TIME SERIES ANALYSIS OF TWITTER DATA

MATTHEW TIGER

1. Introduction

With the advent of social media, the way in which society communicates has evolved. Popular social media networks document these new forms of communications and as a result, a rich set of data surrounding these interactions emerges. In particular, Twitter remains one of the most popular social media platforms to date. Twitter was first launched in July 2006 as a social networking service designed to allow its users to communicate via short 140-character messages called "tweets". These tweets are sometimes affixed by the sender with a meta-label called a "hashtag", denoted by a string leading with the # symbol, that is meant to categorize the information contained in the message. As of May 2015, Twitter's active user base numbers 302 million users sending these categorized messages every second.

In this report, we will analyze data pertaining to a popular television show collected from Twitter's streaming API over the course of three weeks. This analysis will consist of measuring the number of tweets that contain a certain hashtag sent in a given hour over this timeframe. We will then fit a time series model to these measurements and provide a forecasting model of the next week's projected data.

2. Data

In an effort to help researchers glean insight from tweets, Twitter offers a streaming API that streams all tweets that contain a certain string. We leverage this service to gather data surrounding the popular television show *The Walking Dead* and then measure the number of tweets that occur each hour over a time span.

Using the programming language Python and the library `tweepy`, a popular wrapper for the Twitter streaming API, we collected all tweets containing the hashtag "#thewalkingdead" for three weeks from 2015-11-07 21:00 EST to 2015-11-23 21:00 EST. As this television show airs Sundays at 21:00 EST, of particular importance is the tweets occurring around this time. We therefore restrict our measurements to 24 hours prior to and 24 hours after the television show's airing for the above three week time frame.

These measurements give rise to the time series presented in Figure 1. From this plot it is clear that, due to the weekly episodic nature of the television show, there is a seasonal component to this time series. It also appears that there is a trend component. Thus, in order to fit a stationary time series to the underlying data, we will need to first apply transformations to the data.

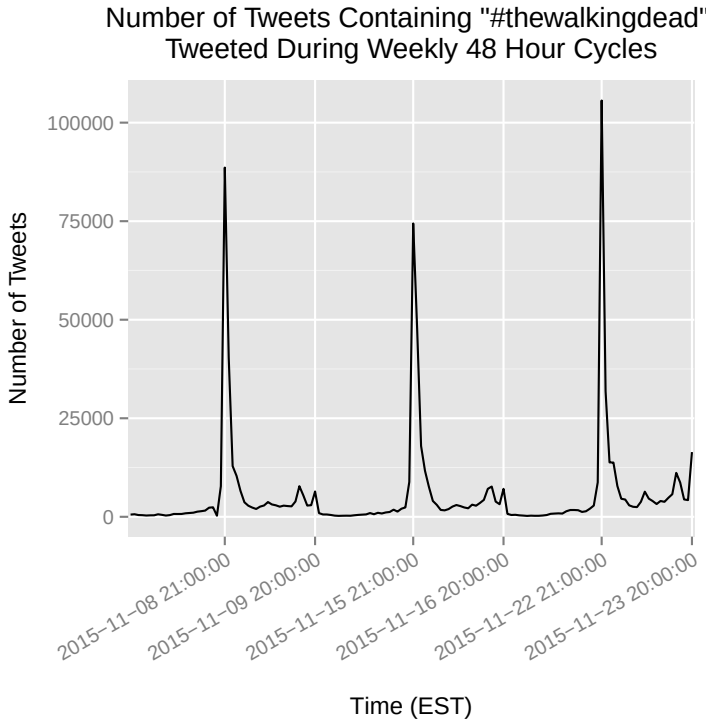

FIGURE 1. Time series plot of number of tweets containing the hashtag #thewalkingdead over three 48 hour cycles.

The details behind the process of fitting a stationary time series model to this data is handled in Section 3.

3. Model Fitting

We now wish to fit a time series model to the data described in Section 2. As can be seen from Figure 1, there are seasonal and trend components present in the underlying data set and what also appears to be non-constant variance.

Thus, we will need to apply transformations to the time series in order to determine the underlying time series model.

3.1. Transformations Let $\{X_t\}$ for $t = 1, 2, \ldots, 144$ denote the observations of the time series described in Section 2. To remove the non-constant variance we apply the Box-Cox transformation log to the observations $\{X_t\}$ to arrive at the mean-corrected data $Y_t = \log(X_t) -$

$E(\log(X_t))$ for $t = 1, 2, \ldots, 144$. As can be seen from Figure 2, this transformation has made the variance of the data constant.

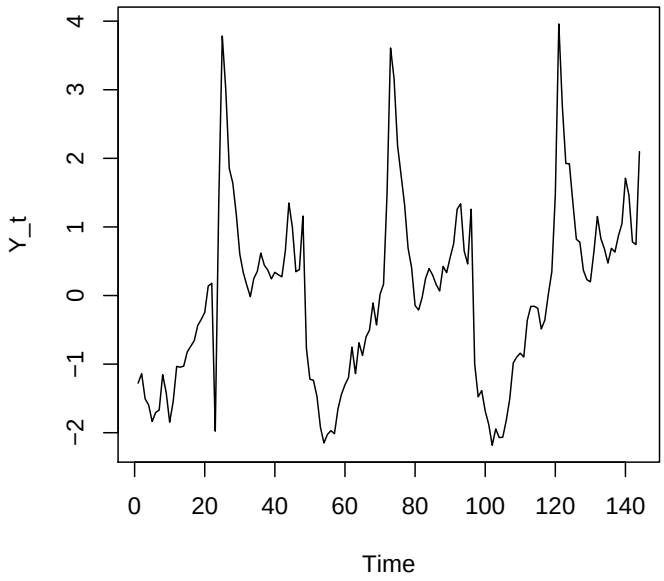

FIGURE 2. Plot of Box-Cox transformed data $Y_t = \log(X_t) - E(\log(X_t))$.

The seasonality and trend noticed with the untransformed data is still present in the transformed data. Through differencing, we hope to characterize the underlying seasonality and trend of the data.

Knowing that the observations come from data with a period of 48 hours, it makes sense to remove the seasonality by applying the differencing operator $(1 - B^{48})$ to Y_t. Applying this transformation results in Figure 3.

As can be seen from the figure, the seasonality has been removed. The plots of the ACF and the PACF of $(1 - B^{48}) Y_t$ are found in Figure 4.

These plots suggest that a seasonal AR model of order 1 or 3 may be appropriate as the ACF slowly decreases to 0 while the PACF stops abruptly after lag 3.

However, as can also be seen in Figure 3, there is still a trend component present. This trend appears linear suggesting that applying the

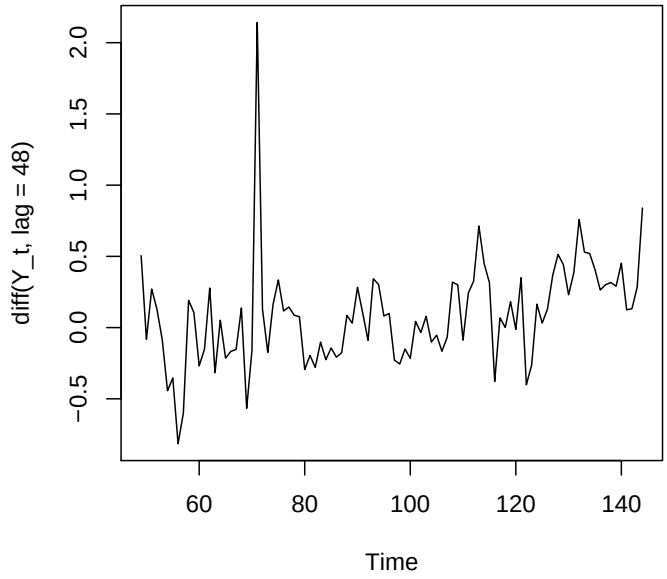

FIGURE 3. Plot of data $(1 - B^{48})\, Y_t$.

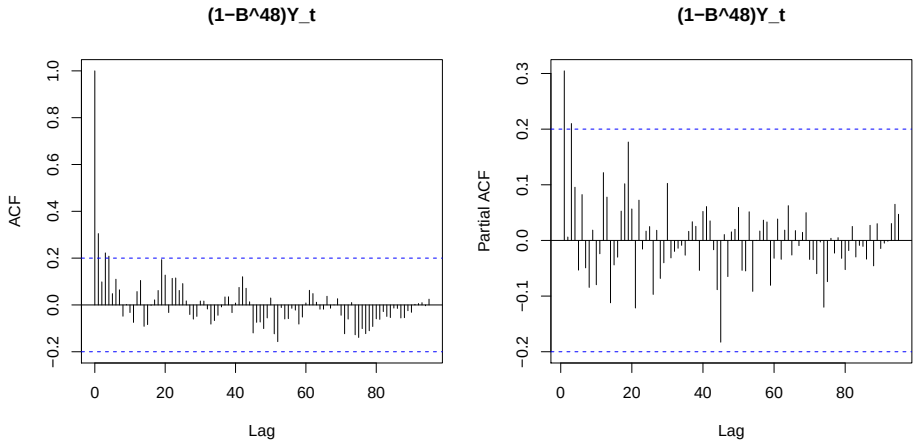

FIGURE 4. ACF and PACF of $(1 - B^{48})\, Y_t$.

difference operator $(1 - B)$ to $(1 - B^{48})\, Y_t$ will remove this trend. The results of applying the operator are found in Figure 5.

The plot in Figure 5 shows that the trend component. The plots of the ACF and the PACF of $(1 - B)\,(1 - B^{48})\, Y_t$ are found in Figure 6.

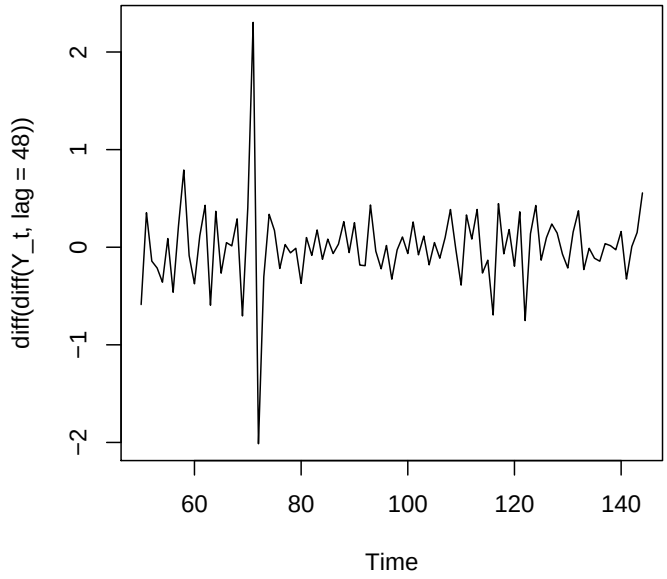

FIGURE 5. Plot of data $(1-B)\,(1-B^{48})\,Y_t$.

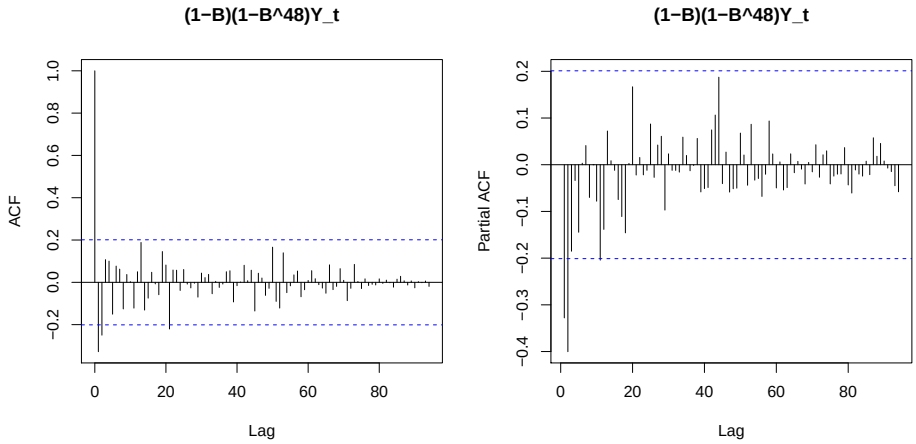

FIGURE 6. ACF and PACF of $(1-B)\,(1-B^{48})\,Y_t$.

These plots suggest a possible AR model of order 2 for the trend component as the ACF slowly decreases to 0 while the PACF stops abruptly after lag 2. It is also possible for an MA component to be

present so we will investigate models of the form ARMA(p, q) for $p, q = 1, 2$ for the trend component.

3.2. Model Selection The findings from Section 3.1 suggest fitting the data $\{Y_t\}$ to a seasonal ARIMA model. Specifically, a model of the form

$$\text{SARIMA}(p, 1, q) \times (P, 1, 0)_{48}$$

where $p, q \in \{0, 1, 2\}$ and $P \in \{1, 3\}$.

Using the programming language R, we examine the AIC statistic associated to each of the different possible models suggested and choose the model that minimizes this statistic. Due to the limitations of the available hardware, we were unable to fit SARIMA models to the data in which $P = 3$ for such a large period. Thus, we present the results of testing for $P = 1$ alone.

p	q	AIC
0	1	77.00203
0	2	75.06627
1	0	99.46614
1	1	`NaN`
1	2	78.04194
2	0	85.23537
2	1	77.33328
2	2	78.92953

TABLE 1. AIC values for different possible SARIMA$(p, 1, q) \times (1, 1, 0)_{48}$ models fitted to the data $Y_t = \log(X_t) - E(\log(X_t))$.

Note that hardware limitations also prevented us from fitting a model of the form SARIMA$(1, 1, 1) \times (1, 1, 0)_{48}$ to the data so that has been omitted from consideration as well. Table 1 suggests that, if we are to select a model based on minimizing the AIC statistic, we should choose to fit the data to a SARIMA$(0, 1, 2) \times (1, 1, 0)_{48}$ model.

Doing so produces the following output in R:

```
Coefficients:
          ma1      ma2     sar1
      -0.6768  -0.2257  -0.1586
s.e.   0.1097   0.1100   0.1441
```

```
sigma^2 estimated as 0.1152:  log likelihood = -33.53,
aic = 75.07
```

The above gives the associated p-values for the coefficients of the model:

```
        ma1          ma2          sar1
6.878613e-10 4.021464e-02 2.709825e-01
```

Using a significance level of $\alpha = 0.05$ suggests that the SAR(1) coefficient is not significant. However, removing the seasonal component and fitting the model to an MA(2) model results in a worse fit. Thus, we elect to keep the seasonal component in the model.

Thus we fit, our data $Y_t = \log(X_t) - E(\log(X_t))$ to the model

$$Y_t = \Phi_1 Y_{t-48} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \quad Z_t \sim \mathrm{WN}(0, \sigma^2)$$

$$\Phi_1 = -0.1586, \quad \theta_1 = -0.6768, \quad \theta_2 = -0.2257, \quad \sigma^2 = 0.1152 \tag{1}$$

By examining the residuals in R, we verify that these residuals are indeed a white noise process by examining the plot of their ACF in Figure 7

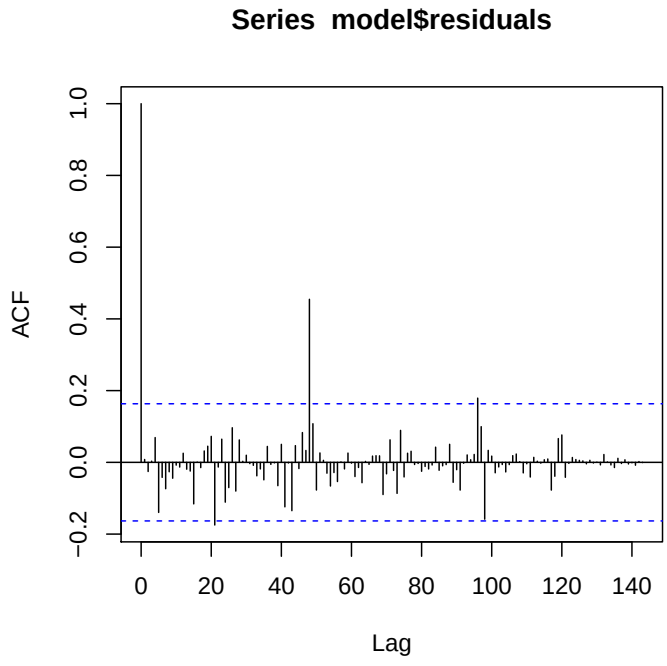

FIGURE 7. Plot of ACF of Z_t in (1).