Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

Problem Statement

- Kaggle competition to predict wikipedia web traffic for individual wikipedia web pages
- The training dataset consists of approximately 145k time series. Each of these time series represent a number of daily views of a different web page
- Predict web page views for a given month using different techniques(auto.arima, sarima,hierarchal, pulse)
- Compare techniques across same Wikipedia page topic. What works best for our time series?

kaggle

WIKIPEDIA
The Free Encyclopedia

Assumptions

- No major news would hit during test period on our subject, there will not be a major pulse during this time
- There could be other datasets closely related to ours that impact this dataset e.g. Microsoft
    - Not modeling or factoring in this data
- The hosting server was never out for parts of any days in the data set

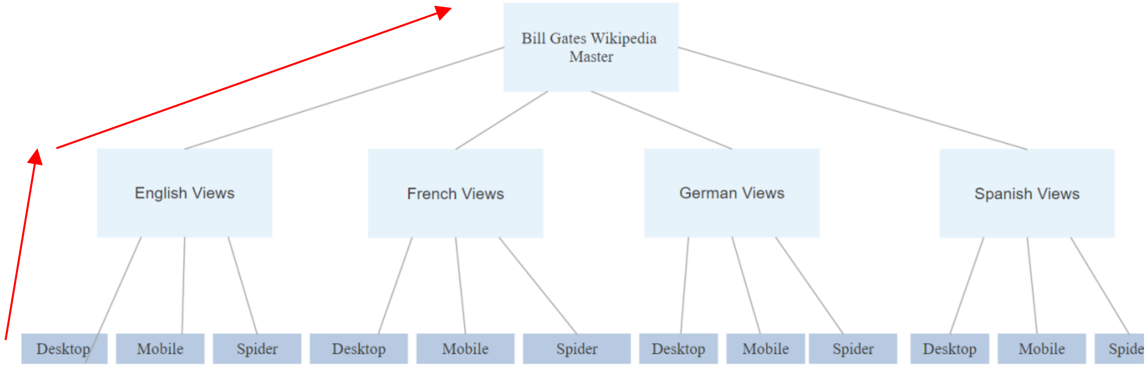
Data Properties & Transformations

- Data organized into separate categories
  - all, mobile, desktop, spider, etc
  - Languages: English, French, Russian, Spanish, etc
- Date range of July 2015 to September of 2017
- Transpose data to get into correct time series format
- Data is spread out; data from same overarching topic at line 5 and 3000
- Created a function used to search dataframe for topics that belong together
- Need to convert selected data into a time series format with a week long frequency

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Page	7/1/2015	7/2/2015	7/3/2015	7/4/2015	7/5/2015	7/6/2015	7/7/2015	7/8/2015	7/9/2015	7/10/2015	7/11/2015	7/12/2015	7/13/2015	7/14/2015	7/15/2015	7/16/2015	7/17/2015	7/18/2015
2	2NE1_zh.wikipedia.	18	11	5	13	14	9	9	22	26	24	19	10	14	15	8	16	8	8
3	2PM_zh.wikipedia.(	11	14	15	18	11	13	22	11	10	4	41	65	57	38	20	62	44	15
4	3C_zh.wikipedia.or	1	0	1	1	0	4	0	3	4	4	1	1	1	6	8	6	4	5
5	4minute_zh.wikipe	35	13	10	94	4	26	14	9	11	16	16	11	23	145	14	17	85	4
6	52_Hz_I_Love_You_zh.wikipedia.org_all-access_spider																		
7	5566_zh.wikipedia.	12	7	4	5	20	8	5	17	24	7	12	11	7	9	6	10	8	13
8	91Days_zh.wikipedia.org_all-access_spider																		
9	A'N'D_zh.wikipedia	118	26	30	24	29	127	53	37	20	32	17	23	47	33	47	58	29	187
10	AKB48_zh.wikipedi	5	23	14	12	9	9	35	15	14	22	8	16	18	12	14	14	7	7
11	ASCII_zh.wikipedia.	6	3	5	12	6	5	4	13	9	15	18	7	8	12	25	23	6	10
12	ASTRO_zh.wikipedia.org_all-access_spider						1	1							1			1	0
13	Ahq_e-Sports_Club_	2	1	4	4	2	6	3	6	9	11	8	8	5	6	5	10	3	9
14	All_your_base_are_	2	5	5	1	3	3	5	3	17	3	9	10	3	8	8	5	4	2
15	AlphaGo_zh.wikipedia.org_all-access_spider																		
16	Android_zh.wikipec	8	27	9	25	25	10	34	22	17	45	27	17	19	32	19	58	19	4
17	Angelababy_zh.wik	40	17	25	42	41	7	18	21	33	15	58	38	39	28	19	4	19	43
18	Apink_zh.wikipedia	61	33	21	10	26	11	39	195	62	18	52	17	41	28	33	52	75	5
19	Apple_II_zh.wikipec	4	8	4	9	7	4	15	9	17	16	10	4	8	17	11	13	12	21
20	As_One_zh.wikiped	13	7	14	11	20	5	32	11	6	4	15	9	29	231	40	45	9	33
21	B-PROJECT_zh.wikipedia.org_all-access_spider																		
22	B1A4_zh.wikipedia.	22	11	23	10	6	12	74	17	38	23	18	17	14	21	16	17	15	19
23	BDSM_zh.wikipedia	25	3	3	4	12	14	16	15	22	23	19	3	42	7	7	12	11	7
24	BEAST_zh.wikipedia	19	6	12	14	13	7	12	64	9	31	23	20	13	48	39	15	19	24
25	BIGBANG_zh.wikip	23	24	31	9	21	27	15	8	50	78	74	35	27	19	11	17	34	14
26	BLACK_PINK_zh.wikipedia.org_all-access_spider																		
27	BLEACH_zh.wikipec	11	5	13	8	6	5	8	5	12	3	10	17	6	6	37	15	29	8
28	BTOB_zh.wikipedia	22	67	26	34	38	13	17	33	43	32	43	26	44	37	17	18	33	47
29	Beautiful_Mind_zh.wikipedia.org_all-access_spider																		
30	Beyond_zh.wikiped	291	64	26	20	28	6	20	10	48	17	14	10	19	17	9	14	12	9
31	Big_zh.wikipedia.or	3	53	11	3	4	3	11	9	5	16	19	3	11	14	14	8	14	2

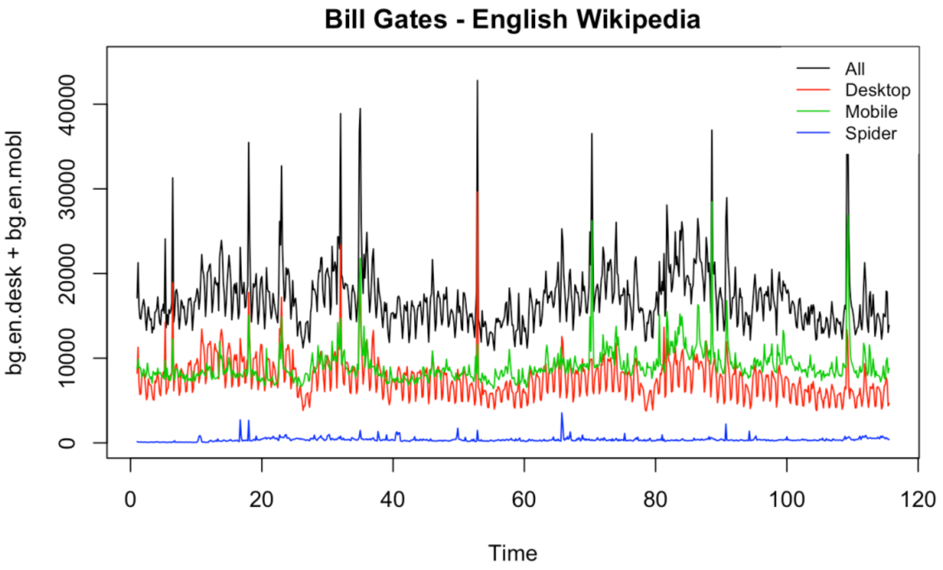
Data Preprocessing

- Check for any missing values
- Convert data into time series objects with a frequency of 7
- Group data by language
- Aggregate all languages to get the total traffic of Bill Gates Wikipedia page



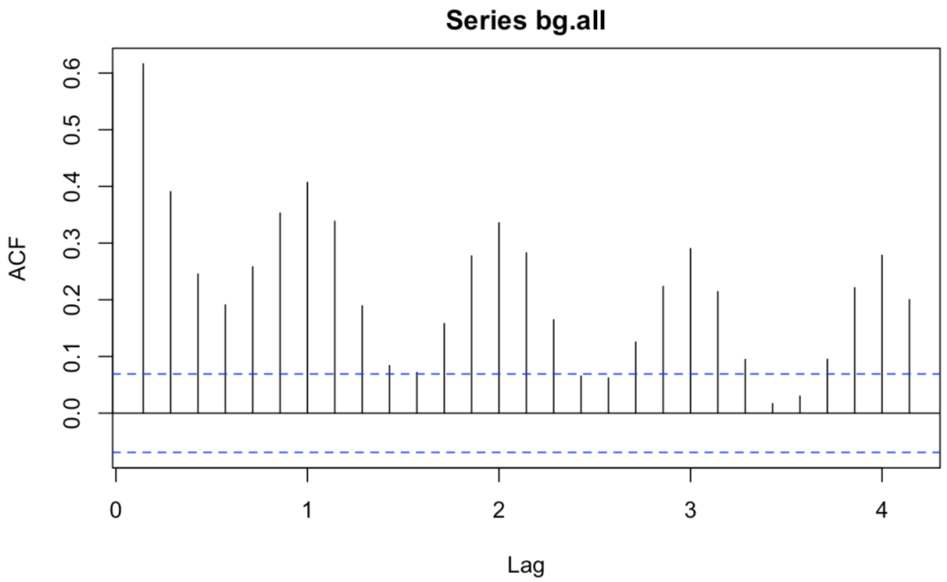
Exploratory Data Analysis

- Red is desktop - can tell that has more weekly activity
- Blue - spider - web crawler is more consistent
- English makes up more than half of the view
- Many spikes could be attributed to news on Gates
    - A few spikes in dataset including in July of 2017 when Gates no longer richest person in the world



Exploratory Data Analysis

- > kpss.test(Bill.Gates.All)
  p-value = 0.01213
- Time series is stationary
- Acf shows seasonality



Proposed Approaches (Models)

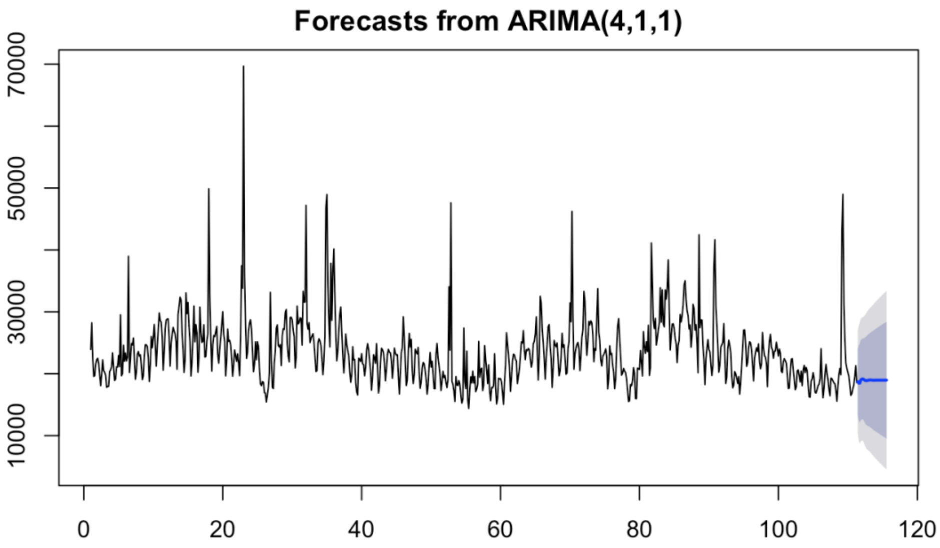
- Bill Gates Wikipedia Page Views
    - Models
        - Nonseasonal auto.arima (baseline)
        - Seasonal auto.arima
        - Hierarchical
        - Pulse
    - Our thesis is bottom’s up hierarchal will produce the best results
    - Testing last 30 days of dataset
    - Analysis of 12 time series of data:
        - French - Desktop, Mobile, Spider
        - English - Desktop, Mobile, Spider
        - German - Desktop, Mobile, Spider
        - Spanish - Desktop, Mobile, Spider

Proposed Solution (Model Selection)

Non-seasonal Arima

> auto.arima(train, seasonal = F, stepwise = F)
ARIMA(4,1,1)

AIC: 15,049
Test RMSE: 3,251
Test MAE: 2,531
Test MPE: 6.81

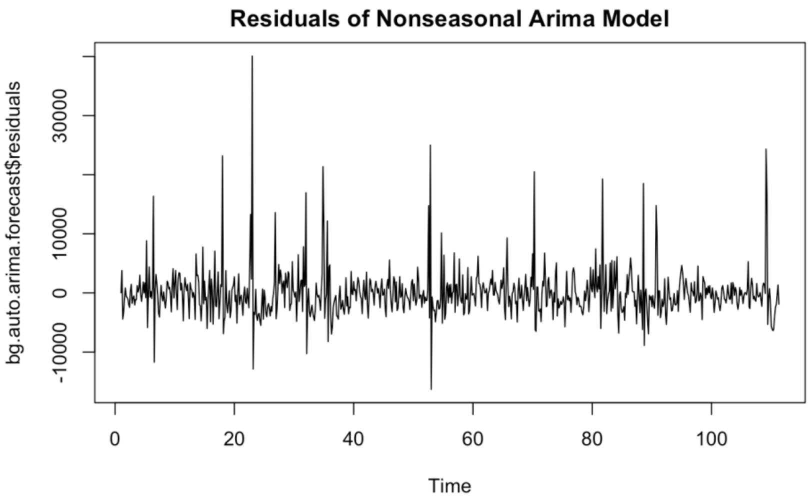


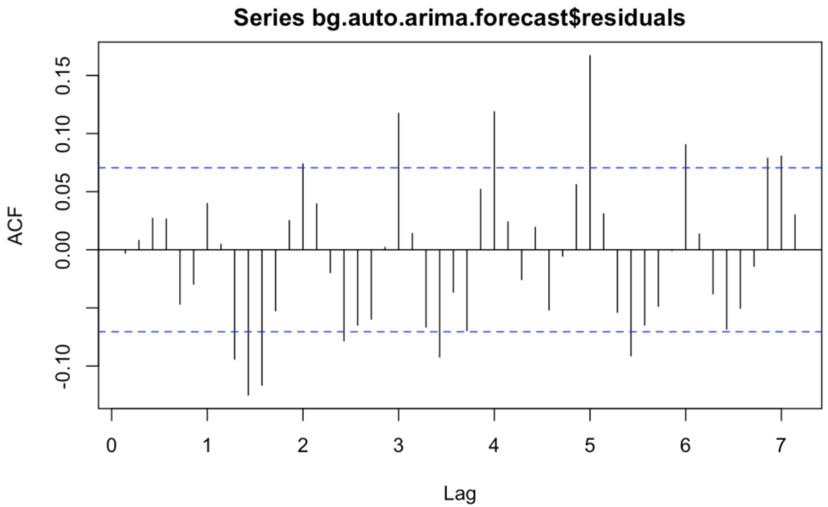
Non-seasonal Arima

>kpss.test(forecast$residuals)
p-value = 0.1

>Box.test(forecast$residuals, lag = 14)
p-value = 0.0001388

>mean(forecast$residuals)
-22.45



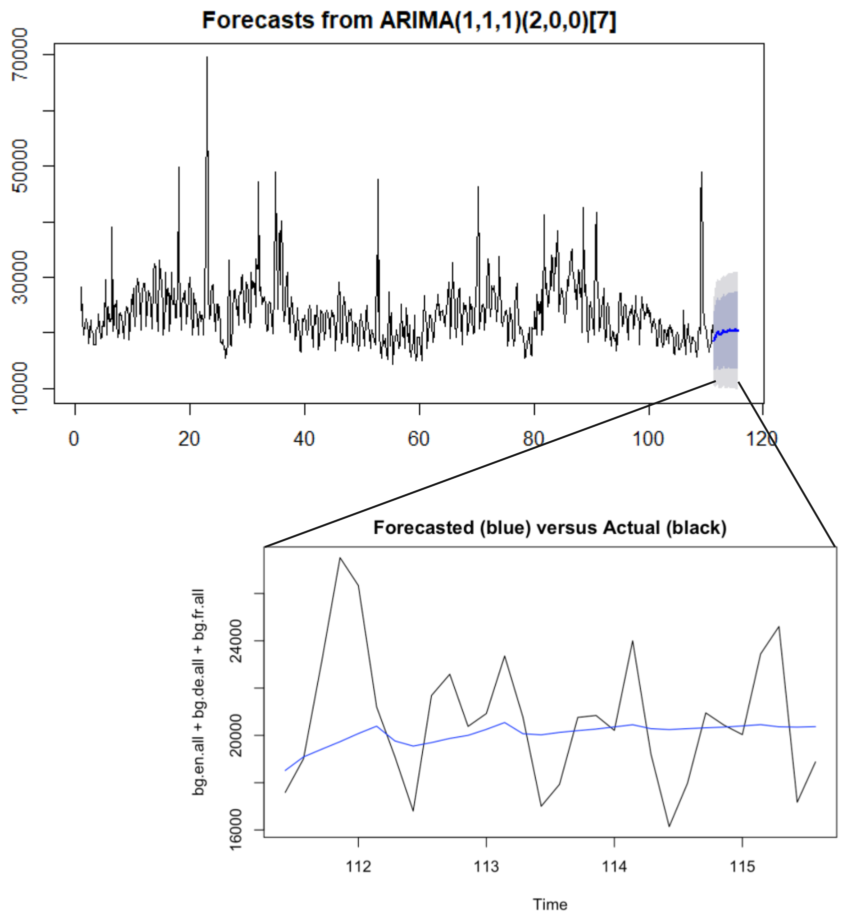


Seasonal Arima

> auto.arima(train, stepwise = F)
ARIMA(1,1,1)(2,0,0)[7]

AIC: 15,023
Test RMSE: 2,791
Test MAE: 2,093
Test MPE: 1.30

The forecasted values in blue are less volatiles than the actual swings in number of views

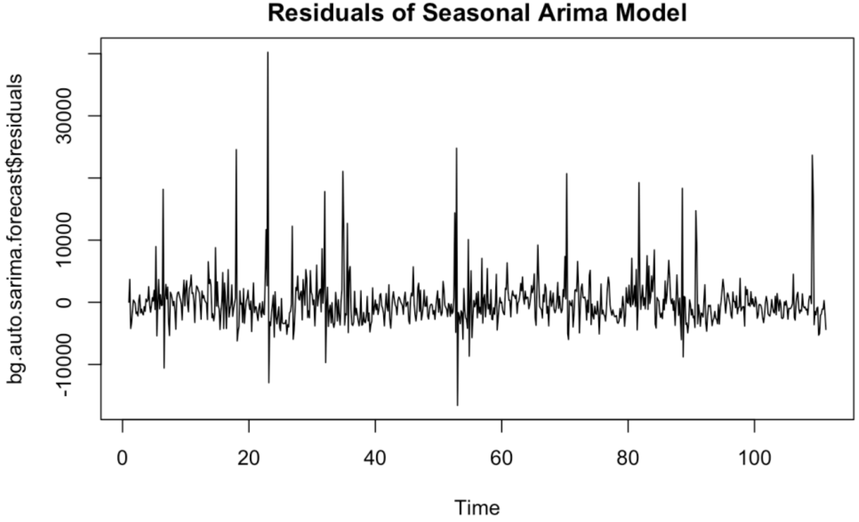


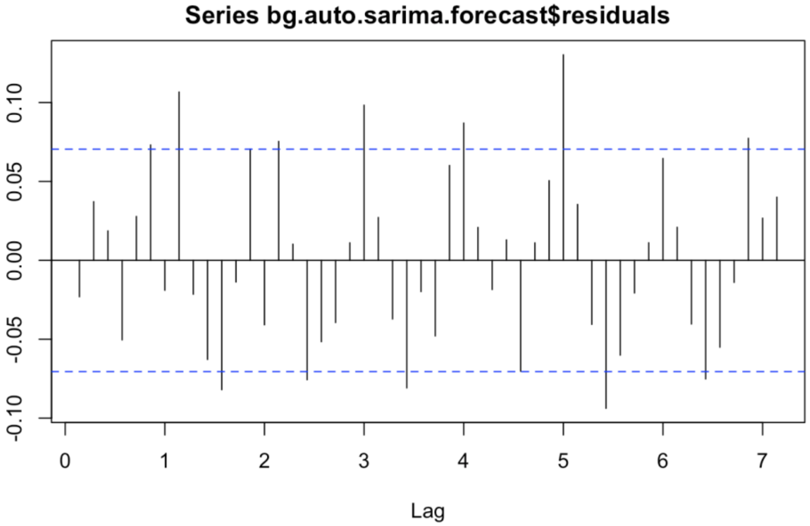
Seasonal Arima

>kpss.test(forecast$residuals)
p-value = 0.1

>Box.test(forecast$residuals, lag = 14)
p-value = 0.004184

>mean(forecast$residuals)
-31.12





Hierarchical

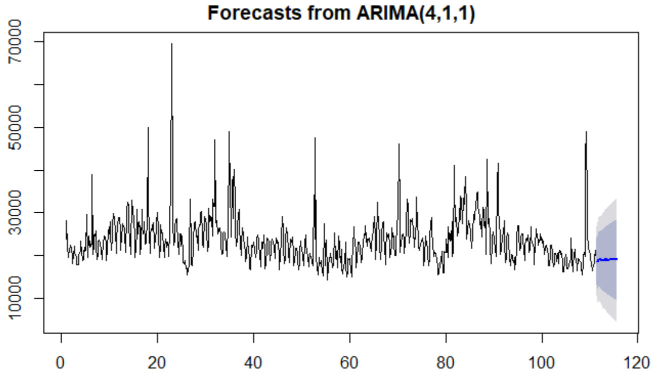
- Bottom-Up Approach gave best results
- library("hts")
- hts(train, characters = c(6,4))
  - Character argument: need to be careful on how the time series are named
- 3 layers
- The bottom layer consists of individual prediction for each languages page from each platform (i.e. mobile, spider, desktop)
- These predictions are then rolled up into their language and then total views

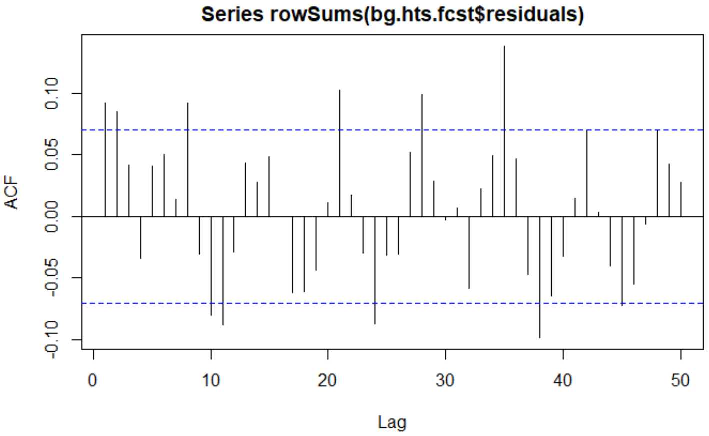
Hierarchical (Forecast and Accuracy)

>accuracy.gts(bg.hts.fcst, test, level = 0)

Level zero refers to aggregated time series

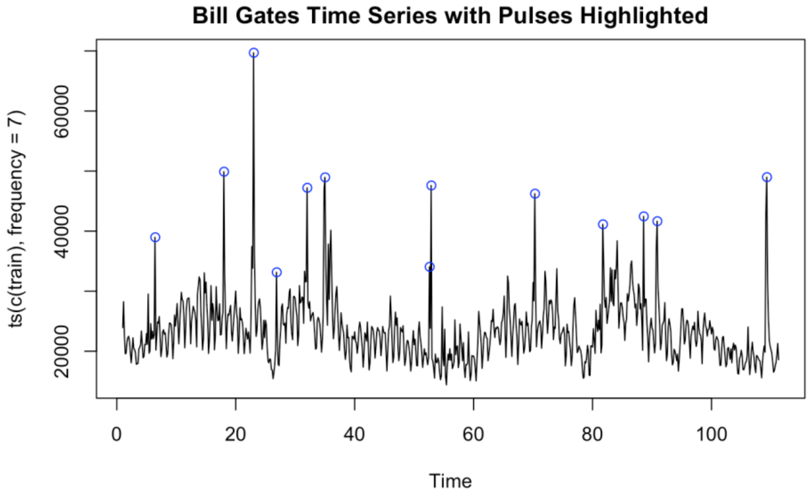
Test RMSE: 2,686
Test MAE: 2,041
Test MPE: -0.12

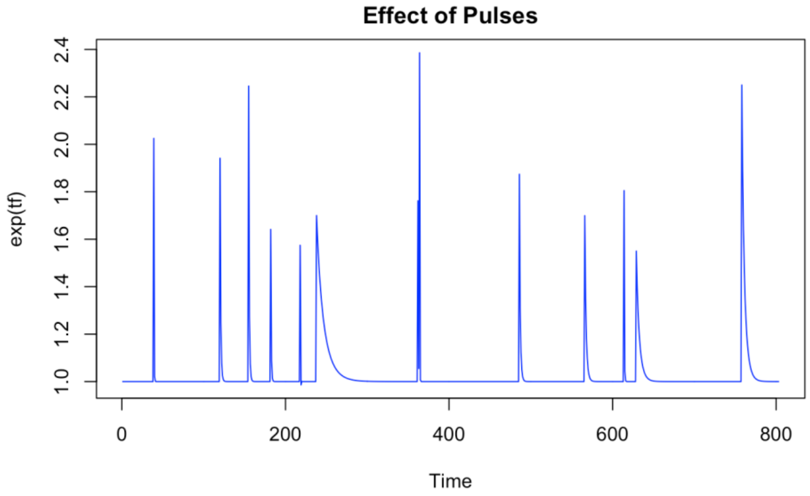




Pulse

- 13 pulses in the time series (train)

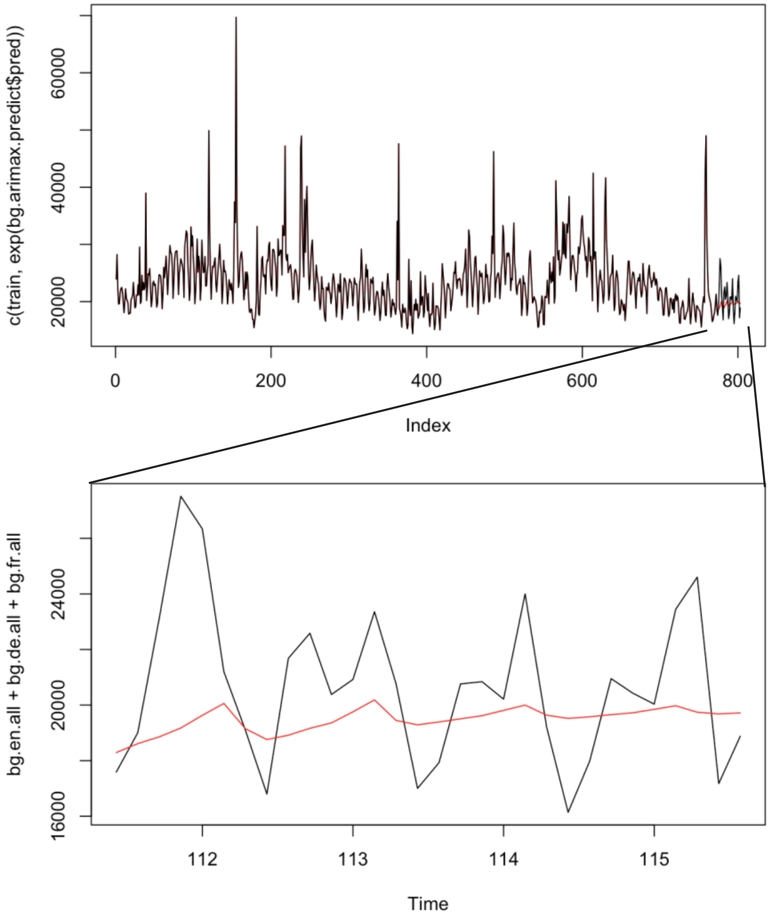




Pulse

> arimax()
> filter()
> Arima()
> predict()

AIC: 1,298
Test RMSE: 3,728
Test MAE: 2,876
Test MPE: 11.9



Results (Accuracy)

- Recommendation: Hierarchical model
- Slightly better than SARIMA, out edges out in all major metrics
- Hierarchical is more time consuming so as dataset gets larger it will be less computationally efficient

Validation of Test/Holdout	Non-Seasonal Arima	Seasonal Arima	Hierarchical	Pulse Intervention
Mean Absolute Error	2530.8	2092.5	2041.4	2876.5
Mean Percent Error	6.81	1.30	-0.12	11.94
Root Mean Square Error	3251.0	2791.2	2686.0	3727.5

Future Work

- Test out the models on different time series in the dataset
  - Fourier model
  - Try various number of pulses
- Are certain models consistently better for this type of problems?
- Compare accuracy of this prediction with a neural net on entire data set
  - Would modeling the rest of the time series create a better prediction? Is there any relationship between time series
- Leader in Kaggle competition used a RNN seq2seq model.