A STUDY ON AUTOMATIC KERNEL BANDWIDTH SELECTOR FOR QUESTIONAIRE-BASED STATISTICS –USING JICA PERSON TRIP DATA IN VARIOUS DEVELOPING CITIES-

Atsuto WATANABE[*a], *Toshikazu NAKAMURA*[b], *Yoshihide SEKIMOTO*[c], *Tomotaka USUI*[d], and *Ryosuke SHIBASAKI*[e]

[a] *Master student, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan; Tel: +81-3-5452-6417; Email: atsuto@csis.u-tokyo.ac.jp*

[b] *Doctor student, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; Tel: +81-4-7136-4307; Email: ki_ki_gu@csis.u-tokyo.ac.jp*

[c] *Associate professor, Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; Tel: +81-4-7136-4307 Email: sekimoto@csis.u-tokyo.ac.jp*

[d] *Assistant professor, Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; Tel: +81-4-7136-4307 Email: usui@csis.u-tokyo.ac.jp*

[e] *Professor, Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; Tel: +81-4-7136-4307 Email: shiba@csis.u-tokyo.ac.jp*

KEYWORDS: Kernel bandwidth, optimization, statistics, questionnaire, person trip data

ABSTRACT: When conducting survey by questionnaire, time in data tends to be biased. Person trip survey has the same characteristic and departure and arrival time of trips tend to be biased to round number. When reconstructing the people flow using person trip data, biases to round number cause the accuracy of reconstruction to be low. In order to reduce the biases and improve the accuracy of the reconstruction, the method has been developed using kernel density estimation. But, in existing method, goodness-of-fit between estimated data and raw data need to be judged by each researcher on their own.
In this study, we experimented and compared methods to optimize parameters for kernel density estimation to reduce biases without any parameters that researchers have to select by themselves in questionnaire-based statistics. We applied this method to experiment on JICA person trip data.

1. INTRODUCTION

Construction of people flow has come important in these days. Center for Information Science (CSIS) at the University of Tokyo launched “People Flow Project” that works on data process technology, data quality using spatio-temporal data acquired from questionnaire-based statistics such as person trip survey. For example, Y. Sekimoto, et al. has done spatio-temporal interpolation by using this kind of data. Among various kinds of statistics, questionnaire-based statistics is valuable statistics because it includes detailed information which sometimes not available from other statistics. But, there are problems with questionnaire-based statistics that there are almost always biases in data. Especially, time in data is biased to the round number such as ○:00 and ○:30. Since The density distribution of time is usually not sequential but discrete, it has to be noted that the density that be estimated is the discrete data. B.W. Silverman (1986) and many researchers have done studies that related to density estimation using kernel density estimation. In these studies, it is noted that bandwidth selection is very critical to kernel density estimation itself. And, in most cases, there are free parameters that researchers have to select by themselves.
In this study, we comprehended the concept overall and method of bandwidth selection and compare them. And, we applied kernel density estimation and bandwidth selector to experiments on Japan International Corporation Agency person trip data (JICA-PT). Since JICA-PT is questionnaire-based statistics, time in the data has strong biases to round number as mentioned in introduction. The characteristics of data affect the result of the kernel density estimation. So, we examined the result in the point of view of the characteristics of data and found the relationships between characteristics of samples and bandwidth selectors.

2. METHODS

2.1 Smoothing departure time using kernel density estimation

Kernel density estimation is originally the method of estimating the true probability density, but also known as good method of smoothing spatio-temporal data. Kernel density estimation estimates the function

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{2.1}$$

where x_i is samples of measured values, n is the number of measured values, h is the bandwidth and k(・) is the kernel function. We used a Gaussian kernel function throughout this study.

2.2 Selection of the bandwidth

Many researchers have done many studies on the kernel density estimation. In these studies there are controversies on the choice of bandwidth because how much smoothed the data should be is different by purpose of uses. B.W. Silverman(1986) explains several bandwidth selectors from i , ii , and iii . Bandwidth selector iv is developed by H. Shimazaki and et al(2010).

i. Reference to standard deviation selector

In the point of view of minimizing the approximate mean integrated square error, bandwidth h_{opt} is optimized where

$$h_{opt} = k_2\{\int K(t)^2 dt\}^{1/5}\left\{\int f''(x)^2\, dx\right\}^{-1/5} n^{-1/5} \quad . \tag{2.2}$$

If true density distribution were assumed as standard family of distribution, and when a Gaussian kernel is used, optimized bandwidth h_{opt} would be

$$h_{opt} = 1.06\sigma n^{-1/5} \tag{2.3}$$

where σ is standard deviation.

ii. Data-based selector

Since (2.3) is based on the assumption that standard distribution is the true density, this can be written as using robust measure of spread by

$$h_{opt} = 0.79Rn^{-1/5} \tag{2.4}$$

where *R* is interquartile range. The best of both (2.3) and (2.4) can be obtained using adaptive estimate spread

$$A = min(\sigma, R/1.34) \tag{2.5}$$

instead of σ in (2.3) and reducing the factor *1.06* in (2.3) to work well if other distributions are assumed. When a Gaussian kernel is used the bandwidth h_{opt} is optimized where

$$h_{opt} = 0.9An^{-1/5} \quad . \tag{2.6}$$

iii. Least-squares cross-validation selector

This method was suggested by Rudemo (1982) and Bowman (1984) and automatically select bandwidth based on the simple principle of minimizing the integrated square error between estimated density and true density.

$$\int(\hat{f} - f)^2 = \int \hat{f}^2 - 2\int \hat{f} f + \int f^2 \tag{2.7}$$

The true density *f* in the second term of (2.7) is estimated from data themselves using cross-validation in this method. This method finds optimized bandwidth which minimizes the function

$$M_1(h) = n^{-2}h^{-1}\sum_i\sum_j K^*\left\{h^{-1}\left(X_i - X_j\right)\right\} + 2n^{-1}h^{-1}K(0) \tag{2.8}$$

where

$$K^*(t) = K^{(2)}(t) - 2K(t) = \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{1}{4}t^2\right) - 2\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\right) \quad . \tag{2.10}$$

iv. Least-squares-Poisson selector

This method is developed by H. Shimazaki and et al. (2010) for estimating spike rate and based on the principle minimizing the mean integrated square error. Here, this method uses Poisson assumption for estimating density function *f* in the second term of (2.7) in this method and finds the optimized bandwidth which minimizes

$$C_n(h) = \frac{1}{n^2}\sum_{i,j}\psi_h\left(t_i,t_j\right) - \frac{2}{n^2}\sum_{i\neq j}k_h\left(t_i - t_j\right) \tag{2.11}$$

where

$$\psi_h\left(t_i,t_j\right) = \int_a^b k_h\left(t - t_i\right)k_h\left(t - t_j\right)dt \quad . \tag{2.12}$$

Therefore, bandwidth selectors above try to suppose that true density is either unimodal (i and ii) or multimodal (iii , iv). And, bandwidth selectors supposing that the true density is multimodal estimate true density either from data themselves (iii) or some discretized density estimation (iv).
And, there are other methods for selecting bandwidth in existing studies. For example, subjective choice plots out curves and chooses estimate that is the most in accordance with the density. Test graph method minimizes maximum of the error between estimated and true density. These methods are introduced by B.W. Silverman(1986). But, there are no approaches to calculate and optimize the bandwidth yet.

3. EXPERIMENT

3.1 Data used in experiment

We used JICA-PT for the experiment applying kernel density estimation explained above. JICA-PT is the data acquired from survey on transportation in the city mostly in developing countries.

3.2 Processing flow of the experiment

First, we extracted the departure time as data sets of time from PT data that has a lot of other attributes. And, we applied the kernel density function (2.1) to the data sets acquired. Just the density is calculated from kernel density estimation, and then we multiply this density with number of samples so that we can get the number of people at each time. In this study, for simplicity, we suppose that the travel time is not changed after the departure time is smoothed. Therefore, we did not smooth arrival time but calculated arrival time from departure time smoothed and travel time.

3.3 Result

We conducted experiment on the data of 4 cities from JICA-PT. Gaussian kernel function and 4 bandwidth selectors were applied. i .Reference to standard deviation selector, ii . Data-based selector, iii . Least-squares cross-validation selector, iv . Least-squares-Poisson selector. (shown as “RtoSD”, “DataBased”, “LSCV”, and “LSP” in Fig.1). We define these selectors selector i - iv from now on. Observed data is drawn in graph as “raw”.

As shown in Fig.1, when selector i , ii , or iv is used, the shape of graph seems not too smoothed but well smoothed. On the other hand, when using selector iii the shape of the graph is jagged for all the cities. It is assumed that selector iii did not smooth the data. The principle of the selector iii supposes that distribution of observed data is close to the estimated distribution. Then, selector iii estimate the distribution close to observed data. In this study, we defined the resolution to be 1 minute, but the resolution of time of the distribution of people’s movement is not necessarily 1 minute. This caused the shape of graph to be jagged. Since selectors i and ii suppose that true density is any of rounded distribution that is either unimodal or multimodal. Thus it is assumed that the density estimated using these selectors is rounded and smoothed well.

Table 1 shows that bandwidth operation time when each bandwidth selectors and number of samples for each cities. Looking at the operation times, when selector iii and iv are used it takes time much longer than selector i and ii . This results from the difference of the calculation process between selector i - ii and selector iii and iv .

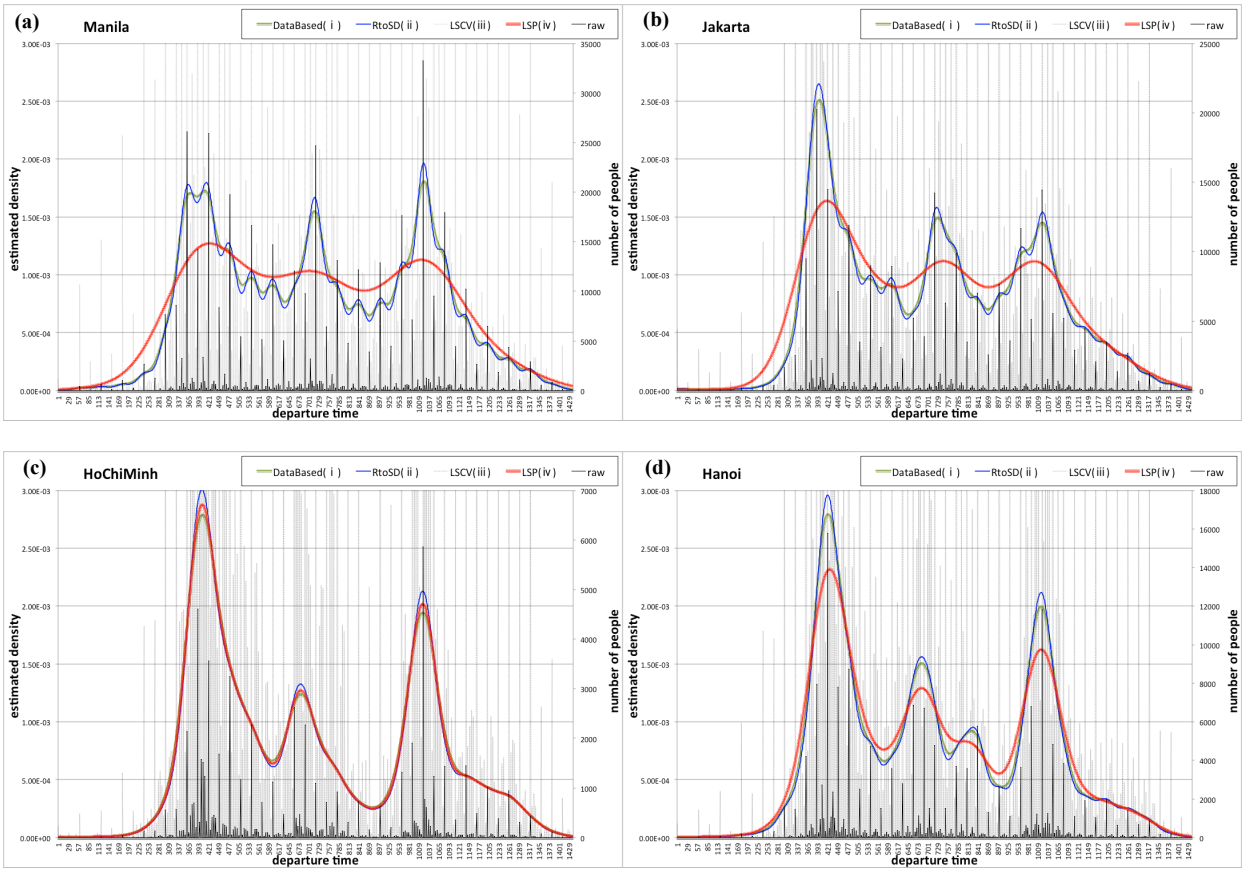

Fig. 1: Estimated density with each bandwidth selectors and number of people of the raw data
(a): Manila, (b): Jakarta, (c): HoChiMinh, (d): hanoi

Table 1: Bandwidth, operation time, and mean square error when calculated using each methods.

City	Number of samples		RtoSD(i)	DataBased(ii)	LSCV(iii)	LSP(iv)
Manila	234293	bandwidth	18.843	22.172	4.0300E-10	81.998
		operation time(ms)	2080	2027	74239	765262
Jakarta	133926	bandwidth	20.285	23.868	4.0279E-10	64.682
		operation time(ms)	1639	1631	62825	649407
HoChiMinh	36757	bandwidth	28.106	33.072	4.0279E-10	30.867
		operation time(ms)	926	879	18405	193713
Hanoi	97244	badwidth	21.441	25.228	4.0279E-10	40.318
		operation time(ms)	1309	1298	40130	419484

Only standard deviation and interquartile range are needed for selector i and ii. On the other hand, there is optimization process finding bandwidth that minimizes or maximizes the cost function when selector iii and iv are used. As shown in Table 1 bandwidths calculated using selector iii are very close to zero and bandwidths with selectors i, ii and iv seems adequate. But, if you see the graphs in Fig.1 and see differences by cities, the estimated distributions with selector i and ii are different even though the bandwidth is not so different. For example, distribution with selector i and ii in Manila or Jakarta has more peaks than 10 peaks or so, but in the HoChiMinh and Hanoi it has only 3 or 4 peaks. When selector iv is used the number of peaks are not different by cities even though bandwidths are not so close. Therefore, it is assumed that when selector i and ii are used the number of peaks are varied due to the characteristics of the sample such as number of samples. But, selector iv is stable regardless of the characteristics of the sample.

4. CONCLUSION AND DISCUSSION

We applied the kernel density estimation with 4 different bandwidth selectors for smoothing departure time in this paper. When using kernel density estimation the true certain density distribution is supposed for each bandwidth selectors and the density at each time is estimated and this leads to smoothing of departure time. The density distribution is different by bandwidth selectors. Here, it should be noted that what the density distribution is estimated in the first place for each bandwidth selectors. In this study, sample data is the family of times that people take trips. The purpose of this study is to smooth the departure time, which means estimating continuous distribution from distribution of the people's movement that is not continuous distribution. Thus, the result is good when the continuous distribution is estimated, selector ⅰ, ⅱ, and ⅳ. Among these selectors, regarding of robustness and versatility of the method, selector ⅳ seems the good bandwidth selector that is not affected by the characteristics of samples in the point of view of peaks of departure time. In addition to that, selector ⅳ provides the adequate estimation for smoothing even though resolution of time in the estimation is different in the real life.

For further study, the variable kernel density estimation might be useful for thin matter because the true density can be estimated differently by time. In kernel density estimation, originally, true density of the sample data are estimated and superimposed and the density distribution overall is estimated. Thus, the true density of the people's movement is estimated at each time before superimpose, and this is done by the same principle at all time. But, in real life of the people movement, the true density is different by time.

ACKNOWLEDGMENT

This research was partially supported from a Grant-in-Aid for Young Scientists (A) by the Ministry of Education, Culture, Sports, Science and Technology(MEXT) and from the Environment Research and Technology Development Fund (RF-1012) of the Ministry of the Environment, Japan.
PT data were lent by JICA in Japan.

REFERENCE

Bowman, A. and Azzelini, A. (1997) "Apllied Smoothing Techniques for Data Analysis" Oxford University Press.
Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Bio-metrika, 71(2), 353.
B.W. Silverman(1986), "Density Estimation for Statistics and Data Analysis" Chapman & Hall/CRC
H. Shimazaki and S. Shinomoto(2010), "Kernel bandwidth optimization in spike rate estimation", J Comput Neurosci(2010) 29:171-182
Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. Scandinavian journal of Statistics, 9(2), 65-78.
Yoshihide Sekimoto, Ryosuke Shibasaki, Hiroshi Kanasugi and Tomotaka Usui, Yasunobu Shimazaki, PFLOW: Reconstruction of people flow by recycling large-scale fragmentary social survey data, *IEEE Pervasive Computing, 2011*(Accepted)
People Flow Project: http://pflow.csis.u-tokyo.ac.jp