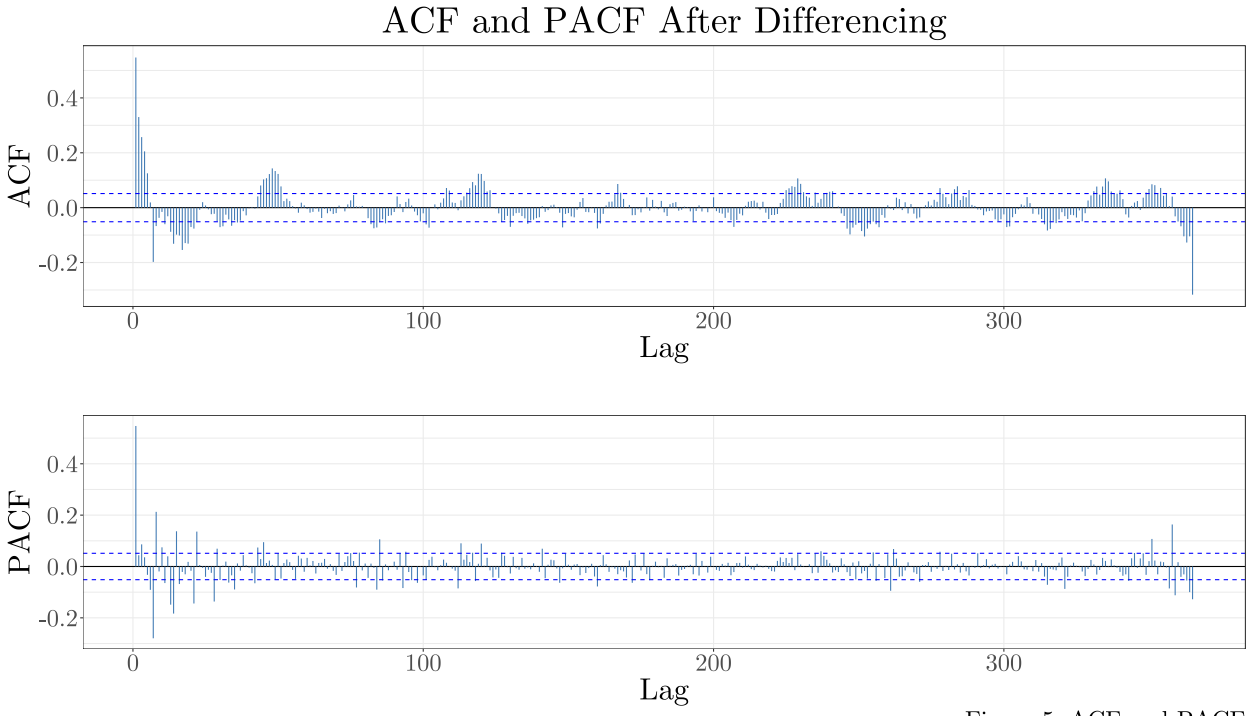


Figure 5: ACF and PACF

$$\tilde{x}_t = \alpha + \phi_1 \tilde{x}_{t-1} + \cdots + \phi_p \tilde{x}_{t-p} + w_t + \theta_1 \omega_{t-1} + \cdots + \theta_p \omega_{t-q} \tag{4}$$

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 SARIMA$(p, d, q)(P, D, Q)_7$ models with small orders, with the final model selected based on AIC and BIC. The model selection criterion for the fitted SARIMA models is shown in the Table 1 below. All of the proposed SARIMA models are fitted using the same order of differencing to ensure that their AIC and BIC are comparable. We decided to choose the SARIMA$(1, 0, 2)(1, 1, 2)_7$ model as it performed best in terms of AIC and BIC. Its parameters are estimated via maximum likelihood and are displayed in (5).

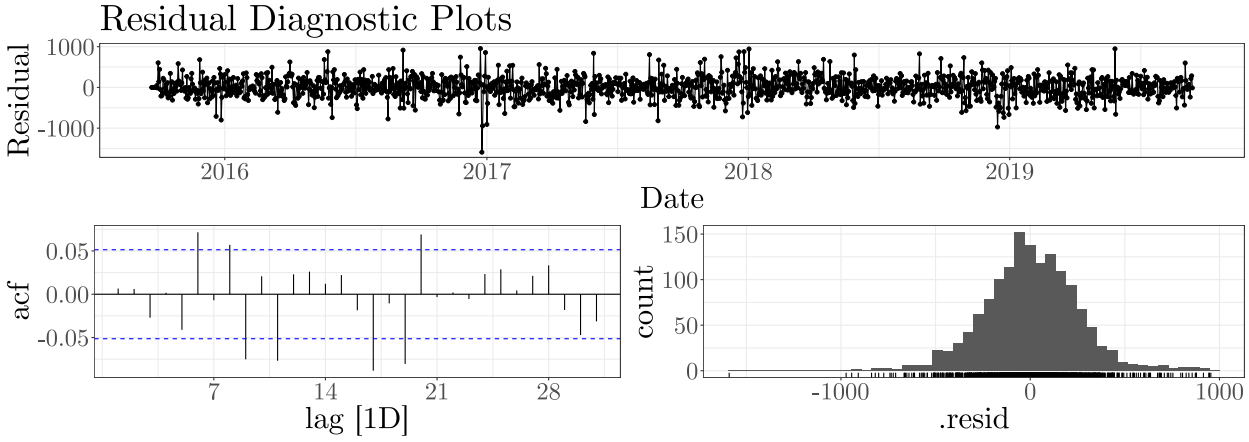
Table 1: Model estimation results.

Model	AIC	BIC
SARIMA(1,0,2)(1,1,2)[7]	20252.29	20289.23
SARIMA(1,0,2)(2,1,2)[7]	20253.21	20295.42
SARIMA(1,0,2)(0,1,2)[7]	20290.06	20321.72
SARIMA(1,0,2)(1,1,1)[7]	20481.26	20512.92
SARIMA(1,0,2)(0,1,1)[7]	20672.43	20698.81
SARIMA(1,0,0)(0,1,0)[7]	22041.21	22051.76
SARIMA(1,0,1)(0,1,0)[7]	22042.83	22058.66
SARIMA(1,0,2)(0,1,0)[7]	22044.43	22065.54
SARIMA(0,0,1)(0,1,0)[7]	22114.65	22125.21

$$\hat{\tilde{x}}_t = 0.94\tilde{x}_{t-1} + 0.23\tilde{x}_{t-7} + \omega_t - 0.33\omega_{t-1} - 0.25\omega_{t-2} - 1.87\omega_{t-7} - 0.87\omega_{t-14} \tag{5}$$

```
## Series: diff
## Model: ARIMA(1,0,2)(1,1,2)[7]
##
## Coefficients:
##          ar1      ma1      ma2   sar1     sma1    sma2
##       0.9381  -0.3345  -0.2502  0.234  -1.8659  0.8659
## s.e.  0.0168   0.0305   0.0300  0.034   0.0244  0.0242
##
## sigma^2 estimated as 66642:  log likelihood=-10119.15
## AIC=20252.29   AICc=20252.37   BIC=20289.23
```

Figure 6 displays of plot of the residuals of the fitted SARIMA$(1,0,2)(1,1,2)_7$ model. Initially, the residuals seem to behave like white noise, being centered around zero with a constant variance. However, further analysis of the autocorrelation plot and formal testing using the Box-Ljung test indicate that the residuals are correlated (i.e., they are not white noise). The autocorrelation in the residuals does appear small for most lags, given how similar the various models are this is likely the best fit we can obtain from the SARIMA$(p,d,q)(P,1,Q)_7$ modeling procedure.



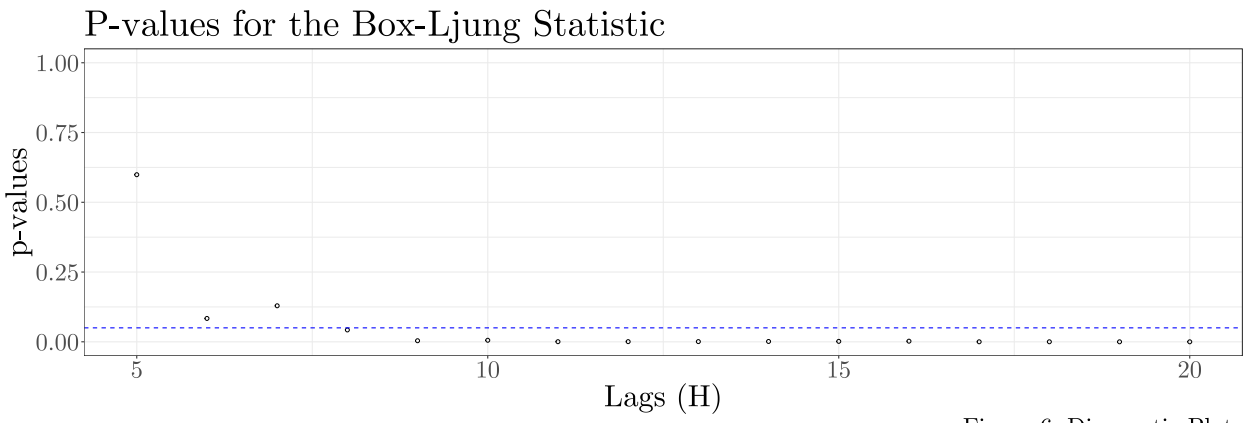


Figure 6: Diagnostic Plots

Figure 7 plots the inverse AR and MA roots of our SARIMA model. A causal invertible model should have all the roots outside the unit circle. Equivalently, the inverse roots should lie inside the unit circle (shown in red). Furthermore, there doesn't appear to be any parameter redundancy, since none of the roots are close to each other.

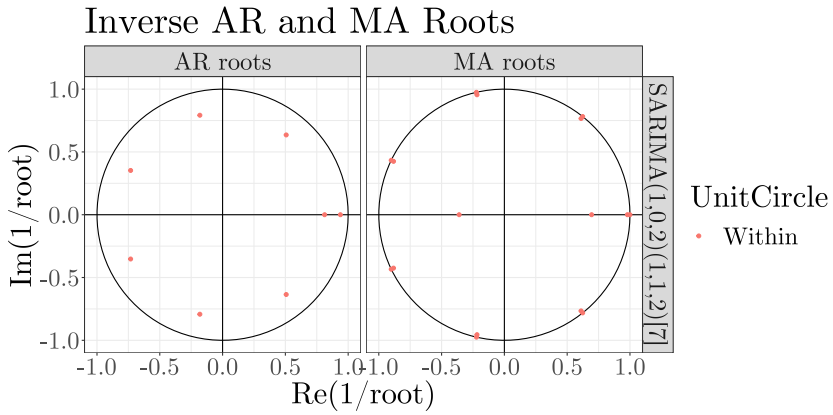


Figure 7: Inverse Roots

Fitted values of the SARIMA$(1,0,2)(1,1,2)_7$ model are transformed to the original scale in order to obtain a fitted plot as seen in Figure 8. The fit of the SARIMA$(1,0,2)(1,1,2)_7$ model (blue) is plotted on top of the number of unique visits (black).

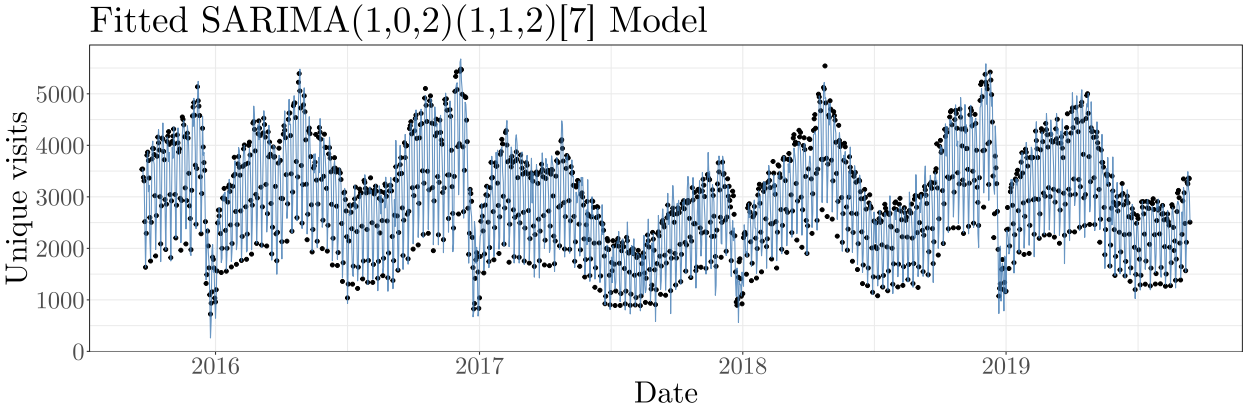


Figure 8: Fitted model (blue) vs Unique Visits (black) using training set

Hyndman and Athanasopoulos [2021] note that seasonal differencing of high order does not make a

lot of sense. Seasonal versions of ARIMA models are designed for shorter seasonal periods such as s = 12 for monthly data or s = 4 for quarterly data. The `Arima()` and `auto.arima()` functions only allow for a seasonal period up to m = 350, but in practice will usually run out of memory whenever the seasonal period is more than about 200. We attempted to work around these issues using the annually differenced time series $\{\tilde{x}_t\}$ and then using the smaller order of seasonal differencing. Our initial results from the ARIMA family of models produced reasonable results, but other models that explicitly handle complex seasonality were considered. We decided to run the Facebook Prophet model, in addition to dynamic harmonic regression models, which are in theory better at handling this type of seasonality.

Facebook Prophet

Prophet is a forecasting tool developed by Taylor and Letham [2018] that is based on an additive regression model with three parts as described in (6).

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{6}$$

where $g(t)$ is the trend function, $s(t)$ is the seasonality function, and $h(t)$ models holiday effect. The authors state that it is designed for data that have strong seasonal effects and/or multiple seasonalities, making it very amenable to our dataset.

Dynamic Harmonic Regression

When a time series exhibits complex seasonality, it is common to model the seasonal component using fourier terms. Dynamic Harmonic Regression (DHR) is based on the principal that a combination of sine and cosine functions can approximate any periodic function. We use a harmonic regression approach where the seasonal patterns are modeled by fourier terms and short-term dynamics are handled by an ARMA error. Thus, the model allows for multiple seasonal components of any length by including fourier terms of different frequencies as can be seen in our proposed model (7).

$$y_t = \beta_0 + s_7(t) + s_{365}(t) + \epsilon_t \tag{7}$$

where

$$s_7(t) = \sum_{i=1}^{3} \left[\alpha_i \sin\left(\frac{2\pi i t}{7}\right) + \beta_i \cos\left(\frac{2\pi i t}{7}\right) \right]$$

$$s_{365}(t) = \sum_{i=1}^{10} \left[\gamma_i \sin\left(\frac{2\pi i t}{365}\right) + \delta_l \cos\left(\frac{2\pi i t}{365}\right) \right]$$

and where ϵ_t is modeled as a non-seasonal ARIMA process. Usually, the number of fourier terms is determined by an iterative approach that minimizes some model selection criterion, such as AIC. However, Taylor and Letham [2018] note that using 3 fourier terms for weekly seasonality, and 10

for annual seasonality works well for most problems. For computational reasons, we decided follow this guideline by picking the number of fourier terms in advance rather than by treating it as a hyperparameter.

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

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)} \tag{8}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

In addition to these accuracy measures, we use a common-sense baseline as a sanity check, which serves as a benchmark for more advanced time series models. Given daily data with annual seasonality, our 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 annual seasonality, which is known as a seasonal naive model (10).

$$\hat{x}_t = x_{t-365} \tag{10}$$

Table 2 and Figure 9 summarize the performance results of our models on the test set. All models outperformed the seasonal naive baseline as measured by MAE and RMSE. Overall, the SARIMA$(1,0,2)(1,1,2)_7$ model performed the best, having the lowest MAE and RMSE.

Table 2: Model errors on test set.

Model	**RMSE**	**MAE**
Top 3 SARIMA		
SARIMA(1,0,2)(1,1,2)[7]	430.64	302.07
SARIMA(1,0,2)(0,1,2)[7]	431.38	302.57
SARIMA(1,0,2)(1,1,1)[7]	432.67	302.89
Dynamic Harmonic Regression	494.46	376.30
Prophet	552.56	429.99
Baseline		
Seasonal Naive	676.70	498.06

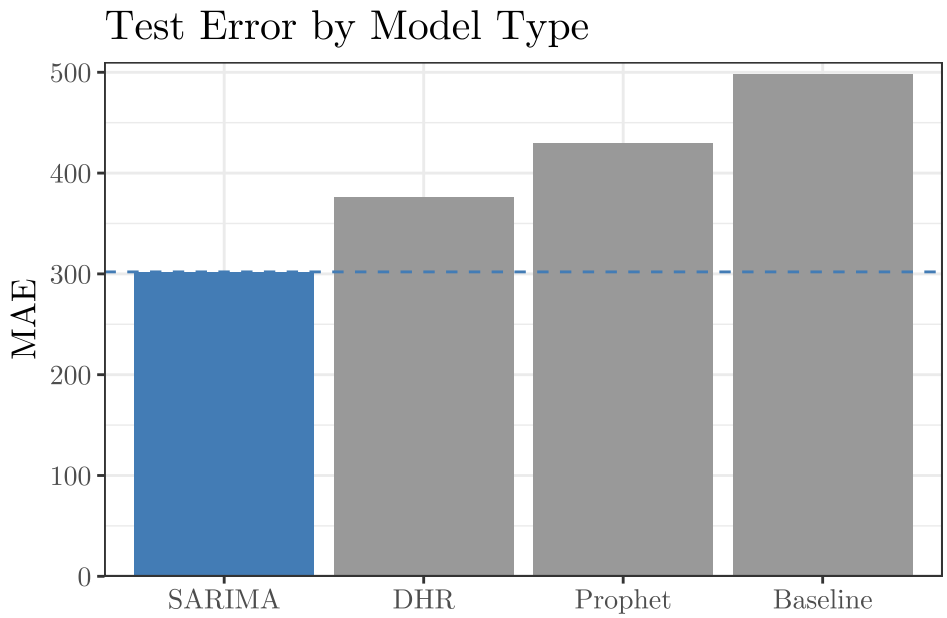


Figure 9: Test error by model type

Figure 10 plots daily forecasts of our best performing model (SARIMA$(1,0,2)(1,1,2)_7$) on the test set.

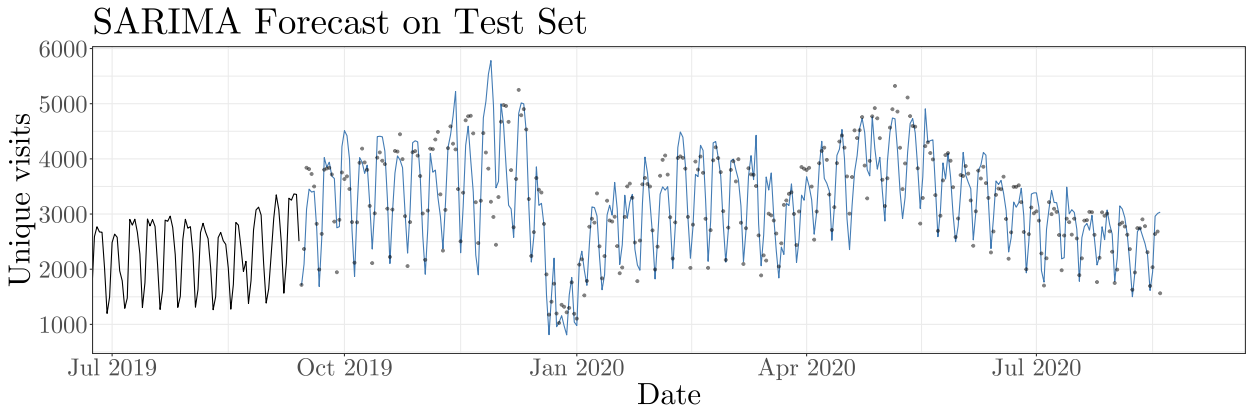


Figure 10: Forecast (blue) vs Unique Visits (black) on the test set

Finally, we are interested in making a forecast for the daily number of unique website visits over the next 30 days. Figure 11 shows such a 30-day forecast in blue, which has been generated using our best performing model. Table 13 uses the same forecast and displays the values for the first 7 days.

Table 3: Sample of 30-day forecast of page visits.

Date	Lower 95% CI	Forecast	Upper 95% CI
2020-08-20	1130	1647	2164
2020-08-21	620	1230	1841
2020-08-22	656	1290	1925
2020-08-23	1125	1780	2436
2020-08-24	1869	2542	3215
2020-08-25	1691	2379	3068
2020-08-26	1673	2374	3076

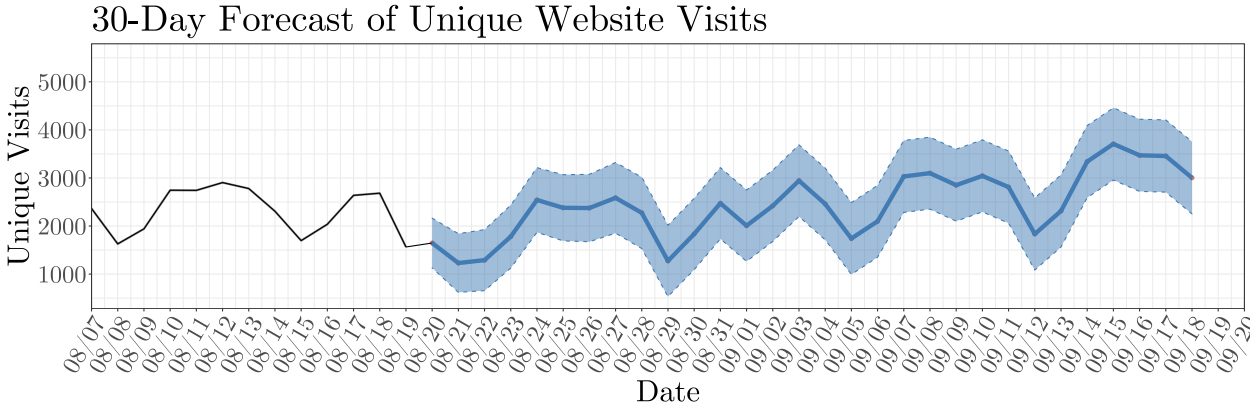


Figure 11: 30-day forecast (blue)

Conclusion

In this analysis we developed and compared different time series models for the task of predicting unique page visits to an academic website. Despite the complex seasonality inherent in the data, the best performing model was a SARIMA$(1,0,2)(1,1,2)_7$ model with an MAE of 302 page visits per day. Provided that a forecast with this error meets the requirements of the website owner, this model could be implemented to predict future trends and better understand user behavior. Furthermore, this model could provide valuable insights for load balancing if necessary.

One major constraint of this analysis is that we only used a univariate time series to make forecasts. Future research can be done to improve model performance by including additional features, as well as recording more data. Furthermore, additional model architectures, such as modifications to SARIMAX models to handle complex seasonality, vector autoregression, and hierarchical models, can be explored to better understand the relationship between different factors and their effects on forecasting daily web traffic.

References

Rob Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. URL OTexts.com/fpp3.

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.