The Researches on Subway Demand Forecast at Station Level: Smart card data, Space Syntax and Points of Interests.

To cite this article: Cheng Cheng Wu and Da Wei Chen 2018 IOP Conf. Ser.: Mater. Sci. Eng. 392 062113

View the article online for updates and enhancements.

Related content
- SMath for Physics: Linear regression
  B V Liengme
- Measuring accessibility of sustainable transportation using space syntax in Bojonggede area
  B A Suryawinata, Y Mariana and S Wijaksono
- Using Space Syntax to Assess Safety in Public Areas - Case Study of Tarbiat Pedestrian Area, Tabriz-Iran
  Kübra Cihangir Çamur, Mehdi Roshani and Sania Pirouzi
The Researches on Subway Demand Forecast at Station Level: Smart card data, Space Syntax and Points of Interests.

Cheng Cheng WU, Da Wei Chen
School of Transportation, Southeast University, Jiangsu, Nanjing 210096, China
Email: niunoumian@sina.com

Abstract. This paper examines the possible ‘attraction’ factors at station level in subway ridership, such as built environment factors and Spatial connectivity. Direct demand models were built to estimate subway ridership based on multiple linear regression. Also, a comparison between predicted values and actual values in subway ridership were applied in order to analyze forecast accuracy. It reveals that energy density on rail scale of service and spatial integration could be an important asset for urban rail demand forecast models at station level. Based on such correlative factors, parameters of direct models are chosen and calibrated by means of multiple linear regressions using actual data from Nanjing subway. The results show that the direct models can be more accurate and efficient in estimate subway ridership at station level in a short circle.

1. Introduction
The subway transportation has played an important role in urban cities due to its large capacity and high efficiency [1]. However, with denser subway stations in China, a challenge for traditional four-steps models of ridership estimation has been proposed, due to the consideration of multiple possible stop-choices by travelers within walking distance [2]. Also, rapid development of urban subway leads to the huge use of new lines and stations, which, on the other hand, leads to the requirement of forecasting ridership on using stations.

In addition, big data can be possible used in traffic prediction due to the wider development of Intelligent Transport Systems(ITS). Smart card data, as one of the most commonly used data, has been widely used in micro ridership analysis, such as forecasting destinations of bus passengers [3], discovering regularities in human mobility [4] or deducing passengers' route choices [5]. However, using smart card data to forecast subway ridership at micro station level has not been studied and direct models without using traditional four-steps models has not been built until now.

Also, Space Syntax theory provides a quantization of spatial relationships [6], which can be applied to quantifies the relationships of different stations in subway network. Public Maps API makes it possible to get points of interests (POI) to examine the relationships of built environment factors and subway station ridership.

2. Factors related to subway ridership attraction

2.1 Spatial factors
With denser subway stations used, pedestrians may meet more than one stations at a distance that they are willing to reach. Thus, diverse selections will be shown and ridership interactions between stations will change. Based on previous research and common experience, Spatial factors in stations has important influence on the choice of pedestrians, which can be delimitated into four types:
Minimum distance: In general, pedestrians may choose the nearest station with minimum walking distance, which shown in Station 1, figure 1;

Shortest path: Space Syntax theory shows that sometimes, travelers may choose stations with shortest path, which means, the most convenient path to reach one of the alternative stations. For example, station A# may be 300m to reach with three turns, while station B# can be 400m to reach without a turn. Then travelers may choose station B# to avoid turns. Shown in Station 2, figure 1;

Best integration: The station which can reach more other stations easily in a subway network may be attractive for some pedestrians as they can be more convenient for travelling in the whole subway transportation. Shown in Station 3, figure 1;

Best choice: The station which can be easier to reach by another transport mode (for example, more comfortable bicycle ways or enough space to park), may be wider used for some travelers. Shown in Station 4, figure 1;

Figure 1. alternative stations with four types.

Based on the four types above, a list of Spatial variables is selected to reflect Spatial influence on station choice. Shown in Table 1. 500m is used as a suitable pedestrian distance to define analysis scope [7], which has also been the standard catchment area for designing transit-oriented development (TOD).

Table 1 List of Spatial variables

| Types            | Variables                                      | Description                                                                 | Data sources                                      |
|------------------|------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------|
| Shortest path    | Integration RN at 500m                         | Reflecting how reachable a station can be in the walking network within 500m | Data collected from public map API                |
|                  | (for single station)                           |                                                                           |                                                   |
|                  | Integration RN                                 | Reflecting how reachable a station can be in the whole network             |                                                   |
|                  | (in the whole network)                         |                                                                           |                                                   |
| Best integration | Choice RN                                      | Reflecting how many times a station used in the whole shortest-path-traveling network | Urban subway axial model                         |
|                  | (in the whole network)                         |                                                                           |                                                   |
|                  | Bus stops at 500m                              | Reflecting how many bus stops within 500m                                 | Data collected from public map API                |
|                  | (for single station)                           |                                                                           |                                                   |
|                  | Parks at 500m                                  | Summation of the minimum physical distances of each parking lot within 500m |                                                   |
|                  | (for single station)                           |                                                                           |                                                   |

2.2 Points of Interests

It’s well known there has been a connection between rail transit ridership and land use patterns. More specifically, population and employment densities, land use mix diversity, and intermodal connectivity
(shown in table 1. Best choice) can be positively related to subway ridership [8]. In this paper, we propose a new variable $P$, which means energy density on rail scale of service, with the built environment data from Point of Interest (POI), to analyze the influence of population and employment densities, land use mix diversity and minimum distance.

The formulae of $P$ are detailed as follows [1], in which the $P_j$ is the energy density on rail scale of service, the $N_{ij}$ means the number of type $i$ POI within the influence area of station $j$, and the $A_{ij}$ represents the ridership generated rate of type $i$ POI within the influence area of station $j$. The ridership generated rates of various POI reference ‘Construction Project Traffic Impact Assessment Standard’ from Ministry of Housing and Urban-Rural Development (MOHURD). $Square_j$ means the influence area of station $j$. If a station has more than one exit ports, we use the mixed 500 meters radius as the influence area of this station, which is shown in figure 2;

$$P_j = \sum_i \frac{(N_{ij}A_{ij})}{Square_j}$$

Figure 2. Definition of the station’s influence area

3. Methodology

3.1 Analytical methods

A direct demand model at station level is built by means of multiple linear regressions, and statistical methods can be divided into 3 steps, as shown below.

First of all, bivariate correlation will be used to identify the correlation between possible ‘attraction’ factors and station ridership per day. Positive relationship will be found among all variables above and the parameters of the direct demand model will be determined. In this step, ridership of each station will be calculated from subway smart card data of more than 4 months. And the values of each variable from every station will be calculated from POI data and urban subway axial model.

Secondly, parameter calibration will be carried out to build a direct demand model by multiple linear regression analysis.

In the end, a series actual data will be provided to compare forecast ridership of each station with actual station data. The values of each variable will be collected and used to forecast station ridership while actual data will be calculated from smart card data.

3.2 Source of data and statistical approach

Three data source have been used and two processes have been executed at the same time.

3.2.1 Process one: values of Independent variables

Values of Independent variables are collected from two data source: POI data from public map API and spatial values from subway axial model.

Various POI are collected from open public map with its type (for example, business point or
education point) and coordinate of latitude and longitude. The amount of Bus stops can be calculated in a certain latitude and longitude range, which can be limited to the exit ports of each station. Also, distances of each parking lot within influence distance and the amount can be collected by coordinate of latitude and longitude.

At the same time, open public map shows the detailed street of each station and the whole subway network, which can be used to create corresponding axial models by using Space Syntax theory. Values of independent variables from Shortest path and Best integration types can be calculated from these models.

3.2.2 Process two: values of dependent variable

Dependent variable estimated in the direct model is ‘ridership per day’ of each station. Smart card data shows entry and exit station name, entry and exit time of every cardholder. Database tools are applied to calculate station ridership of a period time (more than 4 months) in order to get values of dependent variable.

3.3 case study

The case study is the city Nanjing, one of the earliest subway-using-cities in China. The subway in Nanjing meets more than 717 million person-times transport demand every year. Also, the development of subway transportation has been rapid with nearly 2 new lines added each year. The great number of subway ridership and the rapid development make Nanjing an ideal case for this study.

We used smart card data from August to December, 2013 to calculate values of dependent variable (ridership per day). Corresponding values of Independent variables are also collected to build our direct demand model by multiple linear regression analysis. During this period of time (August to December, 2013), subway service includes 2 lines using more than 3 years (Line 1 has been used for 8 years while line 2 has been used for 3 years).

In the comparative analysis of the model results, we used smart card data and POI data from August to December, 2015. During this period of time, a new subway line (line 3) has been used for more than 5 months. So we can estimate station ridership of line 3 with the direct model and compare with actual data from smart card.

4. Results

4.1 Bivariate correlations

The correlation coefficients are presented in table 2, which reveals that energy density on rail scale of service and the amount of bus stops are high correlated with station ridership. Another comments can be done:

- Best choice (bus stops at 500m) shows highest related with ridership of single station, which may point out a twin processes when designers plan the higher-ridership-demand region with more Public transport service points (such as bus stations or a subway station).
- Energy density on rail scale of service \( P \) is high associated with station ridership, which is an expected result, given that higher ridership generated rate leads to more travelers for a single station, especially when most travelers tend to choice the station with minimum distance.
- Spatial factors, especially the factors reflecting one station in the whole network, prove their explanation power on their own right. It’s an expected result as only a few travelers may choose a longer-distance station to convenient their whole subway trip.

| Types            | Variables                          | \( R^2 \) | Descriptions                          |
|------------------|------------------------------------|----------|---------------------------------------|
| Shortest path    | Integration RN at 500m             | 0.522    | It was significantly correlated at 0.1 level (bilateral) |
|                  | (for single station)               |          |                                       |
4.2 Direct forecast model and accuracy of forecast

Table 3 provides the results of parameter calibration for the direct model and the accuracy of forecast are shown on Table 4. Stations of Line 3 (a new-using line in the case study) is the main object of ridership forecasting. Also, the correlation between predicted values are actual values is shown in table 4. The numbers of stations at different degrees of accuracy is shown in table 4 as well.

**Table 3 Results of parameter calibration**

| Types            | Variables                                      | Coefficient |
|------------------|------------------------------------------------|-------------|
| Shortest path    | Integration RN at 500m (for single station)    | 12336.249   |
|                  | Integration RN (in the whole station)          | 13445.780   |
| Best integration | Choice RN (in the whole network)               | -1.369      |
|                  | Best choice                                    |             |
| Bus stops at 500m| (for single station)                           | 507.76      |
| Parks at 500m    | (for single station)                           | 0.218       |
| Points of Interests| (the energy density on rail scale of service) | 0.698       |

**Table 4 Accuracy of forecast**

| Correlation between predicted values and actual values for all stations | 0.829 |
|------------------------------------------------------------------------|-------|
| the margin of error < 10%                                               | 20/38.46% |
| the margin of error < 20%                                               | 30/57.69% |
| the margin of error < 30%                                               | 32/61.54% |
| the margin of error < 40%                                               | 37/71.15% |

Comparison of forecast errors with results by four-steps models in other cities in China

(The deviation between the actual value and the predicted value)

| Subway Line 1# in Nanjing | -31.85% |
| Subway Line 2# in Beijing | 55.96% |

On the whole, the direct forecast model provided more accurate than traditional four-steps models in the station ridership. It’s shown more than 71.15% forecast results of stations has the margin of error less than 40%, while four-steps model can only provide a macro forecast value of the whole subway system with deviation more than 30%. It is clear that forecast results are highly correlated with actual values, which points out a great potential to develop the direct model to get higher accuracy.

5. Conclusion

This paper presents and discusses the relationships among spatial factors, built environment factors and ridership at station level. A new method to build direct forecast model with smart card data and other open public data (POI data and street data) is provided as well. The result shows that energy density on
rail scale of service and spatial integration could be an important asset for forecasting and the direct model has higher accuracy compared with the traditional four-steps method, not to mention the convenience and fast the direct model has. This could be a first step to build up a better ridership forecasting model on a micro level to face the rapid-development subway system.

References

[1] Wang Y, Tang T, Ning B, et al. Passenger-demands-oriented train scheduling for an urban rail transit network [J]. Transportation Research Part C Emerging Technologies, 2015, 60:1-23.

[2] Carpio-Pinedo J. Urban Bus Demand Forecast at Stop Level: Space Syntax and Other Built Environment Factors. Evidence from Madrid [J]. Procedia - Social and Behavioral Sciences, 2014, 160:205-214.

[3] Jung J, Sohn K. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data[J]. Iet Intelligent Transport Systems, 2017, 11(6):334-339.

[4] Zhong C, Batty M, Manley E, et al. Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data[J]. Plos One, 2016, 11(2):e0149222.

[5] Hurk E V D, Kroon L, Maróti G, et al. Deduction of Passengers' Route Choices From Smart Card Data[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1):430-440.

[6] Hillier B, Hanson J. The social logic of space[J]. In press), An Integration of GIS, Internet and Virtual Reality for the visualisation, Analysis, and Exploration of Geospatial Data, International Journal of Geographical Information Science, 1984(4):392.

[7] Mashhoodi B, Pont M Y B. Studying land-use distribution and mixed-use patterns in relation to density, accessibility and urban form[J]. Architecture, 2011.

[8] Jun M J, Choi K, Jeong J E, et al. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul[J]. Journal of Transport Geography, 2015, 48:30-40.