Real-time Estimation of Urban Rail Transit Passenger Flow Status Based on Multi-source Data

Zhengping Tao 1 and Jinjin Tang 1*

1 School of Transportation and Transportation, Beijing Jiaotong University, No. 3 Shangyuan Village, Haidian District, Beijing, 100044, China

*Corresponding author’s e-mail address: jjtang@bjtu.edu.cn

Abstract. The AFC data uploaded in real time for urban rail transit is incomplete and delayed. In order to improve the dynamic management and control level of rail transit, this paper describes the real-time state estimation process of passenger flow and its key issues, and analyses the relationship between real-time uploaded AFC data and mobile phone signal data, establishing a multi-source data fusion model based on gradient descent method; On this basis, combined with the multi-source data after fusion as the basis of real-time estimation, a real-time passenger flow estimation model based on Kalman filter is established. Practice shows that the fusion of multi-source data improves the accuracy of estimating basic data, the error of the real-time passenger flow estimation model based on Kalman filter is less than 10%, and the accuracy of the estimation model is good. The needs of the urban rail transit operation management department to grasp the real-time passenger flow distribution status of the network are met.

1. Introduction
With the development of urban rail transit network in China, the passenger flow of rail transit has grown rapidly, and it has put forward higher requirements for real-time control of the dynamic distribution of passenger flow in the network and ensuring safe operation. Among them, real-time estimation of rail transit is a very important task. It is of great significance for the operation management department to grasp the dynamic distribution of the passenger flow of the network in real time and take corresponding measures. However, the passenger flow data uploaded by the AFC system is delayed and incomplete. The operation management department cannot timely obtain the dynamic distribution of passenger flow in the rail transit network. In this context, it is of great significance to real-time passenger flow state estimation combine with passenger flow data and mobile phone signal data uploaded by AFC.

In the past, the research on AFC data mainly includes two aspects: First, the research on the overall situation of AFC data[1][4], such as the direct extraction of inbound and outbound traffic from AFC data; the second is to obtain the travel habits of specific passengers from AFC data[5][6]. All in all, the current urban rail transit operation department lacks deep excavation of AFC data, and the understanding of the real-time passenger flow and its distribution is relatively lagging and lacks predictability.

2. Real-time Estimation and Analysis of Passenger Flow Status of Urban Rail Transit

2.1. Real time estimation process of passenger flow condition
Real-time estimation of passenger flow status of urban rail transit uses real-time uploaded AFC data and mobile phone signal data as the estimated basic data, analyse the relationship between the two by gradient descent method and generates passenger flow fusion data. Based on this, establish a real-time estimation model of passenger flow state based on Kalman filter.

Figure 1. Real-time estimation process of passenger flow state based on multi-source data.

2.2. Analysis of real-time passenger flow estimation problem

In the real-time estimation process of passenger flow state, it is necessary to combine the characteristics of urban rail transit to sort out the key problems and lay a foundation for constructing a real-time estimation model of passenger flow state. The key issues of real-time estimation of urban rail transit passenger flow state mainly include the following aspects:

(1) Selection of state variables. The passenger travel volume in the urban rail transit network has a similar travel rules for a certain period of time, the average value of the station's historical inbound quantities directly reflects the average travel level of the station, which can more intuitively reflect the trend of station passenger flow. Therefore, it is widely used in the selection of state variables of real-time estimation model of passenger flow.

(2) Determination of inbound quantities threshold. In order to verify whether the estimated inbound data is reasonable, the single-sample K-S test can be used to judge the distribution of the inbound quantities.

Assume that the inbound quantities obey the normal distribution and take the significance level to 0.05. Use $u - 3\alpha$, $u + 3\alpha$ to calculate the lower and upper limits of the thresholds of the same historical period for each station and each time period.
Among them, \( m_i \) is the inbound quantities of station \( i \) in the same period of history; \( u + 3\alpha \) is upper limits of the thresholds; \( u - 3\alpha \) is lower limits of the thresholds.

3. Real-time estimation model of passenger flow state in urban rail transit

Based on the uploaded AFC data and mobile phone signal data, the real-time estimation model of passenger flow state solves the fusion value of multi-source data by gradient descent method, and uses this as the basic data for real-time estimation of passenger flow state. Secondly, using Kalman filter method Perform real-time estimation of passenger flow.

3.1. Basic flow relationship construction

In order to better express the real-time estimation process of urban rail transit passenger flow state, it is necessary to establish a topology diagram of the network traffic relationship. Considering the relationship of closed network traffic conservation, we can get the relational expression between each indicator.

\[
E_i(t) = \sum_{j=1,j\neq i}^{n} f_{ij}(t) \quad \forall i, j \in V, t \in T
\]

Where \( E_i(t) \) is the inbound quantities of station \( i \) at time period \( t \); \( f_{ij}(t) \) indicates the passenger flow from station \( i \) to station \( j \) at time period \( t \); \( T \) represents the real-time estimated time period, \( T = \{1,2,\cdots,t',\cdots,t\} \).

The relational expression of the inbound amount \( E_i \) is as follows:

\[
E_i(t) = \sum_{j=1,j\neq i}^{n} \sum_{t'=0}^{t} E_i(t' - t') R_{ij}(t' - t') \theta_j(t') \quad i \neq j, \forall i, j \in V, t', t', t \in T
\]
period of $t' - t'$, the proportion of passengers who enter the station by station $i$ and exit from station $j$ as a percentage of the total number of stations in station $i$; $\theta_i(t')$ represents the proportion of passengers departing from station $i$ in time period $t$ arriving at station $j$ during time period $t'$. 

3.2. Multi-source data fusion model establishment and solution

Combining AFC data and mobile phone signaling data, generating a data fusion function based on the Logit model; secondly, using the gradient descent method to move from same or opposite direction of the gradient (the step size of each movement is $\lambda$), the iteratively calculated value When the specified error is reached, it means that the optimal value has been reached at this time.

Firstly, combined with AFC data and mobile phone signaling data, a data fusion function is generated based on the Logit model. The Logit model is actually a model for determining the probability that different types of data are selected. The random term $\epsilon_{n}$ is a Gumbel variable, and it is subject to independent distribution, and the selection probability of multi-source data under different weights is calculated.

$$
\hat{\partial}_i = \frac{\exp(-\theta_i x_{i}^{all})}{\sum_i \exp(-\theta_i x_{i}^{all})}
$$

(5)

$\hat{\partial}_i$ indicates the probability that multi-source data $i$ is selected; indicates the adjustment factor of multi-source data $i$; $x_{i}^{all}$ indicates the confidence level of multi-source data $i$.

The following formula is the inbound quantity fusion function:

$$
P(x, y) = F(\alpha) \frac{\exp(-\theta_i x_{i}^{all})}{\sum_i \exp(-\theta_i x_{i}^{all})} + F(\beta) \frac{\exp(-\theta_j y_{j}^{all})}{\sum_j \exp(-\theta_j y_{j}^{all})} = 1
$$

(6)

Secondly, the gradient descent method is used to move from same or opposite direction of the gradient. Its iteration formula is:

$$
x^{(k+1)} = x^k + \lambda d^{(k)}
$$

(7)

Where $d^{(k)}$ is the search direction starting from $x^k$, taking the steepest descending direction of point $x^k$, that is: $d = -\nabla P(x, y) / \|\nabla P(x, y)\|$

$$
\nabla P(x, y) = \left\{ \begin{array}{l} -F(\alpha)(\exp(-\theta_i x_{i}^{all})) \frac{\exp(-\theta_i x_{i}^{all})}{\sum_i \exp(-\theta_i x_{i}^{all})} \\ -F(\beta)(\exp(-\theta_j y_{j}^{all})) \frac{\exp(-\theta_j y_{j}^{all})}{\sum_j \exp(-\theta_j y_{j}^{all})} \end{array} \right\}
$$

(8)

$\lambda$ is the step size in which $\lambda$ starts searching in the direction $d^{(k)}$, which satisfies:

$$
P(x^{(k)} + \lambda d^{(k)}) = \min_{\lambda \geq 0} P(x^{(k)} + \lambda d^{(k)})
$$

(9)

That is:

$$
\min_{\lambda \geq 0} P(x^{(k)} + \lambda d^{(k)}) = P\left( x^{(k)} - \hat{\lambda} \left\{ \begin{array}{l} -F(\alpha)(\exp(-\theta_i x_{i}^{all})) \frac{\exp(-\theta_i x_{i}^{all})}{\sum_i \exp(-\theta_i x_{i}^{all})} \\ -F(\beta)(\exp(-\theta_j y_{j}^{all})) \frac{\exp(-\theta_j y_{j}^{all})}{\sum_j \exp(-\theta_j y_{j}^{all})} \end{array} \right\} \right)
$$

(10)

The $\|d^{(k)}\|$ of the fusion value $P^*(x, y)$ calculated by the iteration (11) is less than the specified error (this is taken as 5%), which is the desired fusion value. That is, the inbound quantity. If there is a
deviation from the actual value, reverse adjustment \(\exp(-\theta y^{all}_j)(\sum_j(\exp(-\theta y^{all}_j)))^{-1}\) or \(\exp(-\theta x^{all}_i)(\sum_i(\exp(-\theta x^{all}_i)))^{-1}\) until it is close to the actual inbound quantity.

3.3. Establishment and solution of real-time estimation model for passenger flow state

The core of the real-time estimation model of passenger flow state is to establish the state transition equation and the observation equation. In this paper, the historical mean value of the inbound quantity is taken as the state variable. The following describes the construction process of the state transition equation and the observation equation.

Figure 3. Kalman filtering method real-time estimation process.

(1) Construct a state transition equation. Based on the similarity characteristics of the passenger flow distribution in the network, this paper takes the historical mean value of the inbound quantity as the state variable and establishes the state transition equation as follows:

\[
X(t) = A(t)X(t-1) + B(t)U(t) + W(t)
\]  

(11)

In the formula, \(X(t)\) and \(X(t-1)\) are the system state variables at time \(t\) and \(t-1\), respectively, and \(X(t)\) is an \(N_s \times 1\) dimensional matrix composed of \(E_i(t)\), where \(X(t) = [E_i(t)]_{N_s \times 1}; N_s\) is the number of stations in the network; \(A(t), B(t)\) represents the system state transition matrix; \(U(t)\) represents the historical state vector of the period \(t\), which is an \(N_s \times 1\) dimensional matrix composed of \(E_i(t); W(t)\) is a Gaussian white noise matrix.

(2) Construct an observation equation. In this paper, the historical average inbound quantity is used to replace the inbound quantity in the future multiple time periods involved in the constraint equation.

\[
Z(t) = H(t)E_i(t) + V(t)
\]  

(12)

In the formula, \(Z(t)\) is the inbound observation vector of system time period \(t\), which is an \(n \times 1\) dimensional matrix composed of \(E_i(t); H(t)\) is the observation matrix of the system time period \(t\), which is the \(n \times N_s\) dimensional matrix determined by \(E_i(t) = \sum_{j=1}^n f_{ij}(t); E_i(t)\) is the \(N_s \times 1\) dimensional matrix composed of the mean value of the historical inbound quantity of the time period \(t\). \(V(t)\) is a Gaussian white noise matrix.
Based on the constructed state transition equation and observation equation, the iterative recursive method is used to realize the real-time update of the inbound quantity matrix. The Kalman filtering method mainly includes a 5-step basic iterative recursive step\cite{7}. The algorithm formula is as follows:

\[
\begin{align*}
\hat{X}^{-}(t) &= A(t)\hat{X}^{+}(t-1) + B(t)U(t) + W(t) \\
\hat{P}^{-}(t) &= A(t)\hat{P}^{+}(t-1)A^{T}(t) + Q(t) \\
K(t) &= \hat{P}^{-}(t)H^{T}(t)\left[ H(t)\hat{P}^{-}(t)H^{T}(t) + R(t) \right]^{-1} \\
\hat{X}^{+}(t) &= X^{-}(t) + K(t)[Z(t) - HB^{-}(t)] \\
\hat{P}^{+}(t) &= \hat{P}^{-}(t) - K(t)H(t)\hat{P}^{-}(t)
\end{align*}
\]

(13)

Where \( \hat{X}^{-}(t), \hat{X}^{+}(t) \) are the a priori estimates and posterior estimates of the system state variables in time period \( t \); \( \hat{P}^{-}(t) \) and \( \hat{P}^{+}(t) \) is the a priori estimation error variance and the a posteriori estimation error variance of the system state variables in the time period \( t \); \( K(t) \) is the Kalman filter factor.

4. Analysis of Chengdu Metro Cases

This article uses the Chengdu Metro (6 lines, 136 stations) AFC data uploaded in real time on the morning peak of 08:00-08:30 on September 3, 2018, and the inbound quantities on the morning peak of 08:00-09:00 statistics by mobile phone signal as an example, real-time estimation of the 5-minute granularity of inbound quantities in station. And uses C# language to calculate through Visual Studio 2017, and finally analyses the accuracy of passenger flow index estimated by Kalman filtering method according to the index error calculation formula.

4.1. Real-time Estimation of Passenger Flow Status in Chengdu Metro

The real-time estimation of passenger flow status is focused on the estimation of the station's inbound quantities in the network. The fusion of multi-source data and real-time estimation will be shown below.

| Station Serial number | Period       | mobile signal statistics | AFC data |
|-----------------------|--------------|--------------------------|----------|
| 1                     | 08:00-08:05  | 111                      | 149      |
| 2                     | 08:05-08:10  | 123                      | 148      |
| 3                     | 08:10-08:15  | 130                      | 155      |
| 4                     | 08:15-08:20  | 127                      | 171      |
| 5                     | 08:20-08:25  | 132                      | 149      |
| Huaxiba               | 6            | 08:25-08:30              | 123      | 135      |
|                      | 7            | 08:30-08:35              | 114      | -        |
|                      | 8            | 08:35-08:40              | 172      | -        |
|                      | 9            | 08:40-08:45              | 164      | -        |
|                      | 10           | 08:45-08:50              | 139      | -        |
|                      | 11           | 08:50-08:55              | 145      | -        |
|                      | 12           | 08:55-09:00              | 162      | -        |

First of all, according to the multi-source data fusion formula (6),(7),(8),(9),(10),(11), the quantities of inbounds after the fusion of the Huaxiba station on Line 1 of September 30, 2018, the morning peak of 08:30-09:00 is as follows:

| Period       | AFC data |
|--------------|----------|
| 08:30-08:35  | 149      |
| 08:35-08:40  | 148      |
| 08:40-08:45  | 155      |
| 08:45-08:50  | 171      |
| 08:50-08:55  | 135      |
| 08:55-09:00  | 139      |
Figure 4 shows the AFC data of Huaxiba station at 08:00-08:30, the inbound traffic data of mobile phone signaling data from 08:00-09:00, and the trend graph of the station's inbound quantities after fusion.

Secondly, based on the quantity of inbound after the fusion of Huaxiba station, the station's inbound quantities is estimated in real time according to the Kalman filtering method. The results are as follows:

**Table 3. Real-time estimated inbound quantity of Huaxiba station.**

| Time   | 09:00-09:05 | 09:05-09:10 | 09:10-09:15 | 09:15-09:20 | 09:20-09:25 | 09:25-09:30 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| Quantity | 205         | 216         | 221         | 194         | 215         | 228         |

The actual inbound capacity of Huaxiba station from 09:00 to 09:30 is as follows:

**Table 4. Actual inbound quantity of Huaxiba station.**

| Time   | 09:00-09:05 | 09:05-09:10 | 09:10-09:15 | 09:15-09:20 | 09:20-09:25 | 09:25-09:30 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| Quantity | 213         | 222         | 205         | 208         | 230         | 222         |

Figure 5. Estimate and actual inbound quantity comparison of Huaxiba station.
4.2. Real-time estimation result analysis
The results of real-time estimation of urban rail transit passenger flow state require certain evaluation indicators to judge the accuracy of the results. In this paper, the mean square error MSE, the mean absolute error MAPE, and the average absolute percentage error MAD are used as the error evaluation indicators.

\[
\begin{cases}
MSE = \frac{1}{N} \left( \sum_{i=1}^{n} (out^i - out'^i)^2 \right)^{1/2} \\
MAPE = \frac{1}{N} \sum_{i=1}^{n} \frac{|out^i - out'^i|}{out^i} \\
MAD = \frac{1}{N} \left( \sum_{i=1}^{n} \frac{|out^i - out'^i|}{out^i} \right) \times 100
\end{cases}
\]

(14)

In the formula, \( out'^i \) is the estimated inbound amount, \( out^i \) is the actual inbound amount, and \( N \) is the estimated number.

According to the results of passenger flow estimation, the mean square error MSE value of Huaxiba station is 4.75, the average absolute value error MAPE value is 10.83, and the average absolute percentage error MAD is 5.04%. Estimating the passenger flow with different granularity in other time periods shows that the MSE, MAPE and MAD values of the Kalman filter estimation value decrease gradually with the estimated time granularity increase, which is mainly due to the small time granularity cause the inbound quantity is small, it is easy to cause large fluctuations in MSE, MAPE and MAD values. Secondly, the MAD value of the passenger flow estimates at different granularities is less than 10%.

5. Conclusion
Aiming at the nonlinearity of rail transit passenger flow, AFC data uploading has the characteristics of delay and incompleteness. This paper mainly studies the hot spot problem of real-time estimation of rail transit passenger flow state. The specific content of the discussion is as follows:

Firstly, the fusion method of AFC data and mobile phone signal as multi-source data is introduced. Secondly, the Kalman filter method is applied as the basis of the estimation model, and a real-time estimation model of rail transit passenger flow is established. The passenger flow data of the Huaxiba station of Chengdu Metro Line 1 was used for real-time estimate. The estimation result error is represented by three indicators: mean square error MSE, average absolute error MAPE and average absolute percentage error MAD. The results show that the error of the real-time passenger flow state estimation model based on Kalman filter is less than 10%, and the accuracy of the estimation model is good. It can provide better support for the urban rail transit operation management department to grasp the real-time passenger flow distribution state of the network.

References
[1] Lchtenen, M., Rosenberg, M., Rasanen, J, Sirkia, A.: Utilization of the smart card payment system (SCPS) data in public transport planning and statistics. In: 9th World Congress on Intelligent Transport Systems,2002.
[2] Chan Joanne. Rail transit OD matrix estimation and journey time reliability metrics using automated fare data [D]. Massachusetts Institute of Technology, 2007.
[3] Shi Zhuangbin, Lu Wenxue, Zhang Ning. Application of Data Mining Technology in Rail Transit AFC System [J]. Urban Rapid Rail Transit, 01:23-27, 2015.
[4] Wang Lina. Research on Passenger Flow Forecast and Scheduling of Urban Rail Transit Network Based on Historical Passenger Flow Data [D]. Master's Thesis of Beijing Jiaotong University, 2011.
[5] Morency, C., Trepanier, M., Agard, B.: Measuring transit use variability with smart-card data. Transp.Policy. 14(3), 193- 203 ,2007.
[6] Cai Changjun, Yao Enjian, Zhang Yongsheng, Liu Shasha. Forecast of Passenger Flow Distribution between Urban Rail Stations Based on AFC Data[J]. China Railway Science, 01:126-132, 2015.

[7] CHUICK, CHENGR. Kalman Filtering with Real-time Applications (fourth edition) [M]. Berlin: Springer-Verlag, 2009: 217-228.