AN ENTROPY-BASED APPROACH FOR ORIGIN AND DESTINATION ACTIVITY FLOW ESTIMATION USING CROWDSOURCED DATA

Teethat Vongvanich*
University of Luxembourg, teethat.vongvanich@uni.lu
Jan-Dirk Schmöcker
Kyoto University, schmoecker@trans.kuciv.kyoto-u.ac.jp
Wenzhe Sun
University of Shanghai for Science and Technology, wzsun@usst.edu.cn
Francesco Viti
University of Luxembourg, francesco.viti@uni.lu
Federico Bigi
University of Luxembourg, federico.bigi@uni.lu
* Corresponding author

Short paper for the 13th Symposium of the European Association for Research in Transportation

Abstract: We developed an approach to estimate OD flows, including activities performed at destinations, using crowdsourced Google Popular Times data and mobile spatial statistics on population presence. Our method enables us to understand the activities that people engage in after their journey, enabling insights into demand sensitivities to urban activities. The study uses data from Kyoto, Japan with a focus on the Kyoto Station area. Results show that GPT data effectively estimate the time-varying population in each zone. Further analysis illustrates the origin of trips, and the activities engaged in.

Keywords: Google Popular Times data, Activity estimation, Origin-destination matrix estimation

1. INTRODUCTION

Despite its importance for transport planning and operation and a vast body of literature, the time-dependent OD estimation problem remains one of the most challenging tasks. The primary hurdle lies in the fact that, especially for expansive networks, the problem is notably under-determined given the usual data available for the analyst (i.e. Van Zuylen & Willumsen, 1980; Cascetta, 1984; Bell, 1991). Traditional methods for estimating OD matrices often rely on survey data or fixed infrastructure metrics, which are usually limited in their temporal and spatial resolution (Mamei et al., 2019; Pinjari & Bhat, 2011). For instance, studies utilizing mobile phone location data have demonstrated the potential for capturing dynamic travel patterns and correlating them with socio-demographic factors, thereby enhancing the accuracy of travel demand predictions (Calabrese et al., 2011; Diao et al., 2016). However, these approaches frequently overlook the rich spatiotemporal interactions that can be derived from crowdsourced data, which provides insights into venue crowdedness and user behaviour in real-time

(Zhang et al., 2021; Peng et al., 2023).

In the context of rapid urbanization and technological advancements, the ability to derive the purpose behind each trip is key. The significance of trip purpose determination extends beyond mere convenience: it plays a vital role in optimizing transportation systems, informing infrastructure decisions, and enhancing overall mobility. The challenges posed by the under-determined nature of the OD estimation problem, coupled with the complexity of trip purpose determination, highlight the need for innovative approaches in transportation research.

In response to these challenges, our study uses aggregated mobile phone and crowdsourced data to establish the relationship between travel demand and trip purpose. We employ the entropy maximization approach, which leverages aggregate constraints from crowdsourced data to estimate OD flows without having to rely on detailed trip chain information. The maximum entropy function is used in the estimation models presented by Bell (1983) among others.

Firstly, we seek to derive zonal activity weights, aligning the presence of people with activities. Secondly, we want to find activity OD matrices, correlating travel patterns with these activities. Lastly, we examine an example zone activity, exploring the variety of activities in the area and their origins. The remainder of this paper is organized as follows: Section 2 introduces the data used in the study, followed by Section 3 detailing the methodology, Section 4 presenting the results and discussion, and Section 5 concluding with key findings and future research directions.

2. DATA

2.1. Mobile Spatial Statistics (MSS)

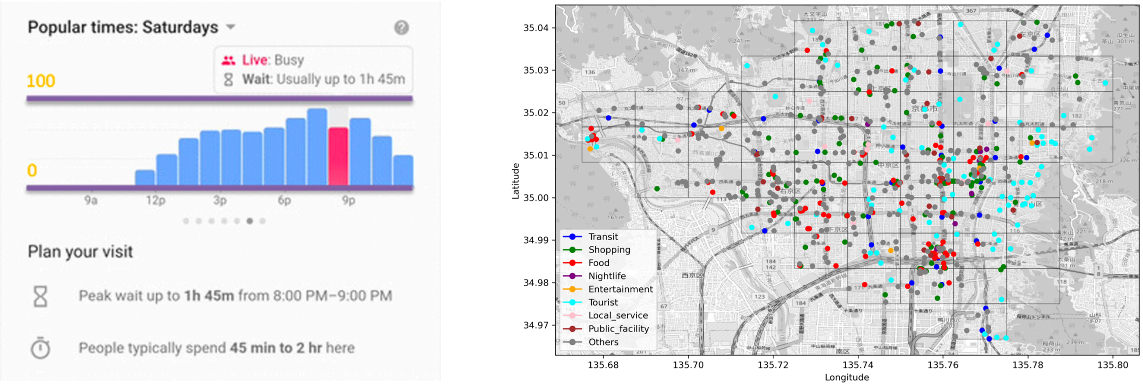
The MSS data is generated from the Nippon Telegraph and Telephone Corporation (NTT) DOCOMO mobile network. It provides information on population counts, age, gender, and residential location in 500m x 500m grid cells at hourly intervals. With over 80 million DOCOMO subscribers, the MSS data offers substantial coverage of Japan's population. To align with the travel time data and GPT zones, we aggregate the 500m MSS grids into 1km x 1km zones.

2.2. Google Popular Times (GPT)

To understand activity patterns, we collect data from the Google API on Points of Interest (POIs). This includes information such as POI name, location, type, and various activity metrics from Google's Popular Times feature. The GPT data, shown in Figure 1a, provides four key pieces of information for each POI: The historical average popularity indicates the typical busyness over past months, relative to the weekly peak. The live visit data shows the current real-time popularity compared to the historical average. The dataset also estimates the average visit duration and expected wait time for service. It is important to emphasize that Google Popular Times (GPT) data is a relative metric indicating the level of activity at a particular POI. The popularity for any given hour is presented relative to the standard peak popularity of the business throughout the week. The data is expressed on a scale ranging from zero to one hundred, where one hundred signifies the typical peak popularity within a one-week period. Live visit data is updated in real-time and can exceed one hundred.

Our case study focuses on Kyoto, where out of the 60,492 POIs identified, 10,121 have GPT data, with 1,524 providing live visitation information. To facilitate trip purpose analysis, we categorize the POIs into 8 activity groups: Transit, Shops, Food, Nightlife, Entertainment, Leisure (Tourist), Local Service, and Public Facility. The distribution of these activity groups varies across the 56 zones, with some areas exhibiting a limited presence of certain activities. This constraint leads us to rely primarily on the GPT data as the source of activity information for our models.

Figure 1b shows our study area, highlighting the 56 zones in Kyoto, each measuring 1km x 1km. Additionally, the POIs from GPT data are color-coded according to their respective activity groups. The distribution of activity groups among the GPT POIs varies across zones; some zones exhibit a presence of only few activity groups. The limited number of GPT POIs in each zone serves as a representation of the zone's activity in our models. Given the constraints, GPT data stands as our primary source of activity data for this study.



Popular times: Saturdays

Live: Busy

Wait: Usually up to 1h 45m

Plan your visit

Peak wait up to **1h 45m** from 8:00 PM–9:00 PM

People typically spend **45 min to 2 hr** here

(a) GPT graph (adapted from Google) (b) MSS zones and GPT POIs in activity groups

Figure 1. GPT data and MSS zones

2.3. Travel time matrix

The generated travel time matrix contains the commuting time between the 56 zones in Kyoto. These travel times are computed by considering the journey from the centroid of one zone to another, incorporating both public transportation (bus and/or trains) and walking. Notably, the travel time matrices are categorized by weekdays, Saturday, and Sunday, with further segmentation for each hour from 8 am to 11 pm.

To create this travel time matrix, information was gathered from web pages, including the coordinates of 1321 bus stops, the order of bus stops in 587 routes, and the frequency of each route (with a focus on the first bus stops). Subsequently, the expected waiting time per 3 hours was calculated to address challenges arising from the frequency data collection. The calculation included factors such as the expected waiting time, distance between bus stops, link generation criteria, and link costs, encompassing travel time, and waiting time. The minimum cost between all nodes was determined using the Dijkstra algorithm.

3. METHODOLOGY

The afore introduced MSS, GPT, and travel impedance data are combined to estimate activity OD matrices. Zonal activity weights link relative GPT data to absolute population measures. The activity OD matrix is then derived using entropy maximization to allocate trips by activity type.

Table 1. Notations

3.1 Zonal Activity Weights

$y_{j,d,t}$	presence of people in zone j on day d at time t
$\beta^{GPT}_{j,k}$	multiplier of activity k in zone j to match the absolute number of people
$x^{GPT}_{j,k,d,t}$	average live GPT data of the POIs categorized as activity type k in zone j on day d at time t
$\beta^{ToD}_{j,t}$	the change in number of people in zone j at time t due to hour variations
$x^{ToD}_{d,t}$	time of day dummy variables
$\beta^{DoW}_{j,d}$	the change in number of people in zone j at time t due to day variations
$x^{DoW}_{d,t}$	day of the week dummy variables
$\beta_{j,0}$	the intercept of multiple linear regression model
3.2 Activity OD Matrix	
$c_{i,j}$	travel impedance from zone i to zone j
o_i	sum of all trips originating from zone i
d_j	sum of all trips destined for zone j
α	Lagrangian multiplier of travel time constraint
C	the total travel expenditure
$q_{i,jk}$	number of trips from zone i to zone j for the purpose of activity k, performed in zone j
λ^{o}_{i}	Lagrangian multiplier of origin constraint for zone i
λ^{d}_{jk}	Lagrangian multiplier of destination constraint for zone j activity k

3.1. Zonal Activity Weights

To bridge the relative nature of GPT data with the absolute measures of MSS data, a multiple linear regression model is applied to establish the relationship between the number of people in a zone and the live visitation data of points of interest (POIs) categorized by activity type. Shown in (1), the model calculates zonal activity weights that scale GPT data into estimated counts of people engaged in specific activities within a zone.

$$y_{j,d,t} = \sum_{k} \beta^{GPT}_{j,k} x^{GPT}_{j,k,d,t} + \beta^{ToD}_{j,t} x^{ToD}_{j,d,t} + \beta^{DoW}_{j,d} x^{DoW}_{j,d,t} + \beta_{j,0} + \varepsilon_{j,d,t} \qquad (1)$$

Activities beyond GPT's scope, such as home-related activities, are accounted for with the baseline dummy variable $\beta_{j,0}$. Temporal variations are incorporated using dummy variables for time of day $\beta^{ToD}_{j,t}$ and day of the week $\beta^{DoW}_{j,t}$. For each zone, $x^{GPT}_{j,k,d,t}$ the average live GPT data for POIs classified under one of nine activity types was computed hourly, based on data from November 5, 2020, to April 30, 2021. The regression was conducted independently for 55 zones (each a 1km² area), yielding the zonal activity weights $\beta^{GPT}_{j,k}$ for activities, time of day, day of week, and the baseline term for each zone.

3.2. Activity OD Matrix

The well-known maximum entropy approach originally proposed in Wilson (1968) is adapted to our data. It relies on the idea that there are many possible trip distributions and that the most probable state of the total OD matrix is the one that maximizes the total entropy, where the entropy is given by the number of possible arrangements of the state. The model can be formulated as an optimization problem as defined in equation (2).

$$E\left(q_{i,jk}\right) = -\sum\nolimits_{i,jk} q_{i,jk}(\ln q_{i,jk} - 1) \qquad (2)$$

We maximize entropy $E\left(q_{i,jk}\right)$ subject to the constraints representing our inputs. The MSS data provide the originating trips o_i. Weighted GPT data provide $\beta_{jk}x_{jk}$, the sum of all trips destined for activity k in zone j. Activities not covered by GPT data are incorporated using intercept terms $\beta_{j,0}$ and adjustments for temporal variations with time-of-day $\beta_{j,t}^{ToD}$ and day-of-week $\beta_{j,d}^{DoW}$ coefficients. Finally, C represents the total expenditure on transportation. Upon formulating the optimization problem, the Lagrangian of this problem can be expressed as:

$$\begin{aligned} L = E\left(q_{i,jk}\right) &- \sum\nolimits_i \lambda_i^o \left(o_i - \sum\nolimits_j \sum\nolimits_k q_{i,jk} - \sum\nolimits_j q_{i,j0} \right) \\ &- \sum\nolimits_j \sum\nolimits_k \lambda_{jk}^d \left(\beta_{jk} x_{jk} - \sum\nolimits_i q_{i,jk} \right) \\ &- \lambda_{j0}^d \left(\beta_j^{ToD} x_j^{ToD} + \beta_j^{DoW} x_j^{DoW} + \beta_j^0 - \sum\nolimits_i q_{i,j0} \right) \\ &- \alpha \left(C - \sum\nolimits_{i,j,k} c_{i,j} q_{i,jk} - \sum\nolimits_{i,j} c_{i,j} q_{i,j0} \right) \end{aligned} \tag{3}$$

where $\lambda_i^o, \lambda_{jk}^d, \lambda_{j0}^d$ and α are Lagrangian multipliers. Solving the optimization problem yields $q_{i,jk}$, which constitutes our activity OD matrix.

4. RESULTS & DISCUSSIONS

4.1. Zonal Activity Weights

We obtained the zonal activity weights for all 56 zones in Kyoto. Figure 2 shows the R^2 values, model fit accuracy, for the zonal activity weights across all zones. Figure 3 presents the activity distribution by time of day for the Kyoto Station zone, which includes the station itself, as well as shops and a mall owned by the railway operator. While not all individuals in this zone are transit users, understanding the activities of people in the station area is important for operators, as many may be shopping or dining at operator-owned establishments. The distribution highlights a significant portion of the population attributed to the "nonGPT" activity group, which represents activities not captured by GPT data. These include home and work activities, as well as activities at POIs without GPT data.

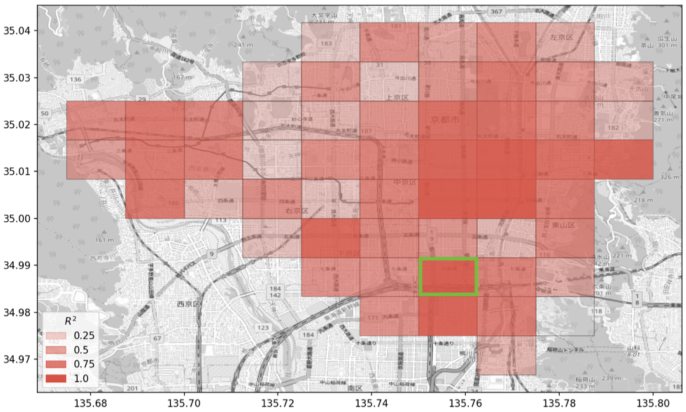


Figure 2. R^2 of zonal activity weights by zone, with the Kyoto station zone highlighted

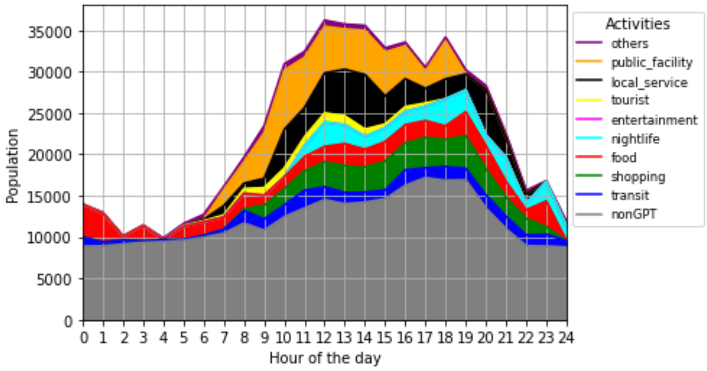
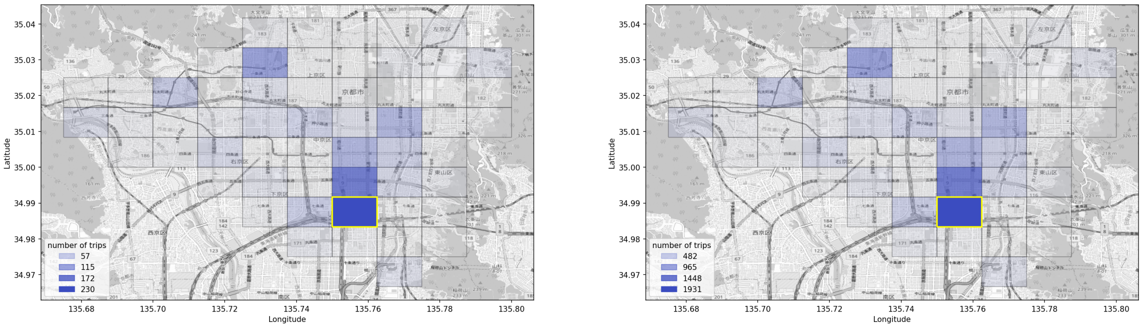


Figure 3. Activity distribution by time of day: Kyoto station zone

The regression analysis for Kyoto station zone has an R^2 value of 0.84, indicating a strong correlation between GPT data and the estimated number of people engaged in activities. However, not all zones have high R^2 values. Zones with higher R^2 values tend to be touristic or have more POIs captured by GPT. This reflects the nature of GPT data, which predominantly represents "non-routine" activities, such as leisure, shopping, dining, and tourism, rather than routine home or work activities. A comparison of “nonGPT” activities during night times with census population of the zones could provide a valuable validation step. If successful, this could allow us to estimate job numbers by analysing the difference between the population and the people accounted for in the "nonGPT" activity group. The high R^2 values in zones like Kyoto station suggest that GPT data, when combined with zonal activity weights, can effectively capture the spatial and temporal distribution of activities in areas with dense POI coverage.

4.2. Activity OD Matrix



(a) Shopping activity (b) Non-GPT activity

Figure 4. Activity OD Matrices – Trips to Kyoto station zone

Figure 4 shows the Activity OD matrices for trips with Kyoto station area as the destination. We can estimate when, how many, and for what purposes people travel to the Kyoto station area. We observe that many zones pairs have no trips at all. This can be attributed to the destination constraints in the optimization problem, which are based on the availability of GPT data. If GPT data indicates a value of zero for a specific activity in a particular zone at a given time (such as when shops are closed), no trips will occur to that activity-zone pair. This phenomenon underscores the influence of GPT data availability on trip destinations, particularly for activities such as entertainment.

Across all activities, individuals tend to remain in the same zone. This trend can be attributed to two primary factors. An important observation is the high tendency of individuals to remain in the same zone. While this is expected, as people often engage in activities within their immediate vicinity, this behaviour can be explained further. It is likely that individuals are either stationary or engaging in activities close to their starting location. However, another possible explanation is related to our input travel time data. The travel time matrix, which computes the distance between zones from their respective centres, indicates that the impedance to stay within the same zone is relatively small compared to other zones.

5. CONCLUSION

This research focused on understanding activity in each zone by estimating origin-destination (OD) matrices with trip purposes, using crowdsourced data from MSS and GPT. We presented two key methodologies: the estimation of zonal activity weights through a multiple linear regression model and the calculation of activity-specific OD matrices using optimality conditions. These approaches allow for a deeper understanding of travel behaviour by incorporating trip purposes, beyond traditional OD estimation methods. The use of crowdsourced data provides a scalable and real-time solution for cities, with potential applications for researchers, urban planners, and policymakers looking to analyse traffic patterns and activity-based travel behaviours. The analysis can be used to predict ridership demand increases; not just considering GPT data in the vicinity of the station as in Vongvanich (2023), but in the whole city.

The novelty of this work lies in its integration of trip purposes into OD estimation, a feature not commonly explored in traditional studies. This framework, while developed for Kyoto, can be extended to other cities, offering a flexible tool for urban analysis. However, challenges remain. The models are more accurate in zones with a higher number of Points of Interest (POIs), and the reliance on GPT data limits the comprehensiveness of the activity types captured. Additionally, the lack of detailed validation data, particularly for activity-specific trips, poses a challenge in assessing the accuracy and reliability of the models.

For future work, improving the validation process with richer datasets, such as detailed activity and land use data, will enhance model robustness. Further refinement of the deterrence function, potentially incorporating additional factors like travel cost and traffic conditions, could improve the realism and accuracy of the models, especially for larger cities where travel time becomes a more significant factor.

REFERENCES

Bell, M. (1991). The real time estimation of origin-destination flows in the presence of platoon dispersion. Transportation Research Part B: Methodological 25, 115–125

Bell, M. (1983) The Estimation of an Origin-Destination Matrix from Traffic Counts. Transportation Science 17(2):198-217. https://doi.org/10.1287/trsc.17.2.198

Calabrese, F., Lorenzo, G. D., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing, 10(4), 36-44. https://doi.org/10.1109/mprv.2011.41

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. Transportation Research Part B, 289–299.

Diao, M., Zhu, Y., Ferreira, J., & Ratti, C. (2016). Inferring individual daily activities from mobile phone traces: a boston example. Environment and Planning B: Planning and Design, 43(5), 920-940. https://doi.org/10.1177/0265813515600896

Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin–destination matrices obtained from cdr data. Sensors, 19(20), 4470. https://doi.org/10.3390/s19204470

Peng, J., Liu, H., Tang, J., Cheng, P., Yang, X., Deng, M., & Xu, Y. (2023). Exploring crowd travel demands based on the characteristics of spatiotemporal interaction between urban functional zones. ISPRS International Journal of Geo-Information, 12(6), 225. https://doi.org/10.3390/ijgi12060225

Pinjari, A. R. and Bhat, C. R. (2011). Activity-based travel demand analysis. A Handbook of Transport Economics. https://doi.org/10.4337/9780857930873.00017

Van Zuylen, H., & Willumsen, L. (1980). The most likely trip matrix estimated from traffic counts. Transportation Research Part B: Methodological, 14, 281-293.

Vongvanich, T., Sun, W., & Schmöcker, J. D. (2023). Explaining and predicting station demand patterns using Google Popular Times data. Data Science in Transportation, 5 (10). https://doi.org/10.1007/s42421-023-00072-z

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research, 1*(3), 253–269. https://doi.org/10.1016/0041-1647(67)90035-4

Zhang, J., Hasan, S., Yan, X., & Liu, X. (2021). Spatio-temporal mobility patterns of on-demand ride-hailing service users. Transportation Letters, 14(9), 1019-1030. https://doi.org/10.1080/19427867.2021.1988439