Simultaneous Urban Region Function Discovery and Popularity Estimation via an Infinite Urbanization Process Model

Bang Zhang
Data61 CSIRO
Sydney, New South Wales, Australia
matt.zhang@data61.csiro.au

Lelin Zhang
Data61 CSIRO
Sydney, New South Wales, Australia
merlin.zhang@data61.csiro.au

Ting Guo
Data61 CSIRO
Sydney, New South Wales, Australia
ting.guo@data61.csiro.au

Yang Wang
Data61 CSIRO
Sydney, New South Wales, Australia
yang.wang@data61.csiro.au

Fang Chen
Data61 CSIRO
Sydney, New South Wales, Australia
fang.chen@data61.csiro.au

ABSTRACT

Urbanization is a global trend that we have all witnessed in the past decades. It brings us both opportunities and challenges. On the one hand, urban system is one of the most sophisticated social-economic systems that is responsible for efficiently providing supplies meeting the demand of residents in various of domains, *e.g.*, dwelling, education, entertainment, healthcare, *etc.* On the other hand, significant diversity and inequality exists in the development patterns of urban systems, which makes urban data analysis difficult. Different urban regions often exhibit diverse urbanization patterns and provide distinct urban functions, *e.g.*, commercial and residential areas offer significantly different urban functions. It is desired to develop the data analytic capabilities for discovering the underlying cross-domain urbanization patterns, clustering urban regions based on their function similarity and predicting region popularity in specified domains.

Previous studies in the urban data analysis area often just focus on individual domains and rarely consider cross-domain urban development patterns hidden in different urban regions. In this paper, we propose the infinite urbanization process (IUP) model for simultaneous urban region function discovery and region popularity prediction. The IUP model is a generative Bayesian nonparametric process that is capable of describing a potentially infinite number of urbanization patterns. It is developed within the supervised topic modeling framework and is supported by a novel hierarchical spatial distance dependent Bayesian nonparametric prior over the spatial region partition space. The empirical study conducted on the real-world datasets shows promising outcome compared with the state-of-the-art techniques.

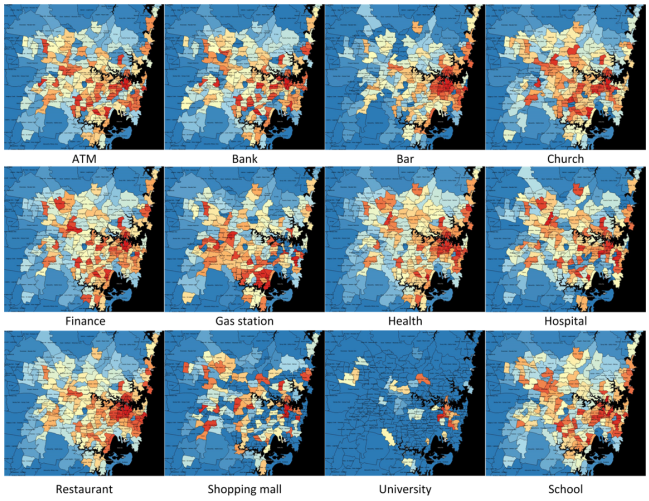

Figure 1: Urban supply density for the greater Sydney area. The statistical area 2 (SA2) defined by Australian Bureau of Statistics (ABS) is used as the region boundary for calculating different urban supplies' densities. Red colour indicates high density and blue colour indicate low density.

CCS CONCEPTS

• **Information systems** → **Location based services**; • **Computing methodologies** → *Supervised learning by regression*; *Topic modeling*;

KEYWORDS

Urban computing, urban function discovery, Bayesian nonparametric, and topic modeling.

ACM Reference Format:
Bang Zhang, Lelin Zhang, Ting Guo, Yang Wang, and Fang Chen. 2018. Simultaneous Urban Region Function Discovery and Popularity Estimation via an Infinite Urbanization Process Model. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19–23, 2018, London, United Kingdom,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3219819.3219987

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219987

1 INTRODUCTION

Urbanization is a global trend that we have all witnessed in the past decades. It has dramatically changed our world and will keep changing it on an unprecedented scale and speed. Fast urbanization also leads to fast-growing demand, *e.g.*, the demand for infrastructure,

transportation, energy, dwelling, education, healthcare, entertainment and communication. As a result, urban system is one of the most complex social-economic systems that provide supplies for meeting all the urban demand. It consists of a large number of subsystems, each of which is responsible for a particular demand and also closely collaborates with others. It requires a comprehensive understanding of the urban system as a whole to plan, operate and maintain our cities efficiently.

Existing research in urban data analysis mainly focus on individual area. For example, traffic condition and transportation capacity data analysis helps understand the equilibrium of urban transportation and detect traffic anomaly [2, 5, 20, 23]. Water quality and air condition data analysis helps estimate and forecast urban environment and resource status [13, 21, 32, 35]. Real estate data analysis [7–9] helps discover the important factors and patterns for property price.

Very few studies focus on integrating urban data from disparate domains [33], discovering correlations among different areas and deriving insights from cross-domain urban data. It is desired to treat a city as as integrated social-economic system and develop the data analysis techniques that can discover underlying cross-domain urban development patterns and predict unknown urban popularity in specified area.

In this paper, we propose a Bayesian nonparametric generative model. It makes the contribution via tackling the following three challenges in the urban data analysis and prediction area: (1) How to define and discover underlying urban development patterns hidden in the cross-domain urban data? (2) How to cluster urban regions based on their urbanization patterns? (3) How to make predictions for target domains given the discovered cross-domain correlations and patterns.

There are two important observations in the real-world urban systems providing helps for tackling the aforementioned challenges. First, significant diversity and inequality exist in urban systems. On the one hand, different urban regions have distinct compositions of urban supplies. For instance, a city's central business area is in high density of restaurants, cafes and hotels compared with residential areas. On the other hand, different urban supplies demonstrate distinct spatial distribution patterns over urban regions. For example, the spatial distribution patterns of restaurant and school are different from each other, as illustrated in Fig.1. Such diversity and inequality reflect the underlying cross-domain correlations and the hidden urban development themes in the different urban regions.

Second, spatial coherence exists for urban regions. As the Tobler's first law of geography states, "everything is related to everything else, but near things are more related than distant things" [30]. In urban data analysis, the urban regions next to each other tend to have similar demographical composition and share similar demand of urban functions.

In this paper, we propose the infinite urbanization process (IUP) model to tackle the three challenges based on the above observation. The IUP model is a Bayesian nonparametric generative process that can describe the generative urban development process. It can discover unique urbanization patterns, predict urban region popularity for specified domains and cluster urban regions into groups automatically without knowing the number of cluster beforehand, which is difficult for traditional parametric generative approaches.

Specifically, the proposed IUP model is a generative supervised topic model governed by a novel hierarchical spatial distance dependent Bayesian nonparametric prior over the urban region partition space. It combine the merits of both supervised latent topic modeling approach and distance dependent Bayesian nonparametric approach.

In this work, we use the term "urban function" to represent the hidden cross-domain urban development theme. Each urban region has its own urbanization theme which expresses itself via an unique combination of urban supplies, and we define the urban function as a distribution of urban supplies, *e.g.*, restaurants, healthcare, *etc.* Different urban regions have different compositions of urban supplies, and hence demonstrate distinct urban functions.

The proposed IUP model captures the analogy between text data analysis and urban data analysis, and utilize the supervised topic modeling technique for discovering urban functions and predicting urban region popularity in specified domains. Analogous to text analysis, urban regions are treated as documents with urban supplies being utilized as words. As a result, urban functions can be modeled as the latent topics in documents.

In such topic modeling framework, an urban region consisting of a large number of urban supplies, might be concisely modeled as deriving from a relatively small number of urban functions, *i.e.*, each urban region can be assigned with a distribution of functions. These urban function representation provide informative statistics for many different tasks, *e.g.*, searching, recommendation, similarity measurement, region segmentation, demand and supply estimation.

Particularly, the IUP model aim to predict urban region popularity in specified domains via utilizing the urban function representation. In this paper, we use real estate price as an example of urban region popularity to explain the model detail and conduct empirical study. But the model is general enough for modeling other urban variables, *e.g.*, region popularity for establishing a new business, region popularity for Uber or taxi drivers making profit, region popularity for commercial and politic campaigns.

To utilize the spatial coherence characteristics in urban data for region clustering, we develop a spatial distance dependent hierarchical Bayesian nonparametric prior over the urban region partition space for governing urban region clustering and urban function generation. Bayesian nonparametric priors are widely used for modeling the data in different structures, *e.g.*, matrices, graphs, arrays [24], trees, relations [14], spatial and temporal events [15–19], images [11] and 3D object surfaces [10]. Similar to the super-pixel concept used in image segmentation, the IUP mode uses the term "the smallest urban statistical area" (SUSA) as the smallest urban unit for extracting descriptive features. The SUSAs are grouped together based on both their spatial coherence and function similarity. In this work, we use the statistical area 1 (SA1) defined by Australian Bureau of Statistics (ABS)[1] as an example for model description and empirical study.

For the rest of the paper, Section 2 gives the technical details of the proposed model with a brief introduction of the related techniques. Section 3 describes the model inference method. Section 4 elaborates the empirical study. Finally, Section 5 concludes the work and discusses the potential future work.

[1] https://goo.gl/BvRcJQ

2 URBAN FUNCTION DISCOVERY AND POPULARITY ESTIMATION

In this section, we first introduce the related techniques. Section 2.1 describes the supervised topic model. Section 2.2 explains the distance dependent Chinese restaurant process, *i.e.*, a Bayesian nonparametric prior for partitioning. Then, we give the details of the proposed IUP model in Section 2.3.

2.1 Supervised Topic Model

Many latent topic models have been developed in recent years [12, 27, 31]. Latent Dirichlet allocation [4] (LDA) is extremely popular among them for text analysis because it provides an efficient way to extract concise and semantic representation of documents in a much lower dimension space compared with the dimension of unique vocabulary. Such latent topic representation is helpful for various text analysis tasks, *e.g.*, information retrieval, text clustering, *etc.*

Although the topics learned in unsupervised topic models can help reduce the dimension of texts, they can hardly be utilized for predicting documents' responses, *e.g.*, document relevance rank, as they express general themes without the ability to take advantage of supervised information.

In the supervised topic models [26, 34], the supervised latent Dirichlet allocation (sLDA) [22] is an extension of LDA for solving supervised learning problems. It overcomes the problem via jointly learning topics and their regression coefficients for the document responses. The response for a document is predicted by regressing on the averaged empirical topic allocations of the document. The generative process of sLDA for generating a document can be described as:

1. Draw topic proportion for document d: $\theta_d \sim Dir(\zeta)$.
2. For each word n in the document:
   (a) Draw topic assignment $Z_n|\theta_d \sim Mult(\theta_d)$.
   (b) Draw topics $\phi_{1:T}|\beta \sim Dir(\beta)$,
   (c) Draw a word $W_n|Z_n, \phi_{1:T} \sim Mult(\phi_{Z_n})$.
3. Draw document response $Y|Z_{1:N}, \eta, \sigma^2 \sim N(\eta^T \bar{Z}, \sigma_2)$.

The variables d and n indicate the indices of document and word respectively. The variable N represents the number of words and the variable T represents the number of topics. The variables θ and ϕ indicate the topic assignments for documents and the topics over the vocabulary respectively with ζ and β as hyperparameters.

2.2 Distance dependent CRP

Chinese restaurant process (CRP) [25] is a distribution over a probability measure. It is one of the three popular representations of Dirichlet process (DP) [1, 6] with emphasis on the clustering nature of DP. As a result, it is widely used as a Bayesian nonparametric prior for clustering methods and mixture models. It avoids the model selection problem that hinders most of the parametric models. It also helps solve the open-ended problems in which the number of components or patterns in the mixture model can grow with incoming data points. The model developed with CRP as a prior can learn the number of clusters automatically from data without knowing it beforehand.

The generative process of CRP can be described via a metaphor. Imaging there is a Chinese restaurant with infinite number of tables serving customers. The new customer comes in and sits at a table with the probability proportional to the number of existing customers at that table. A concentration parameter α of the CRP controls the probability that the incoming customer sits at a new table. We can see from the metaphor that the table with more customers tends to attract more new customers than the table with fewer customers.

Although described in a sequential manner, the CRP generates exchangeable distributions over partitions. In other words, the clustering outcome is invariant to the order of customers. This is a perfect assumption for many real-world applications in which exchangeability is needed. But it is inappropriate for urban region clustering in our case, in which nearby urban regions are more likely to have similar demographical composition and demand for urban functions.

The first Law of Geography, according to Waldo Tobler, states "everything is related to everything else, but near things are more related than distant things". Such spatial coherence nature in urban data analysis breaks the assumption of exchangeability. The distance dependent CRP (ddCRP) [3] modifies the CRP by determining the table assignment via customer links. A new customer tends to sit in a table where her closest friend sits. Formally, the generative process of the ddCRP can be described through the customer assignments c_i, which indirectly determines the table assignment t_i, where i indicates the customer index. The customer assignments are generated according to the distribution:

$$p(c_i = j|\alpha, f, \mathcal{D}) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \alpha & \text{if } i = j, \end{cases} \tag{1}$$

In Equation 1, d_{ij} indicates the predefined distance measurement between customer pairs, α is the probability that a new customer is assigned to herself (*i.e.*, sitting at a new table), and $f(\cdot)$ is a decaying function mediates how the distance between two customers affects their probability of linking to each other. The overall customer assignments indirectly determine the table assignments, which specifies the partition of customers.

Two customers are in the same cluster if and only if a customer can reach the other customer via the customer links. In our case, we treat SUSAs as customers. An example is given in Figure 3 by using SA1 in the greater Sydney area as SUSAs. Arrows indicate customer assignment and curved arrows indicate self-assignment.

We define the spatial distance of SUSAs as the number of hops required to reach each other. We set the decay function as $f(d) = \mathbf{1}(d \leq a)$, where $\mathbf{1}(\cdot)$ is the indicator function. It equals to 1 when the input condition is satisfied, and 0 otherwise. Such decay function forces spatial continuality. Only adjacent SUSAs can be linked (we set $a = 1$) as only the spatially connected SUSAs have non zero probability of linking to each other. However, it doesn't limit the number of linked SUSAs.

With the ddCRP as the prior over the urban region partition space, we can generate a set of contiguous segments (corresponding to tables) having similar urban functions. Although it is valuable, such segmentation has the limit on clustering urban regions with similar functions but far away from each other. For example, if a city

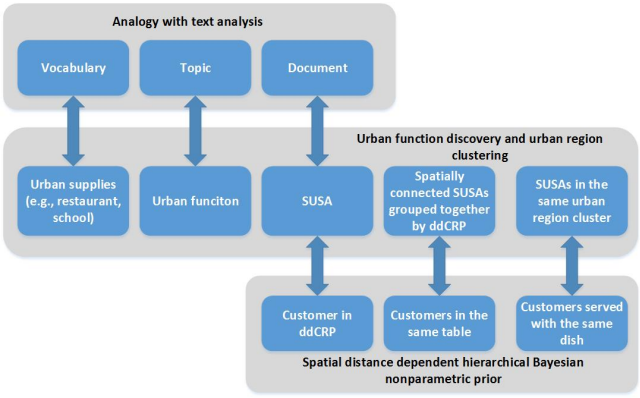

Figure 2: Top two rows: Analogy from urban function discovery to text analysis. Bottom two rows: The correspondence between urban region clustering and the metaphor of spatial distance dependent hierarchical Bayesian nonparametric prior.

have multiple business areas and share similar urban functions, such areas should be clustered as one group because they have similar functions and tend to have similar demand for urban supplies.

Therefore, we introduce another hierarchy layer (a standard CRP) on top of the ddCRP in which the connected SUSAs generated by the ddCRP are further grouped into clusters. Similar to the Chinese restaurant franchise (CRF) [28], the new hierarchy layer (CRP) assigns each table with a dish (as the final cluster index) from a menu consisting of a potentially countably infinite number of dishes.

Now, each cluster can sample the model parameter for generating the urban functions for the SUSAs in it. It is worth noting that this hierarchical ddCRP can be regarded as an extension of the CRF with the ddCRP determining the customer assignments instead of the standard CRP.

Such spatial distance dependent hierarchical Bayesian nonparametric prior can capture both the spatial coherence via the spatial ddCRP and the urban function similarity via the standard CRP layer.

2.3 Infinite Urbanization Process Model

In this section, we give the formal description of the proposed IUP model. Similar to the latent topic models which are utilized for understanding and organizing documents in a low dimensional space, the IUP model provides an interpretable and compact statistical representation for depicting the themes of urban regions.

There is an analogy between latent topic model for text analysis and the IUP model for urban data analysis, as shown in Figure 2. We treat urban supplies, *e.g.*, restaurants, schools, *etc.*, as words arising from a set of latent urban functions which correspond to latent document topics. A latent urban function is defined as a distribution of urban supplies. SUSAs are considered as documents in the IUP model, each SUSA is assigned with a set of urban functions. All the SUSAs share the same set of urban functions.

In the IUP model, each SUSA is also associated with a response representing a particular social-economic characteristics of the SUSA, *e.g.*, dwelling popularity. This social-economic response can be in various types, *e.g.*, ordering, positive integer or real-valued numbers.

Similar to image segmentation in which images are observed as a collection of super-pixels, *i.e.*, the smallest blocks of spatially adjacent pixels generating depicting visual features, we treat cities as a set of SUSAs (*e.g.*, SA1 defined by ABS) providing the smallest spatial granularity for obtaining statistical description. The goal of the IUP model is to cluster SUSAs into groups in which the SUSAs in the same group share similar urban functions.

To cluster SUSAs with the consideration of both spatial coherence and function similarity, we develop the spatial distance dependent hierarchical Bayesian nonparametric prior over the SUSA partition space. As introduced in Section 2.2 and illustrated in Figure 2, SUSAs correspond to the customers in the Bayesian nonparametric prior. Each SUSA links to another SUSA via customer assignment which is governed by Equation 1. A SUSA can only assigned to its adjacent SUSA or itself as we set $a = 1$ in the decay function $f(\cdot)$. It is worth noting that this does not restrict the size of the table as any pair of SUSAs that can reach each other are in the same table. The SUSAs assigned to the same table form a spatially connected urban region in which similar functions are shared among SUSAs.

However, urban regions with similar functions might be far away form each other. The model needs the ability to group similar urban regions that are not adjacent to each other. As in the CRF, a standard CRP prior is used for further grouping tables. The tables that are served with the same dish belong to the same cluster. In such way, the urban regions that are not adjacent to each other but share similar urban functions can be grouped together. Because of its nonparametric nature, the IUP model can generate an infinite large number of clusters. It can determines the number of clusters automatically based on the observed data points.

The key idea of the IUP model is to have the supervised topic model for discovering the latent urban functions and predicting urban region popularity while a spatial distance dependent hierarchical Bayesian nonparametric prior is developed over that urban regions partition space for governing urban region clustering and urban function generation. The generative process of the IUP model is given in the following:

1. For each customer (*i.e.*, SUSA), sample customer assignment, $c_i \sim ddCRP(\alpha, f, \mathcal{D})$. It implicitly determines the table assignment $t_{1:N}$.
2. For each table t, sample dish assignment (table grouping) $k_t \sim CRP(\gamma)$.
3. For each dish (*i.e.*, urban region cluster), sample cluster parameter (topic proportion) $\theta_k \sim Dir(\zeta)$.
4. For each SUSA:
   (a) For each urban supply (*i.e.*, word) with the assigned region cluster index:
      (i) Draw urban function assignment $Z_n|\theta_k \sim Mult(\theta_k)$.
      (ii) Draw urban function $\phi_{1:T}|\beta \sim Dir(\beta)$,
      (iii) Draw urban supply $W_n|Z_n, \phi_{1:T} \sim Mult(\phi_{Z_n})$.
   (b) Draw response variable $Y|Z_{1:N}, \eta_k, \sigma_k^2 \sim N(\eta^T \bar{Z}, \sigma_2)$.

The variable $\bar{Z}$ is defined as the empirical frequencies of the urban functions assigned to the SUSA. Variables η_k and σ_k^2 are

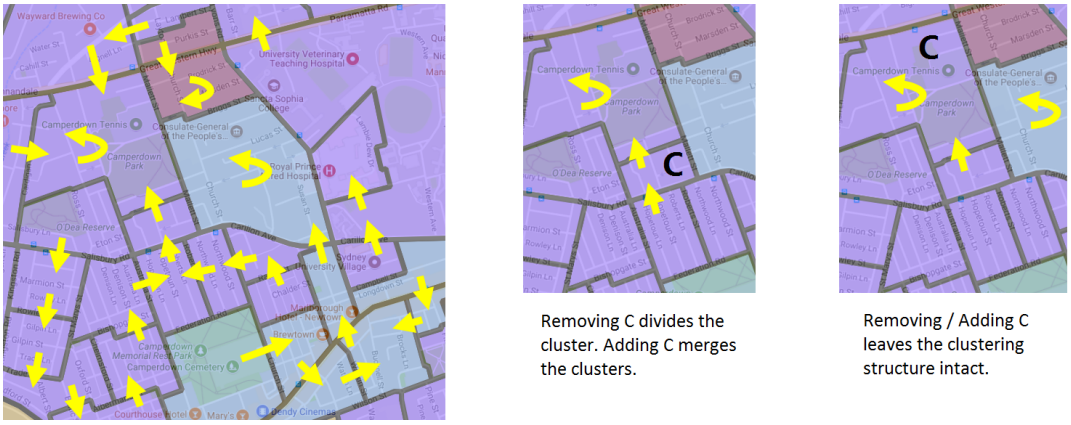

Figure 3: Left: An example of the relationship between customer assignment representation and table assignment representation. Each region is a SUSA, *i.e.*, SA1. Arrows indicate customer assignments. Curved arrow indicates self-assignment. Middle: An illustration of the outcome of reassigning a customer assignment with cluster structure changes. Right: An illustration of the outcome of reassigning a customer assignment without cluster structure changes.

the parameters in cluster k for generating responses. We skip the cluster index k for notation simplicity when no confusion exists.

3 MODEL INFERENCE

In this section, we give the inference method for inferring model parameters from observed data points.

The exact inference is analytically intractable for CRP-based models due to the combinatorial nature in partitions. Hence, approximated inference is used. Particularly, we adopt the collapsed Gibbs sampler due to the fact that hyperparameters are conjugate priors of model parameters in the IUP model.

The key part of the inference is to infer the cluster index and urban function assignment for SUSAs. The original work of the ddCRP [3] provides the sampling details for updating the customer assignment variables. Here, we first introduce the collapsed Gibbs sampler for the ddCRP in terms of the SUSA assignment. Then, we elaborate how to extend this sampler for the proposed hierarchical prior, *i.e.*, a standard CRP layer on top of the ddCRP, in which the CRP prior guides the generation of the table grouping and the ddCRP prior governs the generation of the customer assignments.

To sample the SUSA assignment for the IUP model, the conditional distribution of the SUSA assignment, c_i, on the other parameters, hyperparameters and observations can be derived as the following with $\boldsymbol{Z}, \boldsymbol{W}$ and $\boldsymbol{Y}$ as the data vector.

$$p(c_i|c_{-i}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \mathcal{D}, \alpha, \zeta, \beta, \eta, \sigma^2) \propto p(c_i|\mathcal{D}, \alpha) p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}|\pi(c_{-i} \cup c_i), \zeta, \beta, \eta, \sigma^2) \tag{2}$$

The prior term in Equation 2 is given by Equation 1. For the likelihood term, it can be further decomposed to Equation 3. Function $\pi(\cdot)$ represents the conversion from customer assignments to table assignments.

$$p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}|\pi(c_{-i} \cup c_i), \zeta, \beta, \eta, \sigma^2) = \Pi_{k=1}^{|\pi(c)|} p(\boldsymbol{Z}_{\pi(c_{1:N})=k}, \boldsymbol{W}_{\pi(c_{1:N})=k}, \boldsymbol{Y}_{\pi(c_{1:N})=k}|\pi(c_{1:N}), \zeta, \beta, \eta, \sigma^2). \tag{3}$$

In Equation 3, we define $|\pi(c)|$ as the number of unique clusters (treat them as tables for now, further explanation will be given for the hierarchical prior), and $\boldsymbol{Z}_{\pi(c_{1:N})=k}$ as the set of Z_i (latent function vector) which are generated from cluster k. Similarly, $\boldsymbol{W}_{\pi(c_{1:N})=k}$ are the business supplies generated from cluster k, and $\boldsymbol{Y}_{\pi(c_{1:N})=k}$ are the responses generated from cluster k.

For a particular cluster, the joint distribution can be expressed as:

$$p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \theta, \phi|\zeta, \beta, \eta, \sigma^2) = \prod_{t=1}^{T} p(\phi_t|\beta) \cdot \prod_{d=1}^{D} p(\theta_d|\zeta) \cdot \prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(W_{d,n}|\phi_{Z_{d,n}}) p(Y_d|\bar{Z}_d, \eta, \sigma^2). \tag{4}$$

Here, T represents the number of urban functions, D represents the number of SUSAs, and N represents the number of urban supplies. Variables ζ and β are the hyperparameters, expressing the characteristics of the priors on the model parameters θ and ϕ. We assume the priors are symmetric Dirichlet. Hence, ζ and β are scalars. Due to the fact that (1) the priors are conjugate to the multinomial distributions of θ and ϕ, (2) θ and ϕ are independent to each other, and (3) $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is irrelevant to θ and ϕ, we can calculate the likelihood term as the following:

$$\begin{aligned} p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}|\zeta, \beta, \eta, \sigma^2) &= \int_\theta \int_\phi p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \theta, \phi|\zeta, \beta, \eta, \sigma^2)\, d\phi d\theta \\ &= \int_\theta \prod_{d=1}^{D} p(\theta_d|\zeta) \prod_{n=1}^{N} p(Z_{d,n}|\theta_d) d\theta \cdot \\ &\int_\phi \prod_{t=1}^{T} p(\phi_k|\beta) \prod_{d=1}^{D} \prod_{n=1}^{N} p(W_{d,n}|\phi_{Z_{d,n}}) d\phi \cdot \\ &\prod_{d=1}^{D} p(Y_d|Z_{d,n}, \eta, \sigma^2). \end{aligned} \tag{5}$$

The first term can be computed by integrating out θ as:

$$\left(\frac{\Gamma(T\zeta)}{\Gamma(\zeta)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \zeta)}{\Gamma(n_\cdot^{(d)} + T\zeta)}, \tag{6}$$

where $n_j^{(d)}$ represents the number of times document d's words are associated to topic j and $\Gamma(\cdot)$ is the gamma function. The second term can be commutated by integrating out ϕ as:

$$\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + V\beta)}, \tag{7}$$

where V represents the number of unique words, $n_j^{(w)}$ represents the number of times word w are associated with topic j. The last term can be computed as:

$$\prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_d - \eta^T \bar{Z}_d)^2}{2\sigma^2}\right) \tag{8}$$

As elaborated in [3], to sample c_i from Equation 2, we need to consider different situations when the current c_i is removed and a new c_i is assigned. Firstly, removing the current c_i can either leave the current cluster (urban region cluster) structure intact, or divide the cluster associated with data point i to two clusters. Secondly, reassigning c_i can also lead to two different situations: (1) New c_i value leaves the cluster structure intact. (2) The new c_i joins the cluster associated with data point i with another cluster.

Figure 3 gives an illustration of the different situations in removing and reassigning c_i. The Gibbs sampler can explore the space of the potential urban region partition space via such removing and reassigning process of c_i.

When new value is assigned to c_i, we need to consider the prior probability of such new c_i value and the corresponding changes in the likelihood term. If we use l and m represent the indices of the tables that are joined for indexing k, then the resampling of the customer assignment can be derived as:

$$p(c_i|c_{-i}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \mathcal{D}, \alpha, \zeta, \beta, \eta, \sigma^2) \propto \begin{cases} p(c_i|\mathcal{D}, \alpha)\Delta(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \zeta, \beta, \eta, \sigma^2) & \text{if } c_i \text{ joins } l \text{ and } m, \\ p(c_i|\mathcal{D}, \alpha) & \text{otherwise,} \end{cases} \tag{9}$$

where the Δ function is defined as:

$$\Delta(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}, \zeta, \beta, \eta, \sigma^2) = \frac{p(\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{\pi(c_{1:N})=k}|\zeta, \beta, \eta, \sigma^2)}{p(\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{\pi(c_{1:N})=m}|\zeta, \beta, \eta, \sigma^2)p(\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{\pi(c_{1:N})=l}|\zeta, \beta, \eta, \sigma^2)}. \tag{10}$$

For the proposed IUP model, the sampler remains the same except three changes. Firstly, removing the current c_i can cause a new urban cluster which needs to be sampled from the urban cluster level, *i.e.*, the standard CRP prior. Secondly, the likelihood term in the above equation now depends on the customers assigned to the tables that are assigned to the same cluster instead of the customers assigned to the same table. Finally, the resampling of the urban region cluster assignment (dish assignment) can be performed as:

$$p(k_t = l|k_{-t}, \{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{1:N}, \pi(c_{1:N}), \gamma, \zeta, \beta, \eta, \sigma^2) \propto \begin{cases} s_l^{-t} p(\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_t|\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{-t}, \zeta, \beta, \eta, \sigma^2) & \text{if } l \text{ exists,} \\ \gamma p(\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_t|\zeta, \beta, \eta, \sigma^2) & \text{if } l \text{ is new,} \end{cases} \tag{11}$$

where $\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_t$ indicate the collection of data vectors that sit at table t, $\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_{-t}$ indicate the collection of data vectors that are assigned with cluster l excluding $\{\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y}\}_t$, and s_l^{-t} represents the number of tables that are associated with the cluster l excluding table t.

Sampling urban function assignment $\boldsymbol{Z}$ given cluster index is the same to the sLDA [28].

Once the function assignment of urban supplies are sampled, we can estimate the parameters η and σ^2 via MLE:

$$\hat{\eta_{MLE}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}, \tag{12}$$

$$\hat{\sigma^2_{MLE}} = \frac{1}{D}(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}), \tag{13}$$

where $\boldsymbol{A}$ is the matrix in which rows are the vectors $\bar{Z}_d^T$ and $\boldsymbol{Y}$ is the $D \times 1$ document response vector.

To predict the response for a new SUSA given the fitted model, we sample the function assignment for each word following the above steps for Z_{w_i}. With $\bar{Z}$ defined as $\sum_{w_i} Z_{w_i}$, we can predict the response variable as:

$$\hat{Y} = \hat{\eta}^T \bar{Z}. \tag{14}$$

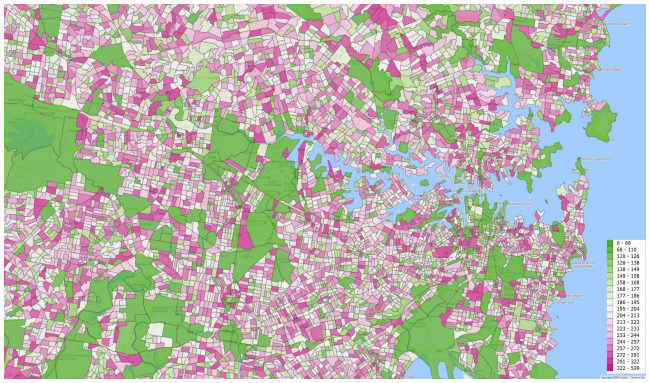

Figure 4: Statistical area 1 (SA1) is defined by Australian Bureau of Statistic (ABS) as the smallest unit for the release of census data. SA1 regions are coloured by their population sizes in the figure. High population SA1 regions are coloured in dark purple while low population SA1 regions are coloured in dark green colour.

4 EMPIRICAL STUDY

In this section, we conduct comparison experiments between the proposed IUP model and several state-of-the-art approaches on the real-world datasets to demonstrate the superiority of the proposed IUP model.

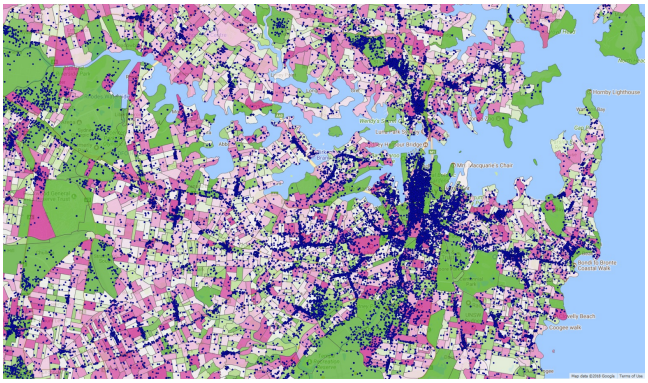

Figure 5: Business location information. Each business is treated as a word in its corresponding SA1 region regarded as a document.

Figure 6: Samples of unique business types and their relative proportions. The sum of all these sample business types' relative proportions equals to 100% **in the pie chart. They take** 12% **business location records in all the** 1100 **business types.**

In this study, we use the greater Sydney area as an example to evaluate the performance of the proposed IUP model on urban region function discovery and urban popularity estimation.

As mentioned, we utilize the Statistical Area 1 (SA1)[2] defined by Australian Bureau of Statistics (ABS) as the smallest urban statistical area (SUSA) for obtaining descriptive features depicting various urban characteristics. It is the finest spatial granularity for our urban data study as the super-pixel defined in the image segmentation study which is the smallest grid or patch for extracting descriptive visual features and the basic element for segmentation. Figure 4 demonstrates the SA1 regions that we used in this study. SA1 regions are coloured by their population sizes in Figure 4.

To study the urban region function of the greater Sydney area, we collect and combine business location information from both Yellow Page[3] and Google Place[4]. Business locations are matched with SA1 regions, as shown in Figure 5. Each individual business location with its business type is treated as a word in the corresponding SA1 region which is regarded as a document.

More than 2500 different business types are collected in the original dataset, and we remove the business types which have less than 20 locations. It generates 1100 unique business types in total. Some popular business types are shown in Figure 6. Here, we use restaurant and cafe as examples to discuss the quality of the dataset for representing the real-world situation. As studied by ABS[5], there were 13987 cafe and restaurant businesses operating in Australia at the end of June 2007 and a large proportion of them operated in the greater Sydney area. In the collected dataset, there are 6381 restaurants and 2999 cafes, namely 9380 in total. It is difficult, if not impossible, to collect all the restaurant and cafe locations for the greater Sydney area. But the collected dataset provides a relatively high-quality representation reflecting the underlying demand and supply nature of urban regions.

Urban region popularity can be measured in different ways, *e.g.*, travel convenience, education quality, *etc.* In this empirical study, we use property price as the measure of urban region popularity. Compared with other popularity measurements, the property price provides a reliable indicator of regions' underlying values, which reflects the value of their urban functions. We collect more than 1 million property location and price information from the Office of the Value General New South Wales[6] for the greater Sydney area, and only the properties that have transaction records in the recent 5 years are kept for study. Figure 7 shows the collected property information. Properties are coloured by their recent transaction prices. Dark purple indicates high transaction prices and white indicates low transaction prices. The median property price in each SA1 is used as the response variable representing region popularity.

To evaluate the performance of the proposed approach, we compare the IUP model with the state-of-the-art approaches that can retrieve latent topics and estimate response variables, *i.e.*, supervised Latent Dirichlet Allocation (sLDA), supervised Hierarchical Dirichlet Process (sHDP). The IUP model is implemented by using the MCMC with 3000 iterations. We tried the number from 5 and 60 as the dimension of θ, the dimension of urban functions assigned to SUSAs.

We also compared the IUP model to the LASSO, *i.e.*, L_1-regularized least squares regression. It is widely used for prediction problems in high-dimension space [29]. We use each SA1's empirical distribution of business types as its LASSO covariates.

The predictive R^2 score is used for measuring the regression performance, which is defined in Equation 15.

$$pR^2 = 1 - \frac{\sum_d (\hat{y_d} - y_d)^2}{\sum_d (y_d - \bar{y})^2}, \tag{15}$$

where y_d is the observed response with d as SA1 index, $\hat{y_d}$ is the predicted response, and $\bar{y} = 1/D \sum_{d=1}^{D} y_d$ is the mean of the observed responses. pR^2 is widely used for measuring the performance of regression models. Its value indicates the proportion of the variability in the data that can be explained by the model. A value of 1 means the model perfectly explains all the observed data points.

[2]https://goo.gl/BvRcJQ
[3]https://www.yellowpages.com.au/
[4]https://developers.google.com/places
[5]https://goo.gl/KAgVTm
[6]https://goo.gl/xKuExX

Figure 7: Property locations and the the recent prices. Properties are coloured based on their recent selling prices. Dark purple indicates high selling prices.

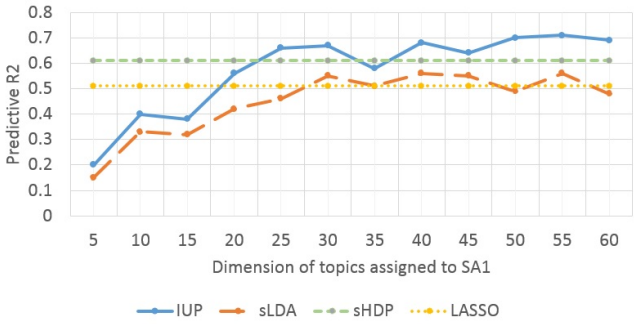

Figure 8: The prediction results of the compared methods.

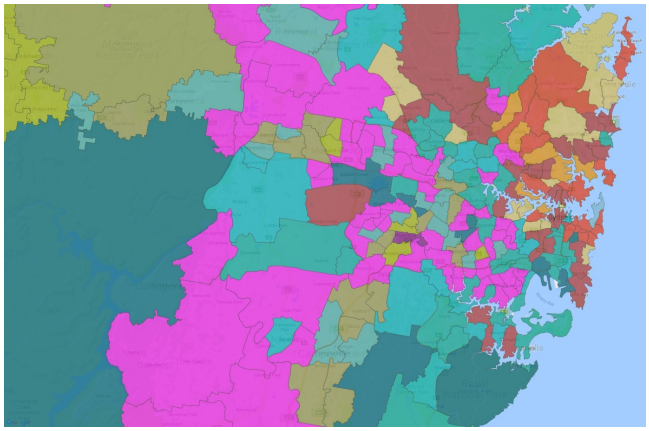

Figure 9: The outcome of the urban region clustering. The regions in the same colour are in the same cluster as the result of the balance between urban function similarity and spatial coherence.

Table 1: The most positive topics obtained in terms of their regression coefficients for predicting property prices.

Top topic 1	Top topic 2	Top topic 3
Restaurants	Medical practitioners	Lawyers & Solicitors
Fast Food	Real Estate Agents	Child Care Centres
Cafes	Dentist	Accountants & Auditors
ATM	Florist	Tuition & Tutoring
Bus Stop	Vet	Supermarket & Grocery

Five-fold cross validation is used for performance measurement. We iteratively use each fold as the testing set and the rest as the training set. The results are averaged over all the tests. The results are shown in Figure 8. As we can see, the IUP model outperforms the other approaches. Supervised HDP achieves the second place. Supervised LDA slightly outperforms the LASSO. The comparison outcome verifies: (1) The latent urban functions as covariates offer better representation for predicting the response variable compared with using the empirical distribution of business types as covariates (LASSO). (2) The proposed spatial distance dependent Bayesian nonparametric prior helps generate more informative latent urban function representation compared with the traditional Bayesian parametric approach (sLDA) and Bayesian nonparametric prior without considering spatial coherence (sHDP).

For the IUP model, the top positive topics in terms of their regression coefficients and their most frequent urban supplies are shown in Table 1. From the frequent urban supplies in each top topic, we can obtain a concise and informative description or summarization for the urban regions. For instance, the top topic 1 depicts the urban regions that are in busy business areas with high-volume crowd flow. The top topic 2 describes the urban regions that have been well developed. The top topic 3 summarizes the urban regions that are developing with strong demand in various professional services.

Finally, we show the urban region clustering outcome from the best setting of the IUP model. The result is visualized in Fig. 9. The regions in the same cluster are in the same colour. We can see that the spatial connectivity is preserved and no scattered small regions in the result. Moreover, the region similarities in urban functions are well captured, *e.g.*, central business areas are grouped together in orange, the western developed areas are grouped in cyan, the western developing areas are in pink. Other developed regions are grouped in dark orange and light brown. Both dark green and dark brown areas are developing areas. Other compared approaches are lack of the capability to partition urban regions considering both spatial coherence and function similarity.

5 CONCLUSIONS

In this paper, we proposed an infinite urbanization process (IUP) model that can discover urban functions from cross-domain urban data and predict urban popularity in specified domain simultaneously. The proposed IUP model has a spatial distance dependent hierarchical Bayesian nonparametric prior over the urban region partition space. It can be regarded as an extension of the Chinese restaurant franchise (CRF), in which the customer (the smallest urban statistical area) assignment is governed by a spatial ddCRP instead of CRP, while the dish assignment (for grouping tables across restaurants) is still governed by a standard CRP. With such hierarchical Bayesian nonparametric prior, the urban regions with similar urban function can be clustered together

Besides, we do not need to preset the number of urban region clusters, which can be difficult and jeopardizes the performance. Instead, the proposed model can automatically learn the number of clusters from the provided data.

The empirical study shows promising outcome, suggesting that the proposed IUP model can well capture the latent urban development themes and use the obtained compact urban theme representation to make accurate predictions for urban popularity in specified domains.

For the future work, a Bayesian nonparametric prior for the dimension of the latent representation space can be added. Another possible extension is to consider the changes of urban region-functions over time. The proposed model can be extended to capture the temporal pattern of urban development.

REFERENCES

[1] Charles E Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* 2, 6 (1974), 1152–1174.

[2] Jie Bao, Tianfu He, Sijie Ruan, Yanhua Li, and Yu Zheng. 2017. Planning Bike Lanes Based on Sharing-Bikes' Trajectories. In *SIGKDD 2017*. ACM, New York, NY, USA, 1377–1386.

[3] David M Blei and Peter I Frazier. 2011. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* 12, Aug (2011), 2461–2488.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] Sanjay Chawla, Yu Zheng, and Jiafeng Hu. 2012. Inferring the Root Cause in Road Traffic Anomalies. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12)*. IEEE Computer Society, Washington, DC, USA, 141–150.

[6] Thomas S Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics* 1, 2 (1973), 209–230.

[7] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. 2014. Sparse real estate ranking with online user reviews and offline moving behaviors. In *ICDM*. IEEE, Computer Society, Washington, DC, USA, 120–129.

[8] Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. 2015. Real Estate Ranking via Mixed Land-use Latent Models. In *SIGKDD*. ACM, New York, NY, USA, 299–308.

[9] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. 2014. Exploiting Geographic Dependencies for Real Estate Appraisal: A Mutual Perspective of Ranking and Clustering. In *Proceedings of the 20th ACM SIGKDD*. ACM, New York, NY, USA, 1047–1056.

[10] Soumya Ghosh, Matthew Loper, Erik B. Sudderth, and Michael J. Black. 2012. From Deformations to Parts: Motion-based Segmentation of 3D Objects. In *NIPS*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Lake TAHOE, Nevada, USA, 1997–2005.

[11] Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei. 2011. Spatial distance dependent Chinese restaurant processes for image segmentation. In *NIPS*. Curran Associates, Inc., Granada, Spain, 1476–1484.

[12] Thomas Hofmann. 2000. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in neural information processing systems*. MIT Press, Denver, CO, USA, 914–920.

[13] Hsun-Ping Hsieh, Shou-De Lin, and Yu Zheng. 2015. Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, ACM, New York, NY, USA, 437–446.

[14] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. 2006. Learning Systems of Concepts with an Infinite Relational Model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1 (AAAI'06)*. AAAI Press, Boston, Massachusetts, USA, 381–388.

[15] Bin Li, Bang Zhang, Zhidong Li, Yang Wang, Fang Chen, and Dammika Vitanage. 2015. Prioritising water pipes for condition assessment with data analytics. (2015).

[16] Zhidong Li, Bang Zhang, Yang Wang, Fang Chen, Ronnie Taib, Vicky Whiffin, and Yi Wang. 2014. Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data. *Mach. Learn.* 95, 1 (April 2014), 11–26. https://doi.org/10.1007/s10994-013-5386-z

[17] Peng Lin, Bang Zhang, Ting Guo, Yang Wang, and Fang Chen. 2016. Interaction Point Processes via Infinite Branching Model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, Phoenix, Arizona USA, 1853–1859.

[18] Peng Lin, Bang Zhang, Ting Guo, Yang Wang, Fang Chen, et al. 2016. Infinite hidden semi-Markov modulated interaction point process. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Barcelona, Spain, 3900–3908.

[19] Peng Lin, Bang Zhang, Yi Wang, Zhidong Li, Bin Li, Yang Wang, and Fang Chen. 2015. Data driven water pipe failure prediction: A bayesian nonparametric approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, ACM, New York, NY, USA, 193–202.

[20] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, ACM, New York, NY, USA, 1010–1018.

[21] Ye Liu, Yu Zheng, Yuxuan Liang, Shuming Liu, and David S. Rosenblum. 2016. Urban Water Quality Prediction Based on Multi-task Multi-view Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York City, New York, USA, 2576–2582.

[22] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. Curran Associates, Inc., Vancouver, B.C., Canada, 121–128.

[23] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. 2017. City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In *SIGSPATIAL*. ACM, New York, NY, USA, Article 1, 10 pages. https://doi.org/10.1145/3139958.3139984

[24] Peter Orbanz and Daniel M Roy. 2015. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence* 37, 2 (2015), 437–461.

[25] Jim Pitman et al. 2002. Combinatorial stochastic processes. *Lecture Notes for St. Flour Summer School*. 1, 1 (2002).

[26] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. ACL, Association for Computational Linguistics, Stroudsburg, PA, USA, 248–256.

[27] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, AUAI Press, Arlington, Virginia, United States, 487–494.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* 101, 476 (2006), 1566–1581.

[29] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288.

[30] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[31] Alexei Vinokourov and Mark Girolami. 2002. A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of intelligent information systems* 18, 2-3 (2002), 153–172.

[32] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting Fine-Grained Air Quality Based on Big Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 2267–2276.

[33] Yu Zheng, Huichu Zhang, and Yong Yu. 2015. Detecting Collective Anomalies from Multiple Spatio-temporal Datasets Across Different Domains. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, New York, NY, USA, 2:1–2:10.

[34] Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, New York, NY, USA, 1257–1264.

[35] Julie Yixuan Zhu, Chao Zhang, Huichu Zhang, Shi Zhi, Victor OK Li, Jiawei Han, and Yu Zheng. 2017. pg-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data. *IEEE Transactions on Big Data* early access (2017).