Divvy Bike Share Usage

Hyejeong Lee, Kunal Shukla, WanQi Tay, Yingkun Zhu

Time Series Analysis & Forecasting

Master of Science in Analytics, University of Chicago

August 23, 2018

Agenda

- Project Overview
  - Problem Statement
  - Dataset Overview
  - Data Pre-Processing
- Modeling
  - Part I - Normalized Series
    - sNaive
    - Dynamic Harmonic Regression
  - Part II - Original Series - XREG
  - Cross-Validation
- Future Work
  - VAR
  - Neural Networks
  - TBATS (Individual Stations)

Problem Statement

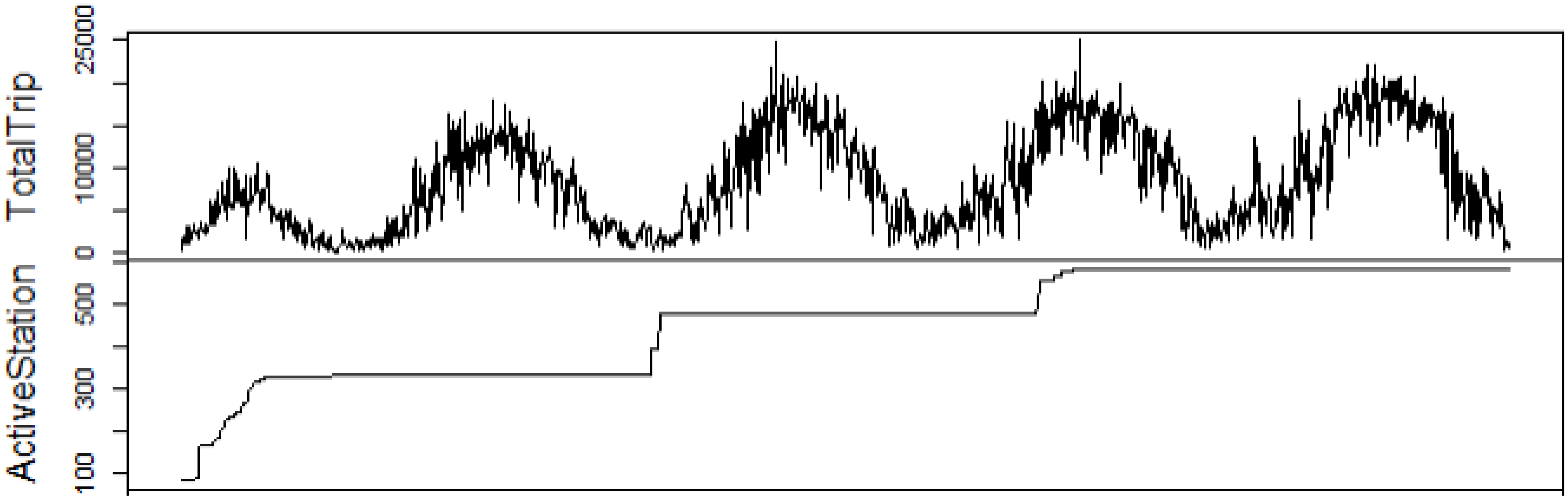
- Enable Divvy to produce detailed annual usage forecasts to assist with business decisions such as
  - Understand key drivers of usage
  - Understand seasonality in usage
  - Number of stations to add during expansion phases

Dataset - Overview

- **Source**
  - https://www.divvybikes.com/system-data
  - https://cran.r-project.org/web/packages/bikedata/vignettes/bikedata.html
- **Description**
  - R package builds database of daily trips from and to each station (matrix)
  - This analysis focuses on total number of outbound trips per day
- **Cleanup**
  - Imputed data for missing dates using average of total trips from adjacent days
  - Removed a single leap year data point to preserve annual seasonality

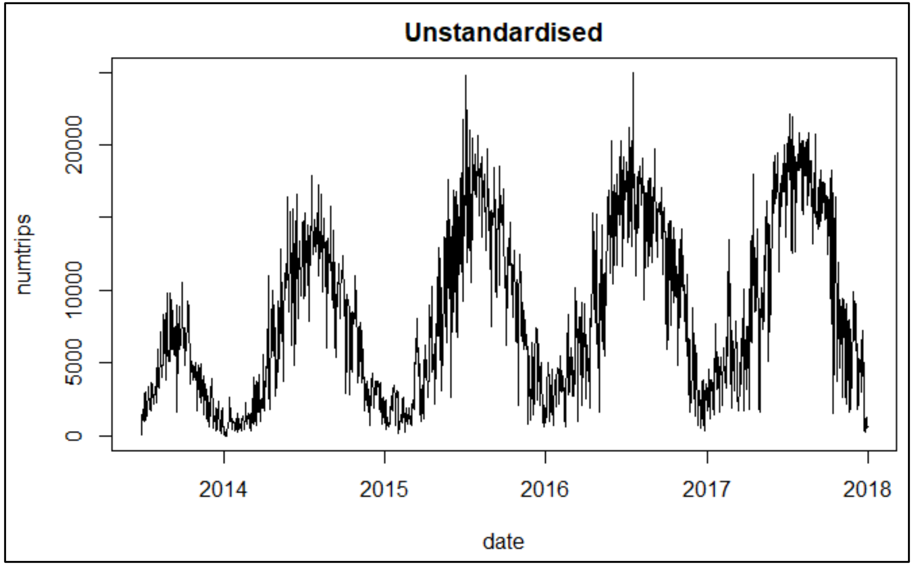
Data Pre-Processing

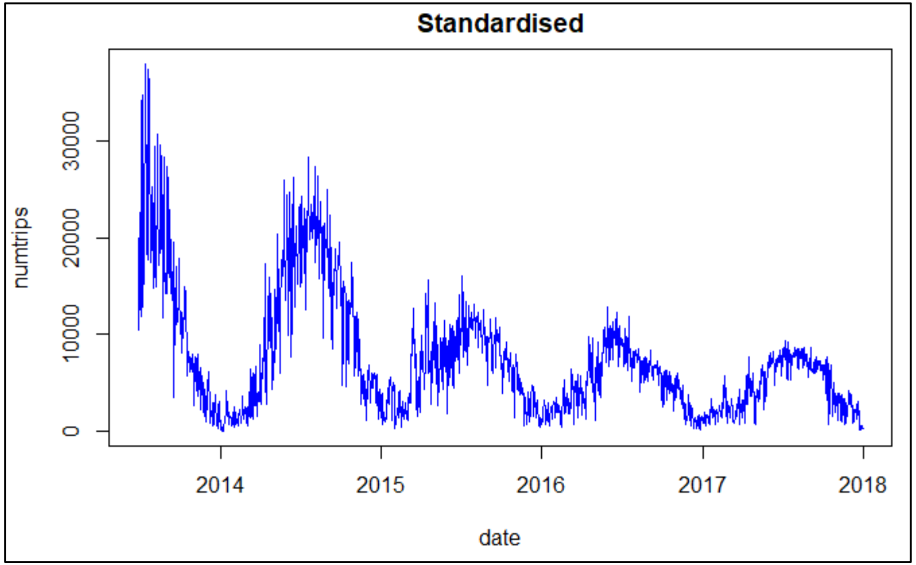
- There is a ***confounding*** effect from the change in number of divvy stations on the time series of total trips



Data Pre-Processing

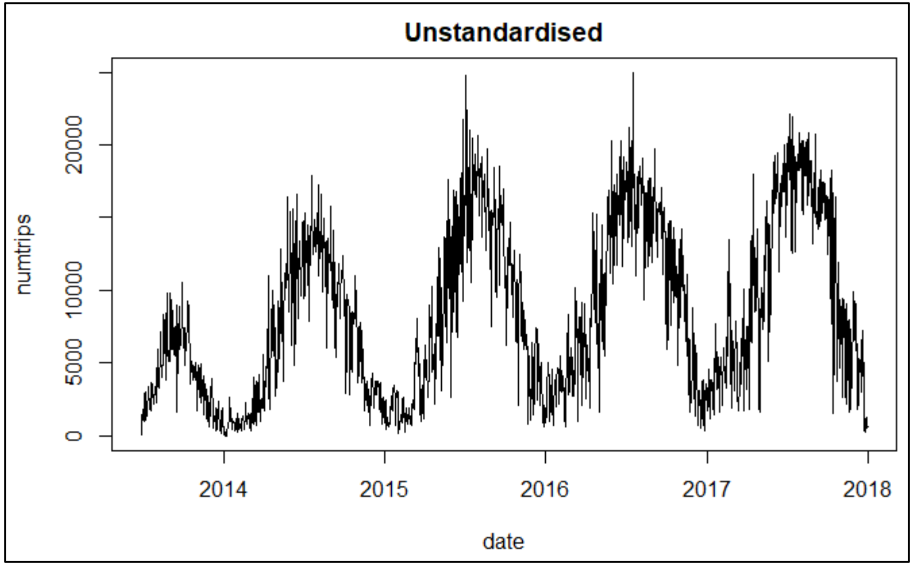
- The R package has a function to generate a "normalized" series, but the detailed mechanics of the normalization process are not fully documented

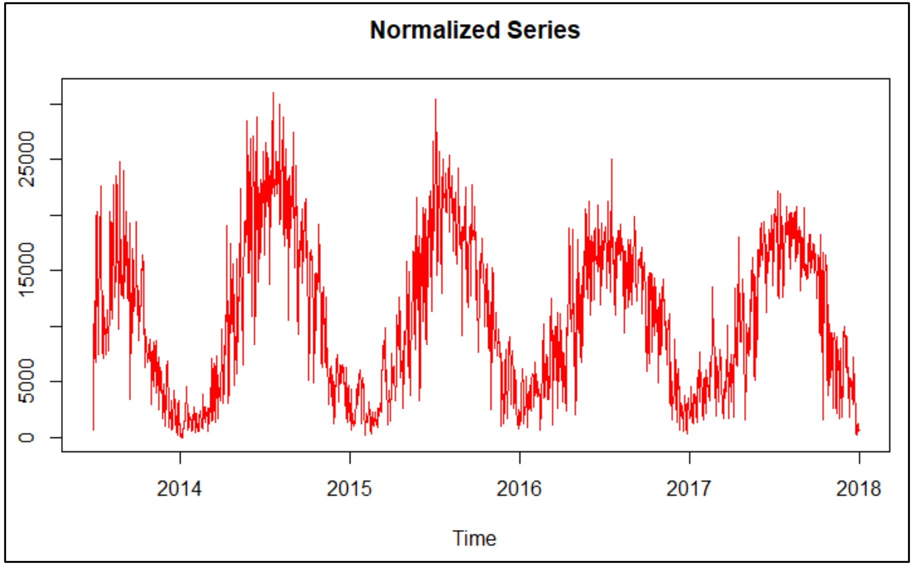




Data Pre-Processing

- We performed our own normalization by scaling each daily trip count by the ratio of the maximum number of active stations to the number of active stations on each specific day



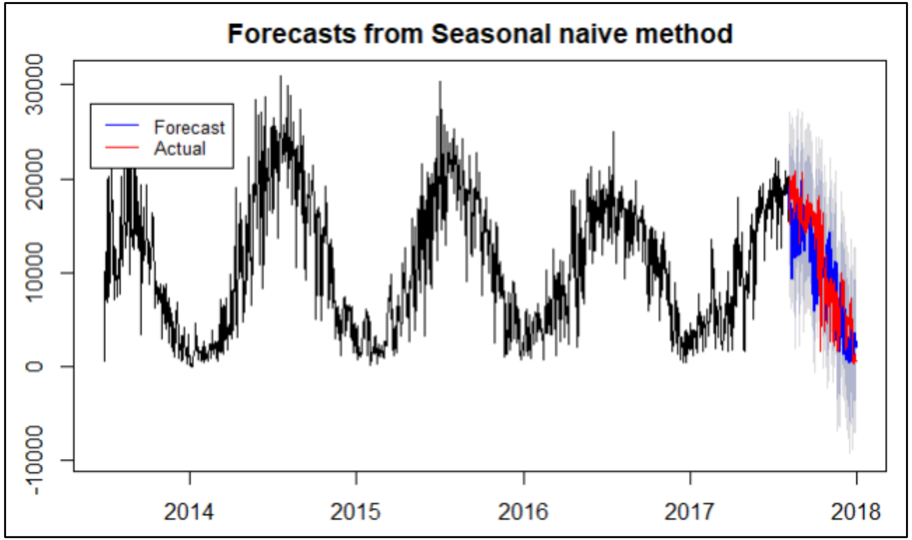


Models - Normalized Series

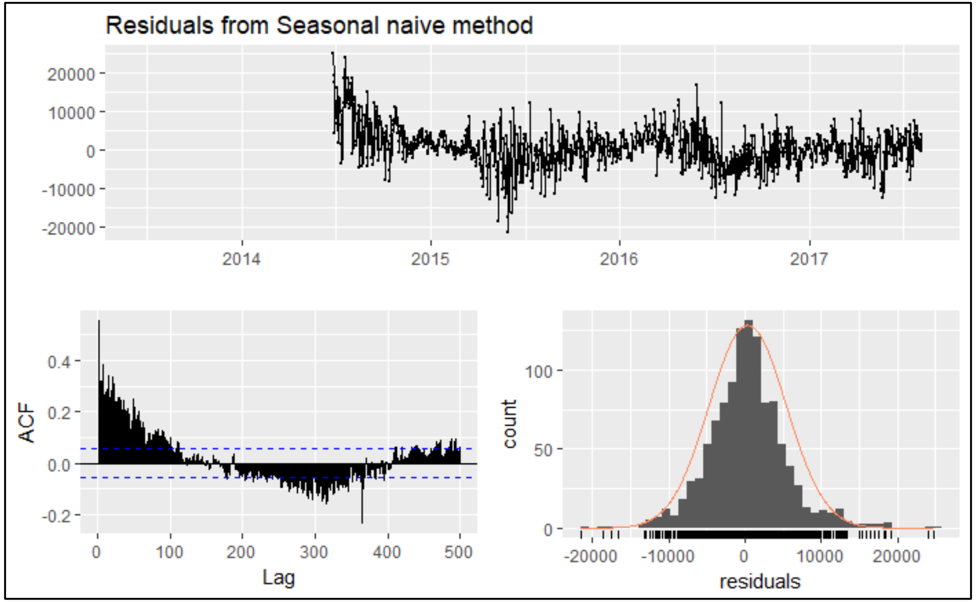
- **Training Period** - June 27, 2013 - August 5, 2017 (1,500 observations)
- **Test Period** - August 6, 2017 - December 31, 2017 (148 observations)
- **Models**
  - sNaive
  - Dynamic Harmonic Regression
  - VAR

Modeling - sNaive

- **Motivation** - Obtain a simple baseline to compare with more advanced models

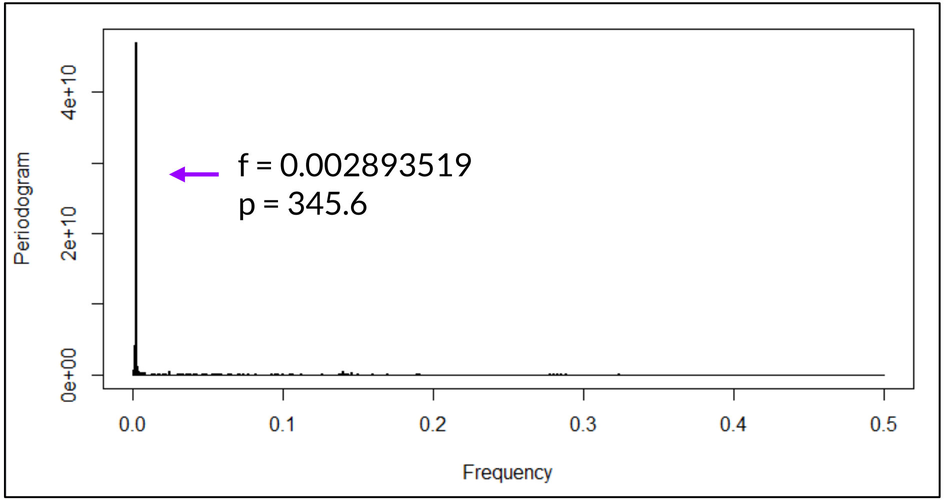


	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	633.186	3960.865	3051.24	-21.61829	53.86553	0.3811809	1.427455



Modeling - Dynamic Harmonic Regression

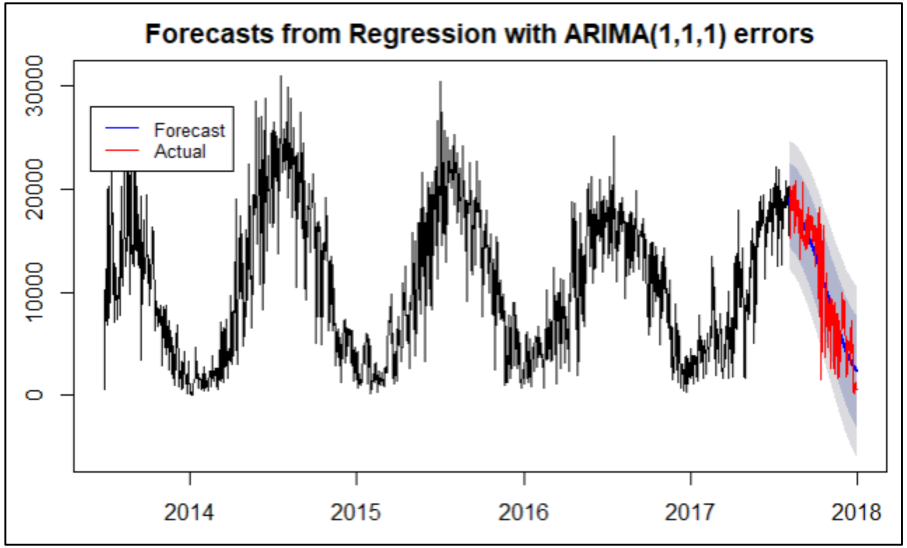
- **Motivation**
  - Initially tried TBATS model, however function was unable to identify trigonometric components
  - Single sharp peak in periodogram suggests single sine-cosine pair
  - Period is close enough to 365

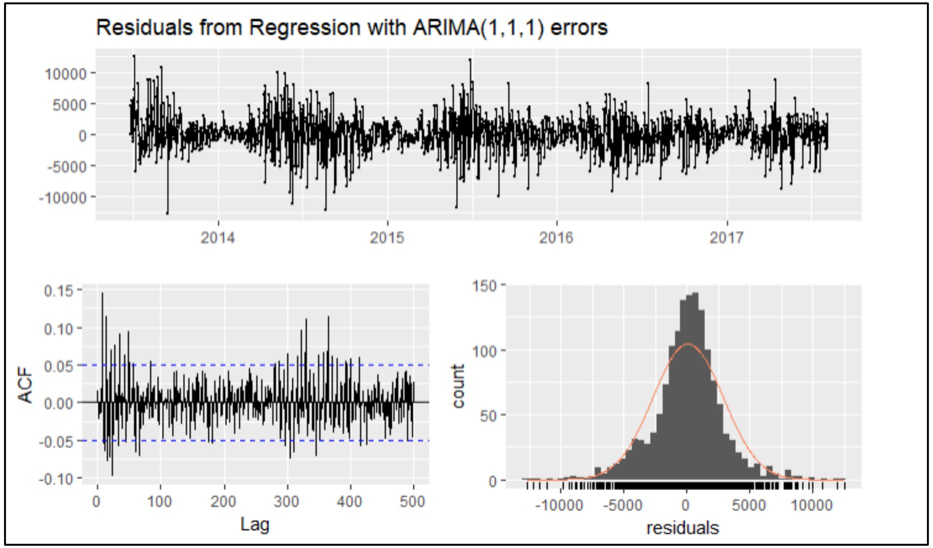



```r
# Dynamic Harmonic Regression
dhg.fit <- list(aicc=Inf)
for (i in 1:25) {
  fit <- auto.arima(train.set, xreg=fourier(train.set, i), seasonal=FALSE)
  if(fit$aicc < dhg.fit$aicc)
    dhg.fit <- fit
}
```

```
Series: train.set
Regression with ARIMA(1,1,1) errors

Coefficients:
         ar1      ma1     S1-365     C1-365
      0.4183  -0.9524  4494.5115  7206.9619
s.e.  0.0265   0.0089   519.0887   518.8199

sigma^2 estimated as 7786130:  log likelihood=-14018.72
AIC=28047.44   AICc=28047.48   BIC=28074

Training set error measures:
                  ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 74.1269 2785.709 2011.926 -29.8146 48.66382 0.5463011 0.01562389
```

Modeling - Dynamic Harmonic Regression - Continued

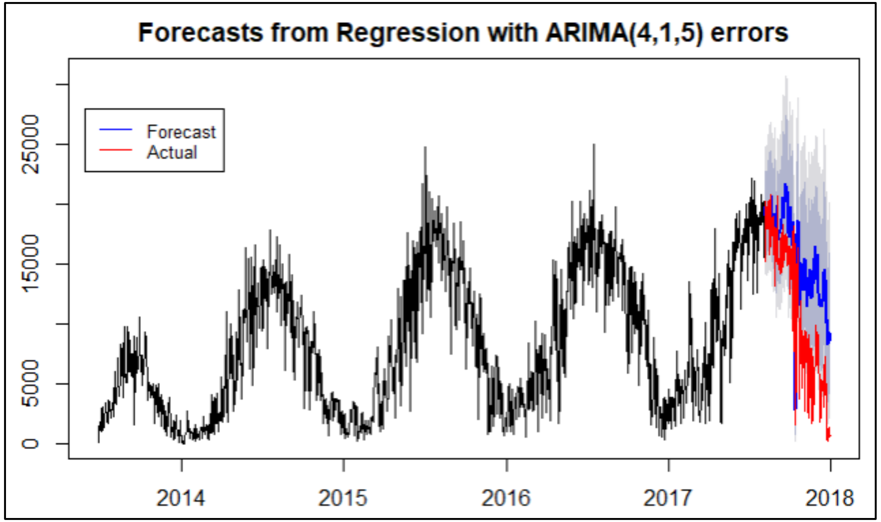




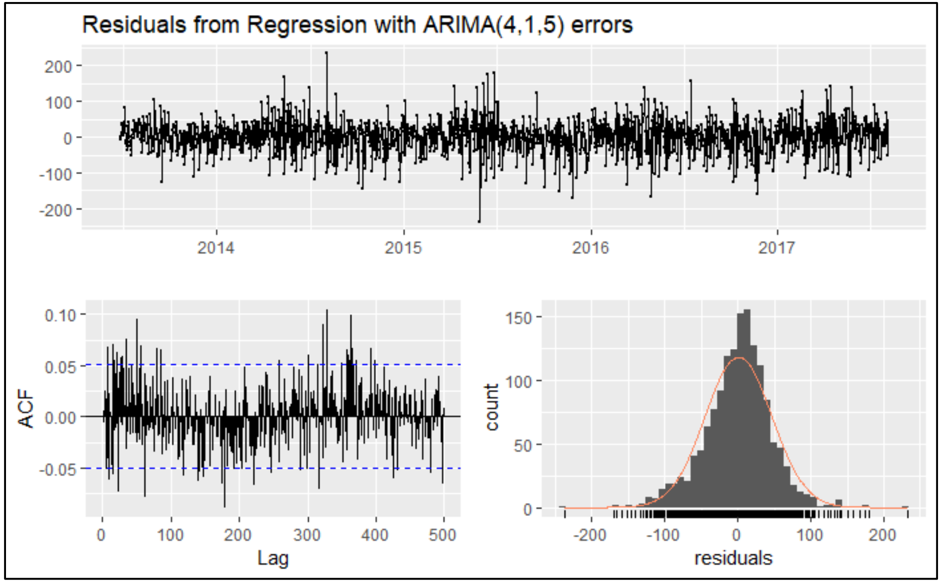
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-34.53916	2691.286	2147.425	-35.51012	54.00101	0.3772269	1.789388

Model - Original Series - XREG w/ ARIMA Errors

- **Training Period** - June 27, 2013 - August 5, 2017 (1,500 observations)
- **Test Period** - August 6, 2017 - December 31, 2017 (148 observations)
- **Motivation** - Forecast the original time series using number of stations as an external regressor
- **External Regressors Used** - Number of Stations, Precipitation, Temperature, Snowfall



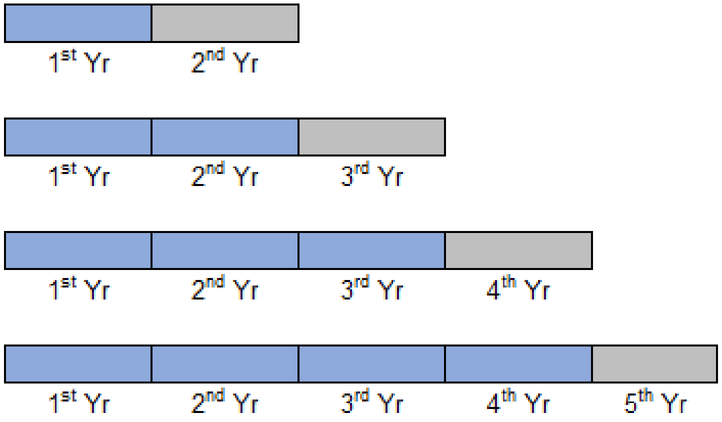
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
-4926.4527898	5905.6195194	5012.6789864	-167.7818331	168.2079261	1.6171297	0.6986942	6.8016341



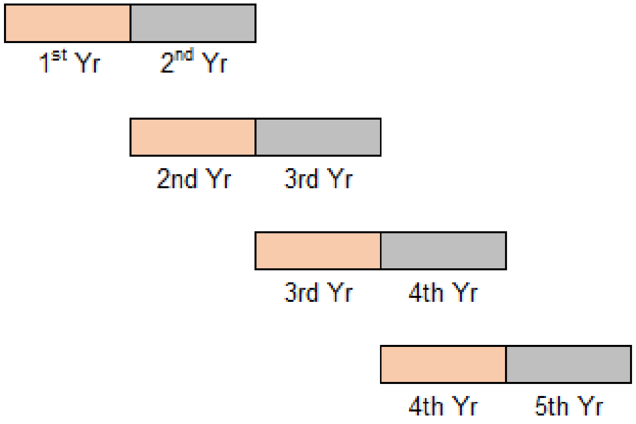
Cross Validation

- **Methodology** - Forecast annual usage using rolling and expanding window
- **Performance Measure** - MAPE

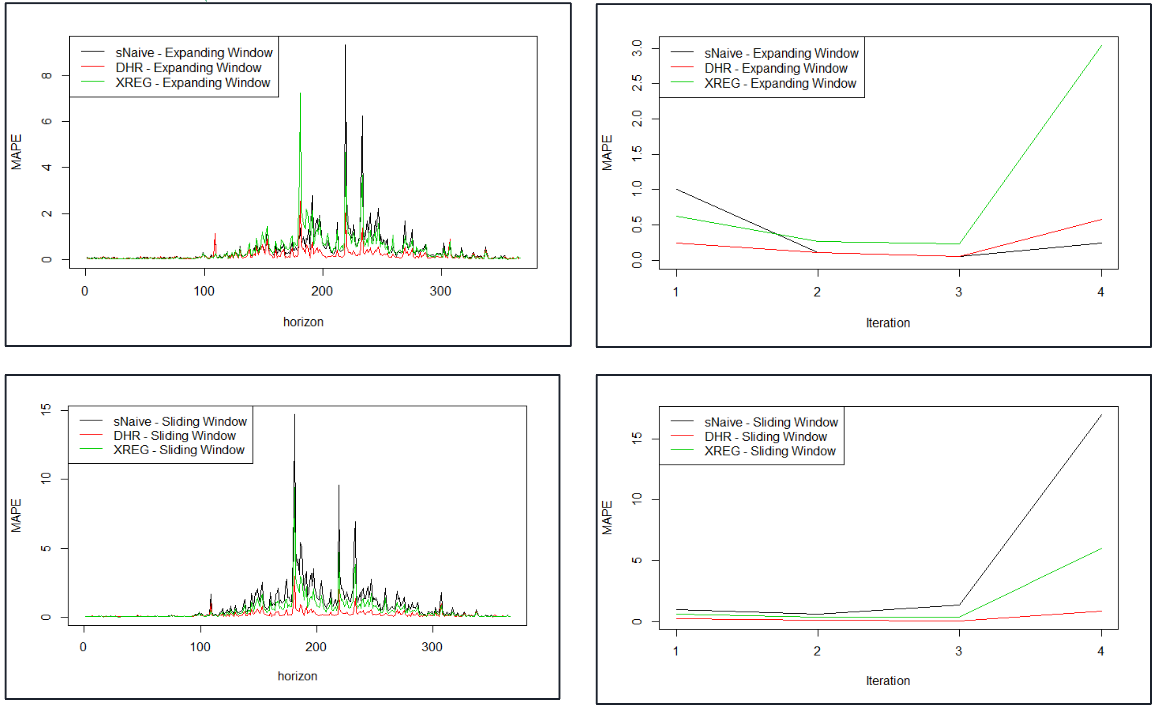
Expanding Window



Sliding Window



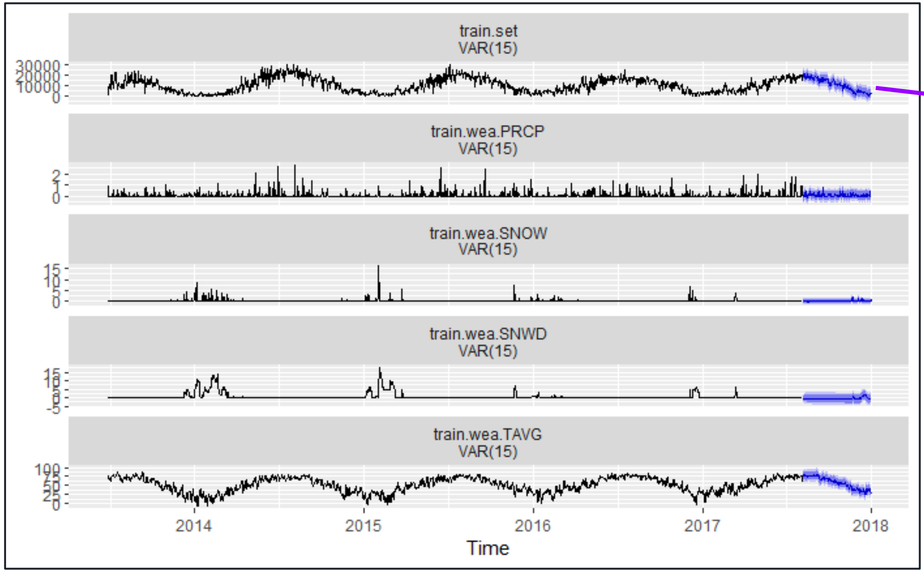
Cross Validation

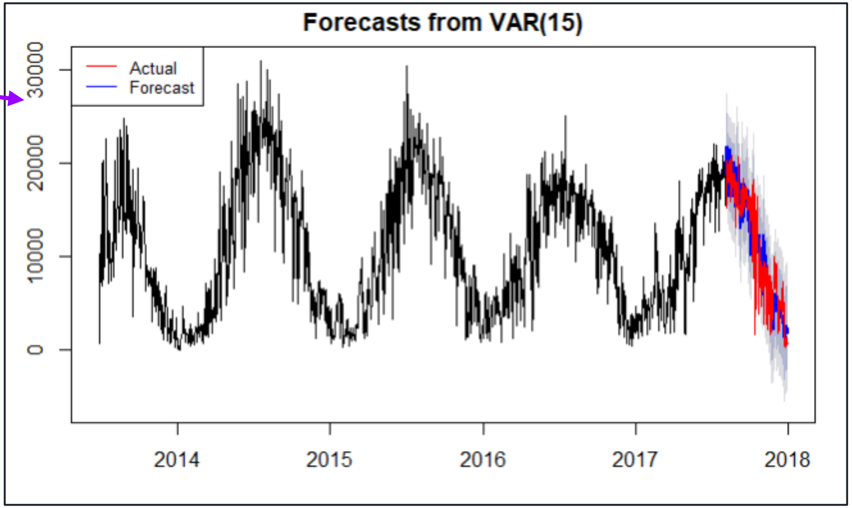


- Dynamic Harmonic Regression consistently yields forecasts with the lowest MAPE across rolling / expanding windows
- Dynamic Harmonic Regression has consistent MAPE across multiple split sizes
- Based on cross-validation performance, and the highly sinusoidal nature of the divvy time series, our **recommendation is the Dynamic Harmonic Regression model**

Future Work - VAR

- **Motivation**
  - **Variables used** – Trips, Precipitation, Temperature, Snowfall, Snow Depth
  - Leverage interdependencies of weather variables to obtain more robust forecasts of daily trips
- **VARselect** gave p=15 with AIC when lag.max=100 -> VAR(15)

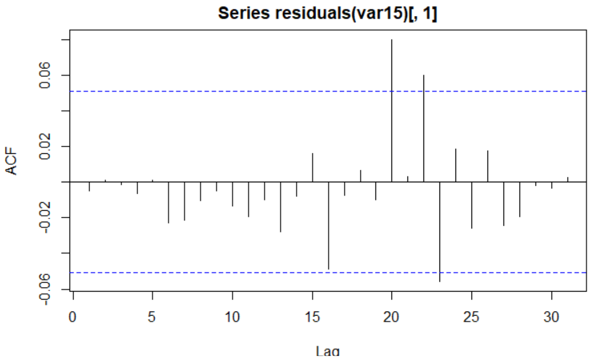




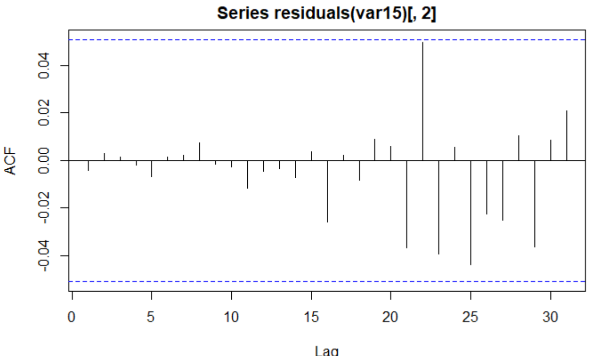
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
-237.1450881	3167.4175365	2460.0438942	-25.7611104	46.4277608	0.6679791	0.3803305	1.2500838

Future Work - VAR Continued

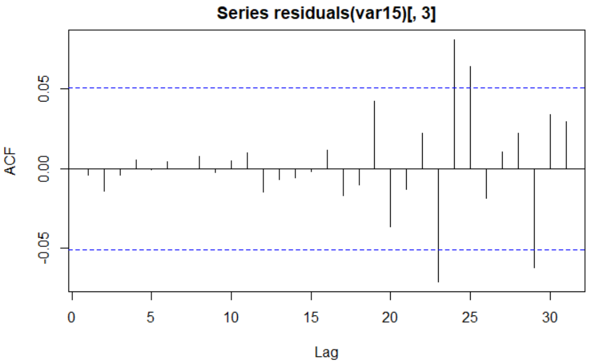
Trips



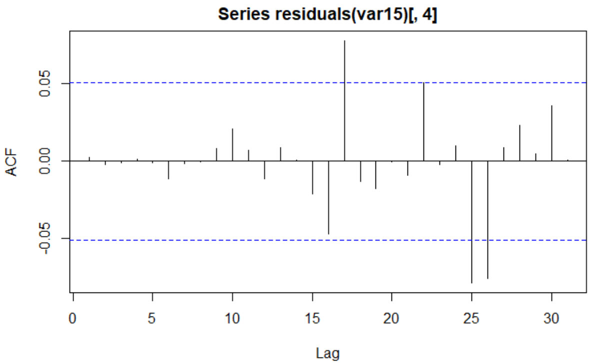
Precipitation

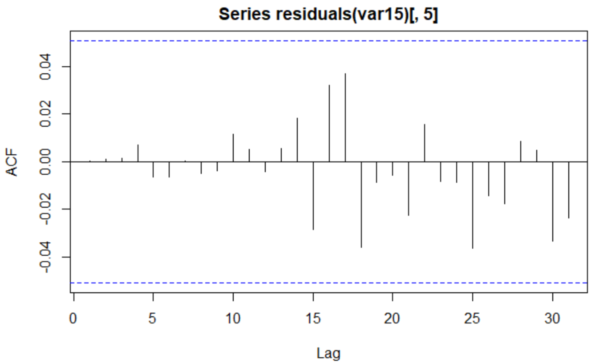


Snow



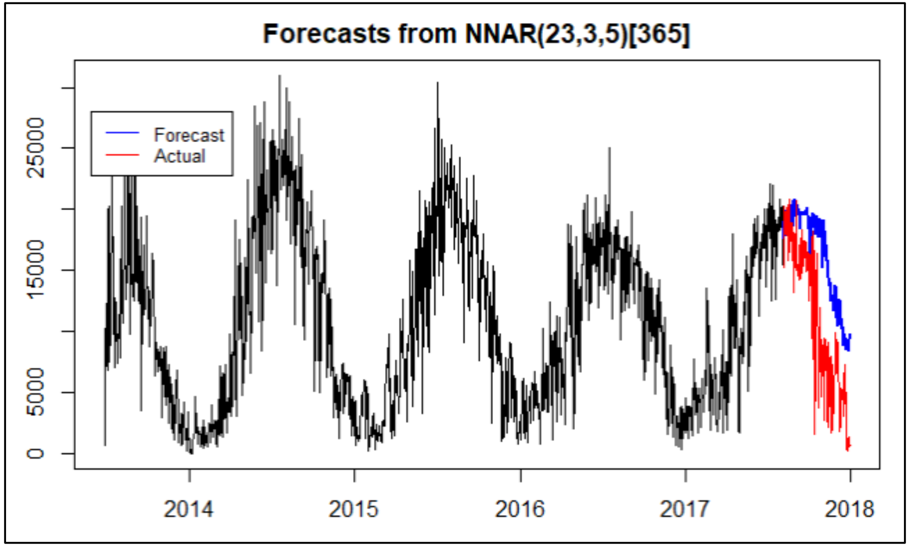
Snow Depth

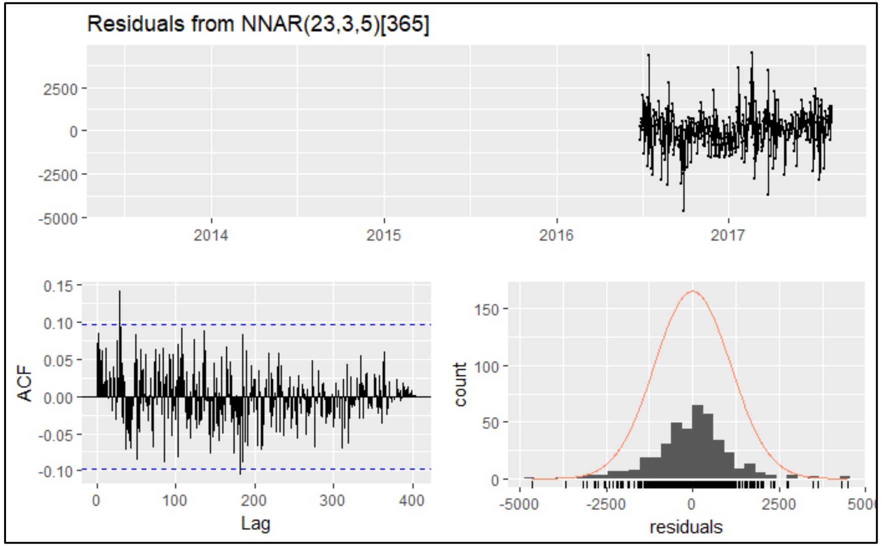




Temperature

Future Work - Neural Networks

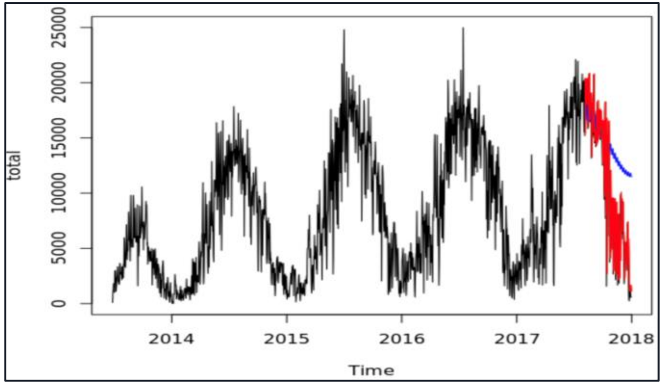




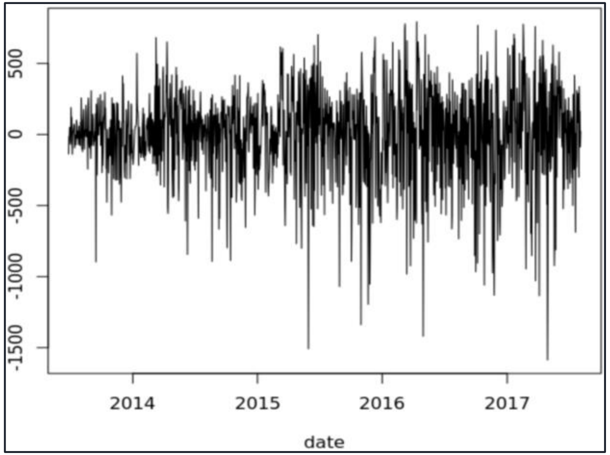
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-4187.294265	5887.264642	4538.766859	-100.5313475	104.2170996	0.7184083126	2.821423848

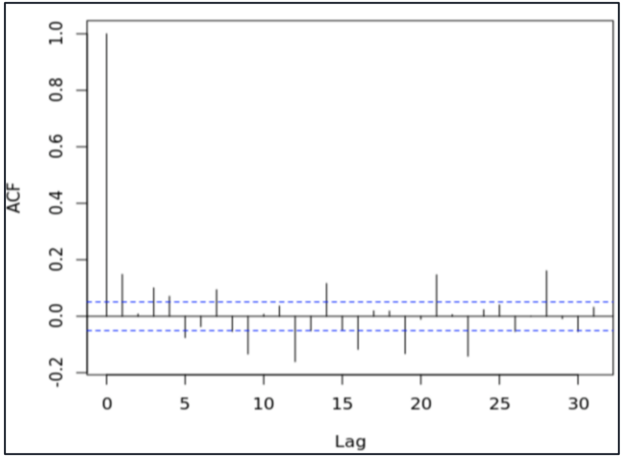
Future Work - Individual Station Forecasts - TBATS

- **Motivation**
  - Optimize bike placement and availability at the station level
- **TBATS – Multiple seasonality (weekly, yearly)**
- **Forecast usage at the station level, then aggregate across all stations to obtain daily total**
  - 584 stations out of 586 stations – 2 stations were built during the forecast period
  - **Run on AWS using multiple cores (~3 hour run time)**



ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
-3111.251	5275.31	4169.82	-98.13128	103.8393	0.7485687	4.167908





Q&A