Final Project

Brian Detweiler

April 24, 2017

Abstract

The city of Chicago publishes a data set on the number of speeding violations captured by cameras posted throughout the city dating back to June 1st, 2014. In this paper, we look for trends over time and attempt to explain the variance in the violations and to forcast future violations for the remainder of 2017.

The Data

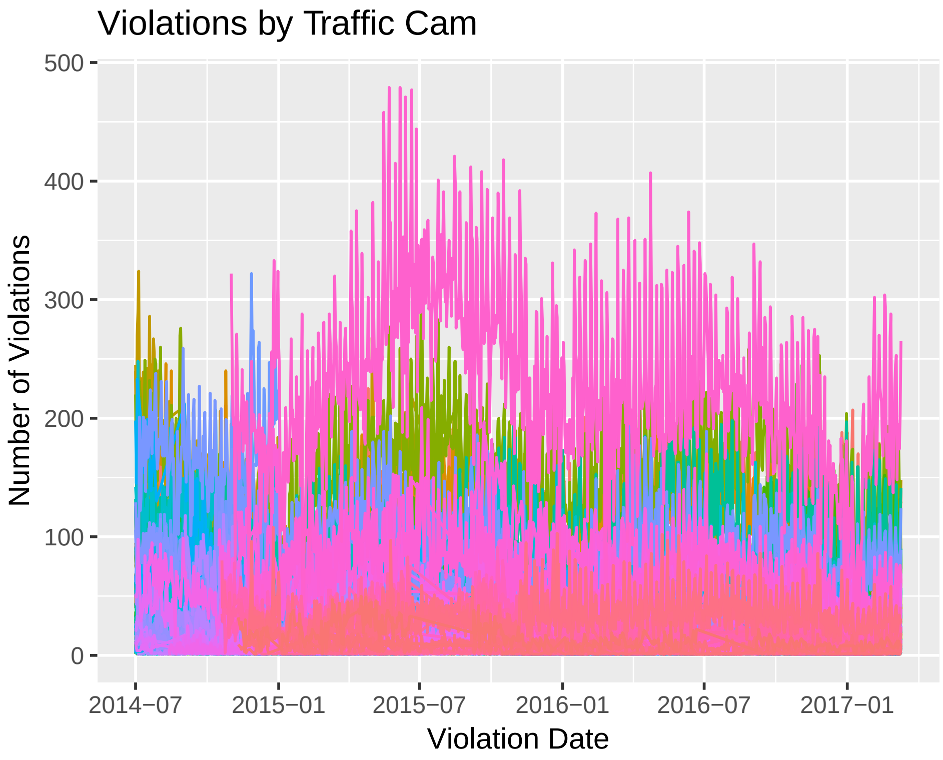
The data set is fairly simple, consisting of the address, camera ID, violation date, number of violations. There are some additional columns for latitude and longitude, but these appear to be sparse. We will repopulate these using the R package `ggmap`.

Although it would be ideal to model both the individual cameras and the overall city-level trend, there seems to be too much missing data on many cameras, so the individual models become more guesswork than empiricism. However, if we aggregate to the city-level, missing and existing data are evenly distributed over all the cameras, and we can have a better shot at building an accurate model.

We can display all violations by traffic camera by representing the different cameras as different colors. There are 150 cameras in the dataset, so the plot does become quite messy.

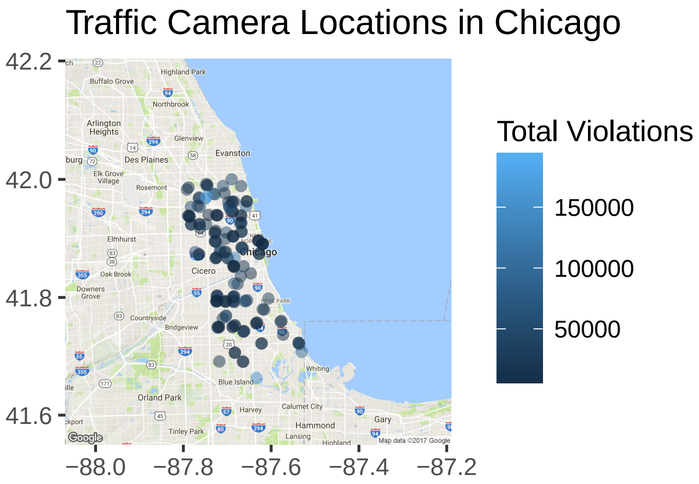
Table 1: Sample of the data

ADDRESS	CAMERA.ID	VIOLATION.DATE	VIOLATIONS
7738 S WESTERN	CHI065	2014-07-08	65
1111 N HUMBOLDT	CHI010	2014-07-16	56
5520 S WESTERN	CHI069	2014-07-08	10
1111 N HUMBOLDT	CHI010	2014-07-26	101
1111 N HUMBOLDT	CHI010	2014-07-27	92



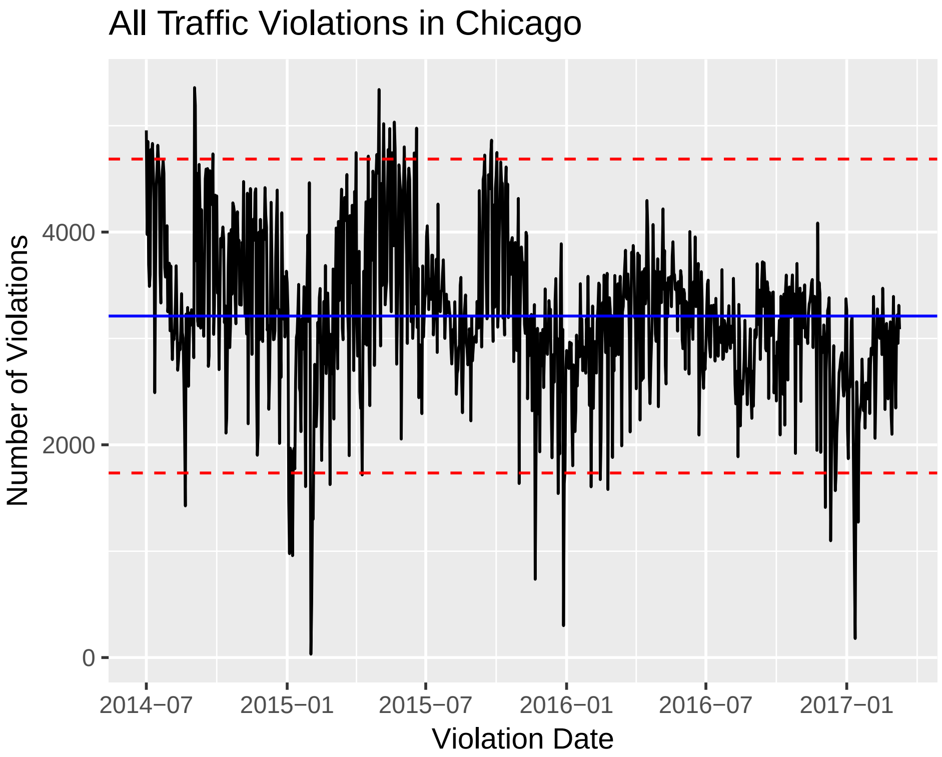
Locations of Cameras

The cameras are scattered througout the city. An interactive version can be found at http://bdetweiler.github.io/projects/chicago-traffic-cameras.html.



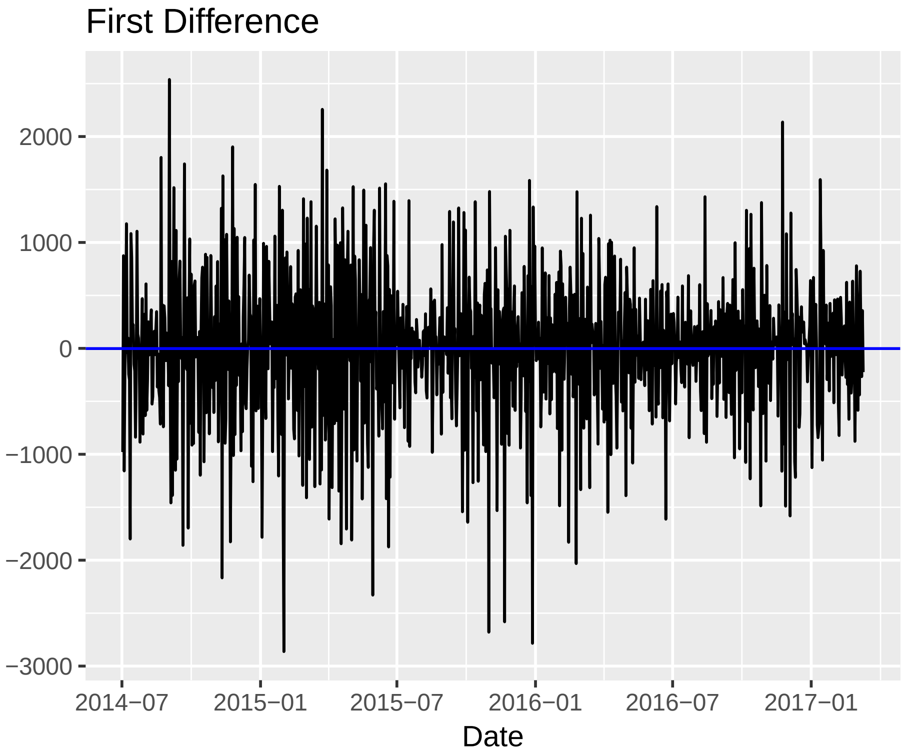
City-Level Analysis

By aggregating the individual camera violations by date, we can get an overview of all the violations in the city as a time series.



The blue line represents the mean, and red dashed lines represent ±$2s$.

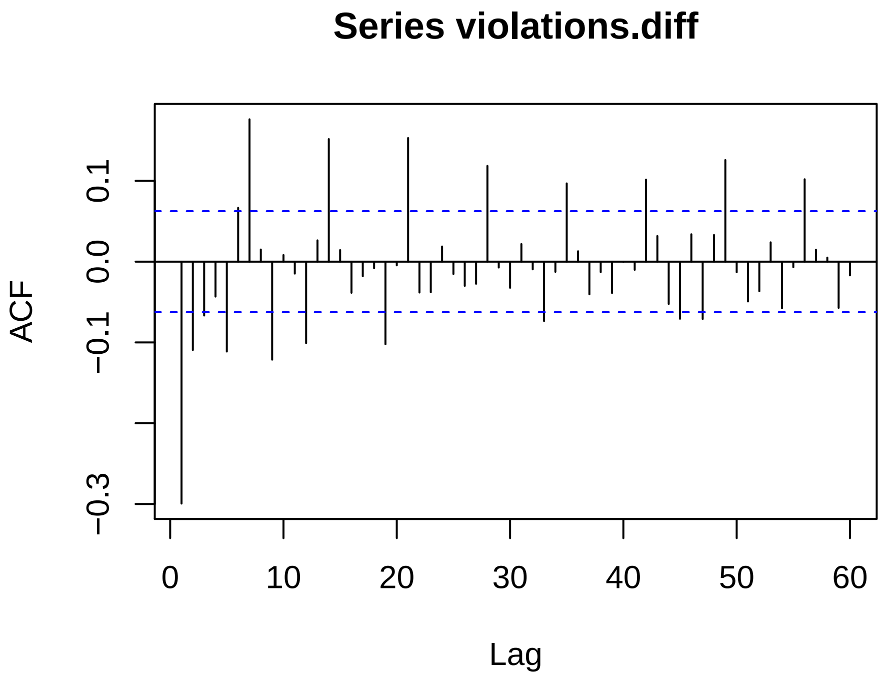
The mean and variance don't appear to be constant over time. The first difference is given by $W_t = \nabla Y_t = Y_t - Y_{t-1}$.

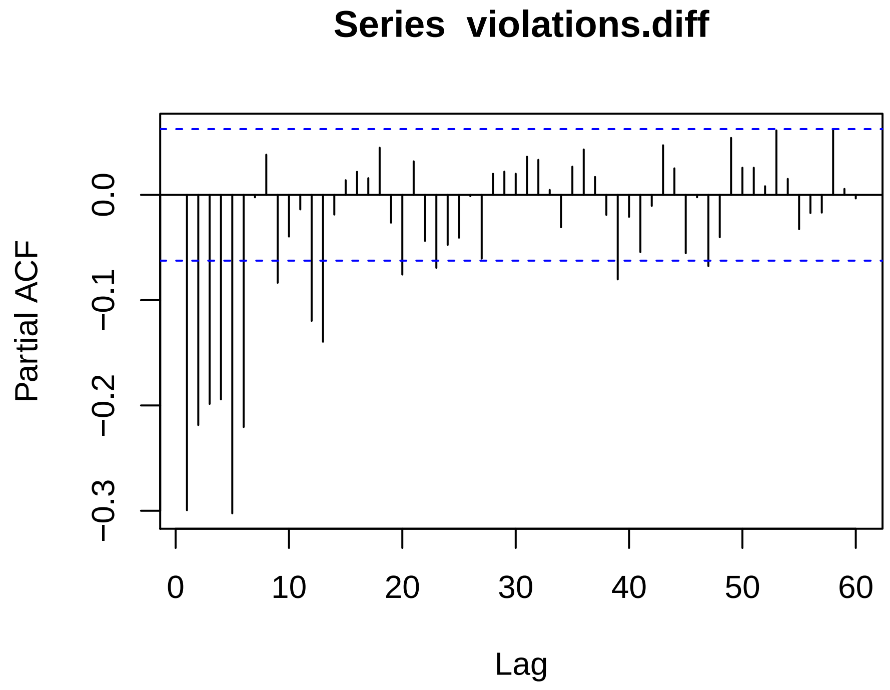


The first difference gives us a constant, zero-mean over time. The variance is still very large.

This looks much more stationary, and an augmented Dickey-Fuller test reinforces this assumption with a p-value less than 0.01.

We will investigate different models and evaluate their residuals, AIC, and BICs to determine the best fit. The first thing we check with any time series is check the ACF, PACF, and EACF.

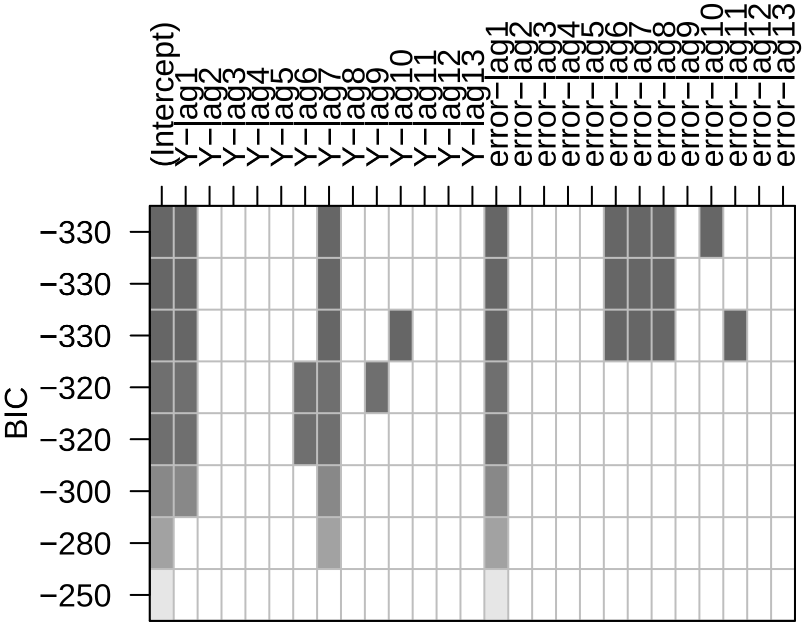




```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x x o x x x o x o o  x  o  x
## 1 x o o o x o x o x x o  x  o  x
## 2 x x o o x o o x x o o  o  o  x
## 3 x o x o x o o o x x o  x  o  x
## 4 x x x x x o o o o o o o  o  o  x
## 5 x x x o o x o o o o o o  o  o  o
## 6 o x x x o x x o o o o o  o  o  o
## 7 x o x x x x x o o o o o  o  o  o
```

The EACF doesn't appear to show any triangle pattern. The ACF and shows a potential seasonal pattern, and the PACF appears to cut off, so we will look to fit an ARMA(p, q) model to the first difference.

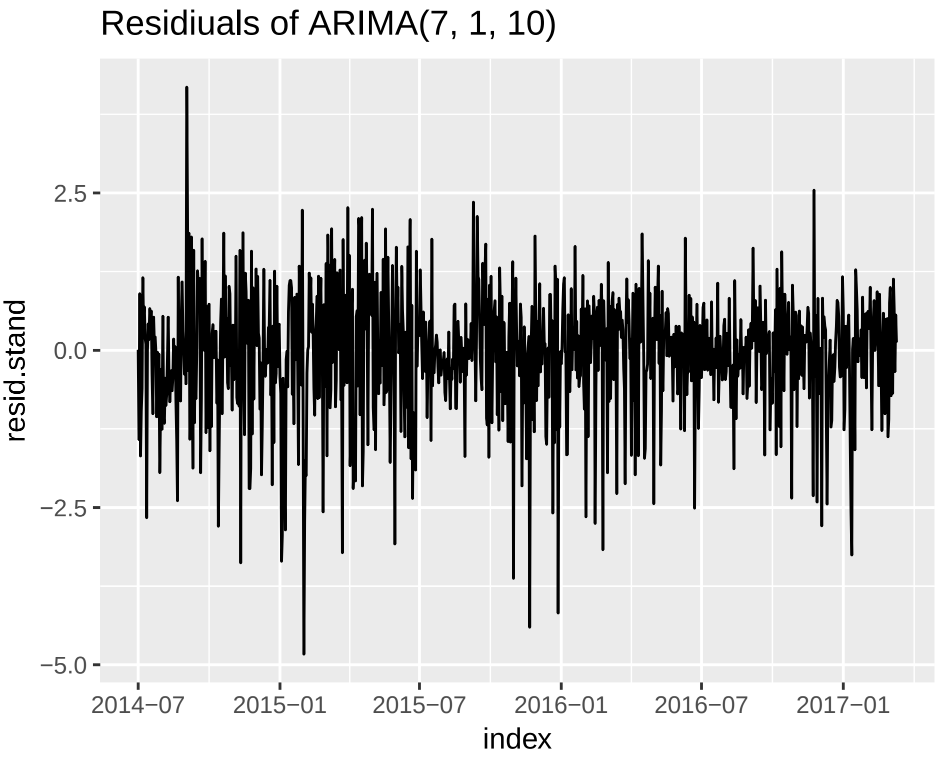
We can check the ARMA subsets as well.

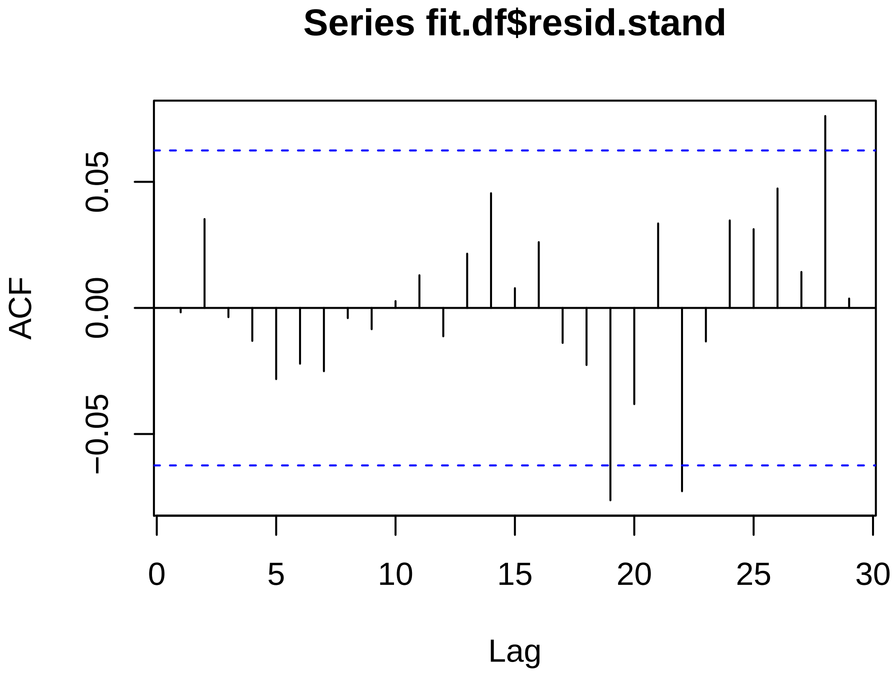


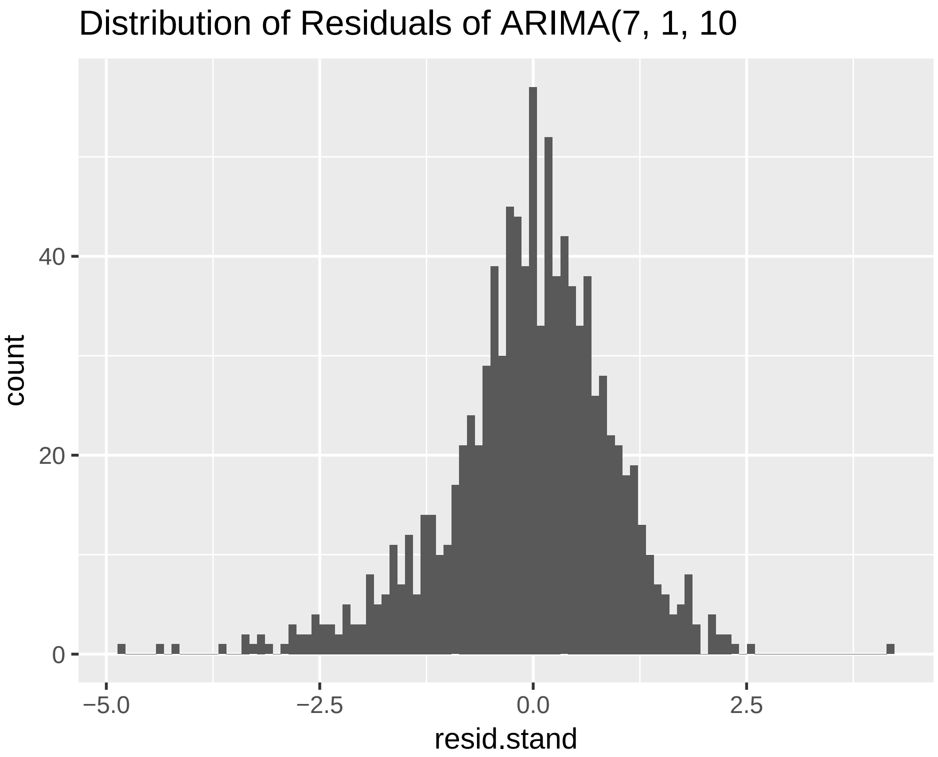
The best model for this, according to the ARMA subsets is an ARIMA(7, 1, 10) given by the equation

$$W_t = \nabla Y_t = Y_t - Y_{t-1}$$
$$W_t = \phi_1 W_{t-1} + \phi_7 W_{t-7} + e_t - \theta_1 e_1 - \theta_6 e_{t-6} - \theta_7 e_{t-7} - \theta_8 e_{t-8} - \theta_{10} e_{t-10}$$

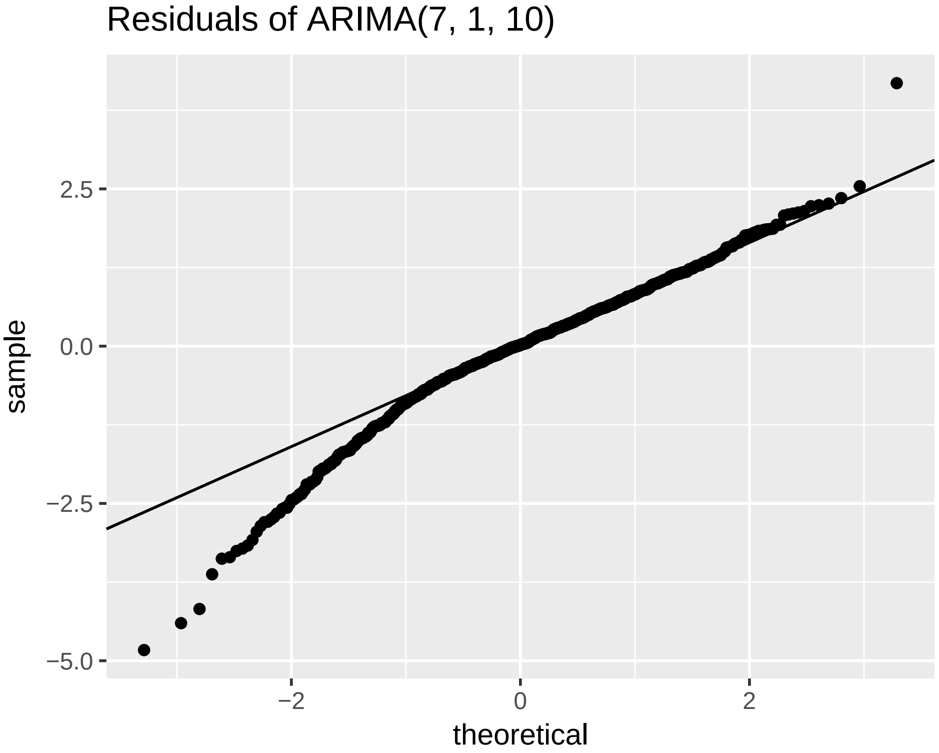
With an AIC of 1.5246818×10^4, this is not an ideal model. However, this is very real world data, so we will likely have to make some consessions regarding our model.


Residiuals of ARIMA(7, 1, 10)
resid.stand
2.5
0.0
−2.5
−5.0
2014−07
2015−01
2015−07
2016−01
2016−07
2017−01
index


Series fit.df$resid.stand
ACF
0.05
0.00
−0.05
0
5
10
15
20
25
30
Lag



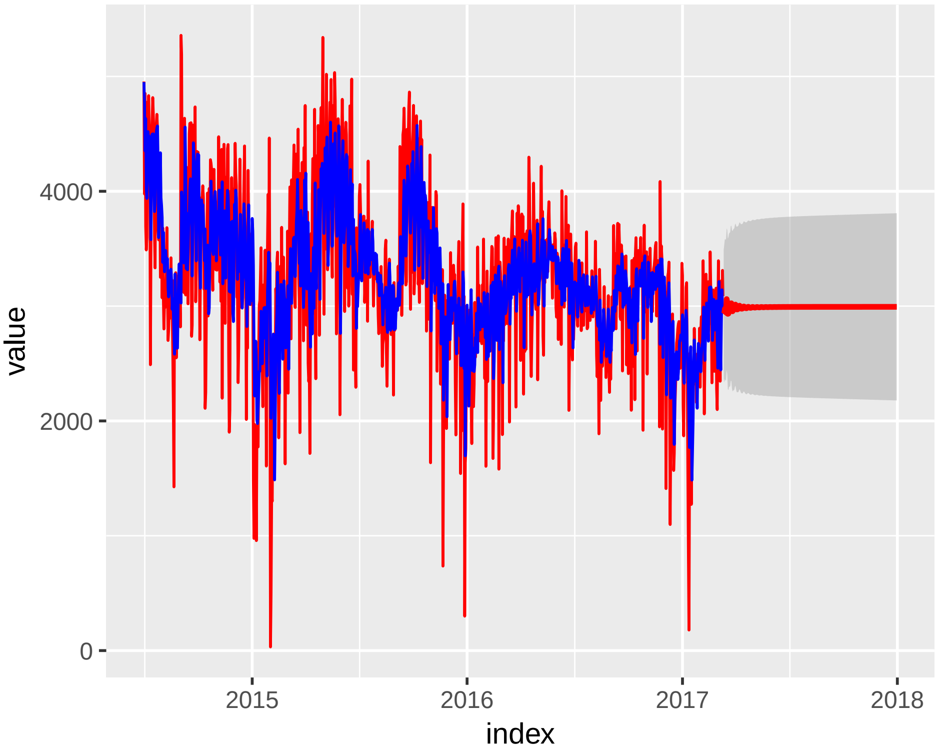
Distribution of Residuals of ARIMA(7, 1, 10



Residuals of ARIMA(7, 1, 10)

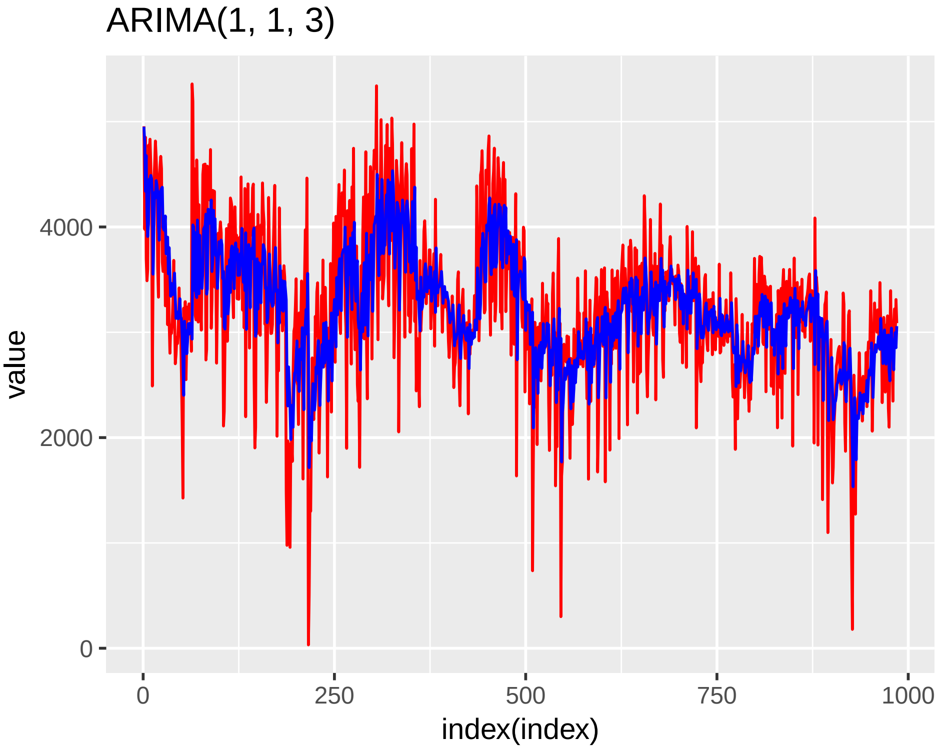
Here we can see the residuals do somewhat resemble white noise, with a few areas of heteroskedascity. The residuals appear to be normally distributed, with a Shapiro-Wilk test returning a p-value of $8.2129341 \times 10^{-14}$.

And a Ljung-Box test of the residuals results in a p-value of 0.9560039, thus we do not reject the null hypothesis that the residuals are independent.

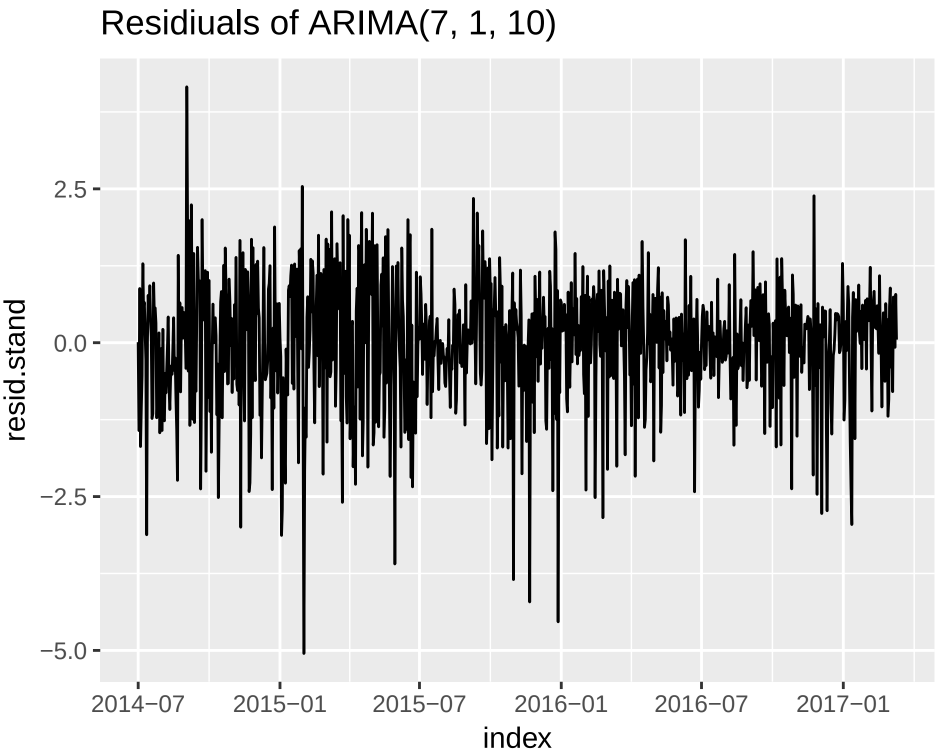


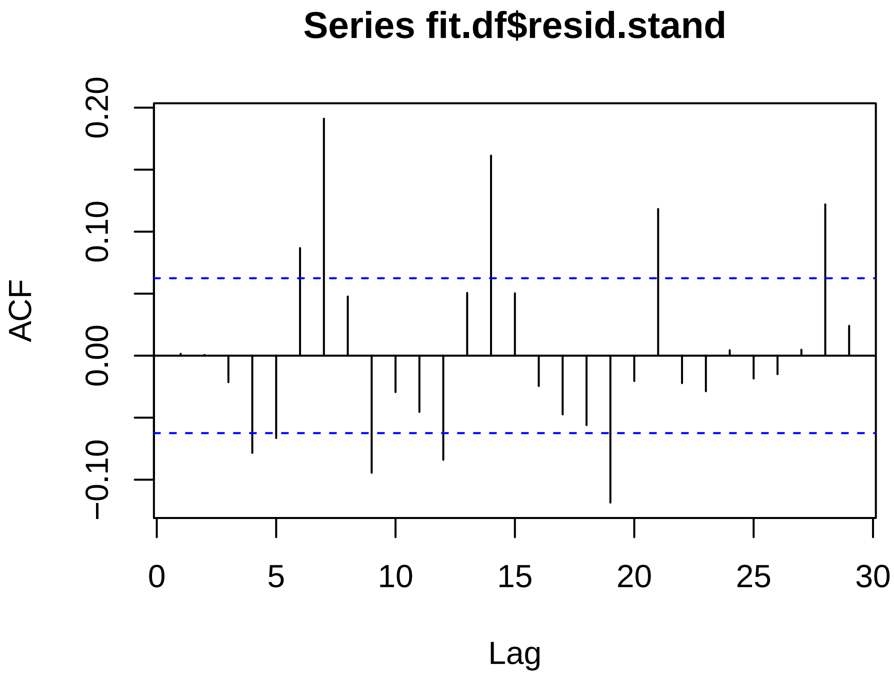
Auto-ARIMA

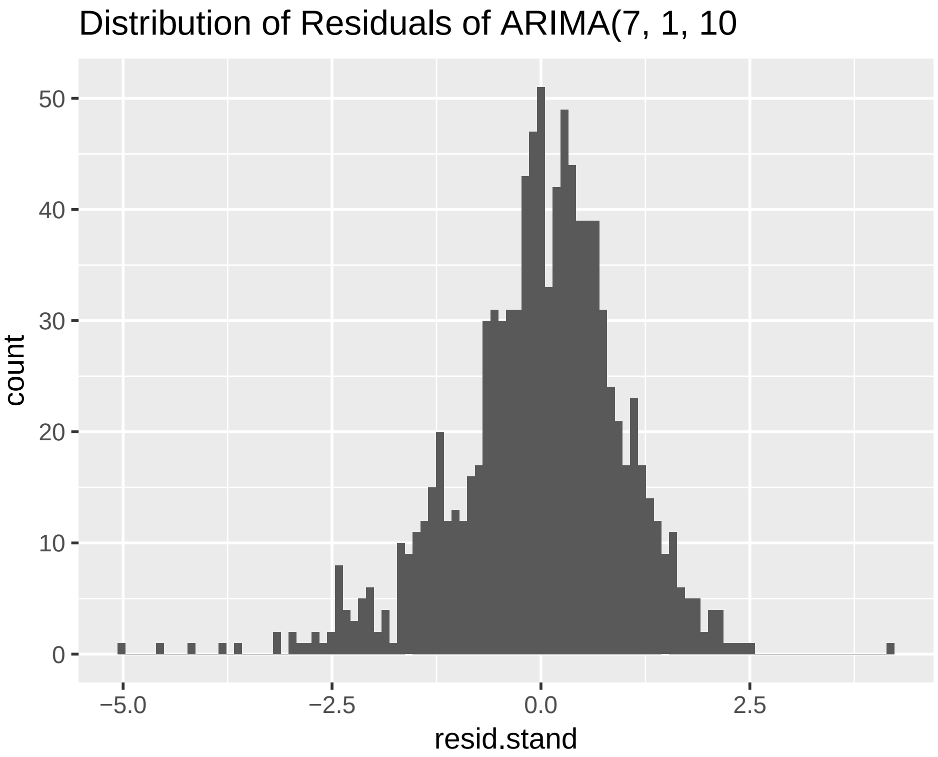
Another way to fit this time series is to use the `forecast::auto.arima()` function.

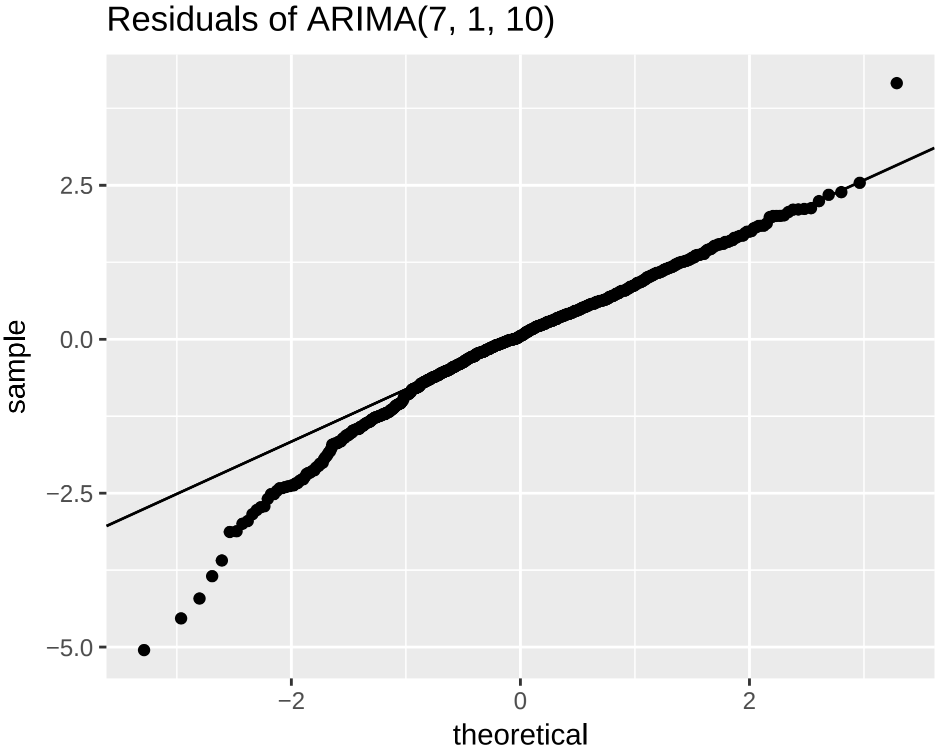


`forecast::auto.arima()` prouduces an ARIMA(1, 1, 3) with an AIC of 1.5331865×10^4. This is worse than our previous model.



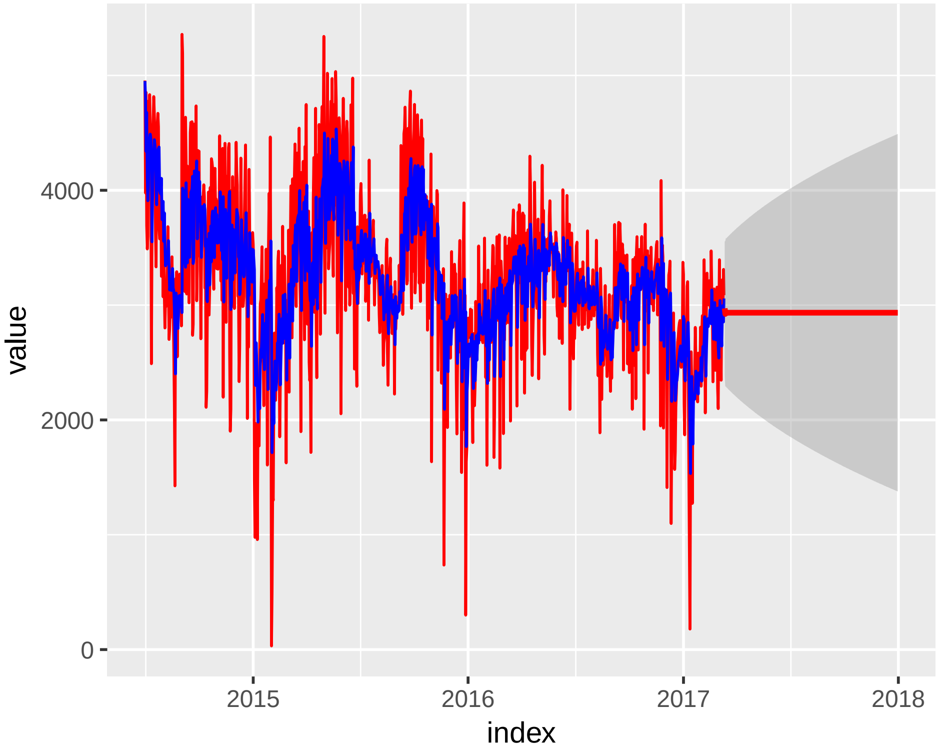






The residuals still resemble white noise, with a few areas of heteroskedascity. The residuals appear to be normally distributed, with a Shapiro-Wilk test returning a p-value of $9.6807843 \times 10^{-13}$. There seems to be some seasonal autocorrelation in the residuals, however, which makes this very problematic.

And a Ljung-Box test of the residuals results in a p-value of 0.9609813, thus we do not reject the null hypothesis that the residuals are independent.



Conclusion

The ARIMA(7, 1, 10) has a better AIC than the ARIMA(1, 1, 3), and the residuals in the second model seem to be strongly autocorrelated, so we would choose the first model over the second. However, the prediction limits for Model 1 seem a bit narrow, and we may feel more comfortable choosing Model 2 for the more reasonable prediction limits, or possibly tweaking the forecast of the first model to obtain something more reasonable.