Statistics and Its Interface Volume 12 (2019) 331–344

A case study for Beijing point of interest data using group linked Cox process

Yu Chen, Rui Pan[*,†], Rong Guan[‡], and Hansheng Wang[§]

We develop in this article a group linked Cox process model for analyzing point of interest (POI) data. We focus on a Beijing POI dataset, which contains more than 22 thousand POIs in Beijing urban area. These POIs have been divided into many small categories (e.g., restaurants, movie theaters, hospitals, universities and subway stations) by the digital map maker (e.g., *Baidu Map*). Empirical analysis provides substantial evidence that POIs across different categories could be highly correlated so that those small categories can be further grouped. To this end, we develop here a group linked Cox process model. Specifically, within each group, we model POI locations by a standard Cox process so that the POI clustering effect can be well described. Furthermore, the idea of bivariate linked Cox process is borrowed and further extended to its multivariate counterpart. Consequently, a more significant number of POI categories can be accommodated within each group. To estimate the model, a minimum contrast type method is developed, and an automatically grouping method is provided. Simulation studies are conducted to validate the proposed methodology. At last, we apply our method to the aforementioned real dataset, and a total of 4 groups are uncovered. This leads to the discovery of some urban-planning-related features.

Keywords and phrases: Cox Process, Group Linked Cox Process, Location Based Service, Point of Interest.

1. INTRODUCTION

Figure 1. A map with different categories of POIs. Mapmakers mark different POIs with different icons (e.g., forks and knives for restaurants) on the map, so that users can quickly find their target places.

A point of interest (i.e., POI) is a specific point location that someone may find useful or interesting, such as hospitals, restaurants, tourist attractions, universities, and many others; see Figure 1 for a quick understanding. POIs are usually generated by digital map makers like *Google Map* and *Baidu Map*. In the meanwhile, users can upload their customized POIs through mobile apps, such as *Facebook* and *Sina Weibo*. A POI specifies, at the minimum, the latitude and longitude of the point location. Names and brief descriptions (e.g., category) for POIs are usually available, and other information such as tags and telephone numbers may also be attached. POI data are one of the most fundamental component for location-based service (LBS), such as restaurant recommendations of *Yelp*.

POI data have received diverse research interests recently, and its applications range broadly across many fields. In the field of urban planning, [21] and [19] use POI data to discover regions of different functions in a city, such as residential and high-tech areas. This provides people with a quick understanding of a sophisticated city. POI data are also powerful in studying human activities. For example, [16], [11] and [3] investigate human mobility using POIs. They combine GPS trace information and POI data to uncover human daily patterns and transitions between different activities. Other than that, POI data are of great use in the study of environmental problems. [23] and [9] use POI data to help infer the urban air quality. The basic idea is that some categories of POIs may have a relationship with air quality, namely, chemical factories and power stations. Furthermore, POI recommendation is a new topic in the mobile Internet era, where restaurants or tourists attractions are recommended to each user based on his/her historical visitations; see [22], [8], [7] and [20]. POI recommendation

[*] Corresponding author.

[†] The research of Rui Pan is supported by National Nature Science Foundation of China (NSFC, 11601539, 11631003, 71771224), the Fundamental Research Funds for the Central Universities (QL18010), and China's National Key Research Special Program Grant 2016YFC0207704.

[‡] The research of Rong Guan is supported in part by National Natural Science Foundation of China (NSFC, 71401192, 71873012, 71771204), the Fundamental Research Funds for the Central Universities (QL18009), and the Program for Innovation Research in Central University of Finance and Economics (011650317002).

[§] Hansheng Wang's research is partially supported by National Natural Science Foundation of China (No. 11831008, 11525101, 71532001, 71332006). It is also supported in part by China's National Key Research Special Program (No. 2016YFC0207704).

is widely used in many LBS applications like *Foursquare*, *Gowalla*, and *Facebook Places*. To summarize, POI data are exploited by many scholars and proved to be useful in many fields.

Typical POI data exhibit several unique characteristics. First, POIs are equipped with latitude and longitude information, so we model them as spatial point processes. Second, POIs are tagged with categories, such as restaurants, hospitals, and many others. In our Beijing POI dataset, more than 22 thousand POIs are assigned to 30 different categories. Third, within the same category, POIs typically exhibit a spatial clustering pattern. For example, restaurants are likely to be clustered together. Fourth, different POI categories are often highly correlated. For example, bus stops (one POI category) are usually close to residential areas (another POI category). There are two challenges to analyze POI data, The first is that the number of categories is relatively large, and there are complex spatial and cross-category correlations. Second, the number of POIs is enormous, which could lead to massive computational cost.

To handle these challenges, we propose here a group linked Cox process (GLCP). GLCP is a generalization of the linked Cox process (LCP) proposed by [5]. The Cox process is a commonly used spatial point process allowing for the clustering effect, and its bivariate extension linked Cox process further provides dependence between component processes. In LCP, the first component follows a univariate Cox process, and the intensity of the second component is the product of a constant and the same realization as the first one. LCP leads to co-located concentrations of both type events, and it is widely used in zoology and botany studies when the presence of one species is of benefit to another [14].

The newly proposed GLCP is a multivariate extension of LCP but with a group structure. For instance in the POI dataset, GLCP assumes that all POI categories can be divided into several *groups*. Within each group, all POI categories are "linked", like LCP. To estimate the model, a minimum contrast type method [4] is developed, and an automatically grouping method is proposed. Simulation studies are conducted to demonstrate the finite sample performance of the methodology. We then apply our method to the aforementioned real dataset, and a total of 4 category groups are uncovered. This leads to the discovery of numerous urban-planning-related features.

In the literature, many scholars have studied the multivariate spatial point process and the cross-process correlation. [15] propose summary statistics to quantify the association between two processes, and we focus on how to simplify the relationship between multiple processes. [12] use group lasso techniques to detect significant interactions among component process, and we try to simplify the interactions through a clustering approach. [17] assume that the component processes are generated by linear combinations of q independent latent random fields. In contrast, GLCP assumes each component process belongs to one latent random field (i.e., we assume group structure) and allow the latent random fields to be dependent. Furthermore, [17] uses cross-validation to determine q and GLCP provides a path of grouping via a hierarchical clustering type method.

We focus on the Beijing POI data and organize the rest of this article as follows. We first introduce the Beijing POI dataset in Section 2. In Section 3 we describe how to develop the GLCP model using the spatial point process technologies. Its statistical properties and estimation methods are discussed. Some simulation studies are conducted in Section 4. Finally, we provide a detailed POI data analysis using the GLCP model in Section 5.

2. THE POI DATASET

2.1 Data description

All data used in this study are obtained from *Sina Weibo*. *Sina Weibo* is a *Twitter* type social media website in Chinese. It maintains a huge POI database. When a user posts a tweet, it will automatically tag the tweet with a POI. If the tagged POI is not satisfactory, the user is allowed to create its own customized POI and report it to *Sina Weibo*, which enables *Sina Weibo* to update its POI database regularly. Also, data are obtained from *Sina Weibo*'s API. API is the abbreviation of Application Programming Interface, which is a set of clearly defined methods for communication between various software components. *Sina Weibo*'s open API makes this POI database publicly available.

Our dataset contains 22,691 POIs in Beijing urban area. The latitude of these POIs ranges from 39°48′N to 40°05′N, and the longitude ranges from 116°18′E to 116°58′E. The whole region is about 667 square kilometers, and the S50 Road in Beijing encircles it. The S50 Road is also called the 5th Ring Road, and it is a highway encircling Beijing. It is about 10 kilometers away from the city center; see Figure 2 for the outermost circle.

The dataset contains the following essential information. For each POI, the latitude, longitude and a category tag are available. The latitude and longitude are with GCJ-02

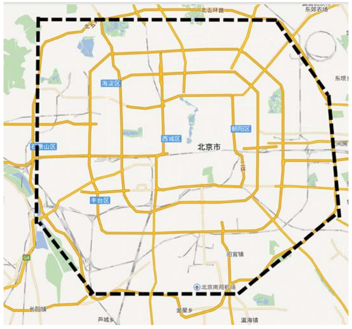

Figure 2. Ring roads in Beijing and the outermost circle marked by the dotted line is S50 Road. The whole region inside the S50 Road is about 667 square kilometers.

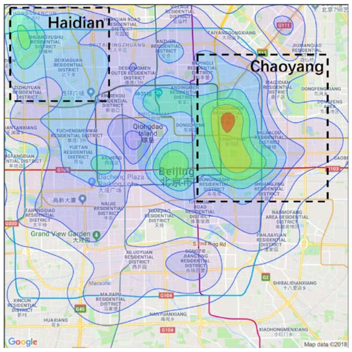

Figure 3. The heat map of POIs in Beijing. The locations of Haidian and Chaoyang District are marked by a black dotted lines.

coordinates, which is formulated by the Chinese State Bureau of Surveying and Mapping (*http://www.sbsm.gov.cn/*). Both latitude and longitude are accurate to 5 decimal places. Holding latitude at 40°N, every 1×10^{-5} difference in longitude is approximately equivalent to 0.8 meters on earth surface. Those POIs have been classified into 30 different categories by the map vendor, and the categories include, for instance, primary schools, banks, companies, and hospitals. See Table 4 for a detailed list together with their frequencies and relative proportion.

2.2 Clustering pattern and cross categories dependence

For a quick understanding, we approximately treat the urban area within Beijing 5th Road as a rectangle. We then split it into a total of $40 \times 40 = 1600$ grids, and plot a heat map based on kernel density estimation of the number of occurrences per grid. As shown in Figure 3, some POI clusters (regions with the bright color) can be easily detected. The largest cluster lies in Chaoyang District in Eastern Beijing. Chaoyang District serves as the Central Business District (CBD) where many international companies are headquartered. The second largest center is in the middle of Beijing, with attractions such as Tiananmen Square. The third largest cluster lies in Haidian District in Northwestern Beijing, where many top Chinese universities and IT companies are located in this area.

In Figure 4, heat maps for $2 \times 3 = 6$ POI categories are provided. It can be found that the distribution patterns of the top three categories are highly correlated since they have similar clustering centers and spreading trends. The bottom three also have similar patterns. We call this phenomenon POI categories grouping effect.

Urban functional zoning might cause the POI categories grouping effect. In many cities, land use is divided by its function [6]. A city may have several functional zones, for example, an industrial zone, a recreational zone, and a residential zone, and the POI categories are different in different functional zones. In order to analyze the POI data, we propose a new point process model based on the above characteristics of POI data, and exploit the grouping effect to simplify the cross categories dependence. In the next chapter, we formally describe this model.

3. THE GROUP LINKED COX PROCESS

3.1 Preliminaries

As shown in the previous subsection, strong clustering and grouping effects are detected in our real data. Thus, it is of interest to develop a statistical model, which can naturally accommodate those interesting patterns. To this end, we develop here a novel model, called group linked Cox process.

We use the spatial point process to represent the POI points on the map. Mathematically, let X be a spatial point process in $\mathbb{R}^2$, where the origin is denoted by O, $\mathcal{B}^2$ be the Borel σ-algebra on $\mathbb{R}^2$, and ν be the Lebesgue measure. For any set $A \in \mathcal{B}^2$, let $X(A)$ be the number of points located in A. In this paper, we assume all point processes are simple, that is $\forall \mathbf{s} = (s_1, s_2) \in \mathbb{R}^2, X(\{\mathbf{s}\}) \in \{0, 1\}$. In other words, we assume that there is at most one POI on the same latitude and longitude. Denote $\lambda(\cdot)$ and $\lambda^{(2)}(\cdot, \cdot)$ to be the first and second order intensity functions.

Let $\{\Psi(\mathbf{s}), \mathbf{s} \in \mathbb{R}^2\}$ be a non-negative random field. A point process X is Cox process [1] directed by Ψ, if conditionally on $\Psi = \psi$, X is an Poisson process with intensity function ψ. Linked Cox process [5] is a bivariate Cox process with linked intensity functions, satisfying $\psi_2 = \omega\psi_1$, where ψ_1 and ψ_2 are intensity functions corresponding to two different component processes, and ω is a positive constant.

3.2 The GLCP model

We next introduce the newly proposed GLCP model for POI data. Let N be the total number of POI categories and M be the number of groups, where $M \ll N$. Let $Z = (Z_1, \cdots, Z_M)^\top \in \mathbb{R}^M$ denote a M-dimensional log-Gaussian random field. Accordingly, for any $\mathbf{s} \in \mathbb{R}^2$, denote the realization of Z at location $\mathbf{s}$ as $Z(\mathbf{s}) = (Z_1(\mathbf{s}), \cdots, Z_M(\mathbf{s}))^\top \in \mathbb{R}^M$. Let $\mu(\mathbf{s}) = (\mu_1, \cdots, \mu_M)^\top = E(Z(\mathbf{s})) \in \mathbb{R}^M$, and $\Sigma = (\sigma_{k_1k_2}) = \text{cov}(Z(\mathbf{s})) \in \mathbb{R}^{M \times M}$. For any $1 \le k_1, k_2 \le M$, we assume $C_{k_1,k_2}(\boldsymbol{h}) = \mathbb{E}(Z_{k_1}(\mathbf{s}) - \mu_{k_1})(Z_{k_2}(\mathbf{s} + \boldsymbol{h}) - \mu_{k_2}) = \rho(\boldsymbol{h})\sigma_{k_1k_2}$, where $\rho(\boldsymbol{h}) = \exp\{-\beta\|\boldsymbol{h}\|\}$ is the spatial covariance function. As one can see, Z is a second-order stationary log-Gaussian process. In addition to that, its cross covariance function is separable [10].

Next, conditional on Z, we generate a total of N spatial point processes (denoted by $X_1, \cdots, X_N$) as follows. Let $\mathcal{J}(\cdot)$ be a mapping from $\{1, \cdots, N\}$ to $\{1, \cdots, M\}$. Simply speaking, $\mathcal{J}(\cdot)$ classifies each component point process (e.g., POI of hospitals) to its corresponding group (e.g., residential functional zone). Conditional on Z, we model a component point process X_i as a Poisson process with conditional intensity function given by

$$\lambda_i(\mathbf{s}) = \omega_i \exp\{Z_{\mathcal{J}(i)}(\mathbf{s})\}, \tag{1}$$

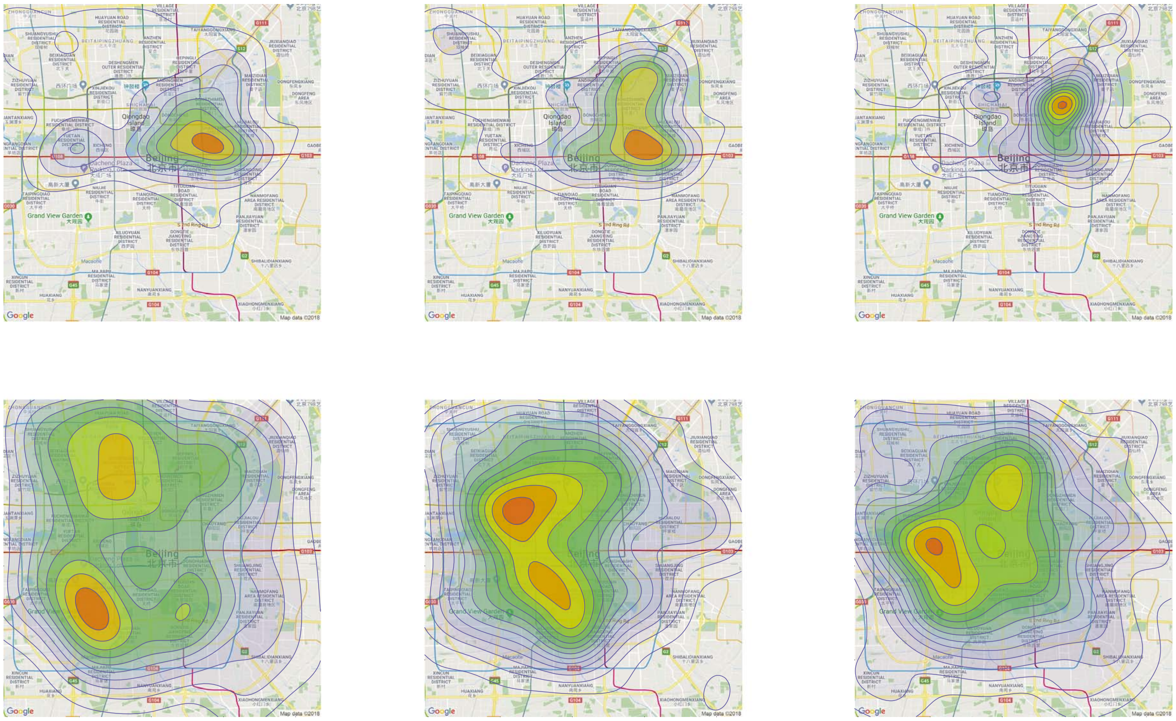

Figure 4. Heat maps of Cantonese cuisine (top left), Japanese cuisine (top middle), western food (top right), kindergarten (bottom left), primary school (bottom middle), middle school (bottom right).

where ω_i is a positive scalar. For convenience define $\mathcal{G}(m) = \{i : \mathcal{J}(i) = m\}$ for $m \in \{1, \cdots, M\}$. Simply speaking, $\mathcal{G}(m)$ collects the indices of all the POI categories within group m. For identification purpose, we require that $\max_{i\in\mathcal{G}(m)} \omega_i = 1$ for any $1 \le m \le M$. As one can see, this model is an extension of the classical bivariate linked Cox process model.

Take POI data as an example. We consider 30 POI categories as 30 different spatial point processes, which are induced by M random fields, where M is to be estimated. In the GLCP model, we link each process to exactly one random field, i.e., we divide the 30 POI categories into different groups, but neither $\mathcal{J}(\cdot)$ nor $\mathcal{G}(\cdot)$ can be observed. The grouping scheme needs to be inferred from the data. In order to automate grouping POI categories, we exploit the following two propositions of GLCP. The first proposition gives the relationship between the log-Gaussian random field and the intensities of the point process.

Proposition 1. *For each i, X_i is stationary with first order intensity function $\lambda_i = \omega_i \exp\{\mu_{\mathcal{J}(i)} + \sigma^2_{\mathcal{J}(i)}/2\}$ and second order intensity function $\lambda_i^{(2)}(\mathbf{s}_1, \mathbf{s}_2) = \omega_i^2 \exp\{2\mu_{\mathcal{J}(i)} + \sigma^2_{\mathcal{J}(i)}\} \exp\{\sigma^2_{\mathcal{J}(i)} \exp(-\beta\|\mathbf{s}_1 - \mathbf{s}_2\|)\}$.*

By Proposition 1, the first and second order intensity of GLCP can be explicitly calculated. We find that larger μ_m and σ_m values lead to larger values for both the first and second order intensities because the mean of the latent log-Gaussian process becomes larger. However, larger β value leads to smaller second order intensity, since the spatial dependence of the latent log-Gaussian process becomes weak. We next investigate the relationships among different processes using cross K-functions [13].

Proposition 2. *We consider two different scenarios.*
Scenario 1. *If the two spatial point processes (indexed by i and j) come from the same group, that is $i, j \in \mathcal{G}(m)$ for some $1 \le m \le M$, we then have*

$$K_{ii}(r) = K_{ij}(r) = K_{ji}(r) = K_{jj}(r) = \int_{B(O,r)} \exp\{\sigma_{kk} \exp\{-\beta\|\mathbf{s}\|\}\}\, d\mathbf{s};$$

Scenario 2. *If the two spatial processes are from two different groups, that is $i \in \mathcal{G}(k_1)$ and $j \in \mathcal{G}(k_2)$ with $k_1 \neq k_2$, we then have*

$$K_{ij}(r) = \int_{B(O,r)} \exp\{\sigma_{k_1k_2} \exp\{-\beta\|\mathbf{s}\|\}\}\, d\mathbf{s}.$$

By Proposition 2, we know that the cross K-functions from the categories belonging to the same group should be identical. This is not true if the categories are from different groups. As we shall demonstrate later, this insightful finding leads to an interesting and effective automatic grouping method.

3.3 Estimation method

In this subsection, we propose a novel algorithm to automatically group POI processes, which is based on the minimum contrast estimation method and hierarchical clustering. We first illustrate the estimation method for the GLCP model, and then we describe the proposed grouping algorithm.

3.3.1 Minimum contrast estimation

Recall that we have a total number of N categories, which are indexed by $1 \le i \le N$. We also assume that those categories can be assigned to different groups by an unobservable $\mathcal{J}(\cdot)$. In this subsection, we assume $\mathcal{J}$ is known and discuss how to estimate $\mathcal{J}$ in the next subsection. We shall develop a minimum contrast type method for estimating μ, Σ, β, and ω_i for every $1 \le i \le N$. We estimate K-function $K_{ii}(r)$ by $\hat{K}_{ii}(r) = \hat{\lambda}_i^{-1} \sum_{\mathbf{s}_1 \in X_i} \sum_{\mathbf{s}_2 \in X_i} I(0 < \|\mathbf{s}_1 - \mathbf{s}_2\| \le r)/X_i(D)$, where $\hat{\lambda}_i = X_i(D)/\nu(D)$. For the cross K-function, we have $\hat{K}_{ij}(r) = (\hat{\lambda}_i \hat{\lambda}_j)^{-1} \sum_{\mathbf{s}_1 \in X_i} \sum_{\mathbf{s}_2 \in X_j} I(0 < \|\mathbf{s}_1 - \mathbf{s}_2\| \le r)/\nu(D)$.

The theoretical solution $K_{ij}(r)$ is given in Proposition 2. We can then compare $K_{ij}(r)$ against its empirical counterpart $\hat{K}_{ij}(r)$ for arbitrary $1 \le i, j \le N$. Specifically, a contrast-type loss function can be constructed for each (i, j) pair as

$$\mathcal{L}_{i,j} = \int_{r_1}^{r_2} \left(K_{ij}^q(r) - \hat{K}_{ij}^q(r) \right)^p dr, \tag{2}$$

where $0 < r_1 < r_2 < \infty$ are two pre-specified constants. Here p and q are two tuning parameters, and $(p, q) = (2, 1/4)$ has been suggested by [18]. Then, by summing over all possible (i, j) pairs, a combined loss function can be obtained as

$$\mathcal{L} = \mathcal{L}(\Sigma, \beta) = \sum_{1 \le k_1, k_2 \le M} \sum_{\substack{i \in \mathcal{G}(k_1) \\ j \in \mathcal{G}(k_2)}} \mathcal{L}_{i,j}. \tag{3}$$

As one can see, $\mathcal{L}$ is a function of Σ and β. Accordingly, they can be estimate by $(\hat{\Sigma}, \hat{\beta}) = \text{argmin}\mathcal{L}(\Sigma, \beta)$, and the elements of $\hat{\Sigma}$ is denoted by $\hat{\sigma}_{ij}, 1 \le i, j \le M$.

Next, we estimate the first order intensity λ_i by $\hat{\lambda}_i = X_i(\mathcal{D})/\nu(\mathcal{D})$, where $\mathcal{D} \subset \mathbb{R}^2$ is a pre-specified bounded region. Then, by Proposition 1, we know that $\mu_m = \log(\lambda_i) - \log(\omega_i) - \sigma_{mm}^2/2$ for every $i \in \mathcal{G}(m)$. By identification assumption, we know that $\max_{i \in \mathcal{G}(m)} \omega_i = 1$. Let i^* be the index associated with the maximum ω_i-value. We then have $\omega_{i^*} = 1$ and $\lambda_{i^*} = \max_{i \in \mathcal{G}(m)} \lambda_i$. We thus further have $\mu_m = \log(\lambda_{i^*}) - \sigma_{mm}^2/2 = \log(\max_{i \in \mathcal{G}(m)} \lambda_i) - \sigma_{mm}^2/2$. Accordingly, we can estimate $\hat{\mu}_m = \log(\max_{i \in \mathcal{G}(m)} \hat{\lambda}_i) - \hat{\sigma}_{mm}^2/2$. Lastly, we can estimate ω_i by $\hat{\omega}_i = \hat{\lambda}_i / \max_{j \in \mathcal{G}(\mathcal{J}(j))} \hat{\lambda}_j$.

3.3.2 The grouping algorithm

Our numerical studies confirm that the estimation method proposed in the previous subsection works quite well. However, it requires an important condition, that is all the categories have been correctly assigned to their corresponding groups. For real data analysis, this is typically unknown. Not only the group mapping function $\mathcal{J}(\cdot)$, but also the number of groups M. Determining the number of groups and how the POIs are grouped becomes an essential issue. To this end, we propose a novel grouping algorithm as follows.

Initialization. The algorithm starts with the most complicated model, where each category belongs to a different group. In that case, the model reaches its greatest flexibility and thus the smallest loss value for $\mathcal{L}$ in (3), which is recorded by $\mathcal{L}^{(0)}$. In this case, we define the mapping function $\mathcal{J}^{(0)}(i) = i$, which is a function from $\{1, \cdots, N\}$ to $G^{(0)} = \{1, \cdots, N\}$.

Hierarchical Merging. After k steps $(0 \le k \le N - 2)$, we assume the mapping function has been updated to be $\mathcal{J}^{(k)}$, which is a function from $\{1, \cdots, N\}$ to $G^{(k)} = \{1, \cdots, N - k\}$, similarly, define the corresponding $\mathcal{G}^{(k)}(m) = \{i : \mathcal{J}^{(k)}(i) = m\}$. In the $(k+1)$-th step, we update $\mathcal{J}^{(k)}$ to $\mathcal{J}^{(k+1)}$, which should be a function defined from $\{1, \cdots, N\}$ to $G^{(k+1)} = \{1, \cdots, N - k - 1\}$. To this end, two groups in $G^{(k)}$ must be merged into a new one. According to which two groups in $G^{(k)}$ to be grouped, the resulting mapping function should be different. For example, let $m_1 < m_2$ be two arbitrary group ID from $G^{(k)}$, if the processes in $\mathcal{G}^{(k)}(m_1)$ and $\mathcal{G}^{(k)}(m_2)$ were merged into a new group in this step, the new mapping function $\mathcal{J}_{m_1 m_2}^{(k+1)}$ would be

$$\mathcal{J}_{m_1 m_2}^{(k+1)}(i) = \begin{cases} \mathcal{J}^{(k)}(i) & \text{if } \mathcal{J}^{(k)}(i) < m_2 \\ m_1 & \text{if } \mathcal{J}^{(k)}(i) = m_2 \\ \mathcal{J}^{(k)}(i) - 1 & \text{if } \mathcal{J}^{(k)}(i) > m_2 \end{cases}.$$

We can then re-evaluate the loss function $\mathcal{L}$ in (3) according to different mapping function. The resulting losses are recorded by $\mathcal{L}^{m_1 m_2}$. Let $(m_1^*, m_2^*) = \text{argmin}_{m_1, m_2} \mathcal{L}^{m_1 m_2}$, $\mathcal{J}^{(k+1)} = \mathcal{J}_{m_1^* m_2^*}^{(k+1)}$ and $\mathcal{L}^{(k+1)} = \mathcal{L}^{m_1^* m_2^*}$. As a result, processes in $\mathcal{G}^{(k)}(m_1^*)$ and $\mathcal{G}^{(k)}(m_2^*)$ are grouped together in the $(k+1)$-th step.

Solution Path. Repeat the hierarchical merging step for $N - 1$ times, which leads to a total of $N - 1$ nested candidate mapping functions (i.e. grouping methods). We then collect those mapping functions by a solution path $\mathbb{S} = \{\mathcal{J}^{(k)} : 1 \le k \le N - 1\}$. To determine the number of groups, define

$$\tau_k = \frac{\mathcal{L}^{(k+1)} - \mathcal{L}^{(k)}}{\mathcal{L}^{(k)} - \mathcal{L}^{(k-1)}}, \tag{4}$$

and let $k_{\max} = \text{argmax}_k \tau_k$. We then estimate M by $\widehat{M} = N - k_{\max}$ and the grouping result is given by $\mathcal{J}^{(\widehat{M})}$.

Below we make some comments on the computational complexity of GLCP. The computational consumption of the GLCP algorithm mainly has the following two parts. The first part is the estimation of the empirical K function required by the minimum contrast method. For N component processes, we need to calculate N^2 times cross K functions, and the computational cost is $O(|\mathcal{D}|N^2)$. The second part is the hierarchical grouping procedure. The algorithm iterates N steps to merge all processes into a single group. In each step t, we need to find the smallest $\mathcal{L}^{m_1 m_2}$, which is an N^2 summation according to (3). So in each step, there are $N^2 \times N^2$ inner loops, and the total computational complexity of GLCP is $O(|\mathcal{D}|N^2 + N^5)$. It is worth mentioning that the GLCP algorithm is suitable for parallel computing. For example, when selecting the optimal (m_1^*, m_2^*), the amount of calculation can be distributed on multiple computing devices.

4. SIMULATION STUDIES

4.1 Model setup

To demonstrate the finite sample performance of the proposed estimation method, we report a number of simulation studies in this section.

The model is simulated on a bounded domain $\mathcal{D} = [0, d] \times [0, d]$, where $d > 0$ determines the domain area (i.e., $|\mathcal{D}| = d^2$). Intuitively, the larger the domain area, the more spatial points are likely to be observed. We consider different (M, N) combinations, where M ranges from 1 to 4 and N ranges from 1 to 12.

For a given M, we set $\beta = 12$, the mean parameter μ of each random field is independently sampled from $\text{Uniform}(5, 6)$, and the covariance matrix is designed to be $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{M \times M}$ with $\sigma_{ij} = \rho^{|i-j|}$, and we try different ρ in $\{0.2, 0.5, 0.8\}$ for low, medium and high correlations. Accordingly, the Gaussian random field $\mathbb{Z}$ can be generated on $\mathcal{D}$, where the size of $\mathcal{D}$ is selected in $\{1, 4\}$. The increase in the size of $\mathcal{D}$ indicates more points, and the performance of the model should be better.

Next, for a given N, we construct a group mapping function from $\{1, \cdots, N\}$ to $\{1, \cdots, M\}$ as $\mathcal{J}(i) = i \bmod M + 1$, where mod is the modulo operation that finds the remainder after division of one number by another. Once $\mathcal{J}(\cdot)$ is given, $\mathcal{G}(\cdot)$ can be defined accordingly. The link parameter ω_i is independently sampled from $\text{Uniform}(0.5, 1)$. For every $1 \le m \le M$, define $i(m) = \text{argmax}_{i \in \mathcal{G}(m)} \omega_i$. We next re-define $\omega_{i(m)} = 1$ so that the identification condition $\max_{i \in \mathcal{G}(m)} \omega_i = 1$ for every $1 \le m \le M$ can be satisfied. Thereafter, the intensity function $\lambda_i(\cdot)$ can be calculated for each X_i according to (1), and the GLCP process X_is can be generated based on the intensity $\lambda_i(\cdot)$. To implement the minimum contrast estimation method in (2), we fix $(r_1, r_2) = (0.01, 0.15)$ and $(p, q) = (2, 1/4)$; see [18].

In order to study the effect of ω on the estimation results and grouping results of the GLCP, we added a set of comparative experiments. In this set of experiments, the size of $\mathcal{D}$ is 4, $\rho = 0.5$, $\beta = 4$, and μ is sampled from $\text{Uniform}(5, 6)$. A total of 12 processes is divided into 4 groups, within group m, $\max_{i \in \mathcal{G}(m)} \omega_i$ is set to be 1 according to the model setup, and the other ω is sampled from $\text{Uniform}(\bar{\omega} - 0.05, \bar{\omega} + 0.05)$. We set $\bar{\omega}$ to be 0.4 and 0.8 to study the influence of ω.

4.2 Performance measurements

For a given parameter setup (i.e., M, N, $|\mathcal{D}|$, β, μ, Σ, and ω), the GLCP data generate process is randomly simulated for $\mathcal{T} = 100$ times and are indexed by $1 \le t \le \mathcal{T}$. We then estimate M by $\widehat{M}$ according to (4), and denote the resulting estimator for the t-th simulation replication as $\widehat{M}^{(t)}$. The following empirical probabilities are then computed. They are, respectively, the under-fitting probability $\text{UFP} = \mathcal{T}^{-1} \sum_t I(\widehat{M}^{(t)} < M)$, the correct-fitting probability $\text{CFP} = \mathcal{T}^{-1} \sum_t I(\widehat{M}^{(t)} = M)$, and the over-fitting probability $\text{OFP} = \mathcal{T}^{-1} \sum_t I(\widehat{M}^{(t)} > M)$. Even if the number of groups is correctly estimated (i.e, $\widehat{M} = M$), whether different spatial point processes are grouped correctly is not clear. In the most ideal situation, those processes belong to the same group should be grouped together empirically. Let $\mathcal{J}^{(t)}$ be the true group mapping function generated by the tth simulation replication, and $\widehat{\mathcal{J}}^{(t)}$ be the resulting estimate. We then compute the true positive grouping probability as

$$\text{TPGP}^{(t)} = \frac{\sum_{i,j} I\Big(\mathcal{J}^{(t)}(i) = \mathcal{J}^{(t)}(j)\Big) I\Big(\widehat{\mathcal{J}}^{(t)}(i) = \widehat{\mathcal{J}}^{(t)}(j)\Big)}{\sum_{i,j} I\Big(\mathcal{J}^{(t)}(i) = \mathcal{J}^{(t)}(j)\Big)},$$

and the false negative grouping probability as

$$\text{TNGP}^{(t)} = \frac{\sum_{i,j} I\Big(\mathcal{J}^{(t)}(i) \neq \mathcal{J}^{(t)}(j)\Big) I\Big(\widehat{\mathcal{J}}^{(t)}(i) \neq \widehat{\mathcal{J}}^{(t)}(j)\Big)}{\sum_{i,j} I\Big(\mathcal{J}^{(t)}(i) \neq \mathcal{J}^{(t)}(j)\Big)}.$$

Lastly, define the overall true positive grouping probability $\text{TPGP} = \mathcal{T}^{-1} \sum_t \text{TPGP}^{(t)}$, and the overall false negative grouping probability $\text{TNGP} = \mathcal{T}^{-1} \sum_t \text{TNGP}^{(t)}$. To verify the computational complexity of GLCP, we recorded the time of each simulation as $\text{TIME}^{(t)}$, and report the average time consuming $\text{TIME} = \mathcal{T}^{-1} \sum_t \text{TIME}^{(t)}$.

Next we demonstrate the performance measurements for the minimum contrast estimation method. We assume that both M and the grouping function $\mathcal{J}$ is given already, so that we can be purely focus on parameter estimation for $\beta \in \mathbb{R}$, $\mu = (\mu_m) \in \mathbb{R}^M$, $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{M \times M}$, and $\omega = (\omega_i) \in \mathbb{R}^N$. Similar to the previous subsection, we use superscript (t) to index the parameter estimates obtained in the tth simulation replication. This leads to the following notations as $\hat{\beta}^{(t)} \in \mathbb{R}$, $\hat{\mu}^{(t)} = (\hat{\mu}_m^{(t)}) \in \mathbb{R}^M$, $\hat{\Sigma}^{(t)} = (\hat{\sigma}_{ij}^{(t)}) \in \mathbb{R}^{M \times M}$, and $\hat{\omega}^{(t)} = (\hat{\omega}_i^{(t)}) \in \mathbb{R}^N$. We then compute the root mean square error (RMSE) for different parameter estimates. Take $\hat{\beta}$ as an example, the RMSE is

Table 1. The empirical probabilities of GLCP group mapping algorithm

M	ρ	$\mathcal{D}$	N	$E(X)$	UFP	CFP	OFP	TPGP	TNGP	TIME
2	0.2	1	4	1453.40	0.00	0.90	0.10	0.97	1.00	6.6
			6	2135.85	0.00	0.91	0.09	0.97	1.00	11.4
		4	4	6176.98	0.00	0.96	0.04	0.99	1.00	5.0
			6	8307.16	0.00	1.00	0.00	1.00	1.00	12.8
	0.5	1	4	1442.03	0.00	0.83	0.17	0.96	1.00	6.0
			6	2063.96	0.00	0.91	0.09	0.97	0.99	12.7
		4	4	5703.52	0.00	0.93	0.07	0.98	1.00	4.9
			6	8538.60	0.00	0.99	0.01	1.00	1.00	12.6
	0.8	1	4	1480.21	0.00	0.70	0.30	0.87	0.92	6.1
			6	2160.31	0.00	0.75	0.25	0.88	0.93	14.3
		4	4	5901.62	0.00	0.87	0.13	0.96	0.99	4.9
			6	8597.55	0.00	0.88	0.12	0.97	1.00	11.1
3	0.2	1	6	2178.67	0.02	0.84	0.14	0.97	0.99	15.7
			9	3136.73	0.03	0.89	0.08	0.97	0.99	56.5
		4	6	8610.80	0.00	0.99	0.01	1.00	1.00	13.4
			9	12477.01	0.00	0.99	0.01	1.00	1.00	54.9
	0.5	1	6	2213.13	0.12	0.74	0.14	0.96	0.96	11.6
			9	3181.36	0.07	0.83	0.10	0.97	0.98	47.0
		4	6	8573.71	0.00	0.97	0.03	0.99	1.00	11.9
			9	12510.61	0.00	0.98	0.02	1.00	1.00	47.3
	0.8	1	6	2252.82	0.36	0.34	0.30	0.89	0.86	11.4
			9	3134.13	0.38	0.43	0.19	0.90	0.84	44.6
		4	6	9124.29	0.04	0.86	0.10	0.97	0.99	13.0
			9	12333.79	0.08	0.83	0.09	0.98	0.97	45.8
4	0.2	1	8	2883.83	0.06	0.82	0.12	0.97	0.98	41.8
			12	4175.91	0.02	0.88	0.10	0.97	0.99	185.0
		4	8	11845.20	0.00	0.98	0.02	1.00	1.00	42.3
			12	16754.99	0.00	0.95	0.05	0.99	1.00	181.6
	0.5	1	8	2866.12	0.20	0.66	0.14	0.96	0.95	34.0
			12	4229.51	0.13	0.75	0.12	0.97	0.97	144.2
		4	8	11886.91	0.01	0.98	0.01	1.00	1.00	32.6
			12	16972.23	0.00	0.98	0.02	1.00	1.00	143.2
	0.8	1	8	2982.89	0.68	0.16	0.16	0.93	0.78	29.9
			12	4368.95	0.64	0.24	0.12	0.92	0.81	130.6
		4	8	11589.62	0.16	0.74	0.10	0.98	0.95	31.1
			12	16961.68	0.11	0.83	0.06	0.99	0.97	136.4

given by $\{\mathcal{T}^{-1}\sum_{t=1}^{\mathcal{T}}(\hat{\beta}^{(t)}-\beta)^2\}^{1/2}$, the RMSE for $\hat{\mu}$ is given by $\{\mathcal{T}^{-1}M^{-1}\sum_{t=1}^{\mathcal{T}}\sum_{m=1}^{M}(\hat{\mu}_m^{(t)}-\mu_m)^2\}^{1/2}$, $\hat{\sigma}$, and $\hat{\omega}$ can be defined similarly.

4.3 Simulation results

We demonstrate the simulation results in this subsection. Proceeding the grouping algorithm for $\mathcal{T} = 100$ times, we obtain Table 1. As one can see, as $|\mathcal{D}|$ increases, CFP, TPGP and TNGP increase, this implies the group mapping estimation can achieve higher accuracy if more data is provided. For a fixed region of $\mathcal{D}$, a larger number of processes N indicates more possible grouping methods, and consequently the lower TPGP. For larger ρ, GLCP trends to under-fit M, because the more significant the correlation, the harder it is to distinguish the latent random fields. The calculation time TIME is measured with a single core of the E5-2680 v2 CPU.

The detailed results of parameters estimation are given in Table 2. As one can see, as $|D|$ increases, the RMSE of all parameters decrease. The RMSE of ω is 0 when $N = M$, this is because the largest ω in each group is set to be 1. It is worth mentioning that for fixed M and $|\mathcal{D}|$, as N increases, the RMSE slightly increase. This is reasonable because, in GLCP, processes in the same group are driven by the same realization of the underlying random field, so in this case, the increase in data is not as informative as increasing the domain size $|\mathcal{D}|$.

From proposition 1 we conclude that the parameters ω, μ, β and Σ together determine the intensity functions of Cox process. Furthermore, the first order intensity function grows linearly with ω. This means that given the other parameters, the larger ω means more data points, which may lead to more accurate estimates. The results in Table 3 also confirm this point.

Table 2. The RMSE for the minimum contrast estimation of GLCP model

M	ρ	$\lvert\mathcal{D}\rvert$	N	$E(X)$	RMSE(β)	RMSE(μ)	RMSE(Σ)	RMSE(ω)
1	0.2	1	1	364.50	5.03	0.21	0.26	0.00
			2	594.00	4.78	0.22	0.29	0.04
			3	969.50	4.71	0.20	0.24	0.05
		4	1	1724.00	1.92	0.12	0.20	0.00
			2	2771.00	1.99	0.13	0.21	0.02
			3	3990.50	2.00	0.15	0.26	0.02
	0.5	1	1	391.50	4.53	0.26	0.30	0.00
			2	648.50	4.61	0.21	0.22	0.04
			3	961.50	5.20	0.21	0.27	0.05
		4	1	1573.50	1.84	0.13	0.21	0.00
			2	2650.00	1.99	0.14	0.20	0.02
			3	4227.00	1.81	0.15	0.21	0.03
	0.8	1	1	367.00	5.21	0.25	0.25	0.00
			2	672.50	6.29	0.19	0.25	0.04
			3	914.50	5.27	0.22	0.22	0.04
		4	1	1565.50	1.89	0.15	0.26	0.00
			2	2753.00	1.89	0.14	0.22	0.02
			3	4302.50	1.94	0.13	0.23	0.02
4	0.2	1	4	1662.00	2.96	0.23	0.22	0.00
			8	2941.50	2.99	0.23	0.20	0.04
			12	4126.50	3.45	0.21	0.21	0.05
		4	4	6490.50	1.17	0.14	0.15	0.00
			8	11898.00	1.15	0.15	0.15	0.02
			12	16583.00	1.13	0.13	0.13	0.02
	0.5	1	4	1659.50	3.28	0.21	0.24	0.00
			8	2783.50	3.15	0.23	0.22	0.04
			12	4175.50	3.13	0.21	0.22	0.05
		4	4	6479.50	1.05	0.15	0.18	0.00
			8	11400.00	1.27	0.14	0.18	0.02
			12	16739.50	1.21	0.15	0.17	0.02
	0.8	1	4	1654.50	3.54	0.21	0.23	0.00
			8	2793.50	3.22	0.22	0.25	0.04
			12	4150.00	3.82	0.21	0.23	0.05
		4	4	6775.00	1.33	0.13	0.19	0.00
			8	11597.50	1.44	0.16	0.21	0.02
			12	16773.00	1.16	0.13	0.17	0.02

Table 3. Simulation results of different ω

E(ω)	E(X)	RMSE(β)	RMSE(μ)	RMSE(Σ)	RMSE(ω)	UFP	CFP	OFP	TPGP	TNGP
0.4	10939	1.230	0.147	0.180	0.013	0.02	0.91	0.07	0.986	0.995
0.8	17454	1.120	0.131	0.163	0.013	0.00	0.99	0.01	0.998	1.00

In practice, the larger the range of data collection, the more likely it is to get accurate results. When analyzing POI data, we recommend taking the city outline as a unit. Because most POIs are highly correlated with human activity, the value of an excessive $\lvert\mathcal{D}\rvert$ will result in a sparse number of points on edge, which will affect the accuracy of the model.

5. APPLY GLCP TO BEIJING POI DATASET

We next provide a detailed analysis for the Beijing POI dataset mentioned above. We focus on the following aspects. First, how different POI categories should be grouped. As we have shown before, some POI categories share very similar spatial distribution characteristics, and thus should be grouped. To this end, the proposed GLCP algorithm is used. Second, we focus on the characteristics of the POI categories within each group. We try to explain the grouping results of the GLCP algorithm from the perspective of urban planning. Third, we examine the relationships between the different groups of POIs and their distributions. By analyzing the Beijing POI dataset using GLCP, we find some interesting phenomenon. Based on these phenomena, one may be able to get some constructive opinions in several fields, including city planning, location choosing.

Table 4. The GLCP grouping result of 30 POI categories in Beijing. There are four POI groups, respectively Commercial, Residential, Diplomatic and High-tech Group. The third column represents the number of occurrences of each POI category, the fourth column is the corresponding proportion, and the fifth column is the estimated link parameter

Group Name	Category Name	Frequency	Proportion	$\hat{\omega}$
High-tech	campus	1252	0.62	1.00(<0.01)
	university	602	0.30	0.48(0.03)
	library	157	0.08	0.13(0.01)
Diplomatic	Western restaurant	781	0.29	1.00(<0.01)
	Cafe	561	0.21	0.72(0.04)
	Cantonese cuisine	554	0.21	0.71(0.04)
	Japanese cuisine	510	0.19	0.65(0.04)
	buffet	242	0.09	0.31(0.02)
Residential	residential area	5533	0.46	1.00(<0.01)
	hospital	1587	0.13	0.29(0.01)
	bus stop	1338	0.11	0.24(0.01)
	hot-pot	1314	0.11	0.24(0.01)
	convenience store	572	0.05	0.10(<0.01)
	middle school	460	0.04	0.08(<0.01)
	Hunan cuisine	357	0.03	0.06(<0.01)
	kindergarten	240	0.02	0.04(<0.01)
	primary school	230	0.02	0.04(<0.01)
	subway	171	0.01	0.03(<0.01)
	seafood restaurant	165	0.01	0.03(<0.01)
	vocational-technical school	87	0.01	0.02(<0.01)
	Hubei cuisine	59	0.00	0.01(<0.01)
Commercial	company	1870	0.31	1.00(<0.01)
	bank	1009	0.17	0.54(0.02)
	Sichuan cuisine	988	0.16	0.53(0.02)
	department store	792	0.13	0.42(0.02)
	mall	631	0.10	0.34(0.02)
	hostel	286	0.05	0.15(0.01)
	bath and massage place	259	0.04	0.14(0.01)
	star hotel	188	0.03	0.10(0.02)
	vacation spot	89	0.01	0.05(0.01)

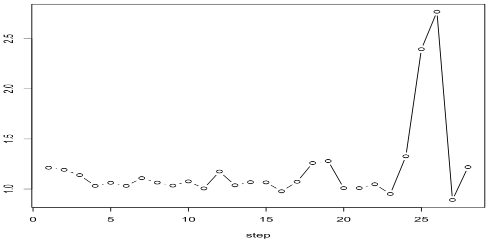

Figure 5. τ_k for Beijing POI dataset, where $\text{argmax}_k \tau_k = 26$, i.e. the algorithm suggests that there should be $30 - 26 = 4$ groups.

5.1 Using GLCP to group Beijing POI categories

In this subsection, we apply the GLCP model to Beijing POI dataset. First we next calculate the τ_k-values according (4), which are then depicted in Figure 5. It suggests that a total of $\widehat{M} = 4$ groups should be formed. Together with the information given in Figure 5, we can classify the 30 POI categories into $\widehat{M} = 4$ groups. Based on the grouping results, we refer to those groups as high-tech, diplomatic, residential and commercial groups, respectively. The detailed results are given in the first columns in Table 4, their distribution patterns are demonstrated in Figure 8.

We next depict the merging path in Figure 6. It shows how the GLCP grouping algorithm combines 30 POI categories into four groups in turn. The four subgraphs represent the four groups obtained. In each subgraph, the letters in the node represent the original POI category, and the number t in the node indicates that its children are merged in step t.

Next, we report the parameter estimation results. At the same time, to evaluate the significance of the coefficients, we apply the parametric bootstrap method [2]. Specifically, we simulate $N = 30$ processes from $M = 4$ groups based on the fitted parameters for $\mathcal{T} = 100$ times. Then we calculate the standard deviation of all coefficients in simulations as the standard error of the real data parameter estimations. The estimated β, μ and Σ is reported in Table 5, and the estimated ω is reported in the fifth column in Table 4. For

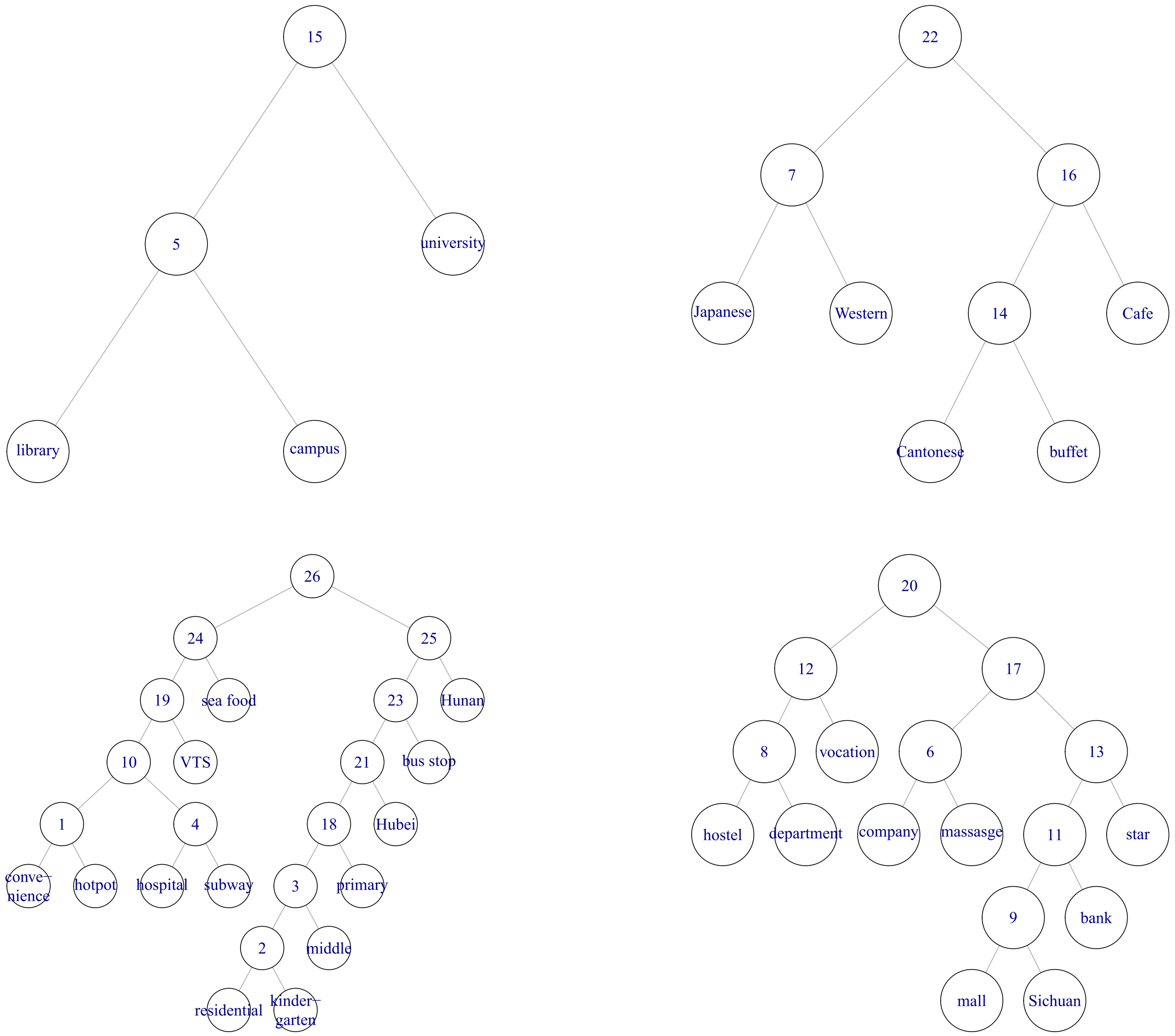

Figure 6. Grouping path of high-tech group (top left), diplomatic group (top right), residential group (bottom left) and commercial group (bottom right). In each binary tree, the letters in the node represent the original component process, and the number t in the node indicates that its children are merged in step t. VTS stands for vocational-technical school, convenience stands for convenience store, residential stands for the residential area, middle stands for middle school, primary stands for primary school, massage stands for bath and massage place, and star stands for star hotel.

Table 5. Estimation result of β, μ and Σ in real data

Group	$\hat{\beta}$	$\hat{\mu}$	$\hat{\Sigma}$			
			1	2	3	4
1	12.38(2.65)	5.72(0.39)	2.82(0.55)	1.47(0.50)	1.14(0.30)	1.27(0.38)
2		5.46(0.35)	1.47(0.50)	2.41(0.50)	1.38(0.30)	1.89(0.40)
3		8.04(0.24)	1.14(0.30)	1.38(0.30)	1.16(0.20)	1.28(0.24)
4		6.73(0.29)	1.27(0.38)	1.89(0.40)	1.28(0.24)	1.60(0.32)

each estimate, the front of the brackets represents real data estimation, and the parentheses are the estimated standard error of the parametric bootstrap method.

We also perform the following sensitivity analysis to evaluate GLCP in this application. For the sake of space, we select 3 categories from each group in the real data analysis, which are listed Table 6. Before applying GLCP, we try to merge the processes i and j into the same group, and then implement the GLCP grouping algorithm for 11 groups. We want to test the stability of the GLCP algorithm with some

Table 6. Selected categories in sensitivity analysis

High-tech (1-3)	Diplomatic (4-6)	Residential (7-9)	Commercial (10-12)
library	Cantonese cuisine	primary school	Sichuan cuisine
university	Japanese cuisine	middle school	company
campus	Western restaurant	hot pot	mall

Table 7. Bias and root mean squared error of parameters in sensitivity analysis

Trial	μ	β	Σ	ω
correct pre-merging	-0.023(0.081)	-0.021(0.074)	0.001(0.003)	0.007(0.025)
incorrect pre-merging	0.185(0.242)	-0.192(0.143)	-0.038(0.038)	0.063(0.040)

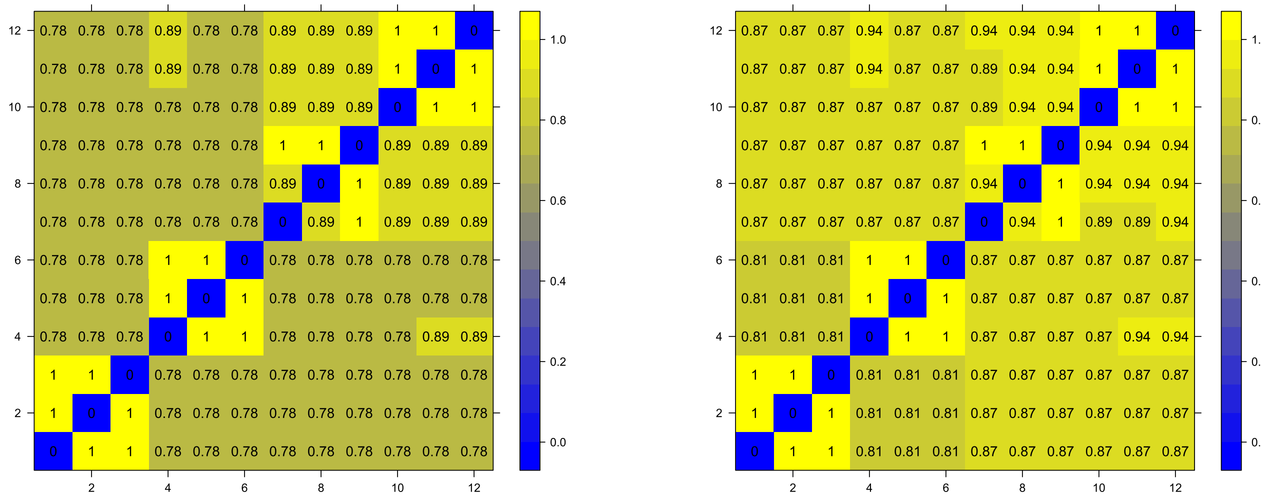

Figure 7. TPGP (left) and TNGP (right) results of the sensitivity analysis.

correct and incorrect pre-merging trials. Formally, let $\hat{\mathcal{J}}(\cdot)$ be the grouping result of real data analysis. For each trial that pre-merges processes i and j, let $\hat{\mathcal{J}}_{ij}(\cdot)$ be the estimated group mapping. The TPGP and TNGP indicators for $\hat{\mathcal{J}}_{ij}(\cdot)$ are shown in Figure 7, and the bias and root mean squared error of parameters in different pre-merging trials are summarized in Table 7. We treat the original estimates of the 12 categories as the ground truth and report average bias and standard deviations of all parameters in the case of correct and incorrect pre-merging trials.

From the results of sensitivity analysis, we find that for correct pre-merging trials, parameter estimations and grouping results are consistent with the ground truth in most cases. For incorrect pre-merging trials, both parameter estimation and grouping results will be affected, but will not lead to systemic disasters. For example, in Figure 7, the minimum value of TNGP and TPGP is 0.78, which indicates that the overall correctness of the 12 categories is still acceptable.

5.2 Model interpretation

We first explain the grouping results of GLCP. According to Figure 5, GLCP divides 30 POI categories into 4 groups. Based on the type of POI in each group, we artificially named them high-tech, diplomatic, residential and commercial group. For example, we found residential areas, schools, and hospitals in the third group, so they were named as the residential group.

The first group is named as high-tech, and it is a simple group contains POIs including university, campus, and library. The high-tech group is relatively independent of other groups, located in the northwest part of the city (the top left subgraph in Figure 8). It is worth noting that the library POIs is also assigned to this group, which means that the reading needs of citizens in other regions may be demanding to meet. The second is the diplomatic group with POIs such as Western restaurant and Cafe. The Diplomatic group POIs are relatively concentrated in Beijing (the top right subgraph in Figure 8), but they still have a strong correlation with the commercial group, which shows that Beijing is becoming an international city.

The residential group is the most complicated one with 13 different POIs and multiple clustering centers. In the group, we saw POI categories that are firmly related to city life, such as housing, education, medical care, transportation, and diet. In the bottom left subgraph of Figure 8, we found

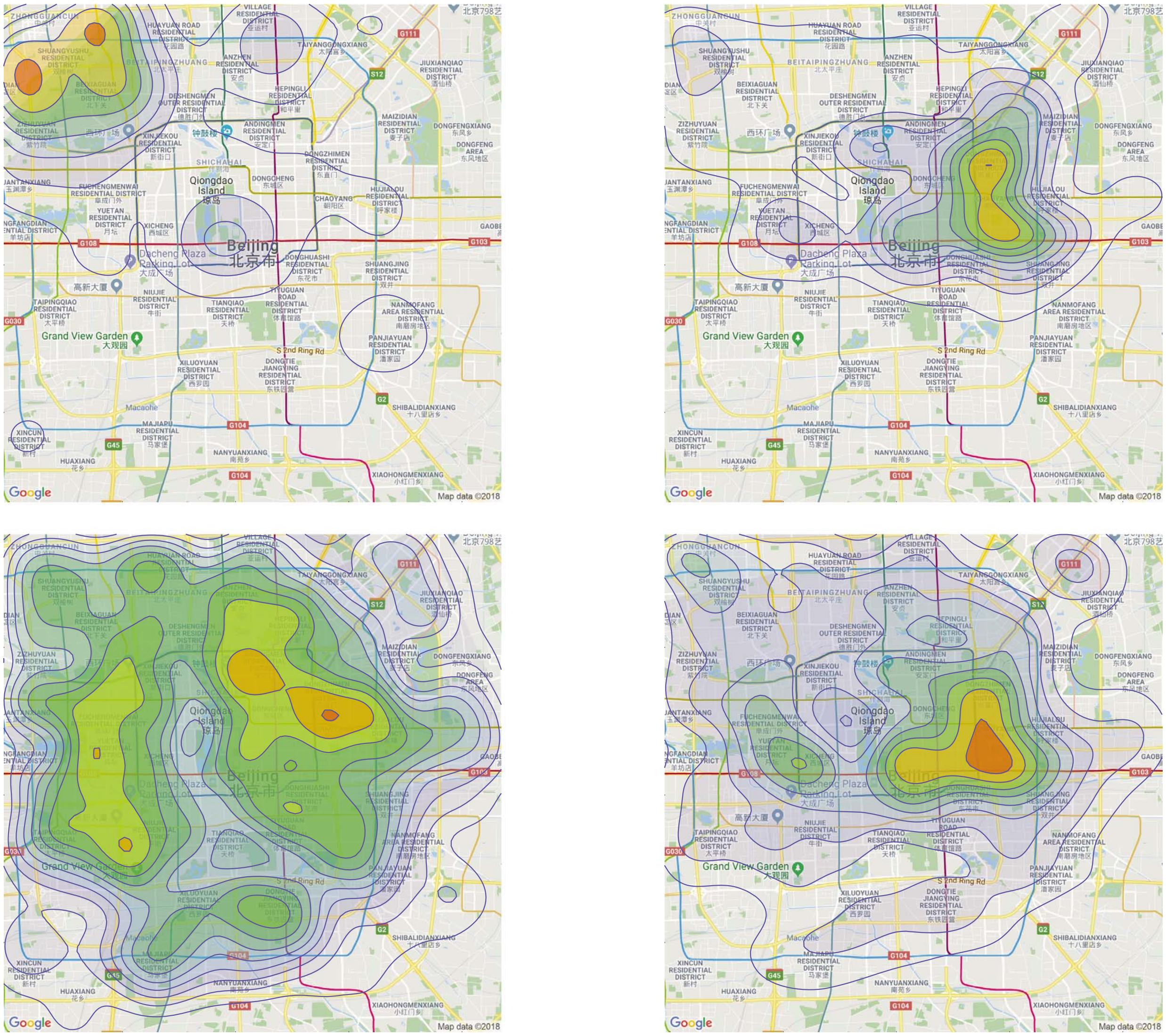
ZHONGGUANCUN
SHUANGYUSHU RESIDENTIAL DISTRICT
HUAYUAN ROAD RESIDENTIAL DISTRICT
VILLAGE RESIDENTIAL DISTRICT
TAIYANGGONGXIANG
JIUXIANQIAO RESIDENTIAL DISTRICT
BEITAIPINGZHUANG
ANZHEN RESIDENTIAL DISTRICT
HEPINGLI RESIDENTIAL DISTRICT
BEIXIAGUAN RESIDENTIAL DISTRICT
DESHENGMEN OUTER RESIDENTIAL DISTRICT
ZIZHUYUAN RESIDENTIAL DISTRICT
XINJIEKOU RESIDENTIAL DISTRICT
ANDINGMEN RESIDENTIAL DISTRICT
MAIZIDIAN RESIDENTIAL DISTRICT
DONGFENGXIANG
DONGFENG AREA
DONGZHIMEN RESIDENTIAL DISTRICT
SHICHAHAI
Qiongdao Island
DONGCHENG
CHAOYANG
HUJIALOU RESIDENTIAL DISTRICT
FUCHENGMENWAI RESIDENTIAL DISTRICT
YUETAN RESIDENTIAL DISTRICT
XICHENG
Beijing
北京市
Dacheng Plaza Parking Lot
DONGHUASHI RESIDENTIAL DISTRICT
SHUANGJING RESIDENTIAL DISTRICT
GAOBEI
TAIPINGQIAO RESIDENTIAL DISTRICT
NIUJIE RESIDENTIAL DISTRICT
TIANQIAO RESIDENTIAL DISTRICT
TIYUGUAN ROAD RESIDENTIAL DISTRICT
NANMOFANG AREA RESIDENTIAL DISTRICT
Grand View Garden
大观园
S 2nd Ring Rd
PANJIAYUAN RESIDENTIAL DISTRICT
XILUOYUAN RESIDENTIAL DISTRICT
DONGTIE JIANGYING RESIDENTIAL DISTRICT
Macaohe
MAJIAPU RESIDENTIAL DISTRICT
SHIBALIDIANXIANG
XINCUN RESIDENTIAL DISTRICT
HUAXIANG
NANYUANXIANG
XIAOHONGMENXIANG
G111
S12
G108
G103
G030
G104
G45
G2
Google
Map data ©2018

Figure 8. Heat maps four POI groups obtained by GLCP, high-tech (top left), diplomatic (top right), residential(bottom left) and commercial (bottom right).

that they were evenly scattered in the city. By comparing $\hat{\omega}$ of hospitals, kindergartens, bus stops, and other POIs, we can assess the allocation of the corresponding resources. For example, bus stop POIs are linked to residential area POIs, and the link parameter is 0.24, which implies approximately four residential area POIs share a bus stop. At the same time, the library category does not enter the residential group, indicating that the government should strengthen the popularity of the library.

Finally, the commercial group contains POI categories such as companies, banks, and hotels. The main center of the business group is in the eastern of the city, almost coincident with the diplomatic group, in addition to several centers in other locations. POIs in the commercial group may be helpful for location selection issues. For example, opening a Sichuan restaurant near other POIs in the group may be a good choice. In contrast, hot pot and Hunan cuisine are more suitable for the POI of the residential group.

Below we interpret the estimated coefficients. For $\hat{\Sigma}$ in Table 5, we find all σ_{ij} are significantly nonzero. There is still a significant correlation between latent random fields since the functions of different POIs in the city are intertwined and complex. This also confirms that the correlation between the latent random field is reasonable. The diagonal element in $\hat{\Sigma}$ is relatively large for high-tech and diplomatic group, which indicates that the distribution of these two groups of resources in the city is more uneven. Beijing universities are concentrated in Haidian District in the northeast corner, and the gathering place of foreigners is in Chaoyang District in the east of the city.

The most significant component of $\hat{\mu}$ comes from the residential group, among categories in the residential group, housing, medical care, and education contributed a large number of POIs. The estimated β is 12.38 with an estimated standard error of 2.65; this confirms the existence of the spatial correlation.

The results of GLCP estimation in the real data indicate that the multivariate spatial point process of POI has complex correlations, namely spatial correlation, within-group correlation and correlation between groups. Direct modeling of 30 spatial point processes with complex dependencies can be involved, while GLCP provides a simplified analytical approach with encouraging results.

6. CONCLUDING REMARKS

This article focuses on the analysis of POI data, in order to simplify the correlation structure between POI processes, a new spatial point process model GLCP is proposed. The proposed model can characterize multiple POI processes with cross-category dependence and group structure. A key contribution of GLCP is that it provides a data-driven way to group POI processes and we allow the latent random field to be correlated. GLCP combines the spatial pattern and the categorical information to analysis multiple POI processes, and it can be quickly scaled up to many cities. GLCP also provides a model-driven way to clustering spatial point processes; specifically, it uses a K function based distance and a minimum contrast estimation loss grouping criteria.

In the Beijing POI dataset, GLCP divides all POI processes into four groups, namely commercial group, residential group, diplomatic group and high-tech group. Based on the grouping method, we can further analyze the different functional areas of the city. GLCP provides an alternative way of dividing the urban functional area, based on POI data, with no significant boundaries between the different functional areas.

Future work may include the comparison of POI grouping in different cities, which may help us understand the differences between cities from several aspects (e.g., economic, social and environmental). From the perspective of simplifying the correlation structure, it is worthwhile to consider the compromise between the grouping model and the factor model. It is also a fascinating topic to analyze POI data in the spatiotemporal framework.

Received 31 October 2017

REFERENCES

[1] David R Cox. Some statistical methods connected with series of events. *J*ournal of the Royal Statistical Society. Series B (Methodological), pages 129–164, 1955. MR0092301

[2] Anthony Christopher Davison, David Victor Hinkley, et al. *B*ootstrap methods and their application, volume 1. Cambridge university press, 1997. MR1478673

[3] Debapratim Das Dawn and Soharab Hossain Shaikh. A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *T*he Visual Computer, 32(3):289–306, 2016.

[4] Peter J Diggle, Julian Besag, and J Timothy Gleaves. Statistical analysis of spatial point patterns by means of distance methods. *B*iometrics, pages 659–667, 1976.

[5] Peter J Diggle and Robin K Milne. Bivariate cox processes: some models for bivariate spatial point patterns. *J*ournal of the Royal Statistical Society. Series B (Methodological), pages 11–21, 1983.

[6] Qinhua Fang, Ran Zhang, Luoping Zhang, and Huasheng Hong. Marine functional zoning in china: experience and prospects. *C*oastal Management, 39(6):656–667, 2011.

[7] Jing He, Xin Li, Lejian Liao, Dandan Song, and William K Cheung. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. In *A*AAI, pages 137–143, 2016.

[8] Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. Learning geographical preferences for point-of-interest recommendation. In *P*roceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1043–1051. ACM, 2013.

[9] Zhao Liu, Meihui Xie, Kun Tian, and Peichao Gao. Gis-based analysis of population exposure to pm2. 5 air pollution—a case study of beijing. *J*ournal of Environmental Sciences, 59:48–53, 2017.

[10] Kanti V Mardia and Colin R Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. *M*ultivariate environmental statistics, 6(76):347–385, 1993.

[11] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *I*CwSM, 11(70-573):2, 2011.

[12] Tuomas Rajala, DJ Murrell, and SC Olhede. Detecting multivariate interactions in spatial point patterns with gibbs models and variable selection. *J*ournal of the Royal Statistical Society: Series C (Applied Statistics), 2017.

[13] Brian D Ripley. Modelling spatial patterns. *J*ournal of the Royal Statistical Society. Series B (Methodological), pages 172–212, 1977.

[14] Dietrich Stoyan and Antti Penttinen. Recent applications of point process methods in forestry statistics. *S*tatistical Science, pages 61–78, 2000.

[15] Maria Nicolette Margaretha Van Lieshout. Nonparametric indices of dependence between components for inhomogeneous multivariate random measures and marked sets. *S*candinavian Journal of Statistics, 2018.

[16] Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento. Urban mobility study using taxi traces. In *P*roceedings of the 2011 international workshop on Trajectory data mining and analysis, pages 23–30. ACM, 2011.

[17] Rasmus Waagepetersen, Yongtao Guan, Abdollah Jalilian, and Jorge Mateu. Analysis of multispecies point patterns by using multivariate log-gaussian cox processes. *J*ournal of the Royal Statistical Society: Series C (Applied Statistics), 65(1):77–96, 2016.

[18] Rasmus Plenge Waagepetersen. An estimating function approach to inference for inhomogeneous neyman–scott processes. *B*iometrics, 63(1):252–258, 2007.

[19] Jingyuan Wang, Fei Gao, Peng Cui, Chao Li, and Zhang Xiong. Discovering urban spatio-temporal structure from time-evolving traffic networks. In *A*sia-Pacific Web Conference, pages 93–104. Springer, 2014.

[20] Zhiwen Yu, Huang Xu, Zhe Yang, and Bin Guo. Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. *I*EEE Transactions on Human-Machine Systems, 46(1):151–158, 2016.

[21] Jing Yuan, Yu Zheng, and Xing Xie. 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*, pages 186–194. ACM, 2012.

[22] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Na-

dia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 363–372. ACM, 2013.

[23] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1444. ACM, 2013.

Yu Chen
Guanghua School of Management
Peking University
Beijing, 100871
P. R. China
E-mail address: yu.chen@pku.edu.cn

Rui Pan
School of Statistics and Mathematics
Central University of Finance and Economics
Beijing, 100081
P. R. China
E-mail address: panrui_cufe@126.com

Rong Guan
School of Statistics and Mathematics
Central University of Finance and Economics
Beijing, 100081
P. R. China
E-mail address: rongguan77@gmail.com

Hansheng Wang
Guanghua School of Management
Peking University
Beijing, 100871
P. R. China
E-mail address: hansheng@pku.edu.cn