Traffic Prediction in a Bike-Sharing System

Team 1
Jiaqi Liu, Shaowei Gong, Qiuyi Hong, Zhenhua Li, Caidan Liu

Background

- Bike-sharing systems are widely deployed in many major cities, providing a convenient transportation mode for citizens' commutes.
- As the rents/returns of bikes at different stations in different periods are unbalanced, the bikes in a system need to be rebalanced frequently.


Solution Overview

Real-time monitoring

- Monitoring the current number of bikes at each station cannot tackle the challenge thoroughly, as it is too late to reallocate bikes after an imbalance has occurred.

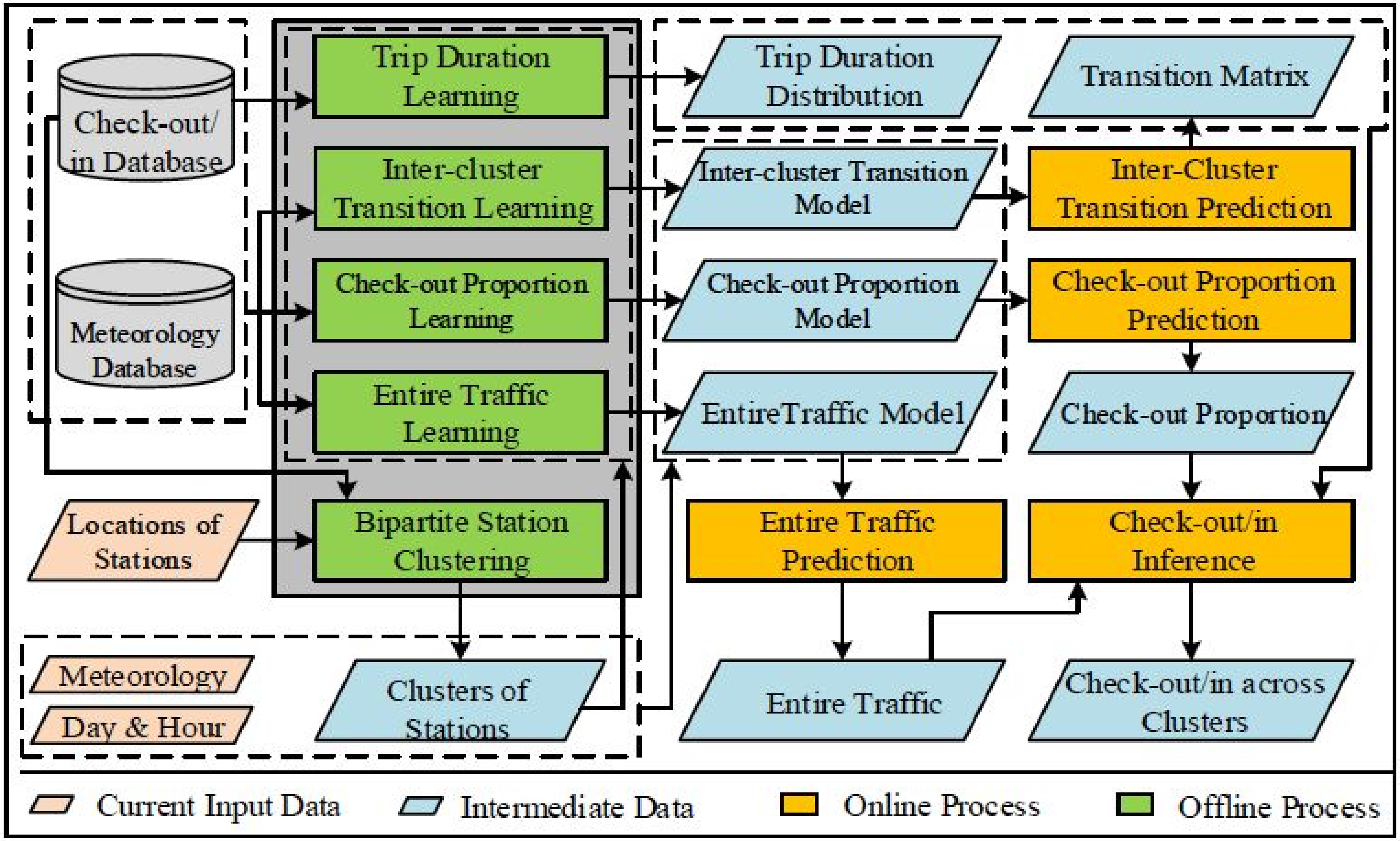
Hierarchical Model

- Bipartite Station Clustering
- multi-similarity-based inference model
- check-in inference algorithm

Challenges:

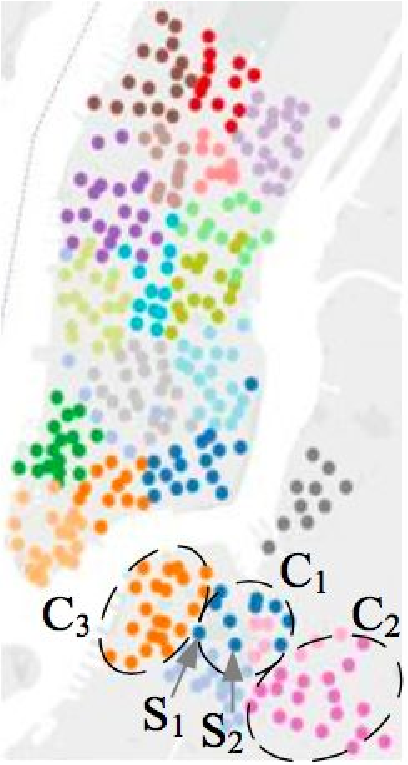
Meteorology	Correlation between stations	Others
•sunny/rainy •cool/warm • rare conditions	• nearby stations • mutual impact	• time of day • day of the week • events,

Framework

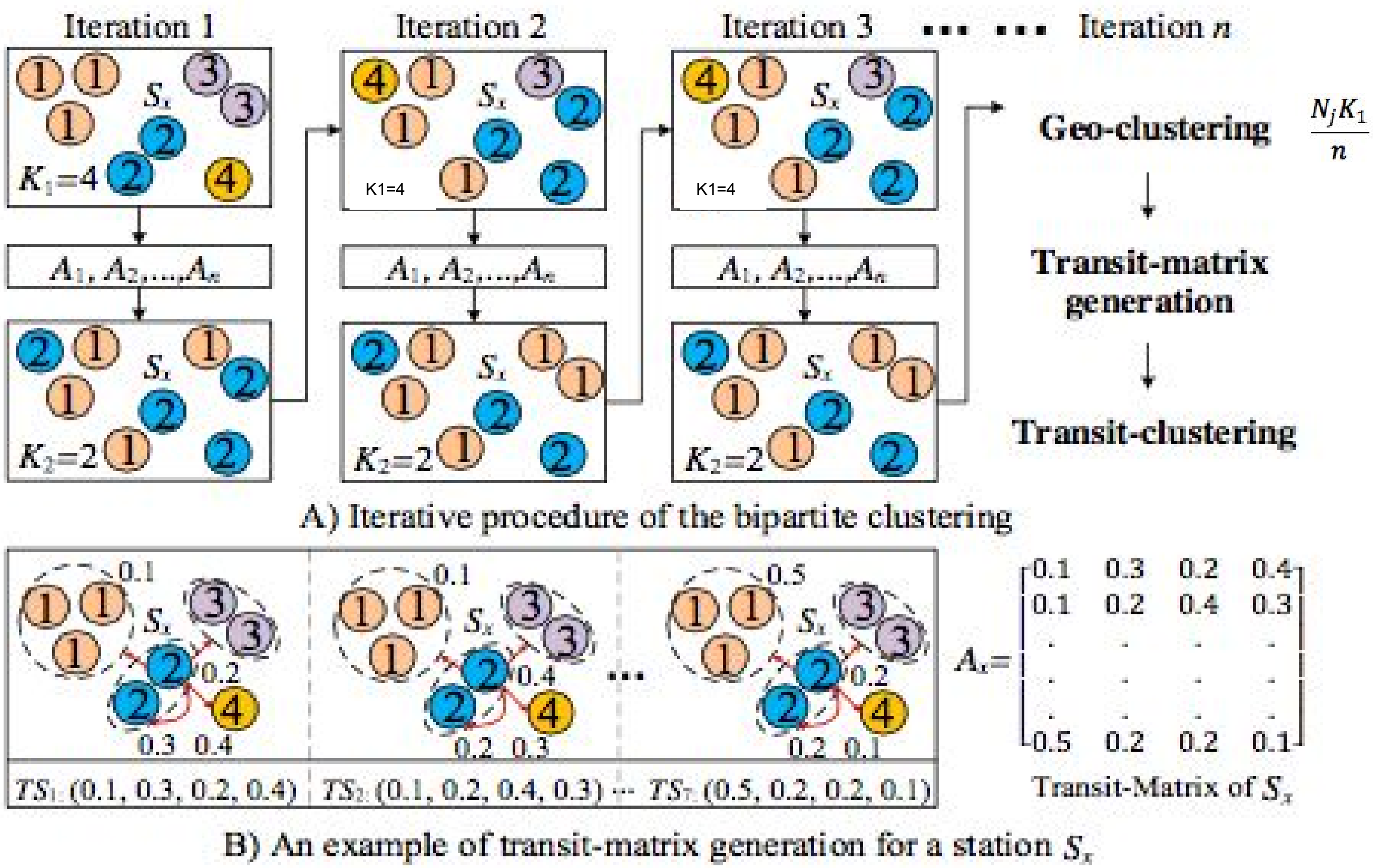

Check-out/ in Database
Meteorology Database
Trip Duration Learning
Inter-cluster Transition Learning
Check-out Proportion Learning
Entire Traffic Learning
Bipartite Station Clustering
Locations of Stations
Meteorology
Day & Hour
Clusters of Stations
Trip Duration Distribution
Transition Matrix
Inter-cluster Transition Model
Inter-Cluster Transition Prediction
Check-out Proportion Model
Check-out Proportion Prediction
EntireTraffic Model
Check-out Proportion
Entire Traffic Prediction
Check-out/in Inference
Entire Traffic
Check-out/in across Clusters
Current Input Data
Intermediate Data
Online Process
Offline Process

Bipartite Station Clustering

- Group individual station into clusters according to their geographical location and transition patterns.
  - a single station's traffic seems too chaotic to predict.
  - It is not necessary to predict the check-out/in of each individual station.



Bipartite Station Clustering



Entire Traffic Learning

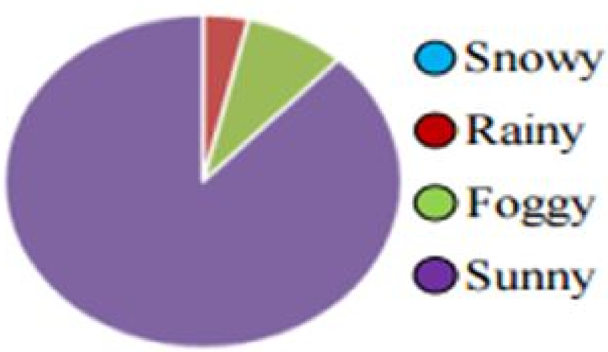
- In hierarchical prediction model, the traffic in the higher level is predicted first.

- Time features
  - the hour of the day
  - the day of the week

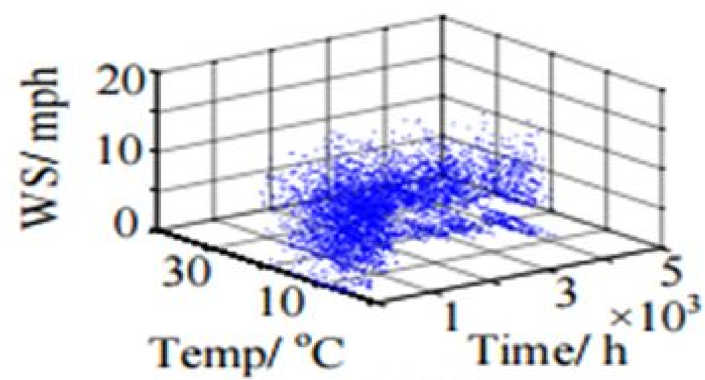
- Meteorology features
  - weather
  - temperature
  - wind speed

Cluster Check-out Proportion Learning

- Insights
  - To allocate entire traffic to each cluster, we predict each cluster's check-out proportion first
  - A multi-similarity-based inference model is proposed.



A) Weather distribution

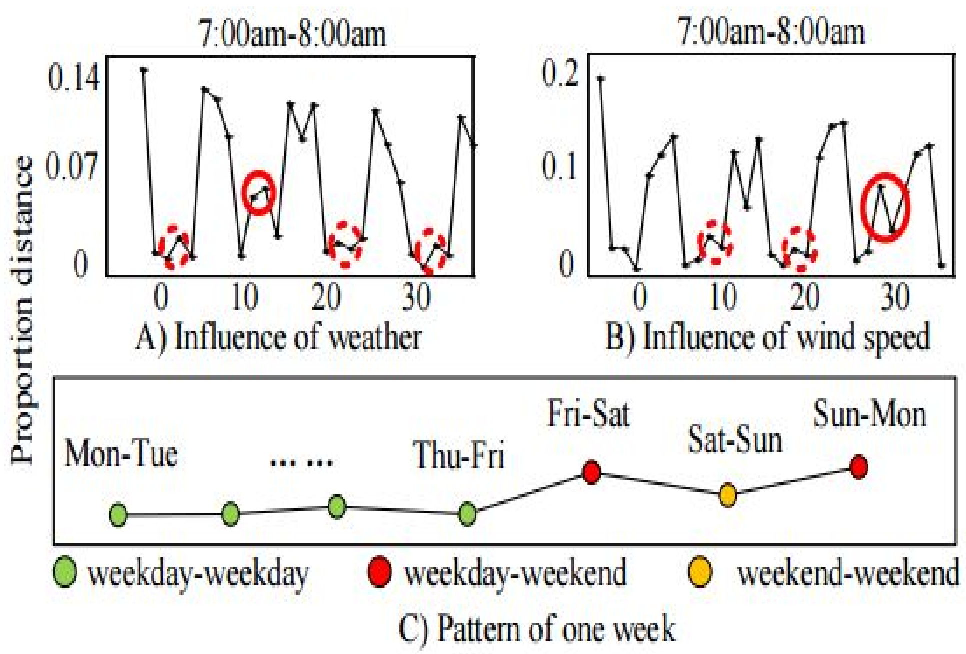


B) Temp & WS sample

- Handle unbalanced meteorology distribution problem
- Guarantee cluster sum of 1 and manage between-cluster difference

Cluster Check-out Proportion Learning

- Insights
  - Our multi-similarity-based inference model integrates 3 similarity functions between features
  - Time similarity
    $\lambda_1 (t_1, t_2)$
  - Weather similarity
    $\lambda_1 (w_1, w_2)$
  - Temperature & wind Speed similarity
    $K ((Pt_1, Vt_1), (Pt_2, Vt_2))$



Cluster Check-out Proportion Learning

- Methodology
  - Assume 1, 2, …, H are the H most recent periods to t
  - Denote corresponding check-out proportions P_1, P_2, …, P_H and P_t
  - Their features are f_1, f_2, …, f_H
  - So that P_t can be predicted by multi-similarity-based inference model

$$\hat{P}_t = \frac{\sum_{i=1}^{H} W(f_i, f_t) \times P_i}{\sum_{i=1}^{H} W(f_i, f_t)}$$

Cluster Check-out Proportion Learning

- Methodology
  - The multi-similarity function, W(fi, fH) is obtained by

$$min_W \sum_{t=H+1}^{T} L(E_t \times P_t, E_t \times \widehat{P}_t)$$

    - T ~ Sample size of historical data
    - Et x Pt, Et x (cap)Pt ~ ground truth and prediction value of check-out across clusters
    - L ~ Loss function used to measure the prediction error
  - The multi-similarity function W has 3 components:

$$W(f_i, f_t) = \lambda_1(i,t) \times \lambda_2(w_i, w_t) \times K((p_i, v_i), (p_t, v_t))$$

Cluster Check-out Proportion Learning

- Methodology
  - Time Similarity
    - Intuitively, check-out proportions corresponding to the same hour or a day are more similar than those corresponding to different hours
    - Additionally, if two proportion vectors both belongs to weekdays or another, the more closed the two days are, the more similar these two vectors should be

$$\lambda_1(t_1, t_2) = 1_{t_1, t_2} \times \rho_1^{\Delta h(t_1, t_2)} \times \rho_2^{\Delta d(t_1, t_2)}$$

$$\Delta h(t_1, t_2) = min\{r(t_1, t_2), 24 - r(t_1, t_2)\}$$

$$r(t_1, t_2) = mod(|t_1 - t_2|, 24)$$

$$\Delta d(t_1, t_2) = \left[\frac{|t_1 - t_2|}{24}\right]$$

Cluster Check-out Proportion Learning

- Methodology
  - Weather Similarity
    - The weather patterns are categorized into four categories: snowy, rainy, foggy and sunny
    - The similarity matrix is symmetric with 6 parameters

	snowy	rainy	foggy	sunny
snowy	1	α_1	α_2	α_3
rainy		1	α_4	α_5
foggy			1	α_6
sunny				1

$$\alpha_1 > \alpha_2 > \alpha_3, \alpha_4 > \alpha_5$$

$$\alpha_6 > \alpha_5 > \alpha_3, \alpha_4 > \alpha_2$$

    - The more different two weather patterns are, the smaller the similarity between them is

Cluster Check-out Proportion Learning

- Methodology
  - Temperature/wind speed domain is continuous, with 'missing' scenarios in historical data
  - 2-D Gaussian Kernel function to measure the similarity between (p_{t1} , v_{t1}) and (p_{t2} , v_{t2})

$$K\left(\left(p_{t_1}, v_{t_1}\right), \left(p_{t_2}, v_{t_2}\right)\right) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\left(\frac{(p_{t_1}-p_{t_2})^2}{\sigma_1^2} + \frac{(v_{t_1}-v_{t_2})^2}{\sigma_2^2}\right)}$$

  - As the prediction errors of successive time periods are not independent, we add an error correction item to the multi-similarity-based inference model.

Cluster Check-out Proportion Learning

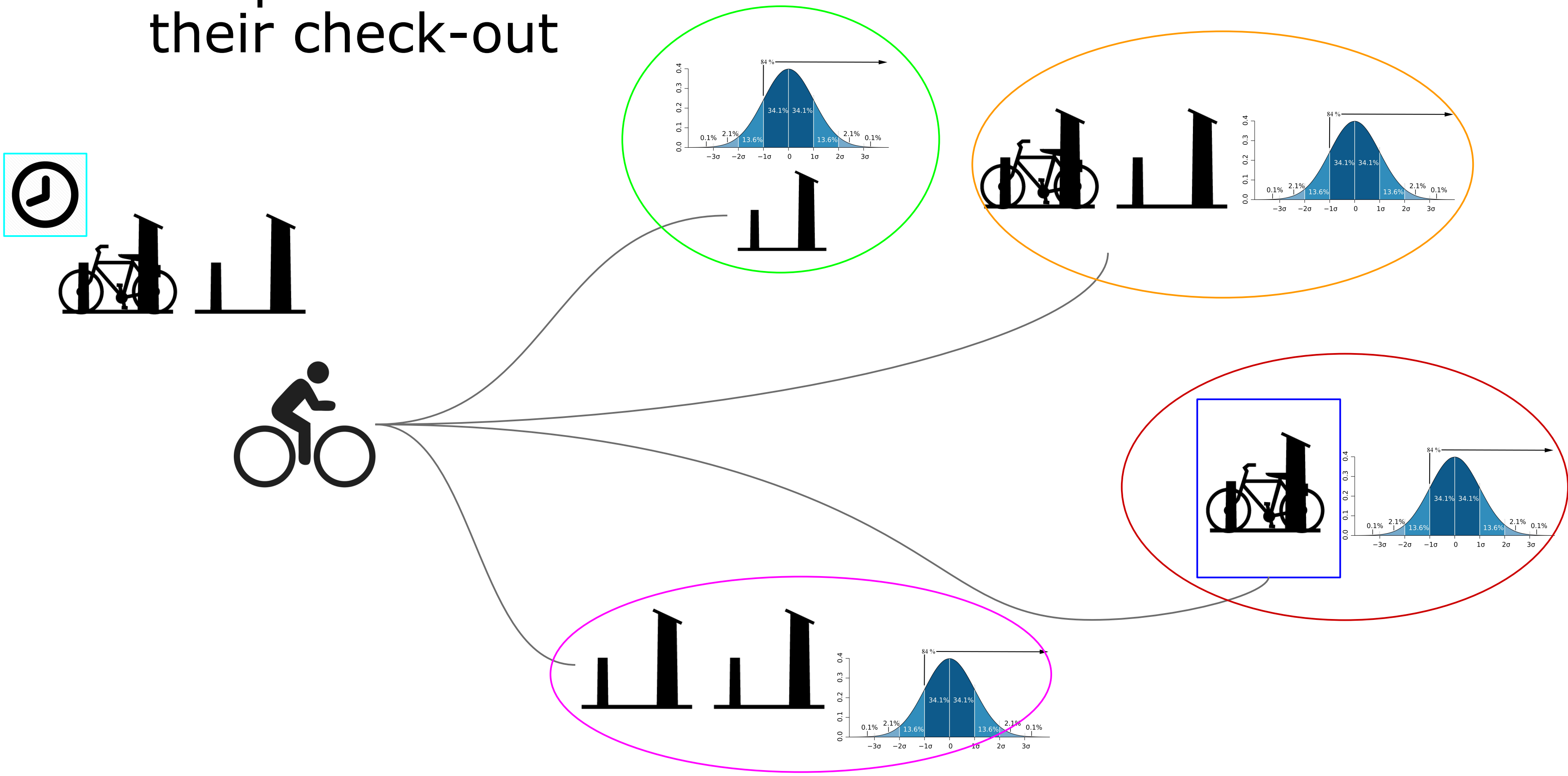
- Methodology
  - The multi-similarity-based model adopted is

$$\hat{P}_t = \frac{\sum_{i=1}^{H} W(f_i, f_t) \times P_i}{\sum_{i=1}^{H} W(f_i, f_t)} + \sum_{j=1}^{J} \psi_j \boxed{e_{t-j}}$$

  - Here, the added items $e_{t-j} = P_{t-j} -$ (cap)$\hat{P}_{t-j}$ are the prediction errors of periods $t - j, j = 1,2, \ldots, J$; J is a threshold of time lag

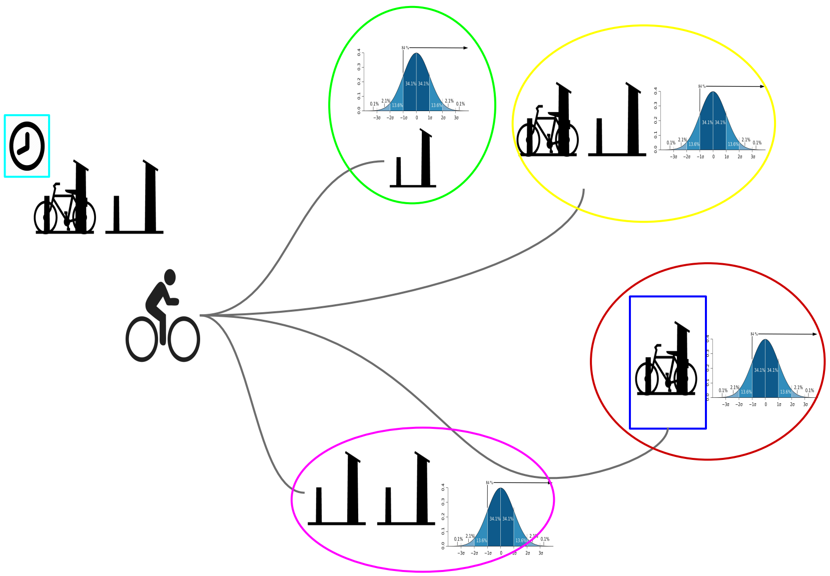
Inter-cluster Transition Learning

- We predict each cluster’s check-in based on their check-out



Inter-cluster Transition Learning

- The inter-cluster transition matrix describe the transition probability between clusters
- Using multi-similarity-based inference model to predict the matrix

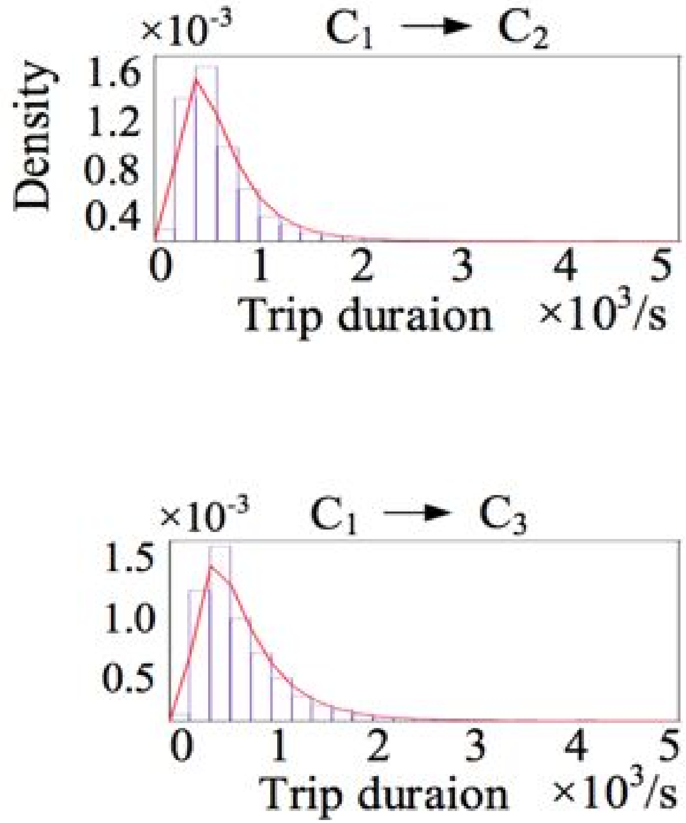


Trip Duration Learning

- In bike traffic, jam is no longer an important factor that affects trip duration
- It is mainly determined by the locations of bike stations
- Duration does not change too much

Trip Duration Learning

- According to NYC's bike data, the trip duration between each pair of cluster
- By maximum likelihood estimation, we obtain symmetric matrix, describing the trip duration between cluster Ci and Cj



Online Prediction Process

- Check-out Inference
  - Entire traffic Prediction Et
  - Check-out proportion prediction Pt
- Calculation
  - Check-out of each cluster Ci is

$$O = Et * Pt$$

Online Prediction Process

- Check-in Inference For Common Scenarios
  - use the same model as calculating check-out
    - Entire traffic Prediction Et
    - Check-in proportion prediction Pt
- Check-in Inference For anomalous Scenarios
  - Update the prediction of target cluster in real time
    For a bike,
      - Original Cluster Ci
      - Check out time
      - Inter-cluster transition matrix and trip duration
    - Get the expectation number of bikes on their way which are going to check in this cluster

Experiments

Data Source:

New York

We use the data of Citi Bike system, which is in NYC, from 1st Apr. to 30th Sep. in 2014 as the bike data. We use the meteorology data of NYC, from 1st, Apr. to 30th, Sep.

D.C

We use the data of Capital Bikeshare system, which is mainly in D.C., from 1st Apr. to 30th Sep. in 2014 as the bike data. we use the meteorology data in D.C., from 1st, Apr. to 30th, Sep., 2014

Baseline & Metric

Methodologies:

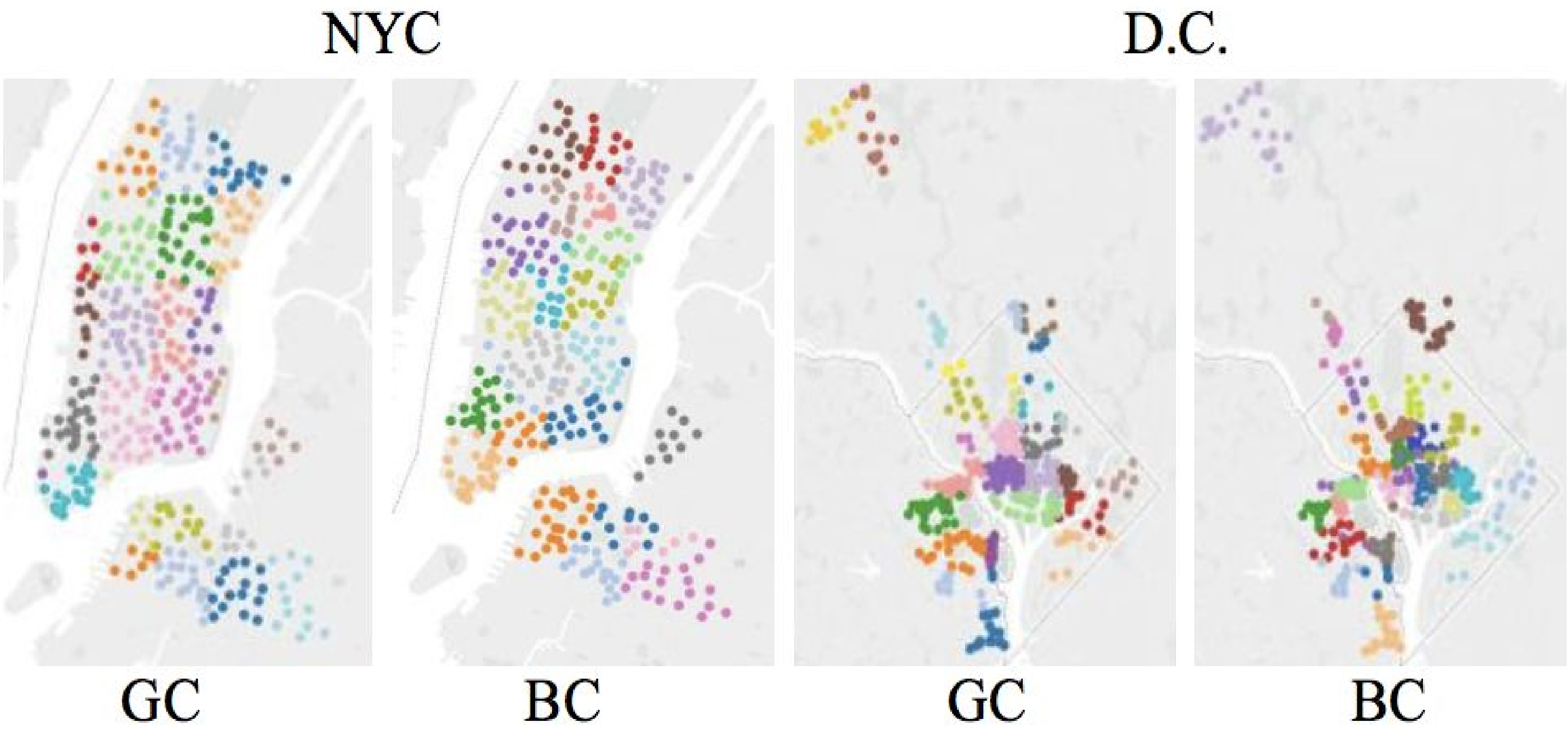
HA, ARMA, GBRT, HP-KNN, GC,

Metric:

RMLSE,ER

Results

Result of clustering



Results (cont.)

Table 3. Prediction error of check-out across clusters

Method	All Hours								Anomalous Hours							
	RMLSE				ER				RMLSE				ER			
	NY		WA		NY		WA		NY		WA		NY		WA	
	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC
HA	0.387	0.372	0.439	0.451	0.353	0.355	0.453	0.489	1.038	1.027	**0.653**	0.715	1.964	1.968	2.111	2.136
ARMA	0.371	0.354	0.413	0.421	0.346	0.346	0.416	0.445	1.114	1.105	0.680	0.722	2.276	2.273	2.245	2.109
GBRT	0.386	0.369	0.423	0.425	0.311	0.314	0.371	0.375	0.647	0.621	0.686	0.670	0.696	0.683	0.830	0.847
HP-KNN	0.377	0.358	0.424	0.410	0.298	0.299	0.364	0.359	0.664	0.642	0.685	0.694	0.692	0.685	0.836	0.838
HP-MSI	0.371	**0.349**	0.421	**0.407**	0.288	**0.282**	0.351	**0.347**	0.646	**0.597**	0.679	**0.664**	0.637	**0.503**	0.794	**0.783**

Table 4. Prediction error of check-in across clusters

Method	All Hours								Anomalous Hours							
	RMLSE				ER				RMLSE				ER			
	NY		WA		NY		WA		NY		WA		NY		WA	
	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC	GC	BC
HA	0.377	0.365	0.435	0.448	0.347	0.352	0.448	0.485	0.954	0.982	0.617	0.672	1.837	1.835	2.201	2.217
ARMA	0.363	0.352	0.409	0.418	0.340	0.344	0.405	0.445	1.025	1.046	0.631	0.700	2.152	2.143	2.123	2.288
GBRT	0.382	0.365	0.420	0.422	0.309	0.309	0.370	0.375	0.624	0.653	0.689	0.701	0.681	0.671	0.834	0.835
HP-KNN	0.375	0.360	0.415	0.411	0.302	0.295	0.367	0.361	0.659	0.647	0.703	0.686	0.694	0.684	0.830	0.830
HP-MSI	0.365	**0.350**	0.408	**0.402**	0.297	**0.290**	0.353	**0.340**	0.646	0.608	0.675	0.660	0.642	0.506	0.810	0.802
P-TD	0.384	0.373	0.425	0.419	0.335	0.302	0.365	0.359	0.626	**0.598**	0.564	**0.558**	0.498	**0.445**	0.802	**0.789**

Conclusion

Our model is better and applicable to different bike-sharing systems

Thank you!