ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

Impact of Sampling Rate and Traffic Zone Numbers on the Validity of Public Transit OD Matrix Estimation in Travel Surveys

Zhang Xuxin[1,a,*], Lei Kefan[2,b]

[1]*School of Traffic and Logistics, Southwest Jiaotong University, Chengdu, China*
[2]*Chengdu Institute of Planning & Design, Chengdu, China*
[a]*zhangxuxin526@163.com*, [b]*lkf_leikefan@163.com*
[*]*Corresponding author*

Abstract: *This study uses conventional bus trip data inferred from the intelligent transit system in the central urban area of Chengdu as the population dataset. Multiple random sampling experiments were conducted under different sampling rates. The impacts of sampling rate on the validity of OD matrix estimation were analyzed using two evaluation indicators: the Relative Error (RE) of Mean Travel Time (MTT) and the goodness-of-fit of the trip time distribution curve. Furthermore, the influence of traffic zone numbers on estimation performance under each sampling rate was investigated. Results show that sampling rate is positively correlated with the validity of OD matrix estimation. When the sampling rate reaches the minimum effective threshold, the estimation results begin to satisfy practical application requirements, and further increases in sampling rate lead to diminishing marginal improvements in validity. Under fixed sampling rates, traffic zone numbers are negatively correlated with estimation validity. Therefore, when sampling rates cannot be increased due to practical constraints, reducing the number of traffic zones appropriately can improve gravity model calibration accuracy and ensure that OD estimation results meet predefined validity requirements.*

Keywords: *travel surveys; gravity model; AFC data; traffic zones*

1. Introduction

Travel surveys are conducted through active inquiry to obtain travelers' characteristics and behaviors and to estimate the overall population using a limited number of samples. The sample size of traditional household travel surveys is usually determined based on statistical principles. Celik et al. [1] investigated the model fitting performance under reduced and optimal sample sizes using the gravity model and concluded that small-sample survey results can be as reliable as large-sample results. In urban transportation planning practice, the recommended sampling rate typically decreases as the population size within the survey area increases.

Horbachov et al. [2] calculated the relative error in estimating travel impedance distributions across cities with different population sizes and found that the sample sizes recommended by existing guidelines (HTM 2005) may lead to estimation errors of 20%–40%. As urban population and spatial extent increase, planners often subdivide urban space into more and smaller traffic zones to "capture" more trips. However, under the combined effect of excessively low sampling rates and excessive traffic zone numbers, a large number of zero-trip cells may appear in the observed OD matrix, preventing survey samples from accurately reflecting actual travel distributions [3].

In recent years, by integrating massive datasets generated from Automated Fare Collection (AFC) systems and Automatic Vehicle Location (AVL) systems in intelligent transit systems, more complete passenger travel records and travel distributions can be obtained. Moreover, the study by Egu et al. [4] showed that AFC data and travel survey data exhibit similar overall demand structures and that some indicators are strongly correlated, indicating that the two datasets are comparable to a certain extent.

This study uses AFC data from the central urban area of Chengdu as the population dataset and adopts gravity model calibration accuracy as the evaluation metric. Through multiple random sampling experiments under different sampling rates, the relationship between sampling rate and the validity of public transit OD matrix estimation is analyzed. Meanwhile, for each sampling rate, different traffic zone schemes are designed to investigate the influence of traffic zone quantity on estimation validity.

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

2. Data Preparation

2.1. Study Area and Traffic Zone Schemes

To facilitate data processing and comparative analysis, this study uses household travel survey data (HTS) from May 2016 in Chengdu and AFC data from May 10, 2016, which has the highest data quality during the same period. To ensure consistency in comparison, five core administrative districts with the highest overlap between the two datasets (Qingyang District, Jinniu District, Chenghua District, Jinjiang District, and Wuhou District) are selected as the study area. Within this area, internal trips account for 72.78% of HTS data and 92.22% of AFC data. To simplify the analysis, only internal trips from both datasets are used in the following sections.

This study is based on the traffic analysis zone and traffic subzone (street-level administrative unit) division scheme defined in the 2016 Chengdu Comprehensive Transportation Survey. Within the study area, 662 traffic analysis zones (Traffic Zone Scheme 1, Zone 662) and 76 traffic analysis districts (Traffic Zone Scheme 4, Town 76), whose boundaries are consistent with street administrative units, are identified. To further analyze the influence of traffic zone quantity on gravity model calibration accuracy, three additional traffic zone schemes are generated by merging adjacent zones according to the proximity principle based on the above two schemes:

Traffic Zone Scheme 2: merged from traffic analysis zones to obtain 500 zones, referred to as Traffic Zone Scheme 2 (Zone 500).

Traffic Zone Scheme 3: merged from traffic analysis zones to obtain 215 zones, referred to as Traffic Zone Scheme 3 (Zone 215).

Traffic Zone Scheme 5: merged from traffic analysis districts to obtain 37 zones, referred to as Traffic Zone Scheme 5 (Town 37).

2.2. Data

2.2.1. AFC Data

Considering that the smart-card payment rate for conventional buses in Chengdu exceeds 90%, this study assumes that the bus trips recorded in AFC data represent the total bus passenger trips. After preliminary filtering and cleaning, a total of 3,550,479 conventional bus smart-card records from May 10, 2016 were extracted as the raw dataset for this study. By integrating onboard GPS data, bus stop GIS data, and IC card transaction records [5], passengers' alighting stops and transfer information were inferred, resulting in 2,533,385 trip records with successfully inferred boarding and alighting stops, corresponding to a success rate of 71.35% and covering 1,979,290 bus trips. After further filtering, only trips with both origins and destinations within the study area were retained, resulting in 1,806,825 internal trips. Assuming that trips within the study area exhibit travel characteristics similar to those represented by the full AFC dataset, these 1,806,825 trips can be considered to account for 71.35% of all conventional bus trips within the study area. Trips with identical boarding and alighting stop pairs were grouped and counted to obtain a stop-based OD table containing all boarding–alighting stop pairs.

Because AFC data can only infer stop-level transit OD (stOD), whereas travel survey data describe zone-level transit OD (ztOD), the two datasets differ in spatial granularity of trip endpoints. To enable comparative analysis, stOD data must be converted into ztOD data. Since bus stops located along roads may lie near traffic zone boundaries, this study follows the method proposed by Jiangping Zhou et al. [6], adopting a 500 m service radius for bus stops [7] to allocate stOD data to corresponding traffic zones. The final ztOD table for Traffic Zone Scheme 1 is generated according to the following conversion rules:

Passenger flow allocation follows a uniform distribution principle, meaning that the passenger volume of each stop is evenly distributed among all zones served by that stop. At the same time, to maintain consistency with HTS data, the indivisibility principle of individual trips is adopted. The passenger volume at each stop is divided by the number of served zones using integer division, and the remainder is further allocated. Because larger service areas generally cover more potential passengers, zones are first sorted in descending order by actual area size when allocating the remainder. The zone ranked first receives 1 additional trip, followed by the second-ranked zone, and so on until all remaining passenger flows are assigned. The final passenger volume for each traffic zone is obtained by summing the integer-division portion and the remainder-allocation portion.

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

To better reflect actual flow patterns and maintain consistency in total trip volume before and after allocation, the allocation process is first applied to the origin side (O). The passenger volume of each boarding–alighting stop pair in stOD is allocated to the service zones of the boarding stop to determine origin zones, and the allocated volumes are aggregated to generate an OD table consisting of origin-zone–alighting-stop pairs. The same process is then applied to the destination side: for each origin-zone–alighting-stop pair, passenger volumes are allocated to the service zones of the alighting stop to determine destination zones. Finally, trips with identical origin–destination zone pairs are aggregated to produce a ztOD table containing all origin–destination zone pairs and their corresponding trip totals. In this table, origins cover 661 traffic zones and destinations cover 662 traffic zones, with 214,764 OD pairs having trips, and the total trip volume remains unchanged at 1,806,825 trips.

Subsequent traffic zone schemes can be regarded as aggregations of one or more traffic analysis zones, and their OD matrices can be derived from the zone-level ztOD table. For example, by grouping and summing trip volumes according to the higher-level traffic subzones to which each traffic analysis zone belongs, OD matrices for traffic subzones can be obtained. This method helps maintain consistency between traffic analysis zones and larger aggregated zones throughout the analysis.

2.2.2. HTS Data

The sampling rate of the 2016 Chengdu household travel survey (HTS) was 2%. To maintain consistency with AFC data, only trips using conventional buses or with conventional buses as the primary travel mode were extracted from the HTS dataset. Because trip origins and destinations in HTS data are recorded as geographic coordinates, they can be directly mapped to corresponding traffic zones to generate ztOD matrices. Ultimately, 26,608 conventional bus trips within the study area were obtained for analysis.

2.2.3. Trip Time

In this study, travel time is used as the sole travel impedance variable. For HTS data, the travel time of each trip is calculated directly from departure and arrival timestamps recorded in the survey, and the inter-zonal travel time is defined as the average duration of all trips between the corresponding zones.

Travel time derived from AFC data consists of three components: access/egress travel time, waiting time, and in-vehicle travel time between stops. Access/egress travel time follows the calculation approach adopted in the Shanghai transportation model. Waiting time is estimated based on the headway of the target bus line [8]. In-vehicle travel time (including potential transfer time) is calculated according to passengers' boarding time and vehicle arrival time at the alighting stop.

Because travel times derived from HTS data within the same period and study area are relatively complete and reliable, linear regression is applied to calibrate inter-zonal travel times obtained from AFC data. Under Traffic Zone Scheme 1 (Zone 662), the Pearson correlation coefficient between the calibrated AFC travel time distribution and the HTS travel time distribution reaches 0.9823, indicating a good fitting performance.

To reduce the influence of extreme values on gravity model calibration parameters, this study sets a minimum travel time threshold of 10 min. After filtering, 1,717,621 trips from AFC data and 26,423 trips from HTS data are retained for subsequent analysis. The retained AFC data account for 67.83% of all trips within the study area. Travel times for other traffic zone schemes are obtained through weighted averaging based on zone inclusion relationships and trip counts.

3. Research Methodology: Model, Framework, and Procedure

3.1. Model Overview

Based on the characteristics of AFC data—namely spatiotemporal continuity and large sample size—and the research objective of analyzing the influence of traffic zone schemes, this study evaluates several mainstream trip distribution models.

Growth factor method: Although computationally simple, it ignores spatial interaction mechanisms and therefore cannot capture the impacts of traffic zone scheme changes on trip distribution.

Entropy maximization model: Despite its solid theoretical foundation, this model involves high

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

computational complexity and parameters lacking clear physical interpretation, making it unsuitable for large-scale iterative sampling experiments.

Random utility model: This approach requires detailed individual-level attributes (e.g., travel purpose and income level), which are unavailable in AFC data [3].

Considering its strong sensitivity to spatial interaction and high computational efficiency, the gravity model is selected as the core analytical approach in this study. Following modeling practices adopted in transportation models of cities such as Shanghai, a singly constrained gravity model is applied [9].

Detailed derivations of the gravity model calibration process are omitted for brevity, as the primary focus of this study is the comparative evaluation framework rather than methodological development.

3.2. Methodological Framework

This study takes the gravity model calibration process as the analytical core and treats traffic zone quantity and sampling rate as variables. By comparing the "model calibration" and "model calculation" results under different variable settings, the impacts of these variables on the validity of public transit OD matrix estimation are quantitatively analyzed. Considering that average travel time and travel time distribution in real-world travel data are easier to obtain and possess reasonable statistical confidence, and that travel time, as a commonly used travel impedance, can effectively reflect transportation infrastructure conditions, this study adopts average travel time and travel time distribution as two indicators to evaluate the validity of OD matrices estimated by the gravity model.

To make the comparison of indicators more intuitive, the Relative Error between the Actual Mean Travel Time (AMTT) and the Estimated Mean Travel Time (EMTT) is defined, and the specific calculation formula is shown in Eq.(1). The error of travel time distribution between the Actual Trip Time Frequency Distribution (ATTD) and the Estimated Trip Time Frequency Distribution (ETTD) is represented by the goodness-of-fit between the ETTD and ATTD curves. Referring to the Shanghai transportation model, when the relative error of mean travel time is less than 10% and the goodness-of-fit of the travel time distribution curve is greater than 0.8, the model results are considered valid and capable of accurately reflecting the observed conditions [10].

$$\delta = |EMTT - AMTT|/AMTT \times 100\% \quad (1)$$

4. Impacts of Sampling Rate and Traffic Zone Quantity on the Validity of Public Transit OD Matrix Estimation

4.1. Sampling Experiments

Under each traffic zone scheme, fifteen sampling rates ranging from low to high are defined. To eliminate the influence of random sampling errors, fifty independent repeated experiments are conducted under each sampling rate, and the average value of these experimental results is taken as the final result for that sampling rate.

A total of 3,750 sampling experiments and gravity model calibrations are carried out in this study. The relationships between relative error and sampling rate under different traffic zone schemes are shown in Fig.1 and Fig.2. In the figures, scatter points represent the indicator values obtained from individual sampling experiments, while solid lines illustrate the average trend of experimental results corresponding to each sampling rate within the same zoning scheme.

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

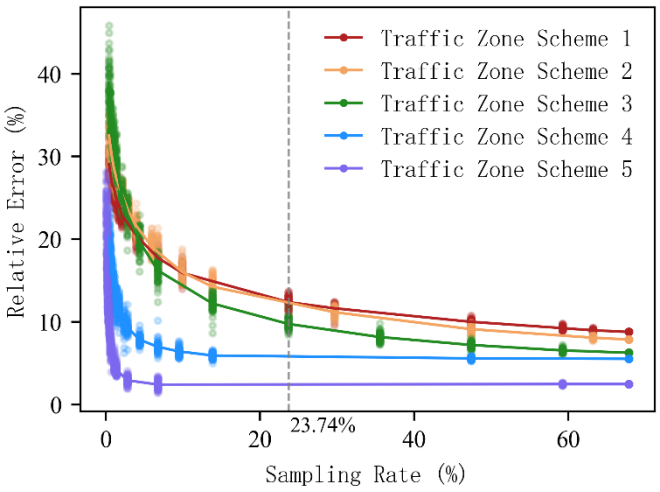

Fig. 1. Trend of Relative Error with Sampling Rate Variation under Different Traffic Zone Schemes

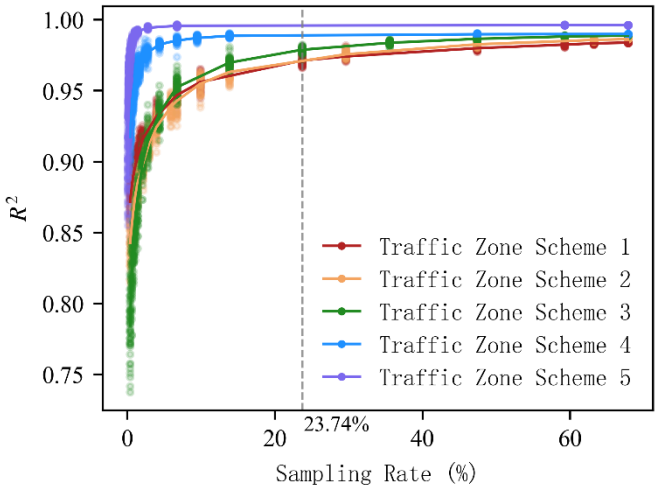

Fig. 2. Trend of R^2 with Sampling Rate Variation under Different Traffic Zone Schemes

4.2. Result Analysis

The following observations can be drawn from the above figures:

(1) Under each traffic zone scheme, sampling rate shows similar trends with relative error and goodness-of-fit. The relative error decreases monotonically as the sampling rate increases, while the mean value of goodness-of-fit increases monotonically with increasing sampling rate.

(2) As the sampling rate increases, the relative error shows a decelerating decreasing trend, whereas the goodness-of-fit exhibits a decelerating increasing trend. That is, when the sampling rate is low, a slight increase in sampling rate can significantly improve the validity of public transit OD matrix estimation results. However, when the sampling rate exceeds a certain inflection point, further increases in sampling rate no longer lead to obvious improvements in estimation validity. This phenomenon reflects the nonlinear relationship between sampling rate and model-result validity as well as the diminishing marginal return effect.

(3) The larger the number of traffic zones, the higher the sampling rate corresponding to the inflection point. This is because an increase in traffic zone quantity leads to sparser sample distribution within individual zones under the same sampling rate, resulting in weaker representativeness.

(4) Under the same sampling rate, a larger number of traffic zones generally leads to larger relative errors. For the example dataset used in this study, when the sampling rate is greater than 23.74%, a strict positive correlation exists between relative error and traffic zone quantity. When the sampling rate is lower than 23.74%, although the strict positive correlation no longer holds, Fig.1 shows that the validity

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

of public transit OD matrix estimation results generally improves as the number of traffic zones decreases. This phenomenon essentially reflects the law of large numbers in statistics. When the sampling rate approaches 0, the sample size becomes too small, resulting in significant statistical fluctuations and strong randomness in OD matrix estimation results. This behavior is consistent with fundamental principles of sampling statistics, indicating that estimation results are unstable under small-sample conditions.

(5) Because the goodness-of-fit results in this study are generally greater than 0.8 and the number of sampling experiments is limited, the sampling rate corresponding to the experimental result closest to a relative error of 10% (with an absolute deviation not exceeding 0.3%) is selected as the minimum effective sampling rate. As shown in Tab.1, the minimum effective sampling rate increases rapidly with the number of traffic zones, and the Spearman correlation coefficient between them equals 1, indicating a significant positive correlation. Moreover, the variance of EMTT under each scheme is relatively small, suggesting that estimation results at this sampling rate are stable.

Tab.1 Minimum Effective Sampling Rates for Different Traffic Zone Quantities

Number of Traffic Zones	Minimum Effective Sampling Rate	EMTT / min	δ	Variance
662	47.39%	47.09	10.00%	0.0163
500	39.49%	46.98	9.75%	0.0618
215	23.69%	46.98	9.74%	0.0777
76	2.37%	46.98	9.75%	0.2080
37	0.35%	47.27	10.42%	0.1587

(6) In summary, increasing the sampling rate or reducing the number of traffic zones can both improve estimation validity. When the number of zones is small, increasing the sampling rate can significantly reduce errors; however, when the number of zones is large, the effect of increasing sampling rate becomes weaker, and the required minimum effective sampling rate may become impractically high. Therefore, when sampling rate is constrained, reducing the number of traffic zones appropriately can help ensure estimation validity, although this comes at the cost of reduced spatial granularity of the OD matrix.

5. Conclusions

This study is based on conventional public transit trip data derived from AFC data in the central urban area of Chengdu. Through sampling simulation experiments, the impacts of sampling rate and the number of traffic zones on the validity of current public transit OD matrix estimation results were investigated. The results indicate that the sampling rate is positively correlated with the validity of OD matrix estimation. When the sampling rate reaches the minimum effective threshold (in this study, the threshold sampling rates corresponding to Traffic Zone Schemes 1–5 are 47.39%, 39.49%, 23.69%, 2.37%, and 0.35%, respectively), the estimated OD matrix satisfies the validity requirements. Beyond this threshold, the improvement in validity exhibits a diminishing marginal return. Under a fixed sampling rate, the number of traffic zones shows a negative correlation with the validity of OD matrix estimation results. Therefore, during gravity model calibration, the validity requirement of OD matrix estimation can be achieved by reasonably selecting the number of traffic zones.

It should be noted that multiple traffic zoning schemes may exist under the same number of traffic zones, and different zoning schemes may lead to certain variations in the calculation results. In addition, the AFC data used in this study represent only an approximation of the full set of conventional public transit trips, which may also affect the accuracy of the results to some extent. Future research can be improved from the following two aspects:

(1) establishing a quantitative relationship between sampling rate, traffic zone quantity, and the validity of OD matrix estimation results based on statistical methods; and

(2) conducting further sampling experiments using datasets that approximate the full population and include multiple travel modes, such as cellular signaling data, to verify whether the above findings are applicable to multimodal travel.

ISSN 2616-5775 Vol. 9, Issue 2: 54-60, DOI: 10.25236/AJCIS.2026.090208

References

[1] SMITH, M.E. (1979) Design of Small-Sample Home-Interview Travel Surveys. Transportation Research Record, 701, 29-35.
[2] HORBACHOV, P., MAKARICHEV, O., SVICHYNSKYI, S., et al. (2022) Framework for Designing Sample Travel Surveys for Transport Demand Modelling in Cities. Transportation, 49, 115-136.
[3] ORTÚZAR, J. de D. and WILLUMSEN, L.G. (2011) Modelling Transport. John Wiley & Sons.
[4] EGU, O. and BONNEL, P. (2020) How Comparable Are Origin-Destination Matrices Estimated from Automatic Fare Collection and Surveys? Transportation Research Part A, 138, 267-282.
[5] CHANG, H. (2019) Identification Method of Bus Transfer Study Based on IC Card and AVL Data. Southwest Jiaotong University.
[6] ZHOU, J., SIPE, N., MA, Z., et al. (2019) Monitoring Transit-Served Areas with Smartcard Data: A Brisbane Case Study. Journal of Transport Geography, 76, 265-275.
[7] LI, M.Y. and LONG, Y. (2015) The Coverage Ratio of Bus Stations and Evaluation of Spatial Patterns of Major Chinese Cities. Urban Planning Forum, 6, 30-37.
[8] DENG, Y.F. (2021) Research on Influencing Factors of Passengers' Route Choice Behaviors Based on Intelligent Transit System Data. Southwest Jiaotong University.
[9] ZHANG, Y. and LI, J.H. (1996) Shanghai Transport Planning Model and Its Applications. Shanghai Urban Transport Planning Institute.
[10] CHEN, B.Z., LU, X.M., DONG, Z.G., et al. (2011) Shanghai Transportation Model System. China Architecture & Building Press.