A Real-Time Passenger Flow Estimation and Prediction Method for Urban Bus Transit Systems

Jun Zhang, Dayong Shen, Lai Tu, Fan Zhang, Chengzhong Xu, Yi Wang, Chen Tian, Xiangyang Li, Fellow, IEEE, Benxiong Huang, and Zhengxi Li

SCH Univ.

Dept. of AI and Bigdata

Sunghun Kim

contents

1. Introduction

1. Providing a $\boxed{\text{comfortable travel experience}}$ for passengers is a key business consideration

$\Downarrow$

Effective bus scheduling

- Definition of Passenger Flow
  - ✓ Number of on-board passengers in public transportation services
  - ✓ Varies over time and space

- Effect of Knowing Passenger Flow
  - ✓ Provide insight into the collective human mobility patterns along a route
  - ✓ Guide the operators to allocate and schedule the bus route and timetable dynamically in fine granularity
  - ✓ New opportunities for using the data-driven approaches to fit the demand of passengers

1. Introduction

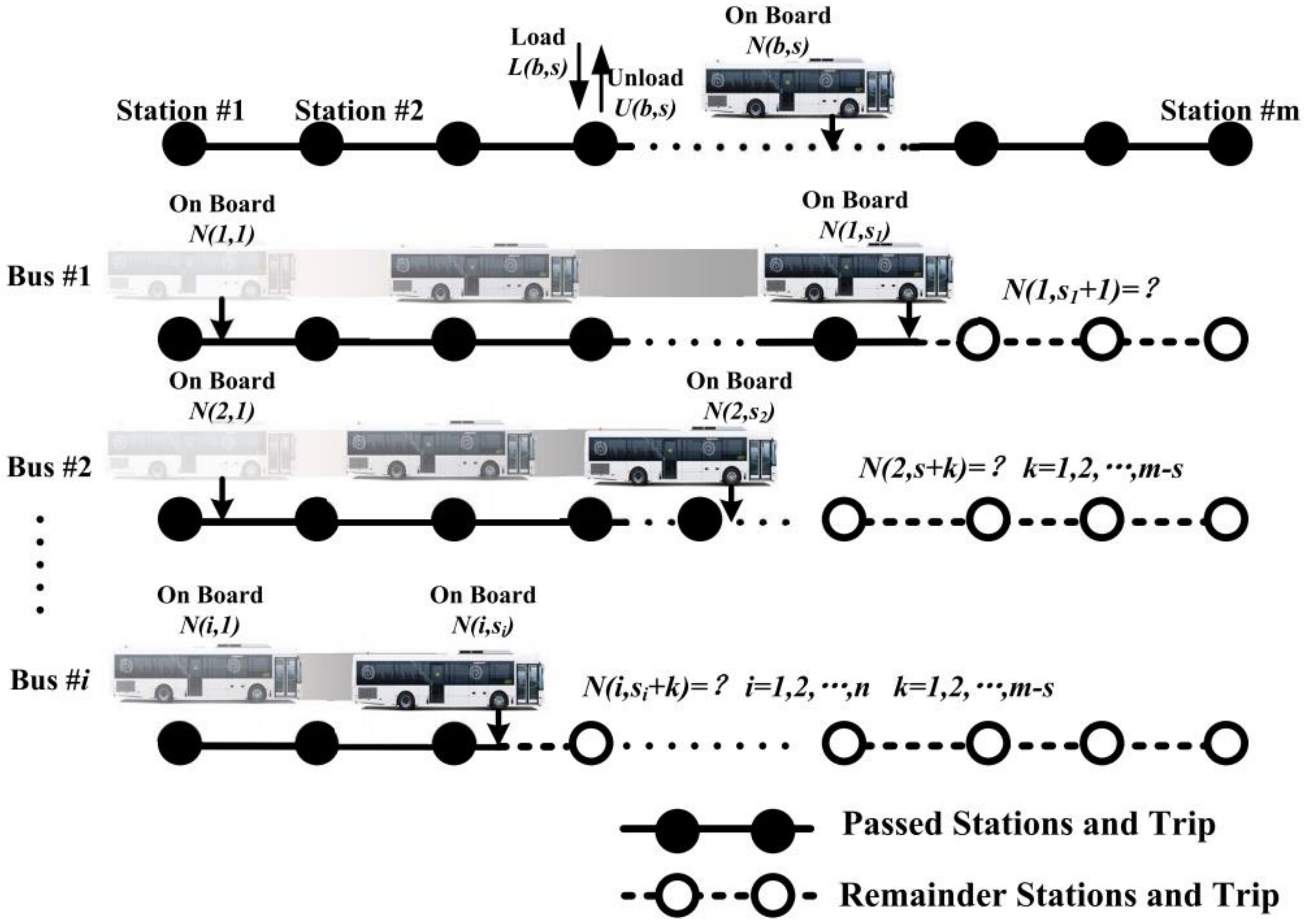
- Bus Transit System (BTS) : Integrated system for operating and managing buses within a city
  - ✓ Manual data-collection efforts are costly and applicable only in small scale

↓

  - ✓ Automatic Fare Collection (AFC) devices : Record payments of riders using smart card
  - ✓ GPS embedded On Board Unit (OBU) : Track the bus location

Estimate and predict the passenger flow of every bus

1. Introduction



1. Introduction

Automatic Fare Collection (AFC) and On-Board Unit (OBU) Data

Q1. How to estimate the number of riders on each bus

Q2. How to predict the number in the remainder of the trip in the near future

Problem 1)

- Bus devices cannot automatically and precisely count the number of the passengers getting on and off the bus
- Impractical to make it widely by human field investigations

Problem 2)

- Passenger's getting off or someone paying by coins cannot be observed directly

Problem 3)

- Due to the uncertainty of people's mobility, challenging to predict the passenger flow of future

1. Introduction

Problem 1)

- Bus devices cannot automatically and precisely count the number of the passengers getting on and off the bus
- Impractical to make it widely by human field investigations

Solution 1)

- Estimate the number of the riders getting on at each station
- Derive the boarding position of a passenger by querying the GPS trace dataset with taping time as key

---

Problem 2)

- Passenger’s getting off or someone paying by coins cannot be observed directly

Solution 2)

- Estimating the alighting stations of passengers based on their historical boarding records.
- Estimate the coin users based on time gap between consecutive smart card user

Problem 3)

- Due to the uncertainty of people's mobility, challenging to predict the passenger flow of future

Solution 3)

- Based on the real-time estimation of the number of passengers on a bus, furtherly predict the number of passengers that will be on the bus upon arrival at its remaining stations.

2. Related work

Short-term transportation forecasting (Short term traffic forecasting / Short-term passenger demand forecasting)

- Parametric
- Non-parametric

Parametric

- Historical average
- Smoothing techniques
- Autoregressive integrated moving average (ARIMA)

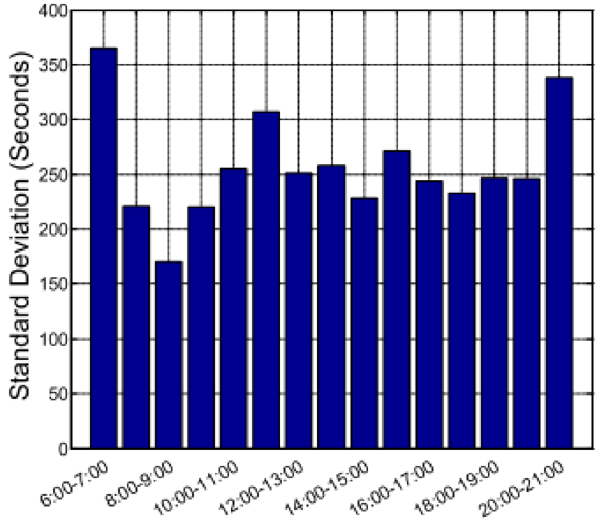
Non-Parametric

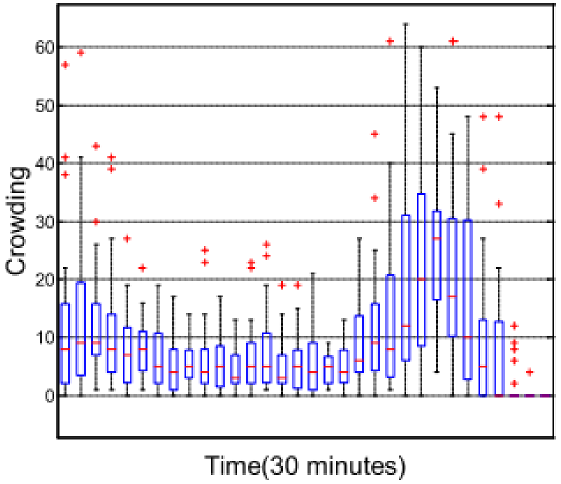
- Neural networks
- Non-parametric regression
- Kalman filtering models
- Gaussian maximum likehood

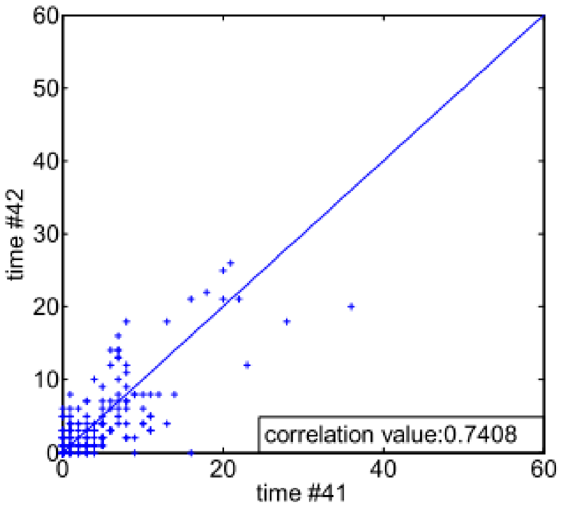
3. Overview

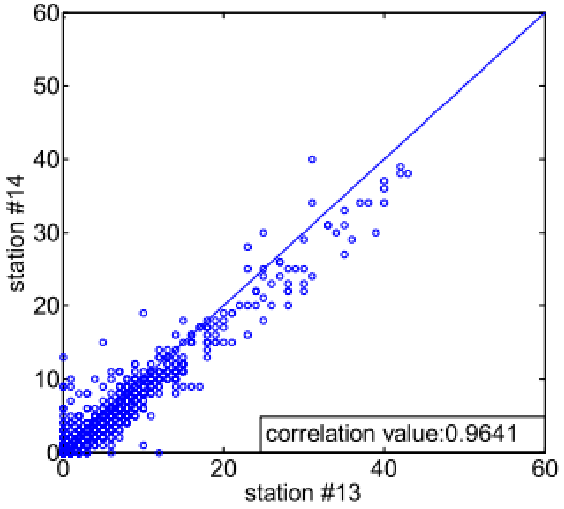
▶ Motivation

- Different periods' standard deviation of a line's arrival time in 15 days
  - ✓ Standard deviation is large → Predicting for that period is difficult, and there is significant uncertainty in the arrival times
- Divide the operating hours of the bus route into 30-minute intervals and investigated the passenger flow
- Correlation between adjacent time slots and stations about passenger's flow









3. Overview

▶ Motivation

- $N(i,j)$: Number of passenger on Bus $\#i$ at Station #j

- $L(i,j)$: Number of passengers boarding the bus
  - ✓ $L_s(i,j)$: Paying by smart card
  - ✓ $L_c(i,j)$: Paying by coin

- $\mathrm{U}(i,j)$: Number of passenger alighting the bus
  - ✓ $U_h(i,j)$: historical trip chain pattern
  - ✓ $U_h(i,j)$: estimated based on a probability model
  - ✓ $U(i,j) = U_h(i,j) + U_p(i,j)$

$$\tilde{\mathbb{N}}(t) = \begin{pmatrix} \tilde{N}_{11} & \tilde{N}_{12} & \dots & \tilde{N}_{1,k} & \hat{N}_{1,k+1} & \dots \\ \tilde{N}_{21} & \tilde{N}_{22} & \dots & \hat{N}_{2,k} & \hat{N}_{2,k+1} & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \tilde{N}_{i,1} & \dots & \tilde{N}_{i,j} & \hat{N}_{i,j+1} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \end{pmatrix}_{B \times S} \quad (1)$$

3. Overview

▶ Motivation

Data set

- SZT card Data : Every smartcard's users boarding data
- BUS Route Map : Passenger boarding station
- BUS GPS Data : GPS coordinates of every bus every 20-40 seconds

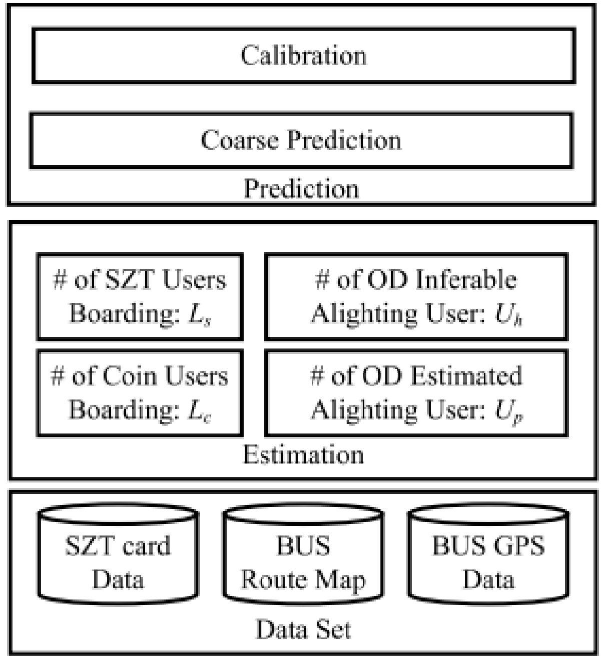
GPS dataset		Smart card dataset	
Content	Remarks	Content	Remarks
OBU ID	On Board Unit ID	Serial number	It is unique for different records
Vehicle ID	Vehicle registration ID.	Card ID	The number of SZT smart card
Line ID	The line number of the bus	FCD ID	Fare Collection Device ID
Position state	Located or un-located	Transaction type	Metro: Get on/off, Bus: Get on
Longitude	The longitude of the vehicle	Time	The time of tapping card
Latitude	The latitude of the vehicle	Name	Metro: station name, Bus: line name
Time	The time of obtaining the location	Vehicle ID	Vehicle registration ID

Estimation

- Estimate the numbers of passengers on buses by estimating the passengers ODs(Origin-Destination)

Prediction

- Build a model to predict the passenger flow



4. Estimation

▶ Time Synchronization and Boarding Event Localization

- No location field in the AFC record → Match the time stamps in AFC records and OBU records

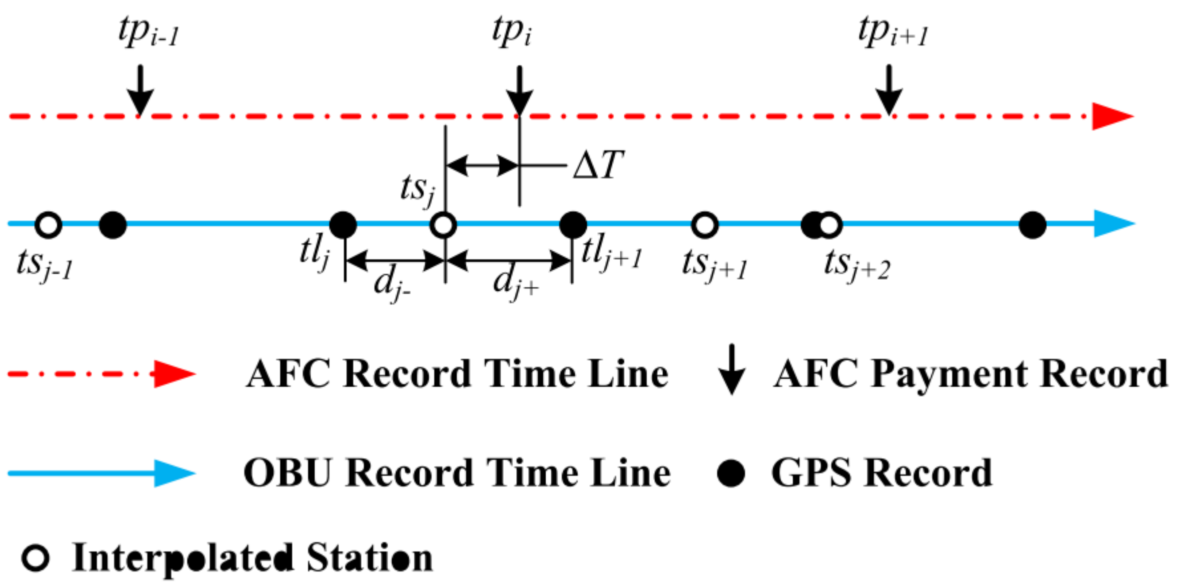
Problem

- AFC device and OBU device work independently → Different exist time
- GPS location sampled every 20-40 seconds

→ Interpolate

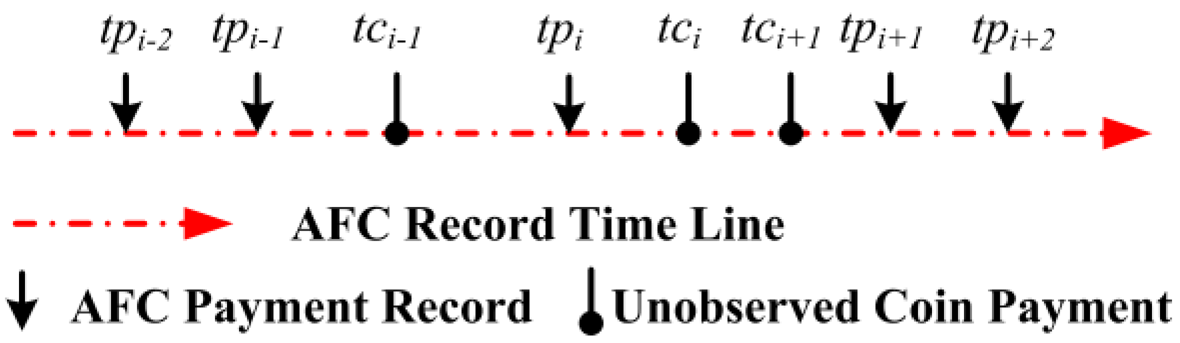
$$ts_j = tl_j + \frac{(tl_{j+1} - tl_j) \times d_{j-}}{d_{j-} + d_{j+}}$$

$$\Delta T = \arg\min_{\Delta T} \sum_i \min_j |tp_i - \Delta T - ts_j|$$



4. Estimation

▶ Estimation of $L(i,j)$



- Estimating the total number of the boarding passengers $L(i,j)$
  - ✓ $L(i,j) = L_s(i,j) + L_c(i,j)$
  - ✓ $L_c(i,j) \leftarrow$ Time gap between two consecutive smart card payment events
- Assumption
  - ✓ Time gap between two consecutive smart card payment events is lager → Coin payment occur

* Poisson Process

Poisson process

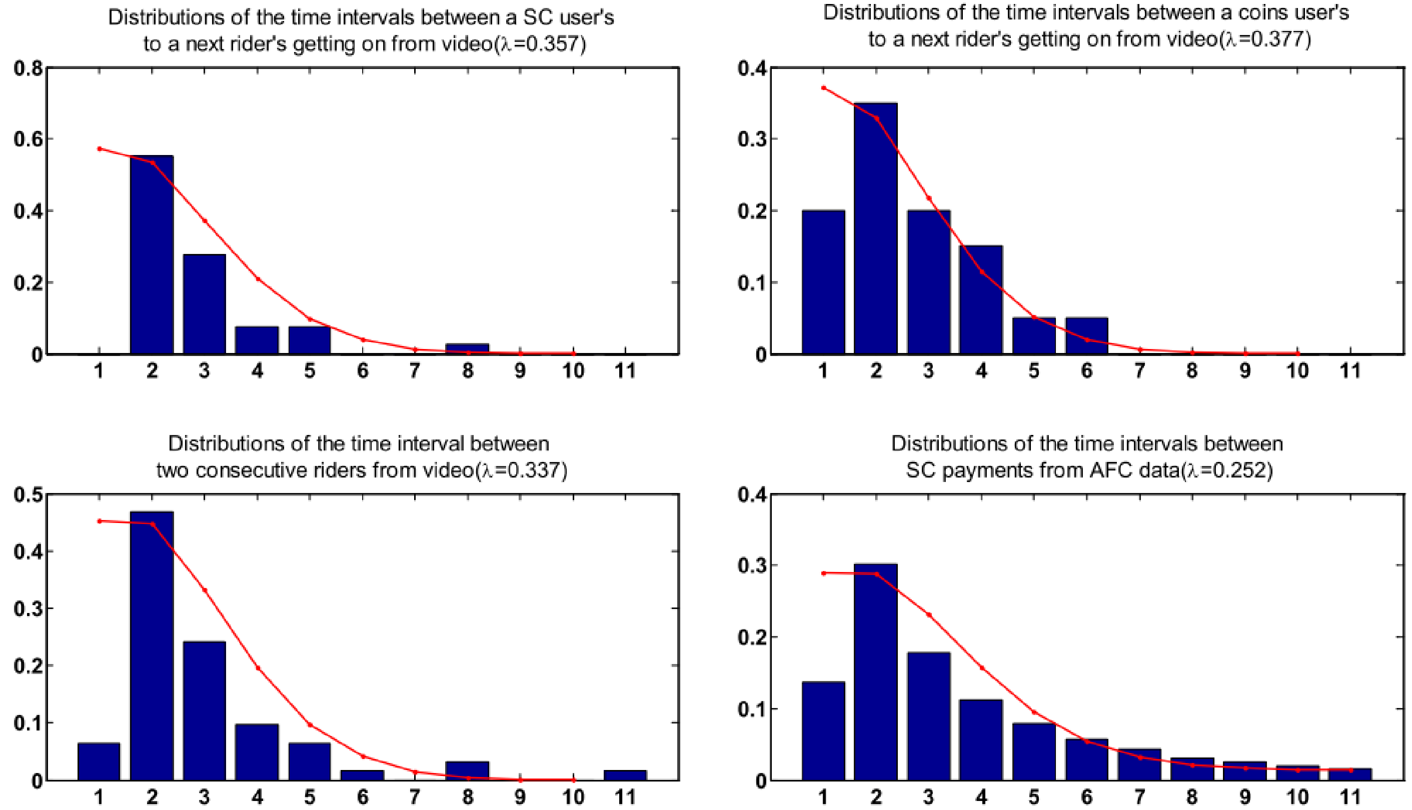
- Discrete probability distribution that expresses how many times an event will occur in unit time and unit space

- Population parameter
  - ✓ Average number of occurrences in unit time or unit space

- Prerequisites
  - ✓ Independent Events : The events in a Poisson process are independent
  - ✓ Constant Rate of Occurrence : The average rate is constant and denoted by λ
  - ✓ Discrete Occurrences : Events in a Poisson process occur discretely
  - ✓ Single Events at a Time : Only one event occurs at a time during a given time or space interval

4. Estimation

▶ Estimation of $L(i,j)$



- λ = 3 → Averagely a passenger takes 3 seconds to get on the bus

$$L_c(i,j) = \sum_{k=1}^{L_S(i,j)-1} \arg\max_n P(n; \lambda(tp_{k+1} - tp_k))$$

4. Estimation

▶ Estimation of $\mathrm{U}(i,j)$

Neither smart card or coin users need extra operations before getting off

⬇

Type 1) $U_h(i,j)$: Number of smart card users that show strong regularity in historical records

Type 2) $U_p(i,j)$: Number of Smart card users except Type 1. and coin users

Type 3) $U_t(i,j)$: Number of Smart card users taking transit ride after alighting current bus

▶ Estimation of $U(i, j)$

Type 1) Estimation Based on Historical Regularity

- Extract trip tuples of $< R_{ID}, O_i, T_i >$
  - ✓ R_{ID} : Smart card ID, exclude coin users
  - ✓ O_i : the origin of ith trip
  - ✓ T_i : the time of the rider paying his ith trip

1. Given an identifiable tuple $< R_{ID}, O_i >$, if O_{i+1} has a larger probability than P_{th} to be one certain station s

   → Regular trip

2. Make estimation of the destination of new trip of R_{ID}

$$D_i \approx arg \max_{O_{i+1}} \{P | P = \mathbb{P}(O_{i+1} | R_{ID}, O_i), P \geq P_{th}\}$$

4. Estimation

▶ Estimation of $\mathrm{U}(i,j)$

Type 2) Dispatch Based on Common Distribution Assumption

- Not enough samples in the historical dataset about smart card and coin users
- The distribution of the destination of a trip is independent to whether the trip is a regular trip

$$\mathbb{P}(D_i | R_{ID}, O_i) \perp \mathbb{P}(< R_{ID}, O_i, T_i > is\ a\ regular\ trip)$$

1. Calculate the empirical distribution of D on condition of O from the observable OD of regular trip in historical data
2. Dispatch the non-regular trips

$$\mathbb{P}(D|O) \quad \longrightarrow \quad U_p(i,j) = \sum_{k=1}^{j-1} L_p(i,k)\mathbb{P}(D=j|O=k)$$

4. Estimation

▶ Estimation of $\mathrm{U}(i,j)$

Type 3) Estimation Amendment Based on Transit Payment

- R_{ID}'s regular trip $< R_{ID}, O_i, T_i >$ $\longrightarrow$ $\widehat{D_i}$ at T_d (Based on historical travel regularity)
- Another payment record (transit occur) : $< R_{ID}, O_{i+1}, T_{i+1} >$ $\longrightarrow$ D_i $(D_i \neq \widehat{D_i})$

- Modify estimates based on observed facts

$$\Downarrow$$

$$U(i,j) = Amendent(U_h) + Amendent(U_P)$$

5. Prediction

Coarse Prediction Based on Historical Data

$$S = \frac{< \tilde{N}, N >}{\sqrt{< \tilde{N}, \tilde{N} >} * \sqrt{< N, N >}}$$

- $\tilde{N}$: Real – time estimation
- N : Passenger flow of one day in historical data
- Operation <,> : Inner product
- S : Similarity of matrix $\tilde{N}$ and matrix N

Assumption

- If current passenger flow pattern is similar with the history, the following passenger flow may change similarly as the pattern on that day

$\{x_1, x_2, x_3, \cdots, x_n\}$: Current passenger flow estimation

$\{u_1, u_2, u_3, \cdots, u_n\}$: Passenger flow patterns from historical data similar with the estimation

- Has similarity during 1~n → output : u_{n+1}

Calibration Based on Extended Kalman Filter

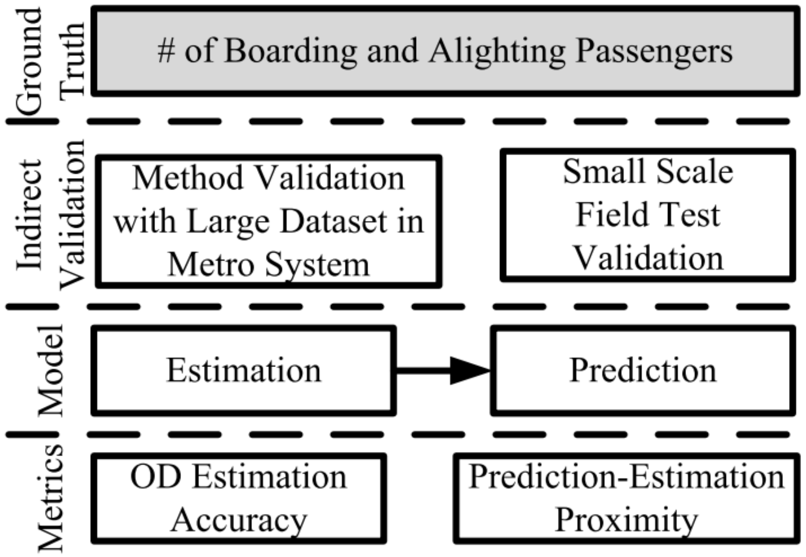
$$f(x_{k-1}, u_{k-1}) = x_{k-1} + \frac{u_k - u_{k-1}}{u_{k-1} - u_{k-2}}(x_{k-1} - x_{k-2}) \qquad h(x_k) = x_k$$

6. Evaluation

▶ The Method and Experiment for Evaluation

Overall performance of the system → The accuracy of estimation and prediction

- Estimation of passenger flow is based on the number of boarding and alighting passengers
  - ✓ Accuracy of OD estimations
- Evaluating the model
  - ✓ Compare the predicted value to the estimated value in the future



↓

- Apply the estimation model in the metro system

- Conduct a small field experiment to evaluate OD estimation for trips where OD cannot be inferred (ex. coin user)

6. Evaluation

▶ Evaluation of the Estimation

The Proportion of the Trip-Chain Inferable ODs

- Using the AFC data in 6 days

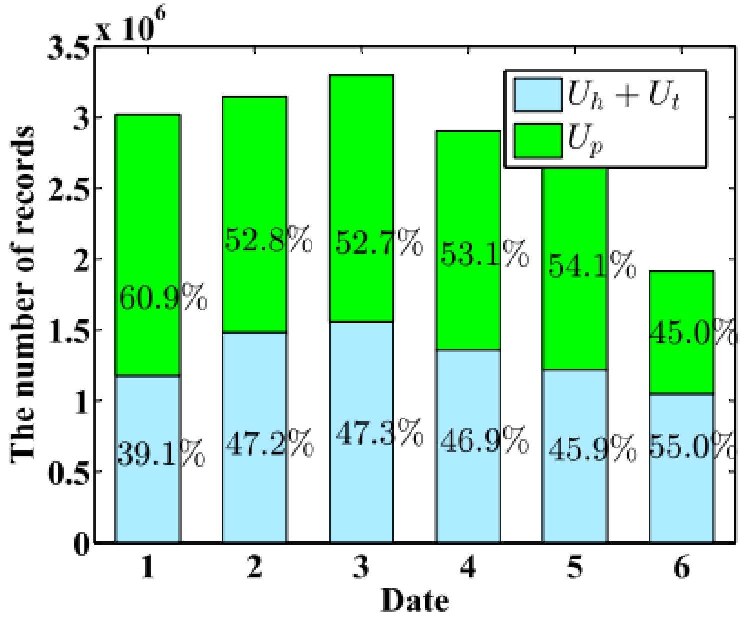
The Accuracy of Destination Estimation

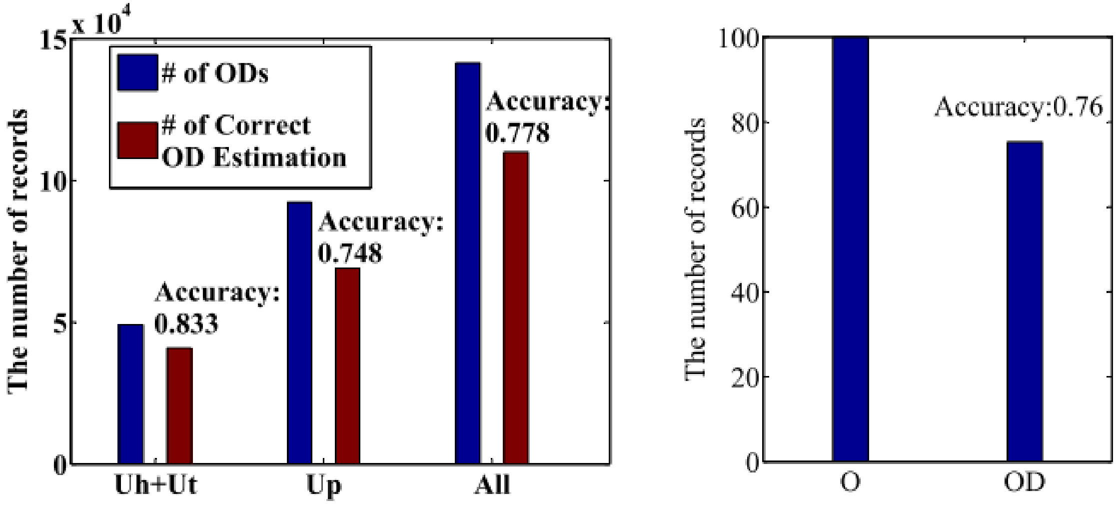
1) Large Scale Metro Data Validation

   ✓ 1.56 million trip samples

2) Small Scale Field Experiments

   ✓ 100 trip of about 20 participants





The Accuracy of the D Estimation in Metro System and Field Experiment

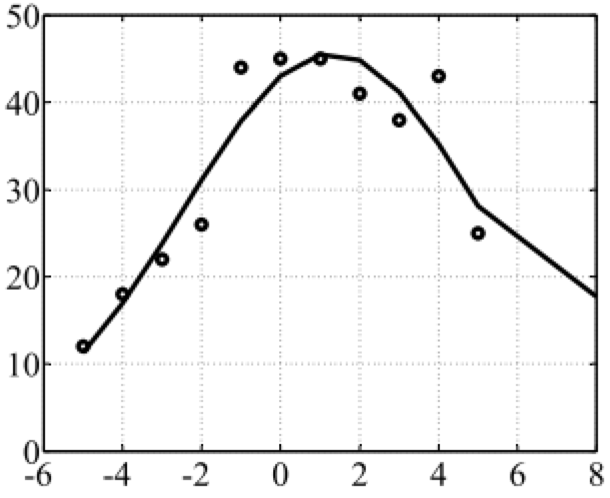
6. Evaluation

▶ Evaluation of the prediction

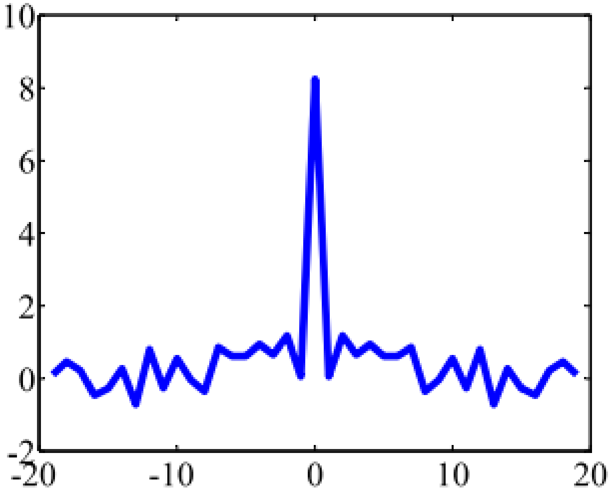
Error Distribution Analysis

- Distribution of the Observation Noise
  - ✓ Error between coarse prediction and true estimation value
  - ✓ Approximately obeys Gaussian distribution

    → The Extended Kalman Filter is handling the noise effectively



- Autocorrelation of the Error Sequence
  - ✓ Autocorrelation values are relatively small except the value at zero
  - ✓ Q-Test result shows confidence level of 85%



White noise characteristics

- Errors are random, unpredictable, and temporally uncorrelated
- Model and filter are not influenced by time-dependent patterns of errors and can effectively handle the noise.

▶ Evaluation of the Prediction

Prediction Results Analysis

- $ARIMA\ (p, d, q)$
  - ✓ p (Autoregressive order) : Number of past values in the time series data that the model takes into consideration
  - ✓ d (Degree of differencing) : Determining the number of times to eliminate patterns. (ex. Trends, Seasonality)
  - ✓ q (Moving-average order) : Remembering past errors to consider them when predicting current values

- Linear Regression
  - ✓ Using different periods of historical data to train the model
  - ✓ Each period has a linear regression model → Create a prediction model specialized for that particular period

6. Evaluation

▶ Evaluation of the Prediction

- Red area : high number of samples with that error value
- Blue area : lower number of samples
- Samples located above the diagonal line indicate that the 2RTP model has smaller prediction errors compared to the baseline models.

TABLE II

THE RMSE OF DIFFERENT MODELS

Model	RMSE
$2RTP$	1.2845
ARIMA(1,1,1)	3.9402
ARIMA(2,1,1)	4.148
ARIMA(2,1,2)	4.9256
Linear Regression	3.1323

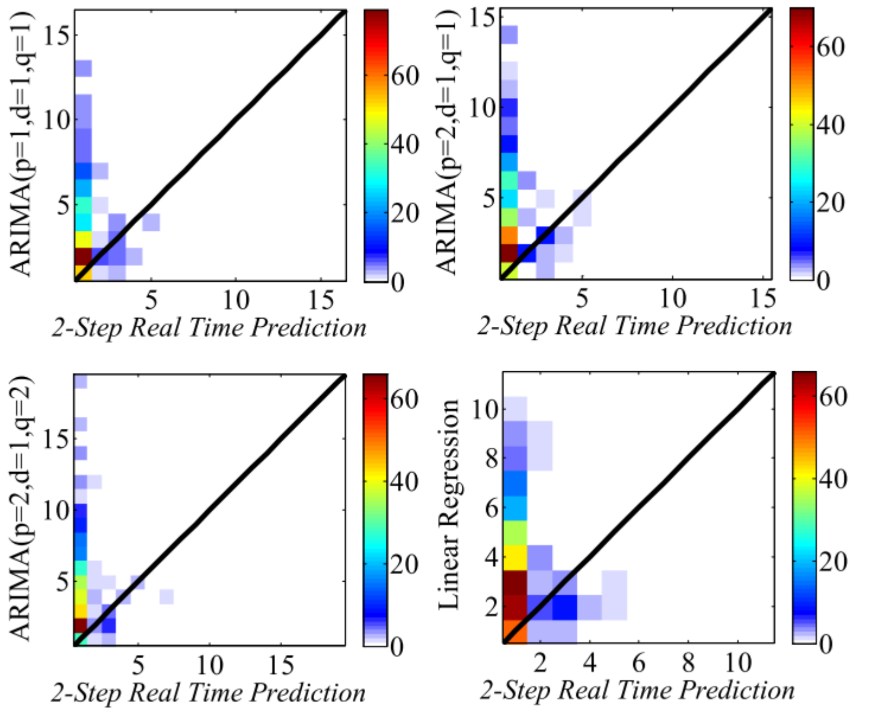


Fig. 14. The Prediction Error Comparison between $2RTP$ and Baseline Models.

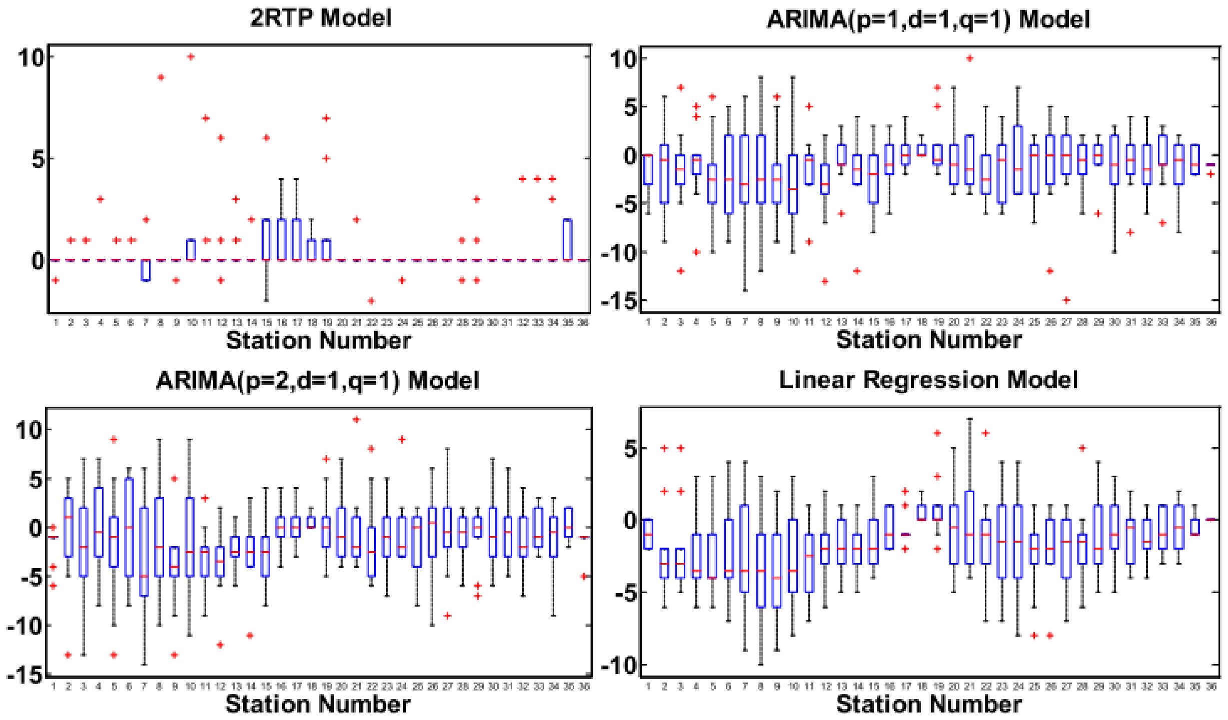


Fig. 15. The Prediction Errors in Different Stations.

6. Evaluation

▶ Evaluation of the Prediction

TABLE III

THE RESULT OF k-MEANS

Crowding Rate	Number of Passengers	Description
1	0-3	Empty
2	3-6	Medium
3	6-14	Full
4	14-26	Crowded
5	More than 26	Very Crowded

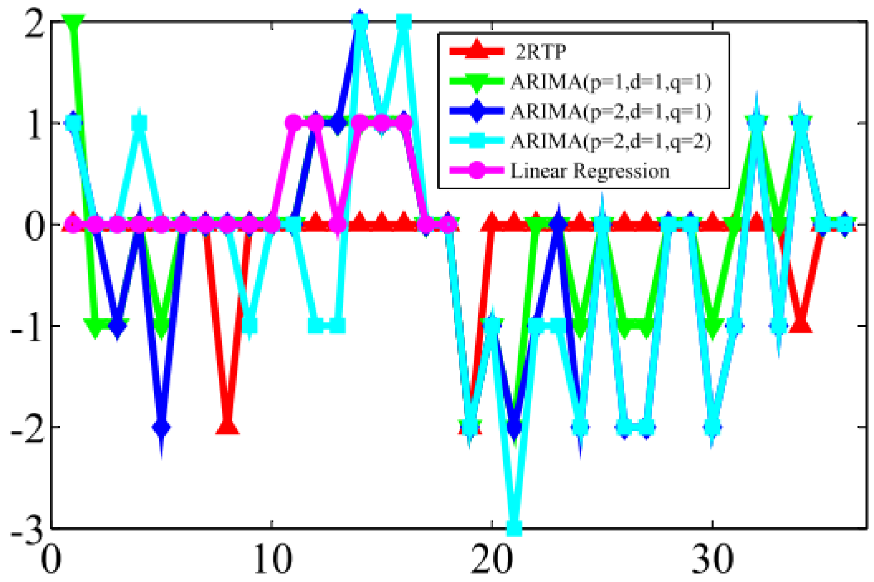


Fig. 16. The Crowding Rate Prediction Errors in Different Stations.

7. Conclusion

- Data : GPS trace and Smart card payment records
- Purpose : Estimating the passenger flow by deriving the origin and destination of passenger
- Comparing with existing prediction model and proposed 2RTP

Outperform in most time and station

Thank You