Hourly Pedestrian Population Trends Estimation using Location Data from Smartphones Dealing with Temporal and Spatial Sparsity

Kentaro Nishi*
the University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan
nishi@ics.t.u-tokyo.ac.jp

Kota Tsubouchi
Yahoo Japan Corporation
9-7-1, Akasaka, Minato-ku,
Tokyo, Japan
ktsubouc@yahoo-corp.jp

Masamichi Shimosaka
the University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan
simosaka@ics.t.u-tokyo.ac.jp

ABSTRACT

This paper describes a pedestrian population trend estimation method using location data of smartphone users. This technique is intended to be an alternative to traffic censuses using tally counters. Traffic censuses using tally counters are still commonly used to survey the number of pedestrians despite their cost and limitations in area and time.

The proposed approach can replace the traffic census by using smartphone users' location data accumulated on Yahoo! Japan. Moreover, it is low cost because it uses location data collaterally acquired from smartphone users, and it has no limits in terms of area or time. This means pedestrian population trends in arbitrary areas and times about which we want to know can be estimated.

The proposed technique is based on the assumption that the number of location data in an area is proportional to the population volume, but it also eliminates some data to increase pedestrian accuracy. In the elimination step, some location data that should not be counted as pedestrians are excluded by estimating transport modes from anteroposterior location data. The supplement step tackles the problem of data shortage when a target area is a small region by using a Gaussian kernel. The Gaussian kernel smoother is also used to deal with data interpolation in the time direction, and it enables us to estimate time-continuous pedestrian volumes in arbitrary areas.

To evaluate the approach, a manual traffic survey was conducted in five areas on 11 days and the ground truth data are acquired. Experimental result shows the approach successfully estimate pedestrian population trends in areas. The proposed method makes less than one-tenth the mean squared errors of hourly pedestrian number estimation than the conventional approach.

Categories and Subject Descriptors

H.2.8 [**Information Systems Applications**]: Database Applications—*Data mining, Spatial databases and GIS*

Keywords

Pedestrian population trends estimation, gaussian kernel, traffic census

1. INTRODUCTION

Knowing the number of pedestrians in a time or place can be an essential data source for market research.

To survey the number of pedestrians, traffic censuses using tally counters are still commonly used. Though the data surveyed this way are very reliable, the traffic census approach has two problems. One is cost and the other is restrictions of research area and time. Finding out the total number of people requires many survey crews and much time. Moreover, we cannot know the population of areas in which and days on which the survey has not been conducted.

Rapid developments in technology of the Internet and smartphones are changing how we communicate with others and exchange information. Location-based services (LBSs) are widely used by smartphone users [8, 2]. LBSs (e.g. maps, navigation tools, train route searching, breaking emergency news) help us to obtain more useful information in accordance with our location. The more LBSs are used, the more location coordinates of smartphones are acquired. Mining data from cell-phones and smartphones is becoming common for developing networks and smartphones [16][3][5].

These data have recently been implemented to analyze urban systems [10]. Previous works studied population movements in time and space from location data of GPS traces or person-trip survey data [6, 12, 14]. Works have been done to visualize population density using cell-phone communications [9, 11].

However, to use GPS data from smartphones for counting

*This work was performed during a research internship at Yahoo! Japan Research, Tokyo, Japan.

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3131-9/14/11 ...$15.00
http://dx.doi.org/10.1145/2666310.2666391

the number of pedestrians, two problems remain.

First is how to pick out only pedestrians. As a related work, ZENRIN Datacom [1] provides a web service that visualizes area-by-area congestion estimated using location data from cellular phones by overlaying colors on a map. This service also provides hourly population histograms by using an about 200-meter mesh.

Pulselli et al. [9], Ratti et al. [10], and Reades et al. [11] also estimated the population density from the phone-call data. The position of users is replaced by the position of a cell phone base station. These researches estimate hourly population density by using an about 20-kilometer mesh.

These services and research cannot estimate pedestrian population volume, which requires estimating total population volume or population density. In terms of commercial applications (e.g. effect measurement of digital signage or market research), pedestrian volume must be estimated while eliminating people in offices or trains. Specifically, the data from this approach cannot estimate the number of pedestrians near train stations or big office buildings.

The second problem is how to set the research area and period at some different angles. Sekimoto et al. [13] visualized people flows in cities with social survey data from JICA (Japan International Cooperation Agency). In this research, the people flows of a developing country are shown by a 1-kilometer mesh. Liu et al. [6] also reproduced the mobility use pattern with GPS data from taxis and used ticket information. The resolution of their study is 1 hourly and each city scale. Calabrese et al. [4] estimated the population volume from phone-call data and the usage data of buses and taxis. The resolution of this study is 1 hourly and a 40-kilometer mesh.

In the above research, different resolutions were decided, meaning how large the mesh is and how long the time scale is. The number of data decreases when a mesh becomes small and the time period becomes narrow. However, market research such as the effect of digital signage and the estimation of the number of shop visitors requires the data that have a free mesh and a free time scale. Thus, the conventional traffic census surveys are still conducted even though they require much cost and much research time.

To overcome these problems, this paper describes a technique for estimating the pedestrian population trends in time and space continuously. Pedestrian population volume is focused on due to requirements of application usage. Moreover, it tackles the problem of data sparsity and limitations of time and area. This means continuous estimation of time and area with relatively little location data.

The technique is expected to be an alternative to conventional traffic censuses using tally counters. The proposed technique estimating pedestrian population volume tackles these problems using location data obtained from smartphone users. This means it is costless and can be applied anytime, anywhere.

A simple way to reconstruct the pedestrian population assumes that the number of records is proportional to the number of people present. However, this assumption does not take into account that the records contain stationary users in offices and passing users on trains.

This problem can be solved by developing an accurate alternative to traffic censuses by considering the records as point processes by each user. By doing this, transportation modes can be also estimated [15]. Non-pedestrian data are detected using estimated velocity and eliminated to estimate the pedestrian number accurately.

Furthermore, it is difficult to estimate pedestrian populations in small areas using the simple approach due to data shortage. Although the data are large in the macro scale, they are very limited in terms of estimating populations in areas scaling, for example, 100 m. Dealing with this sparsity is another problem to be solved.

The proposed technique is evaluated by comparing it with traffic censuses using tally counters, which were conducted in four areas in Japan. This study analyzes the data in an anonymous form and does not specify the individuals. It attempts to analyze and visualize the urban-scale dynamics.

Contributions of this paper are as follows:

- the estimation of the number of ”pedestrians” like traffic census research
- continuous estimation of time and area with relatively few location data
- evaluation by comparing with real traffic census data.

2. DATASET

The location data are provided by smartphone users through services of Yahoo! Japan with their agreement.

The data contain the following information:

- Latitude
- Longitude
- Horizontal accuracy
- Time stamp (at a second rate)
- Anonymized user ID (changes after 24 hours)

Table 1 shows several examples of the location data. Latitudes and longitudes are recorded in degrees. Horizontal accuracies are recorded in meters. Time stamps are recorded at a second rate in format “YYYYMMDDhhmmss”. The user IDs are anonymized and changed after 24 hours to protect privacy of users. It is impossible to identify individuals.

The number of records is approximately 15 million for each day. All data on a certain day are plotted in Figure 1. Note that no map is being overlaid in this figure. This shows that data are acquired from all over Japan.

Table 1: Examples of the location data

Latitude	Longitude	Accuracy	Time stamp	ID
34.8716480098989	135.661309193946	5.00	20130611000000	UID1
38.7213293603050	139.849629867952	30.00	20130611000000	UID2
43.0928880757489	141.371409950081	112.00	20130611000001	UID3
35.5574600559872	139.445981733805	14.88	20130611000000	UID4
35.7329128494678	139.670771645082	65.00	20130611000012	UID5
35.7846891648607	139.899813949837	10.53	20130611000001	UID6
35.6521350751540	140.026954640171	1414.00	20130611000002	UID7
35.6998792435734	139.841407947290	165.00	20130611000002	UID8

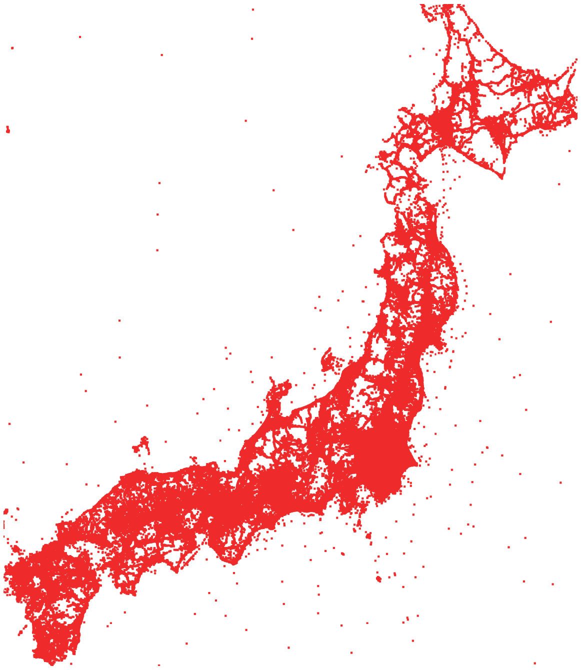


Figure 1: Collected location data points on May 20th, 2013.

The number of data is temporally biased due to the method of data acquisition. The data are acquired mainly when users use applications. In some applications, the data are acquired at regular intervals (e.g. once every hour) or when the connected base stations change. The temporal bias of the number of records is illustrated in Figure 2. It shows that the data are limited from 1 a.m. to 5 a.m. This is because people do not use applications during their sleep. Comparatively, many data are obtained during commuting hours. This must be because commuters use smartphones during travel to and from work.

3. METHODOLOGY

This section describes a method to estimate population as an alternative to traffic censuses using tally counters. First, a simple way to estimate the number of people in an area is explained and its limitations are discussed. Then a method to deal with the limitations using the user IDs and some stochastic techniques are proposed. Finally, parameter tuning using an actual manual traffic census with tally counters is described.

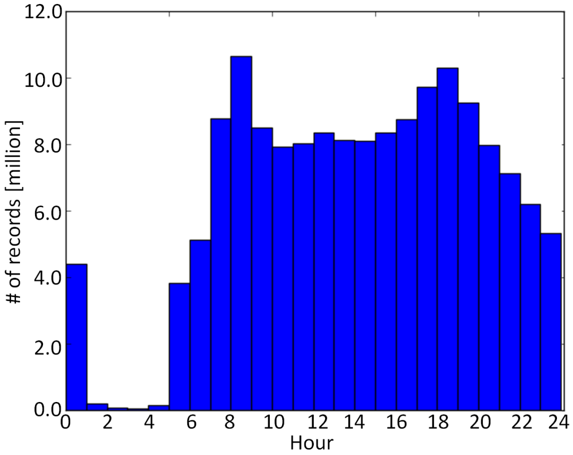


Figure 2: Temporal bias in the number of total records from May 20th through 26th 2013.

3.1 A simple approach

As described in Section 1, a simple method to estimate the population assumes that the number of records in an area is proportional to the people present in the area. In this method, first target areas in which and target days on which the population is estimated are specified for objectives such as market research. The target area is assumed as not a mesh but a polygonal area, and this step is implemented so that the area by specifying vertices (latitudes and longitudes) of the polygon like in Figure 3 can be defined.

As a second step, the number of records existing in the area hourly is counted and then multiplied by a proper factor. The factor is estimated by comparing the number with the actual traffic census described in Section 3.6.

In this approach, low-accuracy data are eliminated. The data that have under 300-meter accuracy (meaning the date has errors more than 300 meters) are eliminated in this research. The threshold of 300-meter is empirically introduced.

Each number represents the amount of traffic in the area and the hour. The number of location data records through a day visualized as a histogram is shown in Figure 4. In this figure, the vertical axis shows the location data count. This lets us know the rough population trend. For example,

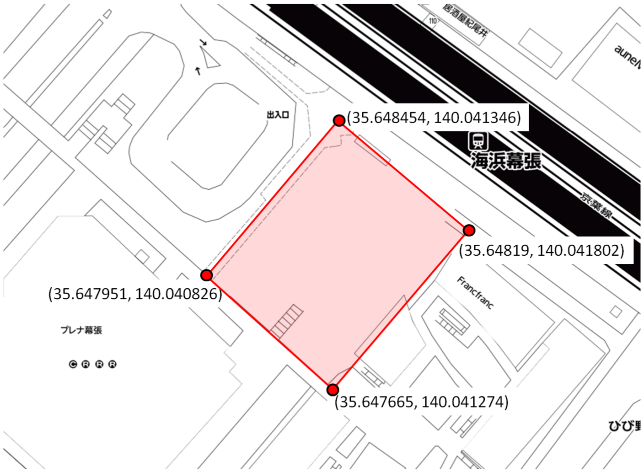


Figure 3: Way to define the target area by specifying the vertices of the polygonal region.

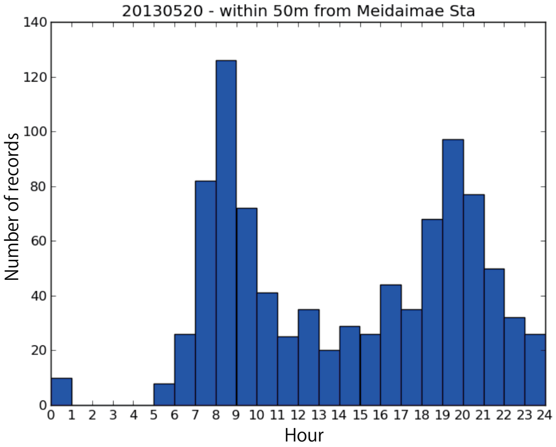


Figure 4: Histogram of number of location data records in an area near Meidaimae Station.

we know Meidaimae station, which is near a university, is mainly used during morning and evening commuting hours.

3.2 Limitation of the simple approach

This simple method has several problems as an alternative to traffic censuses using tally counters. One is that the records contain stationary users (e.g. in offices) and passing users (e.g. on trains). To estimate pedestrian volume, non-pedestrian records must be eliminated.

Another challenge is time continuous estimation. In the simple method, if the bin width of a histogram becomes small, the data become insufficient. Therefore, it is not flexible in terms of time. The number of pedestrians must be able to be estimated in arbitrary time ranges.

The other problem is shortage of the number of the records in accordance with the smallness of target areas. This is similar to the data shortage caused by short arbitrary time ranges. If we intend to estimate the number of people in a relatively narrow area, the number of records is also small. With the simple method, the estimated population volume becomes zero, an unusable result.

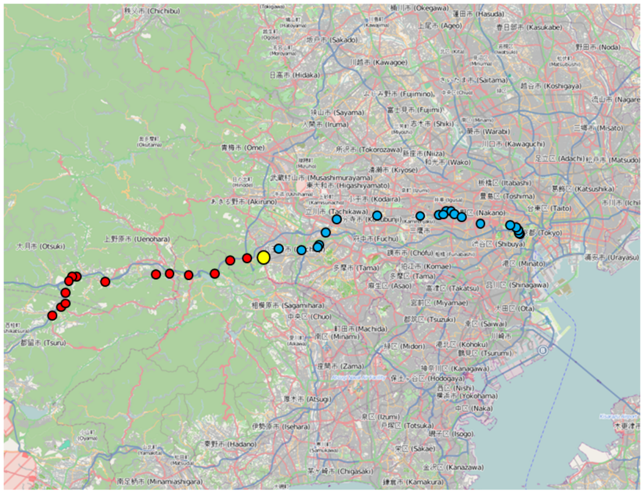


Figure 5: Automatically eliminated records (yellow) and the anteroposterior locations. Blue points and red points indicate data before and after reaching the target area, respectively.

3.3 Improvement method 1: Eliminating non-pedestrian data

To tackle the problem of eliminating non-pedestrian data, the user IDs are focused on and data are extracted user-by-user. In the simple method, data located in target areas are counted. The anteroposterior location of the data is checked using user IDs and timestamps, and approximate mean velocity is estimated an hour before or after a person visits the target area. The records where velocities both before and after are high or near zero are eliminated. The records where velocities both before and after are high are considered as invalid because the people must have been traveling in trains or cars. The records where velocities both before and after are near zero are considered to be of people staying somewhere (e.g. offices).

Figure 5 shows a sequence of records of someone that contain records automatically eliminated by this method. The blue points indicate locations before reaching the target area. The yellow point is the record eliminated, which is in the target area. The red points mean locations after passing the target area.

3.4 Improvement method 2: Realizing time-continuous estimation

The other problem of the simple method is that it cannot estimate the population continuously. The continuous estimation can also solve data sparsity problem in time because the data are supplemented from other nearly data in time. It estimates pedestrian volume with a certain bin width (e.g. one hour) as shown in Figure 4. If the bin width is narrowed, estimation becomes worse due to the number of records decreasing. This means the data are limited in terms of not only space but also time.

Poisson processes [8] are leveraged as a model to represent temporal transition of present pedestrians. Although a common probability model representing counting data is Poisson distribution[7] given by

$$p(m|\lambda) = \frac{1}{m!}\lambda^m \exp -\lambda \tag{1}$$

with a averaged number of counts in a fixed term interval λ, the model assumes the λ is constant. Poisson processes consider λ as a function of time $\lambda(t)$.

If the number of pedestrians is $m \in \mathbb{N}_0$ within t_1 and $t_2(> t_1)$, the likelihood of m can be written as :

$$p(m|\lambda(\cdot)) = \frac{1}{m!}\left[\int_{t_1}^{t_2} \lambda(t)\mathrm{d}t\right]^m \exp\left\{\int_{t_1}^{t_2} \lambda(t)\mathrm{d}t\right\} \tag{2}$$

The expected number of pedestrians within t_1 and $t_2(> t_1)$ is as follows:

$$E[m] = \int_{t_1}^{t_2} \lambda(t)\mathrm{d}t \tag{3}$$

This is the estimated pedestrian number between t_1 and t_2. Therefore, if $\lambda(t)$ can be estimated, so can the pedestrian population in an arbitrary time range.

Then, $\lambda(t)$ is decomposed with $f(t)$ and λ_0 as shown in Equation 4 by introducing $\lambda_0 = \int_0^T \lambda(t)dt$.

$$\lambda(t) = f(t)\lambda_0 \tag{4}$$

$f(t)$ represents the temporal probability density function because it satisfies Equation 5

$$\int_0^T f(t)dt = 1 \tag{5}$$

The continuous distribution of $f(t)$ is estimated using Gaussian kernel density estimation as shown in Equations 6 and 7. $t_i(i = 1, 2, 3...n)$ represents time when data are recorded, and n represents the total number of data points in the target area and on the target day. h (> 0) is a smoothing parameter called the bandwidth.

$$f(t) = \frac{1}{h_t n}\sum_{i=1}^{n} k(\frac{t-t_i}{h_t}) \tag{6}$$

$$k(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \tag{7}$$

λ_0 is decided as $\lambda_0 = Kn$. K is a factor that represents the actual people number per record. K denotes how many pedestrians are represented by a single data point. K can be set as constant or a function of target area.

Figure 6 shows an example of estimation of $\lambda(t)$ when h_t is set as 0.4 hours and K is set as 1.

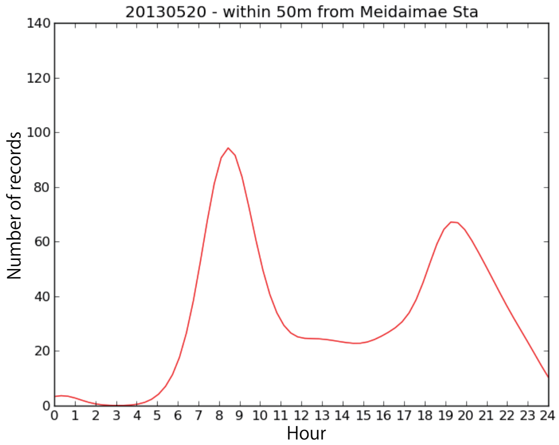


Figure 6: Example of continuously estimated $\lambda(t)$ on May 20, 2013 in an area near Meidaimae Station.

3.5 Improvement method 3: Dealing with data sparsity in space

As well as data shortage in terms of time, if the target area becomes narrow, the data become insufficient in terms of space.

An approach similar to time continuous estimation is proposed. A Gaussian kernel is used to supplement data from around an area. This approach implicitly assumes the pedestrian population trend in a certain area is similar to the pedestrian population trend in its surrounding area. This means the data in surrounding area are taken into account with smaller weight because the data in the surrounding area are less reliable than the data in the target area.

The data's weight w_i is decided as 1 if the data are in the target area where i is the index of the data. If the data are near the target area, the weight w_i is set as follows:

$$w_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(d/h_d)^2} \tag{8}$$

i is the index of data. d denotes the distance between the data point and centroid of the target area. h_d is a constant number that represents the effect of distance. This value is set as 100 meters in the experiment, but the h_d can be changed as the objective of the data usage research.

In this approach, each data point is weighted with w_i $(0 \leq w_i \geq 1)$. Therefore, Equation 6 is changed as

$$f(t) = \frac{1}{h_t \sum w_i}\sum_{i=1}^{n} w_i k(\frac{t-t_i}{h_t}) \tag{9}$$

In this Equation, n denotes all data points around target points. If the point is far from target points, the points are declined because w_i is calculated as 0. λ_0 is decided as $\lambda_0 = K\sum w_i$.

3.6 Parameter selection

Actual traffic censuses using tally counters were conducted for this research. These census results are used as reference data to determine parameters and to evaluate the accuracy of the proposed approach.

The parameters to be determined are K, which is a factor to relate the number of records to the number of people by multiplying, that is, how many pedestrians are represented by a single data point.

The parameter is determined by comparing it with reference data for each area as follows:

$$\hat{K}_p = \frac{M_p}{\left(\sum_i^n w_i\right)\left(\int_{t_1}^{t_2} f(t)\mathrm{d}t\right)} \tag{10}$$

t_1 and t_2 are the census start and end times, respectively. $\hat{K}_p$ is the ideal factor that enables correct estimation of the total number of pedestrians in the day. The denominator is an estimated pedestrian number if K = 1 and the numerator is an actual number.

M_p is total pedestrian number according to traffic census data of the day. p is the index of area.

K is determined as the mean of K_p as follows:

$$K = \frac{1}{l}\sum_p^l K_p \tag{11}$$

l is the total area number with reference data.

4. RESULTS

4.1 Reference data

Traffic censuses using tally counters were conducted in five areas on 11 days in Japan in September 2013 as shown in Table 2. Marunouchi is an urban business district. Meidaimae is near a university. Kaihin Makuhari has a big concert hall. Roppongi Intersection is the most crowded area in these surveyed areas, and it is above a subway station.

There are other methods for acquiring reference data, for example Bluetooth sensing or headcount-cameras, etc. However, these ways have privacy problem and the reference data are inaccurate because of the difficulty of the definition of sensing area.

The obtained information is the number of both inbound and outbound people, their genders, and their ages. The numbers are recorded every 20 minutes. An example of a surveyed area is shown in Figure 7. In the Roppongi Intersection area, 20 survey crews (10 possible comings and goings segments and inbound/outbound) counted the pedestrians passing through the blue segment. Genders and ages of pedestrians are also recorded. Figure 8 shows a snapshot of a pedestrian traffic survey. This shows D and E segments in Figure 7. This region is about 250 m^2. In other areas, the crew number was changed on-demand. The width and shape of research differ by area, as do the data.

Figure 9 shows all reference data obtained manually using tally counters in five areas. Note that the scale of vertical axes is not same and that time-length differs between survey areas. There are no data outside of research times. This does not mean no pedestrians were there in these periods.

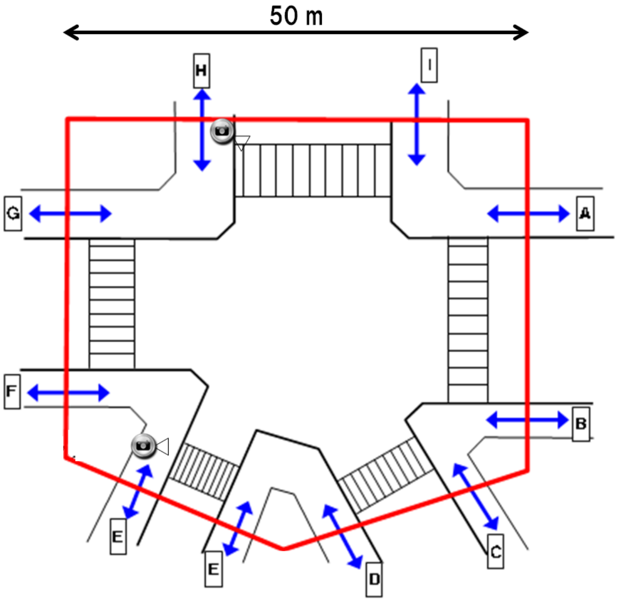


Figure 7: Surveyed region in Roppiongi Intersection area.

4.2 Evaluation metrics

The following three metrics are useed to assess the estimation results.

1. Mean squared error of hourly estimation
2. Absolute error of peak time
3. Absolute error of daily total number

In all approaches, evaluation is calculated hour-by-hour to compare them with non-time-continuous approaches. The first one is the mean squared error of hour-by-hour estimated number represented as

$$\text{metric } 1 = \frac{1}{t_2 - t_1}\sum_{t=t_1}^{t_2-1}(\hat{m}_t - m_t)^2 \tag{12}$$

t_1 and t_2 are the start and end times of manual census in hours. $\hat{m}_t$ and m_t are estimated and actual pedestrian numbers from t to $t+1$.

The second one is about the problem to estimate peak hours in the day and area. Mean absolute errors of estimated peak and actual peak are explained as:

$$\text{metric } 2 = |\arg\max_t(\hat{m}_t) - \arg\max_t(m_t)| \tag{13}$$

The last one is total volume error. The estimated total number of pedestrians in the day and area and actual data are compared. This value is represented as:

$$\text{metric } 3 = |\sum_{t=t_1}^{t_2}\hat{m}_t - \sum_{t=t_1}^{t_2} m_t| \tag{14}$$

Table 2: Date and time of traffic census using tally counters. All censuses were conducted in 2013.

	Sep. 7 (Sat.)	Sep. 8 (Sun.)	Sep. 10 (Wed.)	Sep. 12 (Thu.)	Sep. 14 (Sat.)
Marunouchi	8:00-22:00	8:00-22:00	8:00-22:00	8:00-22:00	-
Meidaimae	8:00-22:00	8:00-22:00	8:00-22:00	8:00-22:00	-
Kaihin Makuhari A area	-	-	-	-	12:00-22:00
Kaihin Makuhari B area	-	-	-	-	12:00-22:00
Roppongi Intersection	-	-	-	8:00-22:00	-

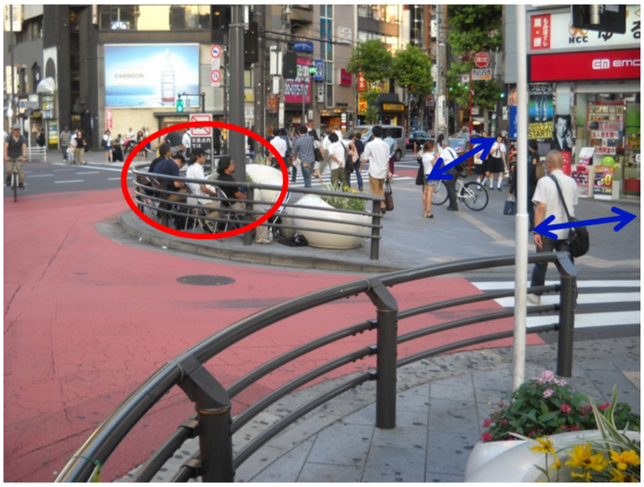


Figure 8: Snapshot of pedestrian traffic census in Roppiongi Intersection area. Survey crews are indicated with red circles. They counted pedestrians passing the blue segment, which are D and E segments in Figure 7.

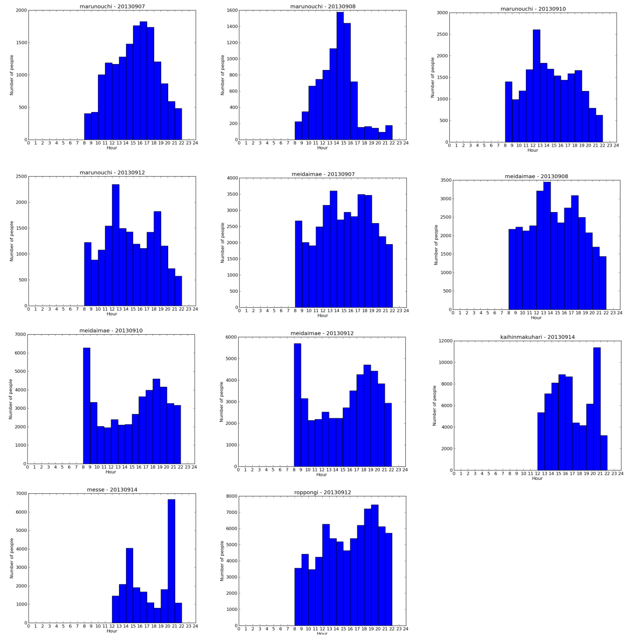


Figure 9: Reference data manually obtained by traffic census using tally counters.

For all metrics, the smaller value indicates the better estimation.

4.3 Proposed pedestrian estimation method

Hourly population trends are estimated with three approaches. The parameter K is determined as Equation 10 and 11 from manual census data without the census data of target area itself because it was used for evaluating results. The results are in Table 3.

The proposed approach uses both spatial and temporal kernels and a data elimination step. The estimated results with this approach are shown in the left side of Figure 10.

To evaluate the effects of the proposed method, two types of comparative approach are prepared.

Approach 1 is the simple approach. This is a method for counting the number of location data in a target area and hour and multiplying K. The estimated results with this approach are shown in the center of Figure 10.

Approach 2 is almost the same as the proposed approach using time continuous method and data supplement using a spatial kernel approach, but a data elimination step is not applied. The estimated results with this approach are shown in the right side of Figure 10.

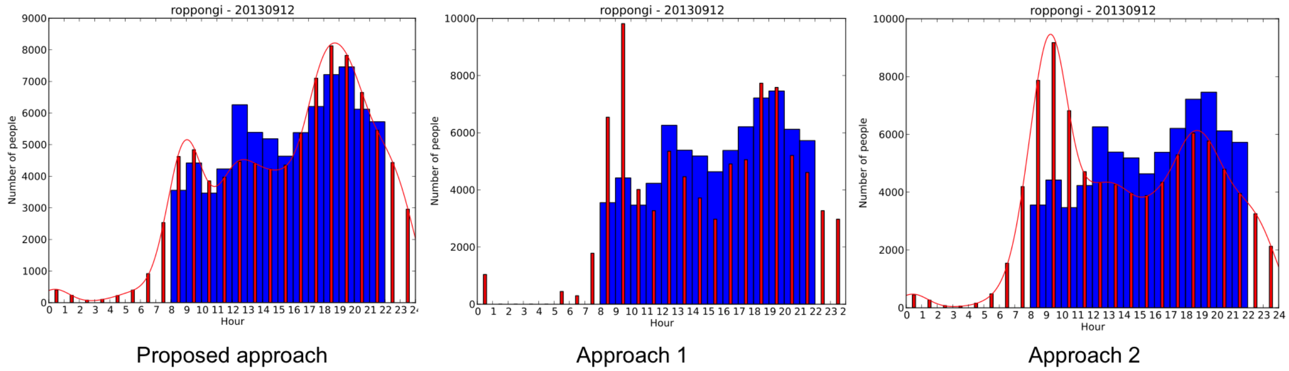
5. DISCUSSION

As shown in Table 3, the proposed method indicates the best performance in terms of all prepared metrics.

In this chapter, the results for each approach are discussed while observing the estimated histogram overlapped in actual census data. Note that in the following figures K is estimated by its own reference data to focus on pedestrian population trends, which means estimation is calculated as the total volume becomes correct.

The limitation of the simple approach (approach 1) is that it cannot deal with data sparsity. In fact, although over 15 million data points are acquired from all over Japan, in the Roppongi Intersection area, there are at most 70 records in one hour. In the Marunouchi area, there are under 10 records in one hour. It is very difficult to obtain reliable estimation results in small area with such little data. Table 3 shows that the simple approach cannot estimate the pedestrian number accurately.

When the pedestrian population in a narrow area or short time period is estimated, it is best to use the proposed approach, which uses kernel density estimation in terms of



Proposed approach

Approach 1

Approach 2

Figure 10: Hourly pedestrian population in Roppongi region estimated by proposed approach, approach 1, and approach 2. Blue histogram is reference data with manual census, and red bars are estimated pedestrian numbers.

Table 3: Evaluation results. Each value shows mean of the metric for all areas.

Approach	simple	elimination	kernel	metric 1 $[10^4]$	metric 2	metric 3
Approach 1	o	-	-	3578.7	4.00	32329
Approach 2	o	-	o	582.6	3.09	19702
Proposed	o	o	o	**284.9**	**2.55**	**12933**

time and space. According to Table 3, we can obtain pedestrian population trends with one-tenth of the mean squared hourly error of the simple approach.

The contribution of the elimination step is shown in Approach 2 (right in Figure 10) and the proposed approach (left in Figure 10). In approach 2, a strange peak is detected. This must be because of data from passengers on subways (there is large subway station under the Roppongi research area). In the proposed approach, the strange peak is not detected. This shows the elimination approach successfully detects the transportation mode of sequential data and eliminates data of non-pedestrians.

Figure 11 shows estimation results using the proposed approach for all areas and days. In these graphs, K is estimated with cross-validation. In most areas and on most days, pedestrian population trends are well extracted, but K is still challenging to estimate accurately.

Estimation results shown in Figure 11 are generally accurate, but the results for Kaihin Makuhari (20130914) and Messe (20130914) are not. The peak during 20 - 21 p.m. does not appear in the estimation. According to post research, a famous musician played a concert at Messe (near Kaihin Makuhari station) on that day. During the concert, most audience members turned off their cell phones so as not to distract from the music. Thus, the original GPS data uploaded by concert goers' cell phones cannot be confirmed on the Yahoo! Japan service. This is why the peaks in two places cannot be reconstructed correctly. This research has limitation in such unusual case that users turn off their mobile phone. To attack to this problem, the data supplement approach from data before and after concert is expected to perform well. This is one of future work.

Another future works will include good estimation of K. In this research, K is determined from reference data except for the target area. If area-by-area factors using features extracted from location data are well estimated, pedestrian populations will be estimated more and more reliably using only location data.

Moreover, to make the estimation accuracy high, the horizontal accuracy in raw data is expected to be useful. In this paper, data over threshold in terms of horizontal accuracy are eliminated. However, a more complicated approach can be considered. For example, w_i can be changed into the function of accuracy since low accuracy data should not be treated the same as high accurate data.

6. CONCLUSION

In this paper, location data from smartphone applications were studied for monitoring the number of pedestrians. Monitoring the pedestrian volume in areas could be a tool for analyzing urban design, i.e., observing daily population trends before and after construction of new infrastructure or stores will let us know the construction's impact. The total number of pedestrians of a day and its hour-by-hour transition gives us useful information to decide where new shops, parking spaces, or ATMs should be put.

The technique is intended to be an alternative to traffic censuses using tally counters. The proposed method has advantages in the cost and flexibility compared to manual censuses because it uses opportunistically acquired location data. Though our approach is based on the assumption that the number of location data in an area is proportional to the population volume, to estimate numbers of only pedestrians, the data from non-pedestrians are eliminated using estimated transportation modes. Moreover, our approach deals with temporal and spatial sparsity of data by introducing non-homogeneous Poisson processes and Gaussian kernels. This enables us to estimate the pedestrian numbers in arbitrary areas and times.

To evaluate the proposed approach, manual traffic surveys are conducted in five narrow areas in Japan. Experimental results showed that the proposed approach successfully estimates the number of pedestrians better than other possible approaches.

7. REFERENCES

[1] ZENRIN Datacom (in Japanese). http://lab.its-mo.com/densitymap/index.html.

[2] P. Adams, G. Ashwell, and R. Baxter. Location-based services - an overview of the standards. *BT Technology Journal*, 21(1):34–43, 2003.

[3] V. Bogorny and S. Shekhar. Spatial and spatio-temporal data mining. In *Proceedings of IEEE 10th International Conference on Data Mining*, pages 1217–1217. IEEE, 2010.

[4] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, March 2011.

[5] V. Frias-Martinez, V. Soto, J. Virseda, and E. Frias. Computing cost-effective census maps from cell phone traces. In *Proceedings of the 2nd Workshop on Pervasive Urban Applications*, 2012.

[6] L. Liu, A. Biderman, and C. Ratti. Urban mobility landscape: Real time monitoring of urban mobility patterns. In *Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management*, pages 1–16, 2009.

[7] D. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[8] J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

[9] R. Pulselli, P. Ramono, C. Ratti, and E. Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *International Journal of Design and Nature and Ecodynamics*, 3:121–134, 2008.

[10] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[11] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, July 2007.

[12] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui, and Y. Shimazaki. PFlow: Reconstructing People Flow Recycling Large-Scale Social Survey Data. *IEEE Pervasive Computing*, 10(4):27–35, Apr. 2011.

[13] Y. Sekimoto, A. Watanabe, T. Nakamura, and T. Horanont. Digital archiving of people flow by recycling large-scale social survey data of developing cities. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIX-B2:101–106, 2012.

[14] Y. Sekimoto, A. Watanabe, T. Nakamura, T. Usui, and H. Kanasugi. Digital archiving of people flow using person trip data of developing cities. In *Proceedings of The First Workshop on Pervasive Urban Applications*, June 2011.

[15] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM, 2011.

[16] H. Yang and S. Parthasarathy. Mining spatial and spatio-temporal patterns in scientific data. In *Proceedings of 22nd International Conference on Data Engineering Workshops*, pages x146–x146. IEEE, 2006.

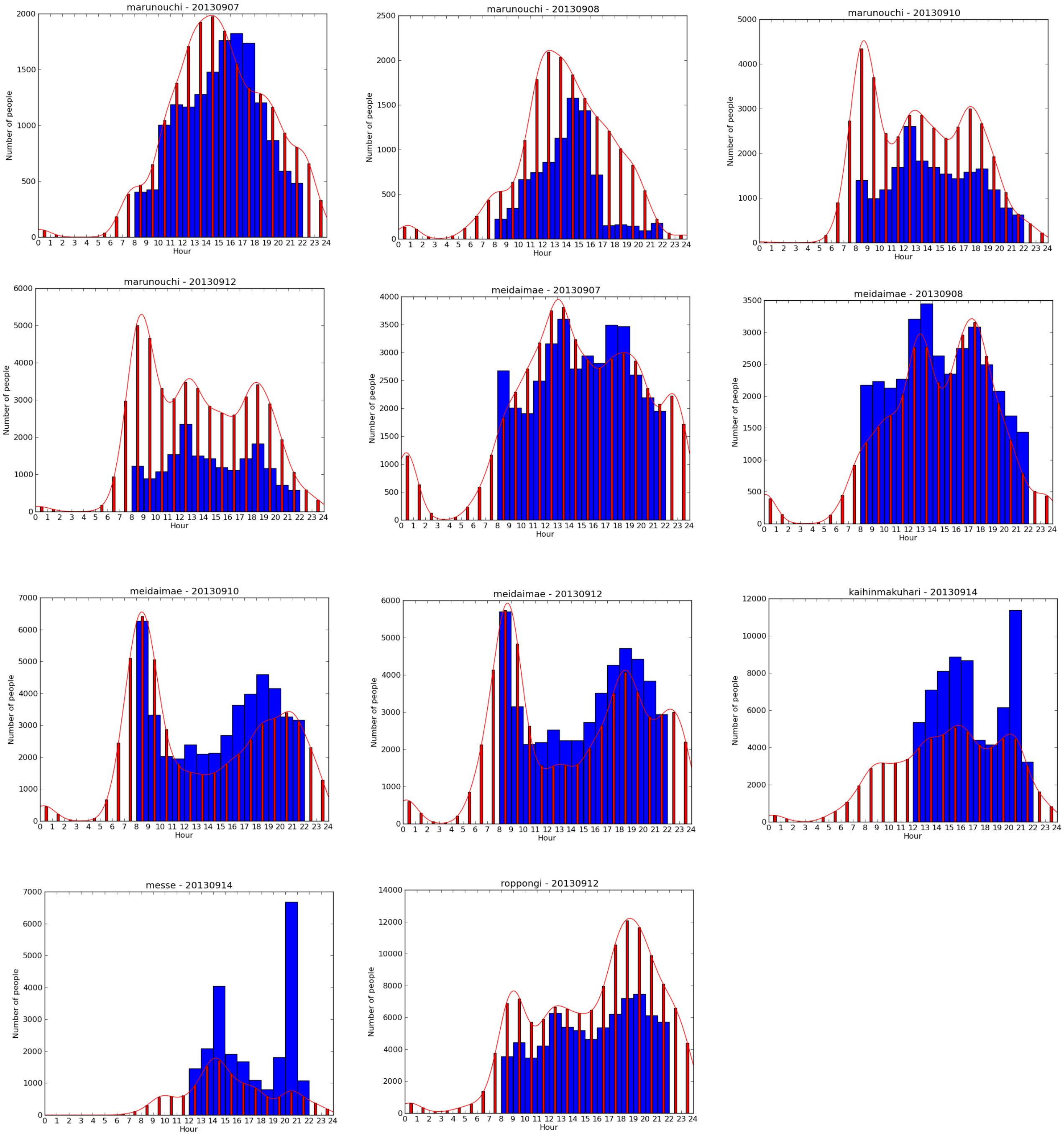


Figure 11: Final estimated results. Blue histograms are reference data from manual censuses, and the red lines and histogram are estimated pedestrian population trends. The estimation results are confirmed as histograms for comparing true data (tally counter data). In application, the estimations are the red lines, and the number of pedestrians is calculated as the area of the curve in the free time window.