Hourly Pedestrian Population Trends Estimation using Location Data from Smartphones Dealing with Temporal and Spatial Sparsity [1]

Kentaro Nishi[1] Kota Tsubouchi[2] Masamichi Shimosaka[1]

[1]the University of Tokyo

[2]Yahoo Japan Corporation

May 18, 2015

[1]Powered by pandoc and XƎLaTeX, Presented by Ziyang Chen.

Outline

Abstract

- Describes a pedestrian population trend estimation method using location data of smartphone users.
- Intended to be an alternative to traffic censuses using tally counters.
- Using smartphone users' location data accumulated on Yahoo! Japan.
- Tackles the problem of data shortage when a target area is a small region by using a Gaussian kernel.

Background

- Knowing the number of pedestrians in a time or place can be an essential data source for market research.
- Triditional method: Traffic censuses using tally counters are still commonly used.
  1. Requires many survey crews and much time.
  2. Temporal and spatial limitation.
- Location-based services (LBSs) are widely used by smartphone users
- two problems remain
  1. How to pick out only pedestrians.
  2. How to set the research area and period.

Outline

Data Set

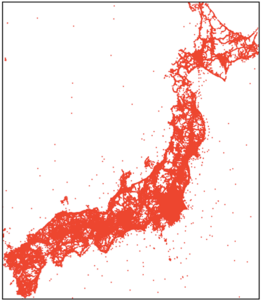
- The location data are provided by smartphone users through services of Yahoo! Japan, contains:
  - Latitude
  - Longitude
  - Horizontal accuracy
  - Time stamp (at a second rate)
  - Anonymized user ID (changes after 24 hours)

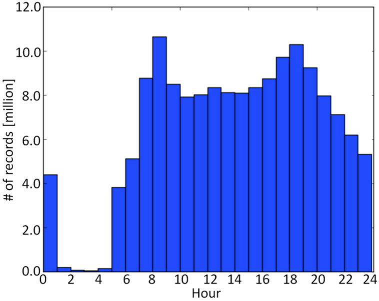
Data Set

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

Table: Examples of the location data

Data Set



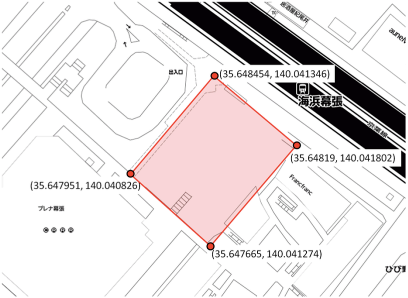

12.0
10.0
8.0
6.0
4.0
2.0
0.0
# of records [million]
0
2
4
6
8
10
12
14
16
18
20
22
24
Hour

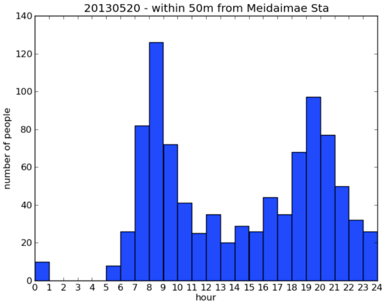
Outline

A simple approach

- Assumes that the number of records in an area is proportional to the people present in the area.
  - Determine target areas(polygonal) and target days
  - Number of records existing in the area hourly is counted and then multiplied by a proper factor.
  - The date has errors more than 300 meters are eliminated.

A simple approach

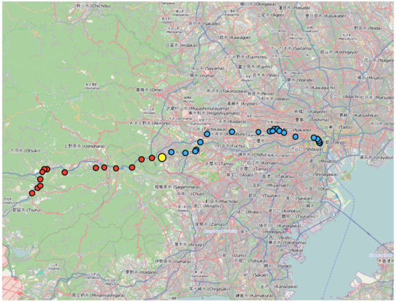




Limitation of the simple approach

- Non-pedestrian records
  - stationary users (e.g. in offices)
  - passing users (e.g. on trains)
- Time continuous estimation
- Sparsity with the smallness of target areas.

Eliminating non-pedestrian data



- Data are extracted user-by-user(by user id).
- 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.

Realizing time-continuous estimation

- Use Poisson distribution to represent counting data

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

- Consider λ as a function of time $\lambda(t)$

$$p(m|\lambda(\cdot)) = \frac{1}{m!}E[m]^m e^{-E[m]}$$

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

Realizing time-continuous estimation

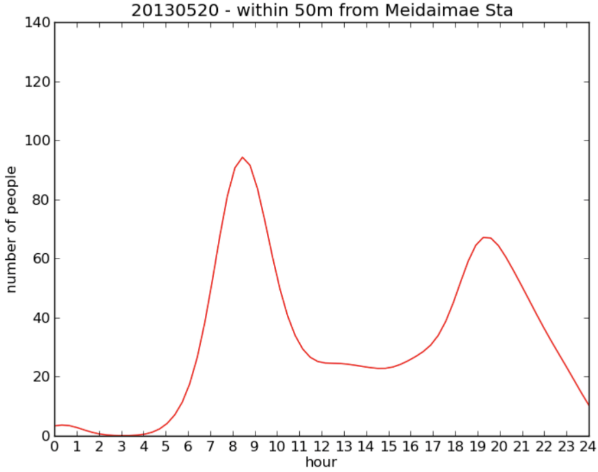
- Decomposed $\lambda(t)$ into $f(t)$ and λ_0

$$\lambda_0 = \int_0^T \lambda(t)\mathrm{d}t \quad \lambda(t) = \lambda_0 f(t) \quad \int_0^T f(t)\mathrm{d}t = 1$$

- $f(t)$ is estimated using Gaussian kernel density estimation

$$f(t) = \frac{1}{h_t n}\sum_{i=1}^{n} k\left(\frac{t - t_i}{h_t}\right) \quad k(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

Realizing time-continuous estimation



Dealing with data sparsity in space

- Data's weight $w_i = 1$ if the data are in the target area
- If the data are near the target area

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

- d is distance between data point and target area centroid
- $f(t)$ changed into

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

Parameter selection

- Traffic censuses using tally counters were conducted for this research.
- The parameters to be determined are K, means how many pedestrians are represented by a single data point.

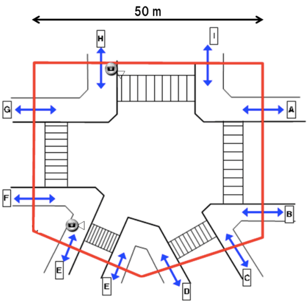
$$\hat{K}_p = \frac{M_p}{(\sum_{i=1}^{n} w_i)(\int_{t_1}^{t_2} f(t)\mathrm{d}t)}$$

- t_1 and t_2 are the census start and end times
- M_p is total pedestrian number according to traffic census data of the day.

Outline

Reference data

- Traffic censuses using tally counters were conducted in five areas on 11 days in Japan in September 2013

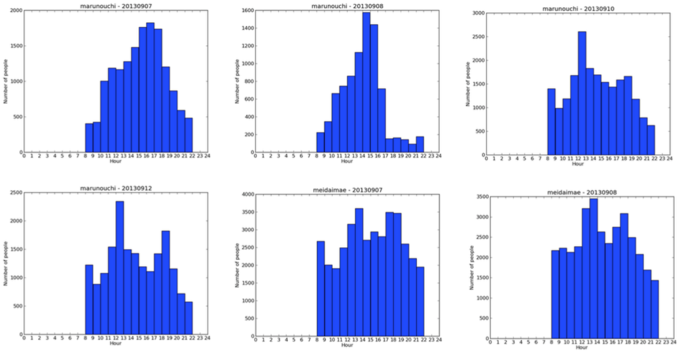


Reference data

	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	-

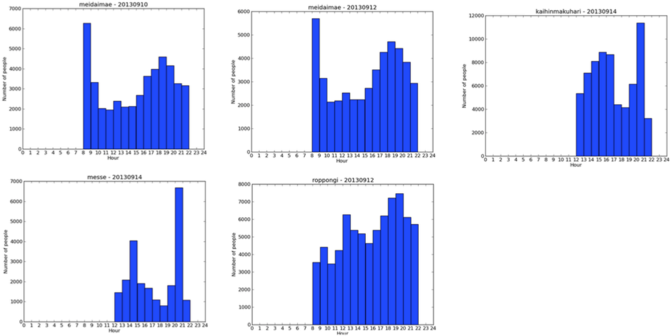
Table: Date and time of traffic census using tally counters. All censuses were conducted in 2013.

Reference data



Reference data manually obtained by traffic census using tally counters

Reference data



Reference data manually obtained by traffic census using tally counters

Evaluation metrics

1. Mean squared error of hourly estimation

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

2. Absolute error of peak time

$$metric\ 2 = |arg_t\, max(\hat{m}_t) - arg_t\, max(m_t)|$$

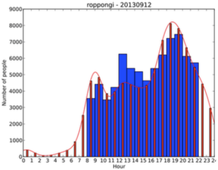
3. Absolute error of daily total number

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

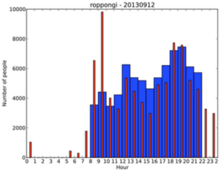
Comparative approach

- **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.
- **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.

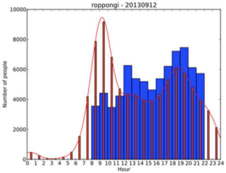
Result



Proposed approach



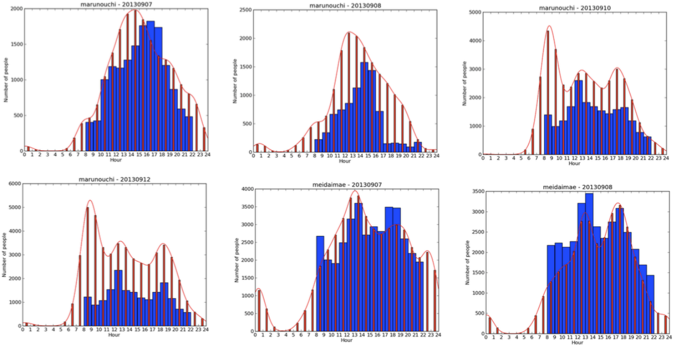
Approach 1



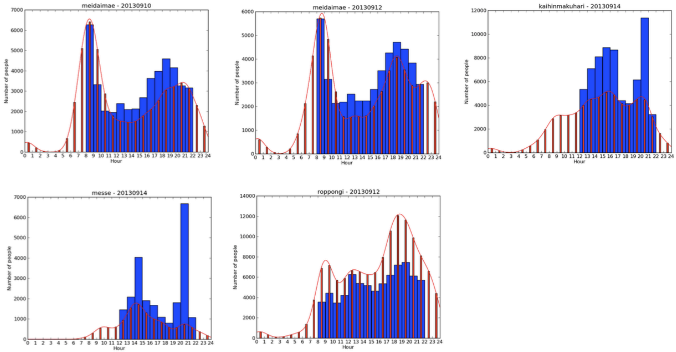
Approach 2

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**

Result


marunouchi - 20130907
marunouchi - 20130908
marunouchi - 20130910
marunouchi - 20130912
meidaimae - 20130907
meidaimae - 20130908
Number of people
Hour

Result



meidaimae - 20130910
meidaimae - 20130912
kaihinmakuhari - 20130914
messe - 20130914
roppongi - 20130912
Number of people
Hour

Outline

Conclusion

- Location data from smartphone applications were studied for monitoring the number of pedestrians.
- Intended to be an alternative to traffic censuses using tally counters.
- Has advantages in the cost and flexibility compared to manual censuses

Conclusion

- Assume that the number of location data in an area is proportional to the population volume
- To estimate numbers of only pedestrians
- Deals with temporal and spatial sparsity of data by introducing non-homogeneous Poisson processes and Gaussian kernels.

Q&A

Any Questions?

End

Thanks For Attention!