21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Statistical Origin-Destination Generation with Multiple Sources

Tetsuro Morimura, Sei Kato
IBM Research – Tokyo
tetsuro@jp.ibm.com, seikato@jp.ibm.com

Abstract

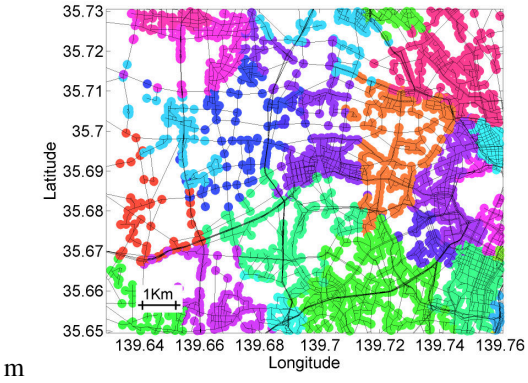
Any trajectory is always generated with its origin and destination. Origin-destination (OD) generation for trips plays an important role in many applications such as trajectory mining, traffic simulation, or marketing. In previous work on traffic pattern recognition, microscopic ODs for limited areas are estimated with probe-car data, while macroscopic ODs for broad areas are usually generated by using road-traffic-census data. In this paper, we propose a microscopic OD determination method for broad areas with the same data and landmark information, which is based on an L1-regularized Poisson regression. We demonstrate performance improvements over baseline methods in numerical experiments with a massive data set from Tokyo.

1. Introduction

In recent years, trajectory mining has become one of the central topics in data-mining and pattern-recognition [4, 2] since it is deeply related to many applications such as marketing [6] or traffic simulation and control [9]. Peculiarly, because any trajectory is always generated with its origin and destination, origin-destination (OD) estimation and generation play an important role in trajectory studies.

In conventional approaches on traffic OD generation [3], an OD table is used, which is typically made by person trip counts of the traffic-census data [1]. However, such traffic-census data is coarse and thus the level of detail in the OD table is at the zone level, as shown in Fig. 1. In the other related work [10], microscopic OD estimation has been proposed with probe-car data, though this approach is often limited to a narrow area being available the probe-car data, as shown in Fig. 2.

In this paper, we propose a microscopic OD-estimation method for broad areas by using three data sources, probe-car, road-traffic-census, and landmark data. Here the landmark data includes various types of landmarks, such as hotels or railway stations, and their locations, as shown in Fig. 3. The proposed OD estimation method is a L1-regularized Poisson regression with an adaptive baseline, in which the probe-car data is propagating with the landmark information. In our OD generation or inference, the traffic-census data is used as prior information.



m

Figure 1. Zones of the traffic-census data for a city center map in Tokyo, Japan: Different colors represents different zones.

2. Traffic OD Estimation Problem

The traffic OD estimation problem is to inference a joint probability function of places for the origin and destination of a trip, Here we assume that the smallest geographical unit to generate OD pairs is a cross-point, *i.e.*, a node of a road-network graph[1].

2.1. Input Data

The cross-point is denoted as

$$c \in \mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\},$$

[1]Our approach can also be directly applied when the unit to generate the ODs is a road, *i.e.*, a link of a road-network graph.

978-4-9906441-0-9 ©2012 ICPR

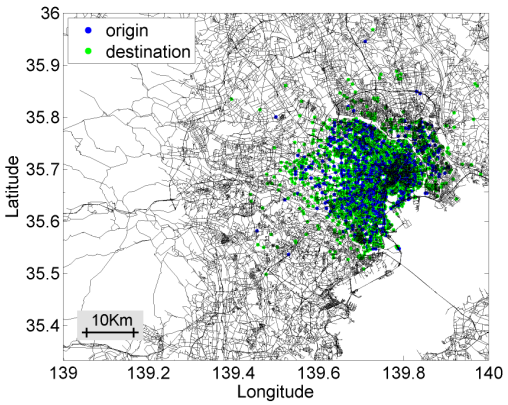


Figure 2. Origin and destination locations extracted from probe-car data in Tokyo.

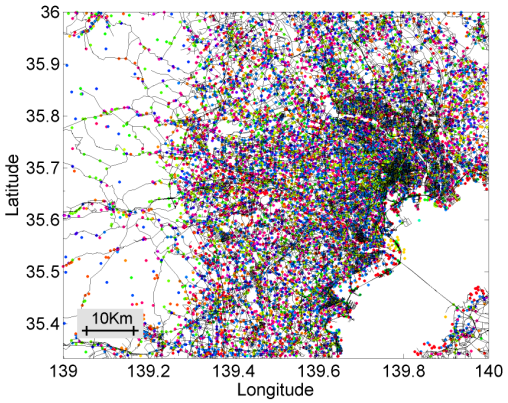


Figure 3. Locations of landmarks in Tokyo: Different colors represents different types of landmarks.

where $|\mathcal{C}|$ is the total number of cross-points in the graph. The each of the cross-points, c_n, is annotated with the OD counts of the probe-car data, o_n and $d_m^n \in \mathbb{N}$, where o_n is the number of probe-car trips departing from c_n and d_m^n is the number of probe-car trips from c_m to c_n. Each cross-point c_n also has landmark information $\boldsymbol{l}_n \in \mathbb{R}^{|\mathcal{L}|}$, where $|\mathcal{L}|$ is the number of landmark types ($\simeq 150$ in our Tokyo data set). The element i of $\boldsymbol{l}_n$ denotes the number of nearby type-i landmarks.

We also assume that a zone-level OD table is available, which was constructed from the person-trip counts of the public traffic-census data [1]. Each entry (i, j) of the table has the number of person-trips from the zone i to the zone j. Each zone is defined by the traffic-census data and includes about a hundred cross-points in our data set, as shown in Fig. 1. The ID of the zone for the cross-point c_n is denoted as z_n. By using this OD table, we can compute the joint probability of the zone-level OD or generate the zone-level ODs.

2.2. Output

The output of the OD estimation problem is a probability function of the cross-points-level OD or a set of the cross-points-level ODs generated by the probability function.

3. Poisson Regression for OD Estimation

3.1. L1-regularized Poisson Regression

The Poisson distribution is one of the standard distributions for the natural number $n \in \mathbb{N}$,

$$\Pr(n \,|\, \mu) \triangleq \frac{\mu^n \exp(-\mu)}{n!}, \tag{1}$$

where μ is the parameter of the Poisson distribution and is also its mean and standard deviation [7]. In a Poisson regression, the parameter μ is represented with a log-linear model as $\mu_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\boldsymbol{\theta}^\top \boldsymbol{x})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$ are the parameter and feature vectors, respectively. The dimension of both is d. The superscript $\top$ denotes the transposition of a vector. Accordingly, Eq. (1) can be rewritten as

$$p(n \,|\, \boldsymbol{x}, \boldsymbol{\theta}) \triangleq \Pr(n \,|\, \mu_{\boldsymbol{\theta}}(\boldsymbol{x})).$$

The objective of L1-regularized Poisson regression is to find the best parameter that minimizes the objective function [5]

$$J(\boldsymbol{\theta}) = \sum_k -\log p(n_k \,|\, \boldsymbol{x}_k, \boldsymbol{\theta}) + \lambda |\boldsymbol{\theta}|_1, \tag{2}$$

where λ (≥ 0) is a regularization parameter and $|\boldsymbol{\theta}|$ denotes the L1-norm of $\boldsymbol{\theta}$. The problem is known to be iteratively solvable with a quadratic Taylor series expansion around the current estimate of $\boldsymbol{\theta}$ and using lasso regression [8, 5].

3.2. OD estimation with adaptive baselines

For the OD estimation, we use the landmark information for the feature vectors of cross-points, *i.e.*

$$\Pr(n_{\mathrm{o}} \,|\, c_{\mathrm{origin}} = c_i) := p(n_{\mathrm{o}} | \boldsymbol{l}_i, \boldsymbol{\theta}_{\mathrm{o}}),$$
$$\Pr(n_{\mathrm{d}} \,|\, c_{\mathrm{origin}} = c_i, c_{\mathrm{destination}} = c_j) := p(n_{\mathrm{d}} | \boldsymbol{l}_i^j, \boldsymbol{\theta}_{\mathrm{d}}),$$

where n_{o} or n_{d} is the number of origin or destination events, respectively, and $:=$ denotes the right-to-left substitution. The feature vector for the destination is $\boldsymbol{l}_i^j \equiv [\sqrt{\mathrm{vec}(\boldsymbol{l}_i \otimes \boldsymbol{l}_j)}^\top, \boldsymbol{l}_j^\top]^\top$, where $\mathrm{vec}(\mathrm{M})$ or $\otimes$ denotes the vectorization of a matrix $\boldsymbol{M}$ or the outer product operator, respectively.

Here we also introduce adaptive baselines to the Poisson regression, in order to reduce the effect that the sampling distribution of the ODs from the probe-car data can be heavily biased, as shown in Fig. 2. Assuming that, given a zone z, the sampling of the probe-car data is uniform, the adaptive baselines of cross-points c_i, c_j for the origin and destination are given as the logarithms of the average origin- or destination-events for zones z_i, z_j,

$$b(c_i) = \log(\mathbb{E}[n_{\mathrm{o}} \mid z_{\mathrm{origin}} = z_i]),$$
$$b(c_i, c_j) = \log(\mathbb{E}[n_{\mathrm{d}} \mid z_{\mathrm{origin}} = z_i, z_{\mathrm{destination}} = z_j]),$$

respectively, where $\mathbb{E}$ denotes expectation operator. These have the corresponding parameters θ^b_{o} and θ^b_{d}. The total feature vector and parameter for the origin[2] will be $\tilde{\boldsymbol{l}}_i = [b(c_i), \boldsymbol{l}_i^\top]^\top$ and $\tilde{\boldsymbol{\theta}}_{\mathrm{o}} = [\theta^b_{\mathrm{o}}, \boldsymbol{\theta}_{\mathrm{o}}^\top]^\top$, respectively. This baseline can be validated by the observation that, if we set $\theta^b_{\mathrm{o}} = 1$ and $\boldsymbol{\theta}_{\mathrm{o}} = \mathbf{0}$, the mean (and standard deviation) of estimated Poisson distribution, $\mu_{\tilde{\boldsymbol{\theta}}_{\mathrm{o}}}(c_i) = \exp(\tilde{\boldsymbol{\theta}}_{\mathrm{o}}{}^\top \tilde{\boldsymbol{l}}_i)$, is equal to $\mathbb{E}[n_{\mathrm{o}} | z_{\mathrm{origin}} = z_i]$. That is, the unbiased center for θ_b or $\boldsymbol{\theta}_{\mathrm{o}}$ is 1 or $\mathbf{0}$. Our objective function can be derived with Eq. 2 as

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{|\mathcal{C}|} -\log p(o_i \mid \tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{\theta}}_{\mathrm{o}}) + \lambda(|\theta_b - 1| + |\boldsymbol{\theta}_{\mathrm{o}}|_1).$$

The optimization problem remains convex and thus can be solved with the same approach as Eq. (2).

3.3. OD generation

Here we use the traffic-census OD table at the zone level as prior information for a stochastic process. First, a pair of zones for an origin and a destination will be chosen from the OD table. Then a cross-point c_i in the chosen origin zone is selected as an origin location with the learned Poisson distribution $p(n_{\mathrm{o}}|\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{\theta}}_{\mathrm{o}})$, in which a cross-point c_i is chosen with probability $\propto \exp(\tilde{\boldsymbol{\theta}}_{\mathrm{o}}^\top \tilde{\boldsymbol{l}}_i)$. Finally a cross-point c_j in the chosen destination zone is selected as a destination location with $p(n_{\mathrm{d}}|\tilde{\boldsymbol{l}}^j_i, \tilde{\boldsymbol{\theta}}_{\mathrm{d}})$. It is noted that the probe-car data is, in our proposed method, required only in the model-training phase, not in the OD generation phase.

4. Experiments

Here we compare the proposed method to the conventional method with a massive data set from Tokyo, which has about a million cross-points and a thousand zones. Figure 4 shows the comparisons between the proposed and traffic-census-based methods. It can be confirmed that, as is consistent with our natural expectations, only the proposed method could infer high origin probabilities around the central stations in Tokyo. To confirm results, we also evaluated the methods in terms of the Area Under Curve (AUC) with 10-fold cross-validation [5] in Fig. 5, where our method (Poisson regression with adaptive baseline) always yields better performance than the baseline methods. For the discovered landmark patterns, the lists of landmark types strongly contributed to the OD estimation as shown in Tables 1 and 2.

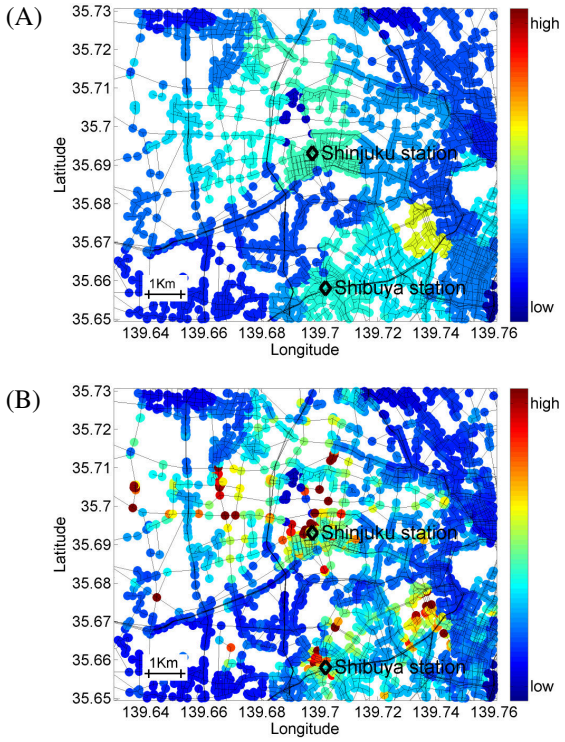


Figure 4. Estimated probability of becoming the origin in a city center map of Tokyo: (A) Conventional method with zone-level OD table from the traffic-census data, (B) Proposed method with multiple sources.

5. Summary

We proposed a microscopic OD estimation method for broad areas by using multiple data sources such as probe-car, road-traffic-census, and landmark data. The proposed method is an L1-regularized Poisson regression with an adaptive baseline, in which the probe-car data from narrow areas is propagated to the entire area along with landmark information. We also showed per-

[2]Because those for the destination are similar for the origin and the size of this paper is limited, we omit the detailed destination descriptions.

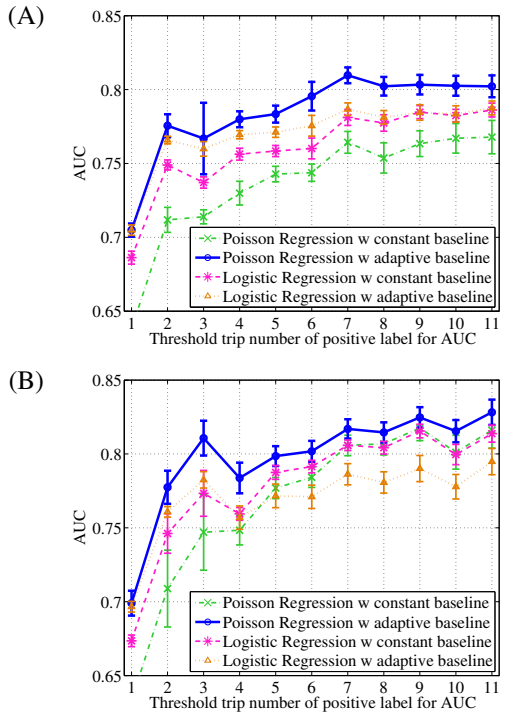


Figure 5. Performance comparisons based on AUC: (A) Origin, (B) Destination given origin.

formance improvements over the baseline methods in experiments with a massive data set for Tokyo.

Acknowledgment

This work was supported by the PREDICT of the Ministry of Internal Affairs and Communications, Japan.

Table 1. List of landmark types for contributing origin events, ($|\theta_i| \geq 0.2$).

Type of Landmark@	parameter θ_i
Hotel	0.414
Interchange, ramp	-0.376
Bank	0.337
Station of Japan Railway	0.279
Convenience store	0.277
Fast-food shop	0.239
Highway junction	-0.237
Embassy	0.229
Shopping center	0.226
Department store	0.208
Subway station	0.204

Table 2. List of landmark types for contributing destination events, ($|\theta_i| \geq 0.2$).

Type of Landmark(s)		parameter θ_i
Origin	Destination	
–	Railway station	0.418
–	Hotel	0.393
–	Interchange, ramp	-0.289
–	Convenience store	0.286
–	Bank	0.279
Hotel	Railway station	0.271
–	Embassy	0.263
–	Department store	0.242
–	Building	0.23
–	Highway junction	-0.22
Railway station	Hotel	0.217
Railway station	Railway station	-0.212
–	Subway station	0.207
–	Fast-food shop	0.201

References

[1] Results from the 4th nationwide person trip survey, 2007. http://www.mlit.go.jp/crd/tosiko/zpt/pdf/zenkokupt_gaiyouban_english.pdf.

[2] A. J. L. amd Y.A. Chen and W. Ip. Mining frequent trajectory patterns in spatial.temporal databases. *Information Sciences*, 179(13):2218–2231, 2009.

[3] E. Cascetta and S. Nguyen. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6):437–455, 1988.

[4] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinellil. Trajectory pattern mining. In *International Conference on Knowledge Discovery and Data Mining*, pages 330–339, 2007.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[6] S. K. Hui, P. S. Fader, and E. T. Bradlow. Path data in marketing: An integrative framework and prospectus for model building. *Marketing Science*, 28(2):320–335, 2009.

[7] A. J.Dobson and A. Barnett. *An Introduction to Generalized Linear Models, Third Edition*. Chapman and Hall, 2008.

[8] R. C. Kelly, M. A. Smith, and R. E. Kass. Accounting for network effects in neuronal responsesusing l1 regularized point process models. In *Advances in Neural Information Processing Systems*, volume 22, pages 1099–1107, 2010.

[9] M. N. Nguyen, D. Shi, C. Quek, and G. S. Ng. Traffic prediction using ying-yang fuzzy cerebellar model articulation controller. In *International Conference on Pattern Recognition*, 2006.

[10] K. Sohn and D. Kim. Dynamic originadestination flow estimation using cellular communication system. *IEEE Transactions on Vehicular Technology*, 57(5):2703–2713, 2008.