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.

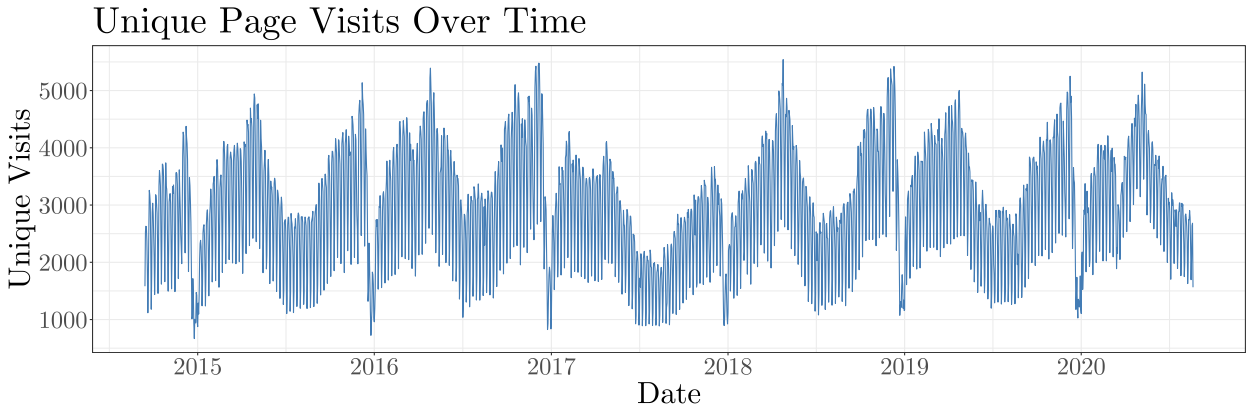


Figure 1

[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.

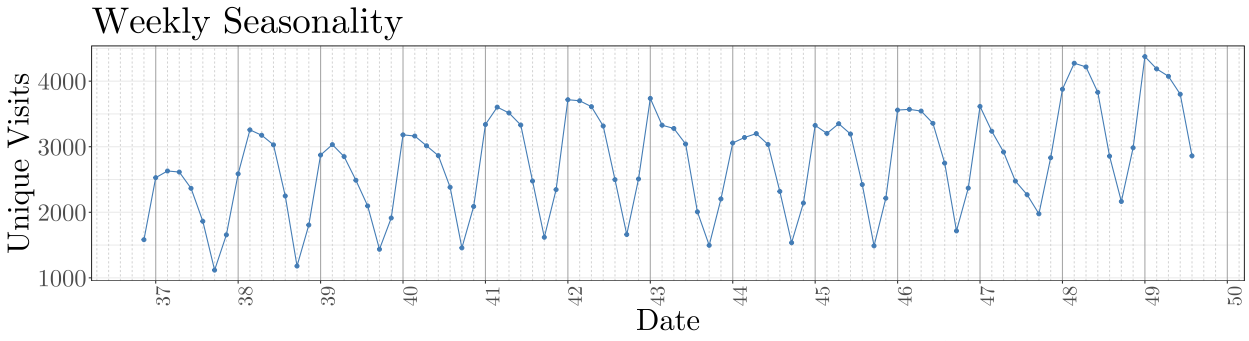


Figure 2: Sample of weekly page visits

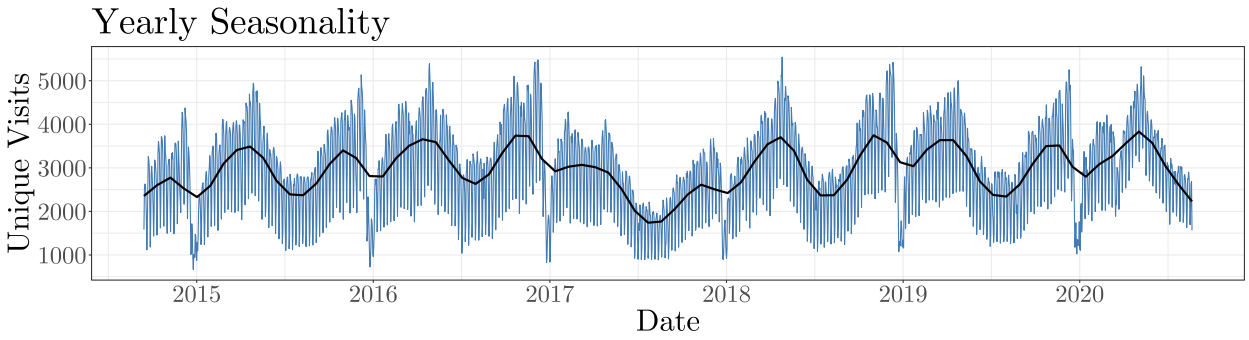


Figure 3: Smoothing via Lowess

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 (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 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 as displayed in (2) by taking the annual difference of the time series, and then taking the weekly difference of the transformed time series from the previous step.

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

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

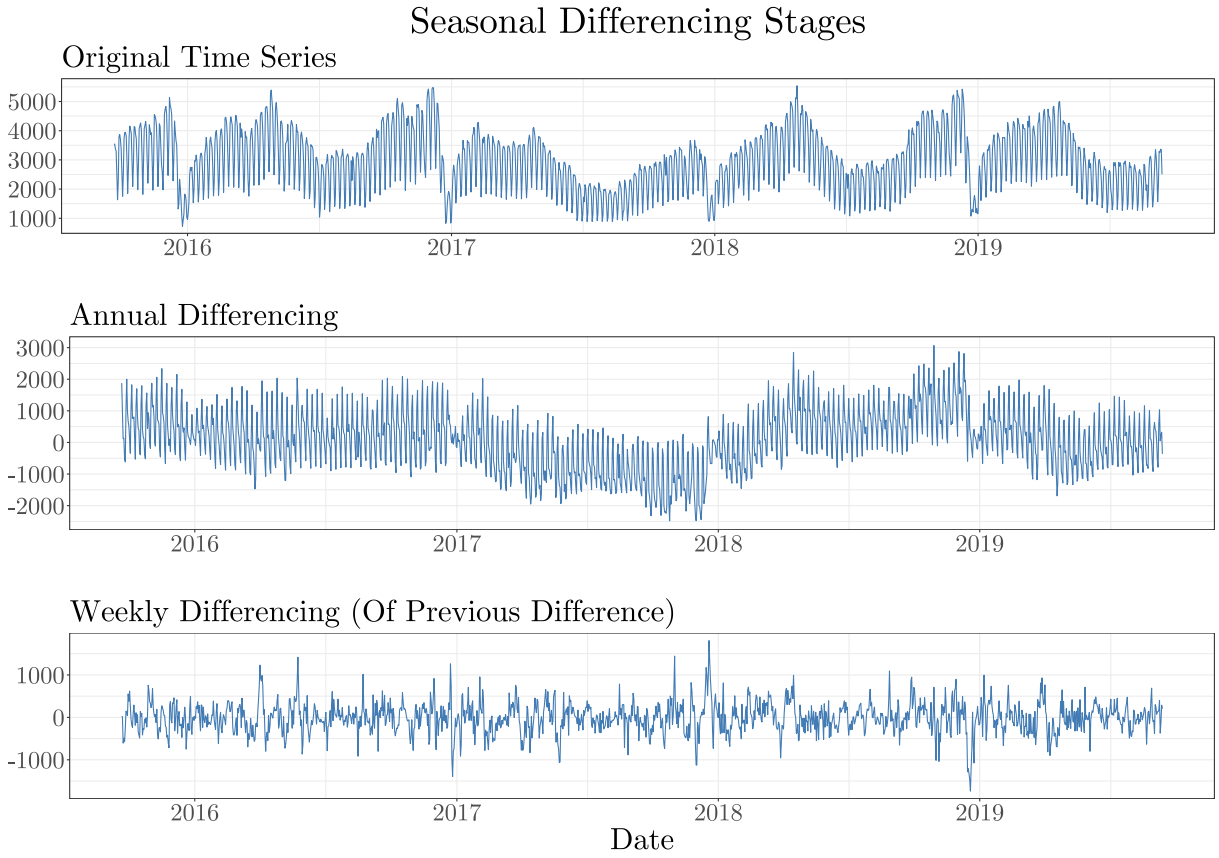


Figure 4: Time plots of differenced series

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 (3).

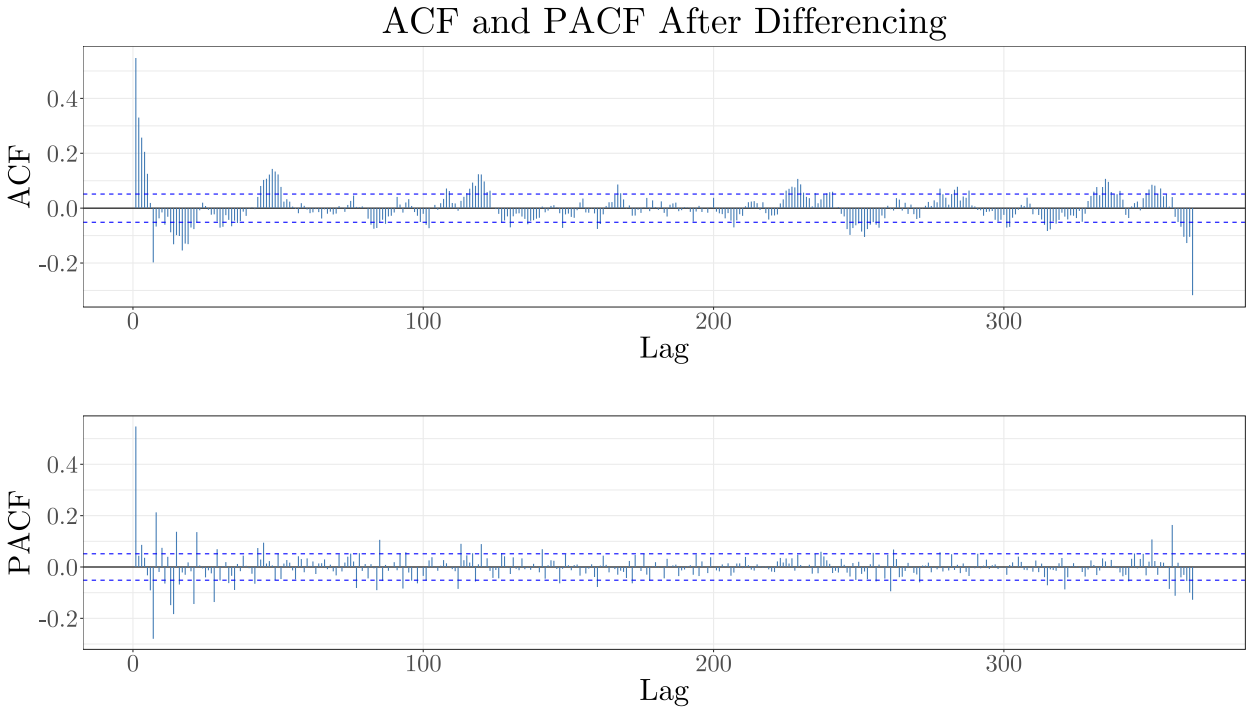


Figure 5: ACF and PACF

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

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 on AIC and BIC. The model selection criterion for the fitted ARMA models is shown in the Table 1 below.

Table 1: Model estimation results.

Model	**AIC**	**BIC**
SARMA(1,2)(0,2)[7]	20287.03	20318.72
ARMA(1,2)	20726.06	20747.19
ARMA(2,2)	20727.44	20753.85
ARMA(2,1)	20729.88	20751.01
ARMA(1,1)	20731.20	20747.04
ARMA(1,0)	20733.32	20743.88
ARMA(0,1)	20856.88	20867.44

We decided to choose the $SARMA(1,2)(0,2)_7$ model as it performed best in terms of AIC and BIC. Its parameters are estimated via maximum likelihood and are displayed in (4).

$$\hat{x}_t = 0.94_{(0.02)}x_{t-1} + \omega_t - 0.34_{(0.03)}\omega_{t-1} - 0.26_{(0.03)}\omega_{t-2} - 0.65_{(0.03)}\omega_{t-7} - 0.18_{(0.03)}\omega_{t-14} \quad (4)$$

Figure 6 displays of plot of the residuals of the fitted SARMA$(1,2)(0,2)_{12}$ 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 ARMA(p,q) modeling procedure.

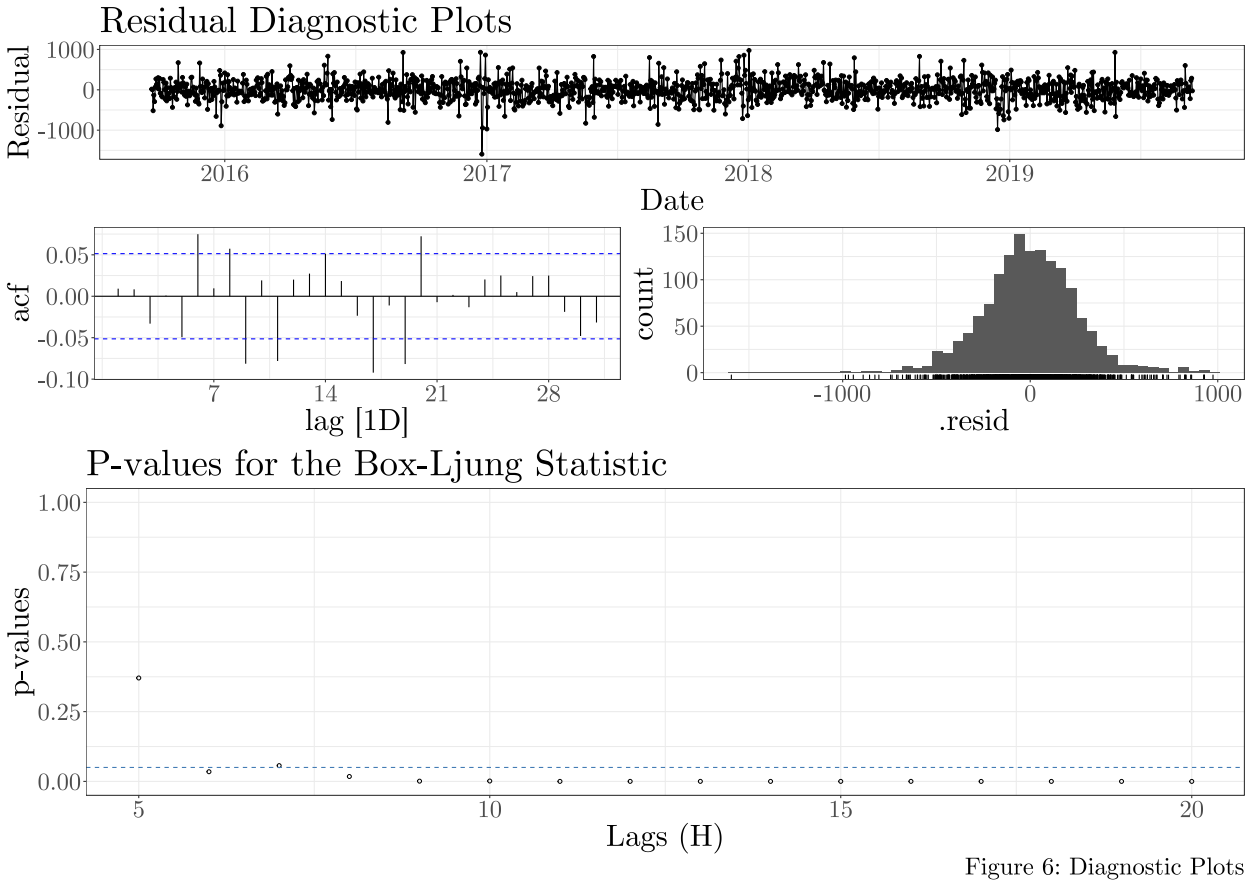


Figure 6: Diagnostic Plots

Figure 7 plots the inverse AR and MA roots of our SARMA 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). Hence, there dosn't appear to be any parameter redundancy.

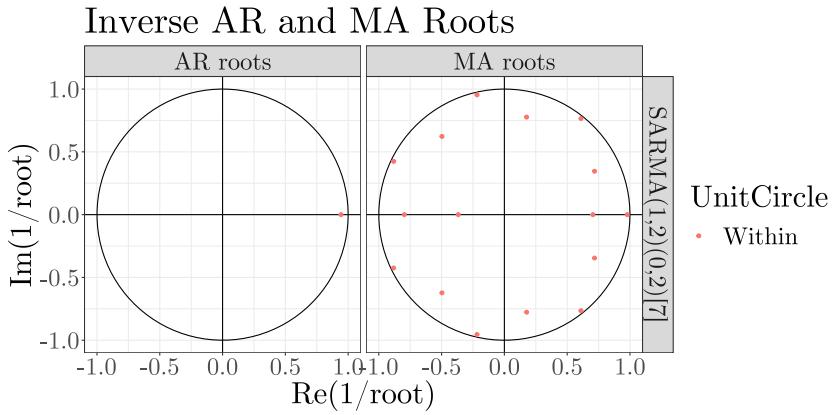


Figure 7: Inverse Roots

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

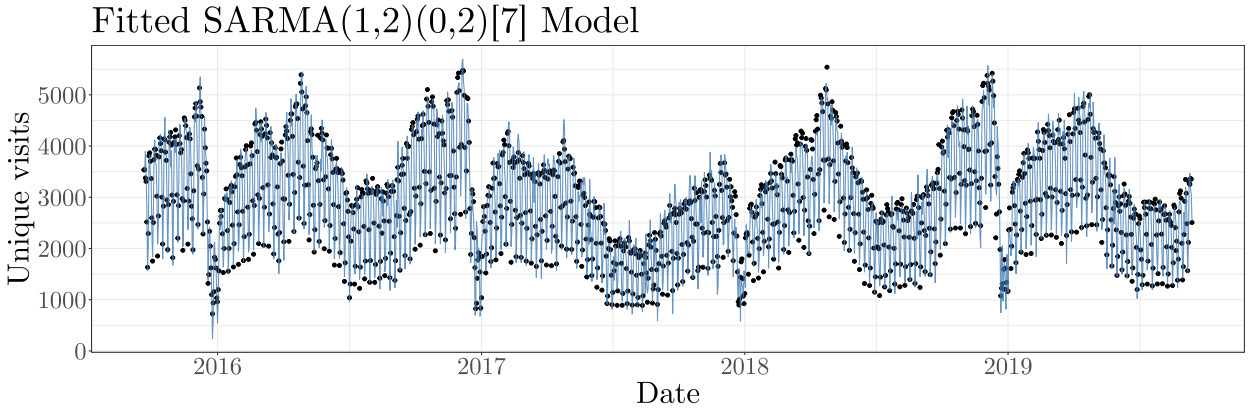


Figure 8: Fitted model (blue) vs Unique Visits (black) using training set

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 periods such as 12 for monthly data or 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. Hence, in addition to the ARMA model, we decided to run Prophet and dynamic linear regression models, which are in theory better at handling this type of seasonality.

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 (5).

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{5}$$

where $g(t)$ is the trend function, $s(t)$ is the seasonality function, and $h(t)$ models holiday effect. It is known to work best with data that have strong seasonal effects and/or multiple seasonalities.

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 seasonalities of any length by including fourier terms of different frequencies. Taylor and Letham [2018] note that using 3 fourier terms for weekly seasonality, and 10 for yearly seasonality works well for most problems. Hence, our DHR model is formulated in (6).

$$y_t = \beta_0 + s_7(t, m) + s_{365}(t, n) + \epsilon_t \tag{6}$$

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

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

where ϵ_t is modeled as a non-seasonal ARIMA process.

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

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

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 yearly 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 yearly seasonality, which is known as a seasonal naive model (11).

$$\hat{x}_t = x_{t-365} \tag{9}$$

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 $SARMA(1,2)(0,2)_7$ model performed the best, having the lowest MAE and RMSE.

Table 2: Model errors on test set.

Model	**RMSE**	**MAE**
SARMA(1,2)(0,2)[7]	430.61	302.12
Dynamic Harmonic Regression	494.46	376.30
Prophet	551.15	429.75
Baseline		
Seasonal Naive	676.70	498.06

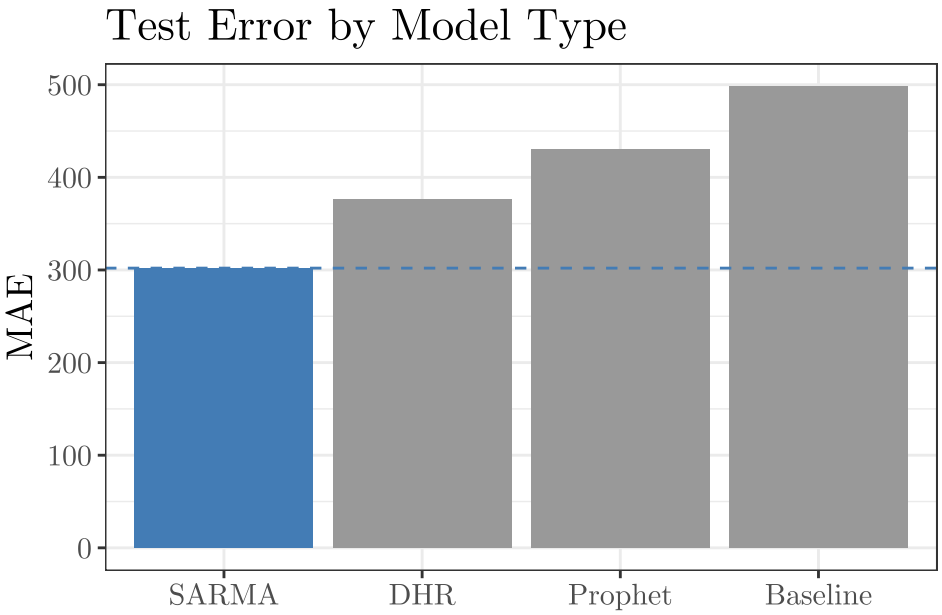
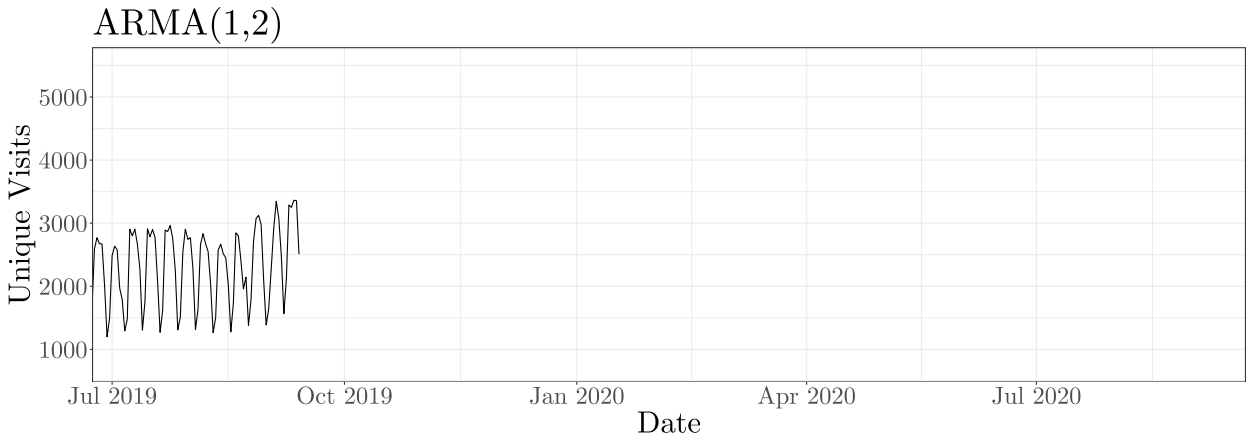


Figure 9: Test error by model type



Conclusion

The d

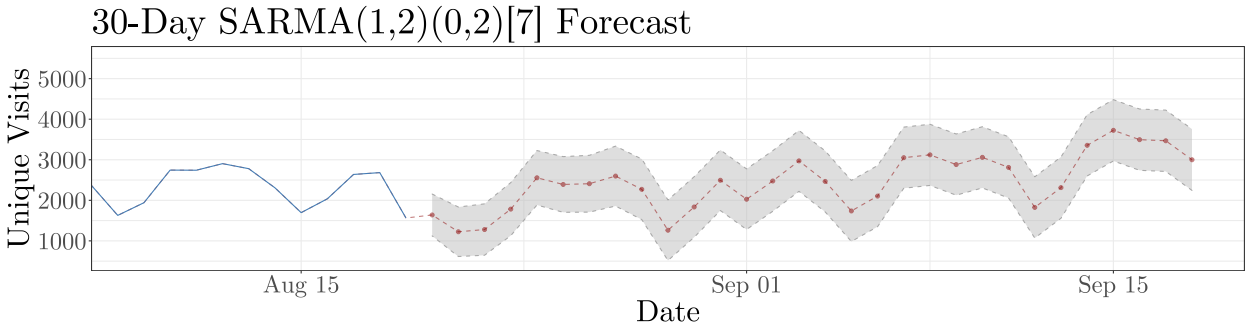


Table 3: 30-Day ARMA(1,2) Forecast.

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

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.