KDD CUP 2017: Volume Prediction Task Solution by CarTrailBlazer

Speaker: Suiqian Luo

Guazi.com

August 16, 2017

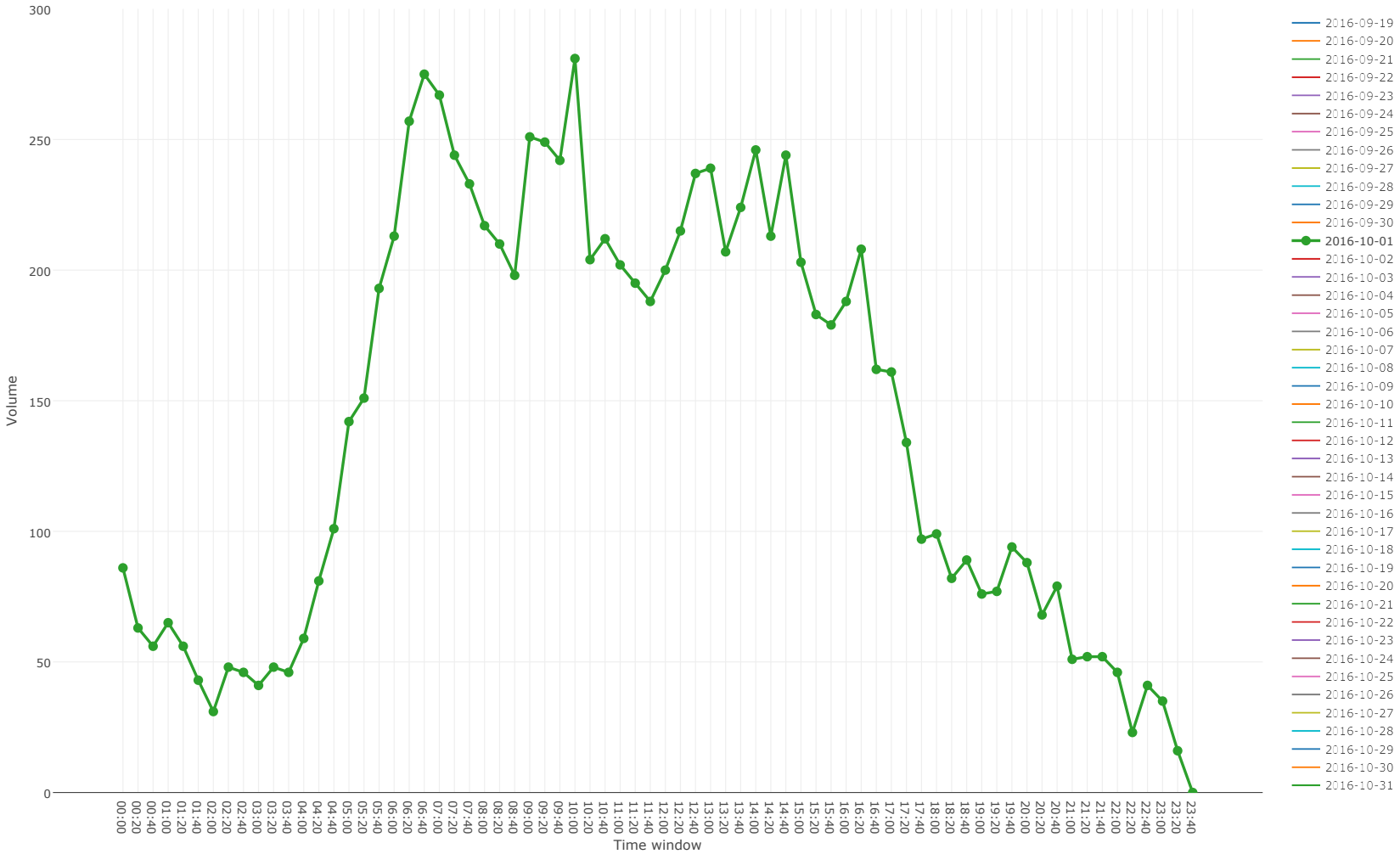
Overview

Overview

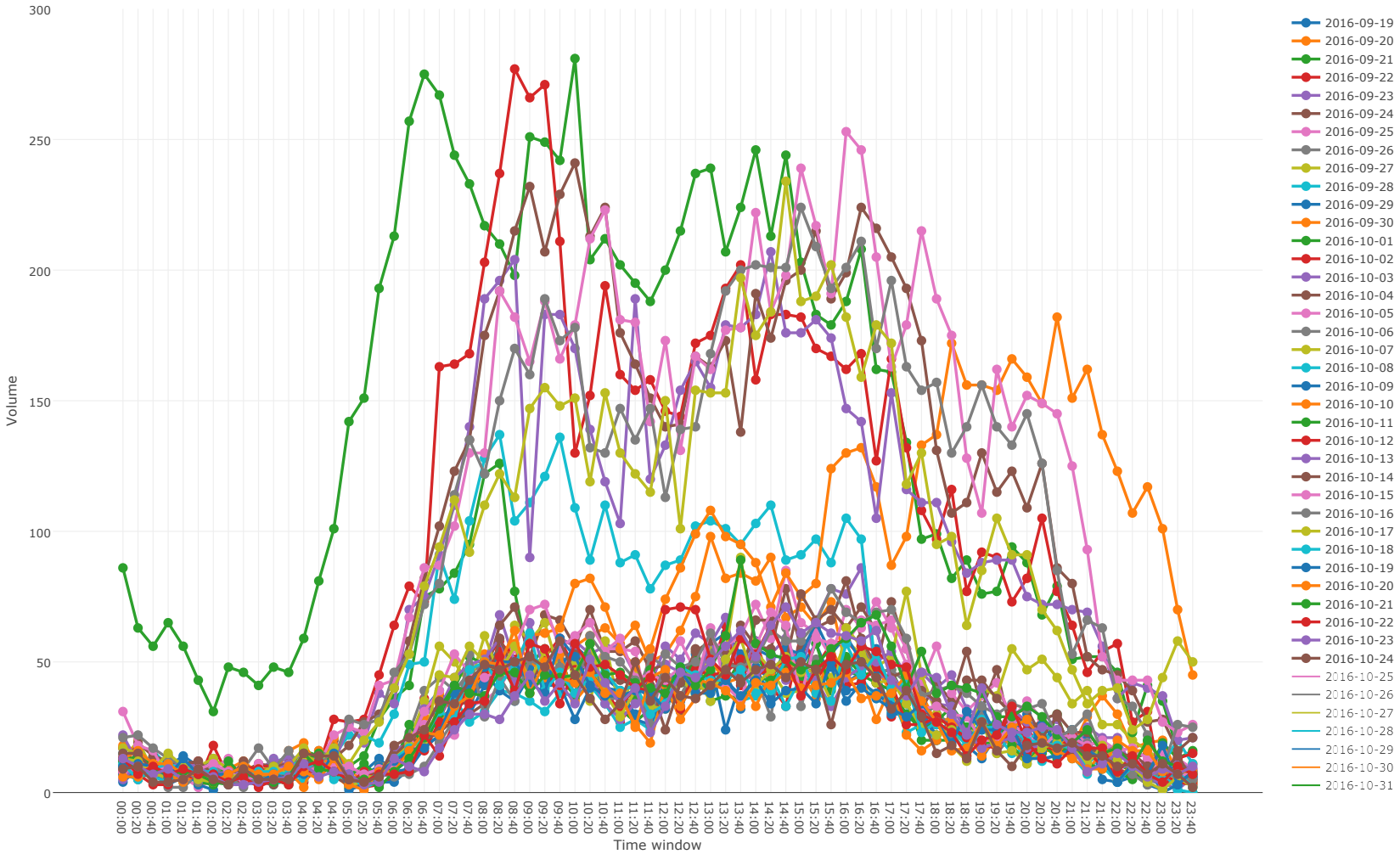
Label

- Generate the labels for each tollgate-direction pair
- For every 20-minute time window in a day
- Draw the graph of Tollgate 1, entry direction

Tollgate 1, entry direction, in 20161001



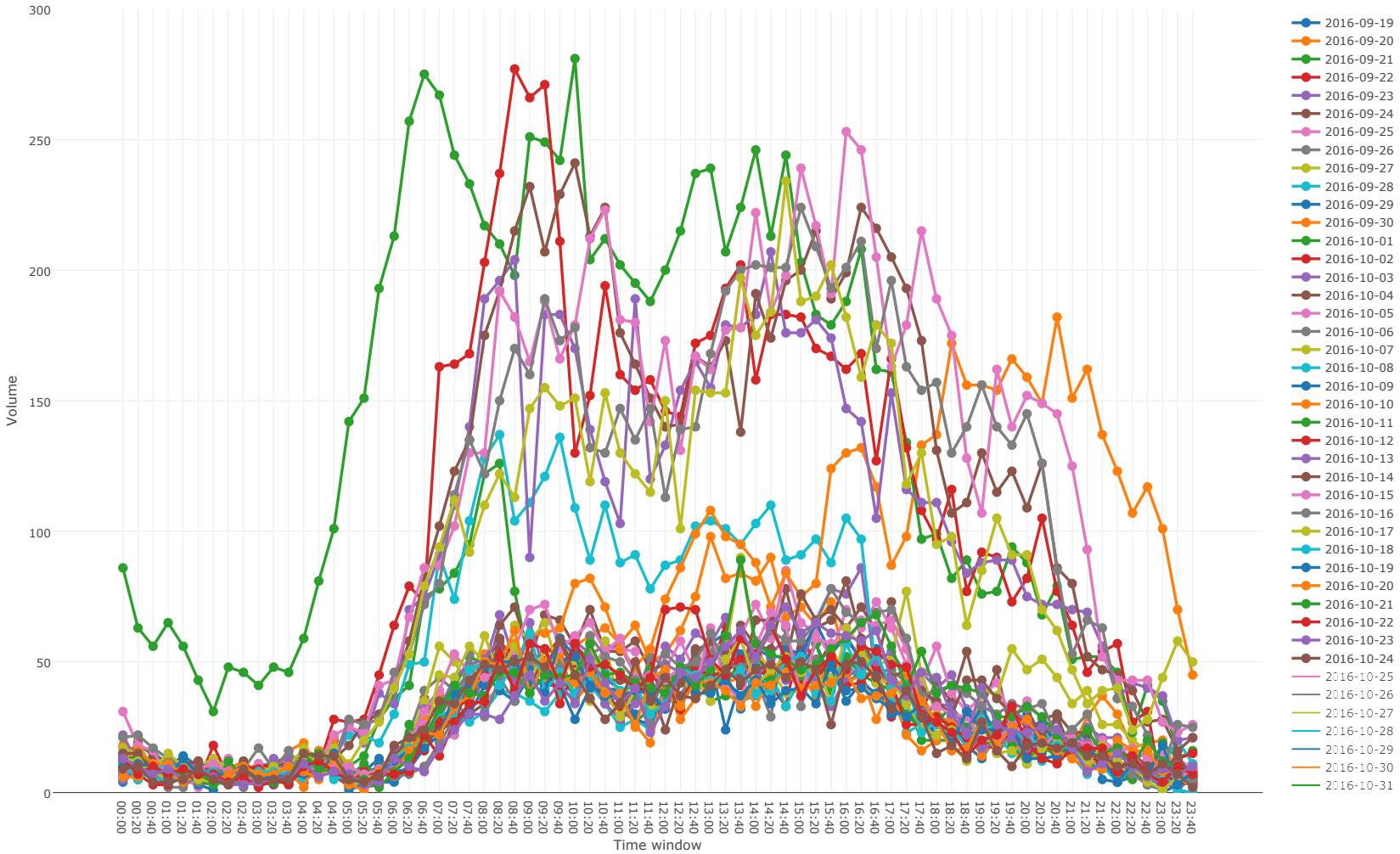
Tollgate 1, entry direction



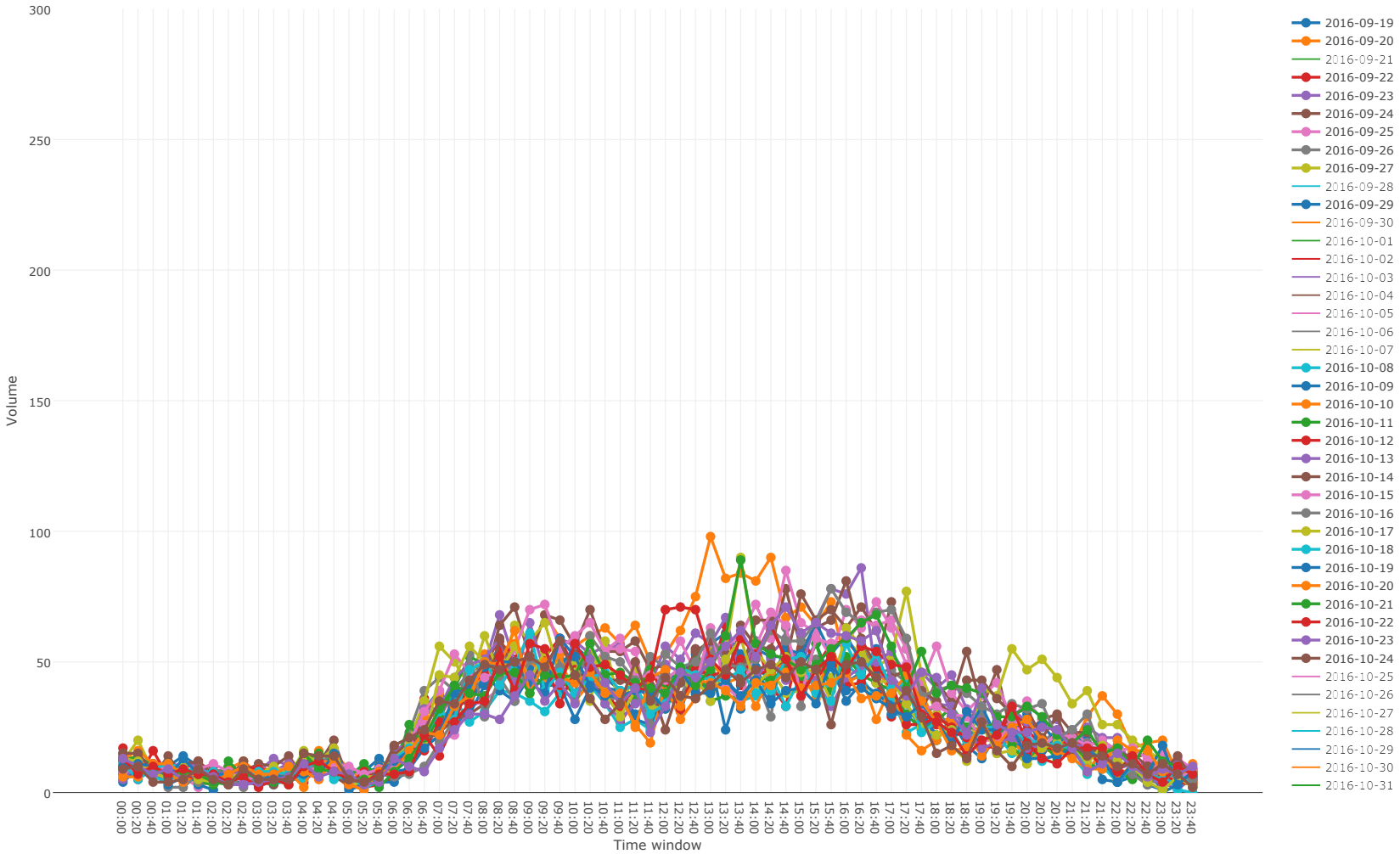
Pre-processing

- Clean the noisy data

Tollgate 1, entry direction



Tollgate 1, entry direction



Classification

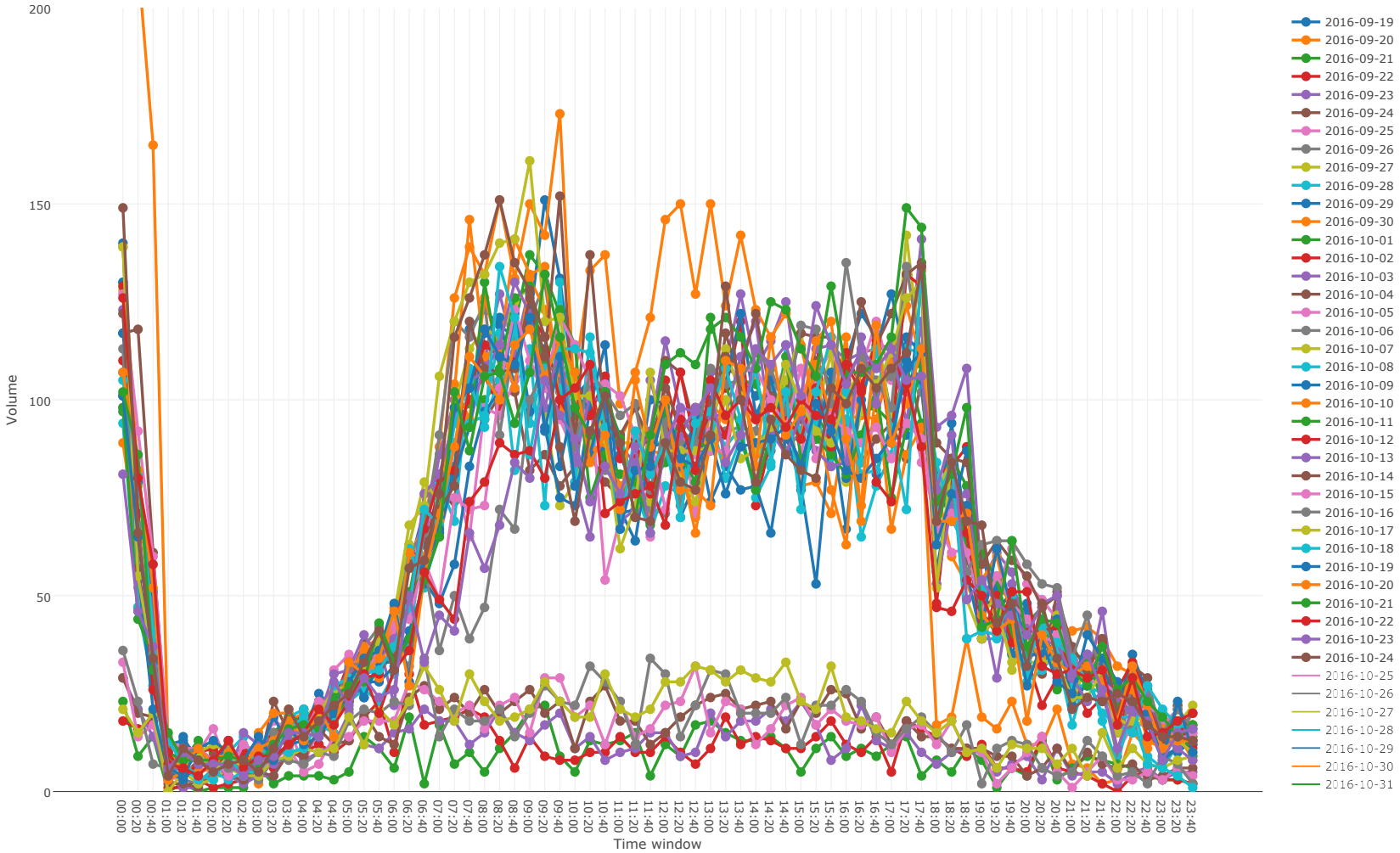
Date	Normal	Noisy
Tollgate 1, entry direction	0919-0920 0922-0927 0929 1008-1024	0921 0928 0930-1007

Table: The classification of the days

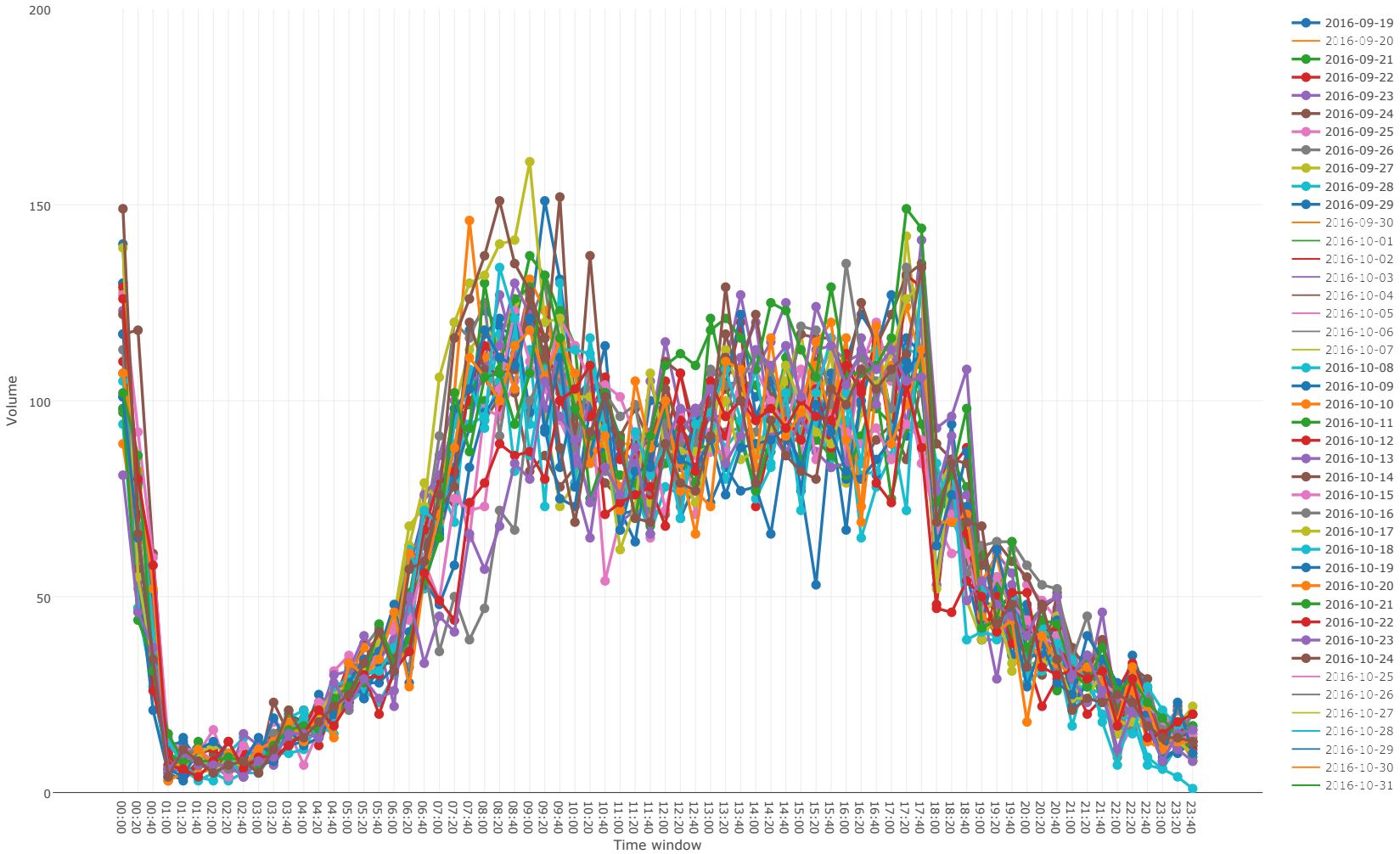
Label

- Draw the graph of Tollgate 1, exit direction in the same way
- Remove the noisy data

Tollgate 1, exit direction



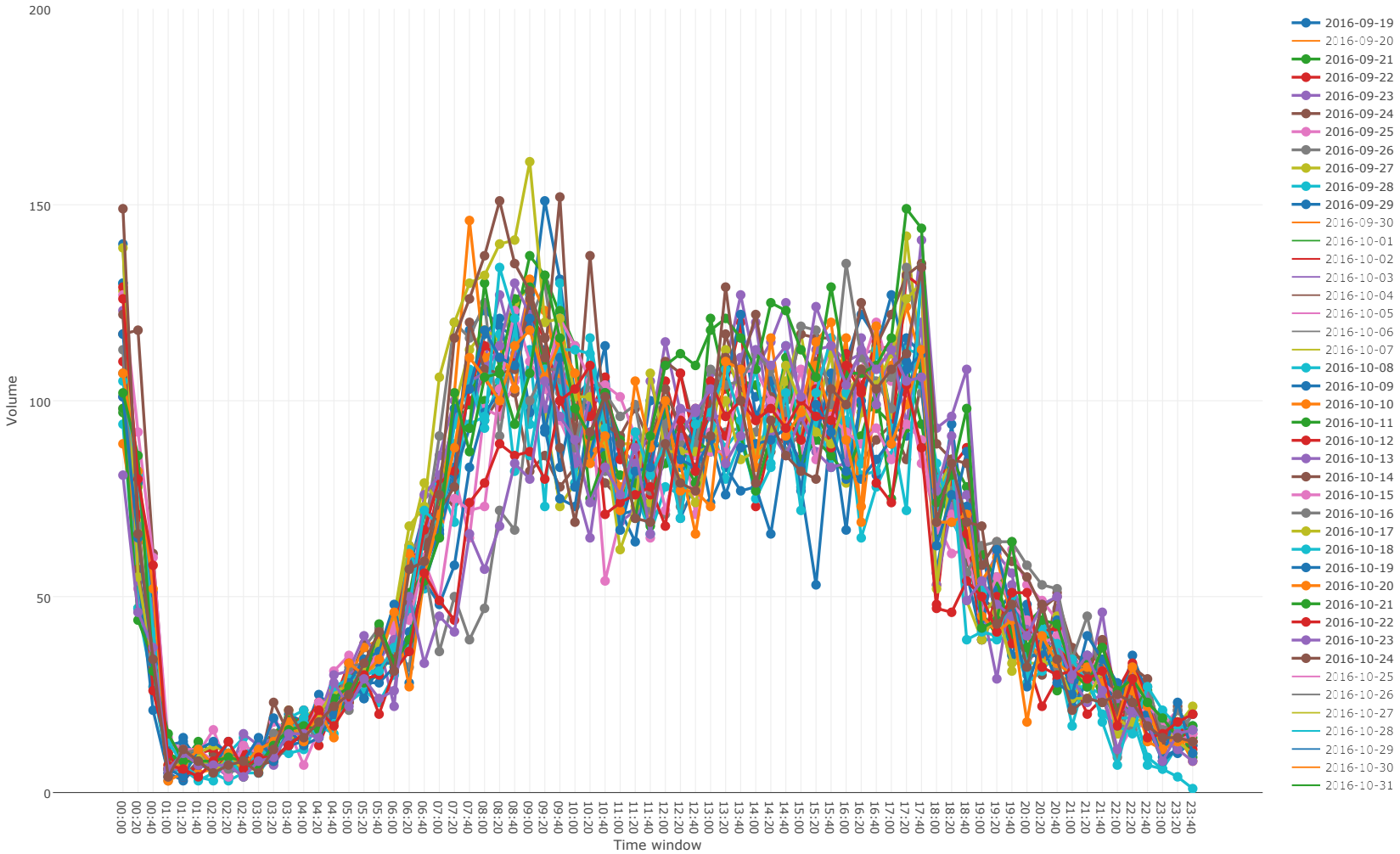
Tollgate 1, exit direction



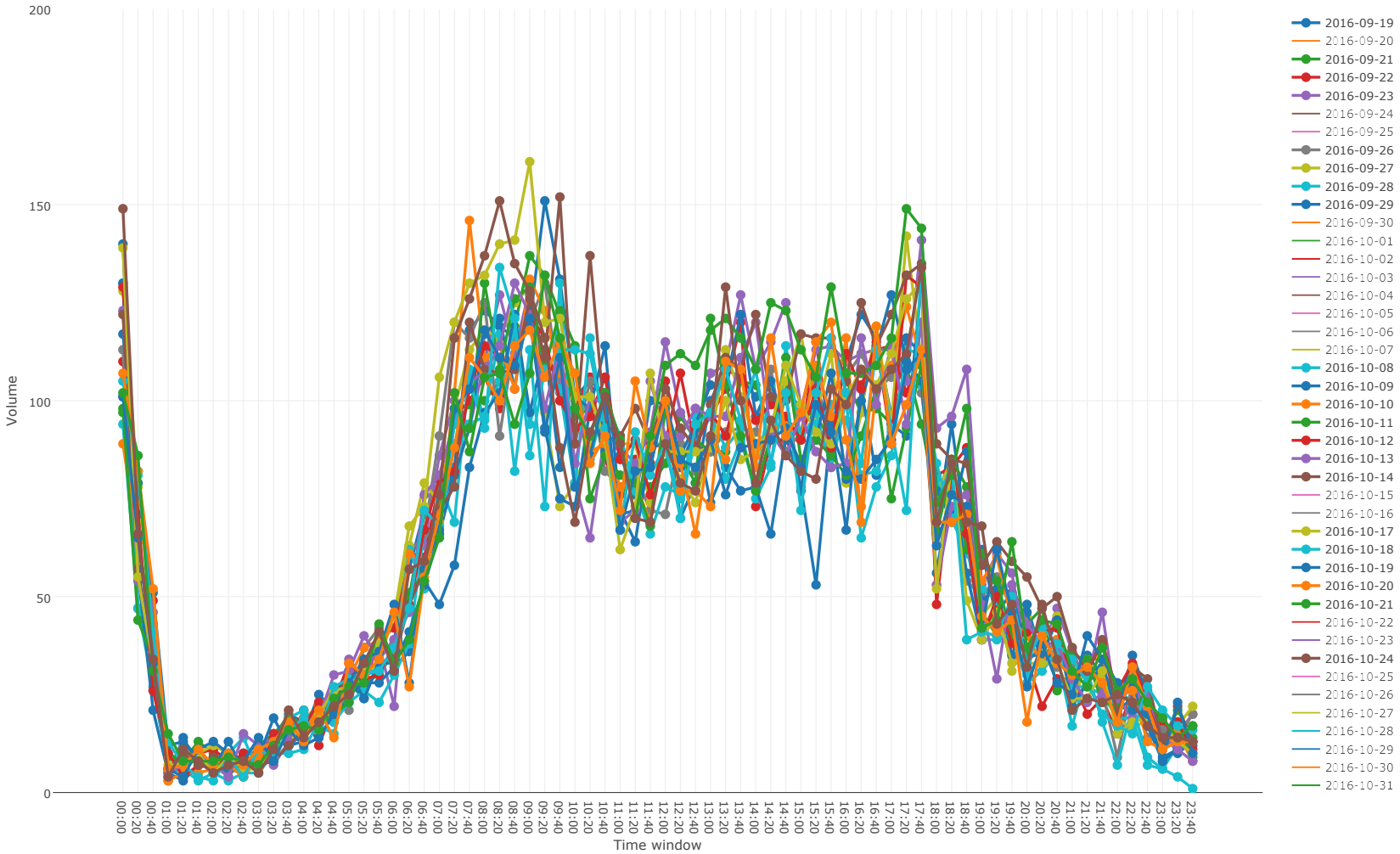
Classification

- After removing the noisy date, classify the days for this tollgate-direction pair.

Tollgate 1, exit direction



Tollgate 1, exit direction



Classification

- In this tollgate-direction pair, whether the day is weekend or not is an important feature.
- In summary, there are 3 different patterns for each day
  - Noisy day
  - Workday
  - Weekend

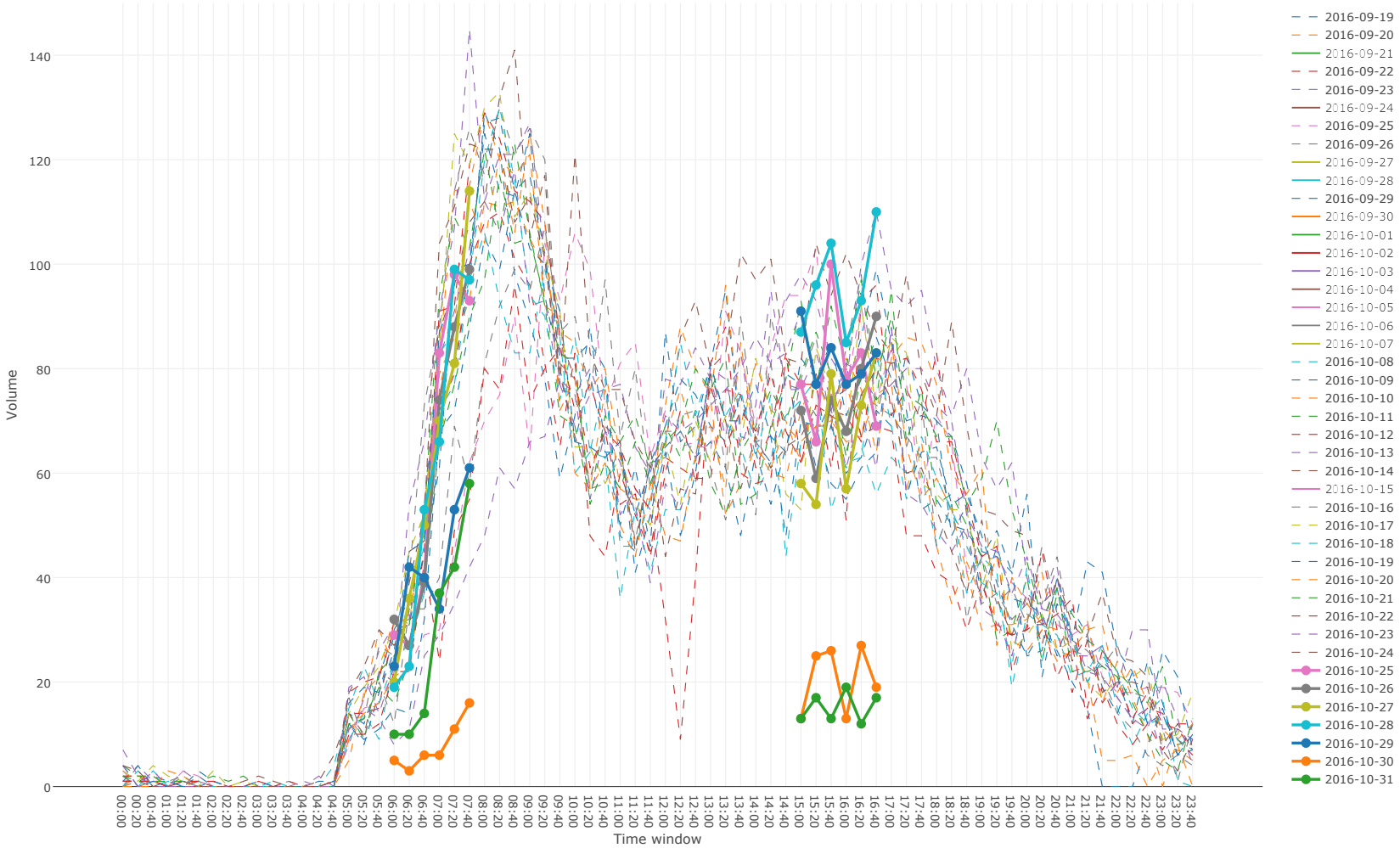
Classification

- In this way, classify the days for each tollgate-direction pair
  - 5 tollgate-direction pairs
  - Workday / weekend / noisy

Cleaning data

- Also, we need to classify the days in test data.
- There are 4 noisy days in the test data.
  - Tollgate 1, entry direction, 20161030
  - Tollgate 1, entry direction, 20161031
  - Tollgate 2, entry direction, 20161030
  - Tollgate 2, entry direction, 20161031

Tollgate 2, entry direction



Cleaning data

- All traffics in tollgate 2, entry direction have **has_etc**=1
- Feature **has_etc** is the key to this task.

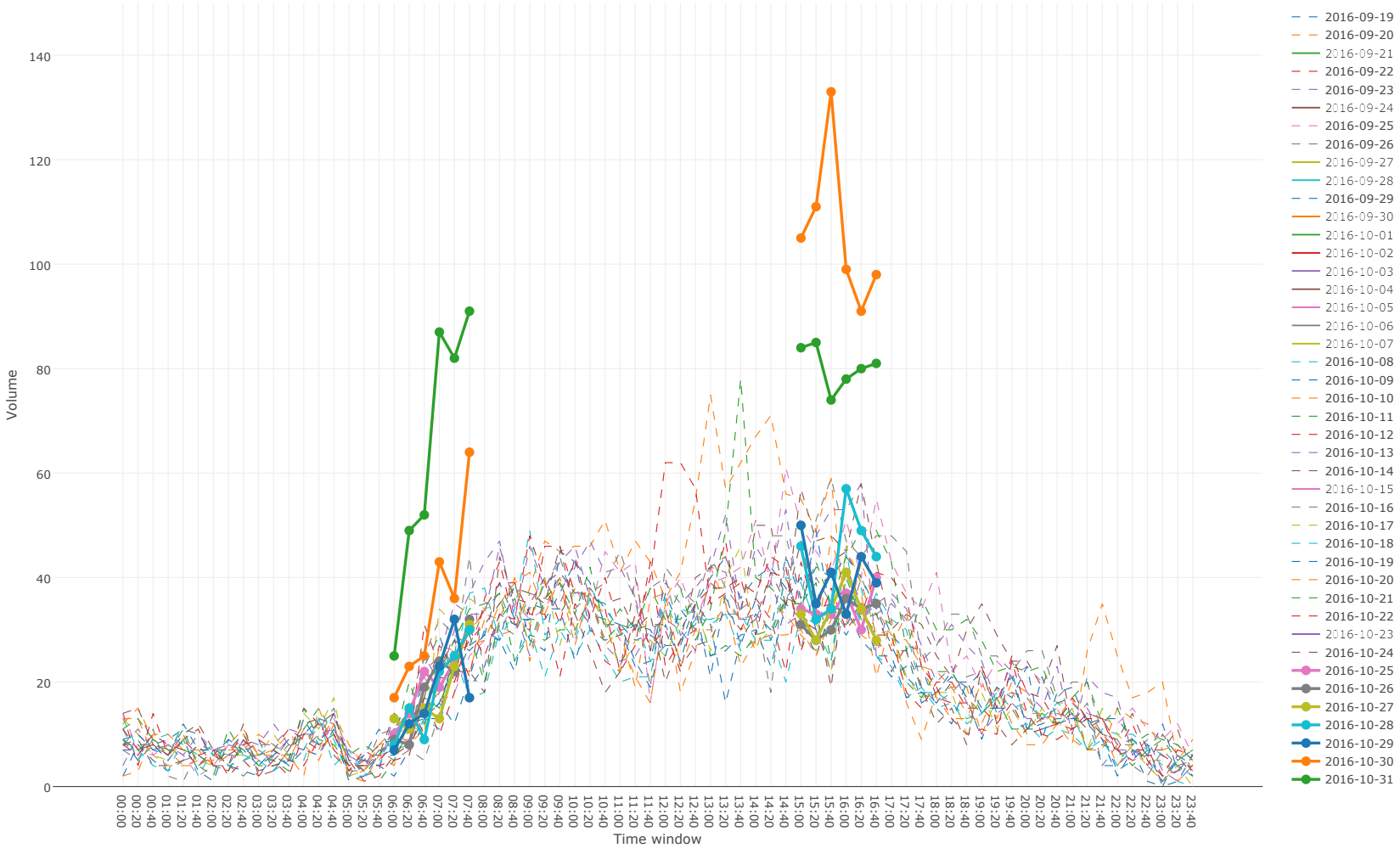
Classification

- Re-classify the days by with ETC and without ETC.
  - 5 tollgate-direction pairs
  - Workday / weekend / noisy
  - With ETC / without ETC

Cleaning data

- There are still two noisy days in the test data.

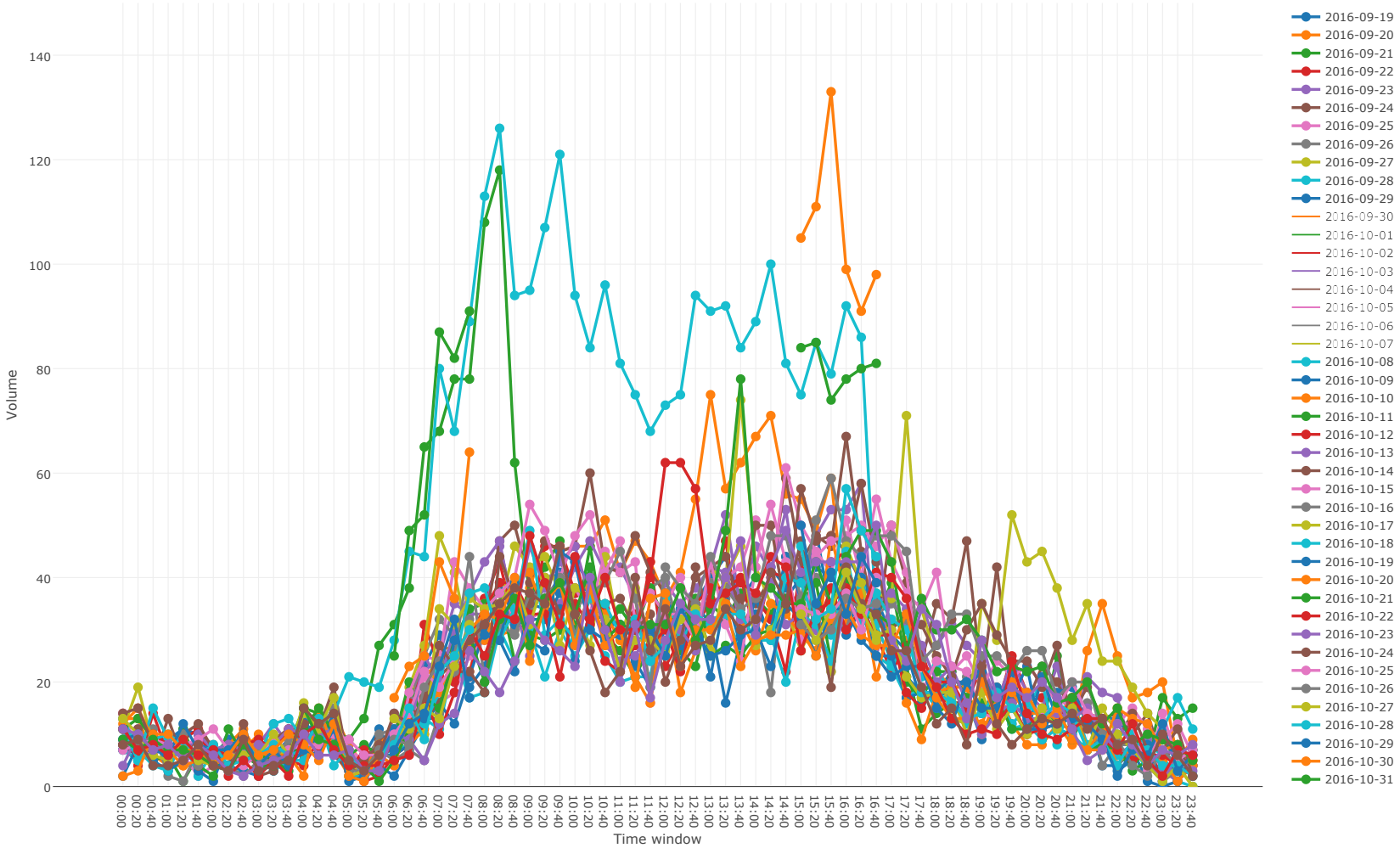
Tollgate 1, entry direction, without ETC



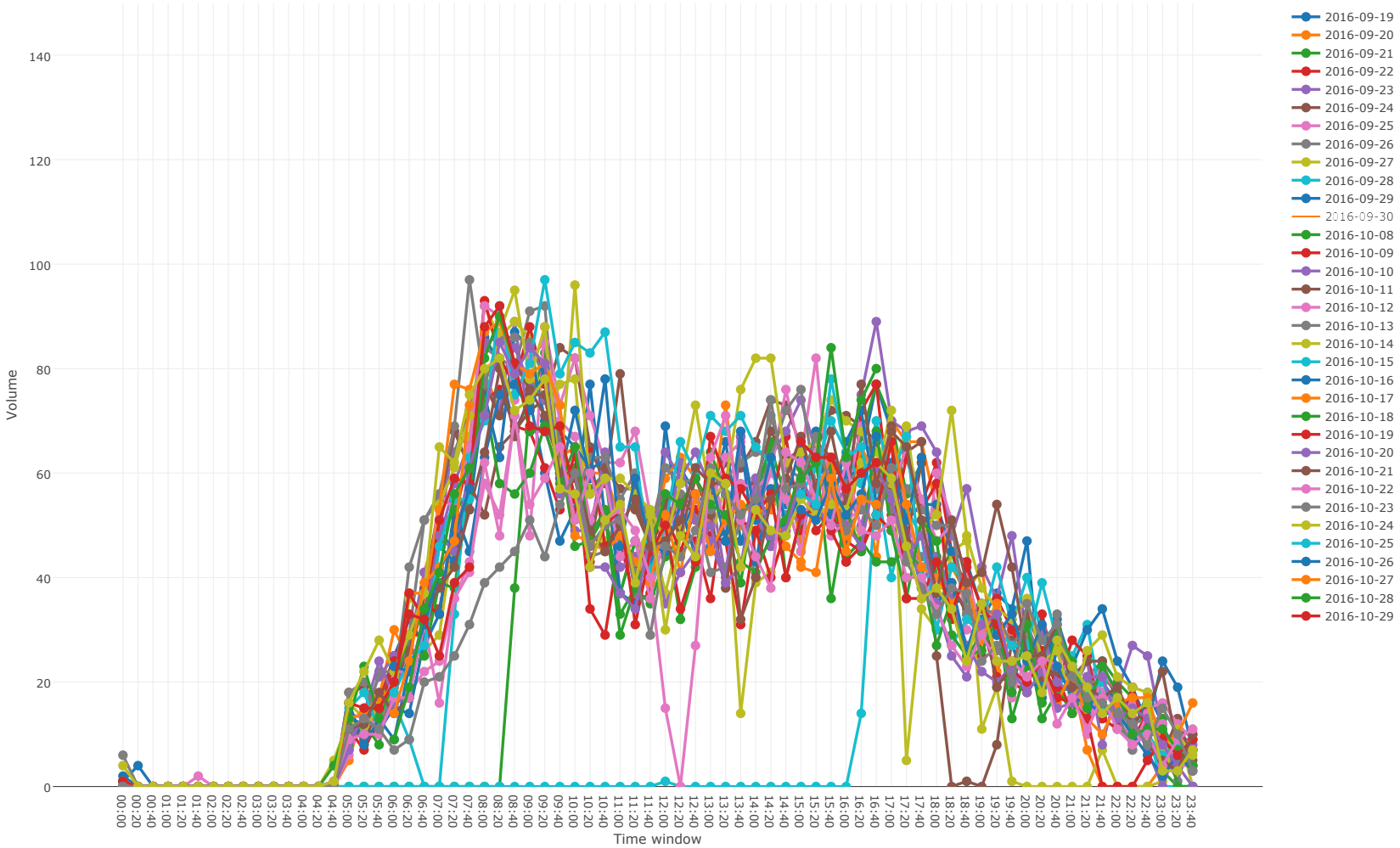
Cleaning data

- Inspect the traffic volume graphs in comparison.

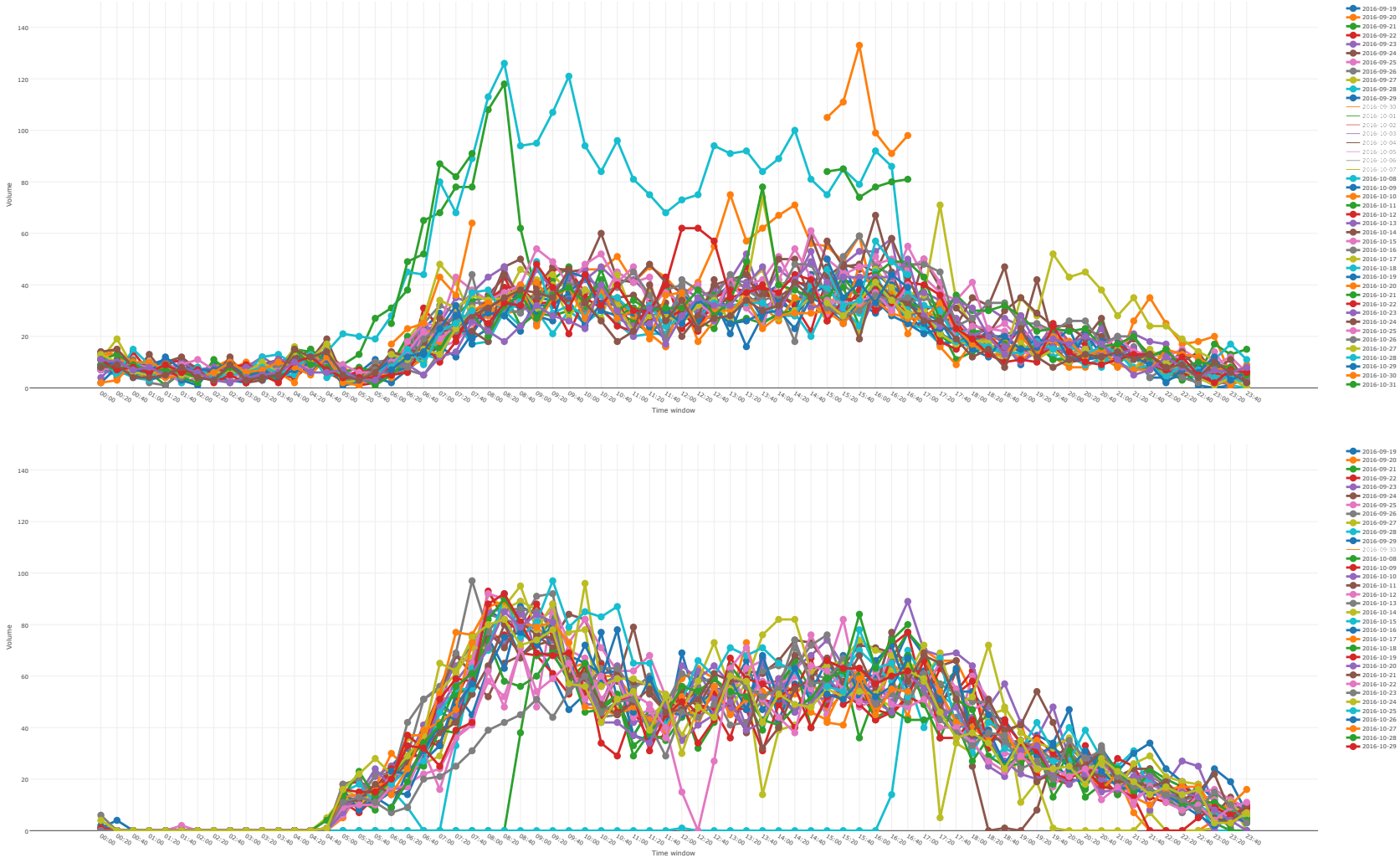
Tollgate 1, entry direction, without ETC



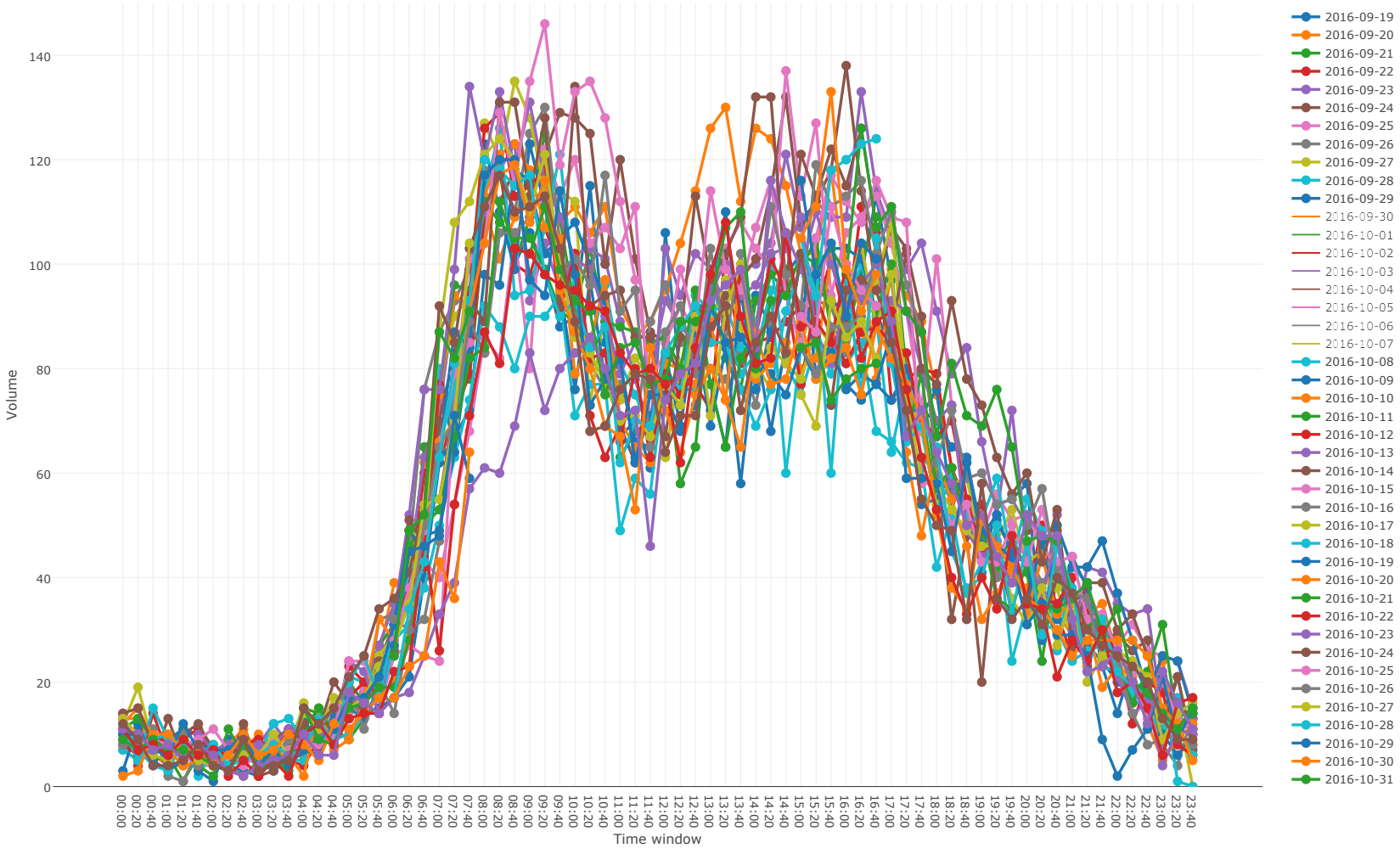
Tollgate 2, entry direction, without ETC



Tollgate 1 & 2, entry direction, without ETC



Tollgate 1 + 2, entry direction, without ETC



Cleaning data

- Important assumption: some traffic volumes are miscounted.
- Some traffic volumes in tollgate two, are actually counted in tollgate one by mistake.

Cleaning data

- Add a fake tollgate indicating the sum of tollgate 1 and tollgate 2.
- Replace the predictions in tollgate 2, entry direction with the fake tollgate in these two noisy days.

Classification

- Classify the days for all the data
  - 5 tollgate-direction pairs and a fake
  - Workday / weekend / noisy
  - With ETC / without ETC
- We will apply the model within the same type of days.

Cleaning data

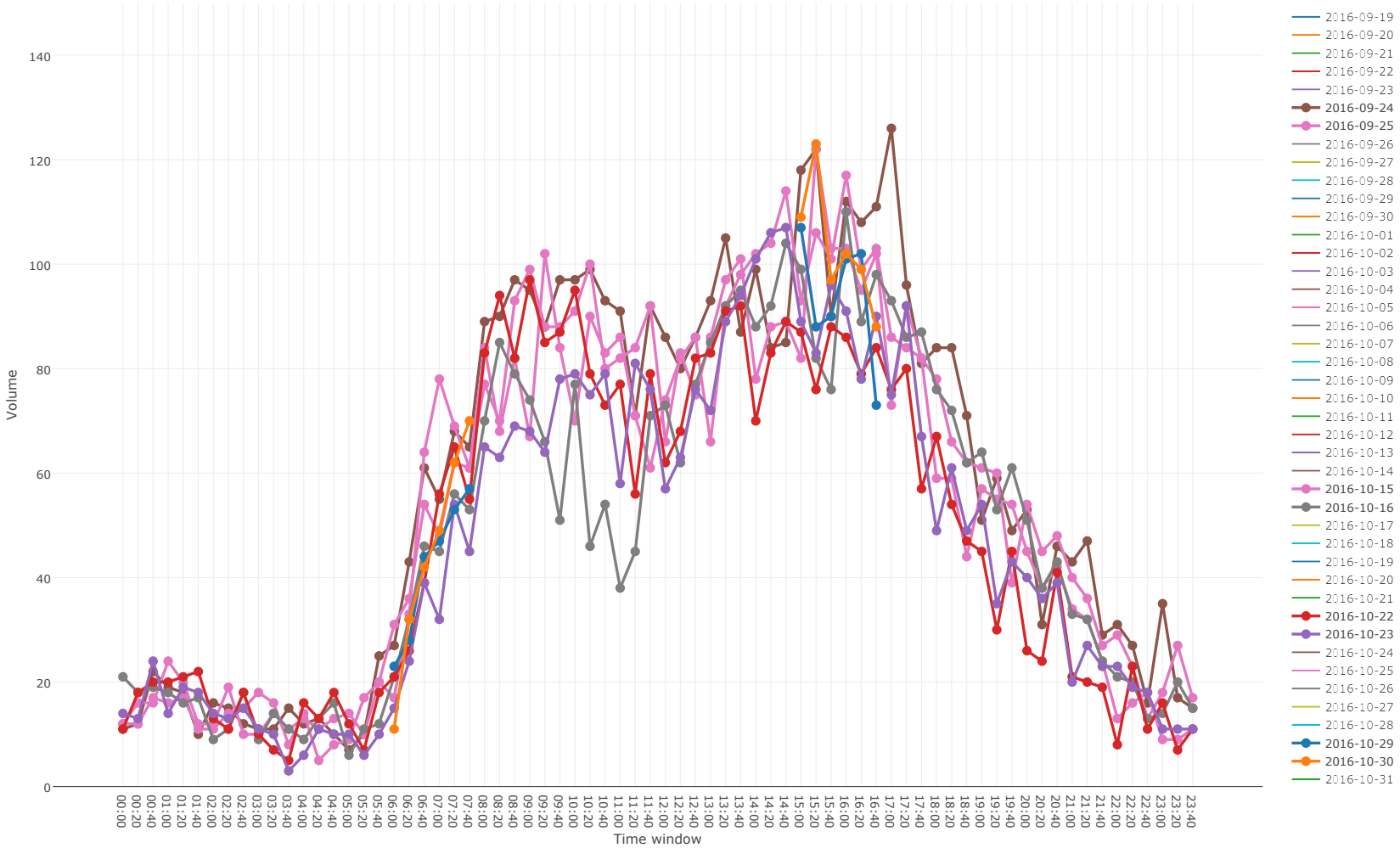
- Finally, we have cleared all the noisy days in the test data.

Overview

Feature

- This task is a regression problem
- The number of training samples is rather small

Tollgate 3, entry direction, without ETC, weekend



Feature

- There are only 6 training examples in this case
- To avoid overfitting, we should use as few features as possible
- With cross validation, we choose only **one** feature
  - The traffic volume right before the predicted time window
  - 80-minute analysis window

Linear regression

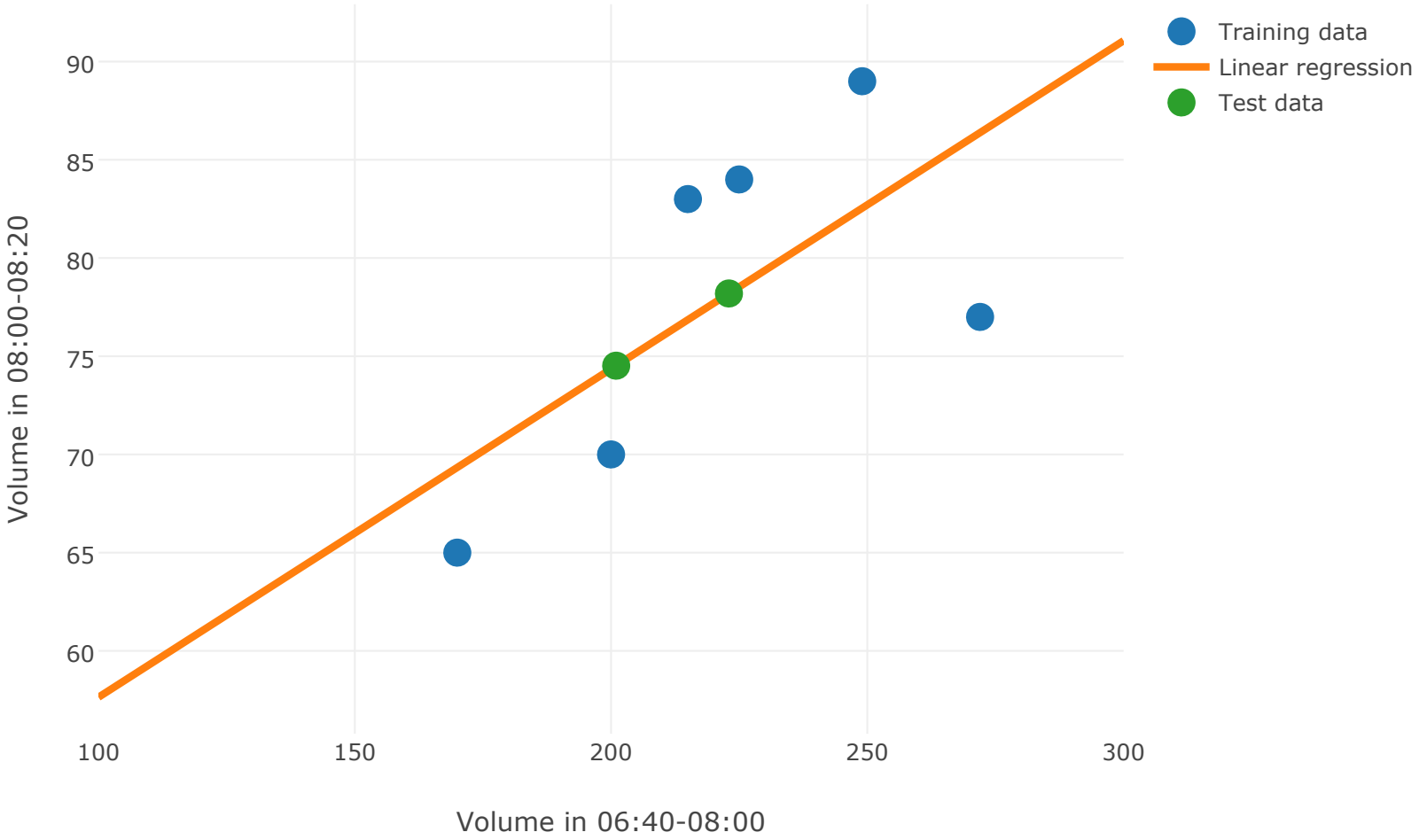
- It is hard to apply an complex model
- We use linear regression to train and predict

Data example

Volume	06:40-08:00	08:00-08:20
0924	249	89
0925	225	84
1015	272	77
1016	200	70
1022	215	83
1023	170	65
1029	201	??
1030	223	??

Table: The data of tollgate 3, entry direction, without ETC, weekend

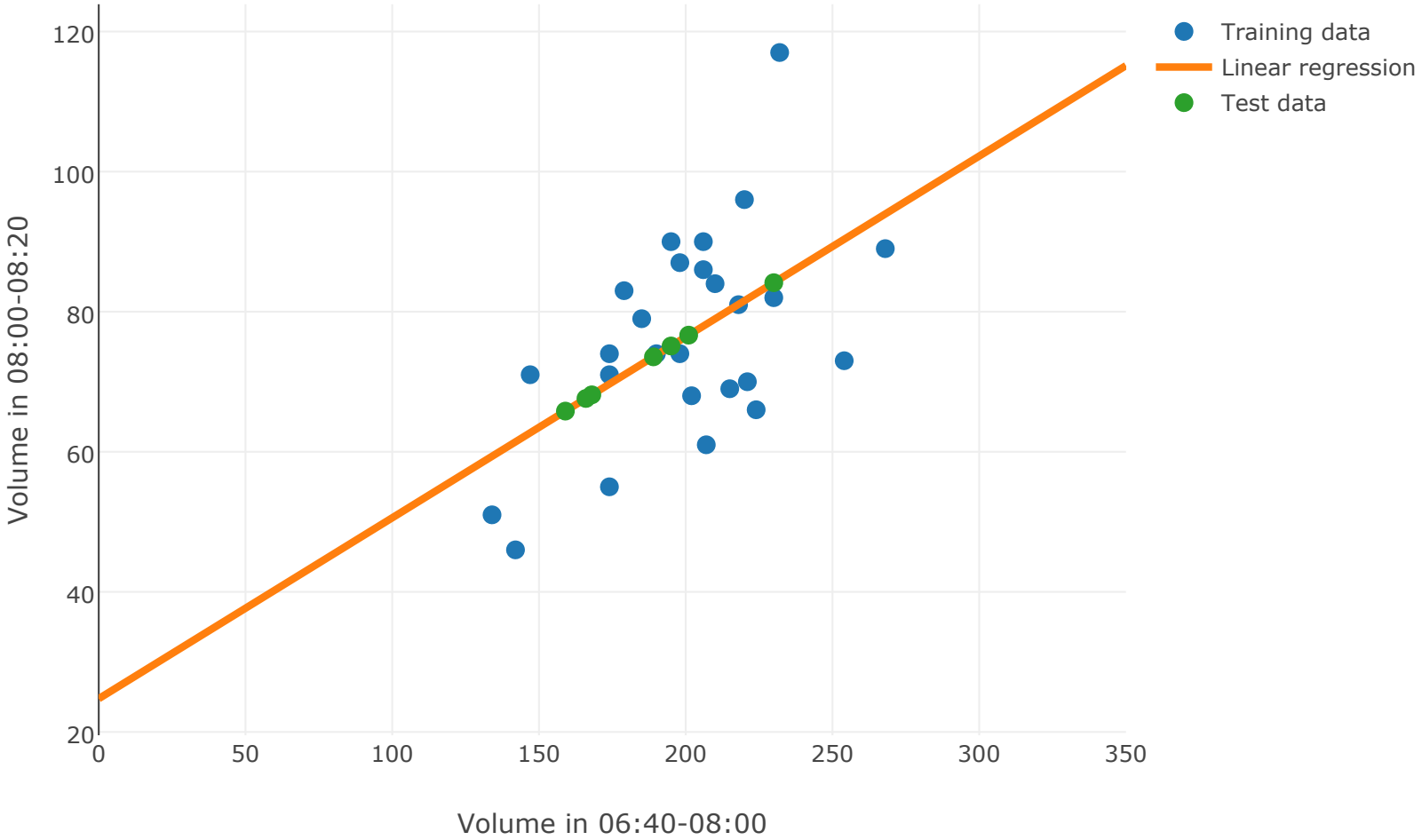
Tollgate 3, entry direction, without ETC, weekend



Linear regression

- Tollgate 3, exit direction, without ETC
- 26 training samples, 7 test samples

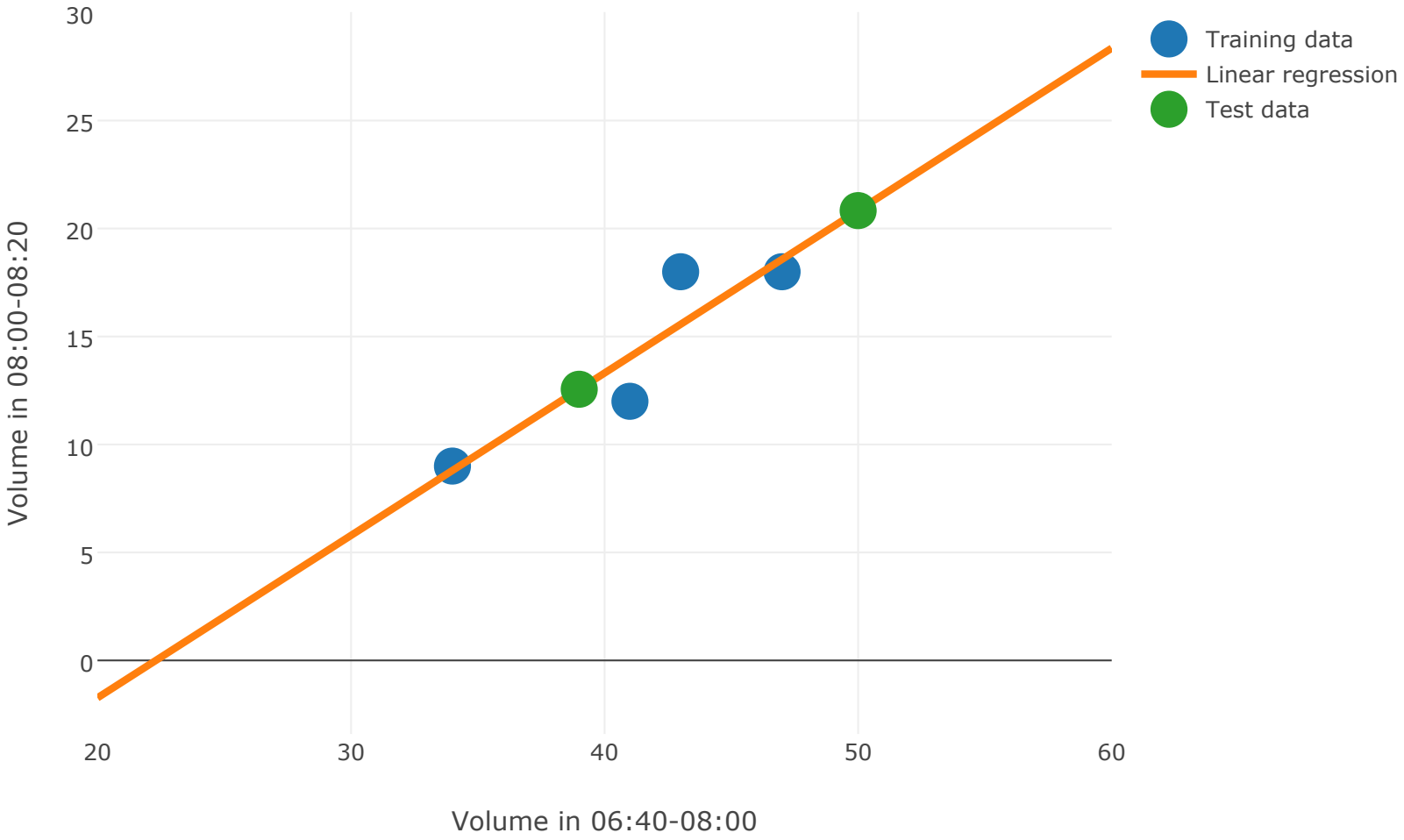
Tollgate 3, exit direction, without ETC



Linear regression

- Tollgate 2, entry direction, with ETC, weekend
- 4 training samples, 2 test samples

Tollgate 2, entry direction, with ETC, weekend



Post-processing

- Let f_i and p_i be the actual and predicted value.
- The cost in linear regression is the root-mean-square error (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - f_i)^2}$$

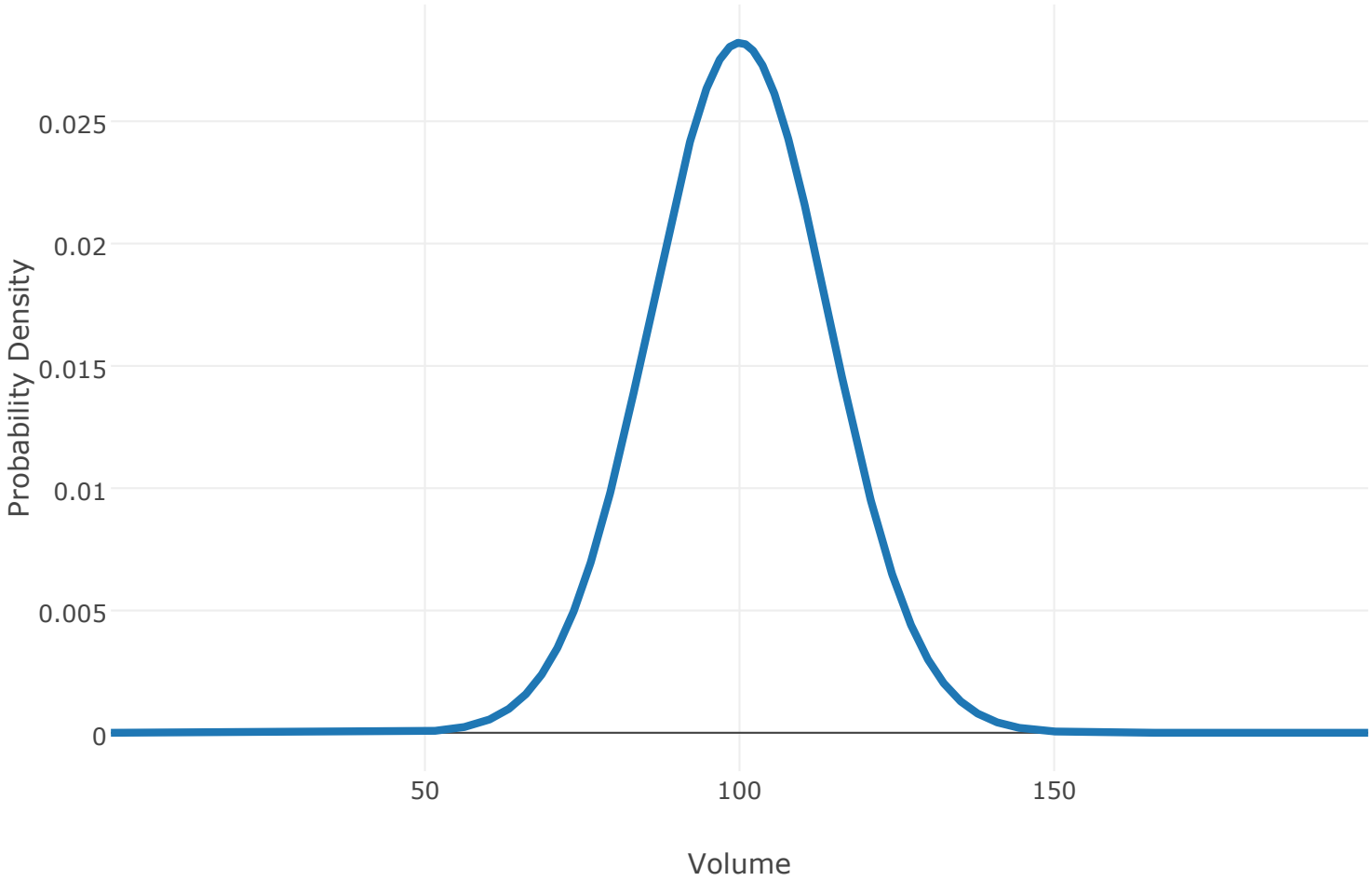
- But in this task, the target is to minimize mean-absolute-percentage error (MAPE).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{p_i - f_i}{f_i}\right|$$

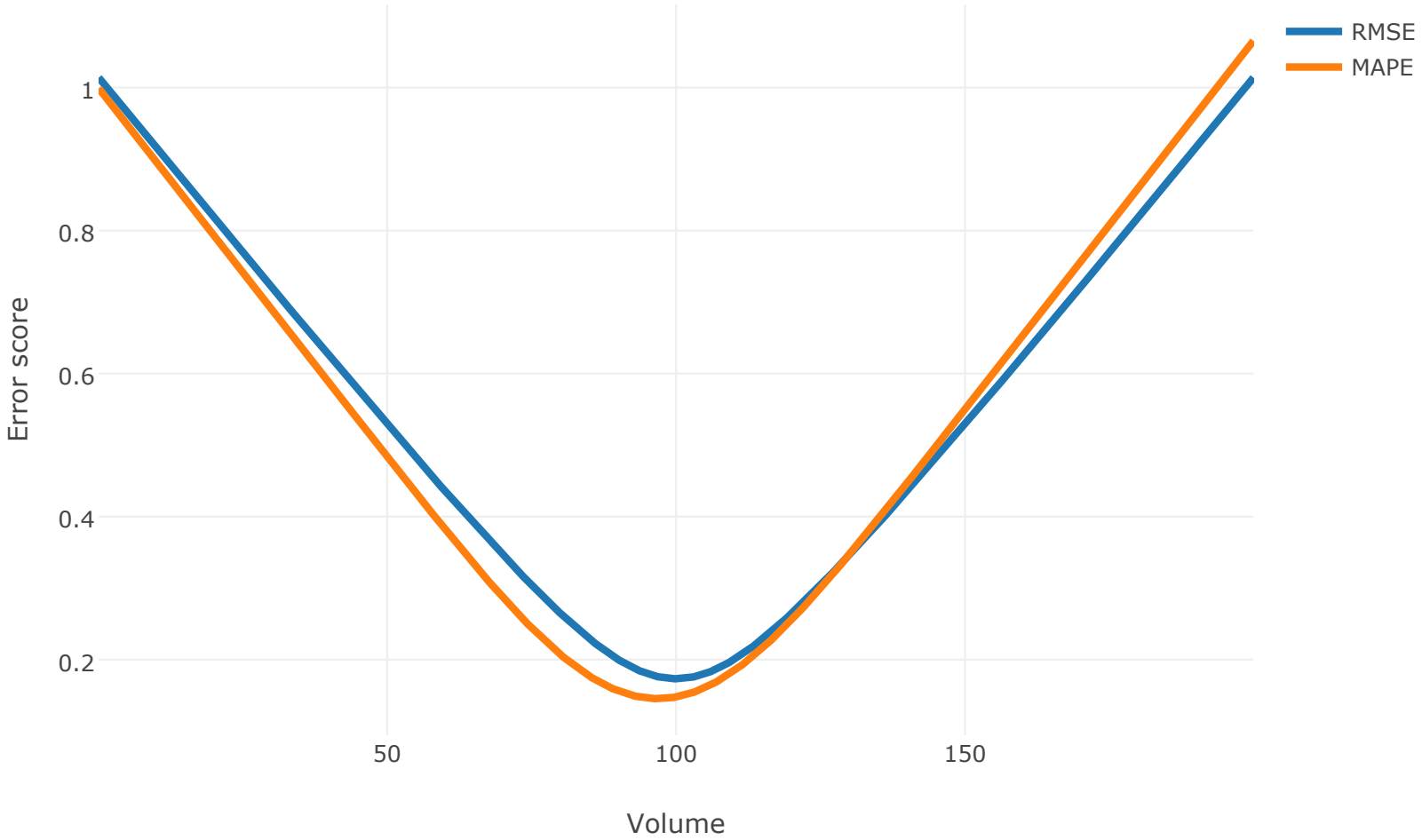
Post-processing

- The MAPE score penalizes more when the ground truth is smaller.
- We can optimize the MAPE score directly in the model.
- However, we choose a simpler method.
  - Let's make a comparison between RMSE and MAPE score.

Gaussian distribution ($\mu = 100, \sigma^2 = 300$)



RMSE & MAPE



Post-processing

- There is slight difference between RMSE and MAPE score.
- The RMSE reaches minimum when $volume = 100$
- The MAPE reaches minimum when $volume = 96.8$

Post-processing

- However, we do not have enough training data to infer the parameter of the distribution.
- All the prediction values scale a constant number.
- Use cross validation to determine that the constant number is 0.96.

Overview

Summary

- Classify the days for each tollgate-direction pair in all the data.
- Use linear regression within the same type of days.
- All the prediction numbers scale 0.96 due to the MAPE score.
- The road network, **vehicle_model**, **vehicle_type** and weather data are not used in our solution.

About Guazi.com

- China's largest used car trading platform
- Covering more than 150 cities in China
- Reform of traditional used car industry by using big-data and artificial intelligence

瓜子
二手车直卖网
Guazi.com

Thank you for listening!