KDD Cup 2017

Speaker：陳袓儒、周鴻汶、王仁緯
Prof.：李漢銘 老師

Divide the work

Name	Data processing	Train model
陳昶儒	v	v
周鴻汶	v	v
王仁緯	v	v

Outline

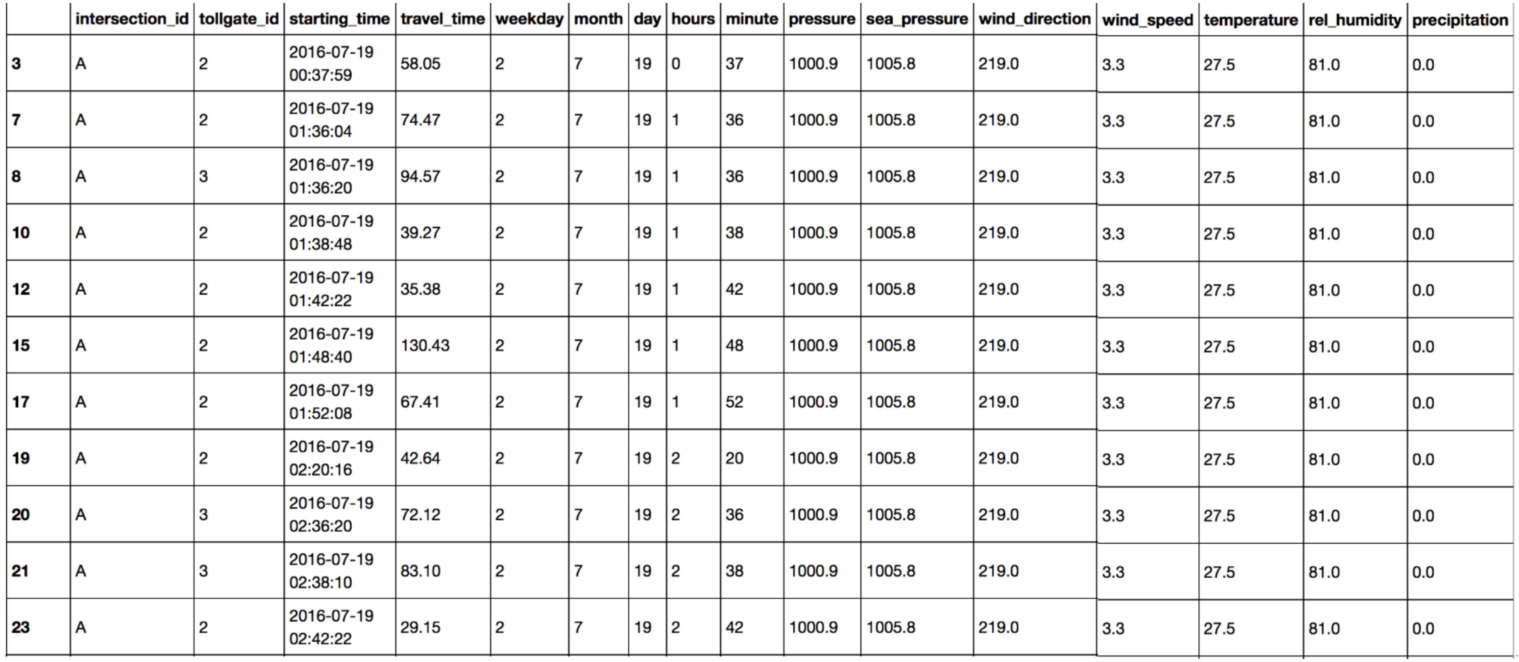
- Task 1 (travel time)
  - Software Platform
  - Data processing
  - Phase1
  - Phase2
- Task 2 (volume)

Task 1 (travel time)

Software Platform

Data processing

- combine the trajectories and weather data
- split dataset and use them to train different models
  - Routes from Intersection A to Tollgates 2 & 3;
  - Routes from Intersection B to Tollgates 1 & 3;
  - Routes from Intersection C to Tollages 1 & 3.



	intersection_id	tollgate_id	starting_time	travel_time	weekday	month	day	hours	minute	pressure	sea_pressure	wind_direction	wind_speed	temperature	rel_humidity	precipitation
3	A	2	2016-07-19 00:37:59	58.05	2	7	19	0	37	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
7	A	2	2016-07-19 01:36:04	74.47	2	7	19	1	36	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
8	A	3	2016-07-19 01:36:20	94.57	2	7	19	1	36	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
10	A	2	2016-07-19 01:38:48	39.27	2	7	19	1	38	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
12	A	2	2016-07-19 01:42:22	35.38	2	7	19	1	42	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
15	A	2	2016-07-19 01:48:40	130.43	2	7	19	1	48	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
17	A	2	2016-07-19 01:52:08	67.41	2	7	19	1	52	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
19	A	2	2016-07-19 02:20:16	42.64	2	7	19	2	20	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
20	A	3	2016-07-19 02:36:20	72.12	2	7	19	2	36	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
21	A	3	2016-07-19 02:38:10	83.10	2	7	19	2	38	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0
23	A	2	2016-07-19 02:42:22	29.15	2	7	19	2	42	1000.9	1005.8	219.0	3.3	27.5	81.0	0.0

Phase1

4-12：use XGboost

- Use feature: weekday,hours,minute,temperature,wind_speed,wind_direction, rel_humidity.
- Drop tarvel_time > 500
- max_depth=2 , n_estimators=250, learning_rate=0.01
- Testing data MAPE : 0.1821
  - Routes A2: 0.1791
  - Routes A3: 0.1935
  - Routes B1: 0.1716
  - Routes B3: 0.1899
  - Routes C1: 0.1445
  - Routes C3: 0.2142
- Predict data MAPE: 0.1846

4-14: use XGboost

- Use feature: weekday,hours,minute,temperature,pressure,rel_humidity.
- Drop tarvel_time > 500
- max_depth=mix , n_estimators=250, learning_rate=0.01
- max_depth=2 , n_estimators=250, learning_rate=0.01
- Testing data MAPE : 0.1796/0.1799
  - Routes A2: 0.1791
  - Routes A3: 0.1900/0.1914
  - Routes B1: 0.1733/0.1737
  - Routes B3: 0.1737
  - Routes C1: 0.1446
  - Routes C3: 0.2172
- Predict data MAPE: 0.1869 / 0.1846

4-18: Mix 4-12,4-14

- 4-12:
  - Model B1,B3,C1,C3
- 4-14
  - Model A2,A3
- Predict data MAPE: 0.1845

4-19: Change model

- Let model A2,A3,B1,B3,C1,C3 change to:
- A2 ： A2M,A2T,A2W,A2R,A2F,A2A,A2S
- A3 ： A3M,A3T,A3W,A3R,A3F,A3A,A3S
- B1 ： B1M,B1T,B1W,B1R,B1F,B1A,B1S
- B3 ： B3M,B3T,B3W,B3R,B3F,B3A,B3S
- C1 ： C1M,C1T,C1W,C1R,C1F,C1A,C1S
- C3 ： C3M,C3T,C3W,C3R,C3F,C3A,C3S

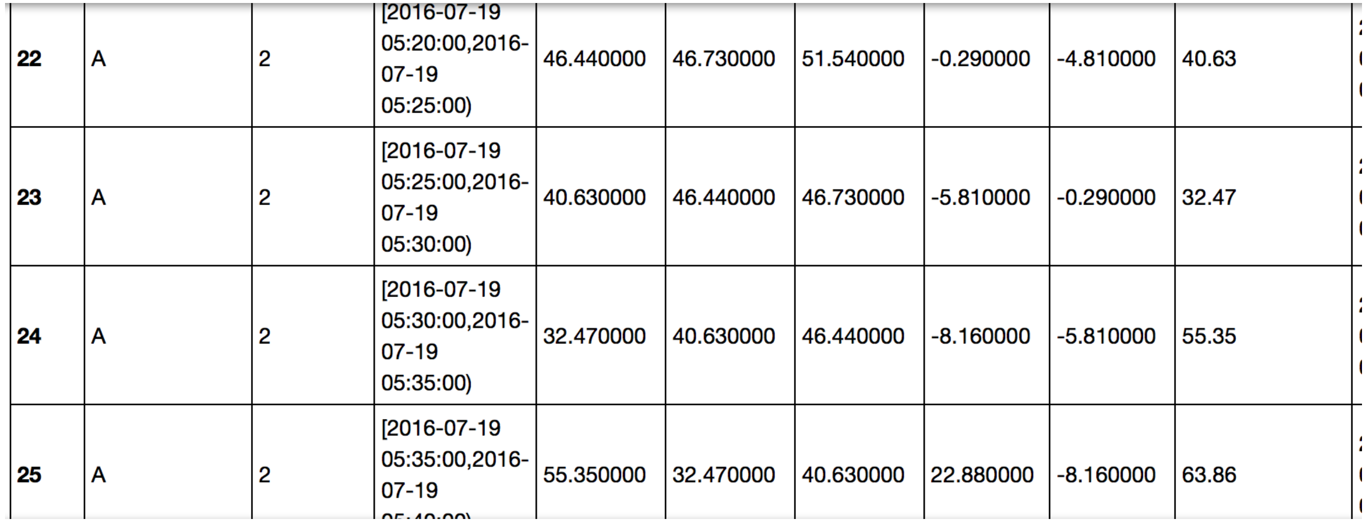
- Predict data MAPE: 0.215

5-3: Cross validation by 4-18 model

- Cv_folds=5
- Metrics=MAE
- Testing data MAPE:0.185~0.19
- Predict data MAPE:0.2005

5-9: Select New Feature

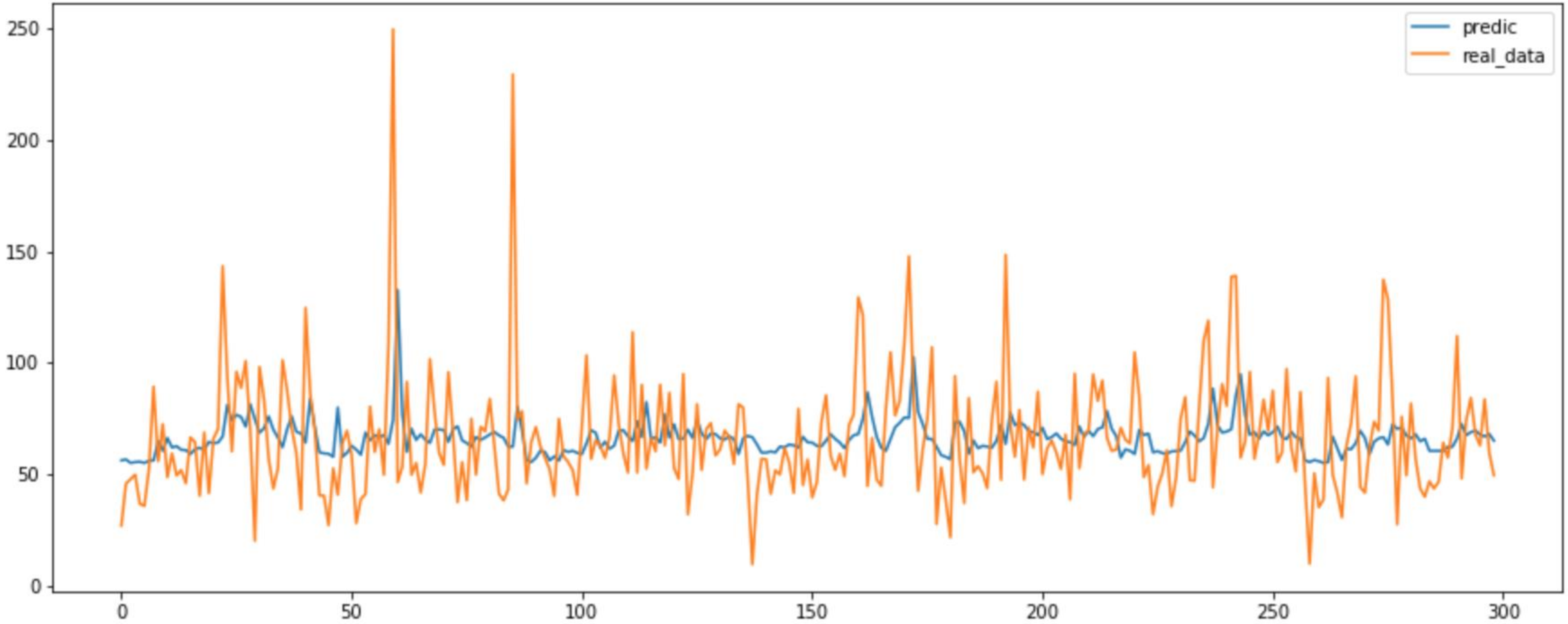
- A gradient boosting method to improve travel time prediction (2015)
- Let travel time data become 5-avg-set example: 6:00,6:05,6:10
- New Feature:
  - T1: before 5 min travel time
  - T2: before 10 min travel time
  - T3: before 15 min travel time
  - Detla1:T1-T2
  - Detla2:T2-T3
  - Time:1~288(5 min 1 set)



22	A	2	[2016-07-19 05:20:00,2016-07-19 05:25:00)	46.440000	46.730000	51.540000	-0.290000	-4.810000	40.63
23	A	2	[2016-07-19 05:25:00,2016-07-19 05:30:00)	40.630000	46.440000	46.730000	-5.810000	-0.290000	32.47
24	A	2	[2016-07-19 05:30:00,2016-07-19 05:35:00)	32.470000	40.630000	46.440000	-8.160000	-5.810000	55.35
25	A	2	[2016-07-19 05:35:00,2016-07-19 05:40:00)	55.350000	32.470000	40.630000	22.880000	-8.160000	63.86

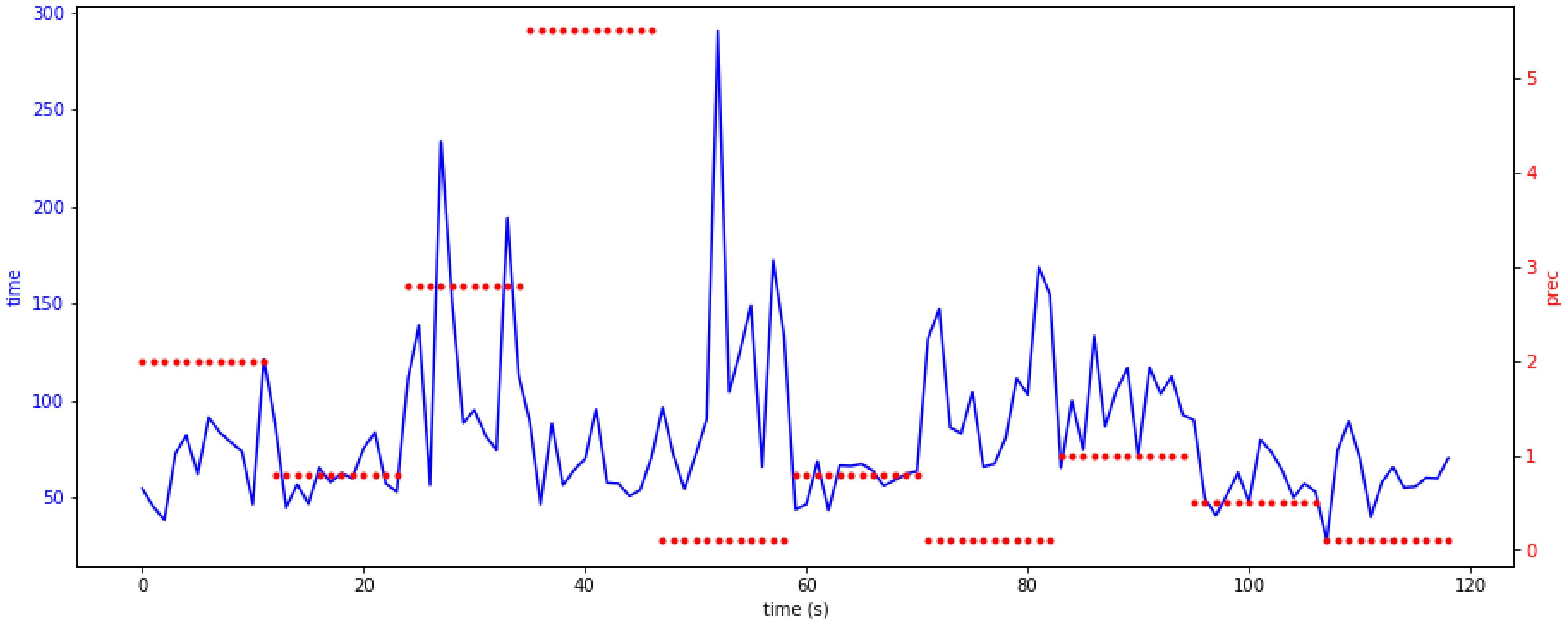
5-9-16: new model

- Use feature:4-18 features + T1,T2,T3,delta1,delta2
- Predicti data MAPE:0.1949 ⟶ 0.1881



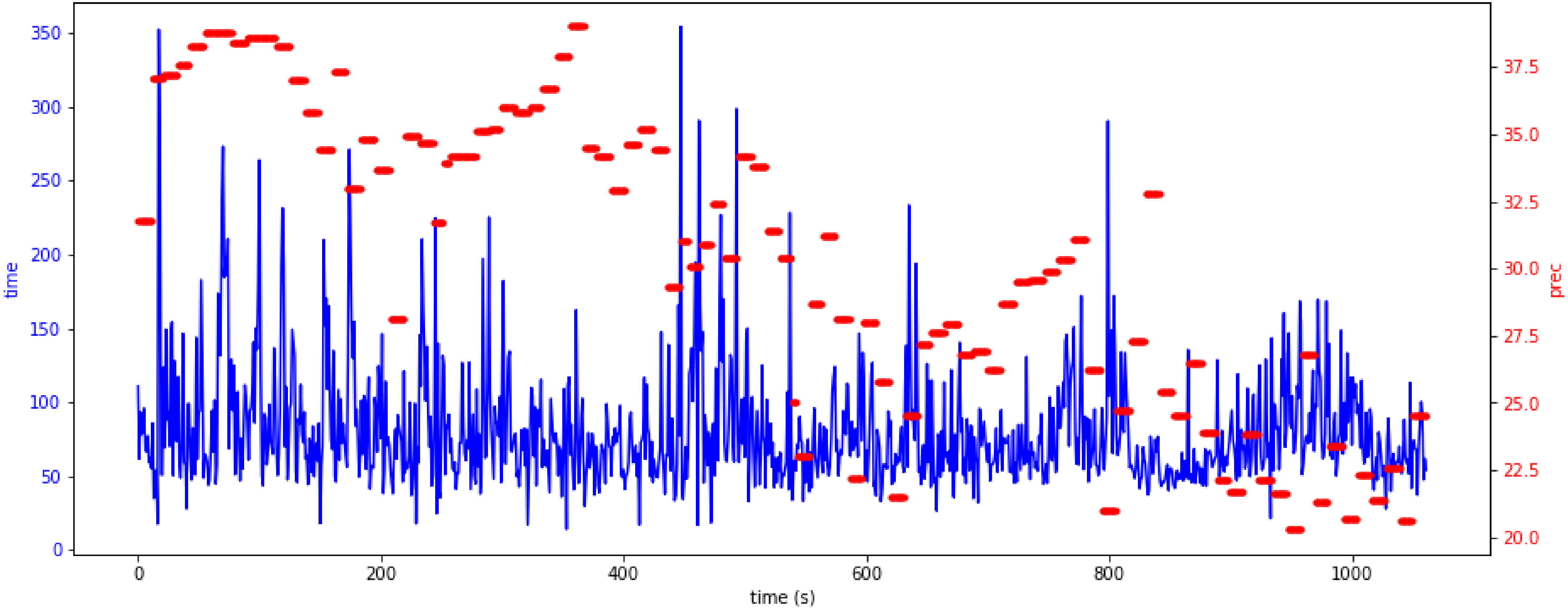
5-19: weather vs travel_time

- Precipitation vs Travel_time (A2 & hour=8)



5-19: weather vs travel_time

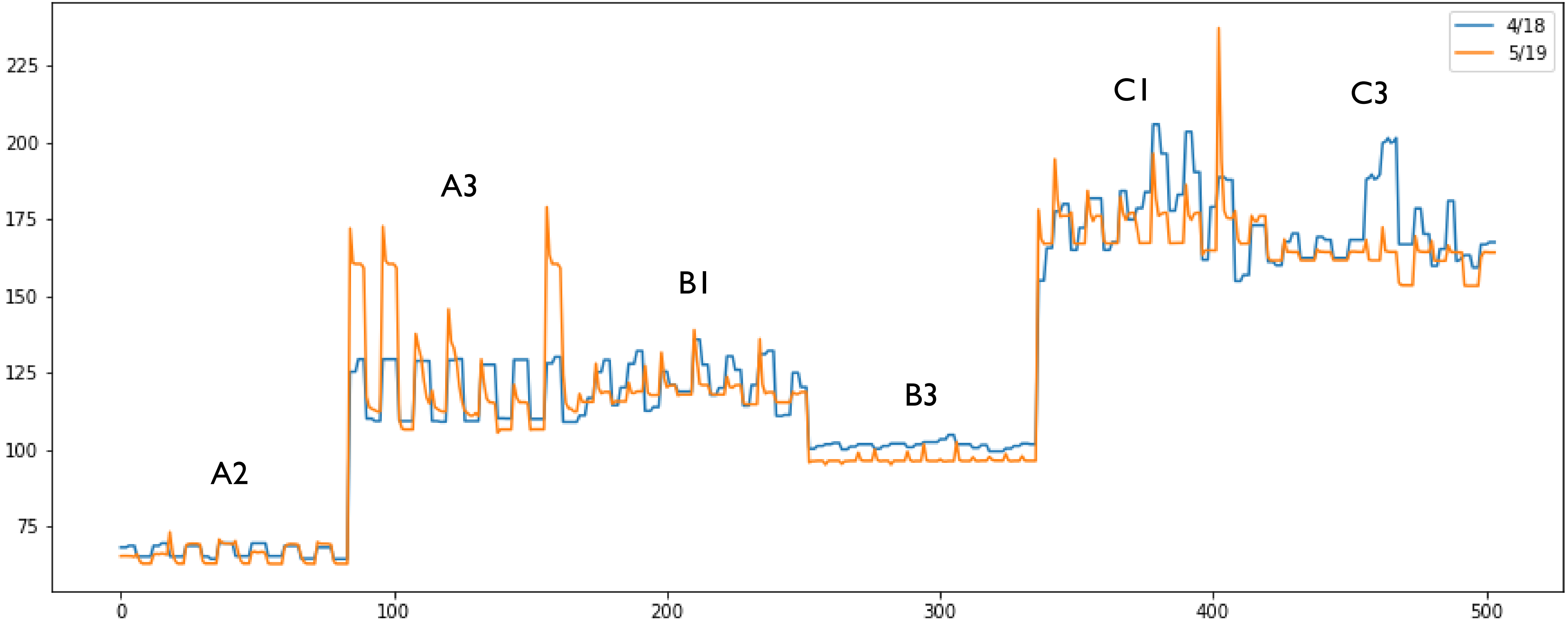
- Temperature vs Travel_time (A2 & hour=8)



5-19: use XGboost

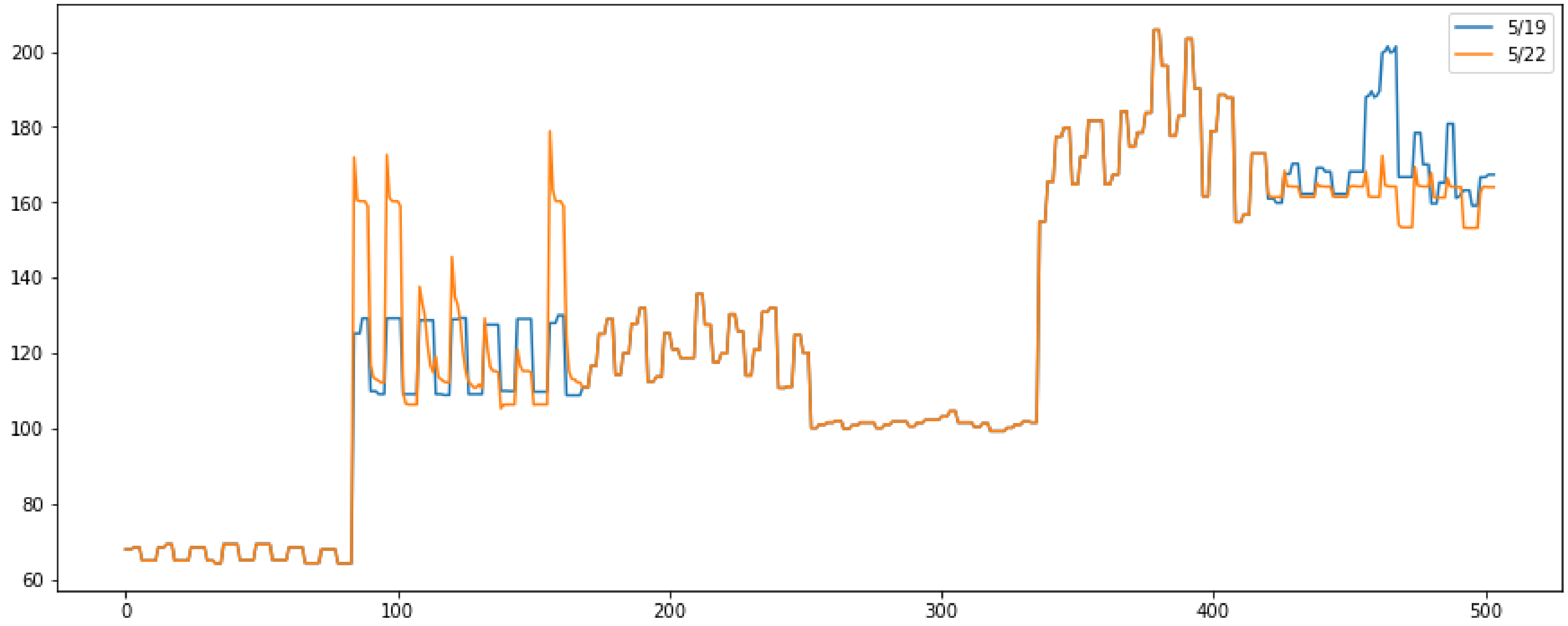
- Use feature: t1,t2,t3,deltat1,deltat2,weekday,hour
- Let tarvel_time > 500 become 500
- Testing data MAPE : 0.1681
  - Routes A2: 0.1704
  - Routes A3: 0.1337
  - Routes B1: 0.1636
  - Routes B3: 0.1874
  - Routes C1: 0.1419
  - Routes C3: 0.2117
- Predict data MAPE: 0.1851

Compare 4-18 and 5-19



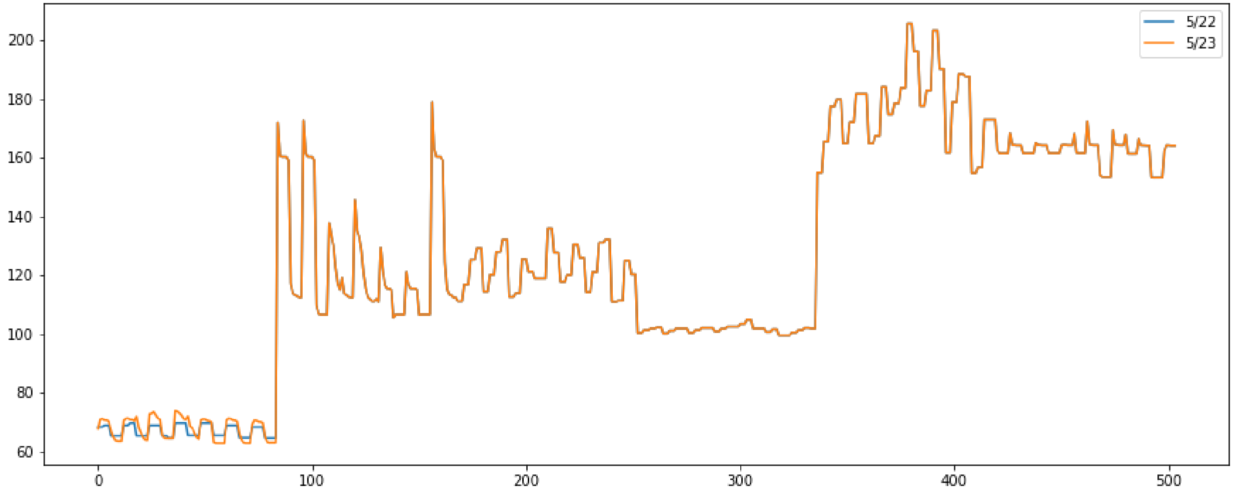
Compare 4-18 and 5-19

- 4-18 VS 5-19
- A2(4-18)
- A3(5-19)
- B1(4-18)
- B3(4-18)
- C1(4-18)
- C3(5-19)

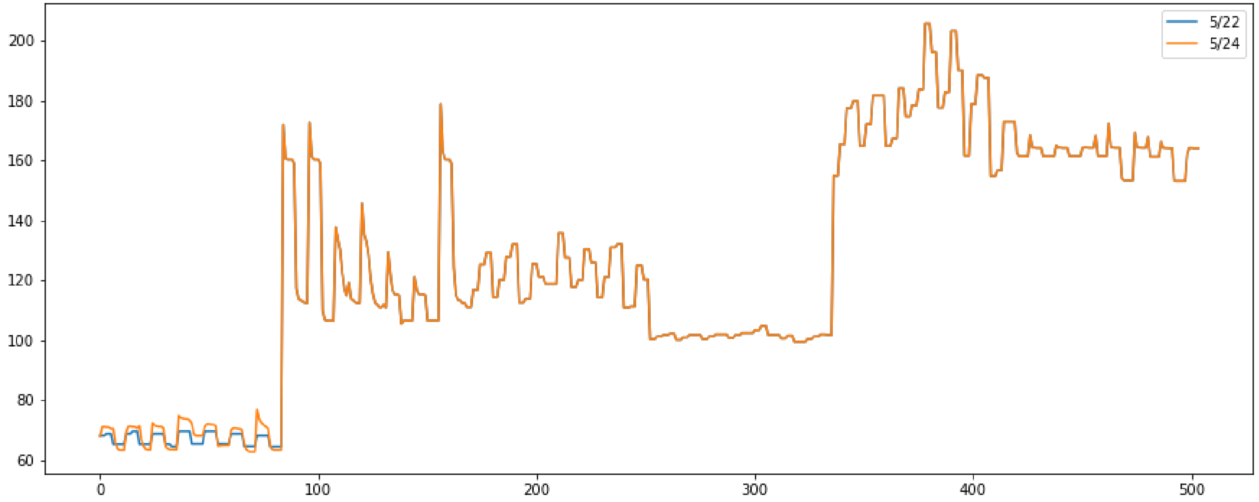


- Predict data MAPE: 0.1786

Turning A2 model 5-22~5-24



Predict data MAPE: 0.1782



Predict data MAPE: 0.1778

5-25 Phase1 over

- Best MAPE: 0.1778
- Rank : 100

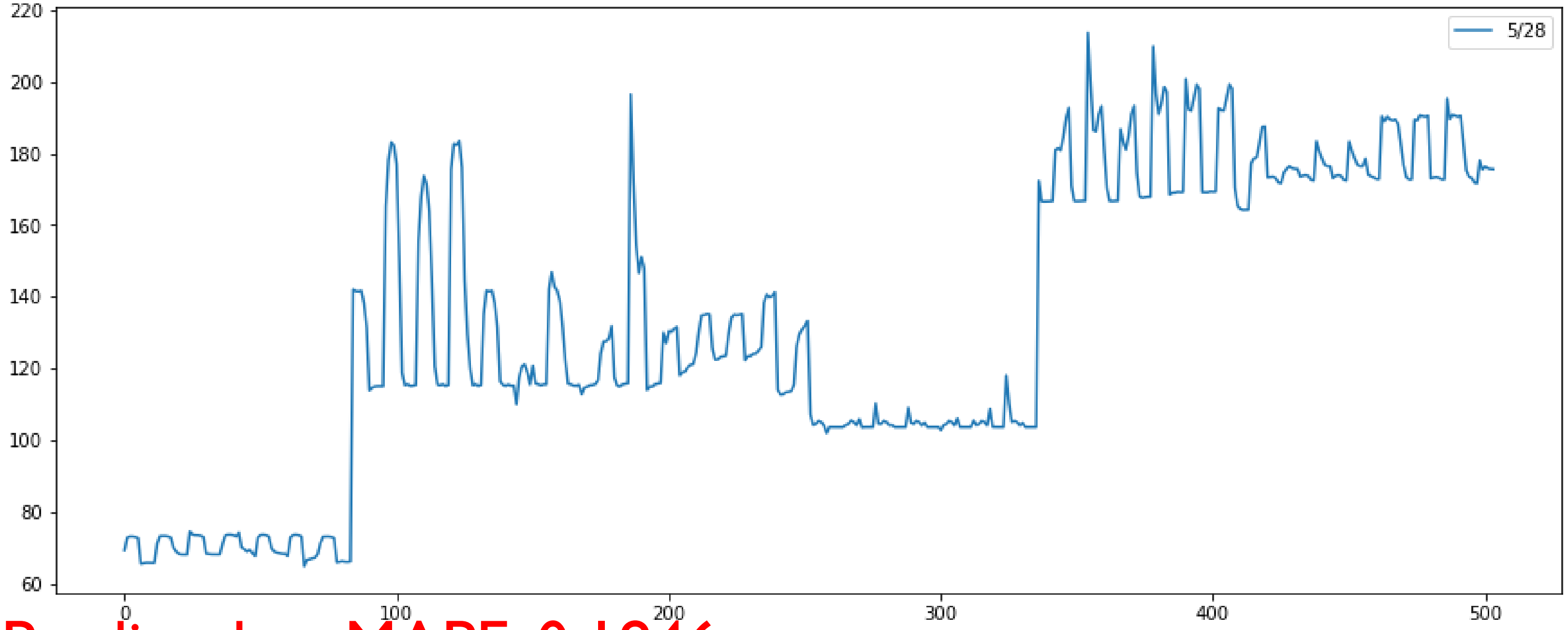
Phase2

Data processing

- Add 10-18 to 10-25 training data into the model
- Let travel time data become 5-avg-set example: 6:00,6:05,6:10
- New Feature:
  - T1: before 5 min travel time
  - T2: before 10 min travel time
  - T3: before 15 min travel time
  - Detla1:T1-T2
  - Detla2:T2-T3
  - Time:1~288(5 min 1 set)
- To predict 10-26 to 10-31 8~10AM and 17~19PM travel time

5-29: Phase2 predict

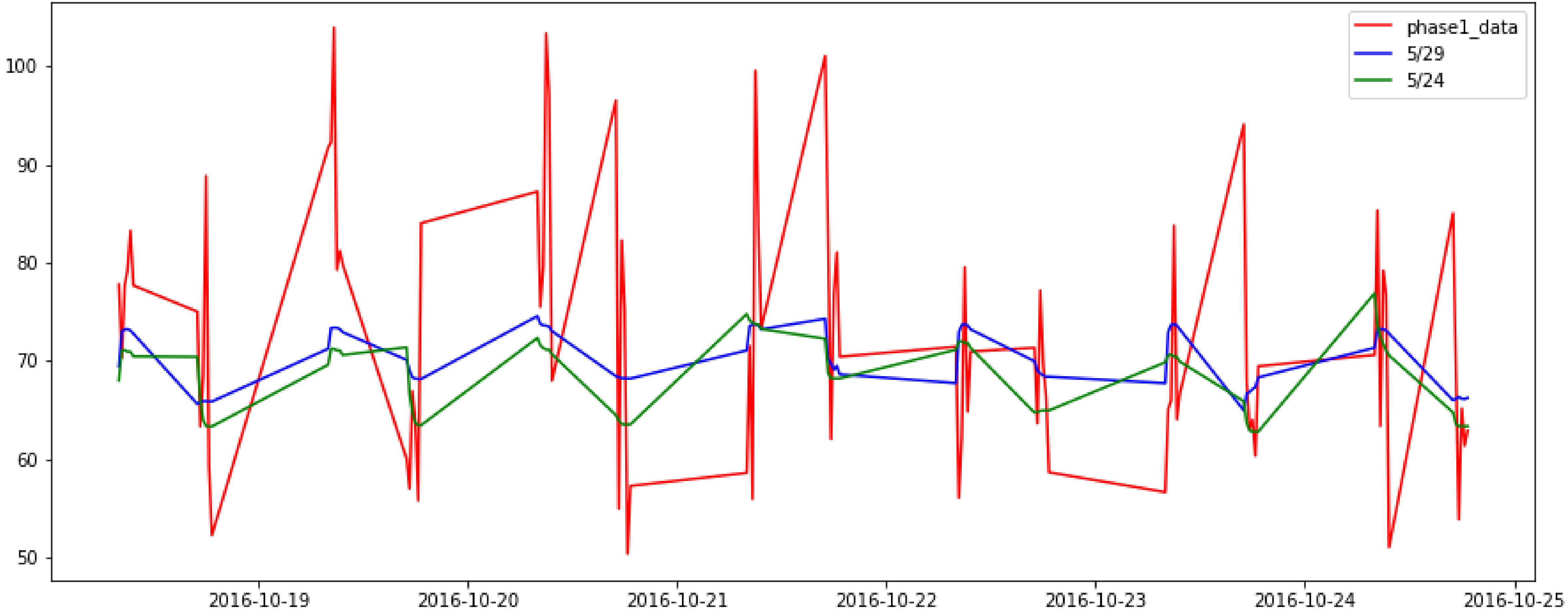
- Use feature: t1,t2,t3,deltat1,deltat2,weekday,hour,check
- Let tarvel_time > 500 become 500



- Predict data MAPE: 0.1846
- Rank:12

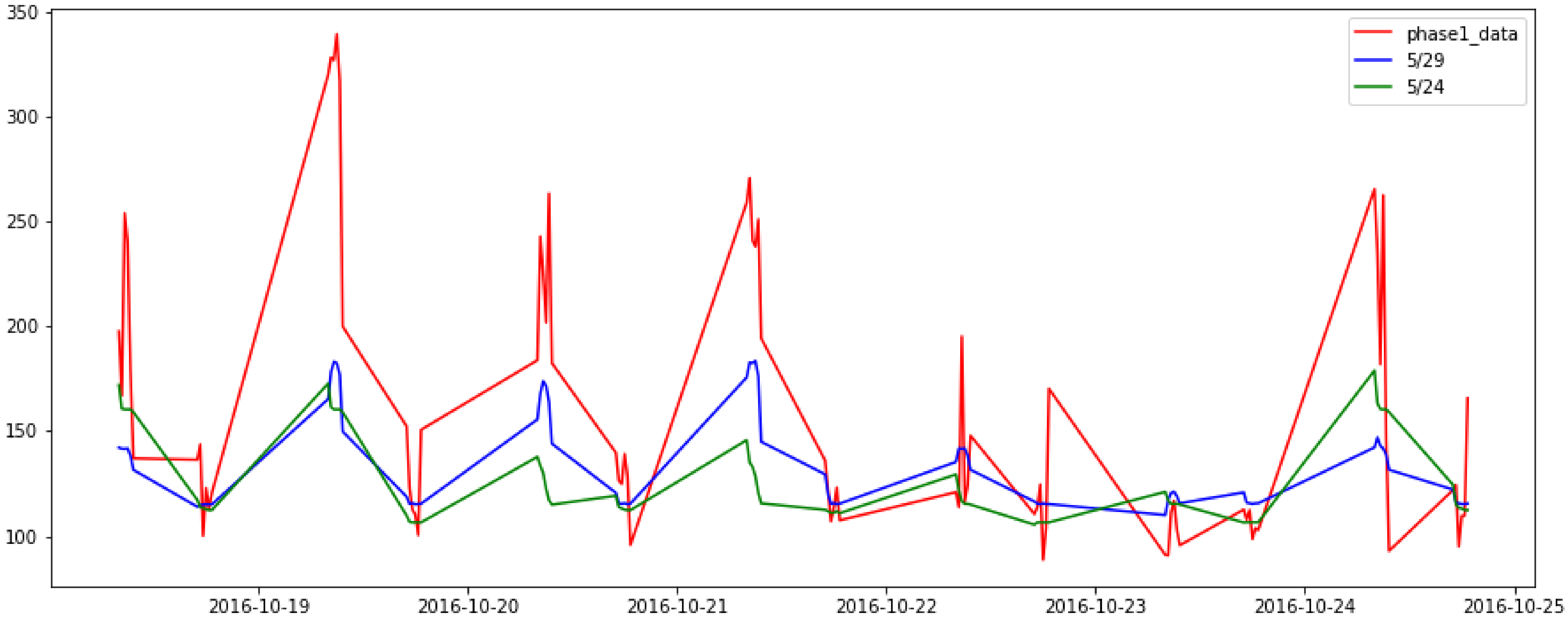
Compare 5/29 and Phase1 real data

- A2



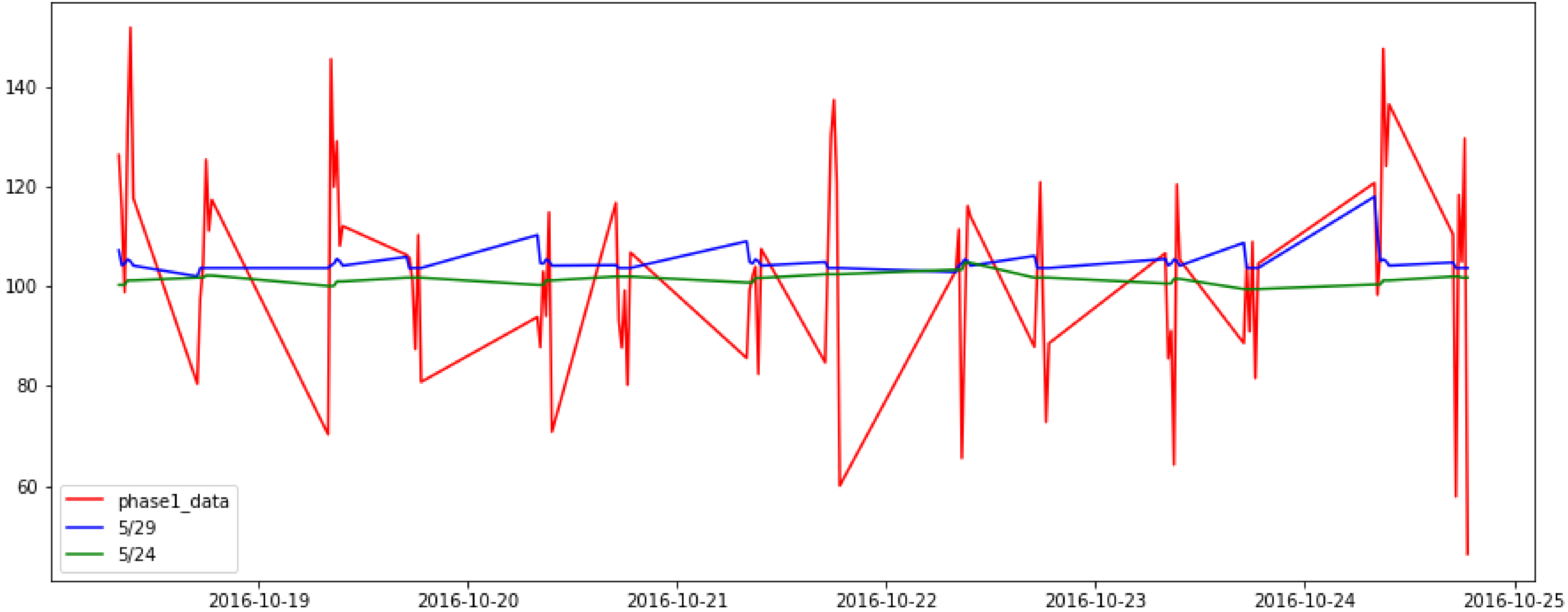
Compare 5/29 and Phase1 real data

- A3



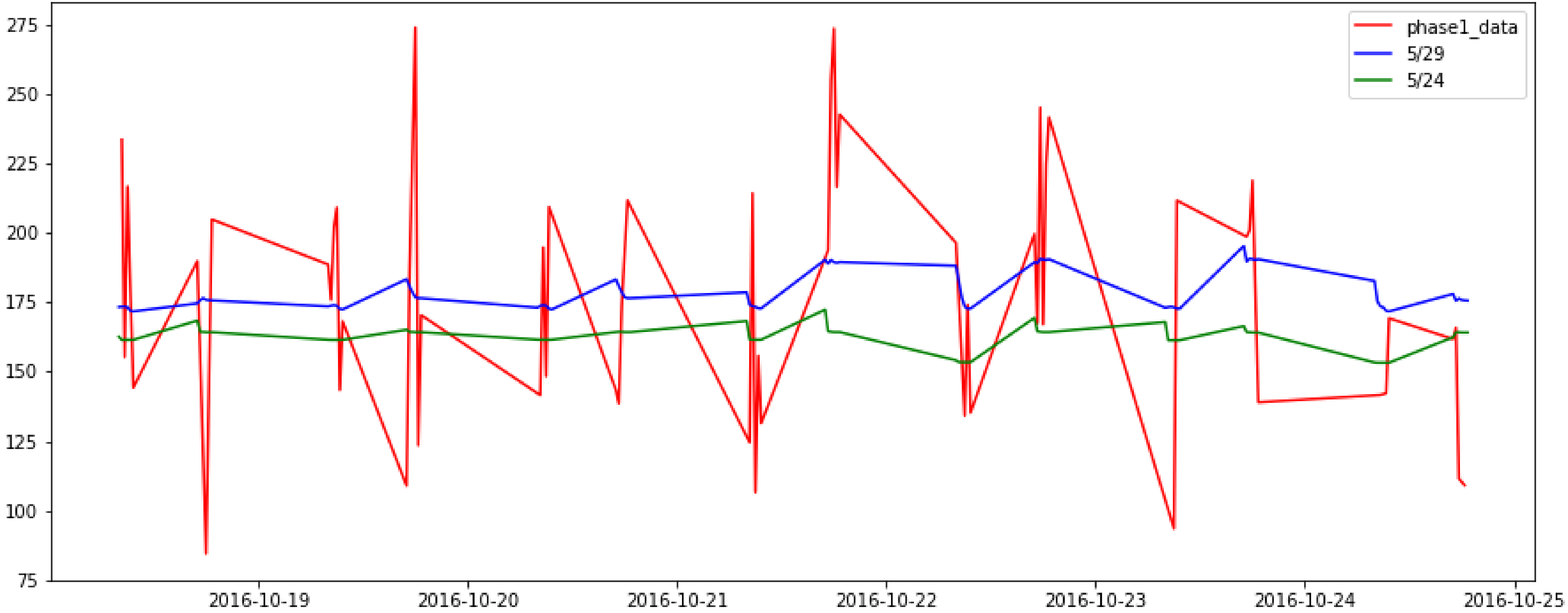
Compare 5/29 and Phase1 real data

- B3



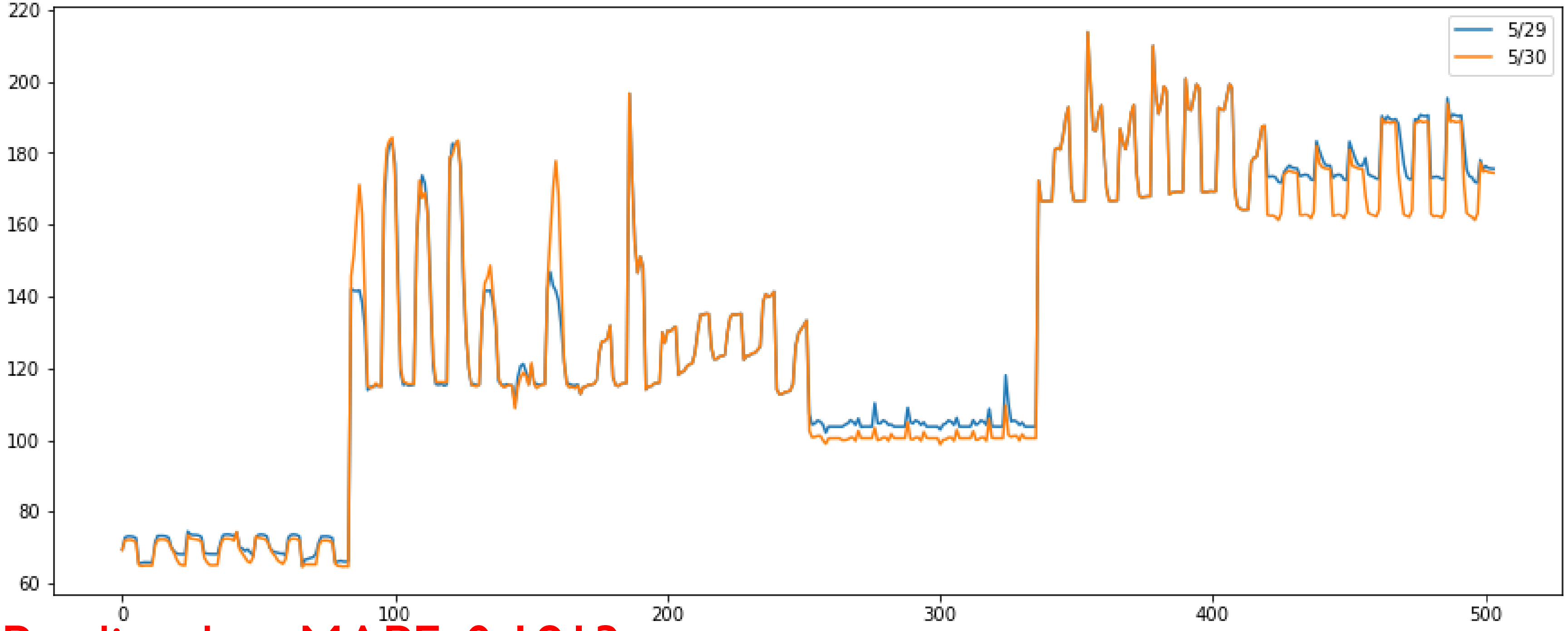
Compare 5/29 and Phase1 real data

- C3



5-30: Phase2 predict

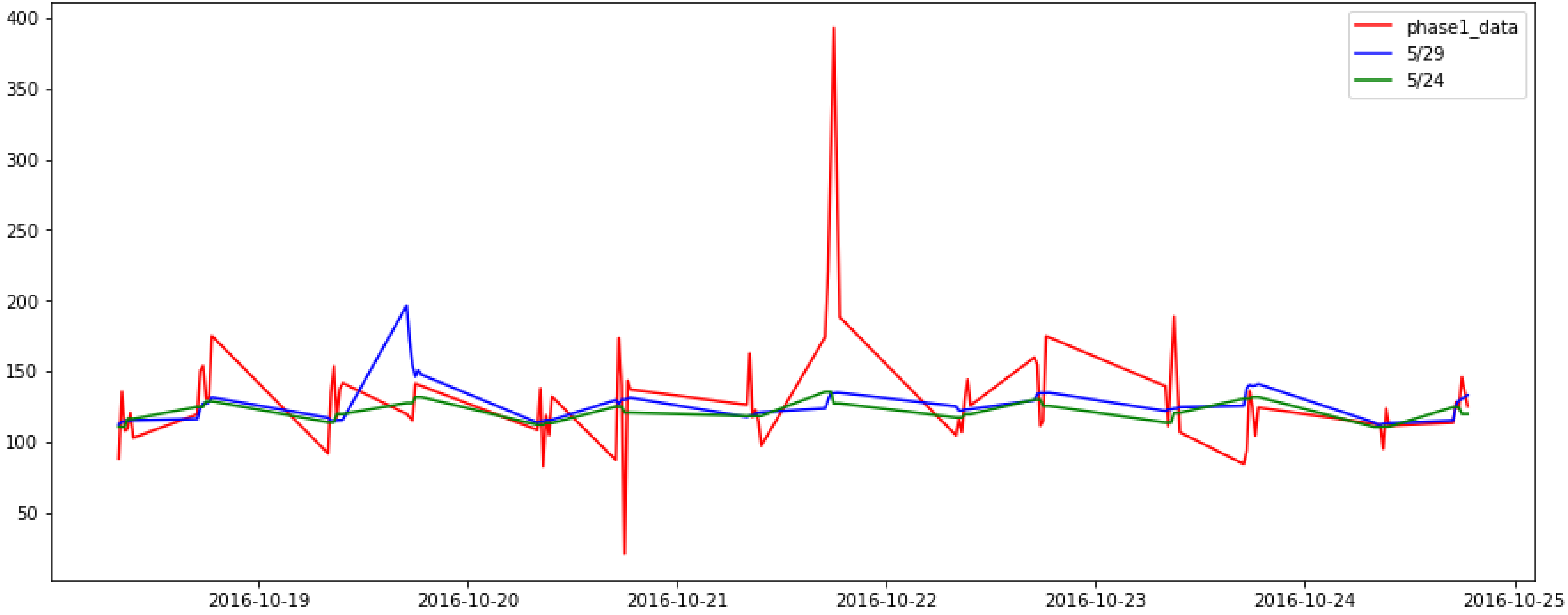
- Use feature: t1,t2,t3,deltat1,deltat2,weekday,hour,check
- Let tarvel_time > 500 become 500



- Predict data MAPE: 0.1813
- Rank:5

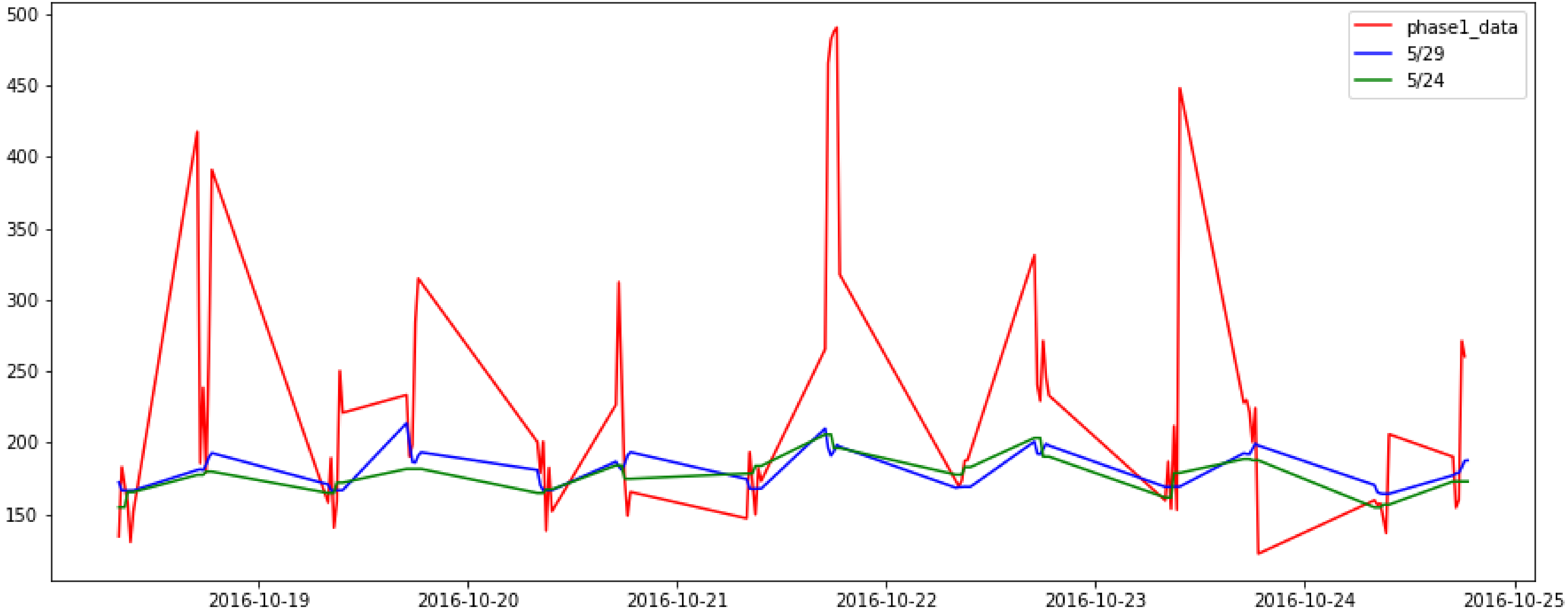
Compare 5/29 and Phase1 real data

- BI



Compare 5/29 and Phase1 real data

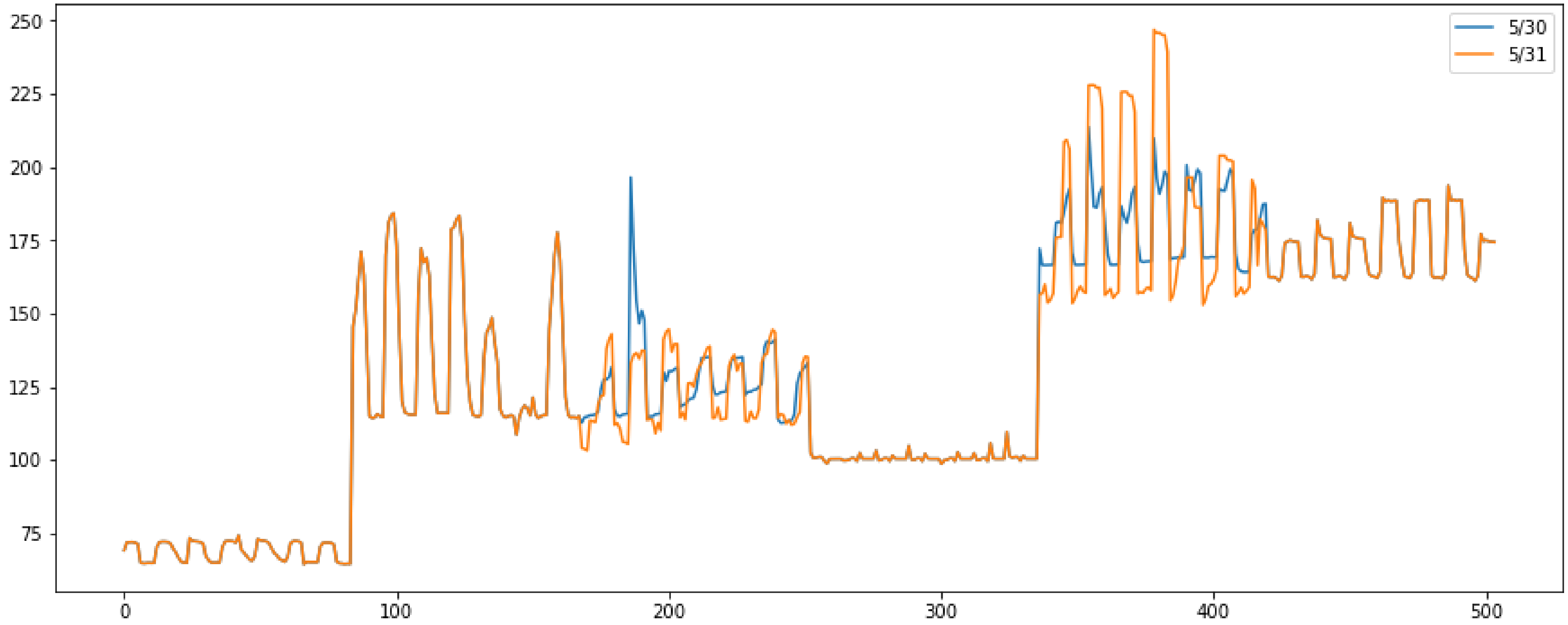
- CI



5-31: Phase2 predict

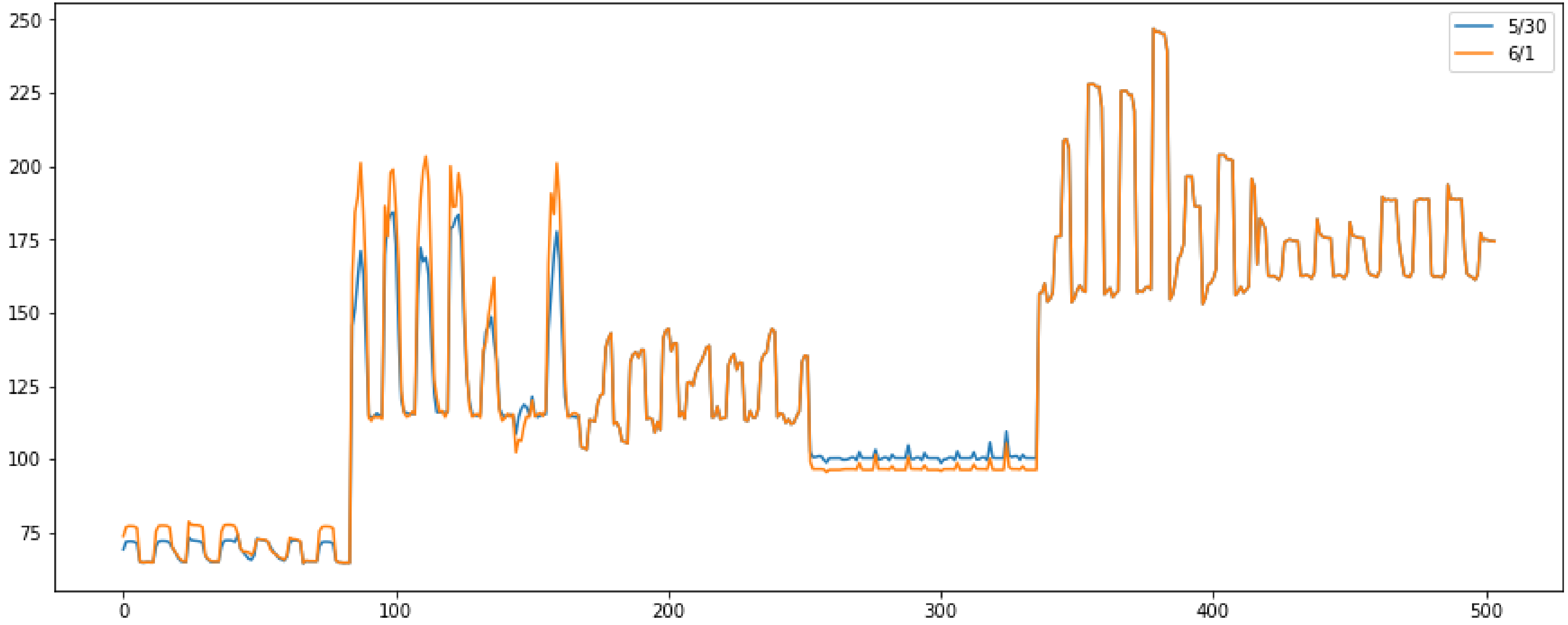
- C1 & B1(4-18 model) both better (5-19 model)
- Use feature: weekday,hours,minute,temperature,pressure,rel_humidity
- Let tarvel_time > 500 become 500

5-31: Phase2 predict



- Predict data MAPE: 0.1789
- Rank:5

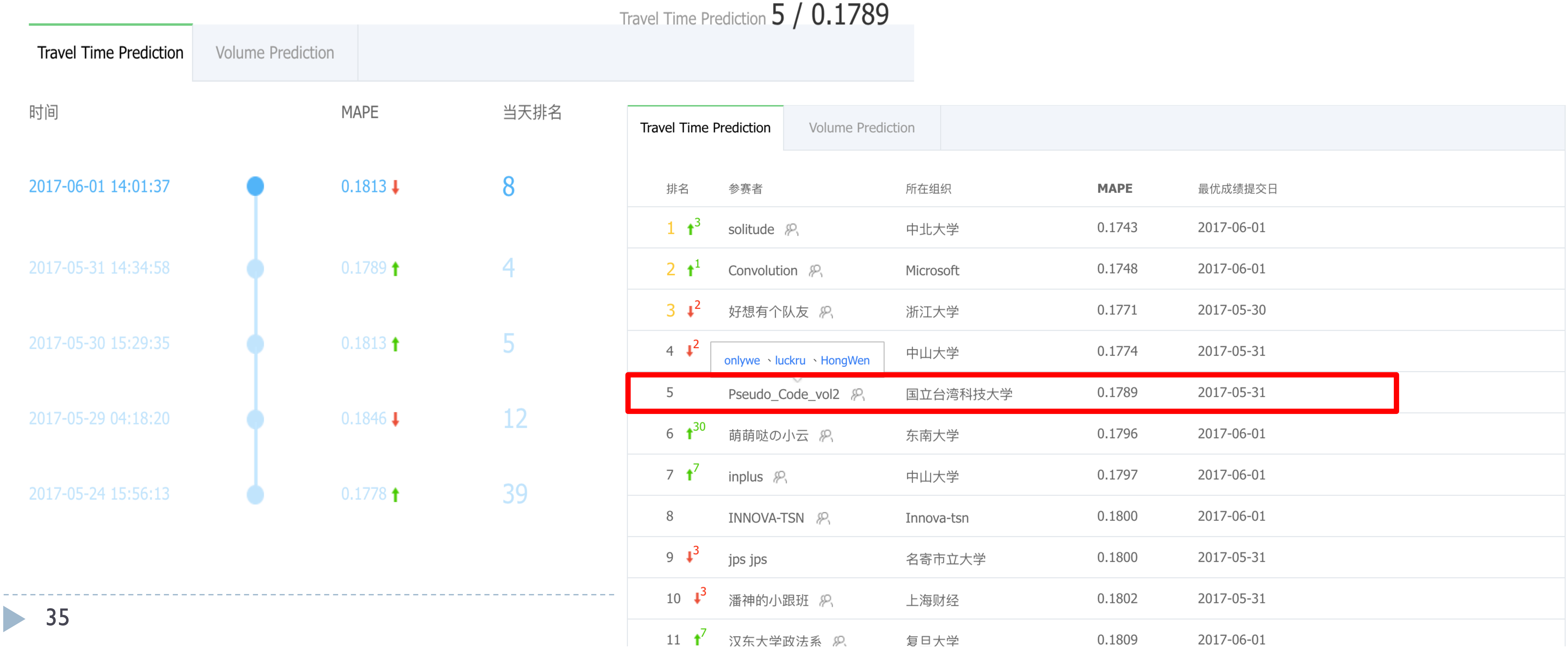
6-1: Phase2 predict



- Predict data MAPE: 0.1813
- Rank:5

5-25 Phase1 over

- Best MAPE: 0.1789
- Rank : 5



Travel Time Prediction 5 / 0.1789

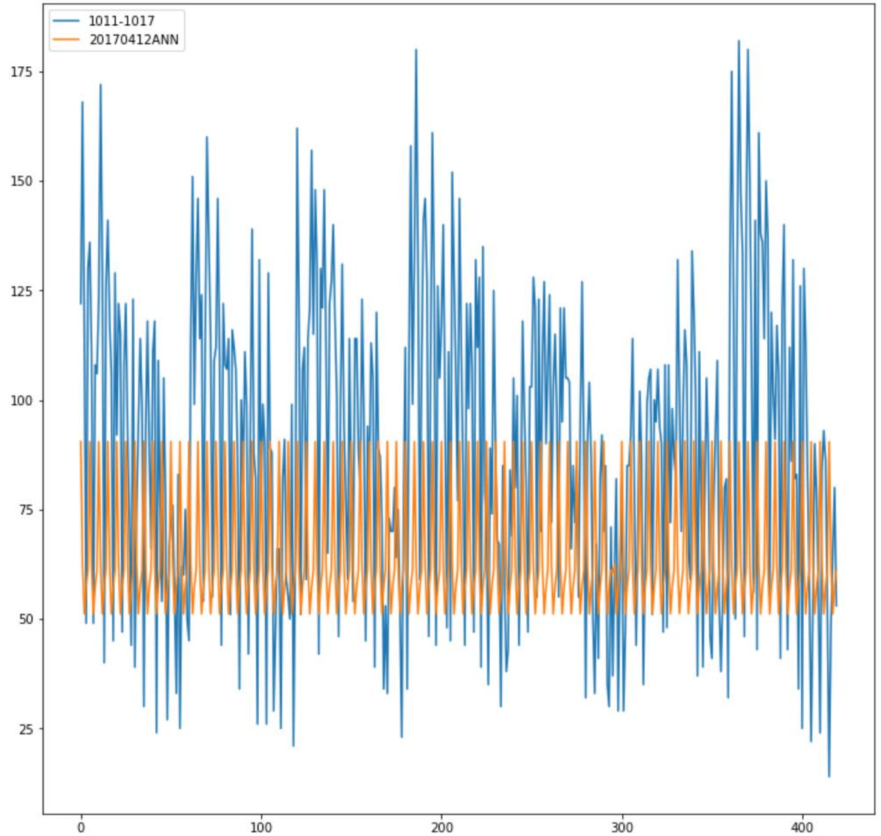
时间	MAPE	当天排名
2017-06-01 14:01:37	0.1813 ↓	8
2017-05-31 14:34:58	0.1789 ↑	4
2017-05-30 15:29:35	0.1813 ↑	5
2017-05-29 04:18:20	0.1846 ↓	12
2017-05-24 15:56:13	0.1778 ↑	39

排名		参赛者	所在组织	MAPE	最优成绩提交日
1	↑3	solitude	中北大学	0.1743	2017-06-01
2	↑1	Convolution	Microsoft	0.1748	2017-06-01
3	↓2	好想有个队友	浙江大学	0.1771	2017-05-30
4	↓2	onlywe 、luckru 、HongWen	中山大学	0.1774	2017-05-31
5		Pseudo_Code_vol2	国立台湾科技大学	0.1789	2017-05-31
6	↑30	萌萌哒の小云	东南大学	0.1796	2017-06-01
7	↑7	inplus	中山大学	0.1797	2017-06-01
8		INNOVA-TSN	Innova-tsn	0.1800	2017-06-01
9	↓3	jps jps	名寄市立大学	0.1800	2017-05-31
10	↓3	潘神的小跟班	上海财经	0.1802	2017-05-31
11	↑7	汉东大学政法系	复旦大学	0.1809	2017-06-01

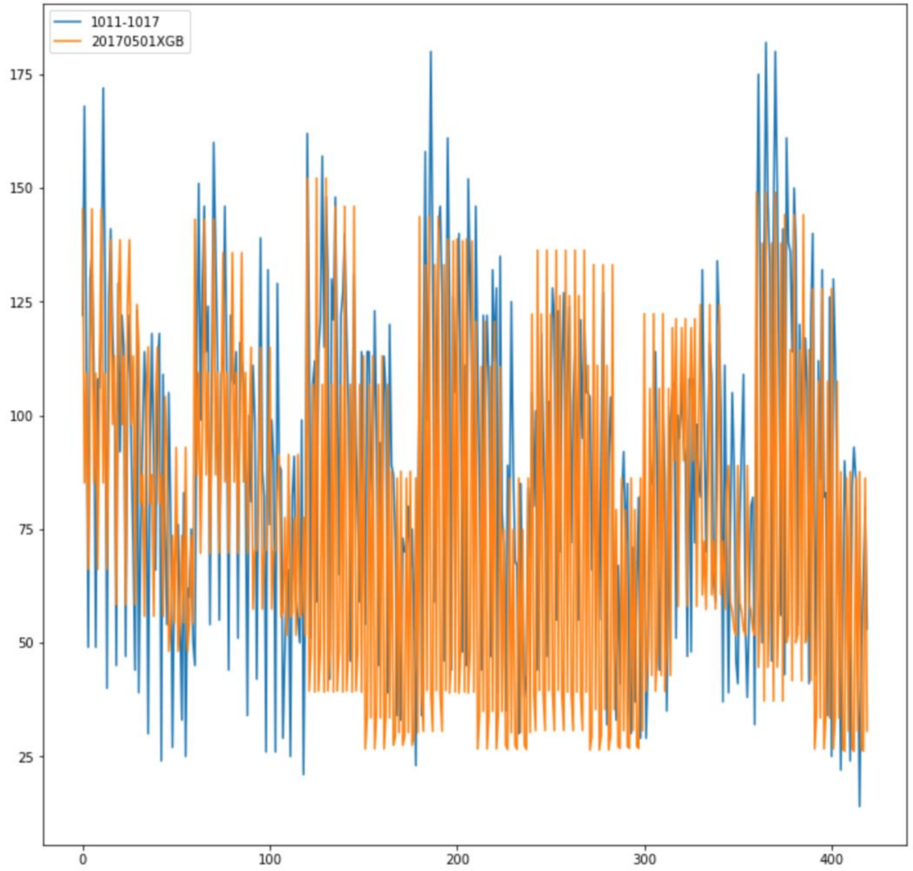
Task 2 (volume)

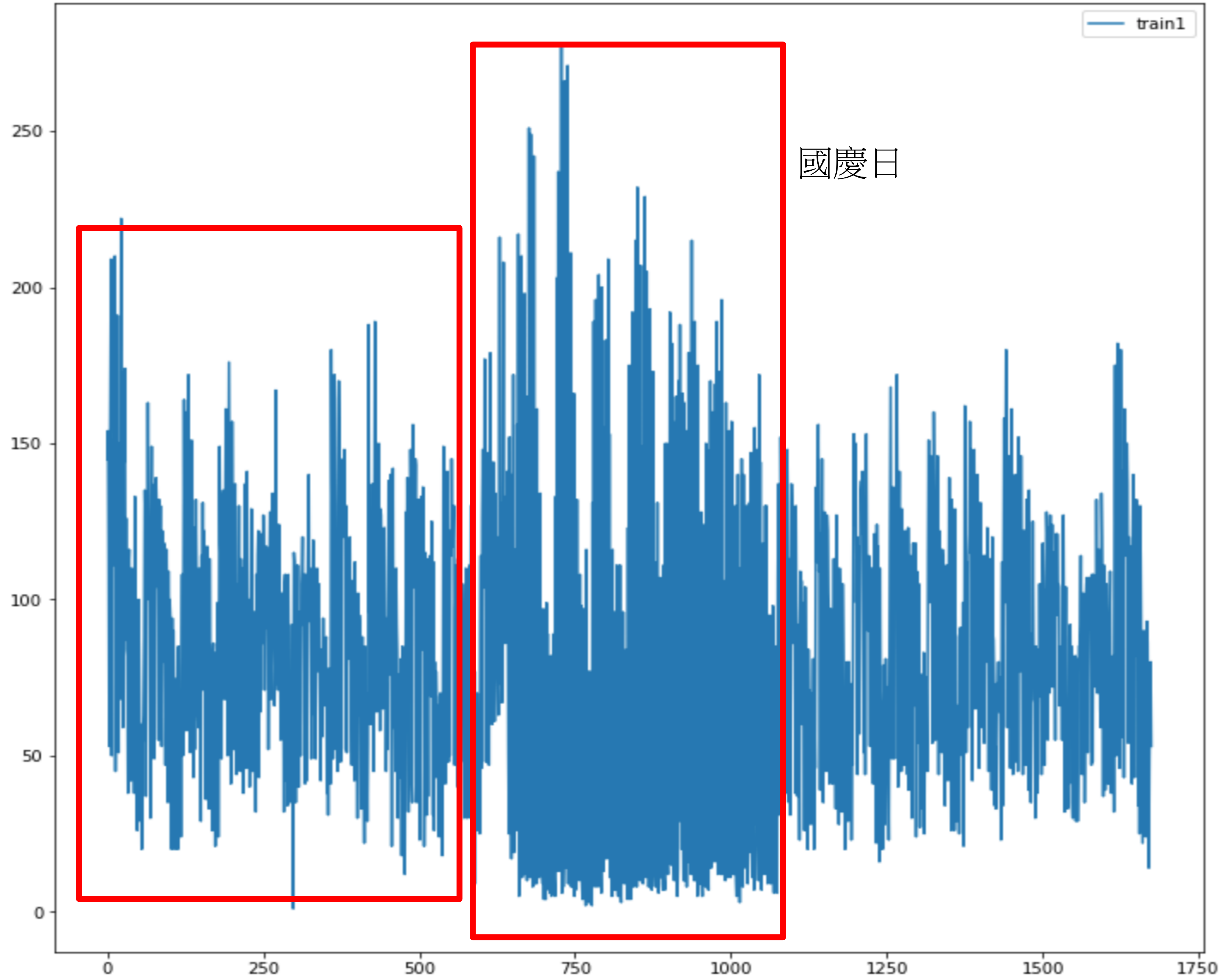
task2 - volume prediction

- phase1 : predict 10/18– 10/24 average tollgate traffic volume.
- at the beginning, I use ANN to train 5 model, and then combine them
  - ANN : 2 hidden layer (10-5)
  - data processing : min-max normalization
  - **features : month, hour, minute, weekday, temperature, rel-humidity**
  - training data : 2016-09-20 to 2016-10-17, every 2 minutes
  - testing data : 2016-10-11 to 2016-10-17
- test MAPE = 0.3848
- real MAPE = 0.4110



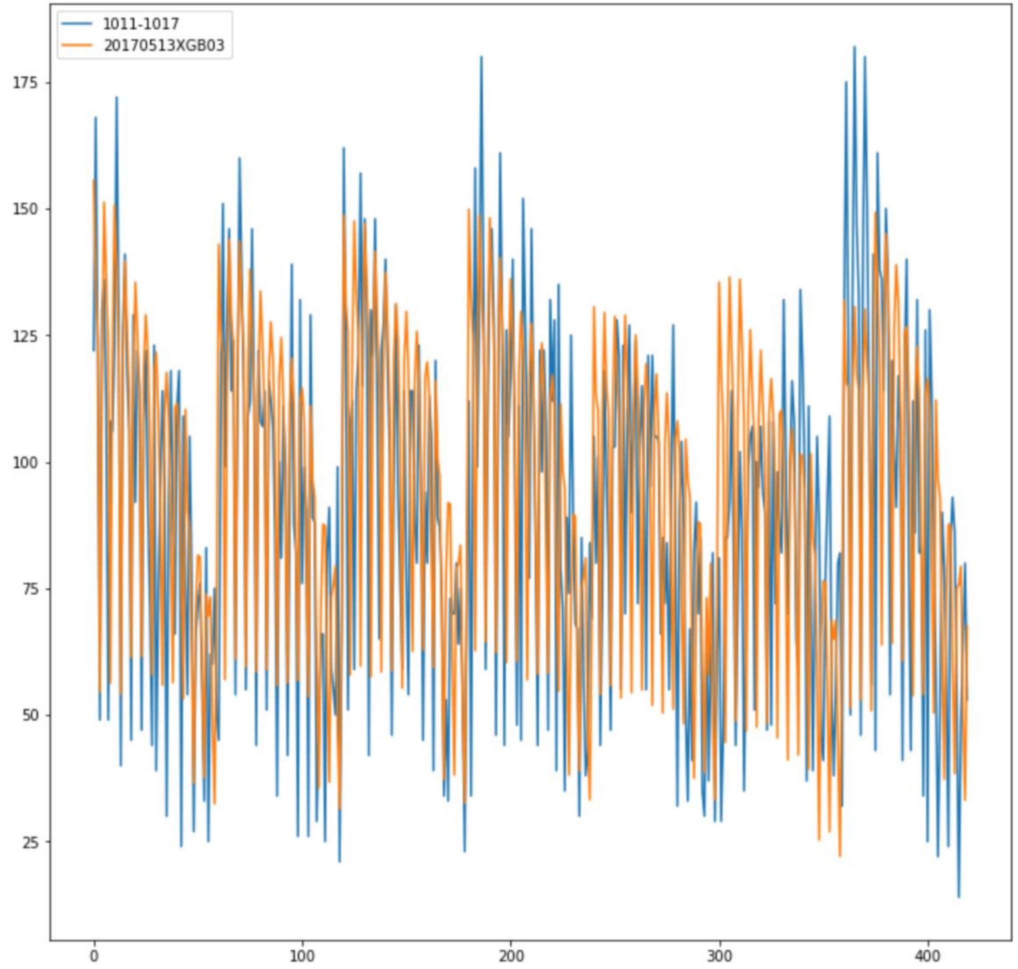
- phase1 : predict 10/18– 10/24 average tollgate traffic volume.
- using xgboost to train 5 model, and then combine them
  - using mse scoring to tune best parameters
  - data processing : min-max normalization
  - **features : month, hour, minute, weekday,**
    temperature, rel-humidity
  - training data : 2016-09-20 to 2016-10-17, every 2 minutes
  - testing data : 2016-10-11 to 2016-10-17
- test MAPE = 0.5721
- real MAPE = 0.4013

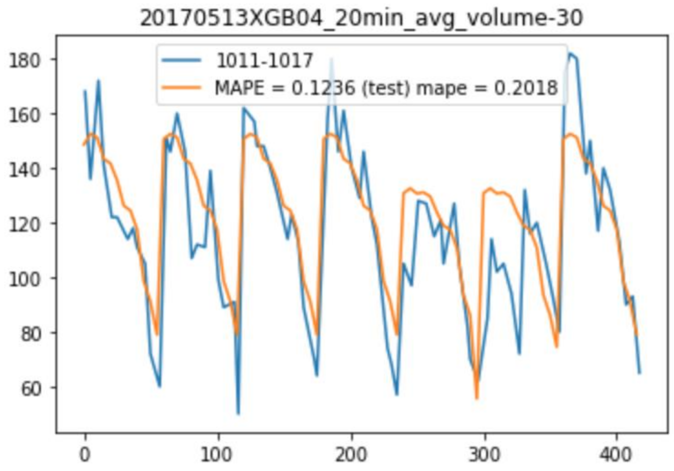




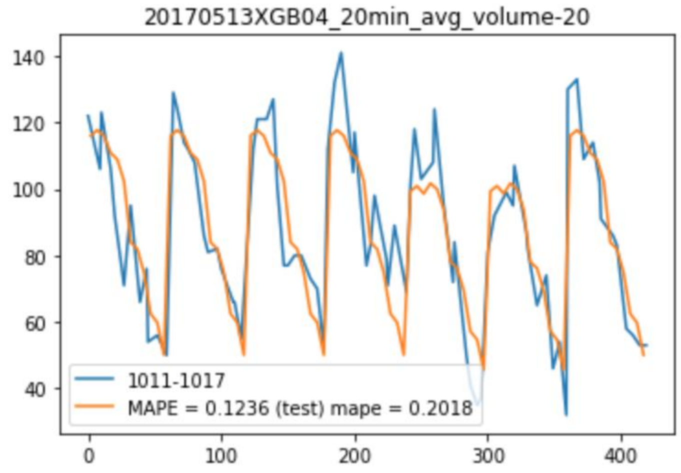
國慶日

- phase1 : predict 10/18– 10/24 average tollgate traffic volume.
- using xgboost to train 1 model
  - using mse scoring to tune best parameters
  - data processing : standardization
  - **features : tollgate_id, direction, hour, minute, holiday, 5min_ago, 10min_ago, weekday**
  - training data : 2016-10-08 to 2016-10-17, every 5 minutes
  - testing data : 2016-10-11 to 2016-10-17
- test MAPE = 0.1236
- real MAPE = 0.2018

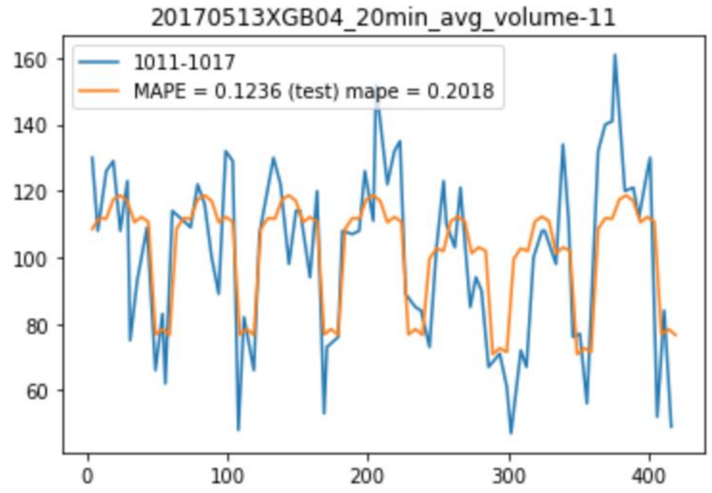




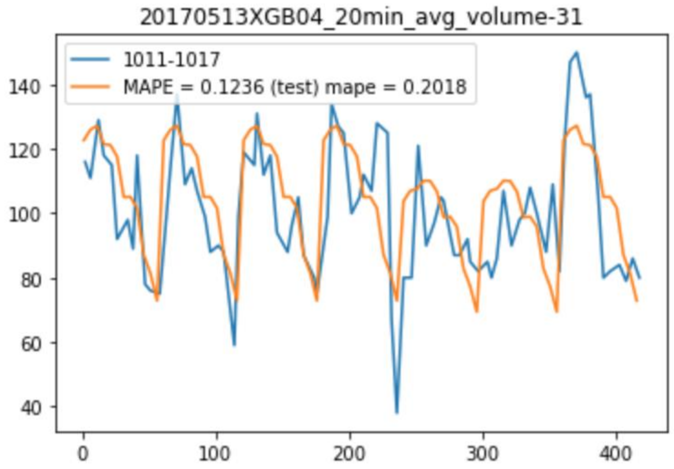
0.04815702009760177



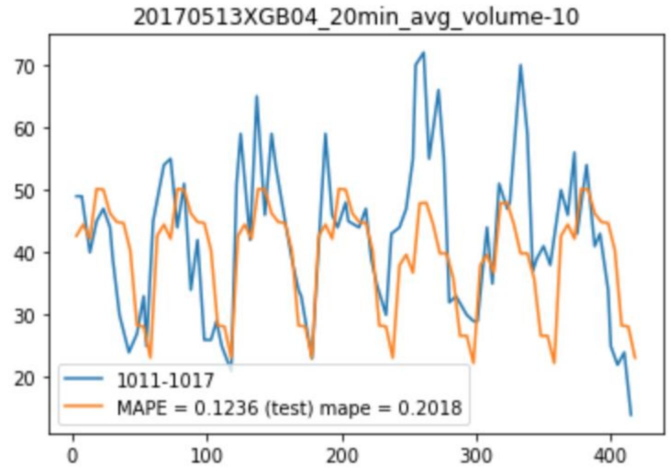
0.03637701789013063



0.06334180251930366

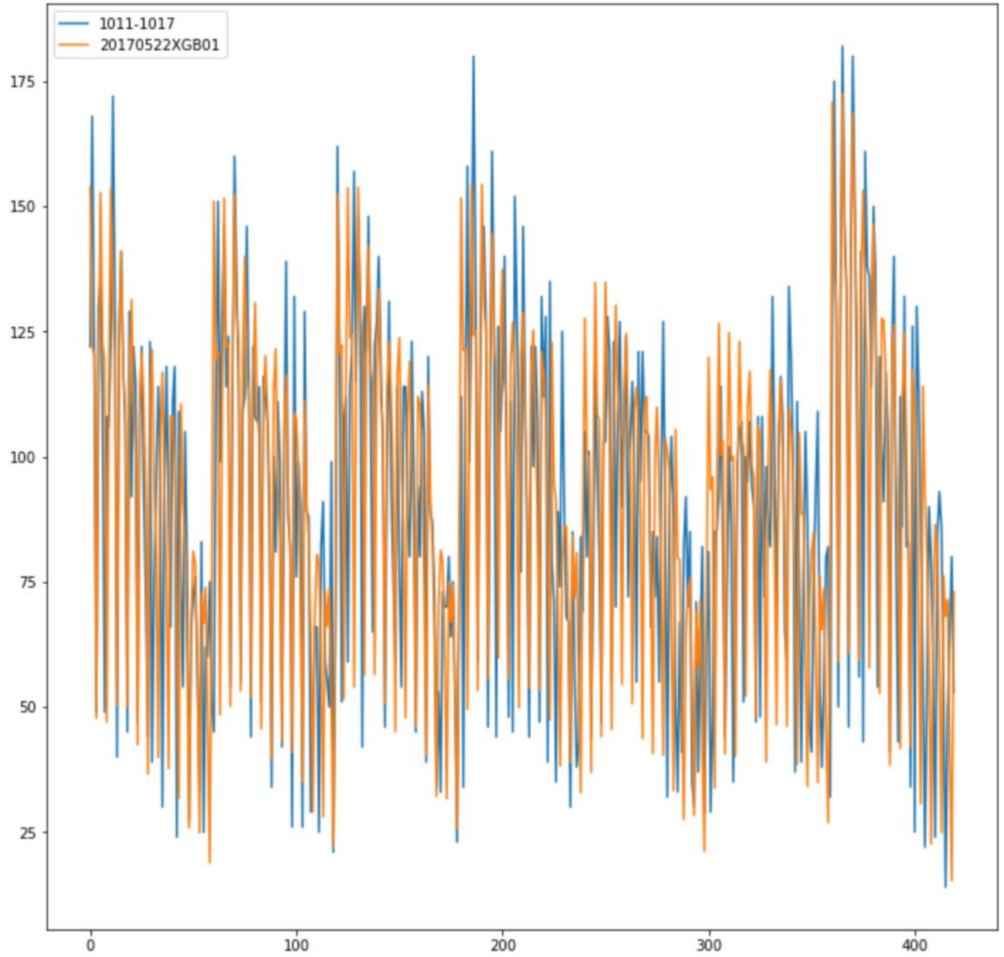


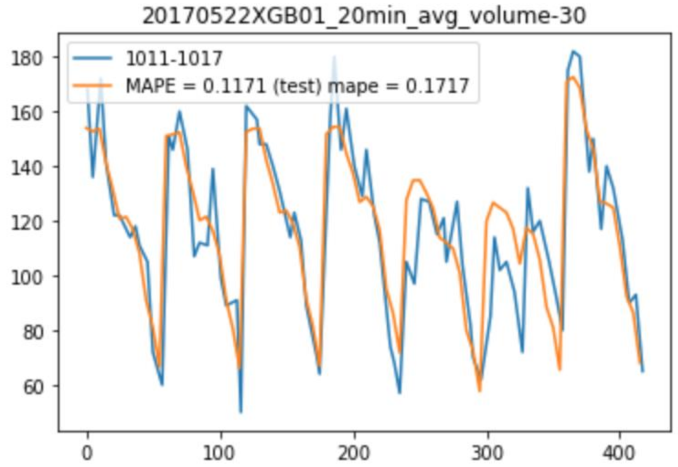
0.055360133671968514



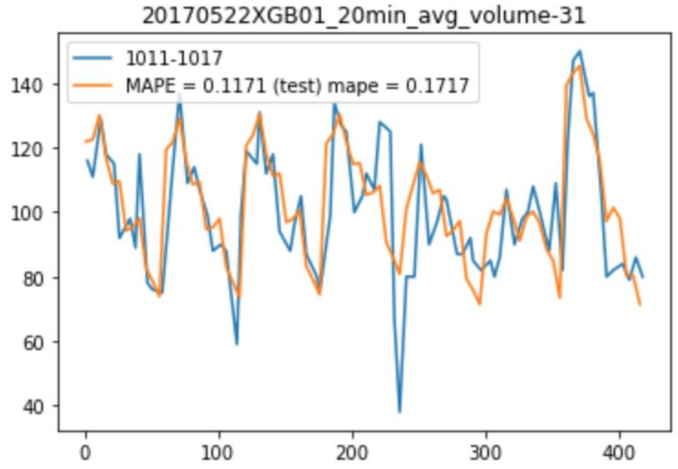
0.07995655244071778

- phase1 : predict 10/18– 10/24 average tollgate traffic volume.
- using xgboost to train 1 model
  - using mse scoring to tune best parameters
  - data processing : standardization
  - features : **tollgate_id, direction, hour, minute, weekday**
  - training data : 2016-10-08 to 2016-10-17, every 20 minutes
  - testing data : 2016-10-11 to 2016-10-17
- test MAPE = 0.1171
- real MAPE = 0.1717

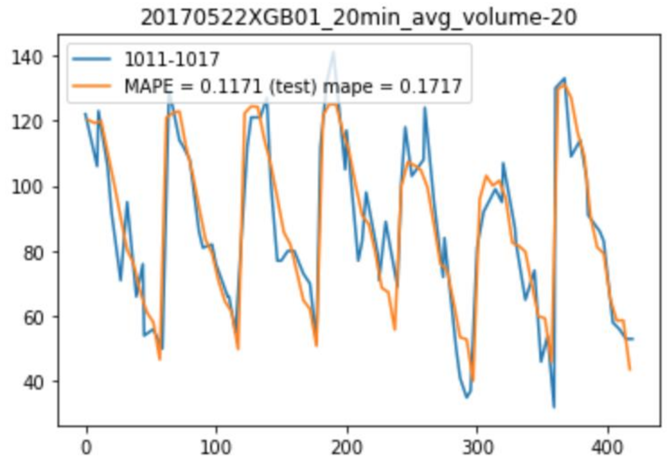



20170522XGB01_20min_avg_volume-30

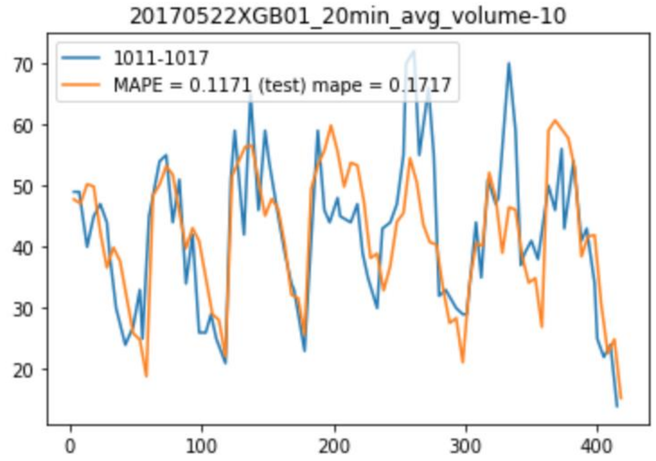
0.030878023808432957


20170522XGB01_20min_avg_volume-31

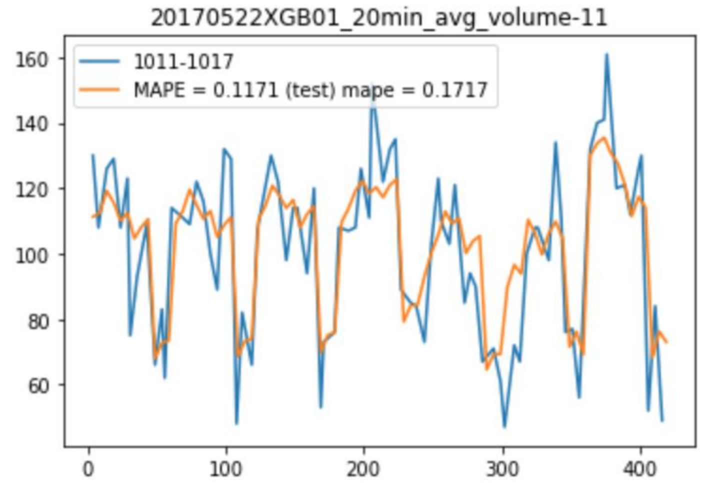
0.046439807806175355


20170522XGB01_20min_avg_volume-20

0.03265662101969155

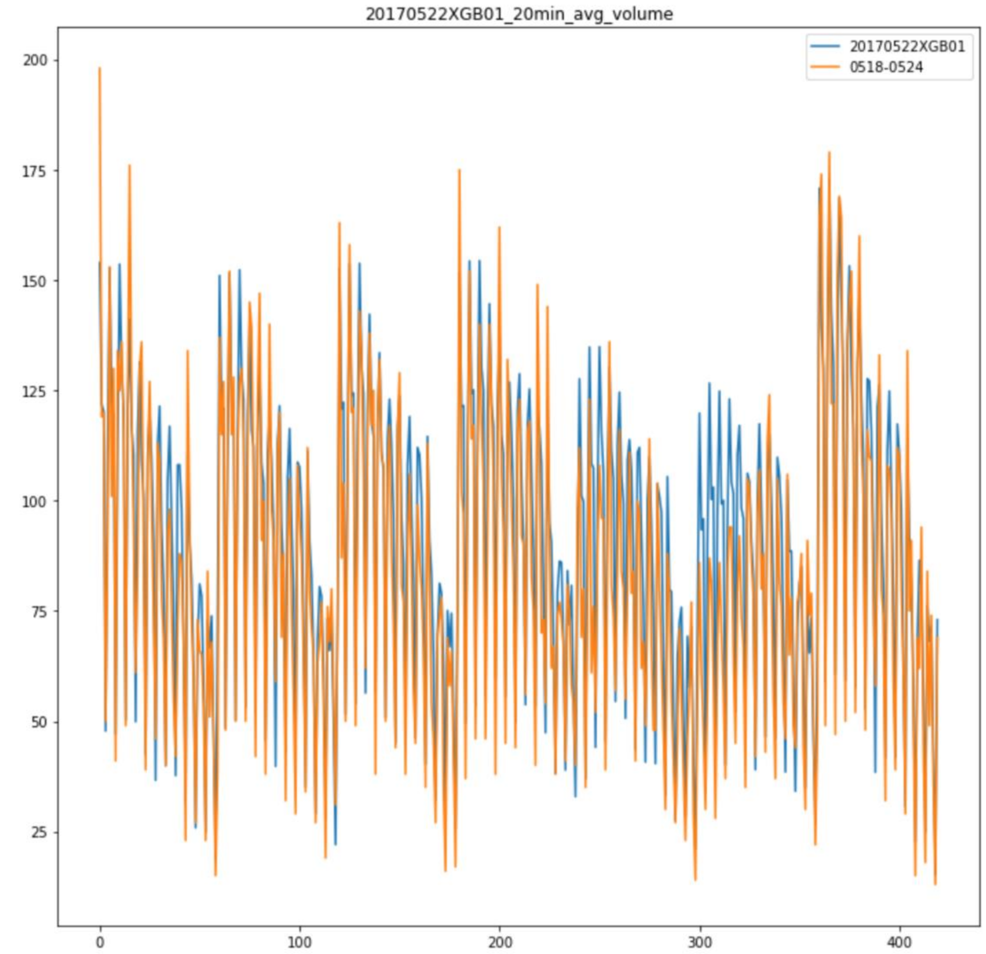

20170522XGB01_20min_avg_volume-10

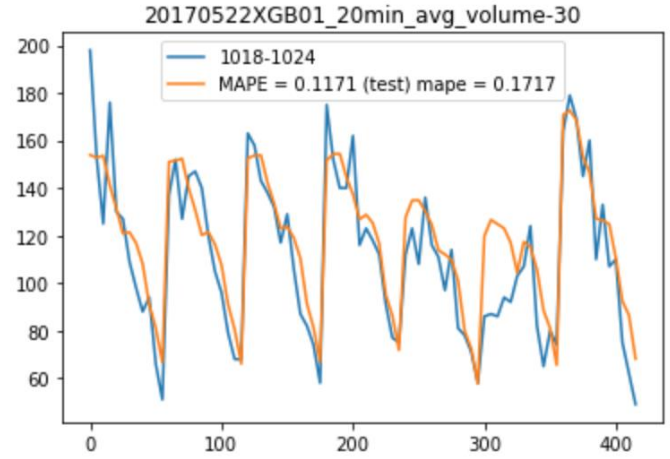
0.07636071839711914


20170522XGB01_20min_avg_volume-11

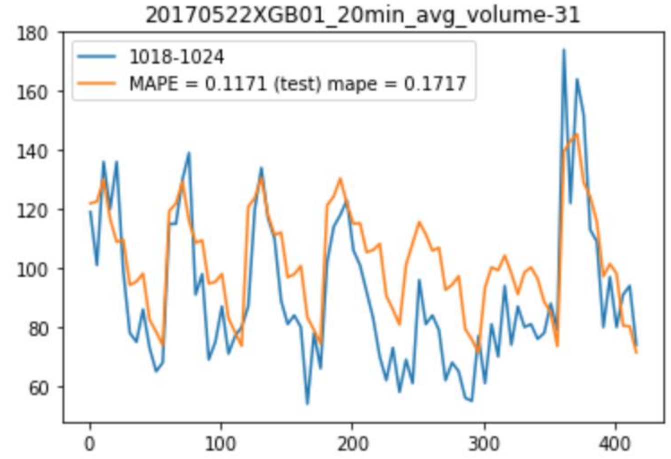
0.047940262555877515

- phase2 : predict 10/25– 10/31 average tollgate traffic volume.
- weekend, low-value

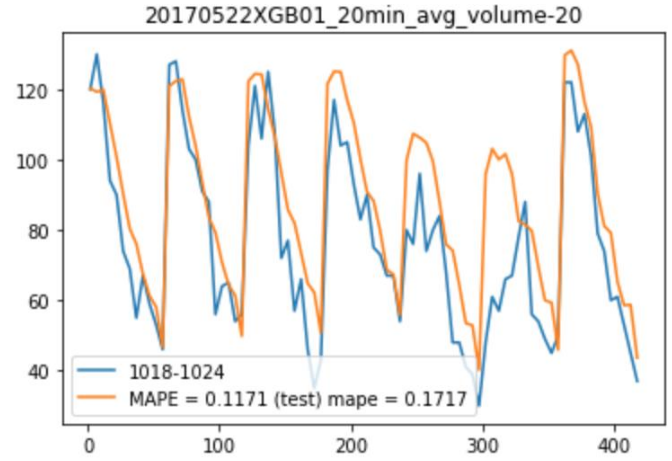




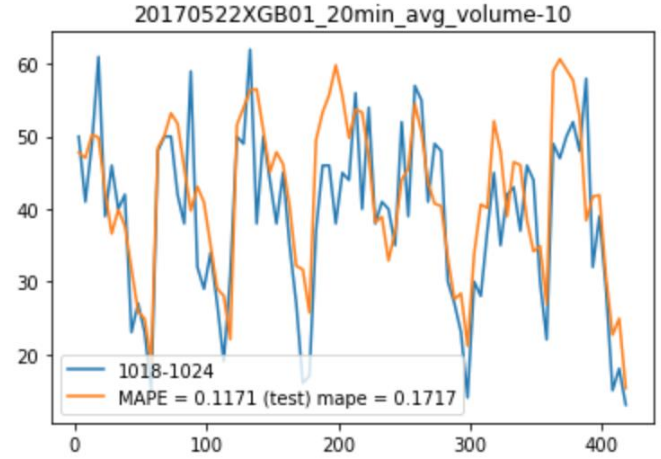
```
0.1253097041218753
```



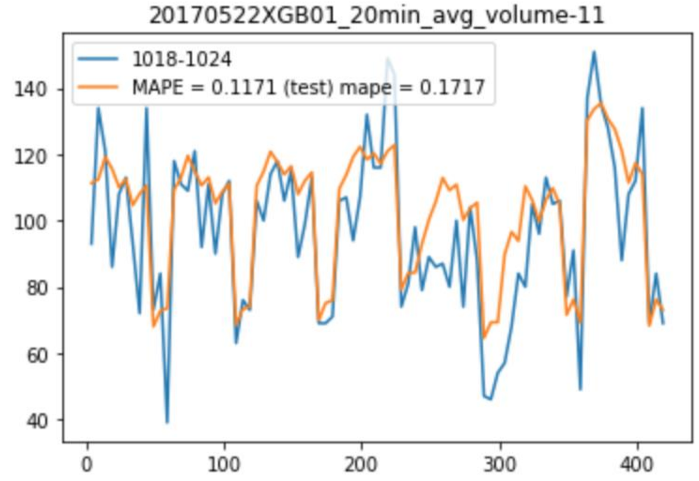
```
0.20051269831804733
```



```
0.09124546125632332
```



```
0.19215886572945845
```



```
0.14172328331091694
```

Volume Prediction 460 / 0.1717

- phase2 : predict 10/25– 10/31 average tollgate traffic volume.
- models :
  - using xgboost
  - features : hour, weekday, minutes, tollgate_id, direction
  - training data : 2016-10-11 to 2016-10-24, every 20 minutes
  - testing data : 2016-10-18 to 2016-10-24
  - test MAPE = 0.1104, real MAPE = 0.3418
  - test MAPE = 0.1358, real MAPE = 0.3446
  - test MAPE = 0.1724, real MAPE = 0.3716

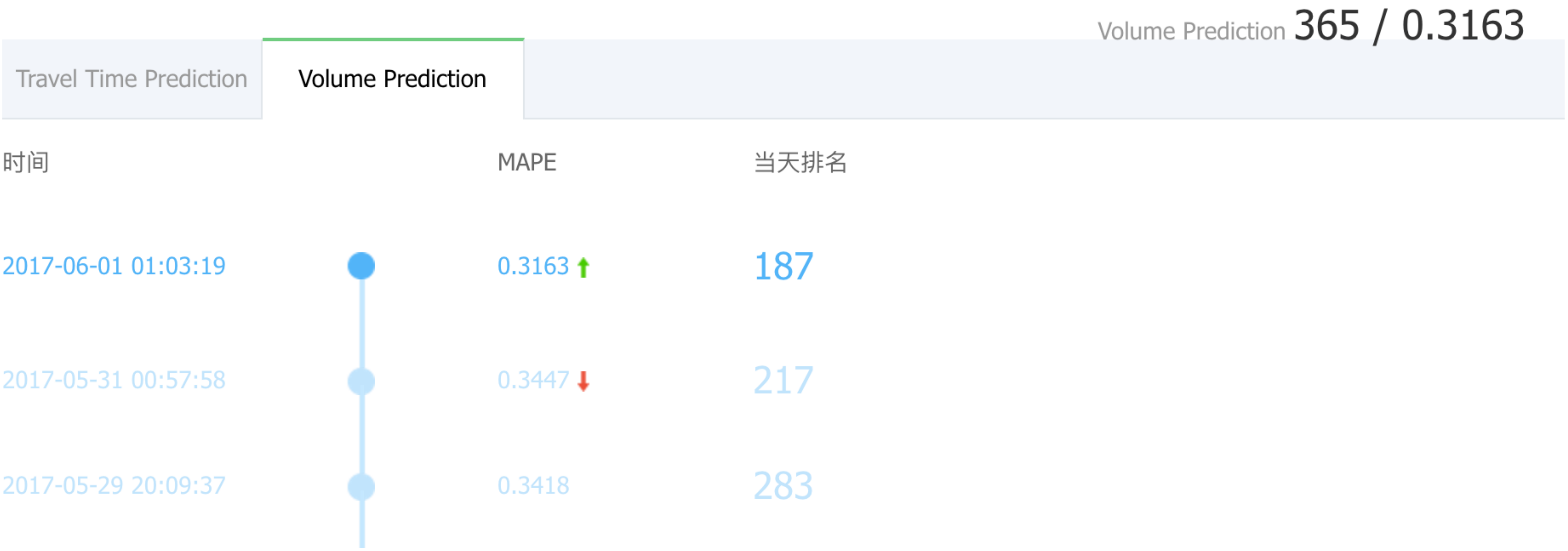
using 09-20 to 09-26data

real MAPE = 0.4324

using 10-11 to 10-17 data

real MAPE = 0.3950

  - training data : 2016-10-18 to 2016-10-24, every 20 minutes
  - testing data : 2016-10-18 to 2016-10-24
  - test MAPE = 0.0748, real MAPE = 0.3163



Volume Prediction 365 / 0.3163

时间	MAPE	当天排名
2017-06-01 01:03:19	0.3163 ↑	187
2017-05-31 00:57:58	0.3447 ↓	217
2017-05-29 20:09:37	0.3418	283

Thank you for listening