Spatiality Preservable Factored Poisson Regression for Large-Scale Fine-Grained GPS-Based Population Analysis

Masamichi Shimosaka,[1] **Yuta Hayakawa,**[1] **Kota Tsubouchi**[2]
[1]Tokyo Institute of Technology, [2]Yahoo Japan Corporation
simosaka@miubiq.cs.titech.ac.jp, hayakawa@miubiq.cs.titech.ac.jp, ktsubouc@yahoo-corp.jp

Abstract

With the wide use of smartphones with Global Positioning System (GPS) sensors, the analysis of the population from GPS traces has been actively explored in the last decade. We propose herein a brand new population prediction model to capture the population trends in a fine-grained point of interest (POI) densely distributed over large areas and understand the relationship of each POI in terms of spatiality preservation. We propose a new framework, called *Spatiality Preservable Factorized Regression (SPFR)*, to realize this model. The SPFR is inspired by the success of the recently proposed bilinear Poisson regression and the concept of multi-task learning with factorization approach and the graph proximity regularization. Given that the proposed model is written simply in terms of optimization, we achieve scalability using our model. The results of our empirical evaluation, which used a massive dataset of GPS logs in the Tokyo region over 32 M count logs, show that our model is comparable to the state-of-the-art methods in terms of capturing the population trend across meshes while retaining spatial preservation in finer mesh areas.

The understanding of the flow of people in a city, which is known as the analysis of urban dynamics, is of great importance in urban planning, emergency management, and commercial activity. Questionnaire-based surveys, such as the person trip survey and the traffic flow survey, have been useful for understanding urban dynamics (Sekimoto et al. 2011). With the spread of smart devices, a large amount of mobility log data, such as the global positioning system (GPS) and cell tower logs, has been accumulated. Therefore, studies have paid attention to the analysis of urban dynamics using mobility logs from smartphones without other additional survey costs.

With the structured property of the population data across regions, time, and days obtained from GPS logs, the analytics based on spatio-temporal tensor factorization (Fan, Song, and Shibasaki 2014) and discriminative approach, such as neural network (Jiang et al. 2018), and Poisson regression (Okawa, Kim, and Toda 2017) have been established as urban dynamics analysis in the last decade. In the former approach, the tensor data, where the population data with respect to locations, time-zones, and days are accumulated, are analyzed by using non-negative tensor factorization (Fan, Song, and Shibasaki 2014). This approach is known to be prominent for analyzing urban dynamics to understand the functionality (Yao et al. 2017) and the population trend of each POI (Nishi, Tsubouchi, and Shimosaka 2014) when the stored data are sufficient in each tensor cell. However, it has a critical drawback when used for predictive tasks, such as congestion forecast (Konishi et al. 2016). This stems from the fact that the approach based on tensor factorization is originally designed for the data compression to intuitively interpret the past GPS records, whereas it provides an inaccurate performance without any near future observation. With regard to the well-known recommendation systems, this issue is akin to the out-of-example extension (i.e., cold start problem for the recommendation systems research) (Lika, Kolomvatsos, and Hadjiefthymiades 2014).

In contrast to the factorization approach, recent advances on urban dynamics based on the discriminative approach using contextual information, such as calendar information, weather information, and urban-scale event information, were reported to provide accurate prediction results (Jiang et al. 2018). However, its POI size in their analyses was prone to be large (e.g., approximately 1 km square with GPS data and $2 \sim 3$ km square with cellular network data). This can be attributed to the performance of the prediction-based approaches being degraded on small-sized regions. This is also analogous to the data sparseness problem in the counting problem (Lichman and Smyth 2018). Furthermore, the increases of the number of POIs were not negligible because of the high spatial granularity. Therefore, the previous work tended to choose the POI sizes, which were, by no means, small. Without the spatially fine-grained analysis (e.g., 100 m square per POI), the population trend in a commercial building, visualization of rush hour in front of the terminal stations, and the functionality of POI with small areas could not be available. In spite of the importance of the analysis with high spatial granularity, to the best of our knowledge, no work on the population analysis with a high spatial granularity, except (Xu et al. 2016), exists in the literature. Indeed, the work provides successful results on visualization and functionality analysis of each POI within a fine-grained way; however, this state-of-the-art work is designed for real-time estimation instead of the prediction task. Prediction tasks are highly significant for real-world applica-

tions. For example, personnel distribution for an event must be predetermined, and real-time estimation cannot be applied for this situation.

This study takes on the challenging new task of constructing a model that predicts urban dynamics at a finer POI scale. Urban dynamics in a fine-grained POI is more important and more practical than the analysis reported in the previous approach with respect to services, applications, marketing, and politics. However, we encounter unexperienced difficulties in performing this task. First, we face the challenge of sparse data caused by far less smartphone GPS logs in small-sized POIs compared to that in urban scales that causes the degradation of the prediction performance. Second, we face a non-negligible computational cost on training with respect to the increase of the number of POIs, thanks to the high spatial granularity.

We propose herein a new framework that provides an accurate prediction result for a large-scale fine grained urban dynamics analysis to mitigate the issues for urban dynamics analytics with spatial granularity: *Spatiality Preservable Factorized Regression (SPFR)*. We extend the idea of a recently proposed work on predictive population, called bilinear Poisson regression (Shimosaka et al. 2015), with the idea of the factorized regression being actively explored in recommendation systems (Xu, Zhou, and Tan 2015) for efficiency and robustness against a large number of POI analyses. Furthermore, when we consider that the target areas are densely distributed with the spatial granularity, we incorporate the idea that the proposed model retains the spatial preservability of population trends on the dense distributed POIs.

The main contributions of this work are summarized as follows:

- We propose a new framework, called *Spatiality Preservable Factorized Regression (SPFR)*, as a novel method to model large-scale urban dynamics at a finer-grained POI densely distributed over large areas. To the best of our knowledge, this is the first work that attempts to model and predict fine-grained urban dynamics incorporating vital domain knowledge: spatiality preservable factorization. The derived statistical model can be simply formalized and optimized by a simple sequence of convex optimizations.

- We use a large-scale real-world dataset and show the proposed scheme's performance compared with that of the state-of-the-art methods in large-sized POIs while the performance is still robust against the data sparseness problem at the spatial granularity scale analysis (e.g., 100 m square mesh size). Furthermore, the proposed model suppresses the computational cost against the increase of the number of POIs in comparison with the state-of-the-art techniques.

- We also show that the proposed model could be a helpful tool for the visualization perspective in terms of capturing the relationship among the meshes and the basic patterns inherited from the demographics of the POIs.

Related work

The large amount of available mobile phone location data has motivated research on urban dynamics. A typical approach in urban dynamics research is the extraction of active population patterns over the course of a day. Researchers often use mixture modeling (Shimosaka et al. 2016), tensor factorization or matrix factorization (Zheng et al. 2014; Takeuchi et al. 2013; Fan, Song, and Shibasaki 2014; Zhang et al. 2015; Yuan, Zheng, and Xie 2012), or eigen decomposition (Reades, Calabrese, and Ratti 2009) to extract basic dynamics patterns in the dataset. Mixture modeling is frequently used as a simple approach for discovering the latent structure of population trends across the meshes (Shimosaka et al. 2016). However, these techniques are not feasible for the use of forecasting congestions of specific areas from the model (Konishi et al. 2016), which is analogous to the cold-start problem on recommendation systems (Lika, Kolomvatsos, and Hadjiefthymiades 2014).

Discriminative methods for active population forecast using features have been proposed as an alternative to the tensor factorization approach (Shimosaka et al. 2015; Zhang, Zheng, and Qi 2017; Jiang et al. 2018; Okawa, Kim, and Toda 2017). Most of these approaches focus on the accuracy of the given POI; however, they tend to choose highly crowded areas, such as stations and amusement areas, and their POI size tends to be large because of the data sparseness problem in a fine-grained POI. (Shimosaka et al. 2015) heuristically chose points of interest by focusing on the number of commuters or popular sightseeing locations. In contrast to this selection, some applications in urban computing requires analytics for densely distributed fine-grained POIs. In this sense, it is not feasible to handle a fine-grained POI densely distributed over wider residential and congestion areas in a unified manner.

From the viewpoint of statistical modeling, the proposed model is inspired by the success of the factorized regression techniques developed in recommendation systems (Yang, Zhao, and Gao 2017; Xu, Zhou, and Tan 2015). However, the performance of population prediction in a fine-grained POI is degraded even if we employ the idea of the factorization approach. In other words, spatiality preservation, which is a brand new concept derived herein, suppresses the performance drawback for this issue. Locality preservable tensor factorizations were proposed (Cai et al. 2009; An, Liu, and Ruan 2017), but they require the approximation of the objective function for the spatiality preservation, which causes instability in learning. Our proposed method optimizes parameters with iterations of convex programming, and its computational cost is at the same level as those in previous factorization methods.

The rest of this paper is organized as follows: Section 3 describes the formalization of the population pattern and the base model considered herein; Section 4 presents our proposed statistical model; Section 5 describes the experiments used to verify our model; and finally, Section 6 presents the conclusions of this study.

Problem setting

We model herein the daily transitions of an active population in certain target areas. For simplicity, we consider the subsection square area (e.g., 100 m $\times$ 100 m) as point of interest (POI) inside the whole area of interest as entire areas for prediction. We could use road network information to segment the whole areas of interest to generate the adaptive size of the POI (Xu et al. 2016); however, we assume that the POIs are arranged in a surface inside the whole areas of interest. For simplicity, we call a POI with the square shape *mesh*. Regarding the problem of active population prediction, we define the number of mobile phone logs within a certain duration in a certain day in a single mesh as an active population.

We divide the d-th day into S segments, then evaluate the number of GPS logs in l-th mesh as the active population number. Let τ be the index of time segment of the day. Given that the population number is influenced by the day of the week and the weather condition, let $\boldsymbol{c}$ denote these conditions as well as l, d, and τ.

In this setting, the main issue to be solved is the precise prediction of the $y^{(d,l)}_{\boldsymbol{c},\tau} \in \mathbb{Z}$ sequence across $\tau = 1 \ldots, S$ from the conditions l, $\boldsymbol{c}$. Note that condition $\boldsymbol{c}$ is a kind of tuple containing the days of the week and the weather conditions. The details of $\boldsymbol{c}$ will be described in the experimental results.

Prediction using bilinear Poisson regression model

As the basis of the model proposed herein, we describe the bilinear Poisson regression proposed in (Shimosaka et al. 2015). This assumes that $y^{(d,l)}_{\boldsymbol{c},\tau}$ is drawn from the Poisson distribution defined as

$$y^{(d,l)}_{\boldsymbol{c},\tau} \sim \mathcal{P}(y^{(d,l)}_{\boldsymbol{c},\tau} | \lambda^{(l)}_{\boldsymbol{c},\tau}) = \frac{\lambda^{y^{(d,l)}_{\boldsymbol{c},\tau}} \exp(-\lambda^{(l)}_{\boldsymbol{c},\tau})}{y^{(d,l)}_{\boldsymbol{c},\tau}!}. \quad (1)$$

This method infers $y^{(d,l)}_{\boldsymbol{c},\tau}$ from $\boldsymbol{c}$, and τ and its relationship can be inferred by using generalized linear regression. The explanatory variable is decoupled into two parts: the time factor $\boldsymbol{\phi}(\tau) \in \mathbb{R}^S$ and the rest of the factors $\boldsymbol{\varphi}(\boldsymbol{c}) \in \mathbb{R}^M$.

As for the representation of the time factor, the feature $\boldsymbol{\phi}(\tau) \in \mathbb{R}^S$ can be thought of as a smoothed variant of the one-hot encoding on the time index τ : $\{\boldsymbol{\phi}(\tau)\}_t = \mathcal{N}(\tau | t, \sigma^2)$, where $\mathcal{N}(\cdot)$ indicates Gaussian distribution. Let $t \in \{1, \ldots, S\}$ be the means and $\sigma > 0$ be the standard deviation. Note that σ serves as a smoothing term for proper population prediction, and is adjusted by empirical evaluation. As for the representation of the rest of the features, $\boldsymbol{\varphi}(\boldsymbol{d}) \in \mathbb{R}^M$ is a kind of one-hot feature encoding handling the weather conditions, the days of the week, and whether or not it is a holiday.

Given that the linear Poisson regression cannot handle the peak shift with respect to the changes in condition $\boldsymbol{c}$ (e.g., commuter congestion is always found on weekdays, but it is rarely found on a Sunday), a bilinear feature representation is employed as one of the simplest ways of handling this issue.

In bilinear representation, the weight parameter $\boldsymbol{W}^{(l)} \in \mathbb{R}^{M \times S}$ is used for inferring the rate parameter $\lambda^{(l)}_{\boldsymbol{c},\tau} \in \mathbb{R}$ in Poisson distribution as

$$\ln \lambda^{(l)}_{\boldsymbol{c},\tau} = \boldsymbol{\varphi}(\boldsymbol{c})^{\mathsf{T}} \boldsymbol{W}^{(l)} \boldsymbol{\phi}_\tau(\tau). \quad (2)$$

This model could be simply optimized via MAP inference. Note that this is independently executed in each mesh in the previous approach.

Problem with large scale finer mesh analytics

In the previous model, the model on each mesh l is treated independently from the other meshes $l' = 1, \ldots, l-1, l+1, \ldots, L$. In other words, many parameter optimizations are handled when the number of regions L is increased or the mesh size is set to be fine (e.g., 100 m $\times$ 100 m square). When each model is learned independently, MSL parameters are required to represent their model. To avoid the overfitting issue raised by the large number of parameters in the bilinear Poisson regression, the previous work employed a low rank approximation of $\boldsymbol{W}$; however, the number of parameters is still very large at $K(M+S)L$, where K is a rank of $\boldsymbol{W}^{(l)}$.

When analytics with large scale urban data and many fine meshes are needed the parameter of one mesh gets close to that of the neighboring meshes. If mesh l becomes neighbors with another mesh l', it is preferable to plug in the assumption $\boldsymbol{W}^{(l)} \approx \boldsymbol{W}^{(l')}$ to avoid overfitting issues. However, the previous work cannot handle this issue.

Another drawback of the bilinear Poisson regression approach is that the model does not leverage common shared population patterns found in the other areas. Given that the previous work describes the discovery of basic population patterns in large-scale population analytics (Fan, Song, and Shibasaki 2014) (e.g., the population pattern in business areas and that in residential areas are quite distinctive), it is preferable to employ the property into the model. Using tensor factorization (Fan, Song, and Shibasaki 2014) and hierarchical Bayesian models (Shimosaka et al. 2016) for finding basic latent population patterns across regions is quite common. However, given that no spatial preservation is plugged in their modeling, their performance drastically worsens when they use a much finer mesh size for large-scale analyses.

In this section, we formalize the model proposed herein. The model heavily relies on the bilinear Poisson regression, but superior to the previous ones in terms of the large-scale/finer mesh analysis. Our model leverages the basic notation described in the previous section and assumes that the active population $y^{(d,l)}_{\boldsymbol{c},\tau}$ depends on the mesh l, time index τ, and other certain explanatory variables $\boldsymbol{c}$, and is drawn from Poisson distribution. In contrast to the previous approach, we pursue both accuracy and robustness in finding the rate parameter $\lambda^{(l)}_{\boldsymbol{c},\tau}$ in finer mesh areas. The model shares the basic population pattern across the regions (e.g., patterns in commercial areas are similar to those in other regions of commercial areas even if they are far away from each other) to ensure that the inference is robust and accurate. In con-

trast, the basic patterns derived from our model should be similar when two meshes are close together.

We focus on the success of the factorization approach for personalized modeling in recommendation systems (Xu, Zhou, and Tan 2015; Yang, Zhao, and Gao 2017) for finding basic pattern structures across regions to employ both of the abovementioned properties for robustness. Moreover, we consider the spatiality preservation of densely arranged and fine-grained POIs via a graph regularization-based multi-task learning approach (Widmer et al. 2012).

Pattern factorization via parameter space

In this section, we assume that the daily active population pattern contains a variety of latent active population patterns across meshes, and that these patterns are shared across meshes. We leverage the basic idea of factorized regression, which has been actively explored in recommendation systems, to formalize this (Yang, Zhao, and Gao 2017). The active population $y_{\boldsymbol{c},\tau}^{(d,l)}$ at the l-th mesh at the τ-th time index with the given explanatory variable $\boldsymbol{c}$ is assumed to be drawn from Poisson distribution, $y_{\boldsymbol{c},\tau}^{(d,l)} \sim \mathcal{P}(\cdot|\lambda_{\boldsymbol{c},\tau}^{(l)})$. We also assume the log of the rate parameter $\ln \lambda_{\boldsymbol{c},\tau}^{(l)}$ as in (2). Although the previous approach independently optimized the rate parameter of each mesh, our model adds the following assumption that the weight parameter in each mesh can be factored as the weighted sum of the basic patterns. $\boldsymbol{W}^{(l)} \in \mathbb{R}^{M\times S}$ can be concretely defined as the linear weighted sum of the $B \ll L$ base weight matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_B$: $\boldsymbol{W}^{(l)} = \sum_{b=1}^{B} z_{l,b}\boldsymbol{Q}_b$, where the weight corresponding to the l-th mesh can be defined as vector $\boldsymbol{z}^{(l)} = (z_{l,1}, \ldots, z_{l,b})^\mathsf{T} \in \mathbb{R}^B$.

We also employ an additional constraint on $\boldsymbol{z}$ as simplex to find a simple interpretation of the resultant model, where $\boldsymbol{z}^{(l)}$ must be $\sum_b z_{l,b} = 1, z_{l,b} \geq 0$. The final form of the weight matrix $\boldsymbol{W}^{(l)}$ can be thought of as a mixture of basic patterns because of this constraint.

The main merit of leveraging the factored approach is that the number of parameters is drastically reduced when compared with the modern models. The previous model of the bilinear Poisson regression requires $LK(M + S)$ parameters because they independently optimize the weight parameter in their analytics. If we employ $M = 64$ explanatory variables and $S = 48$ (used in a previous paper (Shimosaka et al. 2015)) to analyze the $L = 1000 \times 1000$ regions, 560 M parameters are needed, where $K = 5$ is the rank of the weight parameter using low-rank approximation.

In contrast to the previous model, our approach only requires $LB + BMS = B(MS + L)$. Note that $B \ll L$. The total number used in the model is limited to 15 M if we set $B = 15$. This prevents critical overfitting issues even if the low-rank approximation of each weight parameter matrix is not employed. The low rank approximation could be employed in our model; however, we omit this issue to simplify the implementation.

Spatiality preservation via graph proximity matrix

The factored representation brings us the benefit of parameter shrinkage, and is expected to prevent overfitting issues; however, the model could not preserve the spatial relationship across regions because of the factorization approach. In a finer mesh analysis, this sometimes results in a non-smoothed population prediction result. Given that the population number is proportional to the area of the meshes, the issue of parameter sensitivity is raised in a finer mesh analysis.

We overcome this issue by leveraging another perspective of multi-task learning, called graph regularization (Widmer et al. 2012). The model relies on the graph theory, where each node depicts a single mesh, and each edge represents the relationship between two meshes. Thanks to the property of graph regularization, the parameter matrix gets close when two meshes are close together. We simplify the model by employing this idea into the learning phase of $\boldsymbol{Z}$ as a regularization term. We define the following weighted sum of the differentiation of two weight vectors, namely $\boldsymbol{z}^{(l)}$ and $\boldsymbol{z}^{(l')}$, as:

$$\Omega_{\mathrm{G}}(\boldsymbol{Z}) = \frac{1}{2}\zeta_{\mathrm{G}} \sum_{l,l'} \|\boldsymbol{z}^{(l)} - \boldsymbol{z}^{(l')}\|_2^2 A_{l,l'}, \tag{3}$$

where $A_{l,l'}$ denotes an adjacency matrix reflecting the similarity between l and l'. In each $\boldsymbol{A}$ element, we use a Gaussian kernel with a domain-specific distant metric between two meshes: $A_{l,l'} = \exp\left(-\eta \mathrm{dist}^2(l,l')\right)$, where $\eta > 0$, and $\zeta_{\mathrm{G}} > 0$ are the hyper parameters fixed via an empirical performance evaluation. This regularization term is smooth and convex with respect to the latent variable $\boldsymbol{Z}$. Thus, we can simply add this term to the parameter learning process.

As the distance metric $\mathrm{dist}(\cdot,\cdot)$ used in our model, the simple Euclidean distance metric can be employed for its simplicity. However, note that a custom distance metric can be assigned using the knowledge derived from geographical information, such as railway, stations and residential areas. In our experiment, we verify the customized distance derived from the attributes of the areas. In the customized version of the distance metric $\mathrm{dist}_g(l,l') = \mu_0^{[g_l = g_{l'}]}\mathrm{dist}_e(l,l')$, where g_l represents an attribute of the l-th mesh; $\mathrm{dist}_e(l,l')$ depicts the Euclidean distance between the l-th and l'-th meshes; $0 < \mu_0 < 1$ denotes some constant variables; and $[\cdot]$ represents a bracket returning 1 if the argument is true; otherwise, returning 0.

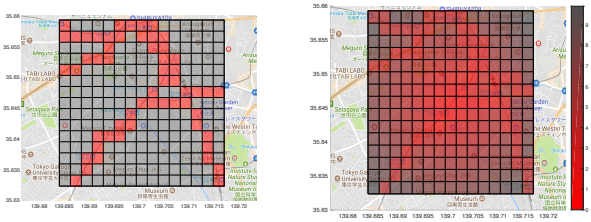

Figure 1: Left: mesh area containing the railway line (red). Right: distance metric from the center mesh considering the railway line. The longer distance gets darkened.

Parameter learning

This section explains the parameter learning process of the proposed model. Let us assume that we have a GPS log dataset ranging from $l = 1, \ldots, L$-th meshes for $d = 1, \ldots, D$-th days. Similar to the previous model, MAP inference is used as regularized training to obtain the optimal parameters from the dataset. We formalize the training process of $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_B, \boldsymbol{Z}$ as the following optimization problem:

$$\begin{aligned} \arg\min_{\boldsymbol{Q}_{1:B}, \boldsymbol{Z}} \quad & -\sum_{d,l} \ln \mathcal{P}(y^{(d,l)}_{\boldsymbol{c},\tau} | \lambda^{(l)}_{\boldsymbol{c},\tau}) \\ & + \sum_b \Omega_Q(\boldsymbol{Q}_b) + \Omega_{\mathrm{G}}(\boldsymbol{Z}), \\ \text{subject to} \quad & |\boldsymbol{z}^{(l)}|_1 = 1, z_{l,b} \geq 0, \quad (l = 1, \ldots, L), \end{aligned} \tag{4}$$

where $\Omega_Q(\boldsymbol{Q}_b) = \xi_Q \|\boldsymbol{Q}_b\|^2_{\mathrm{Fro}}$ is the regularization term for the base latent weight parameter matrices. $\|\cdot\|_{\mathrm{Fro}}$ is known to be a Frobenius norm on matrix.

This optimization is not bi-convex over $\boldsymbol{Q}_{1:B}$ and $\boldsymbol{Z}$, but is convex over $\boldsymbol{Q}_{1:B}$ with the given $\boldsymbol{Z}$ and over $\boldsymbol{Z}$ with the given $\boldsymbol{Q}_{1:B}$. We employ an alternating optimization approach frequently used in non-negative matrix factorization or related recommendation systems (Yang, Zhao, and Gao 2017).

For optimizing $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_B$, we leverage quasi-Newton optimization (Wright and Nocedal 1999) to obtain better $\boldsymbol{Q}_{1:B}$ owing to the smoothness of the objective function with respect to $\boldsymbol{Q}_{1:B}$. We use the projected steepest gradient technique inspired by a similar technique (Lin 2007) to obtain better $\boldsymbol{Z}$ because the domain of $\boldsymbol{Z}$ is simplex. Note that the projection onto the simplex region could be analytically and efficiently solved (Duchi et al. 2008).

As for the implementation, we use the MapReduce framework implemented in Apache Spark for our empirical evaluation because of its efficiency in parallelization. Apache Spark is known to be the next generation of Apache Hadoop. In this framework, we aggregate gradient information with respect to $\boldsymbol{z}^{(l)}$ in each region l and $\boldsymbol{Q}_b$ using a similar word-counting technique in MapReduce.

Experimental results

We conducted the experiment using a population scale mobile phone location dataset with the two settings described below to validate the superiority of our model over the other previous methods.

First, we conducted the experiment to compare the performance of our model with variants of bilinear Poisson models. Note that the results for the mesh-type dataset is not provided in the literature; therefore, we carefully chose a wider area of regions as the whole of areas and divide it into subsequent meshes as a target to be analyzed (Fig. 2). We chose subsequent square areas of 100 m $\times$ 100 m, which are used as meshes with 3 km $\times$ 3 km regions that cover the center of Tokyo region. We adjusted the size of meshes ranging from 100 m $\times$ 100 m, 200 m $\times$ 200 m, 600 m $\times$ 600 m, and 1 km $\times$ 1 km within the fixed 3 km $\times$ 3 km regions, then obtained the predictive performance and the computational cost in comparison with the state-of-the-art techniques.

As for the second experiment, a qualitative evaluation on active population pattern discovery and latent structures

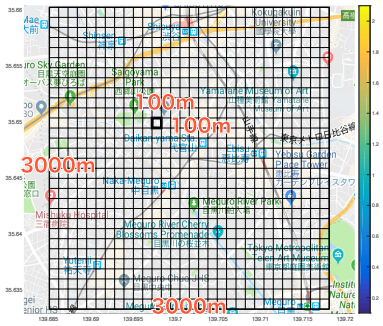

Figure 2: Finer mesh areas used in the 2nd and 3rd evaluations

across the meshes in geographical maps through the visualization of latent variables $\boldsymbol{z}$, $\boldsymbol{Q}$ is described by comparing the effectiveness of the spatial preservation term. In this evaluation, we verified the effects, thanks to the spatiality preservation, by showing the visualization result using the same dataset used in the first quantitative evaluation.

Dataset

We used anonymized large-scale GPS log records collected from the mobile phone application released by commercial companies[1]. Each record has three components: timestamps, latitude, and longitude. We used the data collected over 365 days (from July 1, 2013 to June 30, 2014), which consist of 15 million records per day in the Kanto region of Japan.

As mentioned earlier, the entire areas used in these evaluations are 3 km $\times$ 3 km through three evaluations. In this setting, the model is assumed to predict active population counts $S = 48$ times per day in each mesh. In the first evaluation, we evaluated the performance in $L = 9, 25, 225$, and 900 meshes.

Evaluation Criteria

As for the evaluation, we employed the Mean Negative Log Likelihood (MNLL) as a performance indicator: $\mathrm{MNLL} = -\frac{1}{DT}\sum_{d=1}^{D}\sum_{\tau=1}^{S} \ln p(y^{(l,d)}_{\boldsymbol{c},\tau} | \lambda^{(l)}_{\boldsymbol{c},\tau})$. In this sense, the smaller MNLL indicates the better prediction performance. We also obtained the computational cost of the model training in each condition.

Feature description on Poisson regressors

This section describes the process of encoding the explanatory variable $\boldsymbol{c}$ into $\boldsymbol{\varphi}(\boldsymbol{c})$ of our model. As an external factor, we leverage days of week and holidays in this experiment. The day of week feature can be written as seven distinctive discrete values, whereas the holiday can be defined as Boolean. Let $\boldsymbol{\varphi}_1(\boldsymbol{c}) \in \mathbb{R}^7$ be a one-hot vector of the days of week and $\boldsymbol{\varphi}_2(\boldsymbol{c}) \in \mathbb{R}^2$ be that of holidays. In that manner, the total feature $\boldsymbol{\varphi}(\boldsymbol{c})$ can be written as:

[1]The reference of the dataset is not disclosed for the submission phase because of the double blind review policy; however, it will be explicitly disclosed in the camera-ready version.

$\boldsymbol{\varphi}(\boldsymbol{c}) = \boldsymbol{\varphi}_1(\boldsymbol{c}) \otimes \boldsymbol{\varphi}_2(\boldsymbol{c})$, where the dimension of the resultant feature is 14.

Comparison scheme

To show the validity of our model, we evaluated the performance of the variants of the bilinear Poisson regression models, given that this model is a base of our model. In contrast to the factorization models, bilinear models do not consider the basic population patterns shared across the meshes, but use an explanatory variable $\boldsymbol{c}$ in their prediction stage.

1. **BP 1 for All**: This setting produces a single model with a single parameter (i.e., $\boldsymbol{W}^{(1)} = \ldots, = \boldsymbol{W}^{(L)}$). The model always outputs the same number, regardless of the mesh ID l. The model has a strong bias, but has a small variance, thanks to the small number of parameters. We employed the same feature $\boldsymbol{\varphi}(\boldsymbol{c})$ in this model.
2. **BP 1 for 1**: This setting produces the L models of the bilinear Poisson regression, where the parameter optimization is independently executed across meshes. The total number of parameters to be learned is LSM. This also uses the same feature $\boldsymbol{\varphi}(\boldsymbol{c})$ as the proposed model. Note that this model is quite equivalent to the state-of-the-art model presented in (Shimosaka et al. 2015).
3. **BP with index**: This model produces a single (low rank) bilinear Poisson model, where the mesh ID is directly encoded as $\boldsymbol{c}$, which is in contrast to our model. This work is thought of as a combination of tensor factorization and discriminative models (Tomioka and Suzuki 2013), except that we did not install the nuclear norm regularization for our experiment. The feature $\boldsymbol{\varphi}(\boldsymbol{c})$ in this comparative model is written as follows: $\boldsymbol{\varphi}(\boldsymbol{c}) = \boldsymbol{\varphi}_1(\boldsymbol{c}) \otimes \boldsymbol{\varphi}_2(\boldsymbol{c}) \otimes \boldsymbol{\varphi}_3(\boldsymbol{c})$, where $\boldsymbol{\varphi}_3(\boldsymbol{c}) \in \mathbb{R}^L$ indicates the mesh index as one-hot encoding. With this formulation, the low rank approximation brings an effect of factorization of urban dynamics across meshes even if we did not use nuclear norm minimization.
4. **SPF**($\zeta_{\mathrm{G}} = 0$): This model produces factorized regression, where no spatiality preservation is installed (i.e., this model is equivalent to the proposed model with $\zeta_{\mathrm{G}} = 0$).

Predictive performance in various mesh sizes

The performance of each method was evaluated through the training data with 30 days and the testing data with 180 days to understand the robustness under the severe condition in the setting with the fine-grained POI. The numbers of days for the training and testing datasets were determined by preliminary experiments. The prediction accuracy of the prediction methods was saturated when we used over 30 days of data. As for the testing dataset, the number of days for the testing was determined long enough for better evaluations.

The performance was obtained from a five-fold cross-validation. We also compared this with the performance of the proposed method. In this evaluation, we controlled the size of meshes/the number of the POIs L in the fixed size of regions 3 km $\times$ 3 km in total, then observed the prediction performance of each method and the training cost.

Table 1: MNLL of various mesh sizes

Model	Mesh size			
	100 m	200 m	600 m	1 km
BP 1 for All	1.52 ± 0.15	3.08 ± 0.44	11.2 ± 3.38	23.7 ± 9.22
BP 1 for 1	1.60 ± 0.19	3.07 ± 0.39	9.91 ± 3.21	21.41 ± 8.87
BP with index	1.48 ± 0.14	2.90 ± 0.41	10.2 ± 3.15	22.11 ± 8.69
SPF($\zeta_{\mathrm{G}} = 0$)	1.47 ± 0.15	2.86 ± 0.42	10.1 ± 3.01	21.94 ± 9.14
SPF (proposed)	$\mathbf{1.46 \pm 0.14}$	$\mathbf{2.83 \pm 0.38}$	$\mathbf{9.89 \pm 3.23}$	$\mathbf{21.27 \pm 8.98}$

We employed 100 m ($L = 900$), 200 m ($L = 225$), 600 m ($L = 25$), and 1000 m ($L = 9$) as the POI sizes.

From the table, the experimental results indicated that our method achieved the best performance among the other bilinear Poisson models. BP 1 for 1 obtained a severe performance drawback because of the large number of parameters in spite of the limited number of training dataset. Needless to say, the performance of BP 1 for all did not achieve the best performance because of the variety of population patterns in each mesh. Additionally, the BP with the index behaved as a kind of factorization model, thanks to the low rank approximation; however, this model became overfitted in the setting with the fine-grained POI.

Note that we showed the means and the standard deviations of the MNLL calculated from five means in the five-fold cross-validation. We also confirmed the significant difference between our proposed model and the comparison methods by the p-value ($p < 10^{-3}$) in each iteration calculated from a large number of samples (180 (days) $\times$ 48 (time bins) $\times L$) for the evaluation in each validation. Thus, the performance of the model was clearly better than that of the other models in terms of the MNLL. These results indicated that the spatiality preservation promotes the stability under the limited number of training data, while factorization (in BP with the index and SPF) is essential in achieving a better performance compared with the individual training process.

For the model accuracy, we also calculated the computational cost of our approach and that of the comparison methods. Fig. 3 shows the increase of the computation time according to the number of meshes. The computational time of BP 1 for 1 was $\mathcal{O}(L)$ for the L meshes because it was trained in each city separately, and the total number of parameters also increased by $\mathcal{O}(L)$, as mentioned earlier. Thanks to the reduction of parameters, the increases of the computational time of the proposed model were much smaller than those of BP 1 for 1.

Visualization of the latent parameters

One of the main advantages of the proposed method is providing the functionality to capture the relationship of mesh areas with spatiality preservation and be useful for forecasting the population. In this evaluation, we confirmed the effectiveness of the proposed method by showing the visualization result, where the relationship among the meshes was captured. We also showed the performance of the factorized bilinear Poisson regression without spatiality preservation to ensure the proximity property.

As for the visualization, our model consisted of several parameters with a basic pattern shared across meshes:

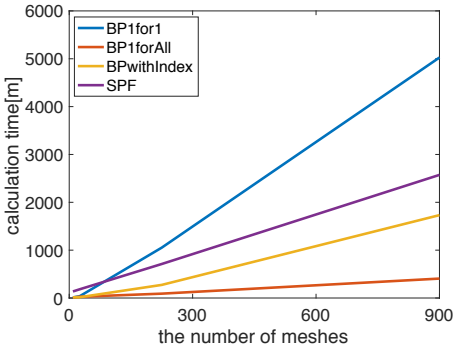

Figure 3: Computation time for training with respect to the number of meshes

$\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_B$, and latent weight vectors $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ that were specific to the l-th mesh. The latter parameters $\boldsymbol{z}^{(l)}$ can be used to understand the similarity between two meshes.

In this experiment, we visualized the relationship of meshes by projecting learned parameters $z_{1,b}, \ldots, z_{L,b}$ in B-times and B basic population patterns governed by $\boldsymbol{c}, \tau$. In our experiment, we used 30 days of data for training as in the previous experiment. The visualization result on $\boldsymbol{Q}_b$ was normalized by scale to intuitively visualize the activity pattern (i.e., the total counts of data per mesh per day were assumed to be 5000).

Fig. 4 represents the $b = 6$-th and $b = 8$-th patterns from the $B = 12$ patterns on $\zeta_{\mathrm{G}} = 0.01$. The $b = 8$ pattern clearly corresponds to the areas related to railway transportation, whereas the $b = 6$ patterns correspond to the activity on residential areas.

Fig. 5 shows the obtained $\boldsymbol{Q}_4$ and $\boldsymbol{Q}_8$ of $\boldsymbol{Q}_{1:12}$ in the second experiment. $\boldsymbol{Q}_8$ also clearly reflects the pattern at the railway/station areas, where two peaks of congestion can be found on weekdays, but no peaks were found on holiday weekends. $\boldsymbol{Q}_4$ can be inferred as population patterns in residential areas. In the future analysis, we will apply *SPFR* to large-scale finer meshes spreading over the nation(s).

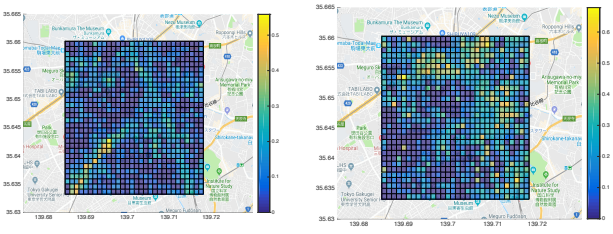

Figure 4: $z_4^{(l)}$ and $z_8^{(l)}$ in each mesh l on $\zeta_{\mathrm{G}} = 0.01$, respectively.

In addition to $\zeta_{\mathrm{G}} = 0.01$, we also tried to obtain another result on $\zeta_{\mathrm{G}} = 0$ to confirm the effect of the spatial preservation. Fig. 6 and Fig. 7 show the visualization results on $\zeta_{\mathrm{G}} = 0$. This result indicated that the larger size of ζ_{G} provides smoothed changes of latent variable $\boldsymbol{z}^{(l)}$ across meshes. This result implied that the usage of graph regularization improves the interpretability of the population pattern analysis as well as avoidance of overfitting issues.

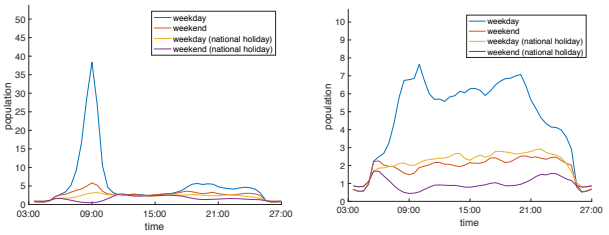

Figure 5: Population patterns $\boldsymbol{Q}_4$ and $\boldsymbol{Q}_8$ on $\zeta_{\mathrm{G}} = 0.01$ obtained in parameter learning in the second experiment

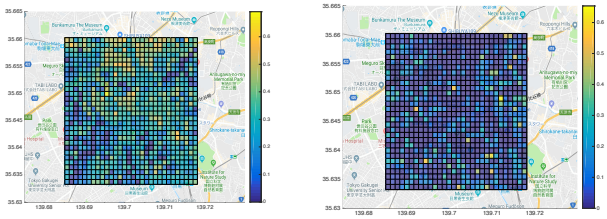

Figure 6: $z_2^{(l)}$ and $z_{11}^{(l)}$ in each mesh l on $\zeta_{\mathrm{G}} = 0$

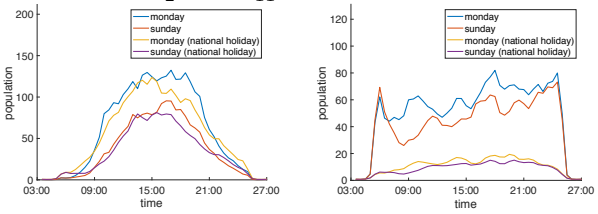

Figure 7: Population patterns $\boldsymbol{Q}_2$ and $\boldsymbol{Q}_{11}$ on $\zeta_{\mathrm{G}} = 0$ obtained in parameter learning in the second experiment

Conclusion

This study presented population pattern modeling using large amounts of GPS data derived from crowd smartphones for urban computing. Our model focuses on the analysis of finer mesh regions densely distributed over large-scale areas of interest. To tackle this issue, we proposed a new framework of urban dynamics analytics, called *Spatiality Preservable Factorized Regression (SPFR)*. Factored modeling enabled us to reduce the number of parameters using basic factored population patterns shared across regions, and a graph regularization term helped the model to preserve spatiality with reasonable optimization. In our empirical evaluation, which used a large dataset of over 32 M GPS logs in the Tokyo region, our model is shown to be superior to the state-of-the-art predictive models.

Future research could address a much larger scale analysis (e.g., nationwide) with the finer mesh analysis and irregularity detection using our model.

Acknowledgement

This work is partially supported by CREST, JST.

References

An, G.; Liu, S.; and Ruan, Q. 2017. A sparse neighborhood preserving non-negative tensor factorization algorithm for facial expression recognition. *Pattern Analysis and Applications* 20(2).

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *Proc. of IJCAI*.

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; and Chandra, T. 2008. Efficient projections onto the l1-ball for learning in high dimensions. In *Proc. of ICML*.

Fan, Z.; Song, X.; and Shibasaki, R. 2014. CitySpectrum: A non-negative tensor factorization approach. In *Proc. of UbiComp*.

Jiang, R.; Song, X.; Fan, Z.; Xia, T.; Chen, Q.; Chen, Q.; and Shibasaki, R. 2018. Deep ROI-based modeling for urban human mobility prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2(1).

Konishi, T.; Maruyama, M.; Tsubouchi, K.; and Shimosaka, M. 2016. CityProphet: City-scale irregularity prediction using transit app logs. In *Proc. of UbiComp*.

Lichman, M., and Smyth, P. 2018. Prediction of sparse user-item consumption rates with Zero-Inflated Poisson regression. In *Proc. of WWW*.

Lika, B.; Kolomvatsos, K.; and Hadjiefthymiades, S. 2014. Cold start problem,recommender systems. *Expert Systems with Applications*.

Lin, C.-J. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation*.

Nishi, K.; Tsubouchi, K.; and Shimosaka, M. 2014. Extracting land-use patterns using location data from smartphones. In *Proc. of the First Intl. Conf. on IoT in Urban Space*.

Okawa, M.; Kim, H.; and Toda, H. 2017. Online traffic flow prediction using convolved bilinear Poisson regression. In *Proc. of MDM*.

Reades, J.; Calabrese, F.; and Ratti, C. 2009. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*.

Sekimoto, Y.; Shibasaki, R.; Kanasugi, H.; Usui, T.; and Shimazaki, Y. 2011. Pflow: Reconstructing people flow recycling large-scale social survey data. *IEEE Pervasive Computing* 10(4).

Shimosaka, M.; Maeda, K.; Tsukiji, T.; and Tsubouchi, K. 2015. Forecasting urban dynamics with mobility logs by bilinear poisson regression. In *Proc. of UbiComp*.

Shimosaka, M.; Tsukiji, T.; Tominaga, S.; and Tsubouchi, K. 2016. Coupled hierarchical Dirichlet process mixtures for simultaneous clustering and topic modeling. In *Proc. of ECML-PKDD*.

Takeuchi, K.; Tomioka, R.; Ishiguro, K.; Kimura, A.; and Sawada, H. 2013. Non-negative multiple tensor factorization. In *Proc. of ICDM*.

Tomioka, R., and Suzuki, T. 2013. Convex tensor decomposition via structured schatten norm regularization. In *Advances in NIPS*.

Widmer, C.; Kloft, M.; Gornitz, N.; and Ratsch, G. 2012. Efficient training of graph-regularized multitask SVMs. In *Proc. of ECML-PKDD*.

Wright, S., and Nocedal, J. 1999. *Numerical optimization*.

Xu, F.; Feng, J.; Zhang, P.; and Li, Y. 2016. Context-aware real-time population estimation for metropolis. In *Proc. of UbiComp*.

Xu, J.; Zhou, J.; and Tan, P.-N. 2015. FORMULA: FactORized MUlti-task leArning for task discovery in personalized medical models. In *Proc. of SIAM SDM*.

Yang, P.; Zhao, P.; and Gao, X. 2017. Robust online multitask learning with correlative and personalized structures. *IEEE Trans. on Knowledge and Data Engineering* 29(11).

Yao, Z.; Fu, Y.; Liu, B.; Hu, W.; and Xiong, H. 2017. Representing urban functions through zone embedding with human mobility patterns. In *Proc. of IJCAI*.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and pois. In *Proc. of KDD*.

Zhang, F.; Yuan, N. J.; Wilkie, D.; Zheng, Y.; and Xie, X. 2015. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility. In *ACM Trans. on Intelligent Systems and Technology*.

Zhang, J.; Zheng, Y.; and Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proc. of AAAI*.

Zheng, Y.; Liu, T.; Wang, Y.; Zhu, Y.; Liu, Y.; and Chang, E. 2014. Diagnosing new york city's noises with ubiquitous data. In *Proc. of UbiComp*.