A CLUSTER-BASED TRIP PREDICTION GRAPH NEURAL NETWORK FOR BIKE SHARING SYSTEMS

By

Bárbara Tavares, Cláudia Soares & Manuel Marques

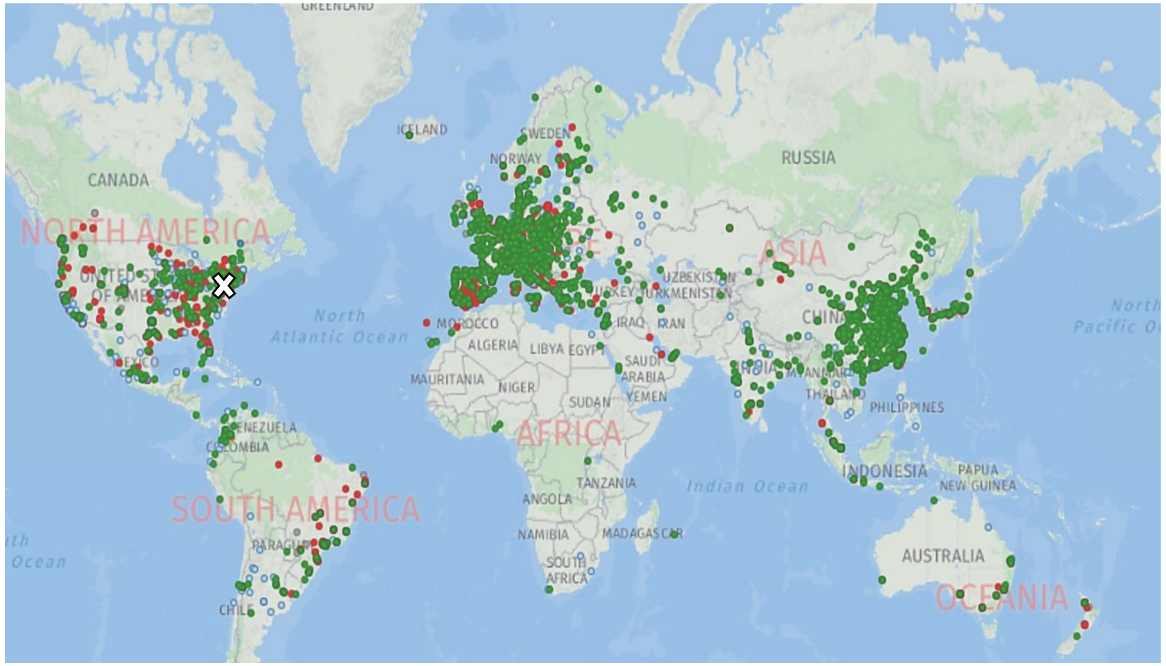
BSS: A MOBILITY SYSTEM EMERGING WORLDWIDE

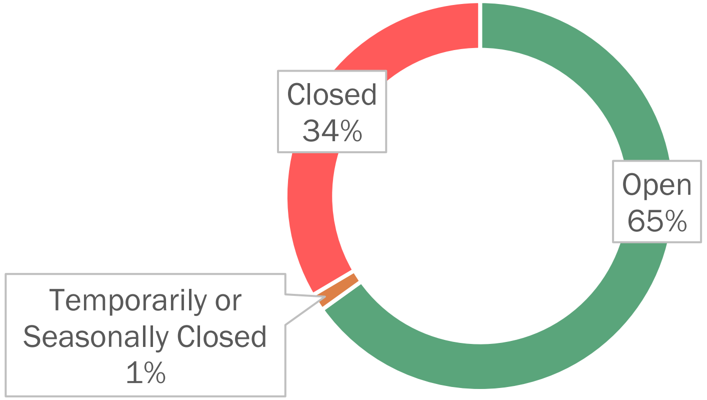
Usage

- Rent a bike on the departure local/station
- Ride to the destination local/station
- Return the bike to the destination local/dock

SBSS

DBSS





THE CURRENT BSS PROBLEM & PRESENTATION OUTLINE

Bike Imbalance

Analogous Transition Patterns →

↓

Customer Loss In The Long Term

System Operation ⇄ **System Prediction**

Rebalancing Strategies

We Propose: An Accurate Prediction of Bike Traffic

citi bike

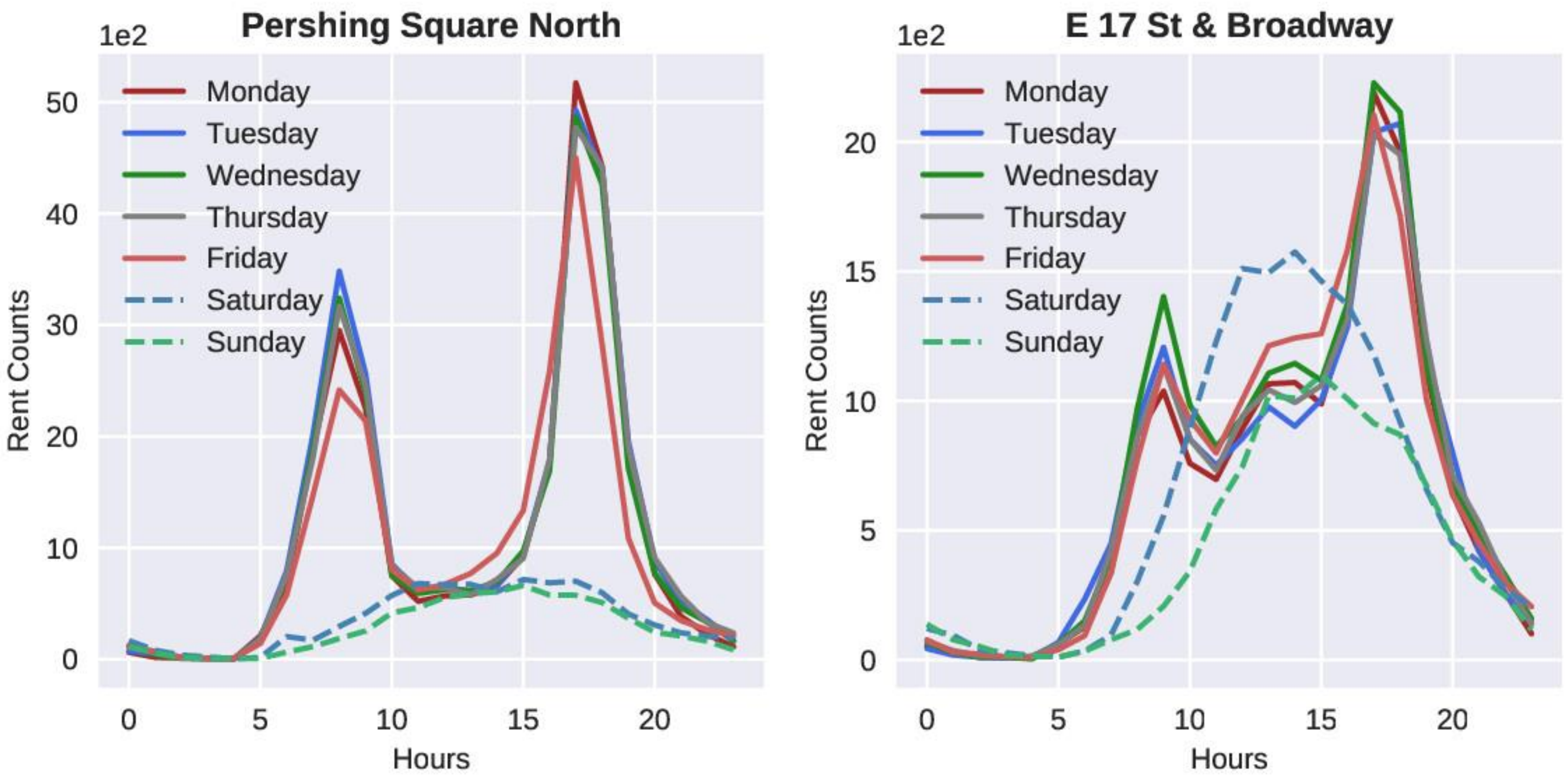
1. Main Challenges
2. **Proposed Solution**
   - **AdaTC** to Cluster the Stations
   - Bicycle Trips Predictor with GNN Embeddings
3. Experimental Results
   - Clustering Pertinence
   - Best Technique for the Problem

Main Challenges

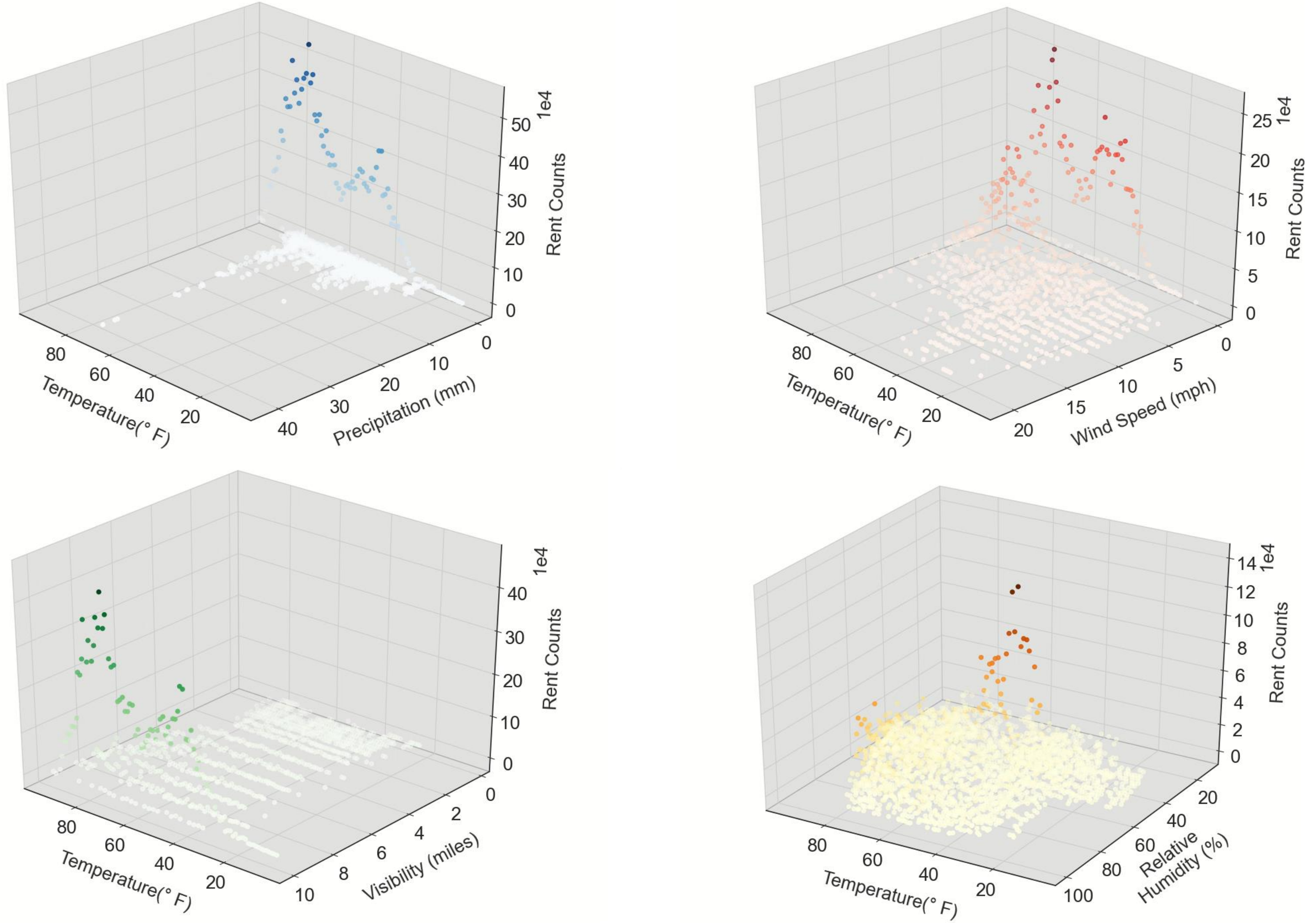
when Predicting Bike Traffic

DEMAND CHANGES TEMPORALLY AND SPATIALLY

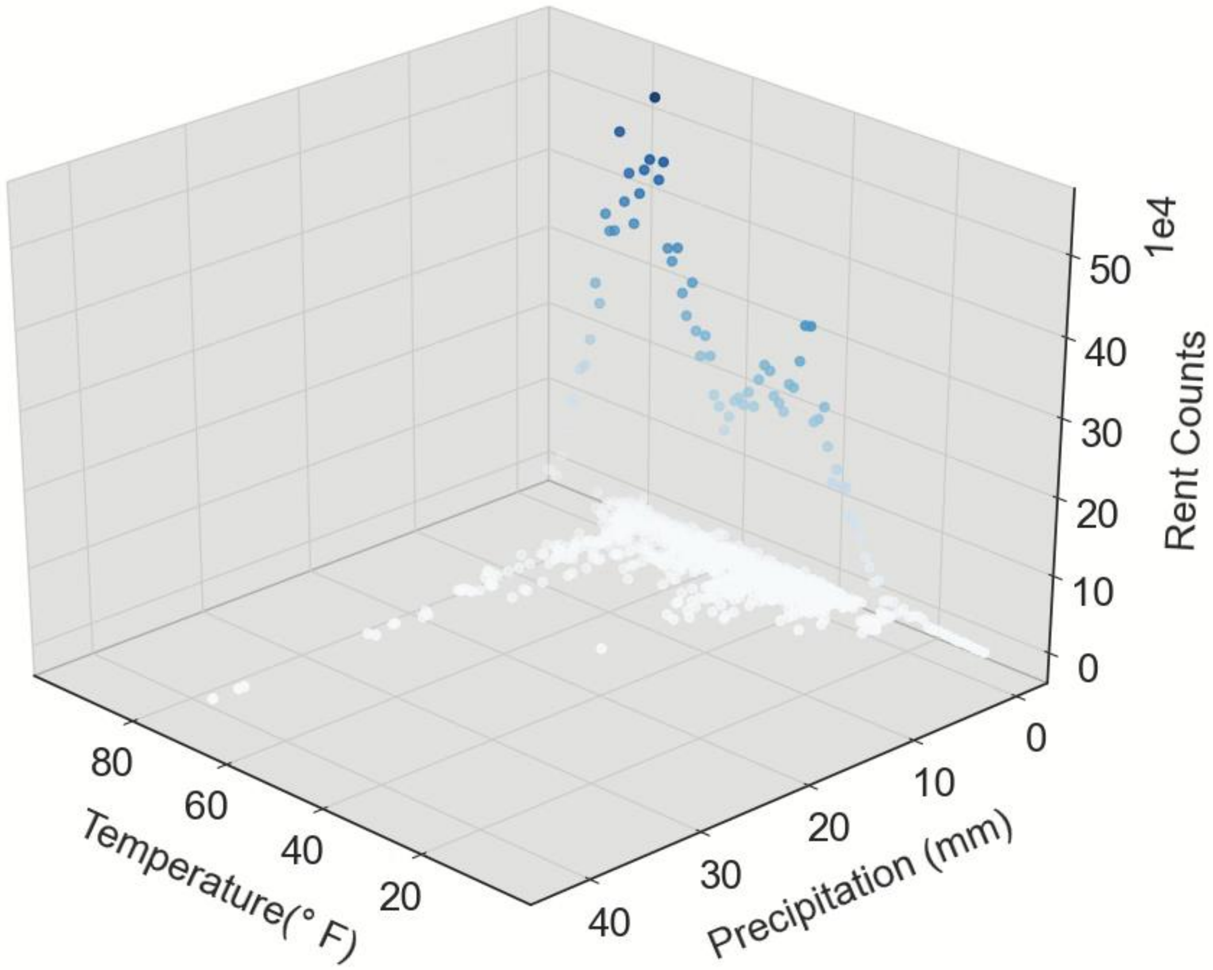
Weekly And Hourly Seasonality In Different Locations



DEMAND IS INFLUENCED BY WEATHER CONDITIONS



DEMAND IS INFLUENCED BY WEATHER CONDITIONS



Temperature and Precipitation

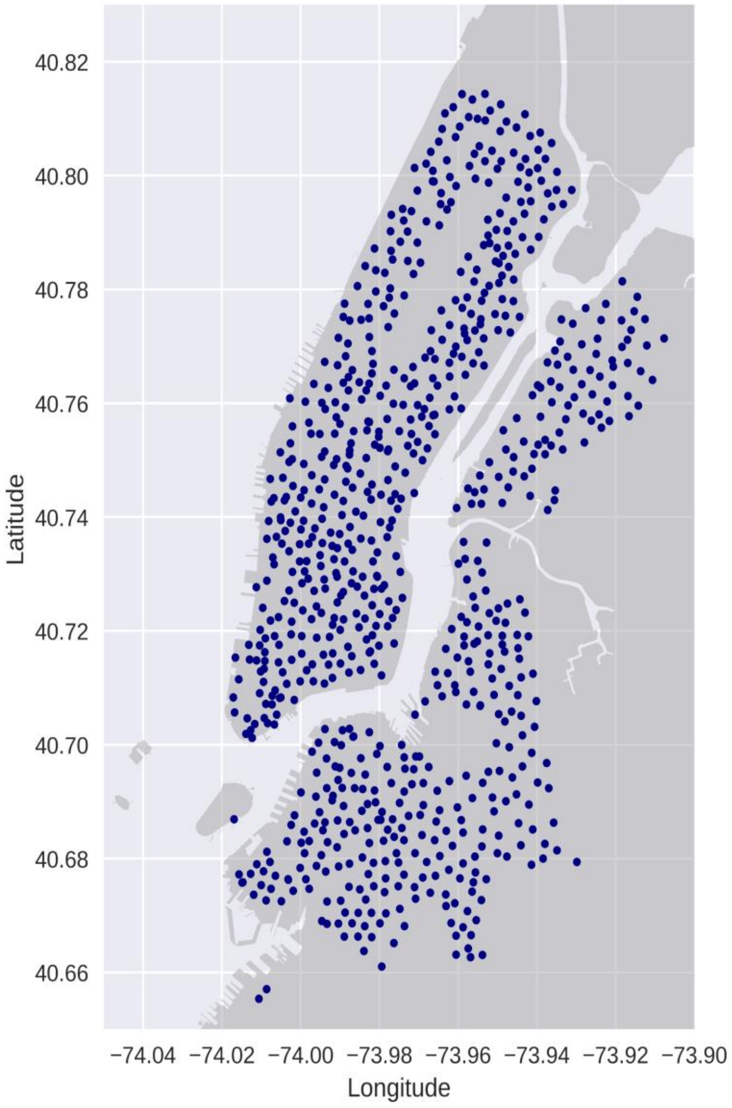
Conducive to Cycling

- Precipitation $\approx 0\ mm/h$
- **Precipitation $\approx 0\ mm/h$ & $\uparrow$ Temperature**

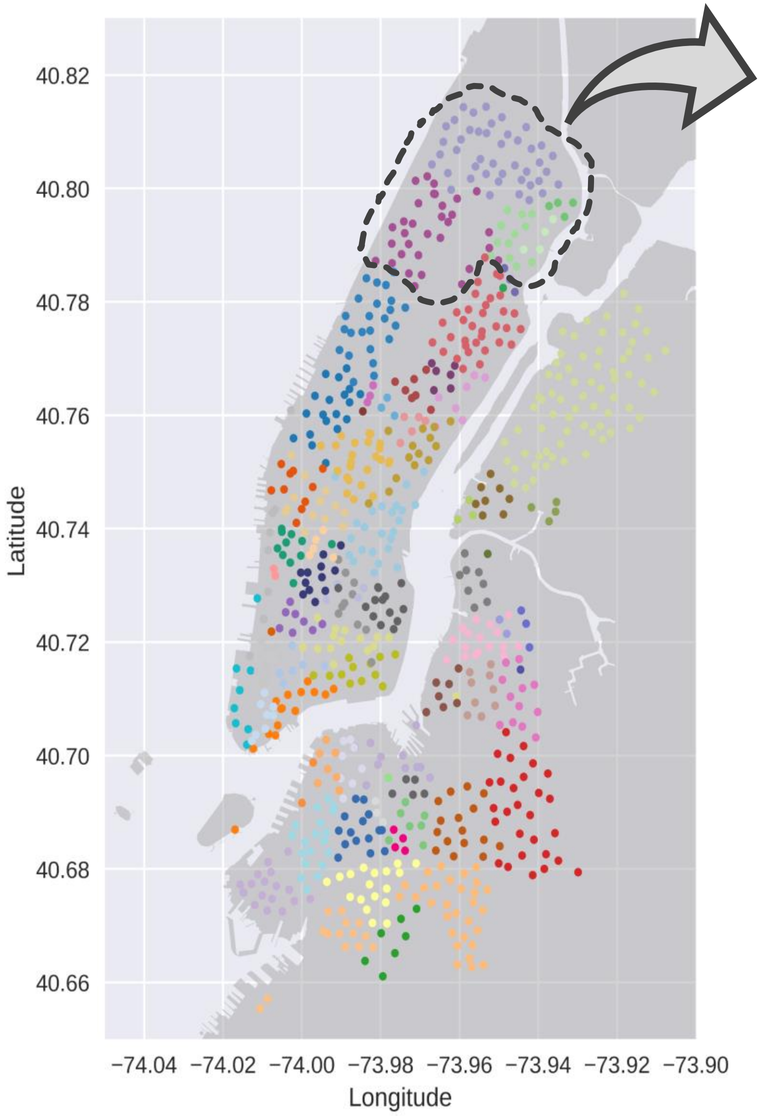
Not Conducive to Cycling

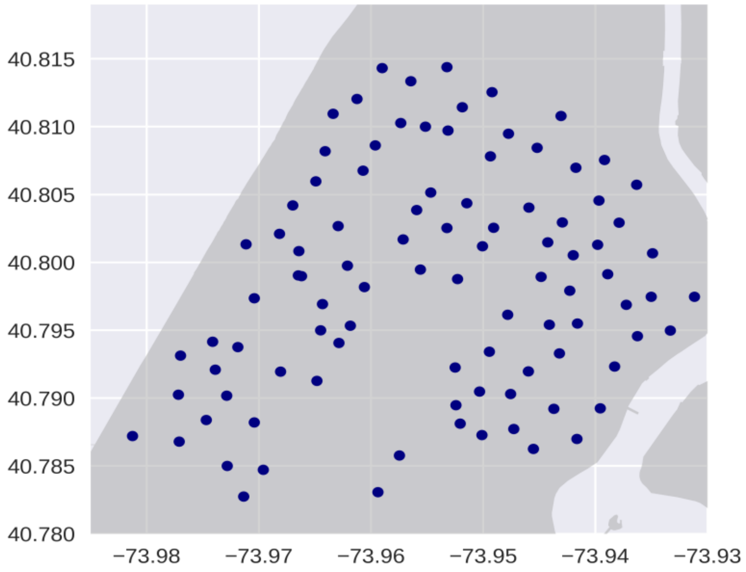
- Precipitation $> 10\ mm/h$
  - Residual # Trips
  - $\uparrow$ Temperature

RIDES RANDOMNESS

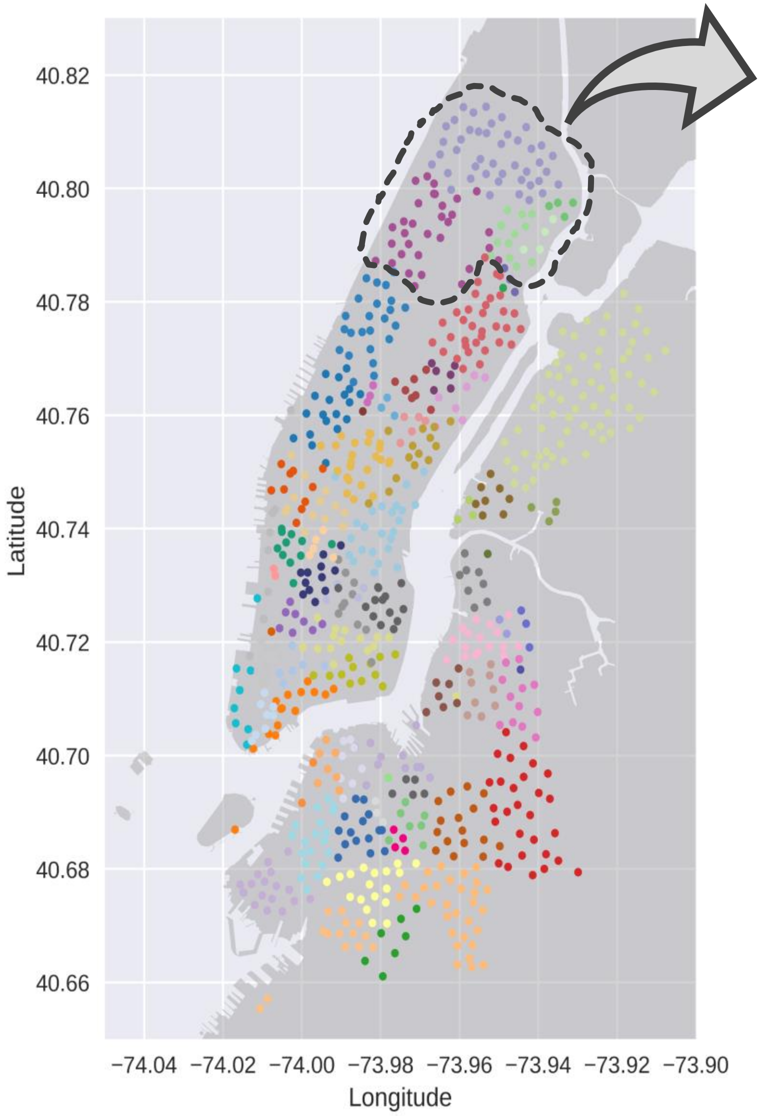


RIDES RANDOMNESS

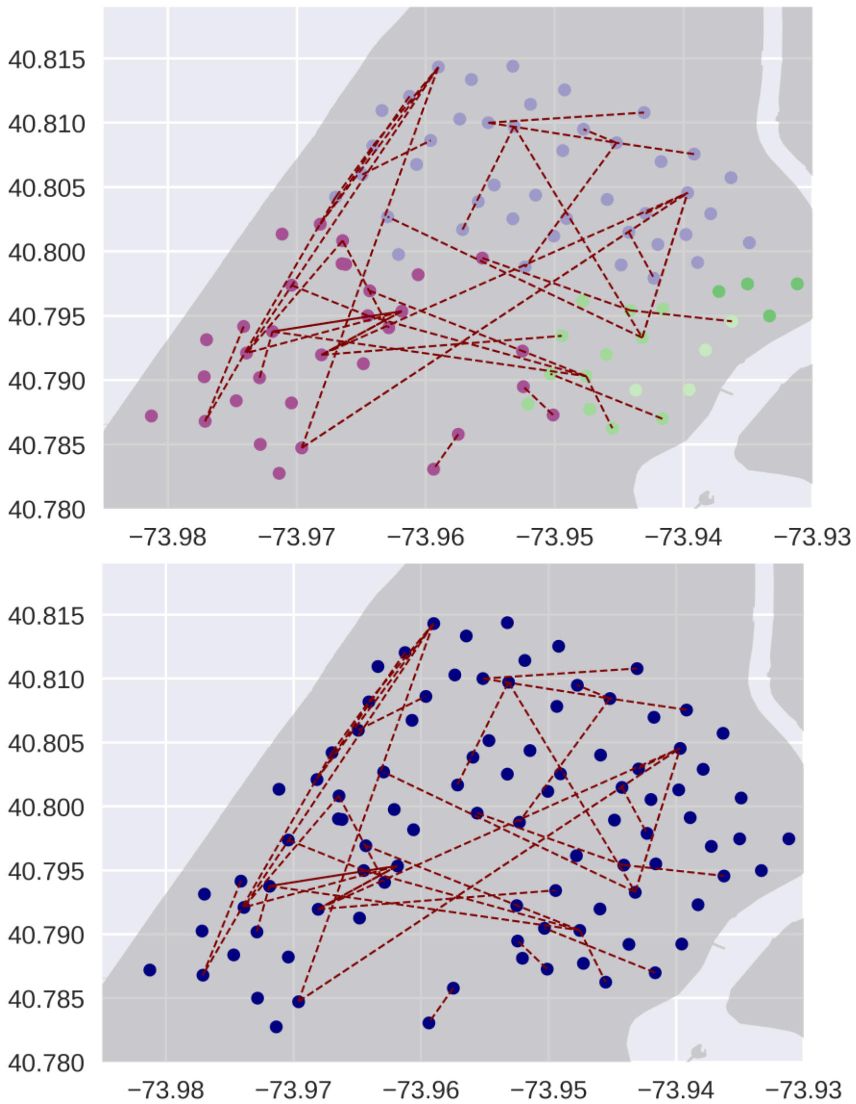


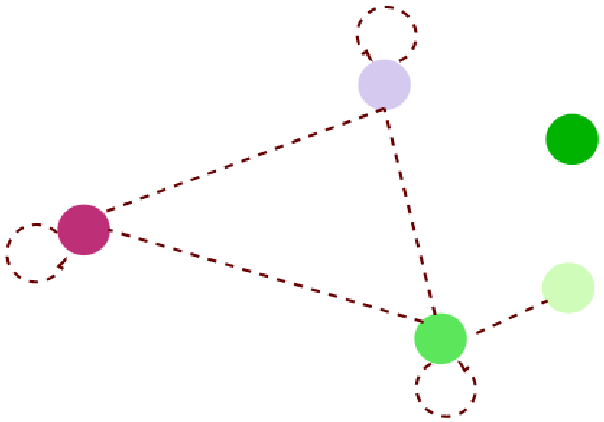


RIDES RANDOMNESS



January 1, 2018
7:00 am – 12:00 pm





Inter-Station Random Trips → Inter/Intra-Cluster Frequent Trips

Apparently Random Trips Between Individual Stations show a Pattern when Stations are Clustered

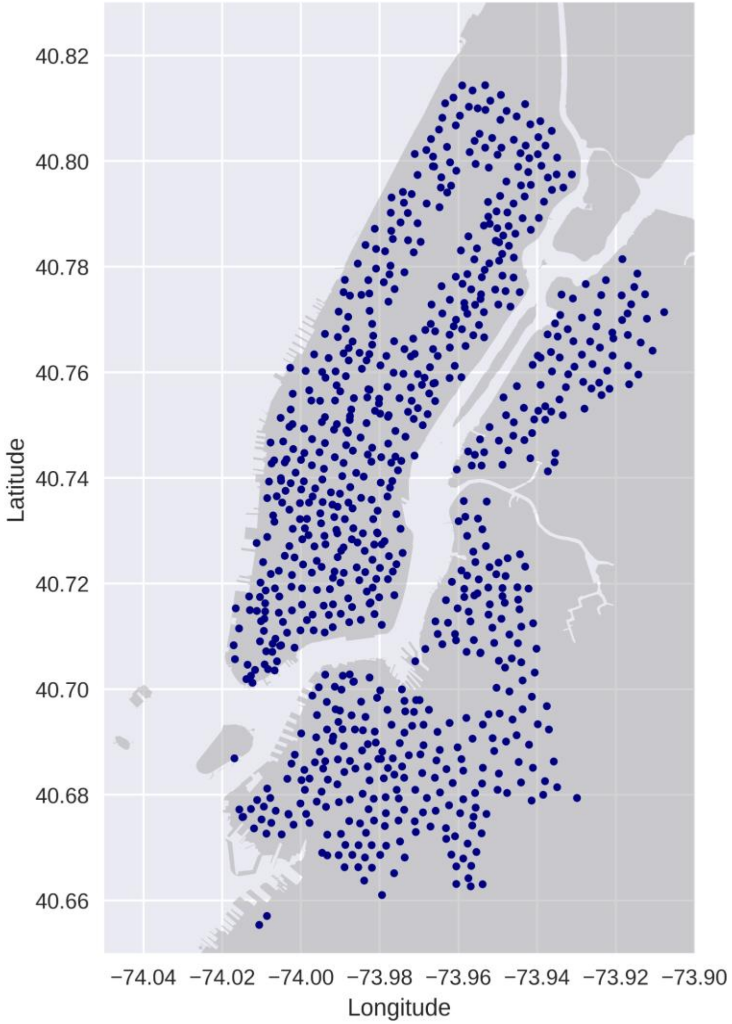
WHY IS CLUSTERING CRUCIAL?

Reduce the Complexity of Random Trips

System Prediction

Traffic Prediction

- Consider the Station Similarity to Neighbor Stations
- Bike Usage more Stable and Regular



System Operation

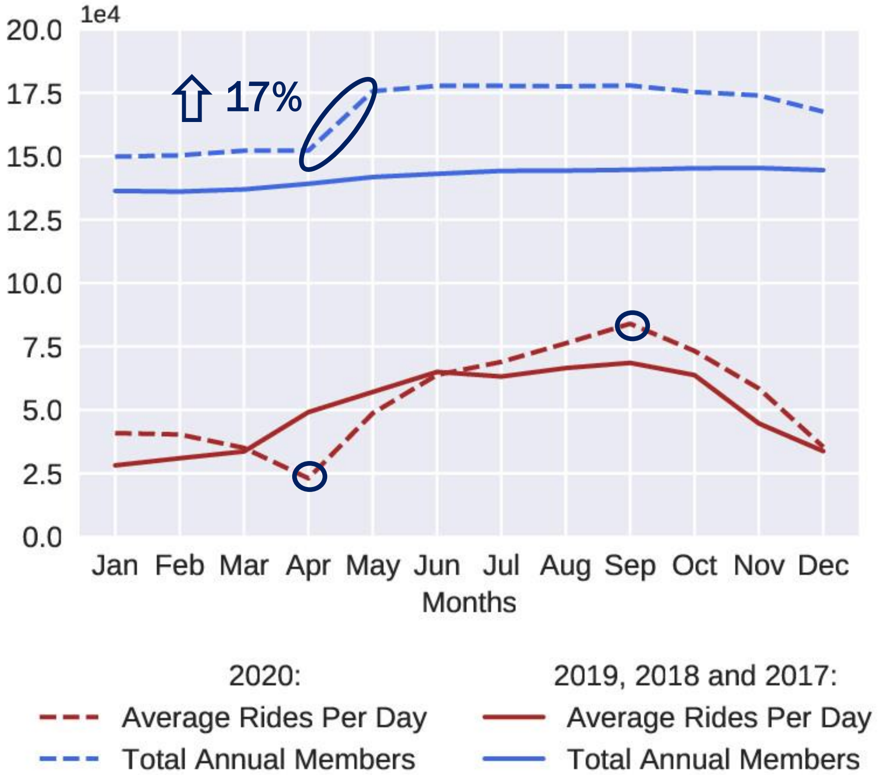
Bike Reposition

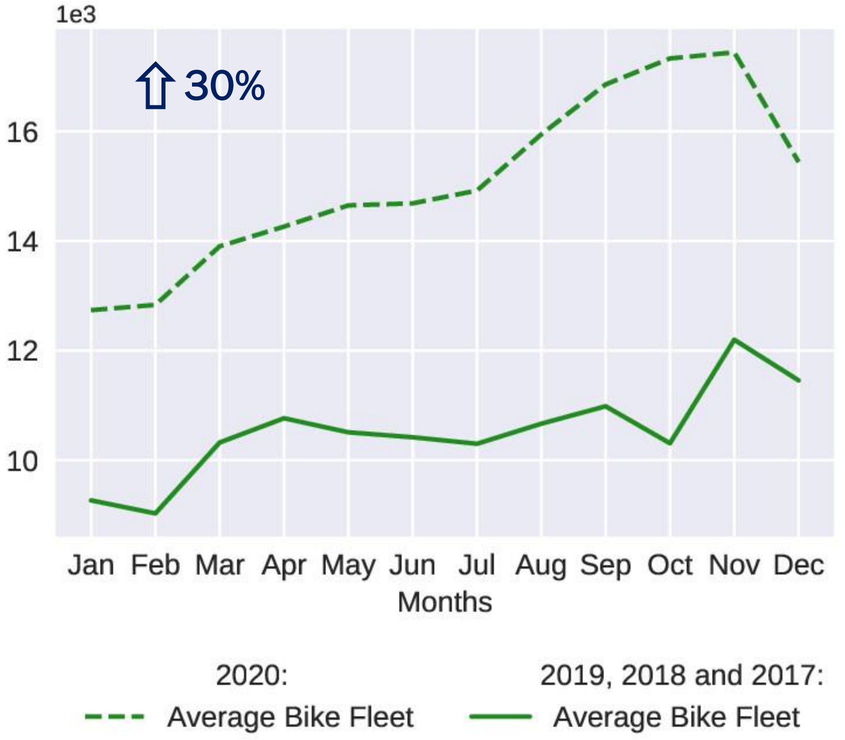
- Inter/Intra-Cluster Dynamics
- Derive Efficient Rebalancing Strategies

COVID-19 IMPACT ON CITIBIKE

2020

March	March – April	April – July	October
WHO declares the COVID-19 Pandemic	Full Lockdown in NY	Four-phase Reopening Plan by Region	Micro-cluster Strategy Introduced





Proposed Solution

Grouping Stations: AdaTC

Before clustering…

- Remove Potential Outlier Stations

$$S^g = \{s_i : s_i \in S, \%\, t_{s_i} \geq 0.001\%\},$$

such that, $\# S^g = 801$

- Assess Cluster Tendency

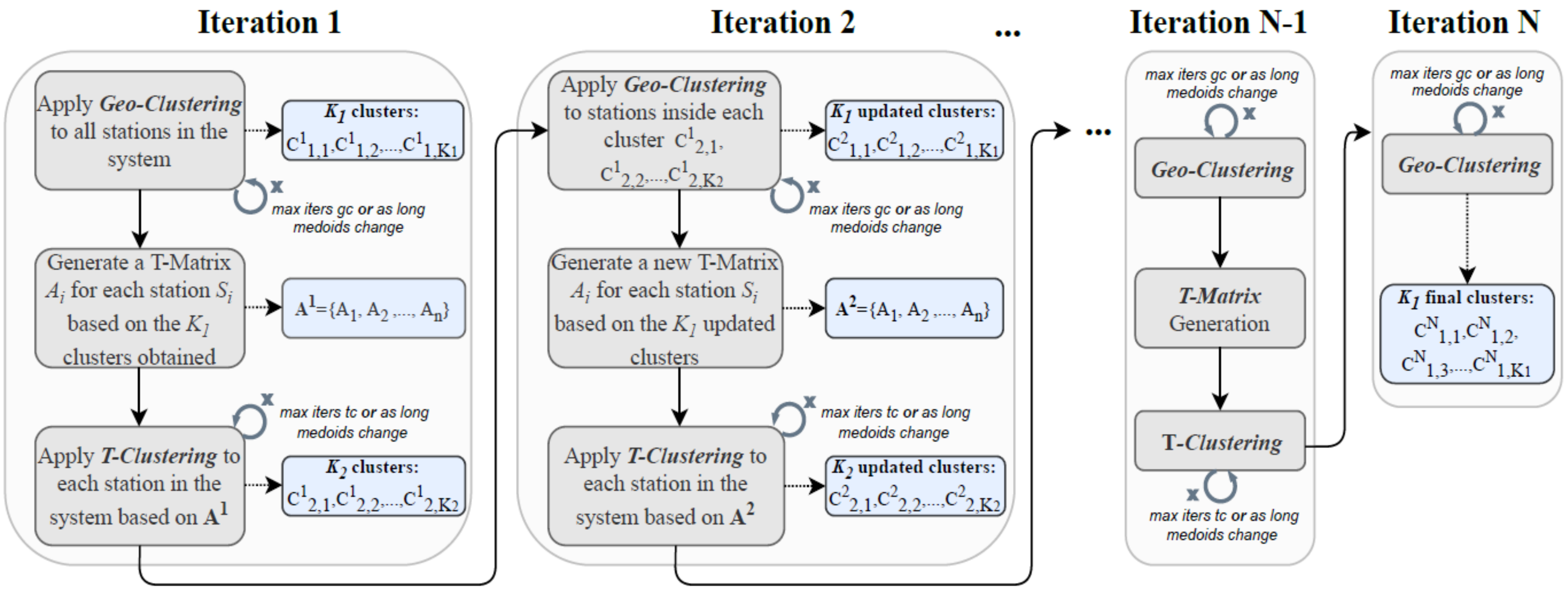
H_0: the data is uniformly distributed

- Define 5 Time Slots

ADAPTIVE TRANSITION CONSTRAINT CLUSTERING

Stations in the same cluster **are expected to:**

- Be Geographically Close
- With Similar Check-outs <u>and</u> Inter-cluster Transitions



AdaTC: GEO-CLUSTERING

$$diss_{GC}(S_h, S_k) = \rho_1 \times gd_{hk} + out_{hk}$$

Trade-off Parameter

Geographical Distance between S_h and S_k

Check-out Difference between S_h and S_k

where

- $$out_{hk} = \|U_h - U_k\|_2$$

where

- $$U_h = \left(\frac{u_1}{\sum_{j=1:3} u_j}, \frac{u_2}{\sum_{j=1:3} u_j}, \frac{u_3}{\sum_{j=1:3} u_j}, \frac{u_4}{\sum_{j=4:5} u_j}, \frac{u_5}{\sum_{j=4:5} u_j}\right)$$ (Check-out Pattern at S_h)

Estimate the Number of Bikes Available at S_h

where

- $$u_i = \frac{1}{TL_i} \times \underbrace{|t|}_{\substack{= \\ \text{60 min.}}} \times r_i$$ (Check-out Pattern at S_h in ts_i)

(Average Time Length that S_h has Bikes Available in ts_i)

$$TL_i = \frac{1}{np_i} \sum_{t=0}^{np_i - 1} \widehat{TL}_{i,t}$$

ts_1

t ... t

$$r_i = \frac{1}{np_i} \sum_{t=0}^{np_i - 1} r_{i,t}$$

(Average Number of Bikes Rented from S_h in ts_i)

AdaTC: GEO-CLUSTERING (cont.)

$\widehat{TL}_{i,t}^{S}$ Estimation

⇩ INFER

<u>Number of Bikes Available</u>:

1

Exact Flow:
Using CitiBike Historical Trips and Open Bus

2

“Offset” Method:
Using only CitiBike Historical Trips

$$\hat{u}_i^S = \frac{1}{\overline{TL_i^S}} \times \underbrace{|t|}_{=\,60\text{ min.}} \times \overline{r_i^S} \qquad (4)$$

(Check-out Pattern at S_h in ts_i)

(Average Time Length that S_h has Bikes Available in ts_i)

$$\boxed{\overline{TL_i^S} = \frac{1}{np_i}\sum_{t=0}^{np_i-1}\widehat{TL}_{i,t}^S}$$

$$\widehat{TL}_{i,t}^S = \begin{cases} \dfrac{1}{\#\,weekdays^{*}} \displaystyle\sum_{weekday\ dates^{*}} \widehat{TL}_{i,t,d}^S & for\ i = 1,2,3 \\[2ex] \dfrac{1}{\#\,weekend\ days^{*}} \displaystyle\sum_{weekend\ dates^{*}} \widehat{TL}_{i,t,d}^S & for\ i = 4,5 \end{cases}$$

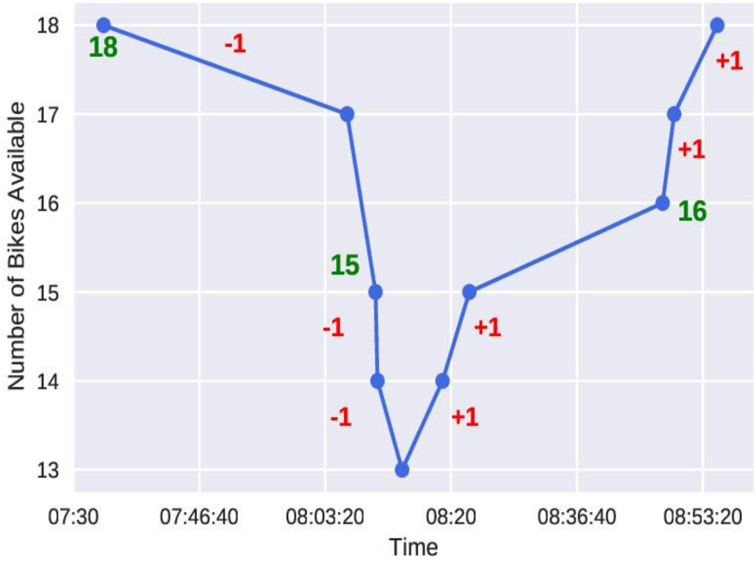
≤ 60 min.

date

≤ 60 min.

* in 2018

1



OpenBus

CitiBike Historical Trips Scores:

- -1 for rents
- +1 for returns

Number of Bikes Available

2

avail_bikes_cum + |min.|

$\underbrace{}_{\text{offset}}$

	station_id	balance	dock_time_time	dock_time_date	avail_bikes_cum	avail_bikes_cum_offset
0	3536.0	-1	09:08:01.358000	2018-01-01	-1	2855
1	3536.0	1	09:18:34.010000	2018-01-01	0	2856
2	3536.0	1	11:23:17.156000	2018-01-01	1	2857
3	3536.0	-1	11:31:55.385000	2018-01-01	0	2856
4	3536.0	1	12:58:13.333000	2018-01-01	1	2857
...	...	...	...	...	...	...
22461	3536.0	-1	13:24:09.036000	2018-12-31	-2854	2
22462	3536.0	1	13:48:14.227000	2018-12-31	-2853	3
22463	3536.0	1	13:48:15.943000	2018-12-31	-2852	4
22464	3536.0	-1	14:07:47.898000	2018-12-31	-2853	3
22465	3536.0	1	16:49:42.485000	2018-12-31	-2852	4

min.

AdaTC: TRANSIT-MATRIX GENERATION

Transit-Matrix: Describes the intra and inter-cluster transitions patterns of a specific station in all time slots

Ride-To-Cluster $(\boldsymbol{rt})$ Return-From-Cluster $(\boldsymbol{rf})$

$$\mathbf{A_s} = \begin{array}{c} \begin{array}{cccccccccc} C_{1,1} & C_{1,2} & C_{1,3} & \cdots & C_{1,K_1} & C_{1,1} & C_{1,2} & C_{1,3} & \cdots & C_{1,K_1} \end{array} \\ \begin{bmatrix} rt_{1,1,1} & rt_{1,2,1} & rt_{1,3,1} & \cdots & rt_{1,K_1,1} & rf_{1,1,1} & rf_{1,2,1} & rf_{1,3,1} & \cdots & rf_{1,K_1,1} \\ & & & \cdots & & & & & \cdots & \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ & & & \cdots & & & & & \ddots & \\ rt_{1,1,5} & rt_{1,2,5} & rt_{1,3,5} & \cdots & rt_{1,K_1,5} & rf_{1,1,5} & rf_{1,2,5} & rf_{1,3,5} & \cdots & rf_{1,K_1,5} \end{bmatrix} \end{array} \begin{array}{c} ts_1 \\ ts_2 \\ ts_3 \\ ts_4 \\ ts_5 \end{array}$$

Time Slots

Dimensions: $5 \times 2K_1$

where for a given $i = 1, \dots, 5$

- $\boldsymbol{rt}_{1,j,i} = \dfrac{\#\ trips\ starting\ in\ S\ ending\ in\ C_{1,j}}{\#\ trips\ in\ ts_i}$, with $\sum_{j=1}^{K_1} rt_{1,j,i} = 1$

- $\boldsymbol{rf}_{1,j,i} = \dfrac{\#\ trips\ ending\ in\ S\ starting\ in\ C_{1,j}}{\#\ trips\ in\ ts_i}$, with $\sum_{j=1}^{K_1} rf_{1,j,i} = 1$

AdaTC: TRANSIT-MATRIX GENERATION (cont.)

Transit-Matrix Generation Step

Iteration $\boldsymbol{n_{it}}$

↓ Geo-Clustering

$C_{1,1}^{n_{it}}, C_{1,2}^{n_{it}}, \dots, C_{1,K_1}^{n_{it}}$

↓

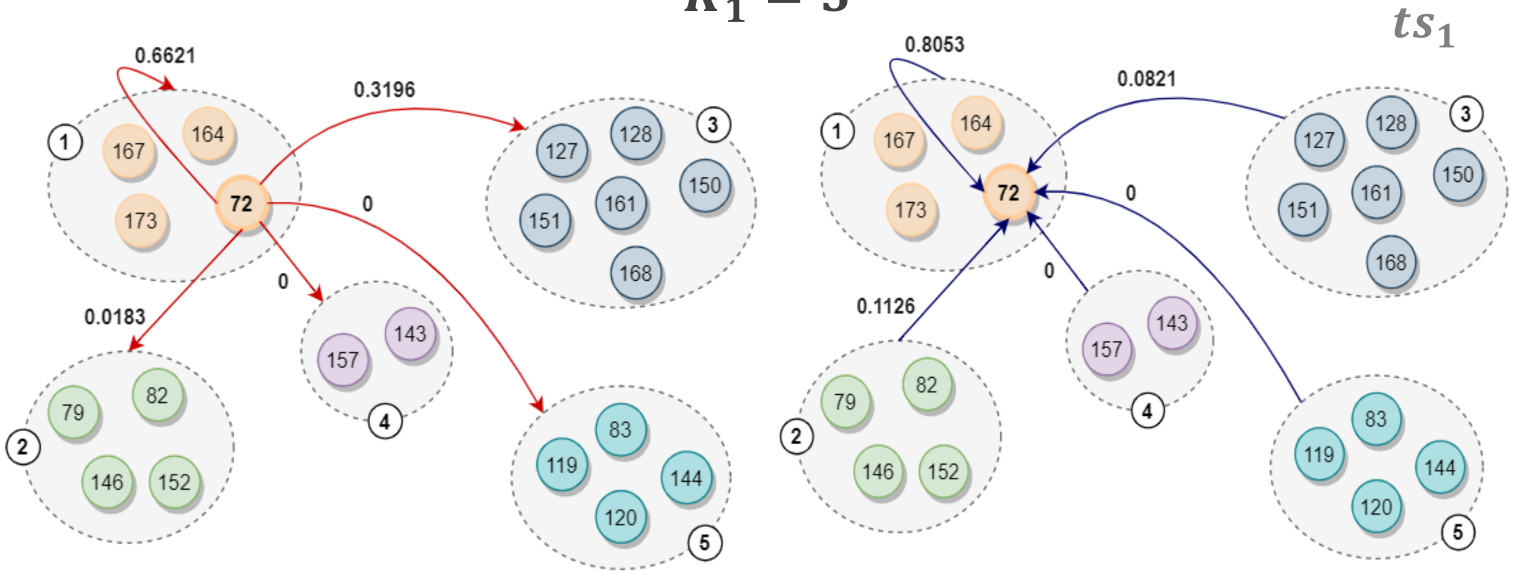
$\mathbb{A}^{n_{it}} = \{A_1, A_2, \dots, A_n\}$, with $n = \#$ stations

↓ Transit-Clustering

W 52 St & 11 Ave $(\boldsymbol{id = 72})$

$\boldsymbol{n = 20}$

$\boldsymbol{K_1 = 5}$



Ride-To-Cluster ⟷ Return-From-Cluster

$$
\mathbf{A}_{72} = \begin{array}{cccccccccc|c}
C_{1,1} & C_{1,2} & C_{1,3} & C_{1,4} & C_{1,5} & C_{1,1} & C_{1,2} & C_{1,3} & C_{1,4} & C_{1,5} & \\
\end{array}
$$

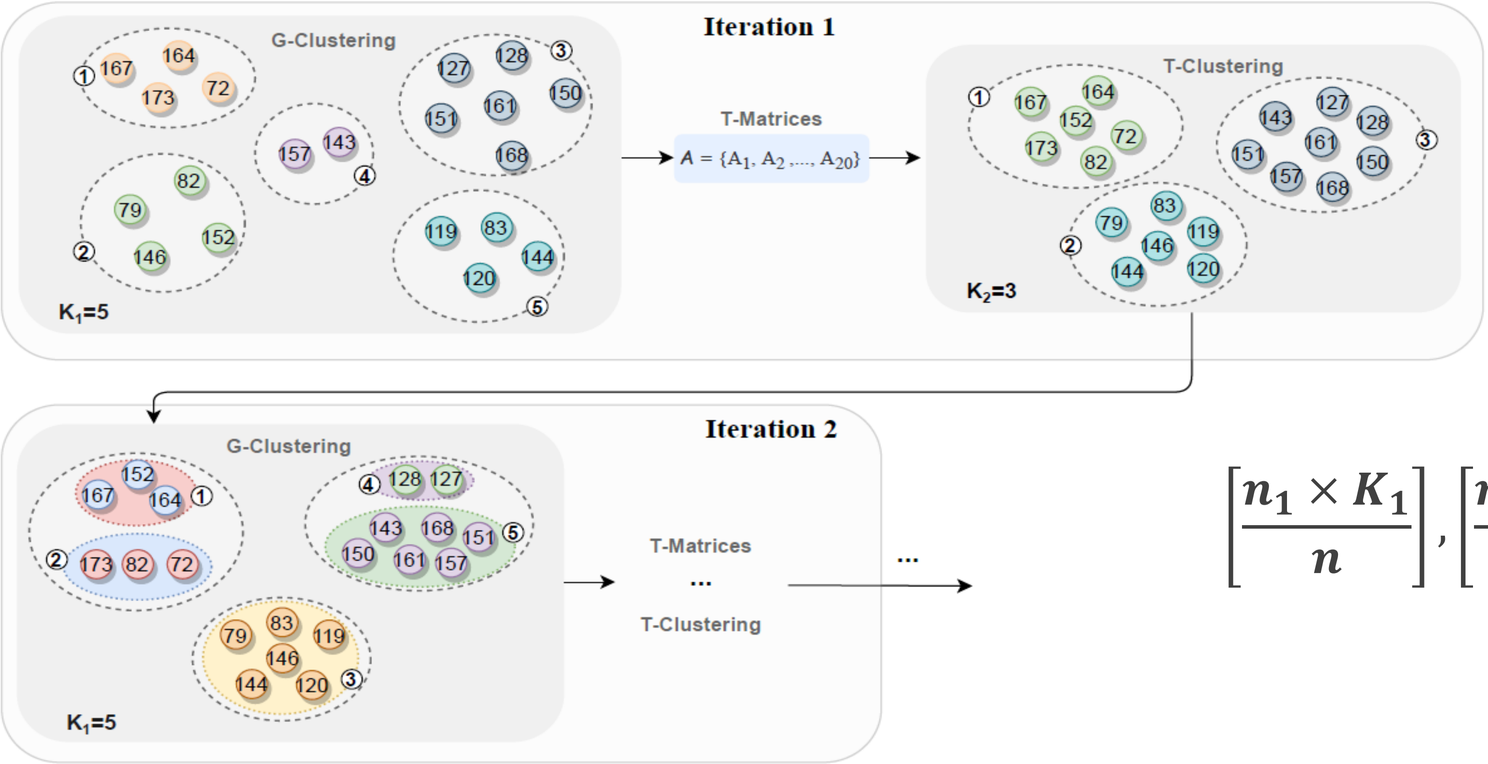
$$
\mathbf{A}_{72} = \begin{bmatrix}
0.6621 & 0.0183 & 0.3196 & 0 & 0 & 0.8053 & 0.1126 & 0.0821 & 0 & 0 \\
0.8360 & 0.0299 & 0.1395 & 0 & 0 & 0.8525 & 0.0354 & 0.1121 & 0 & 0 \\
0.7519 & 0.0611 & 0.187 & 0 & 0 & 0.5443 & 0.085 & 0.3689 & 0.0018 & 0 \\
0.6451 & 0.0444 & 0.3106 & 0 & 0 & 0.6862 & 0.0691 & 0.2394 & 0.0053 & 0 \\
0.8 & 0.0214 & 0.1786 & 0 & 0 & 0.8035 & 0.0173 & 0.1676 & 0.0116 & 0
\end{bmatrix}
\begin{array}{c} ts_1 \\ ts_2 \\ ts_3 \\ ts_4 \\ ts_5 \end{array}
$$

Time Slots

AdaTC: TRANSIT-CLUSTERING

Clusters stations into K_2 groups by *K-Medoids*, such that $K_1 \geq K_2$

$$diss_{TC}(S_h, S_k) = \underbrace{||A_h}_{\text{T-Matrix of Station } S_h} - \underbrace{A_k}_{\text{T-Matrix of Station } S_k}||_F = ||A_{\underbrace{(h-k)}_{\text{Element-By-Element Subtraction of } A_h \text{ with } A_k}}||_F$$



$$\left[\frac{n_1 \times K_1}{n}\right], \left[\frac{n_2 \times K_1}{n}\right], \ldots, \left[\frac{n_{K_2} \times K_1}{n}\right]$$

AdaTC: INTRINSIC PARAMETERS VALIDATION

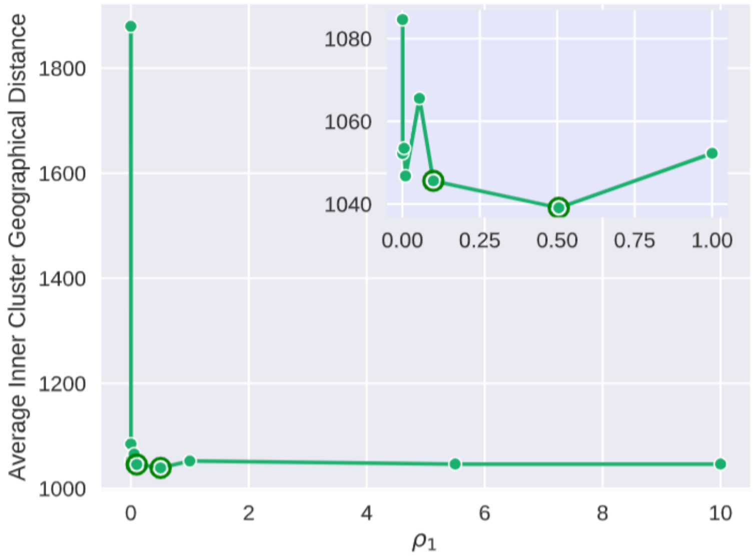
Intrinsic Parameters

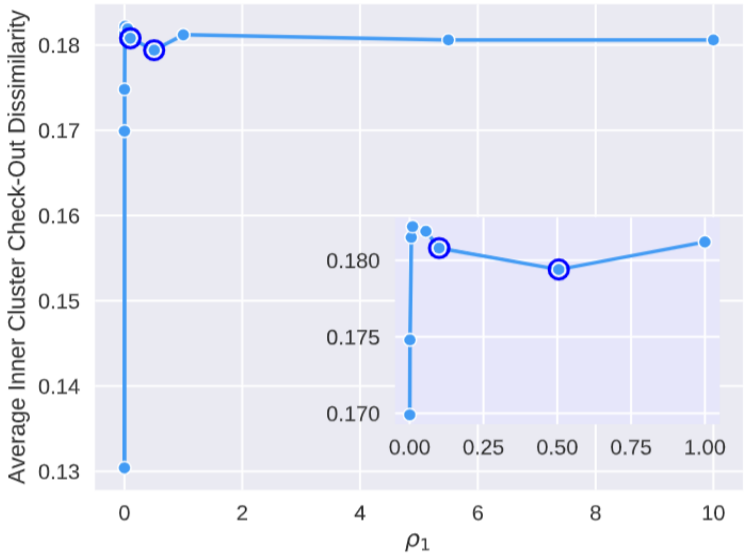
- ρ_1: $\rho_1 \in \{0, 5\times10^{-4}, 10^{-3}, 5.5\times10^{-3}, 10^{-2}, 5.5\times10^{-2}, 10^{-1}, 5.05\times10^{-1}, 1, 5.5, 10\}$
- K_1: $K_1 \in \{50, 60, 70, 80, 90, 100\}$
- K_2: $K_2 \in \{10, 20, 30, 40, 50\}$

VALIDATION METRICS

$\mathbf{AGD}_{\text{inner}}, \mathbf{ACOD}_{\text{inner}}, \mathbf{AGD}_{\text{inter}}, \mathbf{ACOD}_{\text{inter}}$

Average Inner Cluster Geographical Distance	Average Inner Cluster Check-Out Dissimilarity	Average Inter Cluster Geographical Distance	Average Inter Cluster Check-Out Dissimilarity



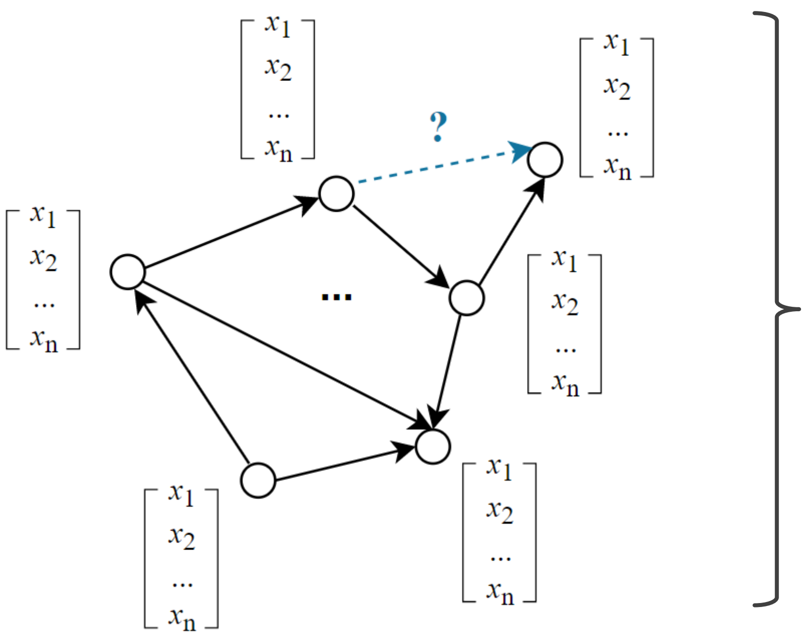


		$K_2 = 10$	$K_2 = 20$	$K_2 = 30$	$K_2 = 40$	$K_2 = 50$
$\rho_1 = 0.1$	$K_1 = 50$	1, 2, 4, 5 *1	3, 4, 2, 2	3, 5, 1, 3	5, 3, 3, 1	2, 1, 5, 4
	$K_1 = 60$	1, 2, 5, 2	4, 5, 3, 1	2, 4, 1, 2	3, 3, 2, 5	4, 1, 4, 3
	$K_1 = 70$	2, 4, 4, 3	1, 1, 5, 2	4, 4, 2, 1	5, 3, 1, 5	3, 2, 3, 4
	$K_1 = 80$	1, 1, 5, 3	5, 5, 1, 5	2, 3, 3, 4	3, 4, 4, 2	4, 2, 2, 1
	$K_1 = 90$	2, 3, 2, 2	4, 1, 3, 1	5, 5, 1, 2	3, 4, 2, 1	1, 2, 5, 3
	$K_1 = 100$	1, 1, 5, 3	3, 4, 4, 2	2, 3, 2, 5	4, 2, 2, 1	5, 5, 1, 3
$\rho_1 = 0.505$	$K_1 = 50$	1, 3, 4, 1	2, 5, 1, 4	2, 4, 2, 2	5, 2, 3, 5	1, 1, 5, 3
	$K_1 = 60$	2, 3, 4, 4	1, 1, 5, 1	5, 3, 2, 3	3, 2, 3, 5	3, 5, 1, 4
	$K_1 = 70$	1, 3, 4, 4	4, 5, 1, 1	3, 4, 3, 3	2, 2, 2, 2	5, 1, 5, 5
	$K_1 = 80$	1, 1, 5, 3	2, 2, 3, 1	3, 5, 1, 1	4, 3, 4, 4	5, 4, 2, 5
	$K_1 = 90$	1, 2, 2, 2	4, 1, 1, 1	2, 2, 3, 1	3, 1, 5, 4	5, 2, 2, 1
	$K_1 = 100$	1, 2, 5, 5	4, 1, 1, 1	2, 2, 4, 1	3, 5, 1, 5	5, 2, 2, 1

Proposed Solution

Bike Trips Predictor

A GNN BICYCLE TRIP PREDICTOR



Ride

Station

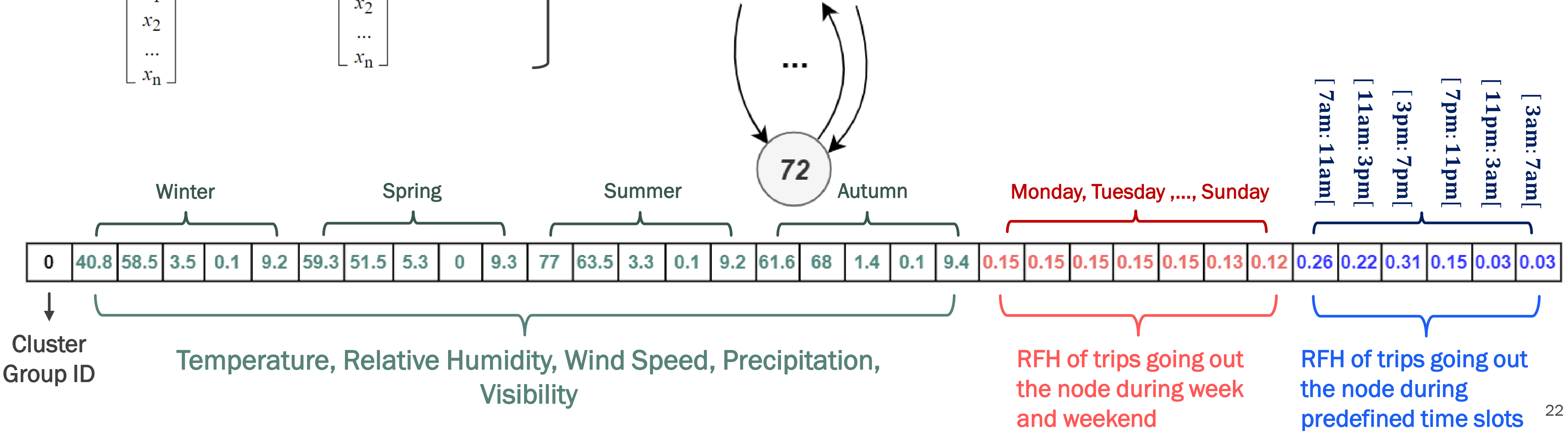
Features

Create the Multigraph Network Structure

+

GraphSAGE Algorithm: an *inductive* node embedding framework

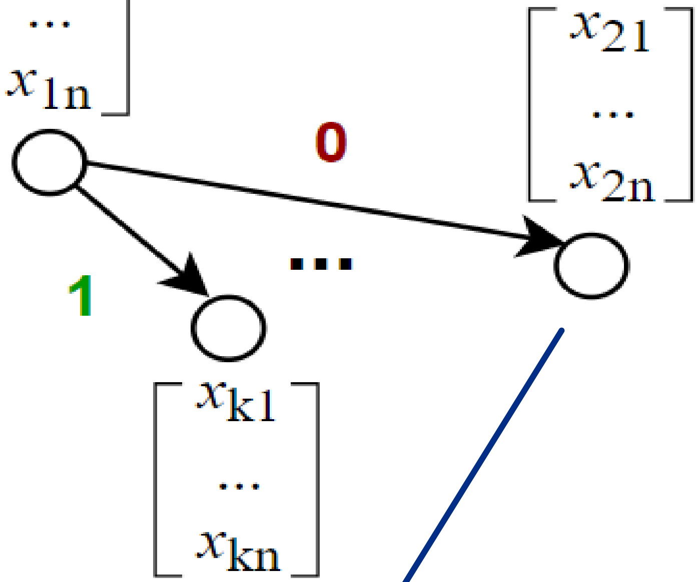
Predict Links (*Bike Trips*) BASED ON → Learning of the Node Features



GNN ARCHITECTURE

Node Embedding

$$\underbrace{\begin{bmatrix} mean(x_{21}, x_{k1}) \\ \dots \\ mean(x_{2n}, x_{kn}) \end{bmatrix} \begin{bmatrix} x_{11} \\ \dots \\ x_{1n} \end{bmatrix}}_{\sigma\,(W \times \mathrm{CONCAT}\,)}$$



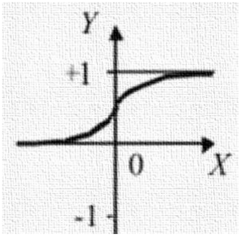
GraphSAGE Embedding Generation

Link Embedding

Link Classification Layer

Inner Product of Node Embeddings

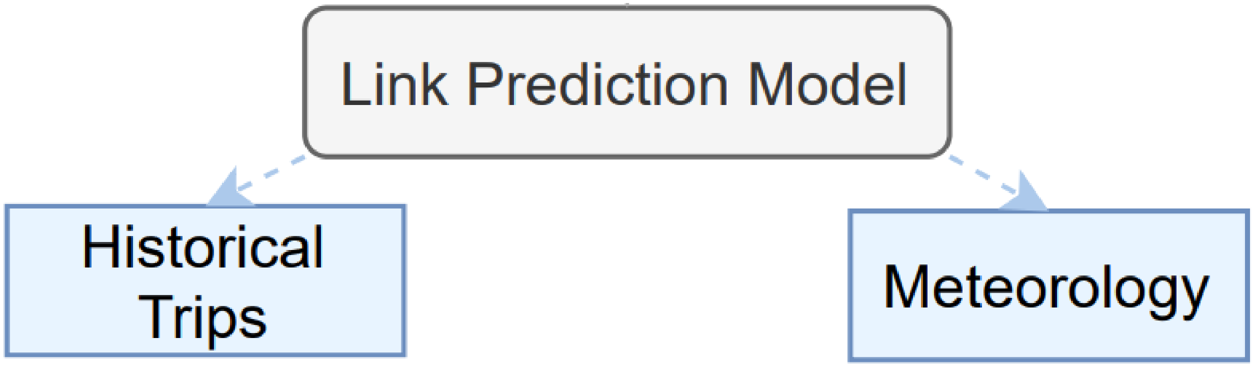
$$ip(u, v) = \sum_i u_i \times v_i$$



Dense Link Classification Layer

PROBLEM SETUP: DATA

2018



Four Clustering Configurations

↓

AdaTC and *three* baselines:

- K-Medoids (**KM**)
- Spectral Clustering (**SC**)
- Geo-Clustering (**GC**)

Train *Five* LP Models

↓

Four Cluster-Based:

- **AdaTC**, **KM**, **SC** and **GC**

One Without Clustering:

- No Clustering (**NC**)

Evaluate Performance on test held out data

2019 NETWORK

2019

Infer Generalization on test data, trained on 2018 data

↓

KEEPING

- Clustering Results
- Weather

UPDATING

- Historical CitiBike Trips

ACCURACY IN THE TEST SET AFTER CALIBRATION. ADATC$_+$ OUTPERFORMS THE BASELINES AND WHEN IN MISMATCH, THE PERFORMANCE DOES NOT DEGRADE SIGNIFICANTLY.

	AdaTC$_+$	GC	SC	KM	NC
2018	88%	87%	86%	83%	83%
2019 (Mismatch)	85%	86%	85%	83%	84%

ENTIRE GRAPH DIMENSIONS IN 2018 AND 2019. $2019_{\text{restricted}}$ REPRESENTS THE 2019 GRAPH RESTRICTED TO THE STATIONS (AND CORRESPONDING TRIPS) IN THE 2018 SETTING.

2018		$2019_{\text{restricted}}$		2019	
# nodes	# links	# nodes	# links	# nodes	# links
801	17.526.058	762	18.674.518	1333	20.551.697

AdaTC + GRAPHSAGE AS A BETTER PERFORMANT MODEL

Best Model ⟶ Performance on the Prediction Task

2018

Cluster-Based				Without Clustering
Our Predictor	Baselines			
AdaTC	GC	SC	KM	NC
88%	87%	86%	83%	83%

- Cluster Relevance
- AdaTC Outperforms all the Baselines in the 2018 setting

AdaTC + GRAPHSAGE MODEL PERCENTAGE ERROR

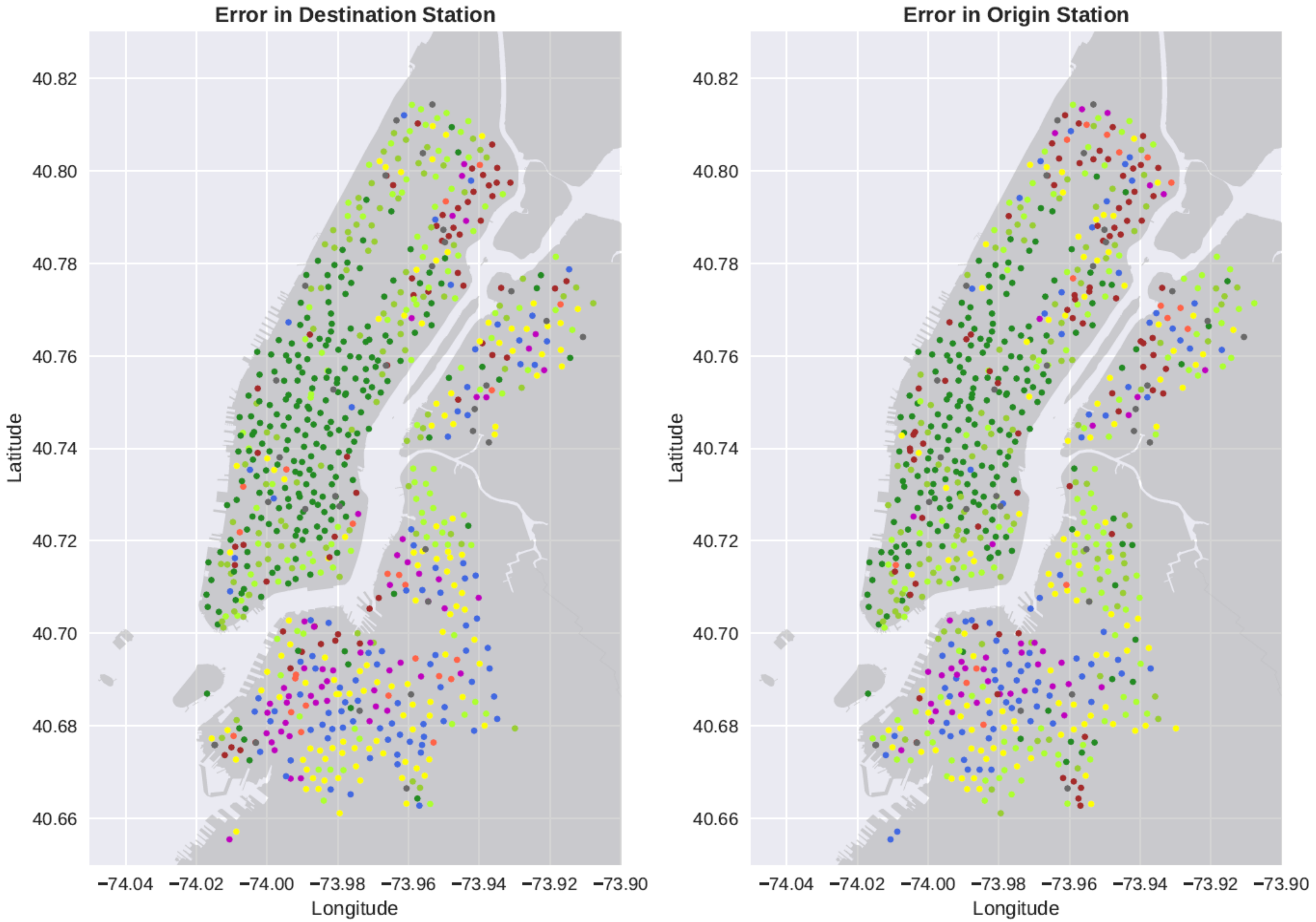
$$PE = \frac{1}{x_t}\left|x_t - x_p\right| \times 100 \ (\%)$$

Number of True Rides

Number of Positive Rides Predicted

citibike_station_id	start_true	start_pred_true	stop_true	stop_pred_true	error_start	error_stop
72.0	4.0	9.0	4.0	6.0	125.00	50.0
120.0	54.0	20.0	0.0	2.0	62.96	inf
127.0	359.0	356.0	8.0	15.0	0.84	87.5

AdaTC + GRAPHSAGE MODEL PERCENTAGE ERROR



$$PE = \frac{1}{x_t}\left|x_t - x_p\right| \times 100\ (\%)$$

Number of True Rides

Number of Positive Rides Predicted

- 0% ≤ Percentage Error ≤ 5%
- 5% < Percentage Error ≤ 15%
- 15% < Percentage Error ≤ 30%
- 30% < Percentage Error ≤ 45%
- 45% < Percentage Error ≤ 60%
- 60% < Percentage Error ≤ 75%
- 75% < Percentage Error ≤ 90%
- Percentage Error > 90%
- Stations Removed

AdaTC + GRAPHSAGE MODEL PERCENTAGE ERROR

\# Trips

↓↑

Error

Stations with Stable Transitions Patterns

↓

PE is on average $\approx 20\%$

12% of Error

↓

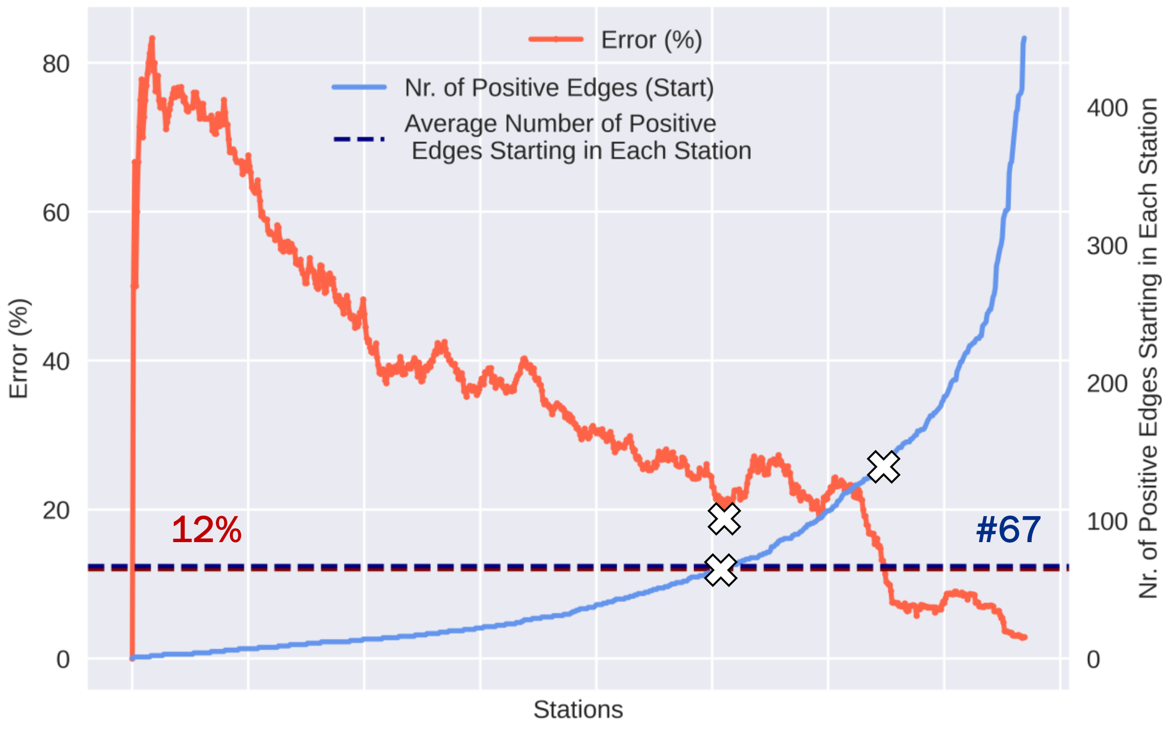
- Stations with ≥ 120 trips starting
- Stations with ≥ 100 trips ending

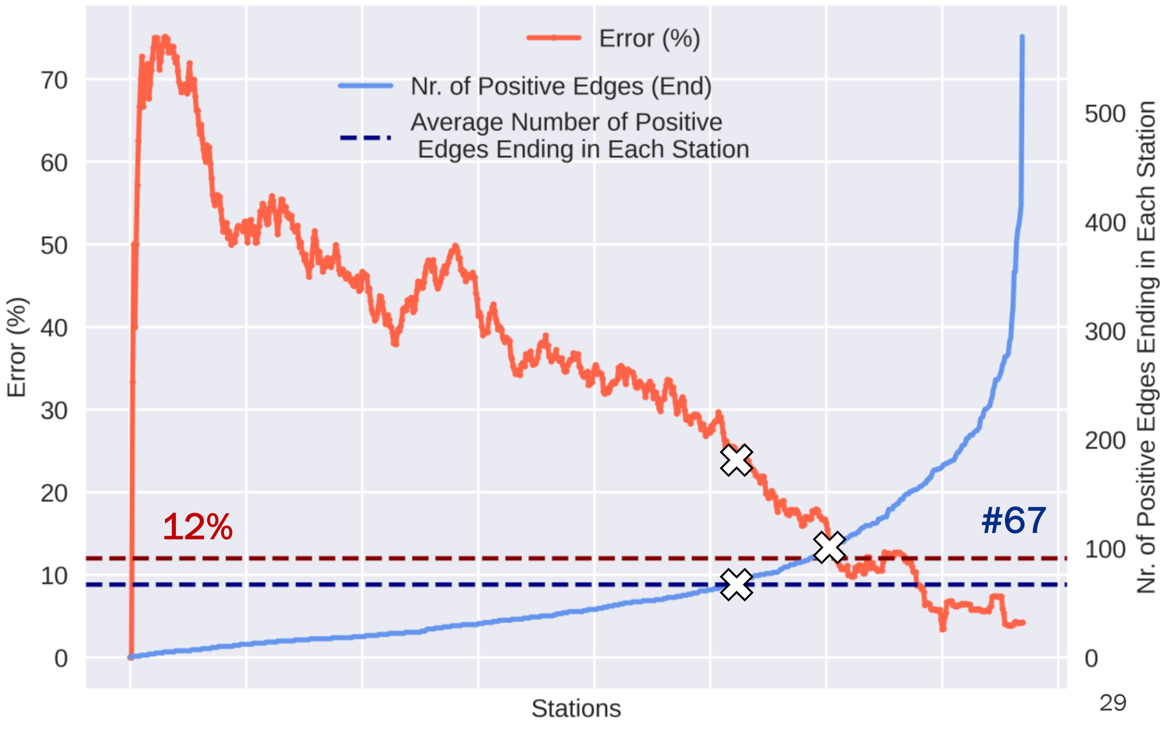
Larger PE → **Outlier Stations**

Stations with a smaller number of trips

64% of Stations are atypical

(below the average)





CONCLUSIONS AND LIMITATIONS

- We Provide a Lower Bound on the Accuracy for the Model in this Predictive Task

- Limitation: We cannot take advantage of the *Indutive Nature* of the Model

Future Work Redefine the Loss or *Continual Learning* Strategy