Estimating People Flow from Spatio-temporal Population Data via Collective Graphical Mixture Models

TOMOHARU IWATA, HITOSHI SHIMIZU, FUTOSHI NAYA, and NAONORI UEDA, NTT Communication Science Laboratories

Thanks to the prevalence of mobile phones and GPS devices, spatio-temporal population data can be obtained easily. In this paper, we propose a mixture of collective graphical models for estimating people flow from spatio-temporal population data. The spatio-temporal population data we use as input are the number of people in each grid cell area over time, which are aggregated information about many individuals; to preserve privacy, they do not contain trajectories of each individual. Therefore, it is impossible to directly estimate people flow. To overcome this problem, the proposed model assumes that transition populations are hidden variables, and estimates the hidden transition populations and transition probabilities simultaneously. The proposed model can handle changes of people flow over time by segmenting time-of-day points into multiple clusters, where different clusters have different flow patterns. We develop an efficient variational Bayesian inference procedure for the collective graphical mixture model. In our experiments, the effectiveness of the proposed method is demonstrated by using four real-world spatio-temporal population data sets in Tokyo, Osaka, Nagoya and Beijing.

CCS Concepts: • **Information systems → Data mining**;

Additional Key Words and Phrases: collective graphical models, mixture models, variational Bayes, spatio-temporal data, population data

ACM Reference format:
Tomoharu Iwata, Hitoshi Shimizu, Futoshi Naya, and Naonori Ueda. 2017. Estimating People Flow from Spatio-temporal Population Data via Collective Graphical Mixture Models. *ACM Trans. Spatial Algorithms Syst.* 9, 4, Article 39 (March 2017), 18 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Analyzing spatio-temporal population data is important in various fields including disaster management [20], marketing [5], public health [3], and urban planning [24]. For example, population information is useful for planning the locations of new stores. Census data have long been used as spatial population data. However, since conducting a census is time-consuming and costly, census data are not updated frequently; e.g. the US census is conducted every ten years. In recent years, thanks to the prevalence of mobile phones and global positioning system (GPS) devices, spatio-temporal population data can be obtained easily. For example, mobile spatial statistics [22] contain the hourly population in 500 meter grid squares in Japan's urban areas, which are calculated

A part of this work has been achieved by "Research and Development on Fundamental and Utilization Technologies for Social Big Data," the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN.

DOI: 0000001.0000001

from mobile network operational data. Similar data can be obtained from the GPS data of navigation apps [23], Wi-Fi [13], geotagged tweets [1] or photos [2] and check-in data of location sharing services such as foursquare [3].

In this paper, we consider the task of estimating people flow from spatio-temporal population data. The estimated people flow can be used for a wide variety of applications, which include simulating people movement in the case of a disaster, detecting anomalous people movement, predicting the future spatial population given the current spatial population, and designing transportation systems. The spatio-temporal population data we use as input are the populations in each grid cell over time as shown in Figure 4. The output is the people flow between grid cells over time as shown in Figure 7. The spatio-temporal population data consist of aggregated information about many individuals. They are aggregated to preserve privacy or because of the difficulty of tracking individuals over time. For instance, the mobile spatial statistics [22] are preprocessed for privacy protection and no one can follow a particular user, which allows the mobile phone operating company to publish population data calculated based on information about 60 million mobile phone users. If the trajectories for each individual are given, people flow can be estimated straightforwardly by counting the number of people who moved between grid cells, i.e. transition population. However, with aggregated data, it is impossible to directly determine the size of the transition population.

To overcome this problem of modeling individual behavior given aggregated data, we propose a *mixture of collective graphical models*. The proposed model assumes that individuals move according to transition probabilities that depend on their locations and time points. Since the transition populations are not given, we treat them as hidden variables. The hidden transition populations relate to observed populations in each cell; the population at a cell is equal to the sum of transition populations from the cell, and the population at a cell in the next time point is equal to the sum of transition populations to the cell. By using these relations that represent flow conservation as constraints, the hidden transition populations and transition probabilities are inferred simultaneously. The proposed model can handle changes of people flow over time by segmenting time-of-day points into multiple clusters, where different clusters have different transition probabilities. For our task of modeling people flow, incorporating time information is crucial. For example, people move from suburbs to the city center to work in the morning, they return to the suburbs from the city center in the evening after work, and they do not travel so often in the middle of the night. The proposed model is an extension of the collective graphical models [17, 18] for handling change over time. Since the proposed model does not use trajectory information, when we have enough amount of population data that approximate the true population distribution, we can estimate people flow.

We develop an efficient variational Bayesian inference procedure for the proposed model. In existing literature on collective graphical models, the EM algorithm is used to obtain point estimates of the parameters [6, 10, 17, 21]. Instead of obtaining point estimates, by estimating distributions of parameters based on the variational Bayesian framework, we can alleviate overfitting especially when data are sparse and models are flexible as in the case of mixture models [1].

The major contributions of this paper include the following:

- It is the first attempt to estimate people flow from spatio-temporal population data without tracking.
- We propose a mixture of collective graphical models for handling behavior change over time.

[1] http://twitter.com
[2] http://www.flickr.com
[3] http://foursquare.com

Table 1. Notation

Symbol	Description
N_{ti}	#people at grid cell i at time t, $N_{ti} \geq 0$
M_{tij}	#people who moved from grid cell i to grid cell j at time t, $M_{tij} \geq 0$
T	#time points, $t \in \{1, \cdots, T\}$
I	#grid cells, $i \in \{1, \cdots, I\}$
$\mathbf{J}_i$	neighbors of grid cell i
S	#time-of-day indices
$s(t)$	time-of-day index of time t, $s(t) \in \{1, \cdots, S\}$
g_s	sth time-of-day $g_s \in [0, 24]$
$r(i, j)$	relative position index of grid cell j from grid cell i
K	#clusters
$z_{s(t)}$	cluster index of time t, $k \in \{1, \cdots, K\}$
θ_{kij}	probability that a person in grid cell i moves to grid cell j during a time-of-day point that belongs to cluster k, $\theta_{kij} \geq 0$, $\sum_{j \in \mathbf{J}_i} \theta_{kij} = 1$
ϕ_k	probability that a time-of-day index belongs to cluster k, $\phi_k \geq 0$, $\sum_{k=1}^{K} \phi_k = 1$
τ_k	time-of-day mean of cluster k, $\tau_k \in [0, 24]$
η_k	time-of-day precision of cluster k, $\eta_k > 0$
$\boldsymbol{\theta}_{ki}$	set of transition probabilities at grid cell i during a time-of-day point that belongs to cluster k, $\boldsymbol{\theta}_{ki} = (\theta_{kij})_{j \in \mathbf{J}_i}$
$\mathbf{N}$	set of population data over time and grid cells, $\mathbf{N} = ((N_{ti})_{i=1}^{I})_{t=1}^{T}$
$\mathbf{M}$	set of transition population data over time and grid cells, $\mathbf{M} = (((M_{tij})_{i=1}^{I})_{j \in \mathbf{J}_i})_{t=1}^{T-1}$

- We develop a variational Bayesian inference procedure for collective graphical models.

The paper is organized as follows: In Section 2, we propose collective graphical mixture models for estimating people flow from spatio-temporal population data, and present its variational Bayesian inference procedure in Section 3. In Section 4, we demonstrate the effectiveness of the proposed method by using real spatio-temporal population data obtained in Tokyo, Osaka, Nagoya and Beijing. In Section 5, we outline related work. Finally, we present concluding remarks and future work in Section 6.

2 PROPOSED MODEL

Suppose that we have population data over time for each of I grid cells, $\mathbf{N} = ((N_{ti})_{i=1}^{I})_{t=1}^{T}$, where N_{ti} is the number of people in grid cell i at time t. We also have neighbor information for each grid cell. Let $\mathbf{J}_i \subseteq \{1, \cdots, I\}$ be the set of neighbor cells of grid cell i, and $\mathbf{J}_i$ always contains grid cell i itself, $i \in \mathbf{J}_i$. We assume that people do not move to its non-neighbor cells, or we can ignore those people because their size is small. Our task is, given aggregated population data $\mathbf{N} = ((N_{ti})_{i=1}^{I})_{t=1}^{T}$, to estimate transition population $\mathbf{M} = (((M_{tij})_{i=1}^{I})_{j \in \mathbf{J}_i})_{t=1}^{T-1}$ in the time period of the given aggregated data, where M_{tij} is the number of people who move from grid cell i to grid cell j at time t. Our notation is summarized in Table 1. We use non-bold typefaces for scalar variables, and bold typefaces for vector and set variables. The time index t and time-of-day index $s(t)$ can be different since the given data might be taken at multiple days. Suppose that the tth time point is Oct 1st 10:00, and t'th time point is Oct 3rd 10:00. Since these time points have the same time-of-day, their time-of-day indices are the same $s(t) = s(t')$, but $t \neq t'$.

We model the probability of people moving to the neighbor grid cells using a mixture of collective graphical models. With the proposed model, the transition patterns are assumed to change

depending on the time-of-day, and the time-of-day points are divided into K clusters according to their transition patterns. Let θ_{kij} be the probability that a person in grid cell i moves to grid cell j during a time-of-day point that belongs to cluster k. The relationship between observed data $\mathbf{N}$ and transition probability θ_{kij} can be modeled by introducing latent variable M_{tij}, which represents the number of people who move from grid cell i to grid cell j at time t. In particular, when the time-of-day of t is assigned to cluster k, the probability of transition population $\mathbf{M}_{ti} = (M_{tij})_{j\in\mathbf{J}_i}$ given the current population N_{ti} and transition probability $\boldsymbol{\theta}_{ki} = (\theta_{kij})_{j\in\mathbf{J}_i}$ is given by the following multinomial distribution,

$$p(\mathbf{M}_{ti}|N_{ti},\boldsymbol{\theta}_{ki}) = \frac{N_{ti}!}{\prod_{j\in\mathbf{J}_i} M_{tij}!}\prod_{j\in\mathbf{J}_i}\theta_{kij}^{M_{tij}}. \tag{1}$$

The population in a cell N_{ti} and the transition population between cells $\mathbf{M}_{ti}$ have the following two relations,

$$N_{ti} = \sum_{j\in\mathbf{J}_i} M_{tij}, \tag{2}$$

which represents the fact that the population at grid cell i is the same as the sum of the transition populations from grid cell i, and

$$N_{t+1,i} = \sum_{j\in\mathbf{J}_i} M_{tji}, \tag{3}$$

which represents the population in grid cell i at the next time point is the same as the sum of the transition populations to grid cell i at the current time point. These relations are used as constraints in the inference as described in Section 3.

For the prior distribution of the transition probability, we use the non-symmetric Dirichlet distribution. With the proposed model, the hyperparameters of the Dirichlet distributions are shared among different grid cells when the relative positions are the same. Let $r(i,j)$ be the index of the relative position of grid cell j from grid cell i. Figure 1 shows an example of the relative position index when the neighbors are its surrounding eight cells with the addition of the cell itself. Then, the Dirichlet prior for the transition probaiblity is given as follows,

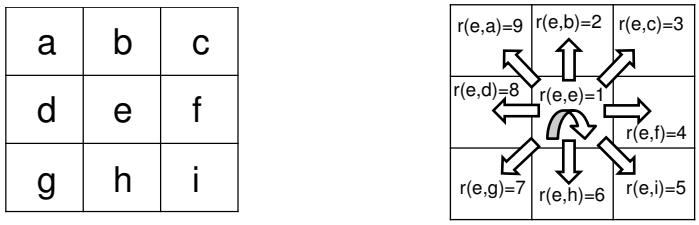
$$p(\boldsymbol{\theta}_{ki}|\boldsymbol{\alpha}_i) = \frac{\Gamma(\bar{\alpha}_i)}{\prod_{j\in\mathbf{J}_i}\Gamma(\alpha_{r(i,j)})}\prod_{j\in\mathbf{J}_i}\theta_{kij}^{\alpha_{r(i,j)}-1}, \tag{4}$$

where $\boldsymbol{\alpha}_i = (\alpha_{r(i,j)})_{j\in\mathbf{J}_i}$ is the Dirichlet hyperparameter, and $\bar{\alpha}_i = \sum_{j\in\mathbf{J}_i}\alpha_{r(i,j)}$. The parameter α_r represents the prospect of moving in the rth relative position, which is shared by all cells and all time points, e.g. $\frac{\alpha_1}{\bar{\alpha}_i}$ is the prior probability of staying at the current point when $r(i,i) = 1$. By estimating the Dirichlet hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots)$ from the given data, we can robustly calculate the population flow especially in cells with sparse data by sharing hyperparameters among different cells.

People flows at temporally close time points are considered to be similar. For example, transition probabilities at 14:00 are similar to those at 15:00. To model this property, we assume that the time-of-day points in cluster k are distributed as Gaussian with mean τ_k and precision η_k, where precision is the inverse of variance, as follows,

$$p(g_s|\tau_k,\eta_k) = \left(\frac{\eta_k}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{\eta_k}{2}|g_s-\tau_k|^2\right), \tag{5}$$

where g_s is the sth time-of-day. For example, when observations are given for every 30 minutes, $g_1 = 0, g_2 = 0.5, g_{48} = 23.5$, and the total number of time-of-day indices is $S = 48$. We use continuous time-of-day g_s for clustering time-of-day points based on the continuous time-of-day values, and



(a) name for each grid cell (b) relative position indices from the center grid cell

Fig. 1. (a) name for each grid cell and (b) their relative position indices from the center grid cell "e", $r(e,\cdot)$.

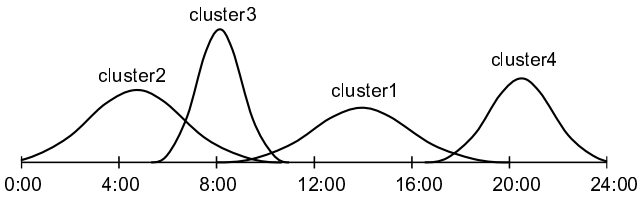


Fig. 2. Example of Gaussian distributions over time-of-day. The mean of cluster k is τ_k, and the precision is $d\eta_k$.

use time-of-day index s for representing the point to be clustered. Figure 2 shows an example of Gaussian distributions over time-of-day. By assuming the Gaussian distribution for each cluster, we can cluster time-of-day points depending on their interval of time-of-day as well as their transition probabilities.

The proposed model assumes the following generative process for the transition population over time $\mathbf{M} = ((\mathbf{M}_{ti})_{i=1}^{I})_{t=1}^{T-1}$,

(1) Draw cluster proportions, $\boldsymbol{\phi} \sim \text{Dirichlet}(\beta)$
(2) For cluster $k = 1$ to K
  (a) For grid cell $i = 1$ to I
    (i) Draw transition probability, $\boldsymbol{\theta}_{ki} \sim \text{Dirichlet}(\boldsymbol{\alpha}_i)$
  (b) Draw precision of time-of-day, $\eta_k \sim \text{Gamma}(a, b)$
  (c) Draw mean of time-of-day, $\tau_k \sim \text{Normal}(f, (d\eta_k)^{-1})$
(3) For time-of-day index $s = 1$ to S
  (a) Draw cluster, $z_s \sim \text{Discrete}(\boldsymbol{\phi})$
  (b) Draw time-of-day, $g_s \sim \text{Normal}(\tau_{z_s}, \eta_{z_s}^{-1})$
(4) For time points $t = 1$ to $T - 1$
  (a) For grid cell $i = 1$ to I
    (i) Draw transition population, $\mathbf{M}_{ti} \sim \text{Multinomial}(\boldsymbol{\theta}_{z_{s(t)}i}, N_{ti})$

Here, $\boldsymbol{\phi} = (\phi_k)_{k=1}^{K}$ is cluster proportions, and $s(t)$ represents the index of time-of-day at t. We assume that the transition probability is the same when time-of-day is the same even if their days are different. We use conjugate priors, i.e. Gaussian-Gamma priors for the Gaussian mean and precision, and Dirichlet priors for Discrete or multinomial parameters, which enables us to analytically derive equations for updating model parameters in the variational Bayesian inference. In particular, the prior of Gaussian mean τ_k and precision η_k is Gaussian-Gamma distribution

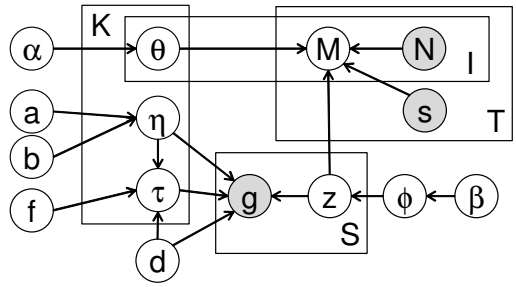


Fig. 3. Graphical model representation of the proposed mixture of collective graphical models.

Normal$(f,(d\eta_k)^{-1})$Gamma(a,b), and the prior of multinomial parameters $\boldsymbol{\theta}_{ki}$ and $\boldsymbol{\phi}$ is Dirichlet distribution Dirichlet$(\boldsymbol{\alpha}_i)$ and Dirichlet(β), respectively, where $f, d, a, b, \boldsymbol{\alpha}_i$ and β are hyperparameters. Given spatio-temporal population $\mathbf{N}$, time-of-day indices $\mathbf{s}$ and hyperparameters $\boldsymbol{\alpha}, \beta, a, b, d, f$, the joint probability of transition population $\mathbf{M}$, time-of-day data $\mathbf{g} = (g_s)_{s=1}^S$, model parameters $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_{ki})_{i=1}^I)_{k=1}^K$, $\boldsymbol{\phi}$, $\boldsymbol{\tau} = (\tau_k)_{k=1}^K$, $\boldsymbol{\eta} = (\eta_k)_{k=1}^K$, and cluster assignments $\mathbf{z} = (z_s)_{s=1}^S$ is given as follows,

$$\begin{aligned} &p(\mathbf{M},\mathbf{g},\boldsymbol{\Theta},\boldsymbol{\phi},\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta}|\mathbf{N},\mathbf{s},\boldsymbol{\alpha},\beta,a,b,d,f) \\ &= p(\mathbf{M}|\mathbf{N},\boldsymbol{\Theta},\mathbf{z},\mathbf{s})p(\boldsymbol{\Theta}|\boldsymbol{\alpha})p(\mathbf{z}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\beta)p(\mathbf{g}|\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta})p(\boldsymbol{\tau}|f,d,\boldsymbol{\eta})p(\boldsymbol{\eta}|a,b). \end{aligned} \tag{6}$$

Figure 3 shows a graphical model representation of the proposed model, where the shaded and unshaded nodes indicate observed and latent variables, respectively. We consider that the time-of-day index s is observed since it is determined from the time of the population data.

3 INFERENCE

We present a variational Bayesian inference procedure for the proposed collective graphical mixture model. The observed variables are spatio-temporal population $\mathbf{N}$ and time-of-day data $\mathbf{g}$. The unknown variables are transition probabilities $\boldsymbol{\Theta}$, cluster assignments $\mathbf{z}$, cluster proportions $\boldsymbol{\phi}$, time mean $\boldsymbol{\tau}$, time precision $\boldsymbol{\eta}$, transition populations $\mathbf{M}$, and transition hyperparameters $\boldsymbol{\alpha}$. We approximate the posterior distributions of all unknown variables except for $\mathbf{M}$ and $\boldsymbol{\alpha}$. We obtain a point estimate of $\mathbf{M}$ and $\boldsymbol{\alpha}$ by maximizing the lower bound of the log marginal likelihood since approximate posterior distributions of $\mathbf{M}$ and $\boldsymbol{\alpha}$ cannot be analytically obtained.

We approximate the following true posterior distributions

$$p(\boldsymbol{\Theta},\boldsymbol{\phi},\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta}|\mathbf{M},\mathbf{N},\mathbf{g},\mathbf{s},\boldsymbol{\alpha},\beta,a,b,d,f), \tag{7}$$

by using the mean-field family

$$q(\boldsymbol{\Theta},\boldsymbol{\phi},\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta}) = \prod_{k=1}^K \prod_{i=1}^I q(\boldsymbol{\theta}_{ki}) q(\boldsymbol{\phi}) \prod_{s=1}^S q(z_s) \prod_{k=1}^K q(\tau_k,\eta_k), \tag{8}$$

which are called variational distributions. The following lower bound of the log marginal likelihood is obtained by applying Jensen's inequality,

$$\begin{aligned} &\log p(\mathbf{M}|\mathbf{N},\mathbf{s},\boldsymbol{\alpha},\beta,a,b,d,f) \geq \mathbb{E}[\log p(\mathbf{M}|\mathbf{N},\boldsymbol{\Theta},\mathbf{z},\mathbf{s})] + \mathbb{E}[\log p(\boldsymbol{\Theta}|\boldsymbol{\alpha})] + \mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\phi})] \\ &+ \mathbb{E}[\log p(\boldsymbol{\phi}|\beta)] + \mathbb{E}[\log p(\mathbf{g}|\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta})] + \mathbb{E}[\log p(\boldsymbol{\tau}|f,d,\boldsymbol{\eta})] + \mathbb{E}[\log p(\boldsymbol{\eta}|a,b)] \\ &+ \mathbb{H}[q(\boldsymbol{\Theta})] + \mathbb{H}[q(\boldsymbol{\phi})] + \mathbb{H}[q(\mathbf{z})] + \mathbb{H}[q(\boldsymbol{\tau})] + \mathbb{H}[q(\boldsymbol{\eta})] \equiv F, \end{aligned} \tag{9}$$

where $\mathbb{E}[f]$ represents the expectation of f with variational distributions q, and $\mathbb{H}[q]$ represents the entropy of q. The complete form of the lower bound is described in Appendix A.

We iteratively optimize each of the variational distributions and unknown variables while keeping the others fixed by maximizing the lower bound (9). Since we use conjugate priors, the variational distributions to be estimated have the same form as their priors as follows,

$$q(\boldsymbol{\theta}_{ki}) = \text{Dirichlet}(\boldsymbol{\alpha}'_{ki}), \tag{10}$$

$$q(\boldsymbol{\phi}) = \text{Dirichlet}(\boldsymbol{\beta}'), \tag{11}$$

$$q(\tau_k) = \text{Normal}(f'_k, (d'_k \eta_k)^{-1}), \tag{12}$$

$$q(\eta_k) = \text{Gamma}(a'_k, b'_k), \tag{13}$$

where $\boldsymbol{\alpha}'_{ki} = (\alpha'_{kij})_{j \in \mathbf{J}_i}$, and $\boldsymbol{\beta}' = (\beta'_k)_{k=1}^K$. Let

$$q_{sk} \equiv q(z_s = k), \tag{14}$$

be the variational probability that the cluster assignment of time-of-day s is k. Its update equation is given by

$$\begin{aligned} \log q_{sk} \propto\ & \Psi(\beta'_k) - \Psi(\bar{\beta}') + \sum_{i=1}^{I} \sum_{j \in \mathbf{J}_i} M_{tij} [\Psi(\alpha'_{kij}) - \Psi(\bar{\alpha}'_{ki})] \\ & + \frac{1}{2}\Psi(a'_k) - \frac{1}{2}\log b'_k - \frac{1}{2d'_k} - \frac{a'_k}{2b'_k}(g_s - f'_k)^2, \end{aligned} \tag{15}$$

where $\Psi(\cdot)$ is the digamma function defined by $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$, $\bar{\alpha}'_{ki} = \sum_{j \in \mathbf{J}_i} \alpha'_{kij}$, and $\bar{\beta}' = \sum_{k=1}^K \beta'_k$. The update equations of unknown parameters in the variational distributions (10)–(13) are given by

$$\alpha'_{kij} = \alpha_{r(i,j)} + \sum_{t=1}^{T-1} q_{s(t)k} M_{tij}, \tag{16}$$

$$\beta'_k = \beta + \sum_{s=1}^{S} q_{sk}, \tag{17}$$

$$d'_k = d + \sum_{s=1}^{S} q_{sk}, \tag{18}$$

$$f'_k = d'^{-1}_k (df + \sum_{s=1}^{S} q_{sk} g_s), \tag{19}$$

$$a'_k = a + \frac{\sum_{s=1}^{S} q_{sk} + 1}{2}, \tag{20}$$

$$b'_k = b + \frac{1}{2} \sum_{s=1}^{S} q_{sk} (g_s - f'_k)^2 + \frac{d}{2}(f'_k - f)^2. \tag{21}$$

We obtain a point estimate of transition populations $\mathbf{M}$ by maximizing the lower bound (9). In the lower bound, only the first term $\mathbb{E}[\log p(\mathbf{M}|\mathbf{N},\Theta,\mathbf{z},\mathbf{s})]$ (30) depends on $\mathbf{M}$, whose approximation is given by

$$\begin{aligned}\mathbb{E}[\log p(\mathbf{M}|\mathbf{N},\Theta,\mathbf{z},\mathbf{s})] &= \int \sum_{\mathbf{z}} q(\Theta)q(\mathbf{z}) \log p(\mathbf{M}|\mathbf{N},\Theta,\mathbf{z},\mathbf{s}) d\Theta \\ &\propto \sum_{t=1}^{T-1}\sum_{i=1}^{I}\sum_{j\in\mathbf{J}_i}\Big(-\log M_{tij}! + M_{tij}\sum_{k=1}^{K} q_{s(t)k}[\Psi(\alpha'_{kij}) - \Psi(\bar{\alpha}'_{ki})]\Big) \\ &\approx \sum_{t=1}^{T-1}\sum_{i=1}^{I}\sum_{j\in\mathbf{J}_i} M_{tij} - M_{tij}\log M_{tij} + M_{tij}\sum_{k=1}^{K} q_{s(t)k}[\Psi(\alpha'_{kij}) - \Psi(\bar{\alpha}'_{ki})] \equiv L, \end{aligned} \tag{22}$$

where $\log N_{ti}!$ is omitted from the second line since it does not depend on $\mathbf{M}$, and Stirling's approximation, $\log M! \approx M\log M - M$, is used at the third line as in [17]. The constraints (2) and (3) might not hold in real-world data because errors in sensing are inevitable and populations would be modified to preserve privacy. We incorporate them as soft constraints, and try to minimize the squared difference between the left- and right-hand sides in each of (2) and (3). Then, the objective function to be maximized becomes as follows,

$$L' = L - \frac{\lambda}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{I}|N_{ti} - \sum_{j\in\mathbf{J}_i} M_{tij}|^2 - \frac{\lambda}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{I}|N_{t+1,i} - \sum_{j\in\mathbf{J}_i} M_{tji}|^2, \tag{23}$$

where the second and third terms in the left-hand side correspond to the soft constraints of (2) and (3), respectively, and $\lambda \geq 0$ is a hyperparameter for controlling the penalty for violating the constraints. Since integer programming problems require a prohibitively large amount of computational time, we allow integer-valued variable $\mathbf{M}$ to take any real value, which enables us to maximize the objective function efficiently. We maximize the objective function L' with respect to $\mathbf{M}$ by using the quasi-Newton method [9] with non-negative bound constraint $\mathbf{M} \geq 0$. The gradient of the objective function L' is given by

$$\frac{\partial L'}{\partial M_{tij}} = -\log M_{tij} + \sum_{k=1}^{K} q_{s(t)k}[\Psi(\alpha'_{kij}) - \Psi(\bar{\alpha}'_{ki})] + \lambda(N_{ti} - \sum_{j\in\mathbf{J}_i} M_{tij}) + \lambda(N_{t+1,i} - \sum_{j\in\mathbf{J}_i} M_{tji}). \tag{24}$$

The objective function L' is convex with respect to $\mathbf{M}$ since the first and third terms in (22) are linear, the second term in (22), $M_{tij}\log M_{tij}$, is convex, and the second and third terms in (23) are convex quadratic functions. Therefore, the global optimum solution for $\mathbf{M}$ is obtained when the other parameters are fixed. Note that $\mathbf{M}_t$ can be optimized in parallel since the objective function L' is decomposed into the sum of $T-1$ terms each of which contains only the transition population at t, $\mathbf{M}_t$.

We estimate the transition hyperparameters $\boldsymbol{\alpha}$ by maximizing the lower bound (9) using the fixed-point iteration method described in [12]. The update rule is given by

$$\alpha_r \leftarrow \alpha_r \frac{\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j\in\mathbf{J}_i}\mathbb{I}(r(i,j)=r)[\Psi(\alpha'_{kij}) - \Psi(\alpha_r)]}{\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j\in\mathbf{J}_i}\mathbb{I}(r(i,j)=r)[\Psi(\bar{\alpha}'_{ki}) - \Psi(\bar{\alpha})]}, \tag{25}$$

where $\mathbb{I}(A)$ is the indicator function, i.e. $\mathbb{I}(A) = 1$ is A is true and $\mathbb{I}(A) = 0$ otherwise.

We can obtain a local optimum for the model parameters. Algorithm 1 summarizes the variational Bayesian inference procedure for the proposed collective graphical mixture model. The end

ALGORITHM 1: Variational Bayesian inference procedure for the proposed collective graphical mixture model.

Input: spatio-temporal population data $\mathbf{N}$, neighbor information $(\mathbf{J}_i)_{i=1}^I$, the number of clusters K, hyperparameters $a, b, f, d, \beta, \lambda$
Output: parameters of variational distributions $((\boldsymbol{\alpha}'_{ki})_{i=1}^I)_{k=1}^K$, $\boldsymbol{\beta}'$, $(f'_k)_{k=1}^K$, $(d'_k)_{k=1}^K$, $(a'_k)_{k=1}^K$, $(b'_k)_{k=1}^K$, $((q_{sk})_{k=1}^K)_{s=1}^S$, transition population $\mathbf{M}$, transition hyperparameters $\boldsymbol{\alpha}$
initialize $((q_{sk})_{k=1}^K)_{s=1}^S$, $\mathbf{M}$, $\boldsymbol{\alpha}$;
repeat
  update α'_{kij} by (16) for $k = 1$ to K, $i = 1$ to I, $j \in \mathbf{J}_i$;
  update β'_k by (17) for $k = 1$ to K;
  update d'_k by (18) for $k = 1$ to K;
  update f'_k by (19) for $k = 1$ to K;
  update a'_k by (20) for $k = 1$ to K;
  update b'_k by (21) for $k = 1$ to K;
  update q_{sk} by (15) for $s = 1$ to S, $k = 1$ to K;
  update $\mathbf{M}$ by maximizing (24);
  update α_r by (25) for all relative position indices;
until;

condition can be based on the number of iterations, or the convergence of the following value

$$F - \frac{\lambda}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{I}|N_{ti} - \sum_{j\in\mathbf{J}_i} M_{tij}|^2 - \frac{\lambda}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{I}|N_{t+1,i} - \sum_{j\in\mathbf{J}_i} M_{tji}|^2, \tag{26}$$

which is the sum of the lower bound of the log marginal likelihood (9) and soft constraints that are used in (24). In the experiments, we fixed the hyperparameters as follows: $\beta = 10^{-2}, a = 1, b = 1, f = 12, d = 1$. It is impossible to tune optimal soft constraint parameter λ for estimating flow, since our task is unsupervised and any people flow data are not given. Therefore, λ is set by using the predictive performance of the population of the next time point in the training data for each data set and for each number of clusters.

The computational complexity for an iteration with the proposed method is $O(TIJK)$; it increases linearly with respect to the number of time points T, the number of grid cells I, the number of neighbors J, and the number of clusters K. Note that it does not depend on the population size $\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}$ since by using the framework of collective graphical models and continuous approximation of population $\mathbf{M}$, we only deal with the sufficient statistics instead of individual behavior data. Therefore, the proposed method is applicable to large quantities of population data.

The transition probability at grid cell i at time t is estimated using estimated $\boldsymbol{\alpha}'_{ki}$ and q_{sk} by

$$\hat{\theta}_{tij} = \frac{\alpha'_{\hat{z}_{s(t)}ij}}{\bar{\alpha}'_{\hat{z}_{s(t)}i}}, \tag{27}$$

where $\hat{z}_{s(t)} = \arg\max_k q_{s(t)k}$ is the estimated cluster assignment of time t. The expected number of people who move from grid cell i to grid cell j at time t is calculated by $\hat{\theta}_{tij}N_{ti}$. Then, the population of the next time point given the current population $\mathbf{N}_{ti}$ is predicted by

$$\hat{N}_{t+1,i} = \sum_{j\in\mathbf{J}_i} \hat{\theta}_{tij}N_{ti}, \tag{28}$$

since the sum of the transition population to i is equal to the population at i at the next timestep.

4 EXPERIMENTS

4.1 Data

We evaluated the proposed method using real-world spatio-temporal population data sets obtained in Tokyo, Osaka, Nagoya and Beijing. Figure 4 shows examples of gridded population at 0:00, 6:00, 12:00 and 18:00 on a day in Tokyo, Osaka, Nagoya and Beijing data. With all of the data sets, grid cells are square, and the neighbors are its surrounding eight cells with the addition of the cell itself. For the grid cells at corners or edges, the outside cells are removed from their neighbor sets.

The original data of the Tokyo, Osaka and Nagoya data sets [4] contained latitude and longitude information for each person obtained every five minutes that were interpolated using railway and road information [15] from geotagged tweets. We created gridded spatio-temporal population data, where the time interval was 30 minutes, the size of each grid cell was 10km $\times$ 10km, and there were 16×14 grid cells. The Tokyo data consisted of data on July 1st, July 7th, October 7th, October 13th, December 16th, and December 22nd of 2013, where the numbers of people were 6,432, 9,166, 6,822, 10,134, 6,646 and 10,338, respectively. The Osaka data consisted of data on August 8th, August 11th, September 16th, September 22th, December 24th, and December 29nd of 2013, where the number of people were 2,256, 3,034, 2,999, 3,569, 2,487 and 3,480 respectively. The Nagoya data consisted of data on July 22th, July 28th, September 16th, September 22th, December 24th, and December 29nd of 2013, where the numbers of people were 929, 1,332, 1,148, 1,460, 975 and 1,570, respectively.

The Beijing data consist of population in Beijing from 3rd February to 7th of 2008, which were obtained from T-Drive trajectory data [25, 26], which contained trajectories of 10,357 taxies. We created gridded spatio-temporal population data and with 15-minute time intervals, 2km $\times$ 2km size grid, and 20×16 grid cells in total. The length of the area was 40km in the north-south direction, and 32km in the east-west direction.

We used the data sets whose original data contain trajectories of twitter users or taxies so that we can evaluate the people flow estimation performance. We did not use trajectory information for the estimation.

4.2 Results

We evaluated the proposed variational Bayesian collective graphical mixture model (VCGMM) in terms of the estimation performance of the people flow. For comparing methods, we used the following four methods: CGM, ICGM, VCGM and STAY. The CGM is a collective graphical model obtained by a maximum likelihood estimation with the same soft constraints with the proposed model, where the transition probability for each grid cell is fixed over time. The ICGM is the inhomogeneous transition probability CGM, which assumes that the transition probability for each grid cell is different across different time points. The VCGM is a collective graphical model obtained with a variational Bayesian inference. When there is one mixture, the proposed method corresponds to the VCGM. CGM and VCGM assume that people flows do not change over time, while VCGMM and ICGM assume that they change over time. The STAY method predicts the population at the next time point from the current population; it assumes that all people do not move, and stay in their current cells. With the proposed VCGMM, we used ten clusters.

[4] SNS-based People Flow Data, Nightley, Inc., Shibasaki & Sekimoto Laboratory, the University of Tokyo, Micro Geo Data Forum, People Flow project, and Center for Spatial Information Science at the University of Tokyo, http://nightley.jp/archives/1954

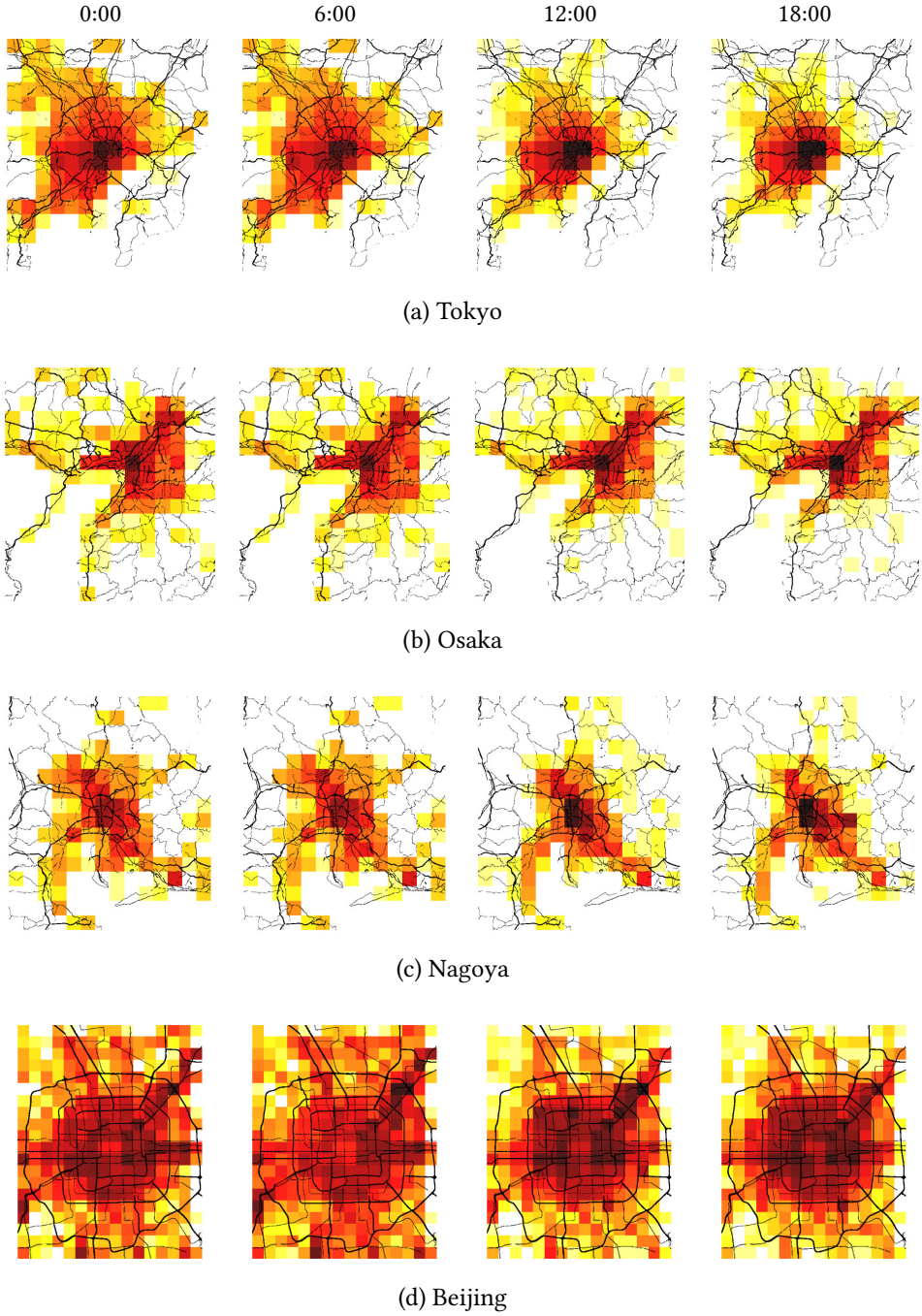


Fig. 4. Spatio-temporal population data (a) in Tokyo on July 1st 2013, (b) in Osaka August 8th 2013, (c) in Nagoya on July 22th 2013, and (d) in Beijing on February 3rd 2008. Darker colors represent higher population densities in each grid cell.

Table 2. Normalized absolute errors when predicting people flow averaged over all time points.

	VCGMM	ICGM	VCGM	CGM	STAY
Tokyo	**0.167**	0.176	0.208	0.208	0.192
Osaka	**0.250**	0.265	0.280	0.280	0.272
Nagoya	**0.250**	0.281	0.292	0.291	0.269
Beijing	**0.470**	0.500	0.479	0.479	0.532

For a evaluation measurement, we used the following normalized absolute error on people flow,

$$\frac{\sum_{i=1}^{I}\sum_{t=1}^{T-1}\sum_{j\in \mathrm{J}_i}|M^*_{tij}-M_{tij}|}{\sum_{i=1}^{I}\sum_{t=1}^{T-1}N_{ti}}, \tag{29}$$

where M^*_{tij} is the true number of transition people who moved from grid cell i to the jth neighbor at time t, and M_{tij} is its estimation. The true number of transition people, which is the test data, was obtained from the original data that contain individual trajectories. Note that for estimation we used population data for each grid cell and each time points, but did not use transition population data between grid cells. Table 2 shows the error of VCGMM, ICGM, VCGM, CGM and STAY methods. The proposed method, VCGMM, achieved the lowest error, which implies its effectiveness on people flow estimation. Since the ICGM estimates the transition probability using only the data at that time point, its performance was worse than the proposed method. The errors by the VCGM and CGM were almost the same. The VCGM and CGM were better than the STAY with the Beijing data, but worse with the Tokyo, Osaka and Nagoya data. It is because that many people stay in the same cell with the Tokyo, Osaka and Nagoya data, and many people move to different cells with the Beijing data, which is reasonable since the Beijing data is taxi trajectories. The proposed method achieved the best performance in both types of data by flexibly changing flows depending on time points.

Figure 5 shows the normalized absolute errors with different numbers of clusters with the proposed VCGMM. As the number of clusters increased, the error decreased. This result indicates the importance of using multiple people flow patterns. The error became relatively steady after seven, three and three with the Tokyo, Osaka and Nagoya data, respectively, and the VCGMM did not overfit the training data even with many clusters.

Figure 6 shows the estimated cluster proportions q_{sk} provided by the VCGMM with ten clusters. For each time-of-day index, only one cluster proportion was estimated as nearly one, and the other cluster proportions were estimated as nearly zero. Consecutive time points were assigned to the same cluster. Nine, seven, six and nine clusters were used with the Tokyo, Osaka, Nagoya and Beijing data sets, respectively. With the variational Bayesian inference, the proportions for unnecessary clusters become zero, and the proportions for necessary clusters become non-zero, which helps to avoid overfitting.

Figure 7 shows people flows estimated by the proposed method. With the Tokyo data, From 0:00 to 6:00, there were few flows; this is reasonable since most people are sleeping at home at this time. From 8:30 to 11:00, people move to the center of Tokyo from the suburbs for work. There were still flows to the city center from 14:30 to 15:00. From 18:30 to 23:00, flows from the city center to the suburbs were estimated since at that time people return home from their offices. With the Osaka data, similar temporal flow patterns were extracted. With the Nagoya data, flows to the city center were discovered in the day time. Since its population size was smaller than that of the Tokyo and Osaka data, estimated flows were not clear. With the Beijing data, the proposed method estimated that people move to the center in the morning.

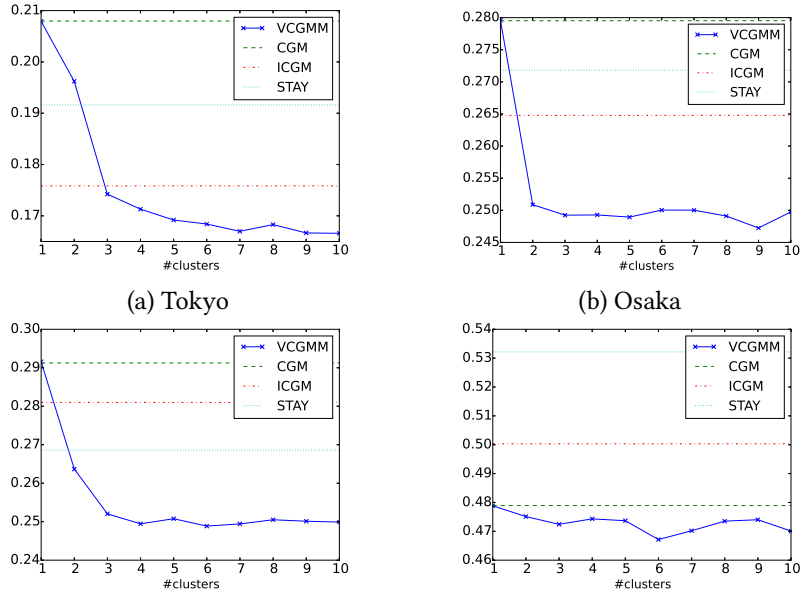


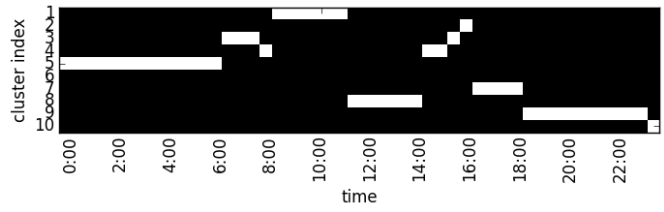
Fig. 5. Normalized absolute errors with different numbers of the clusters with the proposed VCGMM.

5 RELATED WORK

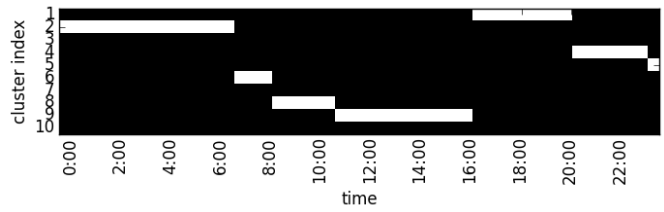
Great interest is being shown in developing methods to analyze people trajectories for a wide variety of applications, such as disaster management [19, 20], marketing [5], public health [2, 3], urban planning [14, 24, 27], traffic forecasting [8], traffic anomaly detection [11], transportation system management [4], and travel route recommendation [7]. These methods require the trajectories of individuals. However, because such trajectories are private information, or because tracking individuals is difficult, only aggregated population data might be available as mobile spatial statistics [22]. The proposed method can estimate people flows using aggregated data without trajectory data, and it would make existing trajectory data mining methods applicable even when trajectory data are unavailable.

Collective graphical models have been proposed and used for modeling contingency tables [18], bird migration [16], and infection [6], but have not been applied to population data. The proposed model is based on collective flow diffusion models [6]. We extend the collective flow diffusion models to mixture models for handling changes in flow patterns over time. To our knowledge, the proposed model is the first mixture model consisting of collective graphical models. To infer hidden variables in collective graphical models, several algorithms have been proposed, such as Markov chain Monte Carlo (MCMC) [18], maximum a posteriori (MAP) [17], the expectation propagation with Gaussian approximation [10], and the belief propagation [21]. The variational Bayesian inference procedure presented in this paper is applicable not only to a mixture of collective graphical models, but also to other collective graphical models.

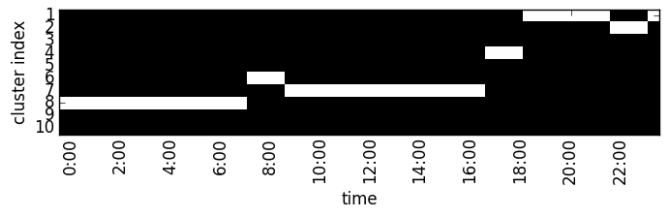
The CGMs in [10, 17, 18, 21] consider noisy observations, where a noise model, such as the Poission distribution, is assumed. The proposed model does not consider noise models for observations. However, by using the soft constraints for flow conservation, the proposed model can handle noisy observations, where flow is not strictly preserved.



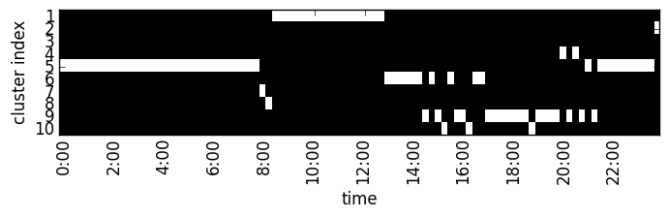
(a) Tokyo



(b) Osaka



(c) Nagoya



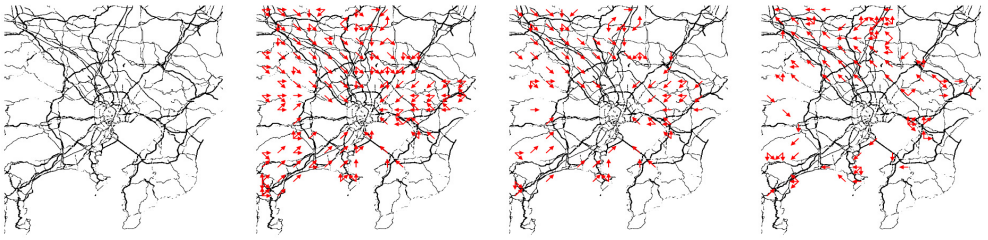
(d) Beijing

Fig. 6. Estimated cluster proportions q_{sk} provided by the proposed method. White represents probability one, and black represents zero. The cluster index is sorted by the start time point.

6 CONCLUSION

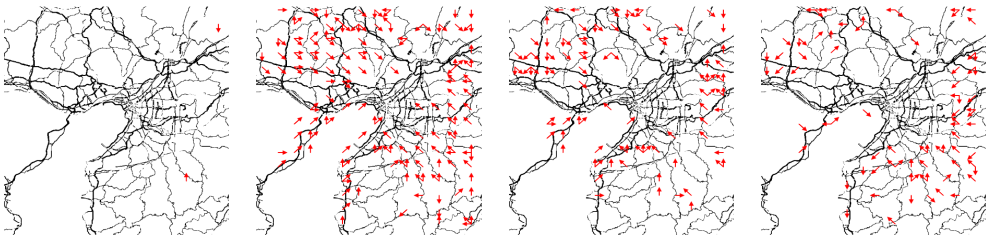
We have proposed a mixture of collective graphical models for estimating people flows from spatio-temporal population data, and developed a variational Bayesian inference procedure for the proposed model. We confirmed experimentally that our proposed method can estimate flow that changes over time, and predict the population at the next time point more precisely than existing methods.

Although our results are encouraging, our framework can be further improved in a number of ways. Firstly, we plan to incorporate the spatial correlation of flows by using location dependent



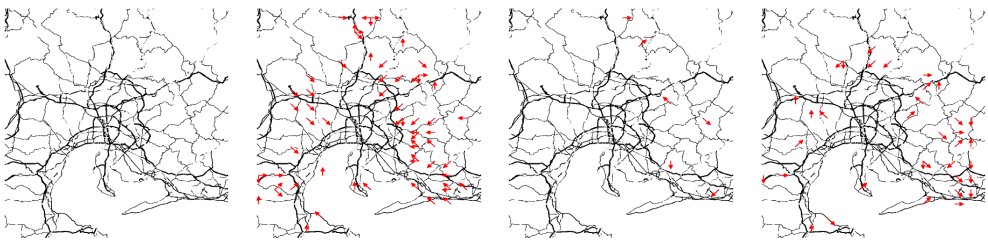
0:00–6:00 8:30–11:00 14:30–15:00 18:30–23:00

(a) Tokyo



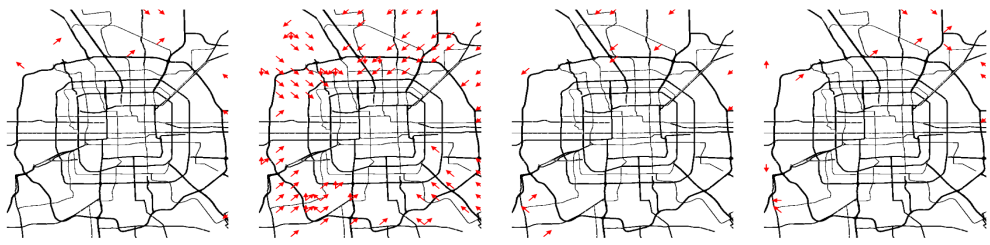
0:00–6:30 8:30–10:30 11:00–16:00 16:30–20:00

(b) Osaka



0:00–7:00 9:00–16:30 17:00–18:00 18:30–21:30

(c) Nagoya



0:00–7:45 8:30–12:45 13:00–16:45 14:30–21:15

(d) Beijing

Fig. 7. Estimated people flows for each time-of-day cluster obtained with the proposed method using the (a) Tokyo, (b) Osaka, (c) Nagoya and (d) Beijing data. When an estimated transition probability is higher than a threshold, an arrow is drawn in that direction.

mixture models as we did to incorporate time correlation in this paper. Secondly, we would like to extend the proposed model to use information about day of week. Thirdly, we will study the effectiveness of the proposed method with different grid sizes, time intervals and neighborhood settings. Finally, it is important to develop a method for selecting neighbors that are appropriate for the given data.

REFERENCES

[1] Matthew J. Beal and Zoubin Ghahramani. 2003. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics* 7 (2003), 453–464.

[2] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 2007. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine* 5, 1 (2007), 34.

[3] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 6988 (2004), 180–184.

[4] Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, and Michael Pazzani. 2010. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 899–908.

[5] Ronald L Hess, Ronald S Rubin, and Lawrence A West. 2004. Geographic information systems as a marketing information system technology. *Decision Support Systems* 38, 2 (2004), 197–212.

[6] Akshat Kumar, Daniel Sheldon, and Biplav Srivastava. 2013. Collective diffusion over networks: Models and inference. In *Proceedings of International Conference on Uncertainty in Artificial Intelligence*.

[7] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. 2010. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 579–588.

[8] Marco Lippi, Matteo Bertini, and Paolo Frasconi. 2010. Collective traffic forecasting. *Machine Learning and Knowledge Discovery in Databases* (2010), 259–273.

[9] Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45, 1-3 (1989), 503–528.

[10] Li-Ping Liu, Daniel Sheldon, and Thomas G Dietterich. 2014. Gaussian Approximation of Collective Graphical Models. In *Proceedings of International Conference on Machine Learning*.

[11] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1010–1018.

[12] Thomas Minka. 2000. Estimating a Dirichlet distribution. (2000).

[13] A.B.M. Musa and Jakob Eriksson. 2012. Tracking unmodified smartphones using Wi-Fi monitors. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 281–294.

[14] Guande Qi, Xiaolong Li, Shijian Li, Gang Pan, Zonghui Wang, and Daqing Zhang. 2011. Measuring social functions of city regions from large-scale taxi behaviors. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 384–388.

[15] Yoshihide Sekimoto, Ryosuke Shibasaki, Hiroshi Kanasugi, Tomotaka Usui, and Yasunobu Shimazaki. 2011. PFlow: Reconstructing people flow recycling large-scale social survey data. *IEEE Pervasive Computing* 10, 4 (2011), 0027–35.

[16] D. Sheldon, M.A.S. Elmohamed, and D. Kozen. 2007. Collective inference on Markov models for modeling bird migration. In *Advances in Neural Information Processing Systems*. 1321–1328.

[17] Daniel Sheldon, Tao Sun, Akshat Kumar, and Tom Dietterich. 2013. Approximate inference in collective graphical models. In *Proceedings of International Conference on Machine Learning*.

[18] Daniel R Sheldon and Thomas G Dietterich. 2011. Collective graphical models. In *Advances in Neural Information Processing Systems*. 1161–1169.

[19] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Teerayut Horanont, Satoshi Ueyama, and Ryosuke Shibasaki. 2013. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1231–1239.

[20] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 5–14.

[21] Tao Sun, Daniel Sheldon, Akshat Kumar, and EDU SG. 2015. Message Passing for Collective Graphical Models. In *Proceedings of International Conference on Machine Learning*. 853–861.

[22] Masayuki Terada, Tomohiro Nagata, and Motonari Kobayashi. 2012. Population estimation technology for mobile spatial statistics. *NTT DOCOMO Technical Journal* 14, 3 (2012), 10–15.
[23] Apichon Witayangkurn, Teerayut Horanont, and Ryosuke Shibasaki. 2013. The Design of Large Scale Data Management for Spatial Analysis on Mobile Phone Dataset. *Asian Journal of Geoinformatics* 13, 3 (2013).
[24] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 186–194.
[25] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. 2011. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 316–324.
[26] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 99–108.
[27] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. 2011. Urban computing with taxicabs. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. ACM, 89–98.

A VARIATIONAL LOWER BOUND

In this appendix, we give the the lower bound of the log marginal likelihood (9). The first term is given by

$$\begin{aligned}&\mathbb{E}[\log p(\mathbf{M}|\mathbf{N},\boldsymbol{\Theta},\mathbf{z},\mathbf{s})]\\&=\sum_{t=1}^{T-1}\sum_{i=1}^{I}\Big(\log N_{ti}!-\sum_{j\in\mathbf{J}_i}\log M_{tij}!+\sum_{j\in\mathbf{J}_i}M_{tij}\sum_{k=1}^{K}q_{s(t)k}[\Psi(\alpha'_{kij})-\Psi(\bar{\alpha}'_{ki})]\Big),\end{aligned}\tag{30}$$

where we used the following expected value of the log of a multinomial parameter under a Dirichlet distribution,

$$\mathbb{E}_{q(\boldsymbol{\alpha}'_{ki})}[\log\theta_{kij}]=\Psi(\alpha'_{kij})-\Psi(\bar{\alpha}'_{ki}).\tag{31}$$

By using the above equation, the second, third and fourth terms become

$$\begin{aligned}\mathbb{E}[\log p(\boldsymbol{\Theta}|\boldsymbol{\alpha})]=\sum_{k=1}^{K}\sum_{i=1}^{I}\Big(&\log\Gamma(\sum_{j\in\mathbf{J}_i}\alpha_{r(i,j)})-\sum_{j\in\mathbf{J}_i}\log\Gamma(\alpha_{r(i,j)})\\&+\sum_{j\in\mathbf{J}_i}(\alpha_{r(i,j)}-1)[\Psi(\alpha'_{kij})-\Psi(\bar{\alpha}'_{ki})]\Big),\end{aligned}\tag{32}$$

$$\mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\phi})]=\sum_{s=1}^{S}\sum_{k=1}^{K}q_{sk}[\Psi(\beta'_k)-\Psi(\bar{\beta}')],\tag{33}$$

$$\mathbb{E}[\log p(\boldsymbol{\phi}|\beta)]=\log\Gamma(\beta K)-K\log\Gamma(\beta)+(\beta-1)\sum_{k=1}^{K}[\Psi(\beta'_k)-\Psi(\bar{\beta}')].\tag{34}$$

The fifth, sixth and seventh terms are given by

$$\mathbb{E}[\log p(\mathbf{g}|\mathbf{z},\boldsymbol{\tau},\boldsymbol{\eta})]=\frac{1}{2}\sum_{k=1}^{K}\Big(T_k[\Psi(a'_k)-\log b'_k]-\frac{T_k}{d'_k}-\frac{a'_k}{b'_k}\sum_{s=1}^{S}q_{sk}(g_s-f'_k)^2\Big)-\frac{T}{2}\log 2\pi,\tag{35}$$

$$\mathbb{E}[\log p(\boldsymbol{\tau}|f,d,\boldsymbol{\eta})]=\frac{1}{2}\sum_{k=1}^{K}\Big(\Psi(a'_k)-\log b'_k+\log\frac{d}{2\pi}-\frac{d}{d'_k}-\frac{da'_k}{b'_k}(f'_k-f)^2\Big),\tag{36}$$

$$\mathbb{E}[\log p(\boldsymbol{\eta}|a,b)] = -K\log\Gamma(a) + Ka\log b + (a-1)\sum_{k=1}^{K}(\Psi(\frac{a'_k}{2}) - \log\frac{b'_k}{2}) - b\sum_{k=1}^{K}\frac{a'_k}{b'_k}, \quad (37)$$

where we used the following expected value of a positive parameter under a Gamma distribution,

$$\mathbb{E}_{q(\eta_k)}[\log\eta_k] = \Psi(a'_k) - \log b'_k. \quad (38)$$

The eighth term is the entropy of the Dirichlet distribution, and is given by,

$$\mathbb{H}[q(\boldsymbol{\Theta})] = \sum_{k=1}^{K}\sum_{i=1}^{I}\Big(\sum_{j\in\mathbf{J}_i}\log\Gamma(\alpha'_{kij}) - \log\Gamma(\bar{\alpha'}_{ki}) - \sum_{j\in\mathbf{J}_i}(\alpha'_{kij}-1)[\Psi(\alpha'_{kij}) - \Psi(\bar{\alpha'}_{ki})]\Big). \quad (39)$$

The ninth term is the entropy of the discrete distribution, and is

$$\mathbb{H}[q(\mathbf{z})] = -\sum_{s=1}^{S} q_{sk}\log q_{sk}. \quad (40)$$

The tenth term is the entropy of the Dirichlet distribution, and is

$$\mathbb{H}[q(\boldsymbol{\phi})] = \sum_{k=1}^{K}\Gamma(\beta'_k) - \log\Gamma(\bar{\beta'}) - \sum_{k=1}^{K}(\beta'_k - 1)[\Psi(\beta'_k) - \Psi(\bar{\beta'})]. \quad (41)$$

The eleventh term is the entropy of the Gaussian distribution, and is

$$\mathbb{H}[q(\boldsymbol{\tau})] = \frac{K}{2}(\log 2\pi + 1) - \frac{1}{2}\sum_{k=1}^{K}\log d'_k - \frac{1}{2}\sum_{k=1}^{K}(\Psi(a'_k) - \log b'_k). \quad (42)$$

The twelfth term is the entropy of the Gamma distribution, and is

$$\mathbb{H}[q(\boldsymbol{\eta})] = \sum_{k=1}^{K}[\log\Gamma(a'_k) - (a'_k - 1)\Psi(a'_k) - \log b'_k + a'_k]. \quad (43)$$

Received February 2007; revised March 2009; accepted June 2009