Estimating the Number of Tourists in Kyoto Based on GPS Traces and Aggregate Mobile Statistics

Tomoki Nishigaki, Jan-Dirk Schmöcker, Tadashi Yamada, and Satoshi Nakao

Abstract A clear understanding of the demand patterns, is one of the key contributors to laying a firm foundation for tourist planning. In pursuit of that aim, we estimated the number of tourists at specific areas and times in Kyoto City using regression analysis and hierarchical linear models (HLM). We first discuss how to extract the tourists' data from a "mesh population" obtained from aggregate mobile network operational data. We then propose that a relatively small sample of GPS tracking data for a population that has been monitored over a longer time than the mesh population can be used as a surrogate. To distinguish tourists from other persons, we find that a specified threshold of visiting a certain number of tourist attractions per day is useful. We also examine the effect of months and time of days by HLM on the model fit and number of tourists. Finally, we show that the accessibility of information such as the level of the attractiveness of particular Points of Interests (POIs) measured in terms of "Google ratings", in conjunction with the GPS records significantly contributes to a better estimation of the number of tourists at specific areas and times in Kyoto City.

Keywords Tourism · Hierarchical linear model · Mesh population · GPS data · Population estimation

T. Nishigaki (✉) · J.-D. Schmöcker · S. Nakao
Graduate School of Engineering, Kyoto University, Kyoto, Japan
e-mail: nishigaki@trans.kuciv.kyoto-u.ac.jp

J.-D. Schmöcker
e-mail: schmoecker@trans.kuciv.kyoto-u.ac.jp

S. Nakao
e-mail: nakao@trans.kuciv.kyoto-u.ac.jp

T. Yamada
Graduate School of Management, Kyoto University, Kyoto, Japan
e-mail: yamada.tadashi.2x@kyoto-u.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

C. Antoniou et al. (eds.), *Proceedings of the 12th International Scientific Conference on Mobility and Transport*, Lecture Notes in Mobility,
https://doi.org/10.1007/978-981-19-8361-0_14

1 Introduction

Over the past few years, Japan's popularity as a tourist destination has been gradually increasing with exception of the sudden interruption by the COVID-19 crisis. Consequently, problems such as traffic congestion and crowding at and around touristic places have become increasingly become more serious issues, causing dissatisfaction amongst both the tourists and residents alike [1]. This study focuses on Kyoto, Japan's old capital, one of the most significant tourist destinations. Before the COVID-19 crisis, the annual number of tourists continuously increased for two decades, reaching nearly 60 million per year, and the tourism consumption reached 1.2 trillion JPY, as illustrated in Fig. 1 [2]. However, along with the rapid rise in tourism, dissatisfaction amongst the tourists with their touristic experience has also gradually increased over the years, congestion being one of the major contributing factors. For instance, between 2011 and 2019, the mention of "crowding" as the main reason for tourist dissatisfaction increased from just over 10 to 20% [2].

For Kyoto City and Japan at large, tourism is an essential stimulant of economic prosperity, and as such improvement of tourists' satisfaction is considered both as a priority and a common fundamental policy objective. Furthermore, in the wake of the COVID crisis and its aftermath, adequate prevention of crowding in touristic areas, after the end of travel restriction policies has become an additional concern. One of the recommended approaches to solving this problem is obtaining a clear understanding of the tourism demand. We aim at contributing to that objective by estimating the number of tourists in specific areas, at specific points in time in Kyoto city. The paper explores if two data sets and transport accessibility measures are suitable to extract and predict tourist numbers.

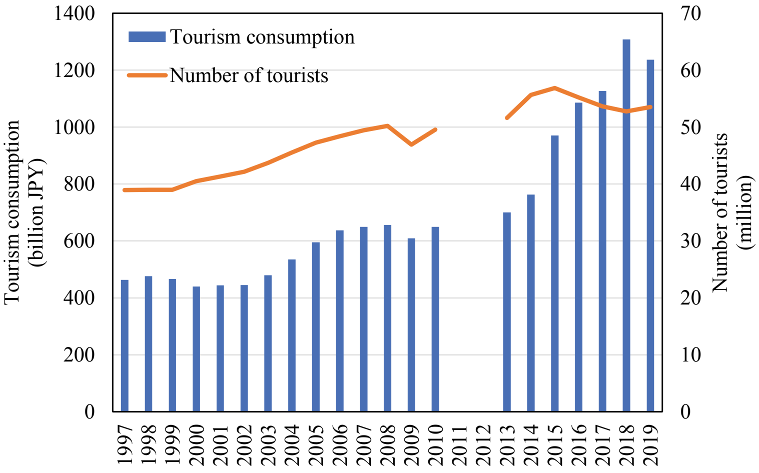


Fig. 1 The trend of tourism consumption and the number of tourists in Kyoto, Japan

Since there is far less data on tourist behavior than residential travel behavior, most travel surveys mainly target residents because obtaining representative samples from tourist surveys is quite difficult. One of the novel approaches to addressing this challenge is the use and reliance on "big data" for tourist travel patterns as discussed by Schmöcker [3]. Though there are several kinds of data such as mobile network operational data, GPS data, Wi-Fi access point data, traffic IC card data, probe car data…etc., all of these require additional analysis to distinguish tourists from other travelers or residents. Furthermore, the spatial and temporal units of a majority of these data sources do not often match those of interest for planning purposes. For instance, in most cities, there could be several areas where tourists tend to gather, at which the planner intends to have a clear estimate of the visitors and crowding.

In this study, the two main data sources used are mobile network operational data and GPS tracking data. The mobile network data provides us with information about the population within predefined mesh areas. This means that for our interest, touristic areas, an estimation based on "interpolation" or other statistical methods is needed. To note is that the data are based on a very large number of mobile phones so that the total number of persons in a mesh can be considered to be a fairly accurate estimate. The GPS data is the location data of users of a travel planning app with a time stamp. This data provides us with detailed individual data of those who agreed to share their location information. Because of this, the sample size is significantly smaller as well as biased towards public transport users. This means that it is more difficult to obtain an accurate population estimate from this data. In addition to these two data sets, we used "Point of Interest" (POI) data and public transportation network data to consider the touristic features of each mesh and its accessibility.

Both, mobile mesh data and GPS data, are hence not the "ground truth data" of tourists in touristic areas. In this paper, we discuss their limitations and show their correlation. In particular, we aim to understand how well the GPS data can be used to estimate tourist numbers. The reasons are twofold. One is that this smaller sample data set is often more available for researchers (or affordable at a lower price). Secondly, if the GPS data can be used for our total tourist number estimates and if the biases in the dataset are understood, this also provides us with more confidence for further analysis of tourist characteristics, such as which places are visited in conjunction and what the typical stay times are, at those touristic areas. Such information is not available from the aggregate mobile phone mesh data.

This motivates the establishment of a model where the mobile mesh data is the dependent variable and the GPS data, accessibility information from POI and public transportation accessibility data are independent variables. We first establish linear regression (LR) models and then also establish models which consider the effect of month or time of day with hierarchical linear models (HLM). Lastly, we apply our model to estimate the population within each area as defined by the Kyoto city government.

2 Literature Review

A relatively rich body of literature exists about the estimation of the static number of persons in specific areas. Some researchers have proposed the use of multiple existing population maps i.e., GPW (Gridded Population of the World), LandScan, WorldPop, GRUMP (Global Rural–Urban Mapping Project), GHS-POP (Global Human Settlement-Population), HYDE (History Database of the Global Environment), census data, etc., to estimate the population within a square of any size as so called "mesh populations" [4, 5], and to evaluate their accuracy [6–10]. Each of these databases is based on census data or information from scanning data of the earth by satellite, etc. and are in general highly reliable.

Notable is that, we found little research on other "shape regions" such as touristic areas because most of them focus on only mesh populations. There are, however, some significant contributions on other population group estimates. Balakrishnan [11] estimated a residential density with a 30 m resolution using street density, building heights, and ward-level data on car ownership. Bakillah et al. [12] estimated the building level population using building footprints and POI data. Shimosaka et al. [13] estimated the population within 100 m square meshes using POI and anonymized large-scale GPS data. There is also research on refining the census population [14, 15]. These focused on the specific areas used in census data, but these only refined the population obtained from census data based on multiple spatial resolutions, optical imagery, or telecommunications data. Kikuchi et al. [16] instead used mesh data to calculate an "expansion factor" of a population estimate obtained from census data. This expansion factor describes the ratio of the mesh data to the census data, so that a population estimate can be obtained from the sample. Otake and Kikuchi [17] used mesh data as the actual value of a population and then refined the origin–destination specific traffic volume based on data assimilation. Similar to what will be done in this study, they redistributed the mesh population into the area used in the census based on the sizes of each area. These contributed to estimating the population within the fine-grained mesh or regions which do not fit the mesh patterns.

There is further research on estimating the number of persons at a specific time in a spatial area. Khodabandelou et al. [18] and Cecaj et al. [19] estimated the urban scale dynamic population densities with mobile network traffic data. Bachir et al. [20] and Aasa et al. [21] estimated the dynamic population densities within an area used in the census. All of the results make methodological contributions, but none is related to the problem of tourism estimation. For tourism applications, we note the work of Ahas et al. [22] who used the anonymized GPS data, extracted foreigners' data based on the information on the mobile phone as the tourists' data from foreign countries, and analyzed the difference in behavior among nationalities. Ahas et al. [23] also analyzed the data with accommodation statistics and showed the distribution of the bias of the data. They emphasized that their data was a sensitive issue due to privacy concerns. If we could access data similar to theirs, our research objective would have been much more quickly accomplished, however, we are unable to access such data due to privacy restrictions in Japan, as they emphasized.

Recent research about Japanese tourism, has mainly focused upon the use of other forms of data other than that from the traditional surveys. Ubukata et al. [24] estimated the features of tourists' behavior using GPS data, i.e., the number of those visiting a specific area, the staying duration, the origin and the traffic mode to visit the site(s). Their data was collected with the smartphone users' consent every 5 min. They reported that the number of users who consented was only about 5% of the Japanese population, implying that there could be a bias in their findings. They extracted the tourists' data by following a 2-step criteria; First, the number of visits to the objective area per year is less than 12. Secondly, the users walked around at least two touristic areas. Kobayashi et al. [25] also used GPS traces to estimate tourist flows but only used the simple criteria of how many days a person is observed as a criterion to distinguish tourists from other persons in their sample. Dantsuji et al. [26] used Wi-Fi packet data and estimated the stay time of tourists. Nakanishi et al. [27] used Wi-Fi packet data and counted the number of persons visiting the facilities with Wi-Fi packet sensors. Kawakami et al. [28] used the same population mesh data that will be used in our subsequent study and OD traffic volume that is published from the same mobile phone provider to estimate the OD traffic volumes. They showed that the data could contribute to understanding the tourists' behavior and also pointed out the need to better estimate the total tourist population. However, they focused on the predefined meshes or areas created by combining the meshes, so their study did not consider the non-uniformly shaped touristic areas. Gao and Schmöcker [29], also suggested that Wi-Fi packet data is a good source to estimate point specific tourist numbers only near the Wi-Fi sensor as well as flows of tourists between specific parts of the city, however, a large number of sensor installations is required which may be nonviable to obtain the total number of tourists.

In conclusion, the majority of the past research on tourist estimation studies is based on data from the traditional surveys, except for some recent contributions [30–32]. There is relatively little research on estimating the number of tourists using other sources of data The main novelty of this study is the estimation of the population from mesh data using the population from small sample GPS data and accessibility information. In addition to this, we also estimate the population within the areas defined by the Kyoto city government. As shown by the literature, land-use and other "map data" also appear to be promising to partly overcome these problems like estimating the number of tourists. The subsequent study aims to explore this further within the touristic areas.

3 Tourism in Kyoto and Data Overview

3.1 Tourism in Kyoto

Tourism behavior in Kyoto is widely varied and dispersed, and the tourists use various transport modes to travel between touristic areas. Figure 2 shows the map of Kyoto city and the 37 tourist areas, and Table 1 indicates the name and size of each area. Shen et al. [33] used the same map's definitions to estimate tourist flows between these areas based on a survey of tourists at public transport stations. The density of public transportation is higher in the south region of Kyoto city than in the north. Kyoto city government strongly recommends tourists to travel without their cars. This policy is called "Arukumachi Kyoto."; in English, "Kyoto, the town for walking around." Thanks to this policy, many tourists use a variety of other transportation modes. These areas shown in Fig. 2 have multiple sizes and features: some include only one famous touristic point, some include a few touristic points, and some cover huge areas like hiking trails. As Fig. 2 and Table 1 illustrate, our challenge is to estimate the number of tourists within the various size areas and characterize them.

Ishigami et al. [34] mentioned that mobile network operational data and GPS data focus on all traffic modes, Wi-Fi access point data mainly targets pedestrians within the vicinity of Wi-Fi spots, traffic IC card data targets only public transportation users, whereas probe car data clearly records only car users. From this perspective, mobile network operational data and GPS data are most suitable for our study, which is why we sought access to these two types of data sources.

3.2 Mesh Data

First, our mobile network operational data are "mobile spatial statistics" from a major Japanese mobile phone service provider [35]. This data is generated based on the following criteria; counting the number of mobile phones around the cellular phone base station, expanding the counted value based on the diffusion rate of the mobile phone provider, and using these values to provide estimates within standardized 1 km square mesh population. This mesh data is only published based on predetermined meshes defined for all of Japan. Adjusting the data to non-uniform meshes is not trivial. Consider a case where half the mesh is covered by inaccessible mountains and the other half of the mesh contains touristic POIs. (A scenario that is common in Kyoto as several temples are located at the gateways to mountains.) Then presuming that half of the mesh population is in the touristic part of the mesh is clearly an underestimation. As will be discussed in Sect. 5–3, we, therefore, use the GPS data to account for such cases.

The total number of persons in a mesh can be considered accurate due to the significant market share of the mobile phone provider. The mesh data also allows us to distinguish between the people from Kyoto, other provinces of Japan and

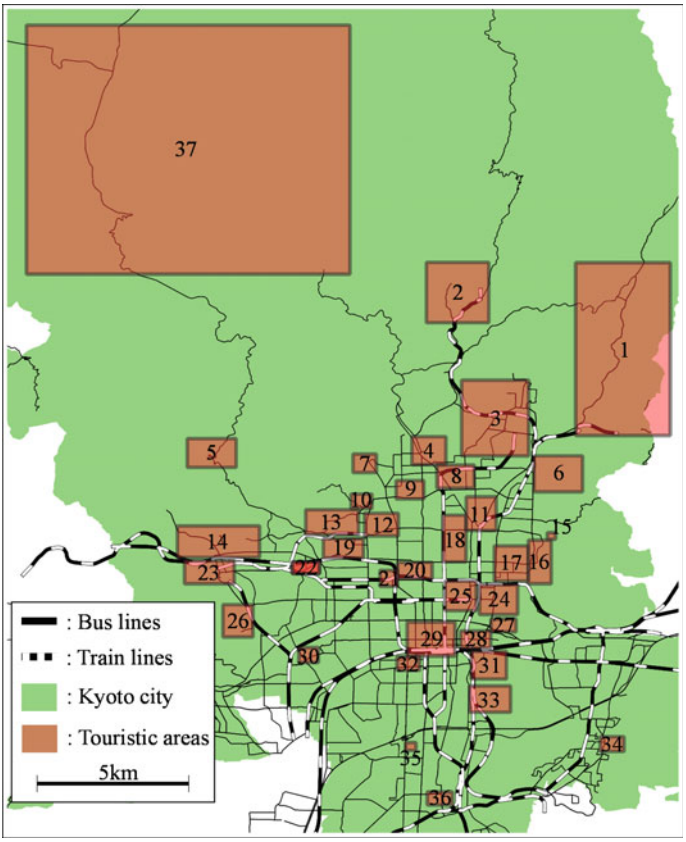


Fig. 2 Touristic areas in Kyoto city (*Source* Survey by Kyoto city government)

foreigners. The data provider is able to do so according to the registered address of the mobile phone. However, for privacy reasons, only aggregated data is available, and the number of persons within a mesh is not published if too small. This particularly implies that, the number of foreign tourists can only be obtained for large space and/or time intervals. In this study, therefore, we focus on Japanese tourists who make up about 90% of the total tourists in Kyoto.

Data was obtained for some Wednesdays and weekends/public holidays for the period (October 2018–January 2019) i.e., from the period before COVID-19. We considered averages for Wednesdays as representative of weekdays and averages for weekends and public holidays as representative of holidays.

Table 1 The name and size of each touristic area

No	Name	Size [km^2]	No	Name	Size [km^2]	No	Name	Size [km^2]
1	Ohara/Yase	21.9	15	Ginkaku-Ji Temple	0.0742	29	Kyoto station Vicinity	2.20
2	Kurama Area	5.03	16	The path of philosophy	1.34	30	Katsura imperial villa	0.396
3	Takaragaike	6.83	17	Heian Jingu Shrine	1.59	31	Tofuku-ji Temple Area	1.31
4	Kamigamo Shrine	1.28	18	Kyoto imperial palace	1.45	32	To-ji Temple Area	0.423
5	Takao Area	1.89	19	Hanazono Area	1.03	33	Fushimi inari Shrine	1.45
6	Shugakuin/Shisendo	2.35	20	Nijo castle Area	0.765	34	Daigo-ji Temple Area	0.476
7	Koetsu-Ji Temple	0.579	21	Nijo station Vicinity	0.382	35	Jonan-gu Shrine Area	0.111
8	Kitayama-dori street	1.17	22	Uzumasa Area	0.431	36	Fushimi Area	0.448
9	Daitoku-Ji Temple	0.685	23	Arashiyama Area	1.54	37	Keihoku direction	108
10	Kinkaku-ji Temple	0.408	24	Gion Area	1.53			
11	Shimogamo Shrine	1.36	25	Kawaramachi	1.25			
12	Kitano Temmangu Shrine	1.03	26	Matsuo Taisha Area	1.30			
13	Kinugasa / Omuro	1.77	27	Kiyomizu-dera Temple	0.433			
14	Sagano Area	3.48	28	Sanjusangendo	0.702			

3.3 GPS Data

Secondly, we have access to GPS data from a public transport planning mobile phone application called "Arukumachi Kyoto." Some users have given their consent to being tracked, and their locations and timestamps are stored mostly while the app is in use. With this individual data, user ID, and using language in users' OS, it is possible to distinguish between Japanese and foreigners. To match our analysis with the mesh data, we use only data from those presumed as Japanese based on the language settings. Mainly due to the necessity of the users' consent to obtain the data, the sample size of this GPS data is much smaller than the mesh data. The data will be further biased towards those using public transport as car users have less need to use this app. Because of this, even if we aggregate the data, it does not provide the actual value of the number of persons within an area and we need to consider it in conjunction with other data. Our data covers all days for the period between October 2018 and January 2019.

We note, that GPS data also has accuracy issues. However, we suggest this is a minor issue in our case, since we used GPS data only for counting the number of users per hour within each mesh. Specifically, we consider a traveler was in the mesh at each hour when at least one GPS record is inside the area. Hence errors can be made only if all records of a user over an hour (if there are multiple ones) are continuously outside the touristic area which the traveler visited. That case can be made only when a user stayed or walked around near the mesh edge for over one hour. This is, however, not likely the case because almost all tourists walk around in various directions for tourism. Clearly, the aforementioned problem, of missing records, is a more significant one.

3.4 Point of Interest (POI) Data

We also used “POI data” collected from Google maps API. We collected the information on the objects labeled “tourist_attraction.” The information includes the latitude, longitude as well as the average rating by visitors of the POI and the number of ratings. Figure 3 shows the number of POIs per mesh, the average rating, and the total number of ratings within each mesh. As can be seen, 165 meshes have no POI. Their distribution is shown in Fig. 3. The figure shows that most meshes do not have POIs and the ratings are concentrated within a few meshes. Furthermore, most of the POIs are rated highly.

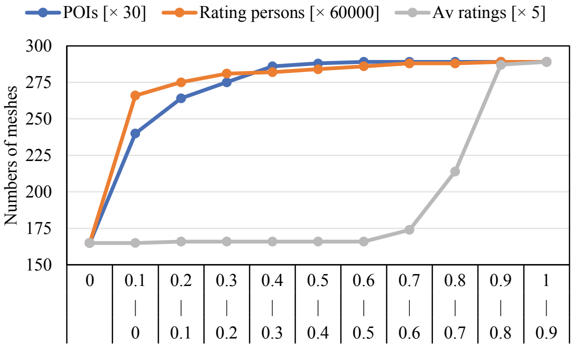


Fig. 3 The distribution of the three types of POI values per mesh

3.5 Public Transportation Data

We also used public transportation data to explain in how far the accessibility of an area explains the tourist presence. We estimated this accessibility by the number of stations or bus stops and travel costs. Since GTFS data was not available for Kyoto, the information on routes, their frequencies and fare information was gathered from the operators' web pages. The average waiting time of each line at its terminal during 3 h time intervals was used as the frequency of the line as a representative. We used the fare table for Kyoto city subway [36] as the fare for all links on the train as fares for each train link are not published and since there is no fixed fare per km. The fare for all buses was set to 230 JPY as this is the Kyoto city bus flat fare, though some operators charge in some parts of the city slightly different (distance-depending) fares that are not published on web pages. The generalized travel cost, therefore, includes travel time, waiting time, and fare. We used the time value of 29.8 JPY/min suggested in the VOT meta-analysis of Kato and Hashimoto [37]. Frequency is converted into waiting time, assuming regular service arrivals and random passenger arrival.

We used the above public transport information as direct indicators of the accessibility in conjunction with the POI data as shown by Eq. (1).

$$W_{it} = \sum_{j} \mathrm{w}_{j} exp\left(-C_{ijt}\right) \tag{1}$$

where C_{ijt} is the generalized cost from station or bus stop i to mesh j in time of day t. This includes the travel time, waiting time, and fare. w_j is the weight of mesh j based on the POIs in the mesh. We consider four types of POI weights to reflect their attractiveness to tourists: w_j^p is the average number of ratings (number of rating persons) and w_j^r the average rate within a mesh. The product of ratings and the number of ratings can be considered as a more comprehensive measure of attractiveness, so that we further define $\mathrm{w}_j^a = \mathrm{w}_j^p \mathrm{w}_j^r$ as well as a logarithmic version of $ln(\mathrm{w}_j^a)$. Our rationale for testing the logarithmic value of w_j^a was that it is closer to a normal distribution than w_j^a. In the regression we tested all versions as will be described in Sect. 5.

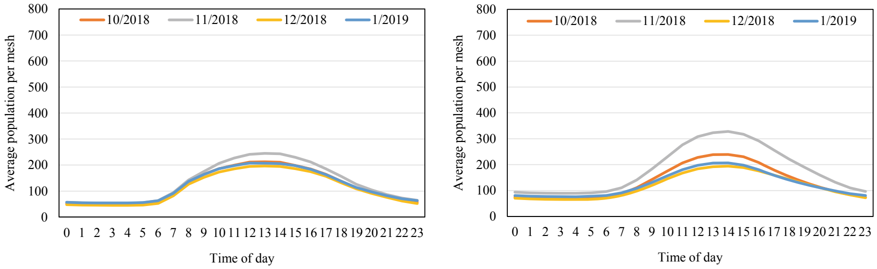
4 Tourist Number Estimation from Mobile Spatial Statistics

Since our objective is to extract the tourists in the various areas shown in Fig. 2, the mobile spatial statistics need to be adjusted. For one, not all non-residents will be tourists and, secondly, the areas of interest do not match those for which data is available.

To address these problems, we tested two approaches. First, we extracted the visitors' data within the mesh, overlapping with the touristic areas. There are 890 meshes within Kyoto city out of which 289 overlap with the touristic areas shown in Fig. 2. The number of persons for each month within the 890 meshes is shown in Fig. 4, and the number of persons within the 289 “touristic meshes” in each month is shown in Fig. 5.

The figures illustrate the concentration of visitors on the touristic meshes, but clearly there are also non-tourists amongst the visiting population. We refer to this estimate of tourists as $\hat{P}$, noting that it will overestimate the number of tourists. The figures also illustrate that November is the busiest tourist month and December the least busy month. November is the month of autumn foliage in Kyoto, usually attracting large numbers of tourists. Instead, December, due to weather conditions and working and school schedules, is generally a month with relatively little domestic tourism. Hence as a conservative (underestimate) of the tourist population, we extracted the part of the tourists’ data in November by removing the number of visitors in December and refer to this as $\check{P}$ and show it in Fig. 6

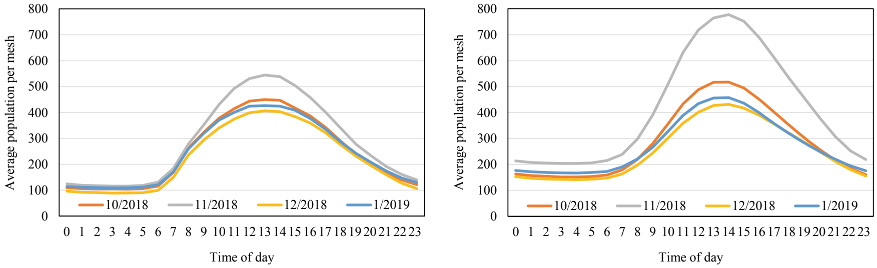
Instead of taking a mean of the two estimates, we aim at obtaining evidence as to which approach is more appropriate. First, we note that Kawakami et al. [28] used the same data as ours for tourist flow estimation and compared their approach with survey data and suggested that results from $\hat{P}$ matched well with the survey result.



(a)

(b)

Fig. 4 Number of visitors in all Kyoto meshes **a** on weekdays (left) **b** on holidays (right)



(a)

(b)

Fig. 5 $\hat{P}$ in each time of day **a** on weekdays (left) **b** on holidays (right)

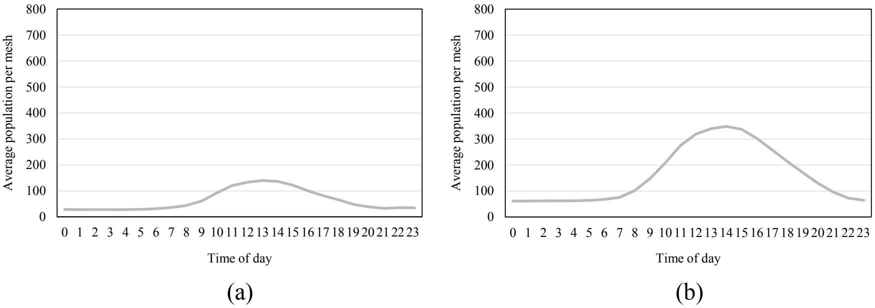


Fig. 6 $\check{P}$ in each time of day **a** on weekdays (left) **b** on holidays (right) in November

To obtain further evidence, we return to our second data set, the GPS traces. Also here, the problem of distinguishing visitors from the tourists remains. As an indicator of whether a person is likely a visitor, we consider the number of touristic places visited by the respondents over the period for which we have (infrequent) tracking records. Our hypothesis is that the number of recorded tourist areas per day referred to as μ, will tend to be larger for tourists. To find a suitable threshold, we conducted a sensitivity analysis comparing mesh data in the form $\hat{P}$ and $\check{P}$ as the dependent variable and GPS data as the independent variable with a LR. The results for different thresholds ranging from $\mu = 0$ to 2 with steps of 0.1 are shown in Fig. 7. We find that the best fit with $\hat{P}$ is achieved with $\mu = 0.3$. For $\check{P}$ the R^2 continuously increased, but the gradient of this became negligible when $\mu > 0.4$. A value of 0.3 might seem low, but our GPS data is sampled mostly only when the travelers used the app. Therefore, especially visits to neighboring, walkable attraction areas might be missed. Overall, considering these points, 0.3 appears to be reasonable. The number of persons based on GPS data with and without the threshold is shown in Fig. 8. In particular, we note that for weekdays the estimate with μ appears to be more realistic as clearly tourists tend to populate the touristic places during the day hours.

Overall, based on this analysis and the aforementioned research of Kawakami et al. [28], we concluded that $\hat{P}$ is a better estimate than $\check{P}$ for the number of tourists and that 0.3 tourist areas visited per day appears to be a suitable threshold to extract tourists from the longer-term GPS tracking data.

We conclude this section by noting the small number of observations we have in Fig. 9 for the GPS data, implying that if one wants to use the GPS data as a basis for tourist number estimation, additional information for appropriate scaling is required, which is our topic of discussion in the next section.

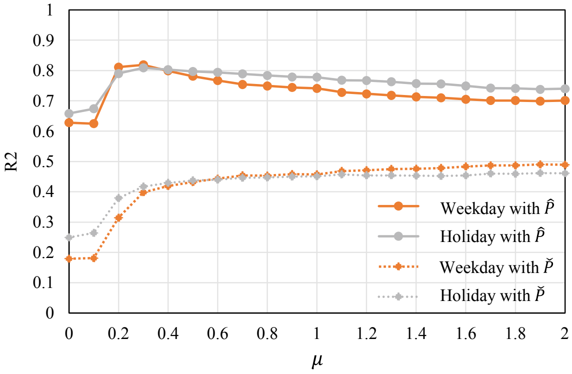


Fig. 7 Sensitivity analysis result with $\hat{P}$ and $\check{P}$

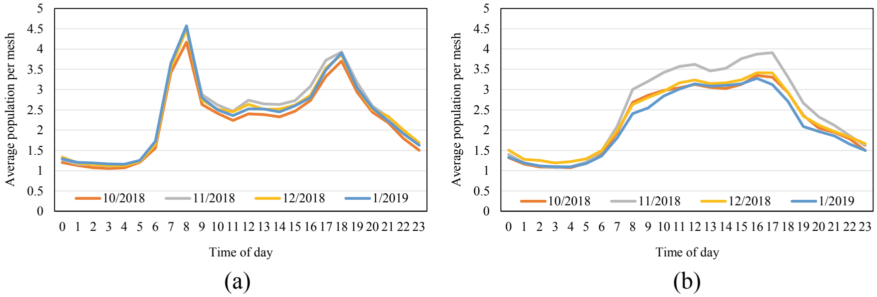


Fig. 8 Number of persons recorded in tourism areas with μ = 0 based on GPS records **a** on weekdays (left) **b** on holidays (right)

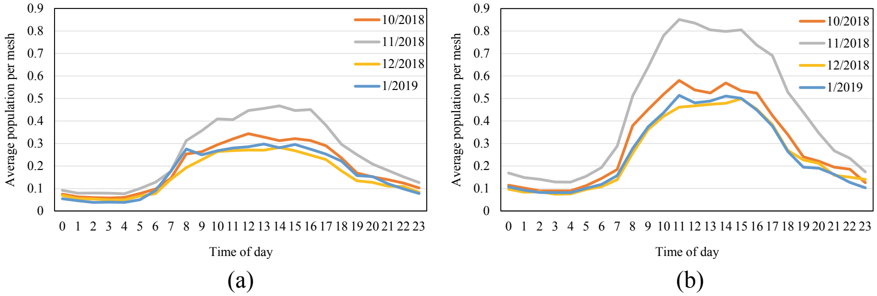


Fig. 9 Number of persons recorded in tourism areas with μ = 0 based on GPS records **a** on weekdays (left) **b** on holidays (right)

5 Tourist Number Estimation from GPS Traces and POI Information

5.1 *Linear Regression Model*

We first conducted LR to estimate the number of tourists. The dependent variable is the number of tourists from the mesh data per day, month, and time of day in each mesh, i.e., we consider this data as "true." The independent variables are the GPS records, the estimates of attractiveness based on POI numbers and ratings as well as the aforementioned accessibility indices. We selected the independent variables following the stepwise forward method. To avoid multicollinearity, we did not use variables with absolute correlation coefficients exceeding 0.4 simultaneously.

As a consequence, our preferred models all only have two significant uncorrelated remaining variables from the set of independent variables. One is $g_{i,m,t}$ the number of tourists from GPS data in mesh *i*, month *m,* and time of day *t* per day. The other is $s_{i,t}$ the weighted number of stations (WNS) in mesh *i* and time of day *t* weighted by type mesh attractiveness w_j^a. The LR results for weekdays are shown in Table 2. We keep the model with GPS data only in the table as it shows the correlation between the two data sets. We note that we tested additional models with a constant but found this constant to be insignificant. Both variables have the expected sign with the GPS data clearly being of more significance. Our composite attractiveness measure is, however, also highly significant and can contribute to explaining the differences between the two data sets.

We also show the LR results for holidays in Table 3. The results were mainly the same as on weekdays, but β of GPS data was smaller, and β of WNS was larger than the result on weekdays. The result suggests hence that the two data sets are less related on weekends and that the attractiveness of the POIs is more important in explaining the tourist number on weekends. Alternative interpretations could be related to different app usage which triggers the recording of a GPS location. For example, most tourists on weekdays probably live nearer Kyoto city and are more familiar than tourists on holidays. If so, most tourists on weekdays use the app only

Table 2 LR results on weekdays

	Model 1				Model 2			
	B	S.E	B	T	B	S.E	β	t
$g_{i,m,t}$ [10^3]	1.29	3.49	0.911	368^{b}	1.22	3.50	0.865	348^{b}
$s_{i,t}$ [10^{-2}]	–	–	–	–	1.64	2.91×10^{-4}	0.140	56.4^{b}
RMSE	424				402			
AIC	414,479.3				414,477.4			
N	27,744				27,744			

B: Non-standardized coefficient, S.E.: standardized error, β: standardized coefficient, ^{a}: p < 0.05, ^{b}: p < 0.01

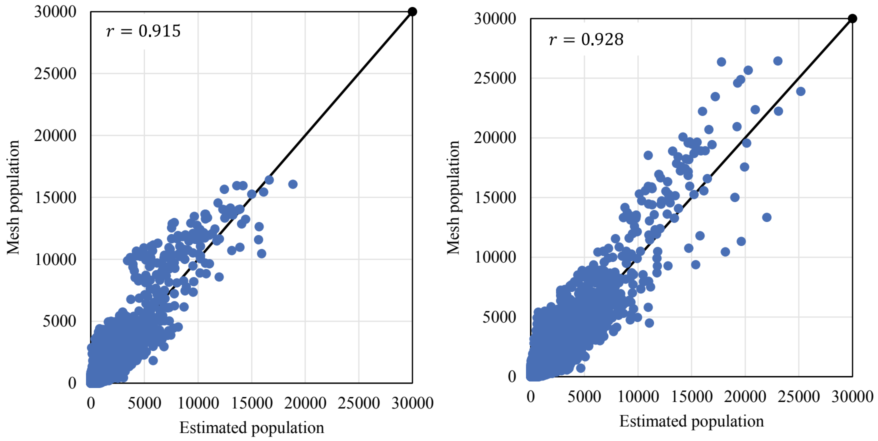
around the station just to know when trains arrive at and leave the station because they have a clear understanding of the transportation system. On the other hand, tourists on holidays utilize the app frequently to search for the best way to their next destination because they are not familiar with transportation system in Kyoto city. In this case, the amount of GPS data could converge on weekdays, and be dispersed on holidays, which clearly explains our obtained result.

The scatter plot, whose x-axis is the estimated population and the y-axis is the mesh population for Model 2, is shown in Fig. 10a, and the fit of Model 4 is shown in Fig. 10b. As can be seen, the results are satisfactory with correlation coefficients above 0.91.

Table 3 LR results on holidays

	Model 3				Model 4			
	B	S.E	B	T	B	S.E	β	T
$g_{i,m,t}$ [10^3]	0.979	2.76	0.905	355[b]	0.892	2.50	0.824	356[b]
$s_{i,t}[10^{-2}]$	–	–	–	–	3.54	3.35×10^{-4}	0.237	103[b]
RMSE	556				474			
AIC	429,466				429,465			
N	27,744				27,744			

B: Non-standardized coefficient, S.E.: standardized error, β: standardized coefficient, [a]: p < 0.05, [b]: p < 0.01



(a) (b)

Fig. 10 LR results **a** on weekdays (left) **b** on holidays (right)

5.2 Hierarchical Linear Models

Tourism behavior in Kyoto changes with the seasons. In particular red leaves in autumn and winter scenery with snow in January or February tend to attract tourists to different sites. In addition to this, the time distribution of GPS data can be different from that of mesh data because GPS data were collected when the app was used. Considering these, we test if the coefficient values vary by month and time of day using HLM.

HLM is a way of considering the fixed and random effects within groups of the whole sample. Consideration of random effects means that group-specific variables are estimated, whereas the assumption of fixed effects means a "global" variable for the whole data set. In HLM, the variances of each coefficient among the group are considered if the coefficient has a random effect. The coefficient of each group is assumed to follow the normal distribution with the fixed effects as average and the variance. All coefficients including constant values can have a random effect, so that we must decide for which coefficient it is more suitable to estimate a random effect when applying the HLM. We test different specifications and select the best model based on terms of minimal AIC [38].

Criteria for the appropriateness of using an HLM approach are the intra-class correlation coefficient (ICC) and the design effect (DE). ICC follows Eq. (2), and DE Eq. (3) [38].

$$ICC = \frac{\tau}{\sigma} \tag{2}$$

$$DE = 1 + (k - 1) \times ICC \tag{3}$$

τ is the variance among groups, and σ is the variance of all samples. k is the average number of samples in each group. The larger the ICC, the more significant the effect of dividing the samples. However, the ICC becomes small if the average number of samples in each group is large. On the other hand, the DE can consider the impact of the average number of samples in each group. It is commonly accepted that an ICC > 0.1 or a DE > 2.0 indicates the suitability of using HLM. Moreover, even if these criteria are not satisfied, applying HLM is reasonable when the application decreases the AIC [38].

We divided our samples into 4 groups based on months (October, November, December, and January) and 24 groups based on the hour of the day. When we divided our samples into monthly groups, we obtained ICC $= 0.0004$ and 0.0005 and DE $= 3.5$, DE $= 30$ for weekdays and holidays respectively. For the hourly models, the ICC values were 0.0172 and 0.011 as well as 20 and 14 respectively. Considering these results, our adoption of the HLM is reasonable.

The monthly results are shown in Table 4. The AIC became the smallest when the coefficient of GPS data had a random effect, and the coefficient of WNS data did not have a random effect. The comparison of the RMSE of the LR and HLM models is shown in Fig. 11. In this figure, for a fairer comparison, we divided RMSE by the average population per month because a higher average population is associated with a bigger improvement in RMSE. Based on this, we can know for each month the improvement of using the HLM. RMSE values slightly decreased compared to the LR results, but the difference in RMSE is not large. Considering these results, GPS data represents the effect of the month well. The range improvement is slightly more significant on weekdays than on holidays. The fact that only GPS data has a random effect suggests that the inconsistency of mesh data and GPS data is more significant on weekdays than on holidays but that this effect depends on the month. An explanation is that mesh data on weekdays includes more non-tourists than on holiday and that the size of this effect has a seasonal dependence.

The hourly results are shown in Table 5. Also here we find that adding random effects for the GPS records but not for $s_{i,t}$ is the preferred model.

Table 4 HLM results considering the month's effect

	Model 5(Weekdays)				Model 6(Holidays)			
	B	S.E	β	t	B	S.E	B	t
$g_{i,m,t}$ [10^3]	1.27	58.0	0.892	21.9[b]	0.906	16.3	0.831	55.5 [b]
$s_{i,t}$ [10^{-2}]	1.58	2.86×10^{-4}	0.138	55.8[b]	3.49	3.45×10^{-4}	0.240	101[b]
Variance of coefficient								
$g_{i,m,t}$ [10^3]	13.3				1.04			
RMSE	393				472			
AIC	410,249				420,398			
N	27,744				27,744			

B: Non-standardized coefficient, S.E.: standardized error, β: standardized coefficient, [a]: p < 0.05, [b]: p < 0.01

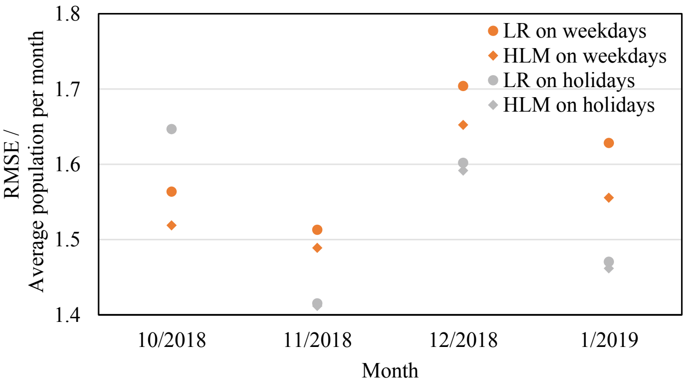


Fig. 11 Comparison of RMSE in Models 2, 4, 5, and 6 for the months Oct 2018–Jan 2019

Table 5 HLM results considering time of day effects

	Model 7 (Weekdays)				Model 8 (Holidays)			
	B	S.E	B	t	B	S.E	B	t
$g_{i,m,t}$ [10^3]	1.19	27.5	0.840	43.4[b]	0.963	36.7	0.883	26.2[b]
$s_{i,t}$ [10^{-2}]	1.63	2.84×10^{-4}	0.143	57.5[b]	3.32	3.15×10^{-4}	0.228	105[b]
Variance of coefficient								
$g_{i,m,t}$ [10^3]	17.5				32.1			
RMSE	390				429			
AIC	409,941				415,172			
N	27,744				27,744			

B: Non-standardized coefficient, S.E.: standardized error, β: standardized coefficient, [a] $p < 0.05$, [b] $p < 0.01$

The RMSE values decreased more significantly compared to the LR results for both weekdays and holidays. The comparison of the RMSE of the LR and HLM is shown in Fig. 12.

It can be observed that the difference in RMSE in the early morning was more significant than for other times of the day. The reason could be that many tourists used the app at that time of day. Further, comparing weekdays and holidays, the range of improvement on holidays is bigger than on weekdays possibly for the same reason. As additional evidence, we find that on weekdays the improvement is larger in earlier times of the day (predominantly 6–8 am.) than on holidays (8–10 am.) as presumably there is a larger proportion of non-tourists in the early morning weekday data. These effects are reflected in Fig. 13 in comparison to Fig. 10.

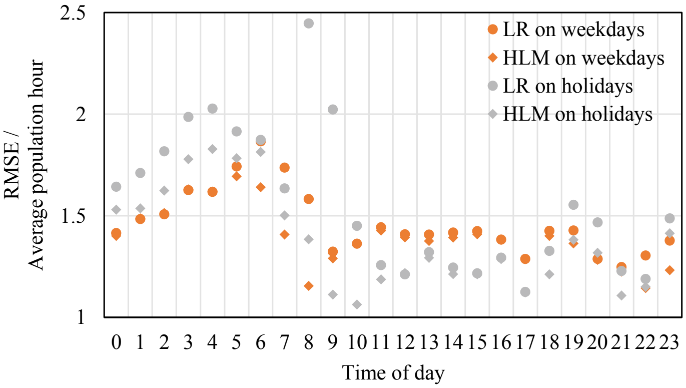


Fig. 12 Comparison of RMSE in Models 2, 4, 7, and 8 by time of day

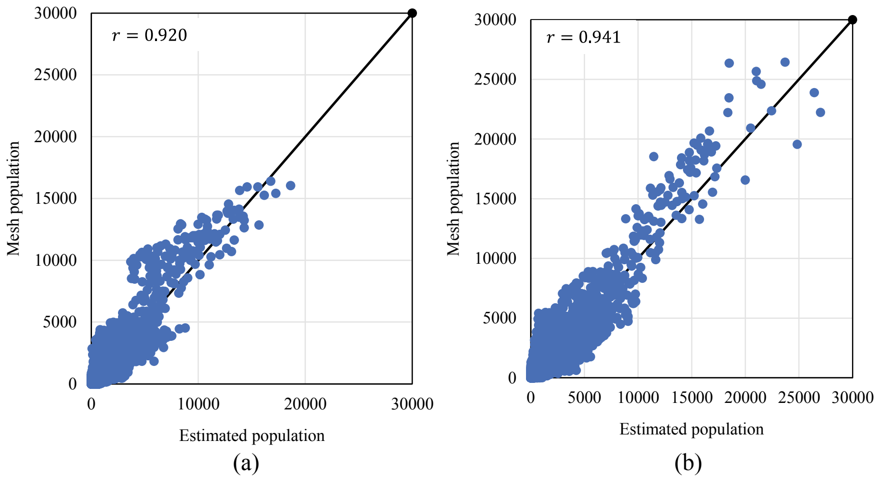


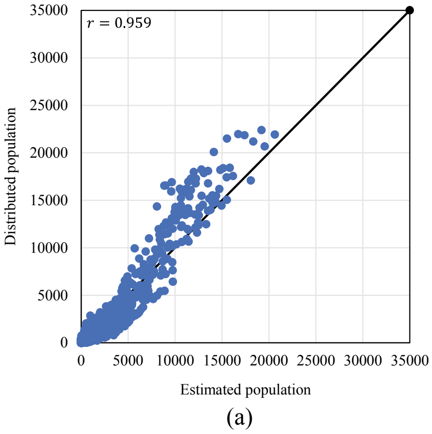
Fig. 13 HLM results **a** on weekdays (left) **b** on holidays (right)

5.3 Estimation of the Number of Tourists Within the Touristic Areas

In previous subsections, we estimated the LR based on the standard 1 km^2 meshes that are provided by the data provider. We now apply the preferred models found in previous subsections to the comparison between the GPS data (plus attractiveness and accessibility) models and the mesh data by considering the actual tourist areas shown in Fig. 2. Hence both data sets are adjusted to fit the revised areas.

The results of using the HLM models are shown in Fig. 14. The figure shows the comparison between the "estimated population" based on the HLM model and the "distributed mesh population" for each area, month and time of day. There are hence 3,552 $(= 37 areas \ \times 4 \ months \times 24h\ )$ points in each graph. The distributed mesh population is obtained by a simple estimate of multiplying the population of each mesh that overlaps with the target area with a "GPS overlapping ratio". This ratio is defined as the percentage of the GPS data that are within the target area part of the mesh compared to all GPS records found in this mesh. The "GPS overlapping ratio" is used instead of simply taking the area overlap itself to correct for cases where, for example, half of the mesh area is mountainous and not accessible. In that case most GPS records will be found in the target area part of the mesh and hence the ratio will be near one so that also all the mesh population will be assumed to be in the target area.

Due to various assumptions discussed in this and previous sections, we acknowledge hence that both "distributed" and "estimated" values could be different from the actual ones. However, our matching shown in Fig. 14 gives us some confidence that our values are not too far from the ground truth. In general, we observe a good model fit for the two methods that need to overcome very different data limitations.



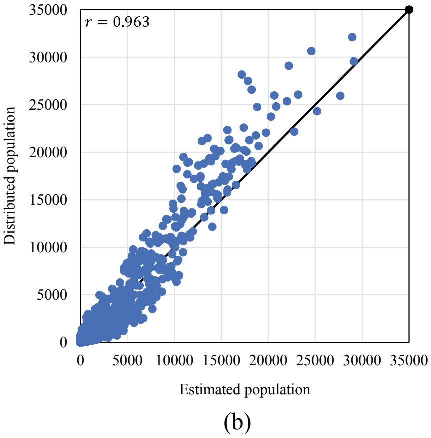


Fig. 14 Comparing the estimated and the distributed populations **a** on weekdays **b** on holidays

Among the areas with worse fit, that is a larger RMSE, are Areas 25 and 29 (refer to Fig. 2 and Table 1). These areas include stations with interchange between different train lines. Our estimated values might hence miss some tourists who only traverse this area but do not stay there for longer term. We further observe some errors in Areas 2, 35 and 37. These are far from the public transportation services and have a low number of tourists. Random under-sampling in these areas as well as systemic under-sampling of tourists coming by car to these areas–who are hence less likely captured with the GPS data from our travel planning app—might contribute to these errors.

6 Conclusion

Our study discussed the problem of estimating the number of tourists in specific areas and times given that most commonly available data sets are inadequate for this purpose. We suggest that, this is not only a common problem for many cities but also an often under-researched area by the transportation research community. Alternative methods to estimate the tourist population in specific areas at specific times are based on counting, tickets sales, hotel bookings, etc. However, in this paper, we showed that a range of map data in conjunction with relatively "small big data" could be a potentially useful alternative.

In the case of Kyoto City, as well as other cities, it might be possible to obtain visitor versus residential data, but clearly not all visitors are tourists. We first discuss how the mobile spatial statistics might be adjusted accordingly and then how an adjusted set of GPS data plus POI and accessibility data can be used to estimate the number of tourists. The mobile spatial statistics of non-residents, in general, appear to give a good estimate of tourist numbers, however, the estimate unsurprisingly appears to be better on holidays than on weekdays.

From the GPS tracking data, we found that taking a threshold of 0.3 tourist attraction visits per day is a reasonable threshold to distinguish tourists. As afore discussed, the value might seem low, but important to remember is that our sampling frequency is quite low such that a fairly large number of attraction visits are likely to be missed and that the data is recorded over a longer period such that also days without tourist activities could be included. One reason why the sampling frequency is quite low is that the GPS data is sampled mostly only when travelled used the app. Also, the utilization frequency of the app is for many tourists low, among others, because the attraction areas in Kyoto city are easily walked to and from. Therefore, if applied to other GPS tracking data sets, the threshold might have to be revised.

Our linear regression models matching the mobile statistics with GPS records plus additional information from the POIs and accessibility information generally, showed a good fit as illustrated by the R^2 values as well as the plot of the two estimates. The standardized coefficients of GPS data and the weighted number of stations differed between weekdays and holidays. This might suggest that the tendency to use the app which triggers the recording of the GPS locations is different on weekdays and holidays. The difference might also be due to different touring patterns on weekdays with a different weight also attached to accessibility. Yet another interpretation is again related to the different proportion of non-tourists in the data on weekdays.

The results improved further when using HLM. Dividing the samples by month also improved the model fit, though the increase was mostly insignificant. Instead, by dividing the samples based on the time of day, the model accuracy was improved more significantly. In both models, variable from GPS data have a random effect so that GPS data can have a bias based on month or time of day, and we could establish the model considering this bias. The range of improvement in the early morning, from 6 a.m. to 8 a.m. on weekdays and from 8 a.m. to 10 a.m. on holidays, was higher than that of other times during the day. These results also suggested different touring patterns on weekdays and holidays and different proportions of non-tourists in the data on weekdays and holidays.

We then applied the HLM model to estimate the tourist population within the touristic areas predefined by the Kyoto city government. Since there is no ground truth data for this estimation result, we compared the estimations by the two data sets. As a result, we gained some confidence that our estimation results are reasonable. Nevertheless, in future work, we certainly hope to obtain some observed data to obtain more evidence to affirm our present conclusions. In ongoing work, we are further utilizing the forementioned data to obtain additional information such as the stay duration in touristic areas and characteristics of tourist movements in the city.

Acknowledgements The first author was supported by JST SPRING, Grant Number JPMJSP2110. The data purchases were supported by JST SICORP projects, Grant Numbers JPMJSC1805 and JPMJSC20C4.

References

1. Imaizumi H (2020) Sustainable tourism for the resilience of vulnerable regions: Pro-poor tourism and over-tourism. J Econ 60(5・6):91–106. (in Japanese)
2. Kyoto city (2019) Kyoto sightseeing overall research, annual reports from 2001 to 2019. https://www.city.kyoto.lg.jp/menu2/category/22-6-0-0-0-0-0-0-0-0.html (in Japanese)
3. Schmöcker J-D (2021) Estimation of city tourism flows: challenges, new data and COVID. Transp Rev. Editorial 41(2):137–140
4. Bai Z, Wang J, Wang M, Gao M, Sun J (2018) Accuracy assessment of multi-source gridded population distribution datasets in China. Sustain 10(5)
5. Gao P, Wu T, Ge Y, Li Z (2022) Improving the accuracy of extant gridded population maps using multisource map fusion. GIScience & Remote Sensing 59(1):54–70
6. Bustos MFA, Hall O, Niedomysl T, Ernstson U (2020) A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015. Popul Environ 42(2):255–277
7. Calka B, Bielecka E (2019) Reliability analysis of landScan gridded population data. The case study of Poland. ISPRS Int J Geo-Inf 8(5)
8. Chen R, Yan H, Liu F, Du W, Yang Y (2020) Multiple global population datasets: Differences and spatial distribution characteristics. ISPRS Int J Geo-Inf 9(11)
9. Mattos ACH, Mcradle G, Bertlolotto M (2020) Assessing the quality of gridded population data for quantifying the population living in deprived communities. arXiv preprint arXiv:2011.12923
10. Seike T, Mimaki H, Hara Y, Odawara T, Nagata T, Terada M (2011) Research on the applicability of mobile spatial statistics for enhanced urban planning. J City Plan Inst Jpn 46(3). (in Japanese)
11. Balakrishnan K (2020) A method for urban population density prediction at 30m resolution. Cartogr Geogr Inf Sci 47(3):193–213
12. Bakillah M, Liang S, Mobasheri A, Arsanjani JJ, Zipf A (2014) Fine-resolution population mapping using OpenStreetMap points-of-interest. Int J Geogr Inf Sci 28(9):1940–1963
13. Shimosaka M, Hayakawa Y, Tsubouchi K (2019) Spatiality preservable factored Poisson regression for large-scale fine-grained GPS-based population analysis. In: Proceedings of the AAAI conference on artificial intelligence. pp 1142–1149
14. Azar D, Graesser J, Engstrom R, Comenetz J, Leddy RM Jr, Schechtman NG, Andrews T (2010) Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. Int J Remote Sens 31(21):5635–5655
15. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. EPJ Data Science 4(4):1–3
16. Kikuchi M, Iwadate K, Hato E, Mogi W, Kato M (2018) Practical method to update master data of parson trip survey in metropolitan areas using the transportation big data. Proc Jpn Soc Civ Eng 74(5):667–676 (in Japanese)
17. Otake T, Kikuchi A (2019) Development of a simulator system for travel demand forecasting with data assimilation. Proc Jpn Soc Civ Eng 75(5):607–613 (in Japanese)
18. Khodabandelou, G., Gauthier, V., El-Yacoubi, M., Fiore, M. (2016). Population estimation from mobile network traffic metadata. In: 2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM). pp 1–9
19. Cecaj A, Lippi M, Mamei M, Zambnelli F (2020) Forecasting crowd distribution in smart cities. In: IEEE international conference on sensing, communication and networking (SECON Workshops). pp 1–6
20. Bachir D, Gauthier V, El-Yacoubi M, Khodabandelou G (2017) Using mobile phone data analysis for the estimation of daily urban dynamics. In: IEEE 20th international conference on intelligent transportation systems (ITSC). pp 626–632
21. Aasa A, Kamenjuk P, Saluveer E, Šimbera J, Raun J (2021) Spatial interpolation of mobile positioning data for population statistics. J Locat Based Serv 15(4):239–260
22. Ahas R, Aasa A, Mark Ü, Pae T, Kull A (2007) Seasonal tourism spaces in Estonia: Case study with mobile positioning data. Tour Manage 28(3):898–910

23. Ahas R, Aasa A, Roose A, Mark Ü, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tour Manage 29(3):469–486
24. Ubukata Y, Sekimoto Y, Horanont T (2013) Availability as tourism statistical data of large scale and long term human mobility tracks by GPS: a study of Ishikawa Pref. Proc Jpn Soc Civ Eng 69(5):345–352 (in Japanese)
25. Kobayashi H, Zhang C, Schmöcker J-D, Nakao S, Yamada T (2021) Markovian analysis of tourist tours based on travel app data from Kyoto, Japan. In: Presented at 25th international conference of the Hong Kong society for transportation studies (HKSTS). December 12–14
26. Dantsuji T, Sugishita D, Fukuda D, Asano M (2017) Analysis of the properties of tourists' dwell time using Wi-Fi packet data a case study of the approach to Hase-Dera temple. J City Plan Inst Jpn 52(3). (in Japanese)
27. Nakanishi W, Kobayashi H, Tsuru T, Matsumoto T, Tanaka K, Suga Y, Kamiya D, Fukuda D (2018) Understanding travel pattern of tourists from Wi-Fi probe requests: a case study in Motobu Peninsula, Okinawa. Proc Jpn Soc Civ Eng 74(5):787–797 (in Japanese)
28. Kawakami R, Schmöcker J-D, Uno N, Nakamura T (2020) OD matrix estimation utilizing mobile spatial statistics with Kyoto tourism case study. Proc Jpn Soc Civ Eng 75(6):379–391 (in Japanese)
29. Gao Y, Schmöcker J-D (2022) Distinguishing different types of city tourists through clustering and recursive logit models applied to Wi-Fi data. Asian Transport Studies 8:100044
30. Takahashi K, Igarashi H (1990) Study on the recreation activity by recreation spot attractive index. Proc Jpn Soc Civ Eng 8:233–240 (in Japanese)
31. Kobayashi K, Sekihara Y (1991) Estimating the number of tourist visitors with destination-based surveys. Proc Jpn Soc Civ Eng 9:101–108 (in Japanese)
32. Mizokami S, Mogisugi H, Fujita M (1992) Modelling on the attraction of sightseeing area and excursion behavior. J City Plan Inst Jpn 27:517–522 (in Japanese)
33. Shen K, Schmöcker JD, Sun WZ, Qureshi AG (2022) Calibration of sightseeing tour choices considering multiple decision criteria with diminishing reward. Transportation. https://doi.org/10.1007/s11116-022-10296-7
34. Ishigami T, Kikuchi M, Inoue T, Iwadate K, Morio J, Ishii R (2017) Expectations and problems of traffic-related big data from a stand point of urban transport practical work. Jpn Soc Civ Eng 55. (in Japanese)
35. NTT Docomo Mobile spatial statistics. https://mobaku.jp/ (19 May 2022) (in Japanese)
36. Kyoto city. Fare table for Kyoto city subway. https://www.city.kyoto.lg.jp/kotsu/page/0000163782.html (in Japanese)
37. Kato H, Hashimoto T (2008) Mata-analysis on value of travel time savings in Japan. Jpn Soc Civ Eng 38. (in Japanese)
38. Simizu H (2017) Multilevel modelings for individual and group data. Nakanishiya Shuppan. (in Japanese)