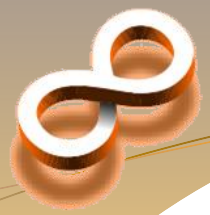


Data Analysis for CDN Server Population Estimation with CRM Model

HSU, CHENG 許誠
Network and Systems Laboratory
Graduate Institute of Communication Engineering
National Taiwan University

Outline

- Introduction
- Data Analysis
- Clustering
- CRM Model Result
- CJS with Different Data
- Computation Time of CJS

Introduction

- Twitch
- Capture-Recapture Model
- Problem Statement
- Contribution

Twitch

- A platform hosts game streams and eSport events
- Dominate the live video traffic
  - more than 70% of the game streaming market
- Statistics (in 2021)
  - average concurrent viewers: 2,778,000
  - average concurrent streamers: 105,000

Twitch's CDN

- One-time experiment have been done (2017)
- Properties
  - Large scale
  - Rapid growth
- Remain large unknown to public today

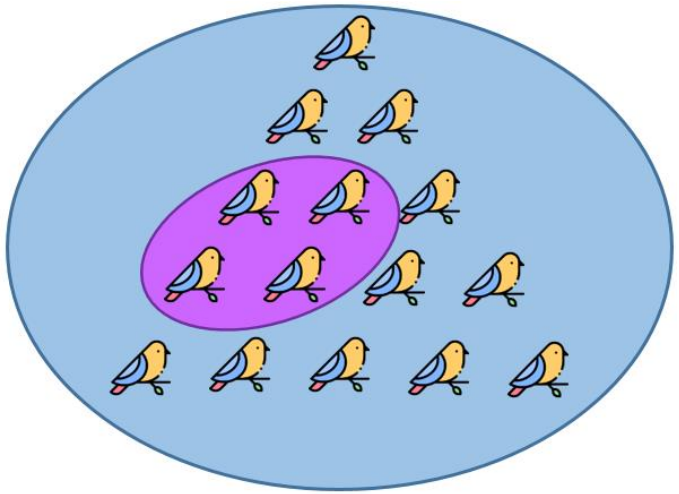
Our goal is to detect CDN continuously with lower traffic

Capture-Recapture Model (CMR)

- Usually use in biology for population estimation
- Population is hard to get ground truth
- CRM models:
  - Lincoln-Petersen Model (LP model)
  - Cormack-Jolly-Seber model (CJS model)
  - CRM model with Heterogeneity

LP Model

- The simplest CRM model
- Steps
  1. Capture a few animal, mark, and release
  2. Capture again
  3. Use the ratio of marked animal to estimate the population



CJS Model

- Open population
  - Birth, death, move in, move out
- Parameters
  1. Time-dependent survival rate
  2. Time-dependent capture probability

CRM Model with Heterogeneity

- Allow Individual Heterogeneity
  - Individual from different classes has different survival rate and capture probability
- Use Maximum Likelihood Estimation

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \sum_{c=1}^{C} \sum_{d=\ell_i}^{K} \left\{ \pi_c \left(\prod_{j=f_i}^{d-1} \phi_{jc} \right) (1 - \phi_{dc}) \times \left(\prod_{j=f_i+1}^{d} p_{jc}^{x_{ij}} (1 - p_{jc})^{1-x_{ij}} \right) \right\}.$$

p: Capture Probability
φ: Survival Rate
$\boldsymbol{\pi}$: Ratio of Class c

Problem Statement

- Previous work:
  - Use the CJS model to estimate servers population
  - In the CJS model, all servers share the same time-dependent survival rate and capture probability
- However, the different servers may have different survival rates and capture probabilities.

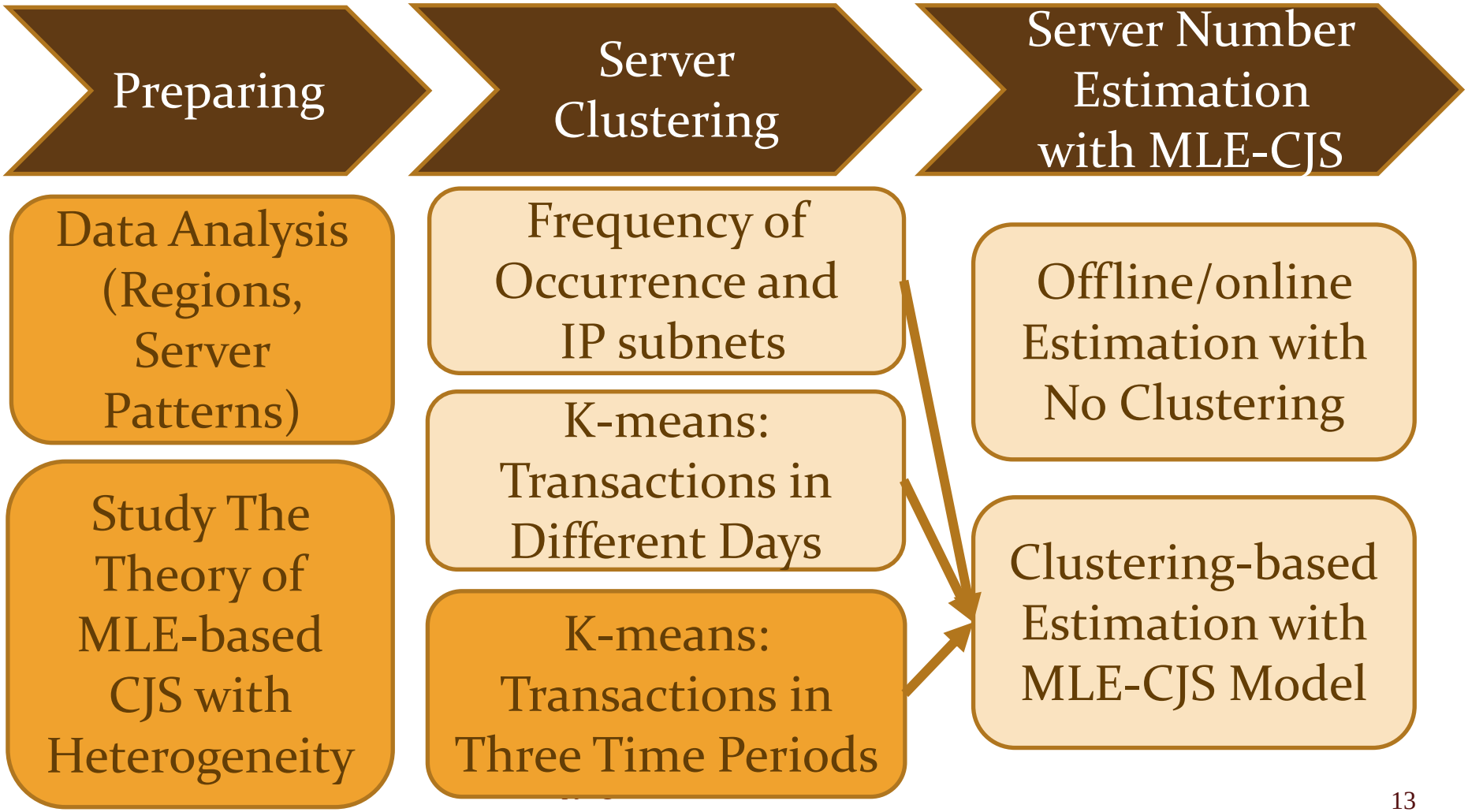
Problem Statement

- Most general model:
  - Each servers has its own time-dependent survival rate and capture probability
- Disadvantage:
  - High computation overhead
- Solution:
  - Use clustering on servers to lower computation overhead of the CRM model with heterogeneity.

Problem Statement

- Improve the CDN servers estimation model by using data analysis to find a clustering way for **the CRM model with heterogeneity**

Co-work with Wung Wei-Shiang



Contribution

- Study the theory of CRM model with heterogeneity
- Divide the Twitch's CDN servers into several groups by transaction counts in different periods for the CRM model with heterogeneity
- Survey the relationship between clustering results and the CRM model performance.

Data Analysis

- Dataset Introduction
- Data in Different Regions
- Continuous Data
- Hour-Count Distribution

Dateset Introduction

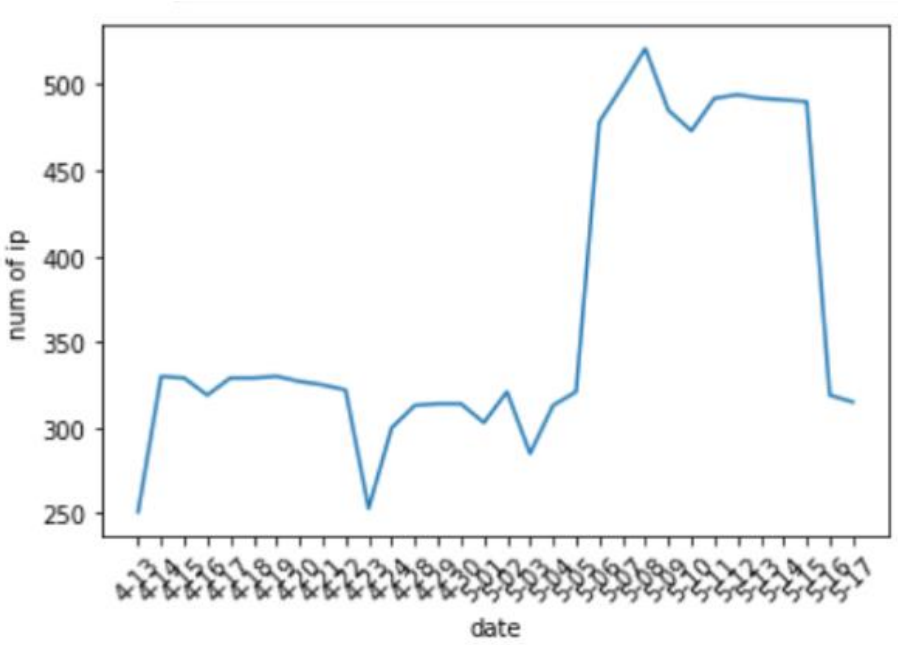
- Collected by Caleb Wang in 2021 April ~ 2021 May
- Goal: choose top K channels that account for 80% of total viewers
- Result: Wang selected VPN servers in 18 different countries that contain about 75% traffic on Twitch.

Dateset Introduction

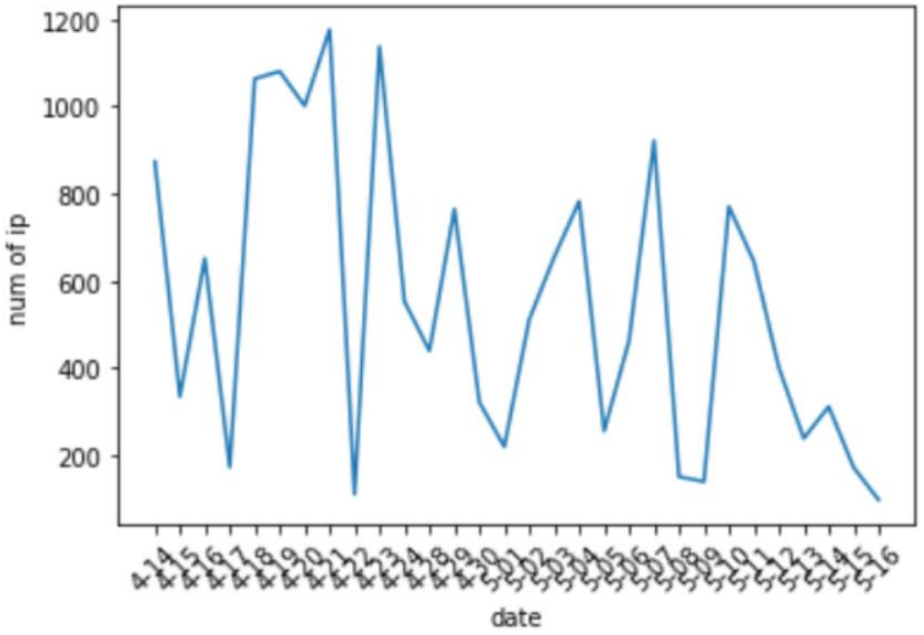
- Data Structure
- ['_id', 'vpnServerId', 'channel', 'language', 'serverPool', 'start', 'end', '**transactionList**', 'addrPool']
- 'transactionList': This attribute is the record of probing. It contains a list of times and the corresponding server ips.

  e.g. {'2020-10-19T14:56:08': '52.223.247.211',
  '2020-10- 19T15:02:04': '45.113.128.160'}

Date in Different Regions

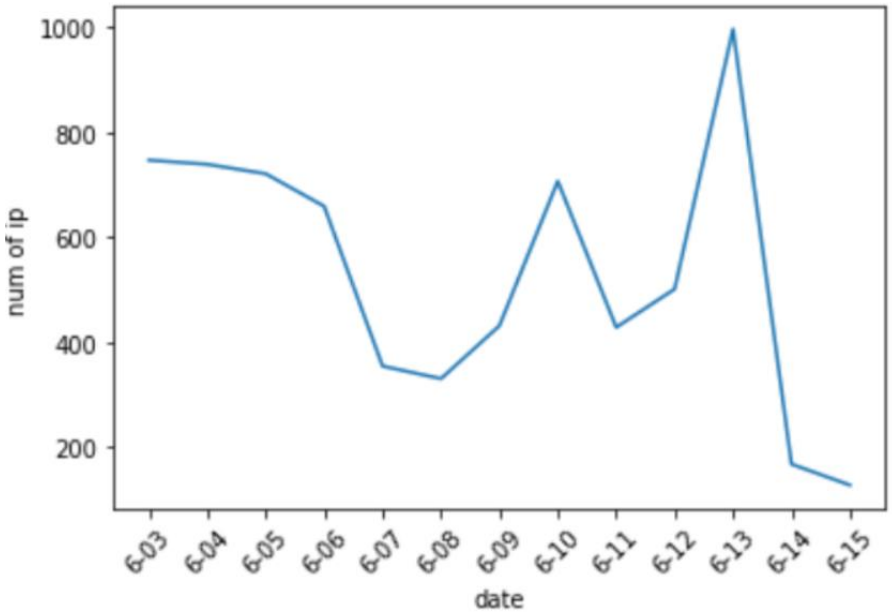


United States

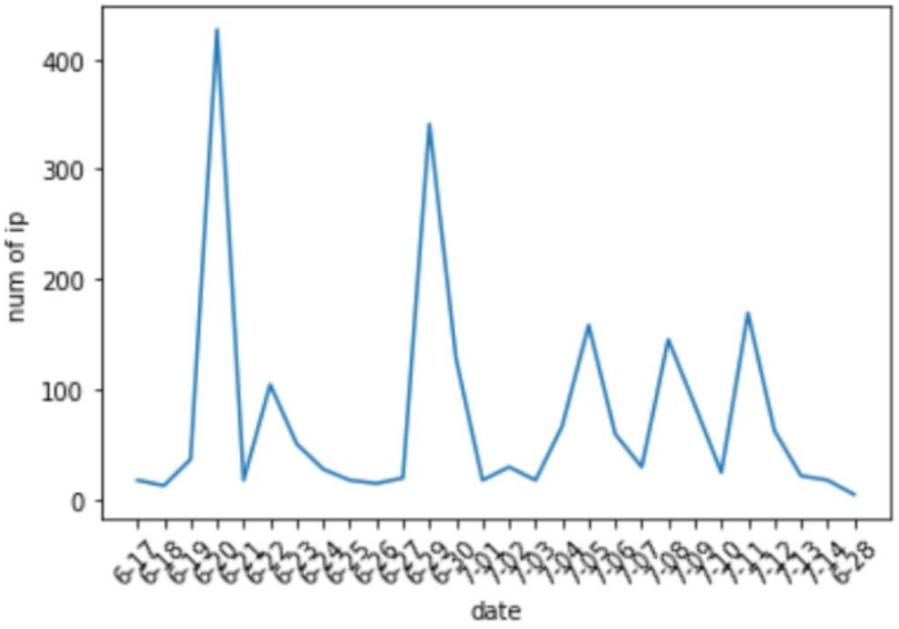


United Kingdom

Date in Different Regions



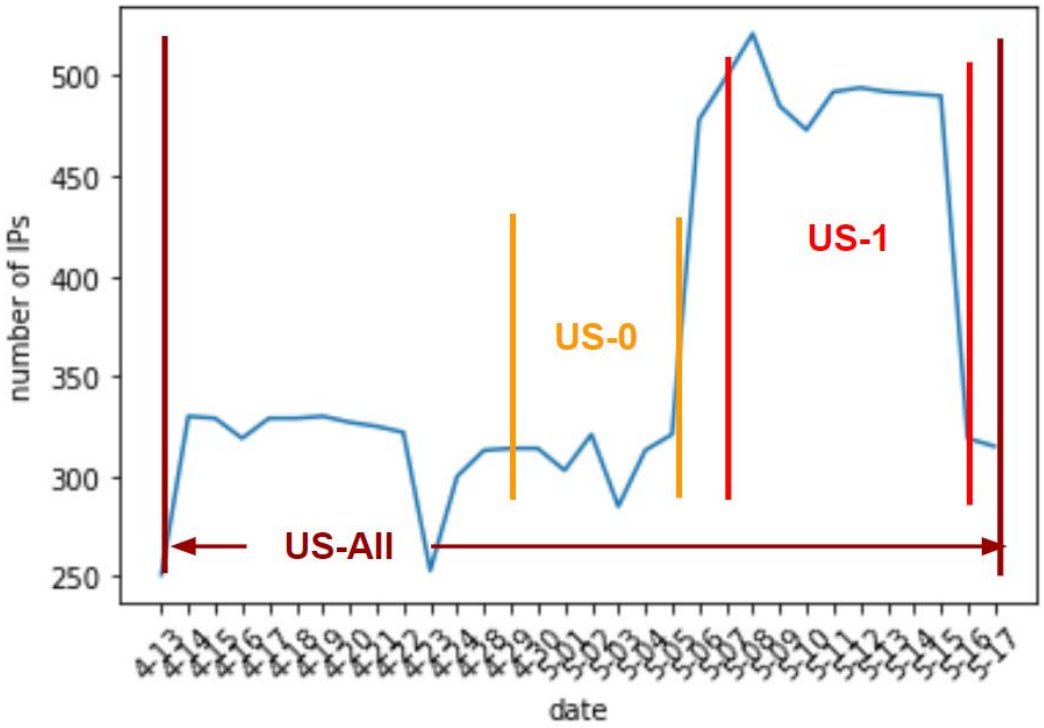
France



Netherlands

Continuous Data

- Definition: the crawler functioned 24 hours for more than 7 continuous days



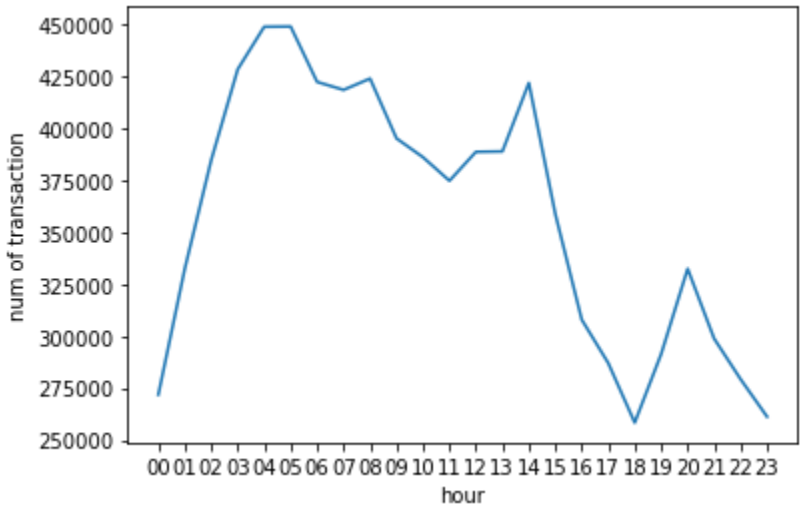
Continuous Data – US-1

- Period:
  - May 7 ~ May 16
- In US-1, the crawler collected the data for 24 hours for 10 continuous days.
- I use the data, US-1, to do the following experiments.

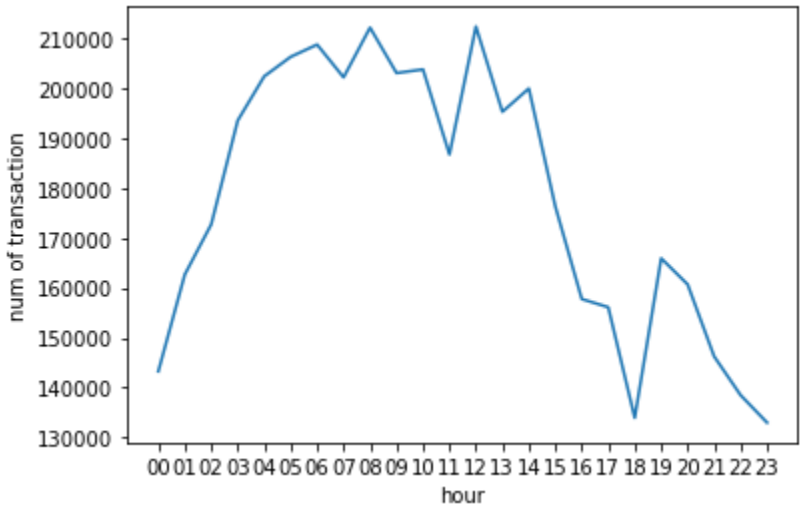
dates	working hours	transaction counts	mini hour count
May-6	12	51894	0
May-7	24	436552	13101
May-8	24	511106	13064
May-9	24	460176	9374
May-10	24	362727	10053
May-11	24	403603	11393
May-12	24	438546	7760
May-13	24	444512	8894
May-14	24	473601	13575
May-15	24	487068	10961
May-16	24	255964	5757
May-17	7	40057	0

Hour-Count Distribution

- Hour-Count Distribution in different periods



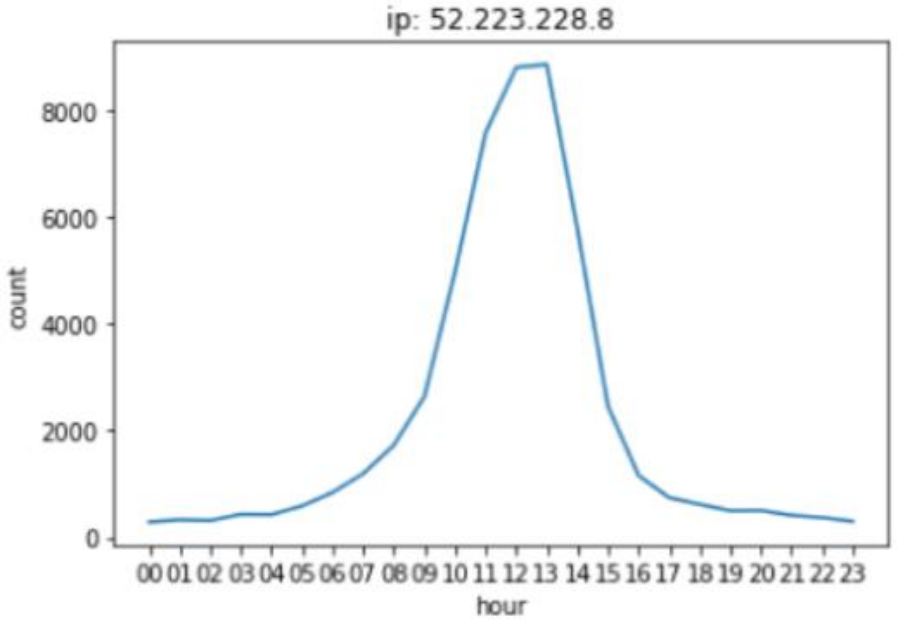
US-All

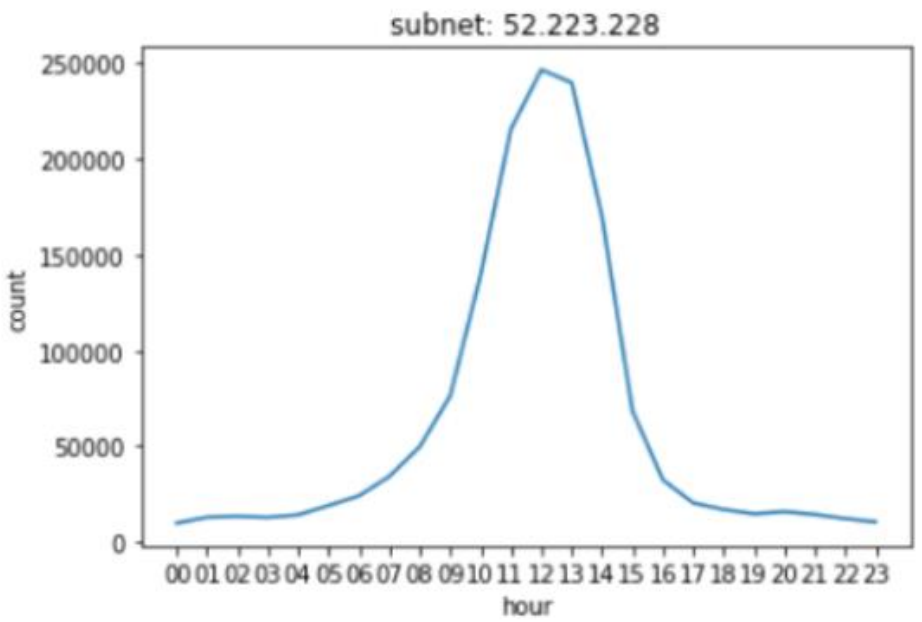


Us-1

Hour-Count Distribution

- Hour-Count Distribution in same subnet (24bit)



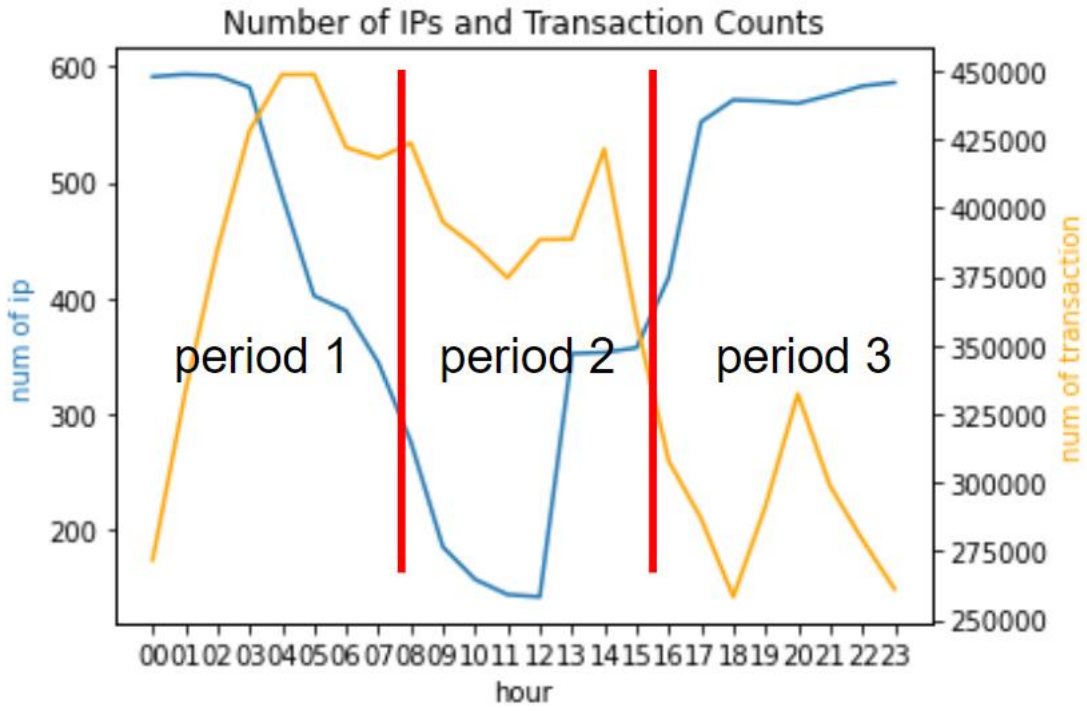


Clustering

- Attributes Selection
- Clustering Method
- Clustering Results Evaluation

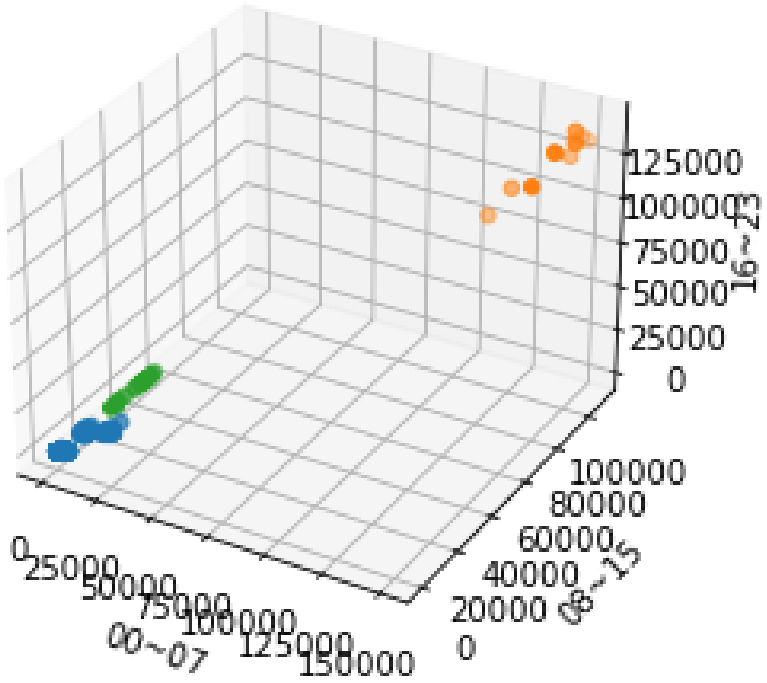
Attributes Selection

- Use transaction counts (orange line) in 3 periods to be attributes for clustering



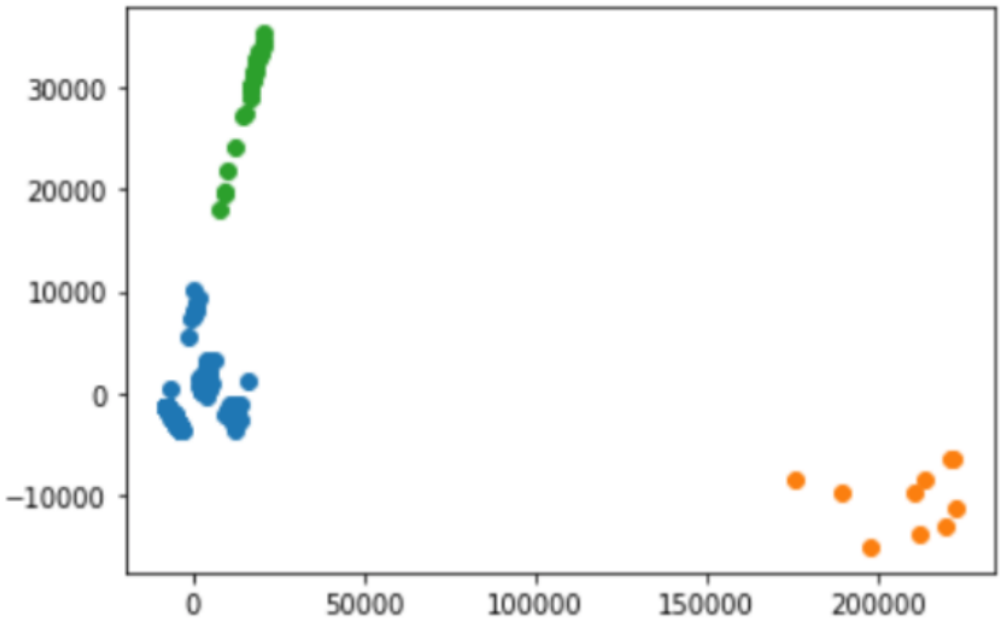
Clustering Method

- Clustering Algorithm:
  - Mini-Batch K-Means
- Number of Clusters: 3



Clustering Method

- Dimension Reduction:
  - PCA
- Dimension: 3 -> 2



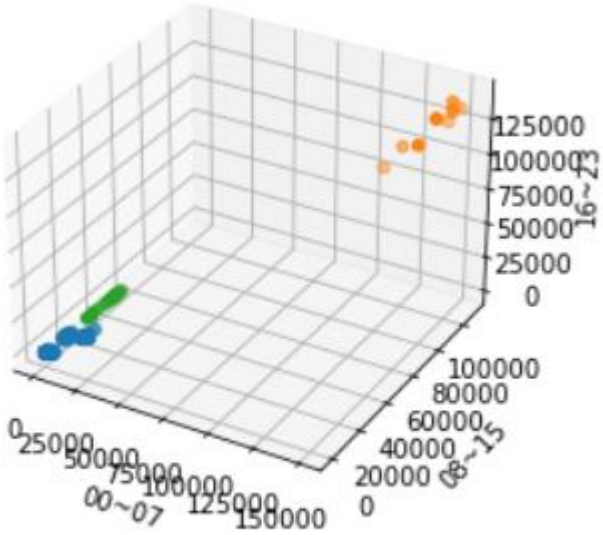
Clustering Results Evaluation

- S_Dbw score
  - Internal evaluation: no ground truth
  - Consider both inter-cluster density (Dens_bw) and intra-cluster variance (Scat)
  - Lower value represents better performance

$$S\_Dbw\ (c) = Scat(c) + Dens\_bw(c)$$

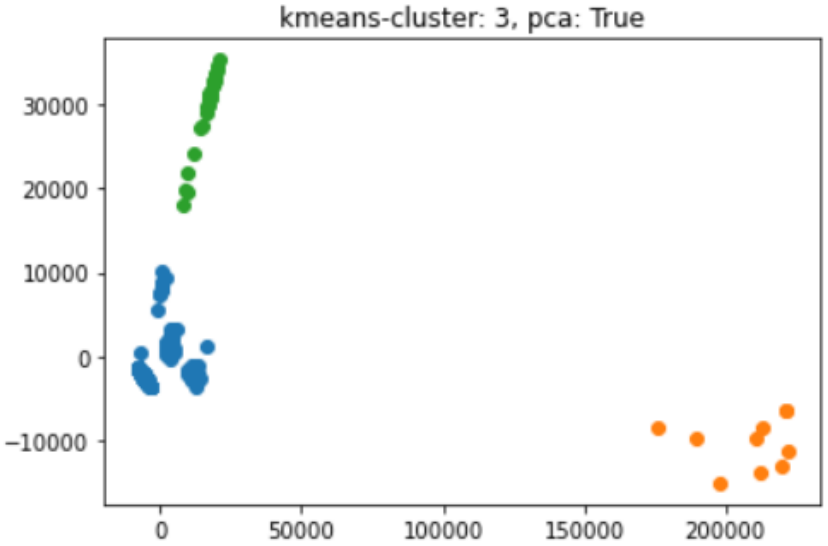
Clustering Results Evaluation

- Result with and without dimension reduction
- Data: US-All



S_Dbw: 0.36733253592159115

Without PCA

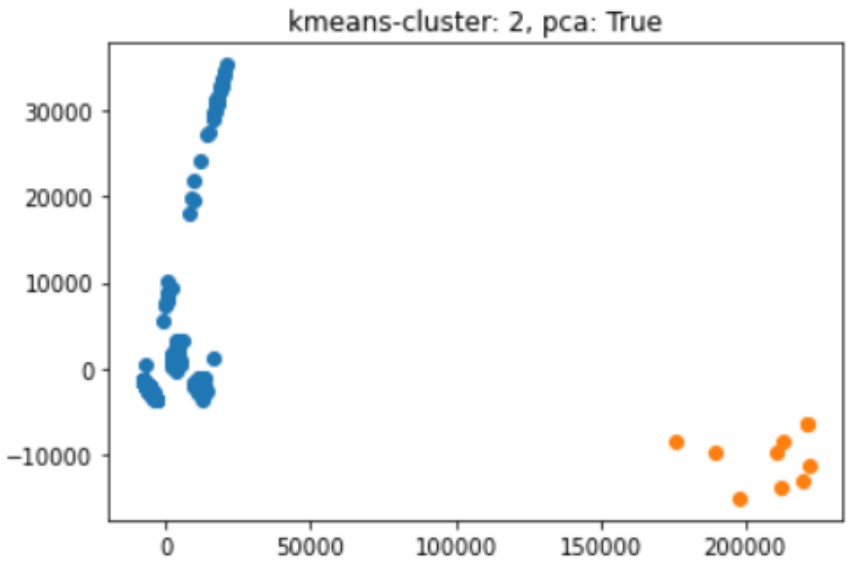


s_dbw score: 0.3601130751852925

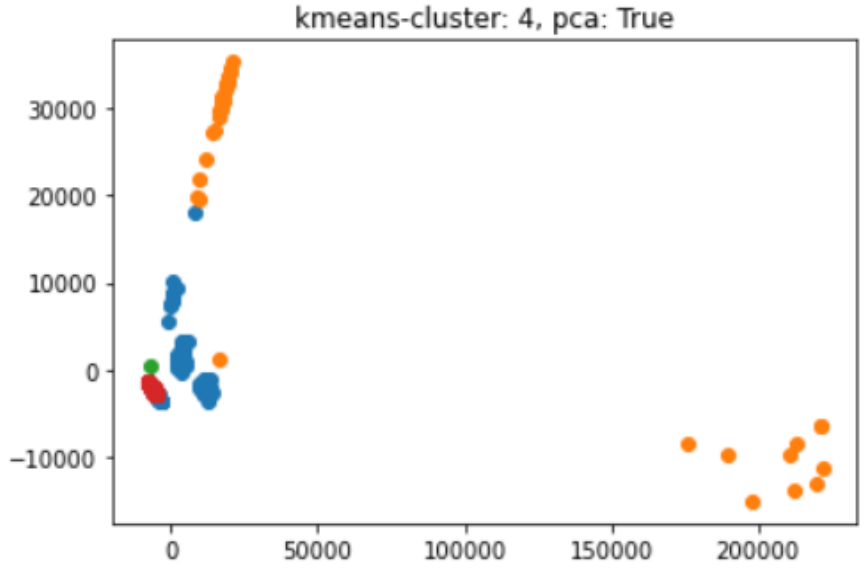
With PCA

Clustering Results Evaluation

- K-means with different numbers of clusters
- Data: US-All



s_dbw score: 0.5267170413404887



s_dbw score: 1.0620020470782143

Clustering Results Evaluation

- Data: US-All
- Number of Clusters = 3
- Number of Periods = 3
- Example:

n_periods: 3, slide_hour:2

Transaction count in

[02~09, 10~17, 18~01]

n_periods	slide_hour	S_Dbw score
3	0	0.3601
3	1	0.3549
3	2	0.3504
3	3	0.4792
3	4	0.5270
3	5	0.4227
3	6	0.3818
3	7	0.3635

Clustering Results Evaluation

- Number of Periods = 4

n_periods	slide_hour	S_Dbw score
4	0	0.3946
4	1	0.3584
4	2	0.3591
4	3	0.3598
4	4	0.4601
4	5	0.4480

- Number of Periods = 6

n_periods	slide_hour	S_Dbw score
6	0	0.3586
6	1	0.4549
6	2	0.4653
6	3	0.4695

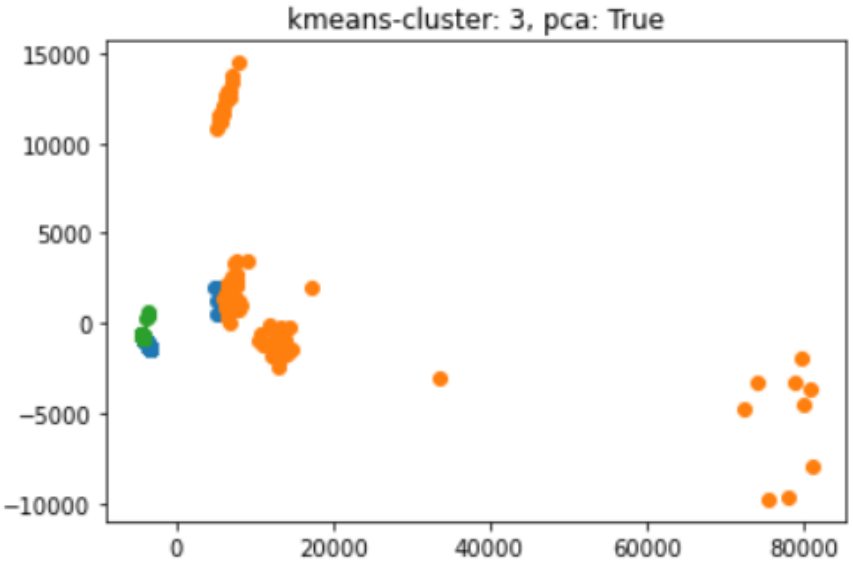
Clustering Results Evaluation

- After I check all the clustering results, I find that the results with S_Dbw score less than 0.37 are actually the same as each other.

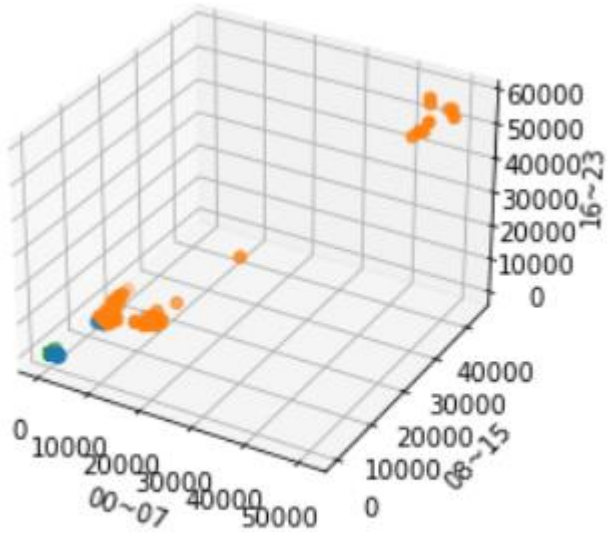
They have the same servers' distribution of 3 clusters.
I choose one result to apply in Preliminaries

Clustering Results Evaluation

- K-means with US-1 data



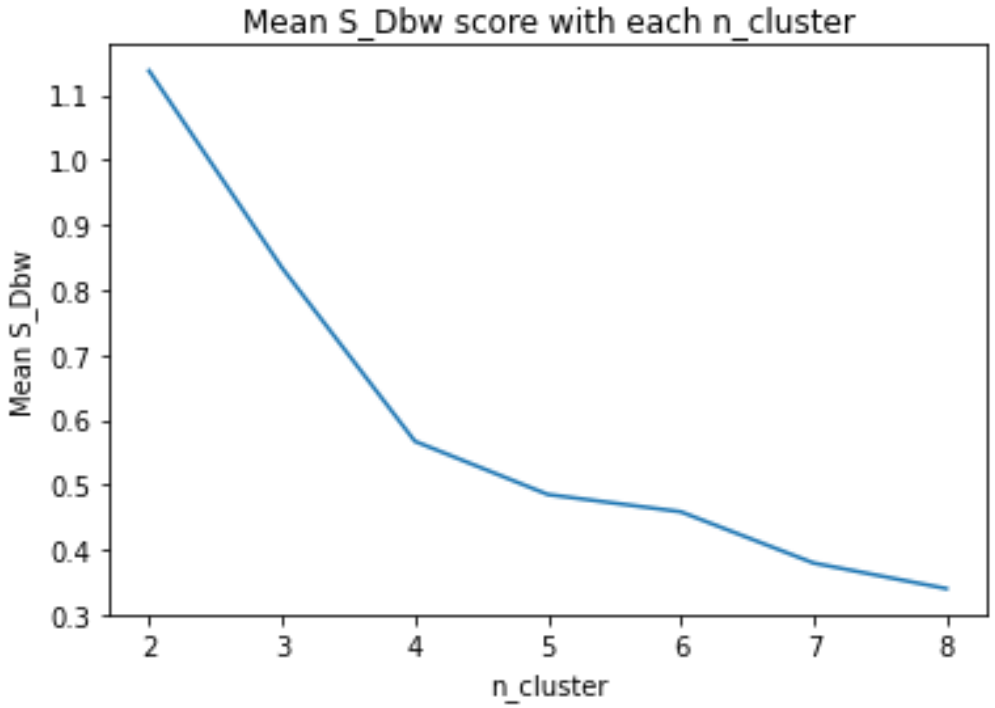
S_dbw: 1.1467056156867095



S_dbw: 1.1467056156867095

Clustering Results Evaluation

- Run Mini-Batch K-Means 10 times for each n_cluster



Clustering Results Evaluation

- Alternative Metric
  - For each cluster
    - Std/Avg: Standard Deviation/Average Number
      Std and Avg are calculated by sample number on each date
    - Cluster Size: the number of IPs in one cluster
  - For each clustering result (contains n cluster)
    - Mean Std/Avg: mean Std/Avg of all the cluster in one clustering result
    - Min Cluster Size: the minimum cluster size in one clustering result

CRM Model result

- Preliminaries
- CJS Model Analysis – K-Means
- CJS Model Analysis – Random Clustering

Preliminaries

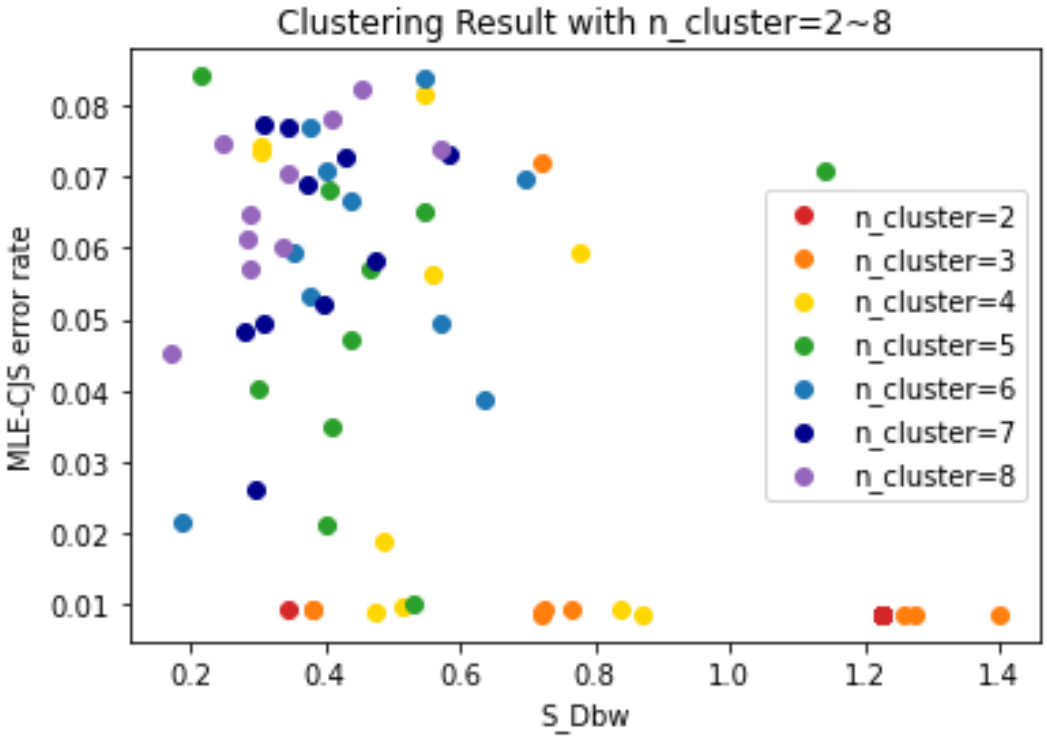
- Error rate = |baseline – estimation| /baseline * 100%
- Baseline: number of servers observed in one day (24 hours)

Preliminaries

- Error Rate in the CJS model on US-1 (sample at 12am)
  - No clustering: 0.82%
  - K-Means (3 clusters): 1.92%
    * Use the clustering result with US-All

CJS Model Analysis – K-Means

- The scatter plot for S_Dbw score and Error Rate of MLE-CJS (US-1)



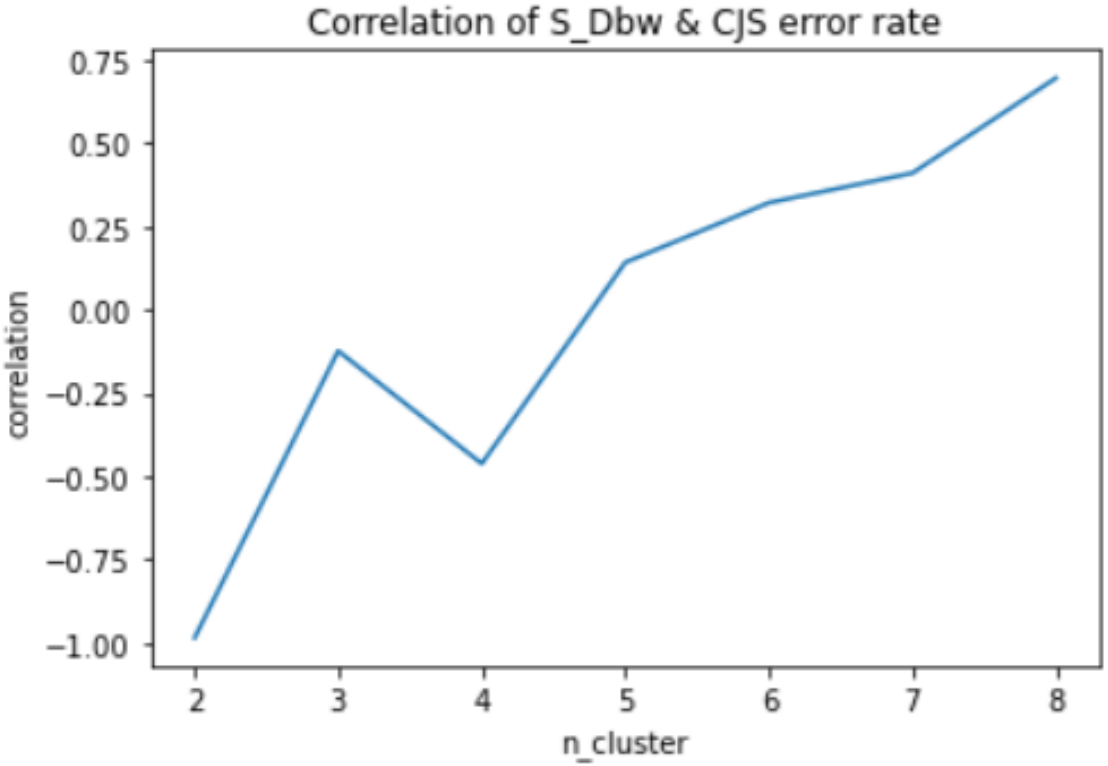
CJS Model Analysis – K-Means

- Correlation Matrix

	n_cluster	S_Dbw	err	std
n_cluster	1.000000	-0.714362	0.719124	0.678303
S_Dbw	-0.714362	1.000000	-0.571475	-0.544296
err	0.719124	-0.571475	1.000000	0.994293
std	0.678303	-0.544296	0.994293	1.000000

CJS Model Analysis – K-Means

- Correlation of S_Dbw & Error Rate with different n_cluster



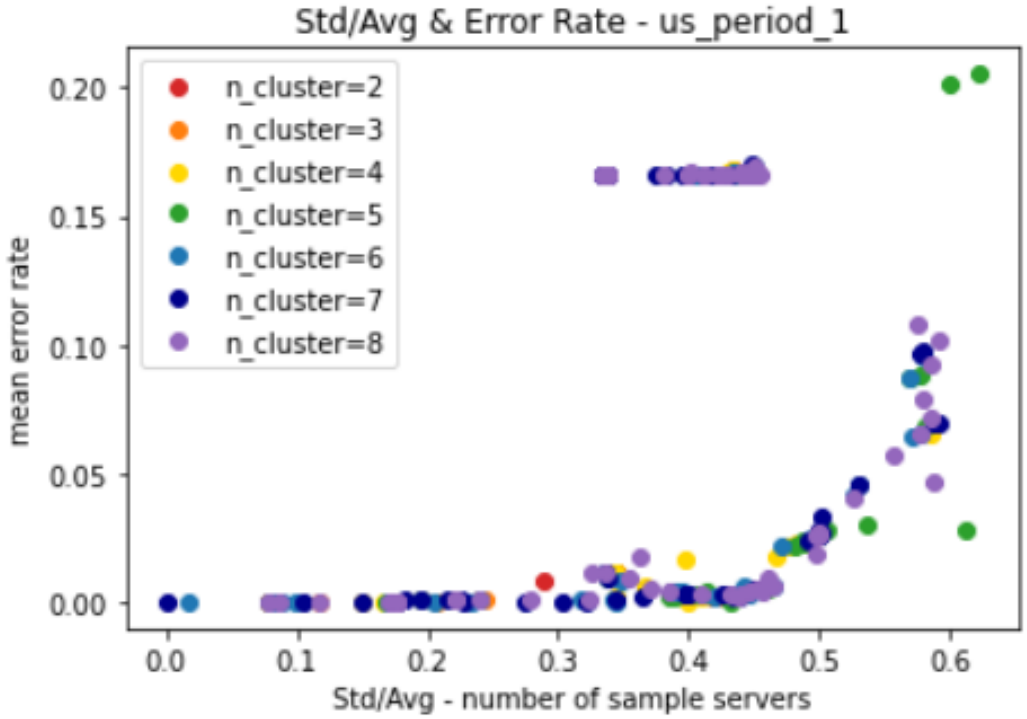
CJS Model Analysis – K-Means

- Mean of S_Dbw and Error Rate

n_cluster	S_Dbw mean	MLE-CJS error rate mean (%)
2	1.1379	0.8709
3	0.8343	1.5248
4	0.5669	4.0055
5	0.4849	4.9931
6	0.4583	5.9059
7	0.3792	6.0445
8	0.3398	6.6806

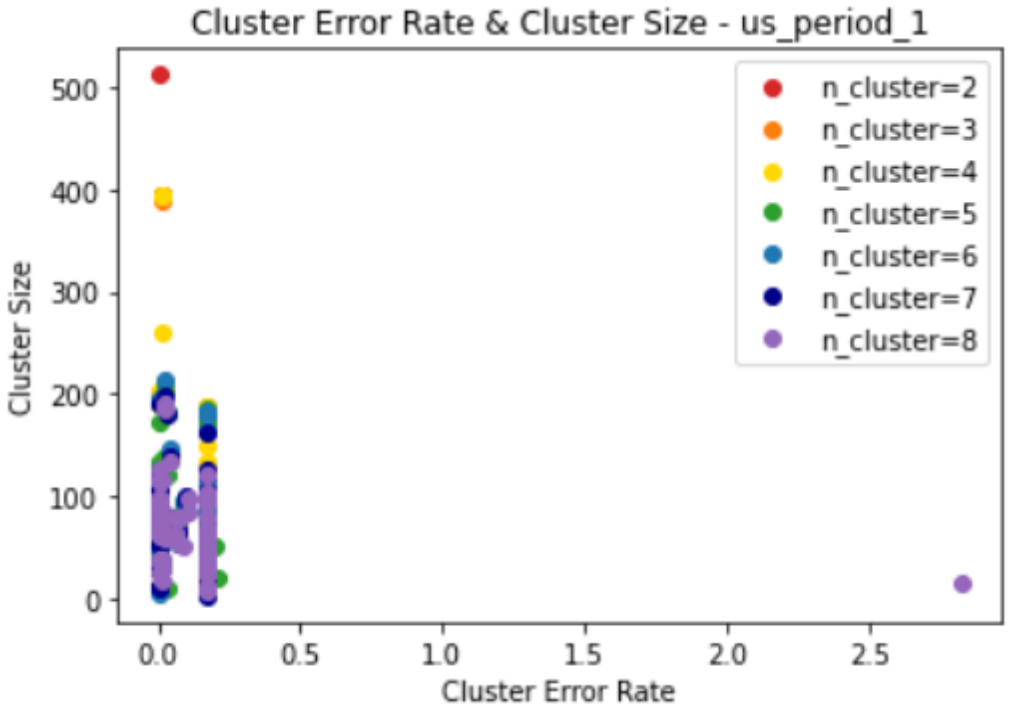
CJS Model Analysis – K-Means

- The scatter plot of Std/Avg & Cluster Error Rate (for each cluster)



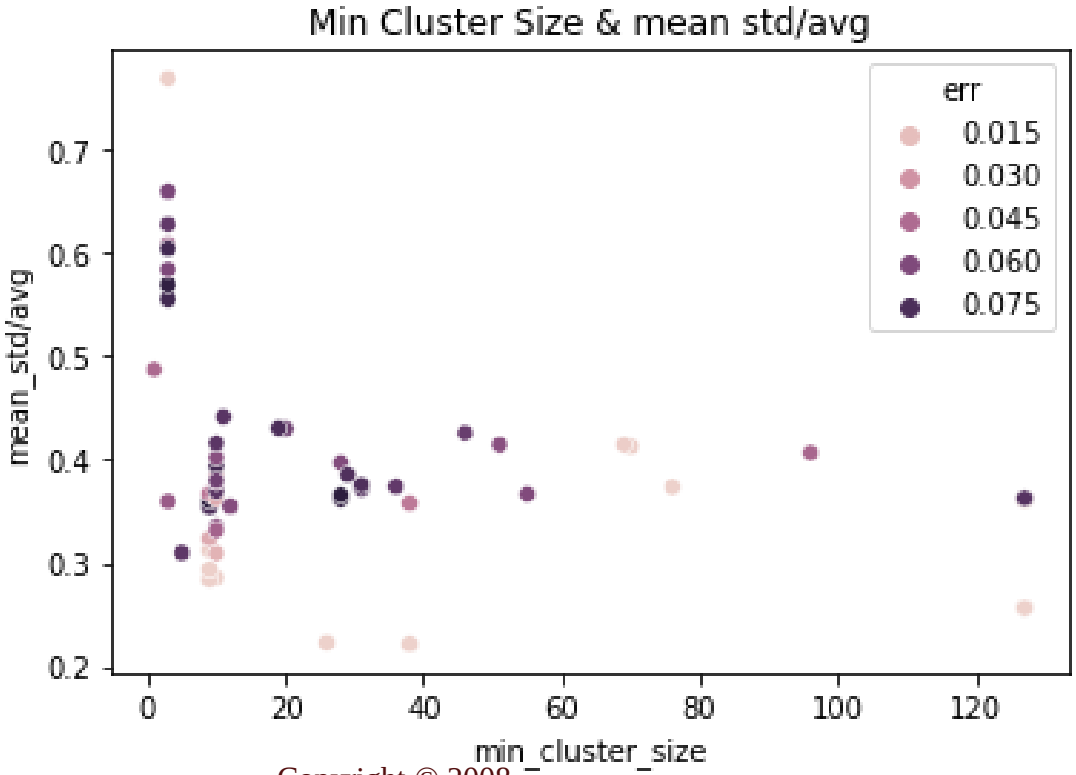
CJS Model Analysis – K-Means

- The scatter plot of Cluster Size & Cluster Error Rate (for each cluster)



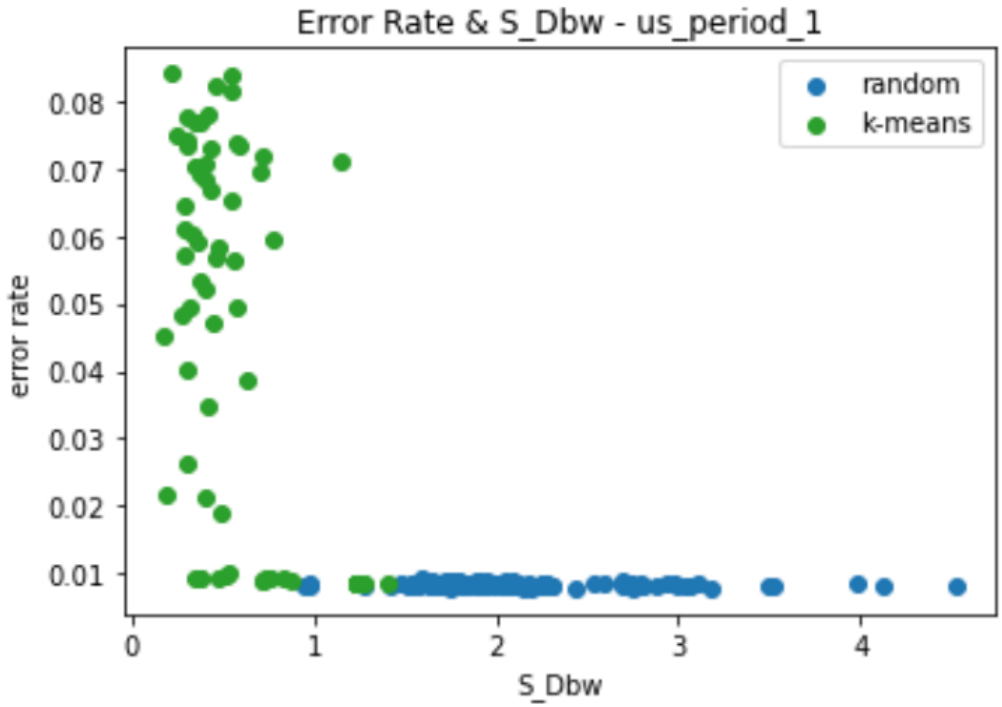
CJS Model Analysis – K-Means

- The scatter plot of Min Cluster Size & Mean Std/Avg & Error Rate (for each clustering result)



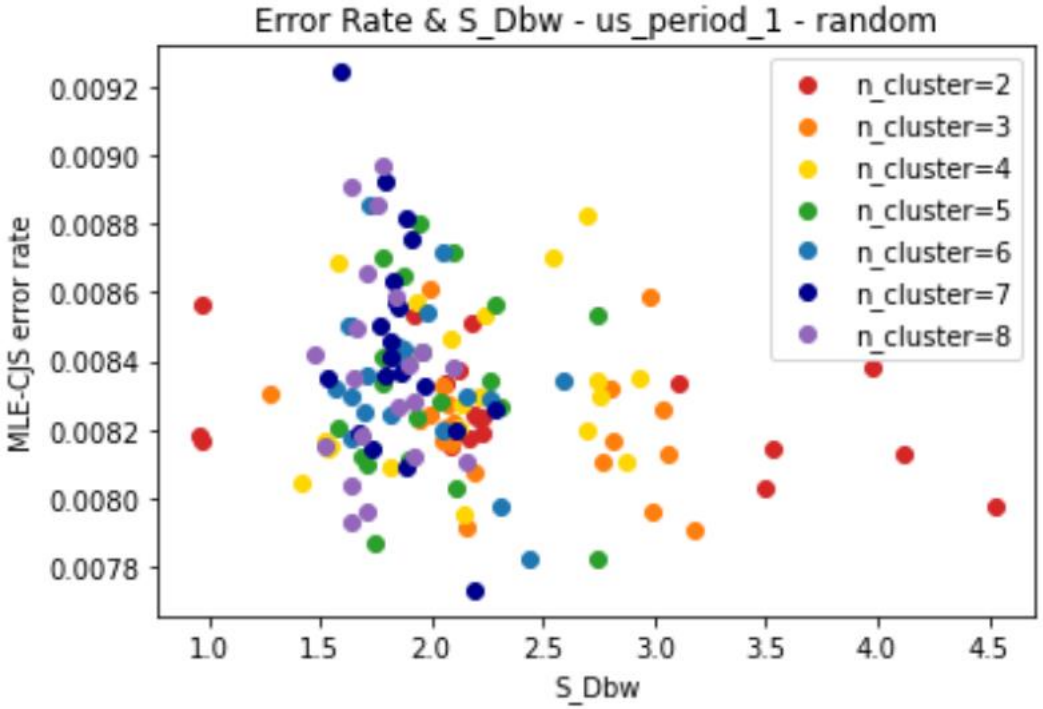
CJS Model Analysis – Random Clustering

- To be the control group, I randomly divide IPs into n clusters with similar size 20 times for each n_cluster



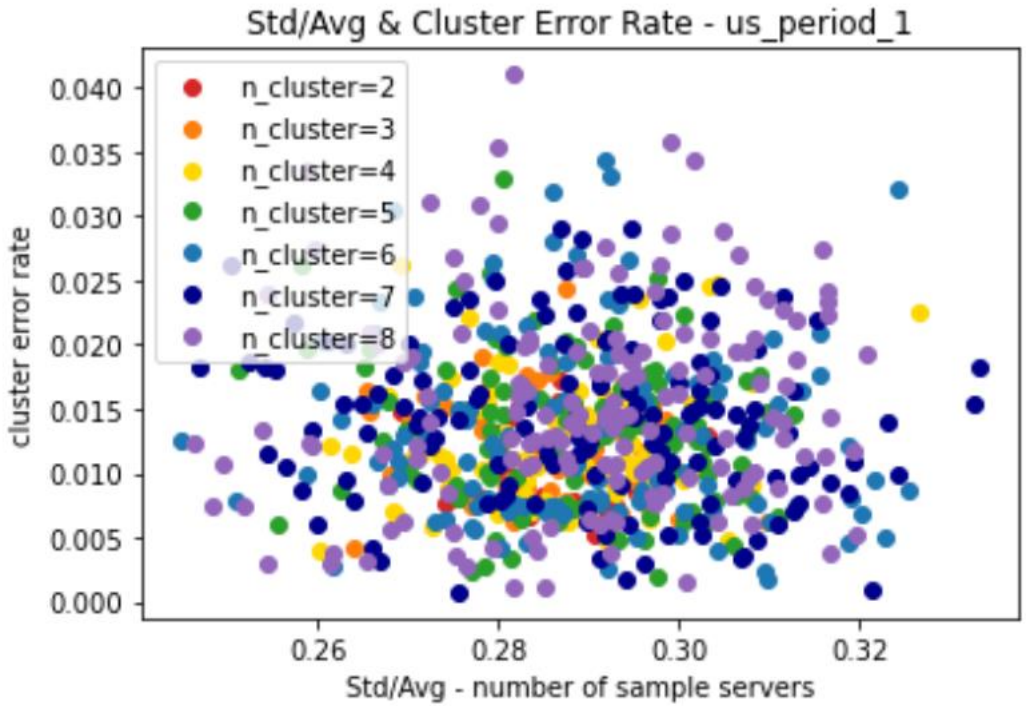
CJS Model Analysis – Random Clustering

- The scatter plot of Error Rate & S_Dbw (for each clustering result)



CJS Model Analysis – Random Clustering

- The scatter plot of Std/Avg & Cluster Error Rate (for each cluster)

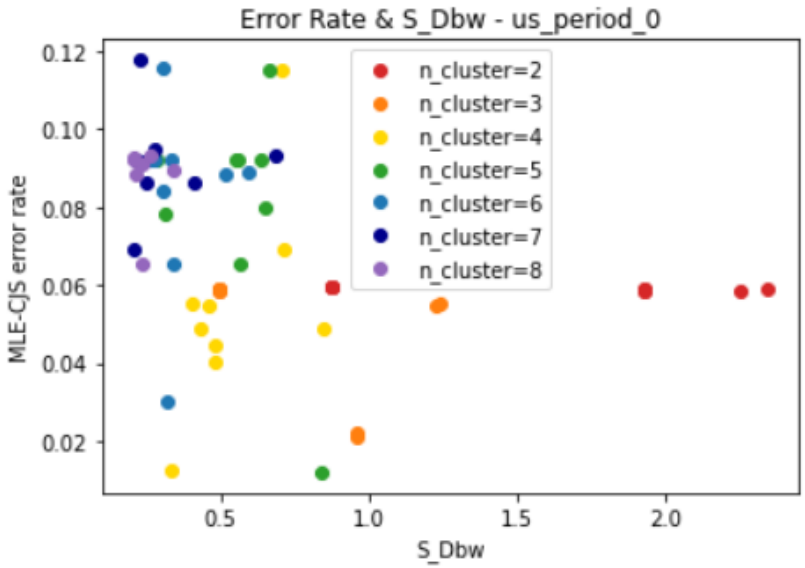


CJS with Different Data

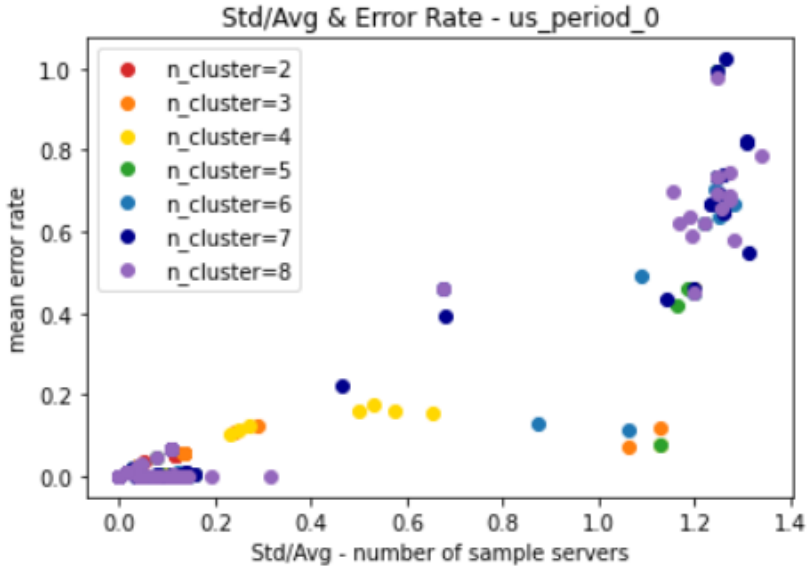
- US - o
- The UK
- France
- The Netherlands
- Germany

US-0 (April 29 to May 05)

- K-Means Clustering



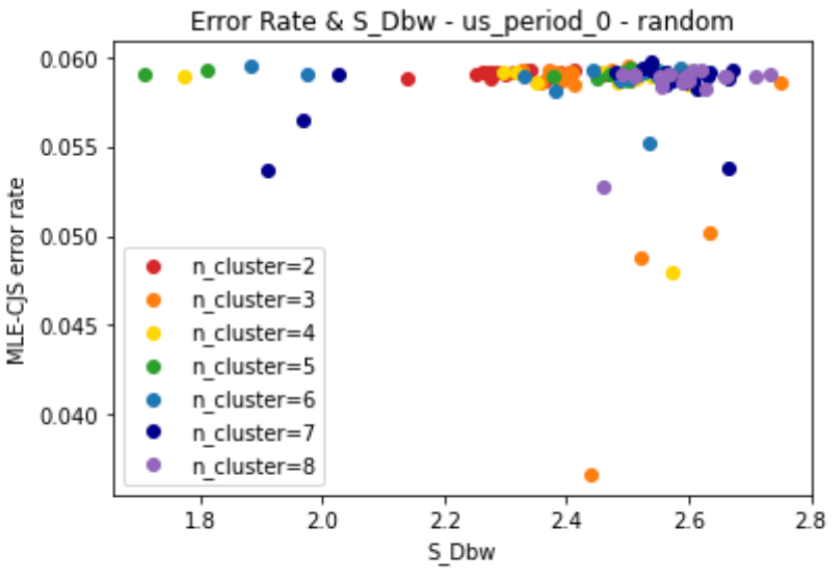
S_Dbw & Error Rate



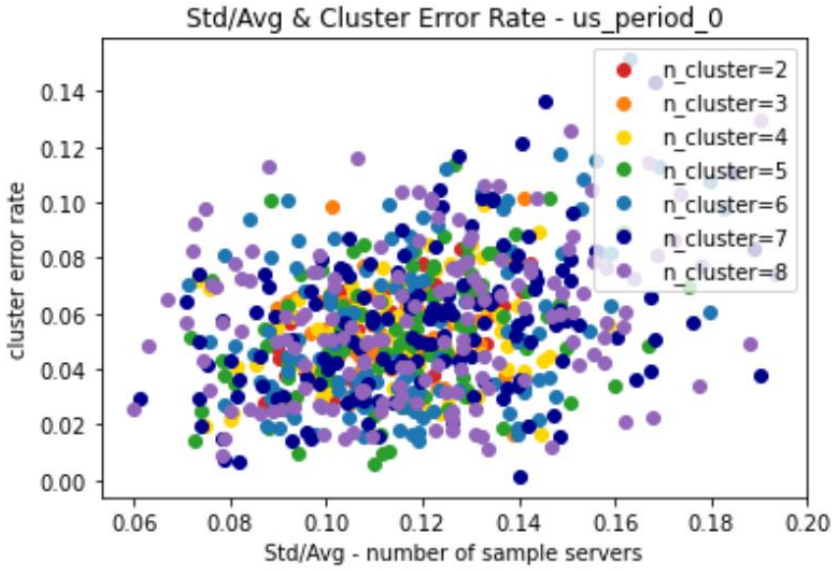
Std/Avg & Cluster Error Rate

US-0 (April 29 to May 05)

- Random Clustering



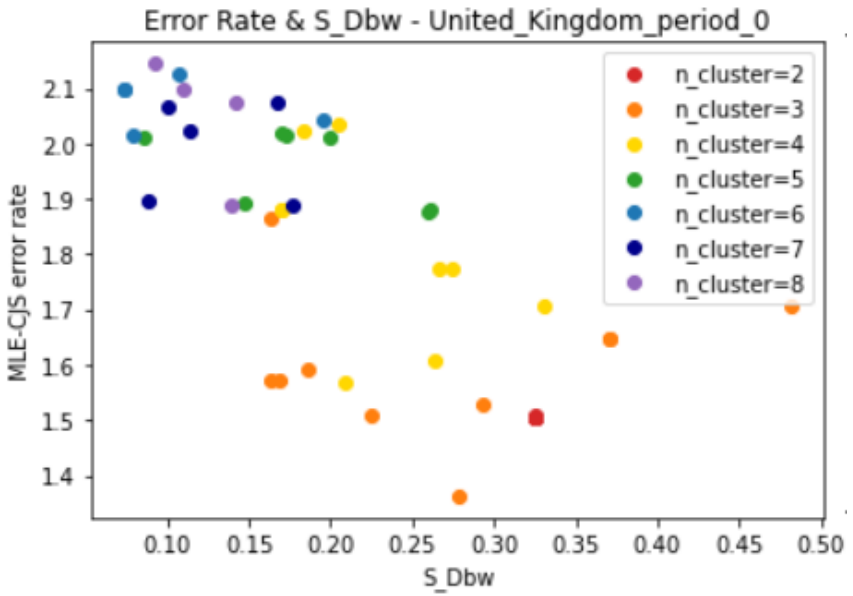
S_Dbw & Error Rate



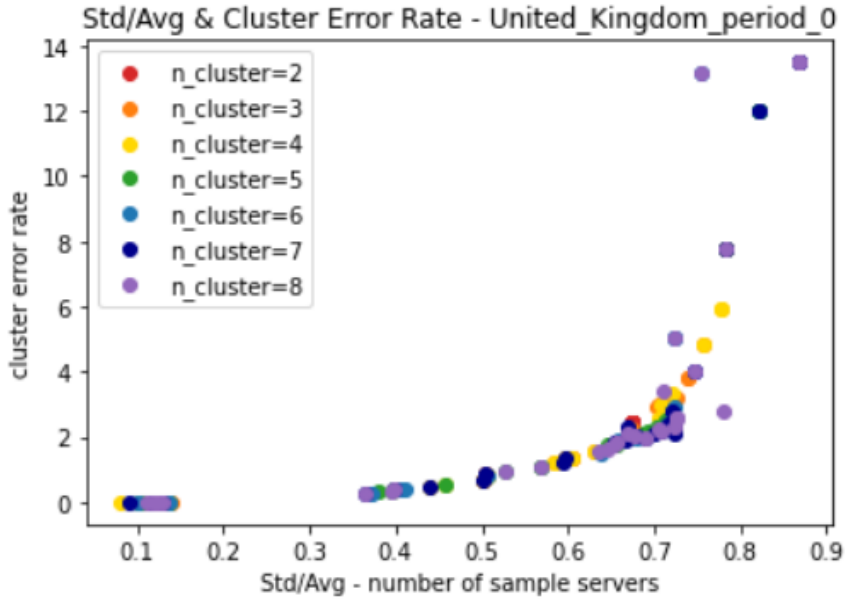
Std/Avg & Cluster Error Rate

UK-0 (April 29 to May 05)

- K-Means Clustering



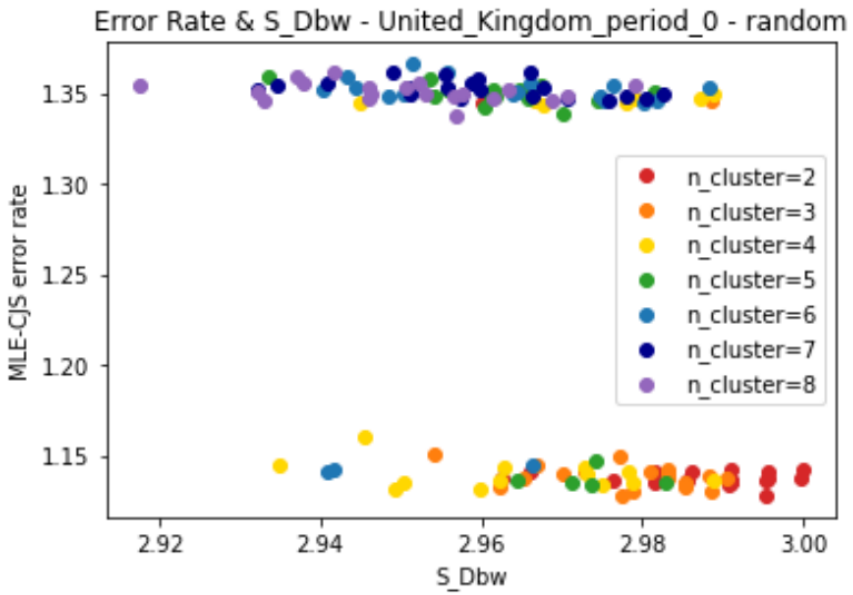
S_Dbw & Error Rate



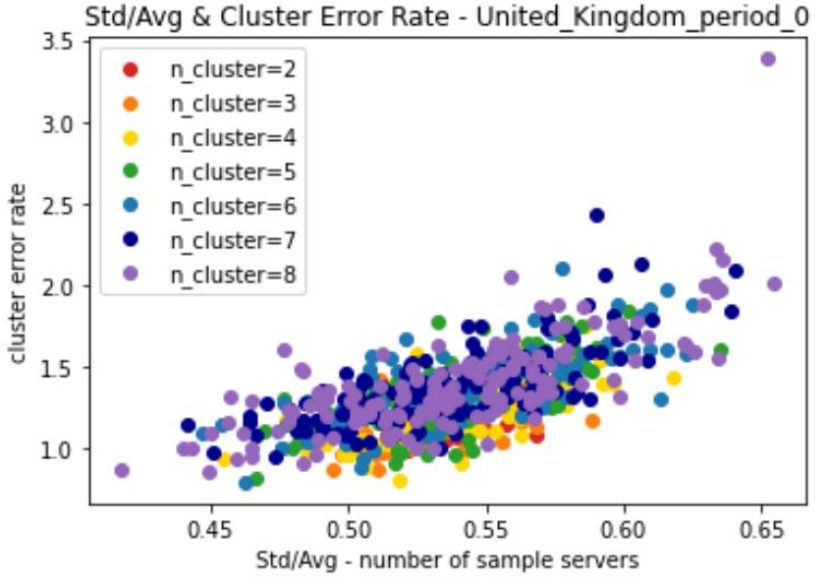
Std/Avg & Cluster Error Rate

UK-0 (April 29 to May 05)

- Random Clustering



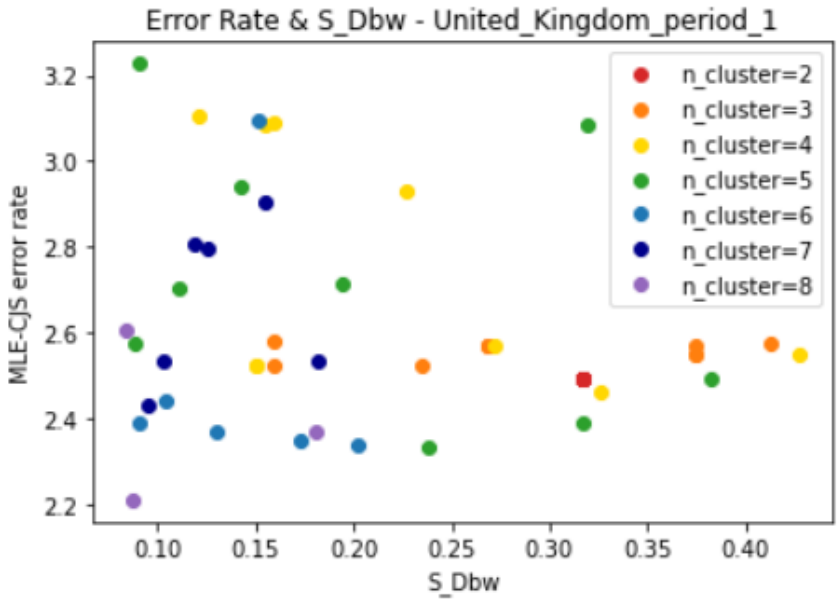
S_Dbw & Error Rate



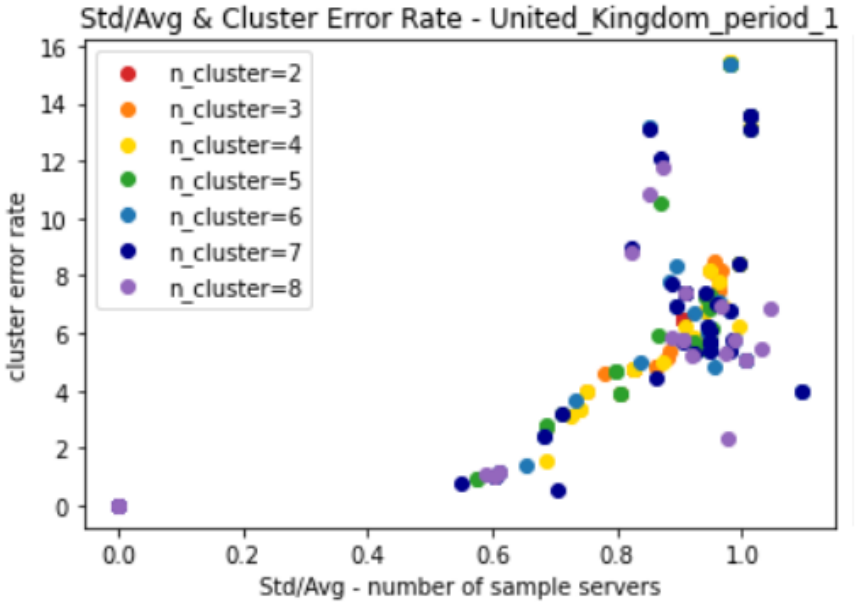
Std/Avg & Cluster Error Rate

UK-1 (May 07 to May 15)

- K-Means Clustering



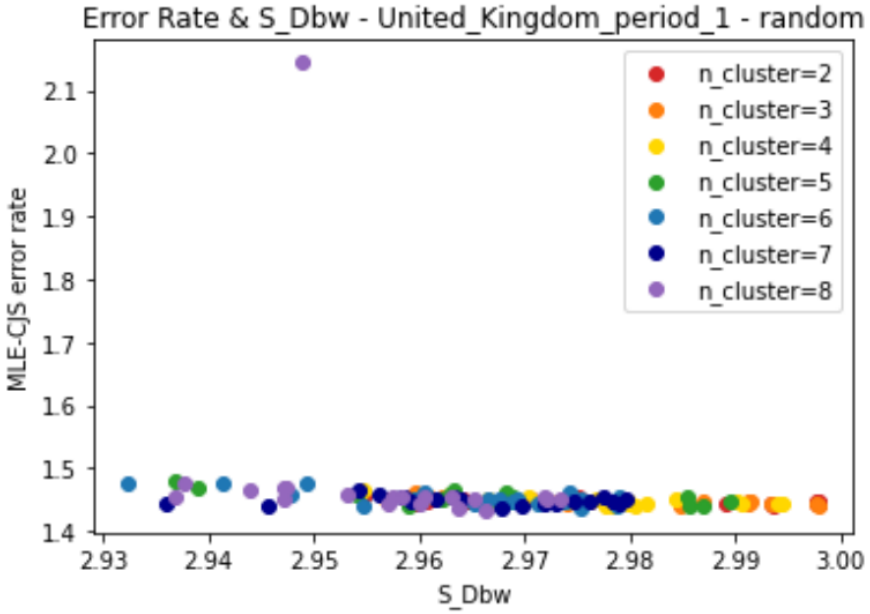
S_Dbw & Error Rate



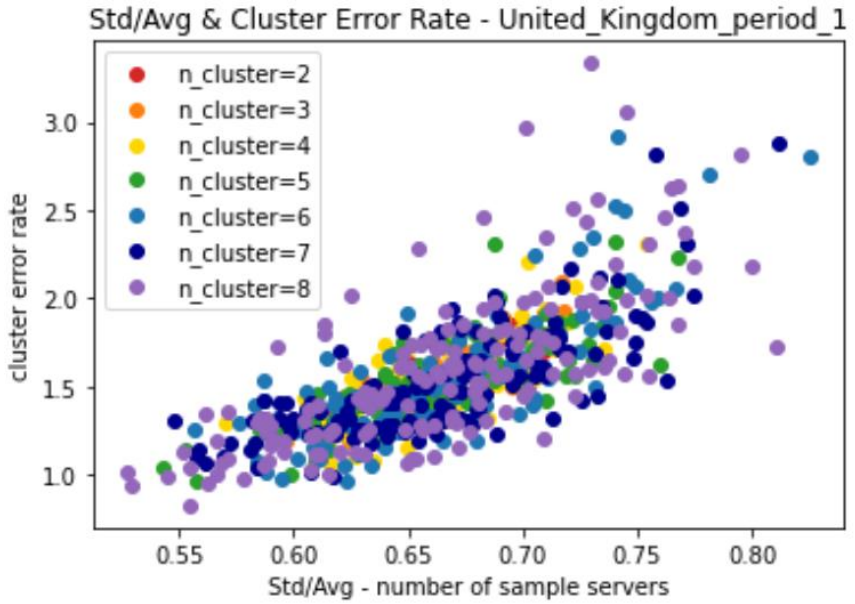
Std/Avg & Cluster Error Rate

UK-1 (May 07 to May 15)

- Random Clustering



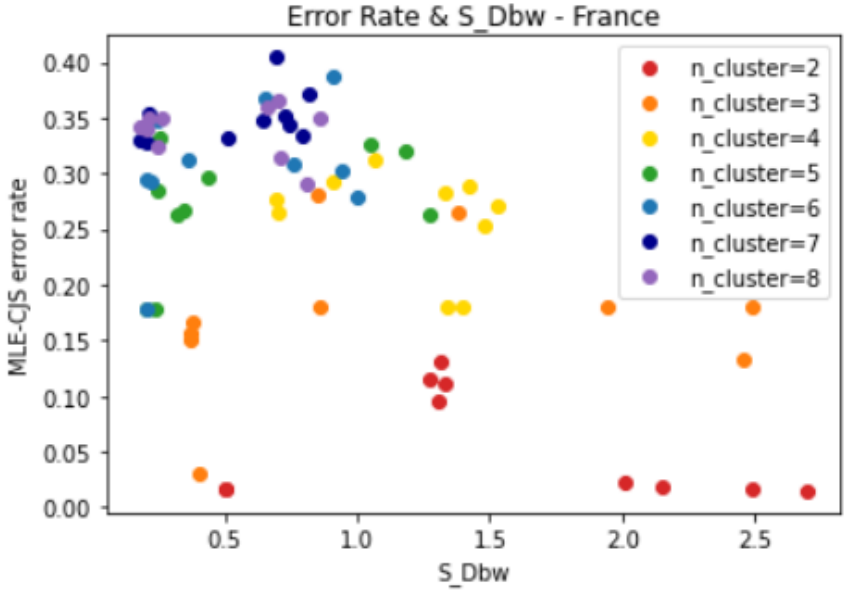
S_Dbw & Error Rate



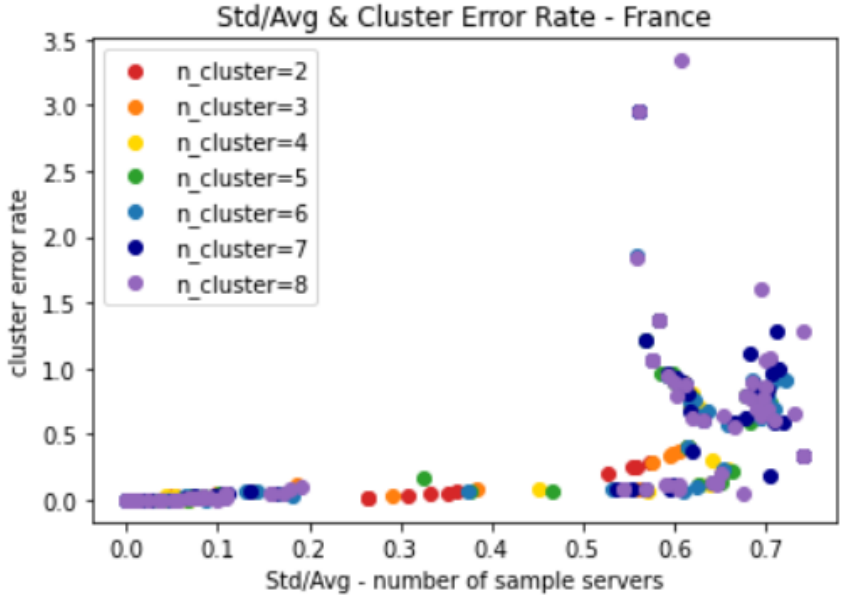
Std/Avg & Cluster Error Rate

France (April 29 to May 05)

- K-Means Clustering



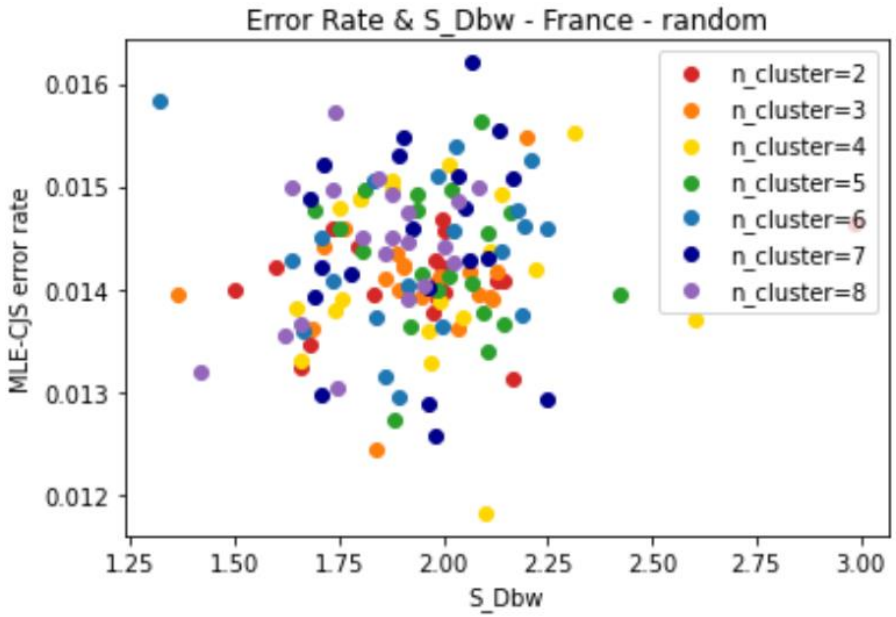
S_Dbw & Error Rate



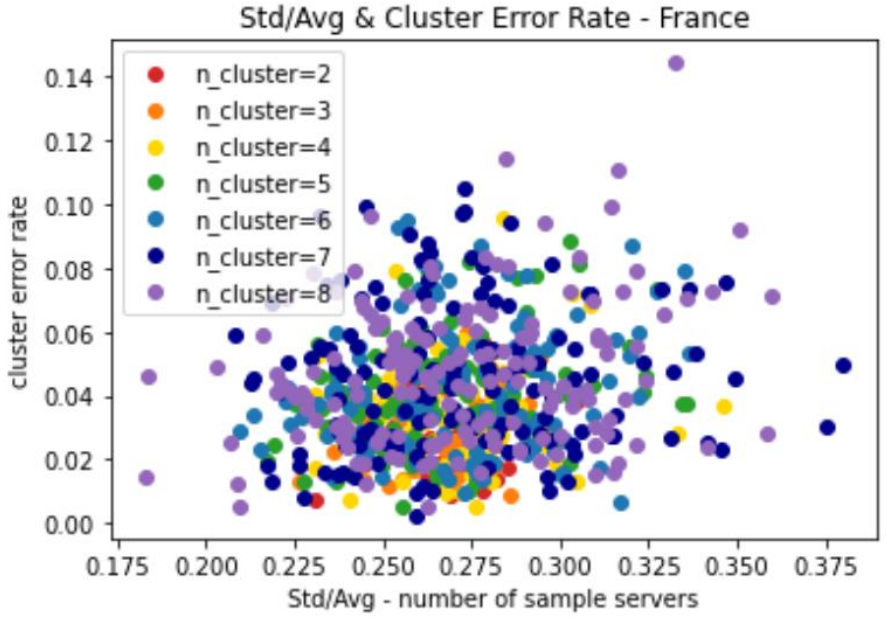
Std/Avg & Cluster Error Rate

France (April 29 to May 05)

- Random Clustering



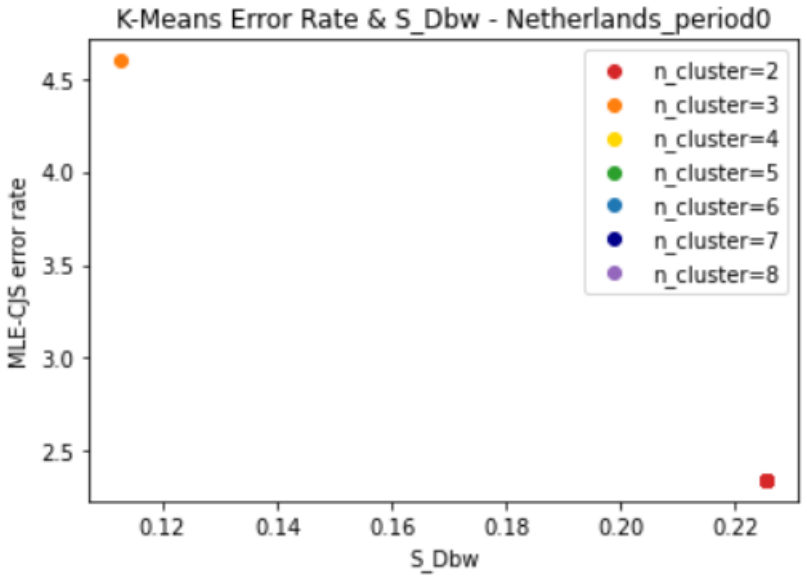
S_Dbw & Error Rate



Std/Avg & Cluster Error Rate

Netherlands-0 (June 18 to June 27)

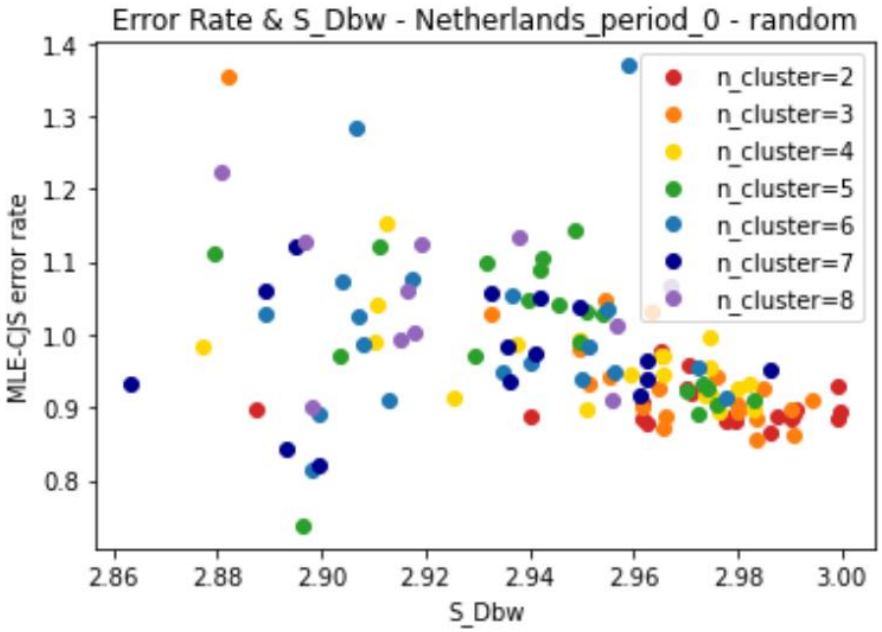
- K-Means Clustering



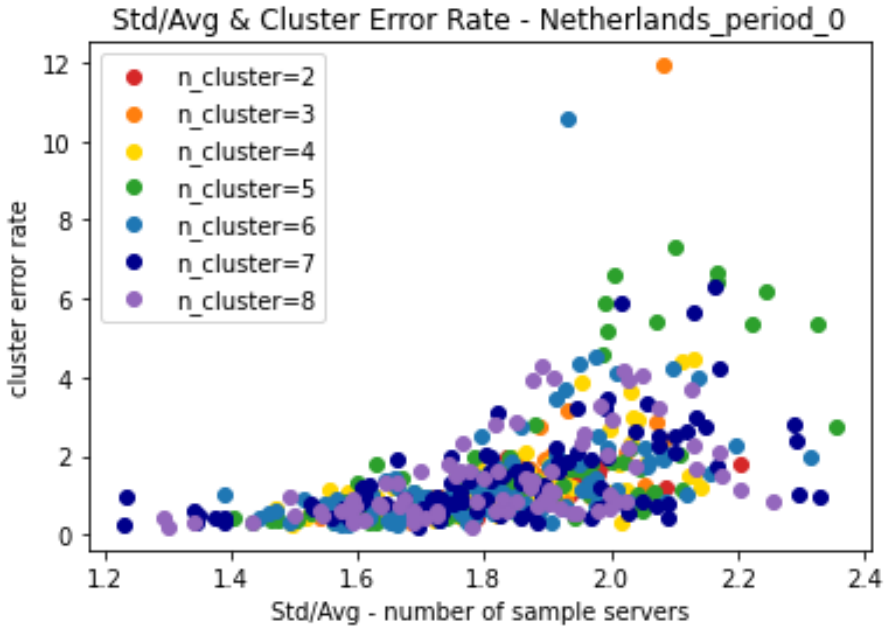
S_Dbw & Error Rate

Netherlands-0 (June 18 to June 27)

- Random Clustering



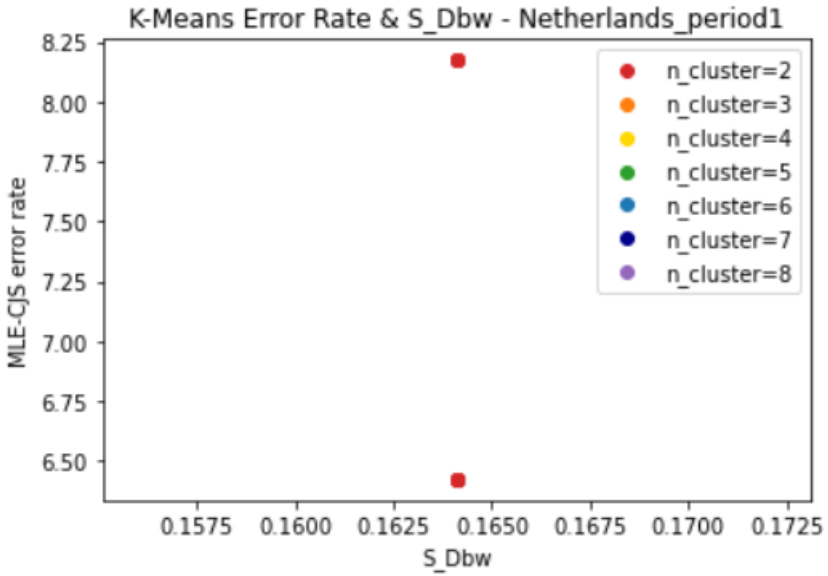
S_Dbw & Error Rate



Std/Avg & Cluster Error Rate

Netherlands-1 (June 30 to July 13)

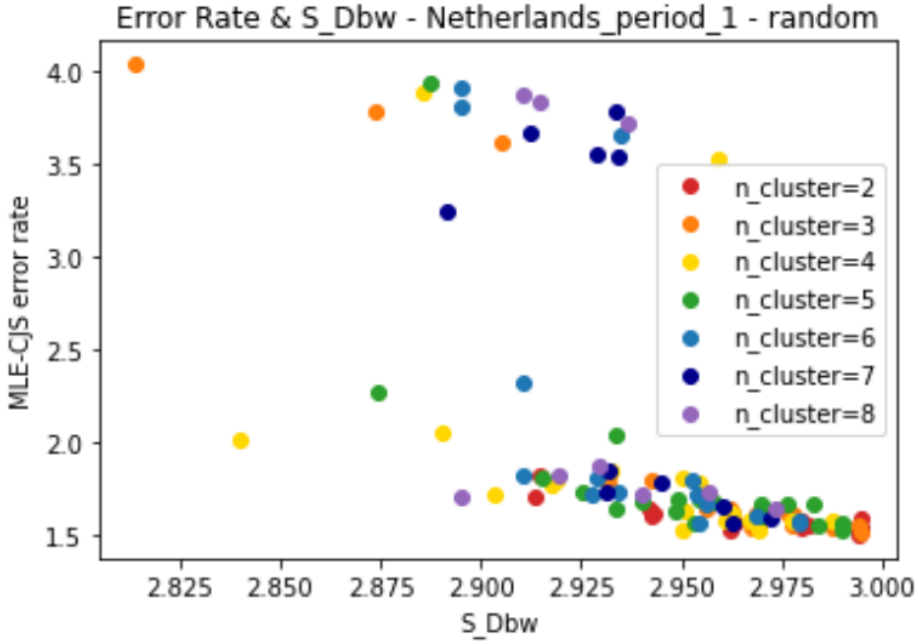
- K-Means Clustering



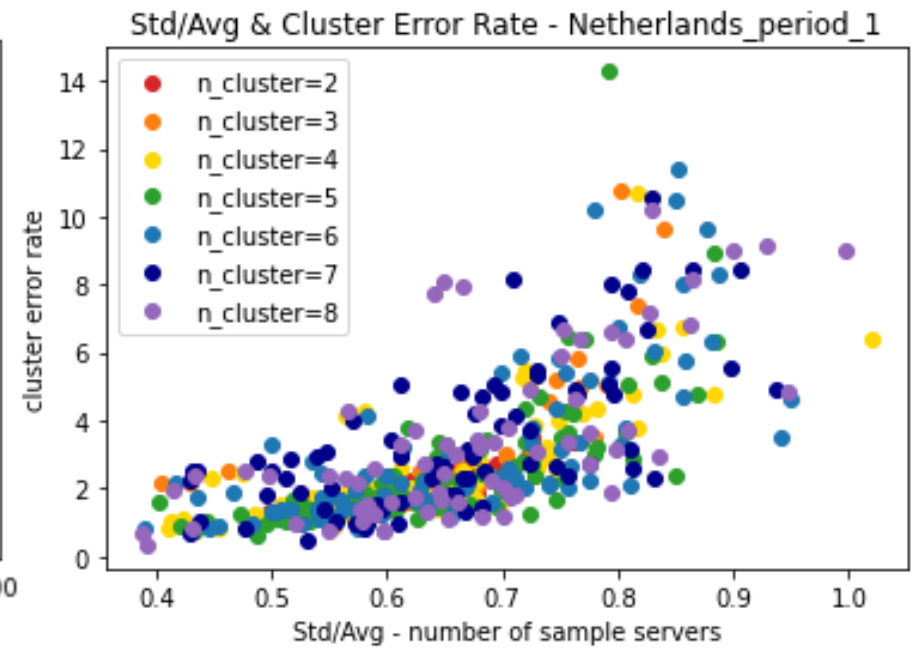
S_Dbw & Error Rate

Netherlands-1 (June 30 to July 13)

- Random Clustering



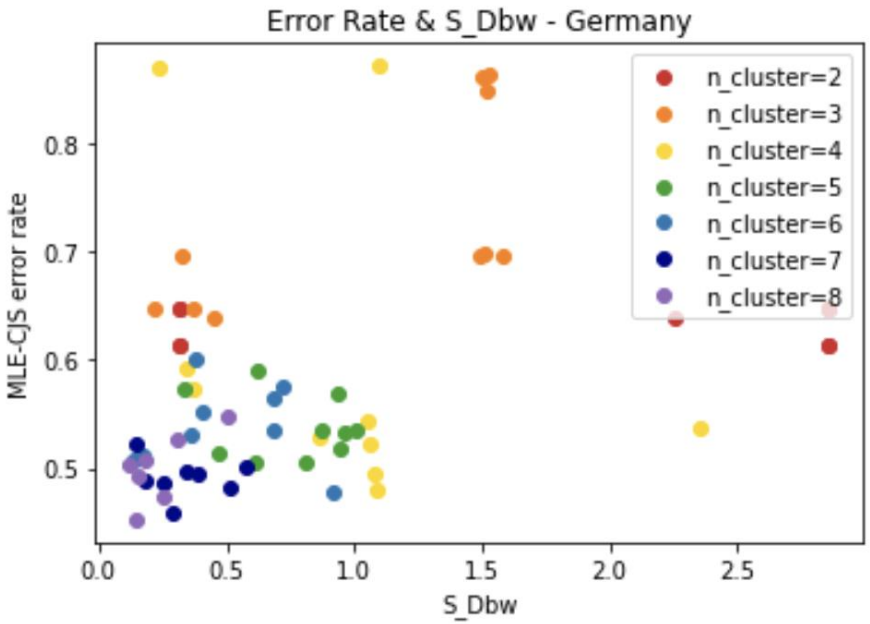
S_Dbw & Error Rate



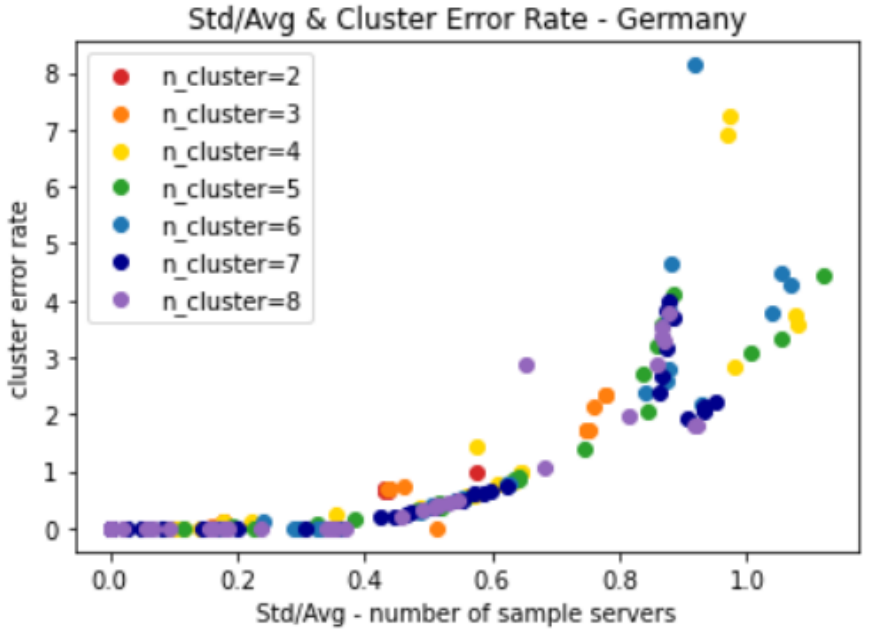
Std/Avg & Cluster Error Rate

Germany (June 04 to June 14)

- K-Means Clustering



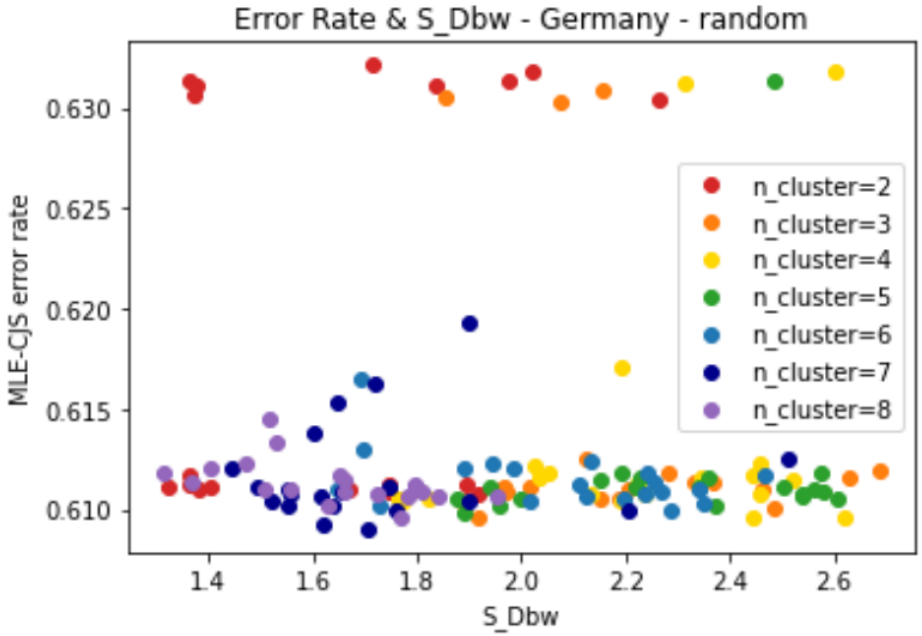
S_Dbw & Error Rate



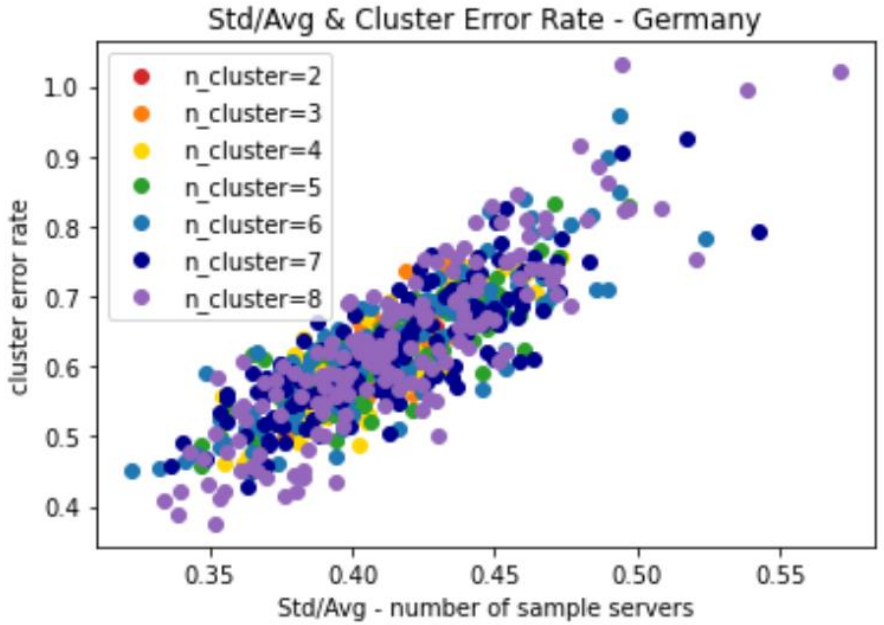
Std/Avg & Cluster Error Rate

Germany (June 04 to June 14)

- Random Clustering



S_Dbw & Error Rate



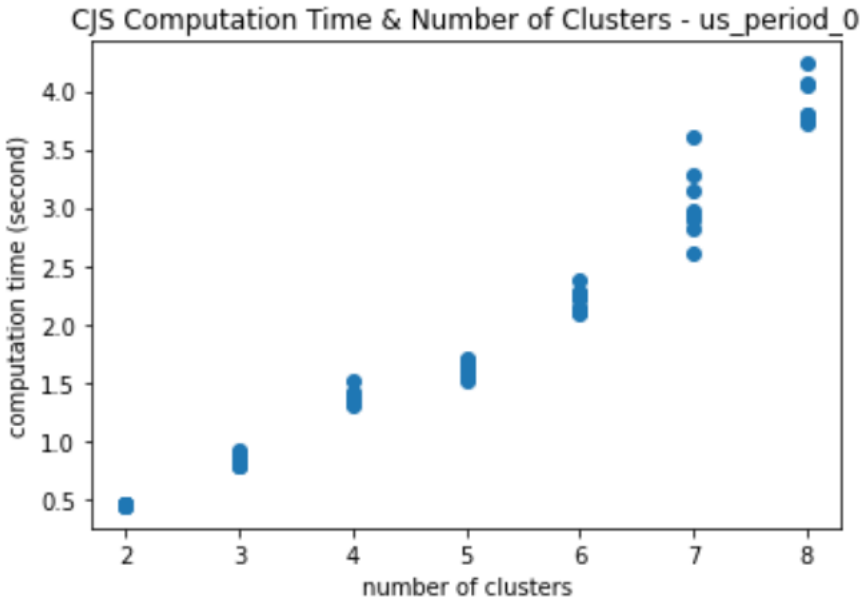
Std/Avg & Cluster Error Rate

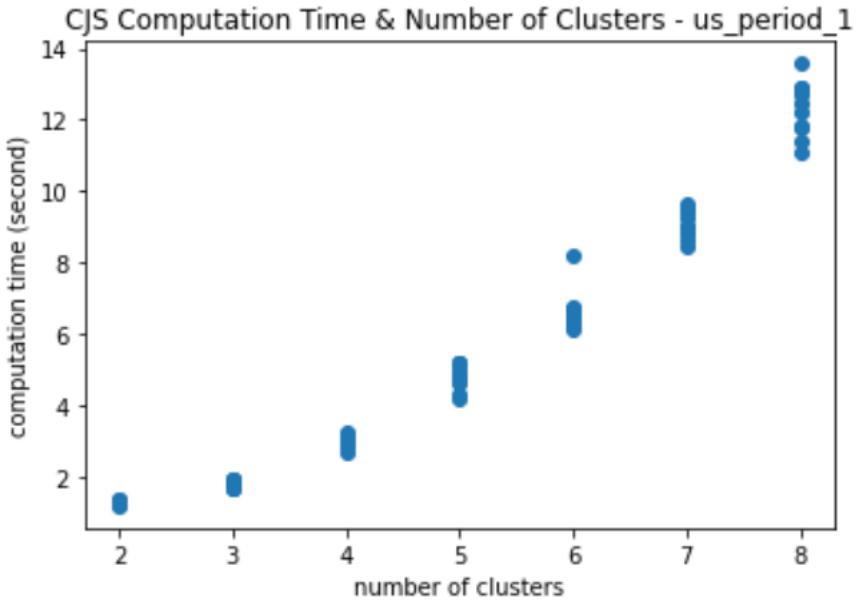
Computation Time of CJS

- Computation Time in the US
- Computation Time in Different Region

Computation Time in the US

- Computation Time & Number of Clusters





Computation Time in Different Region

- The Correlations between n_cluster & computation time are all close to 1.
- As the number of clusters gets larger, the computation time of the CJS will tend to have a linear growth.

data	period length (day)	correlation	max time (second)
US-0	7	0.9755	4.24
US-1	10	0.9708	13.58
UK-0	7	0.9688	6.15
UK-1	9	0.9562	15.23
France	7	0.9584	6.97
Netherlands-0	10	0.9751	8.82
Netherlands-1	14	0.9472	41.21
Germany	11	0.9585	45.15

Conclusion

- The goal of this research is to build the CRM model with heterogeneity for CDN servers population estimation.
- CDN servers from same subnet (24bit) has similar hour and transaction count distribution.
- S_Dbw is not a good metric for the CRM model.
  Instead, Std/Avg could be a good metric for predicting the performance of the CRM model.
- Computation time is close to have a linear relationship with number of clusters.

Q&A