Real Estate Ranking via Mixed Land-use Latent Models

Yanjie Fu[‡], Guannan Liu[◇], Spiros Papadimitriou[‡], Hui Xiong[‡], Yong Ge[§], Hengshu Zhu[+], Chen Zhu[†]
[‡]Rutgers University, USA, {yanjie.fu, s.papadim, hxiong}@rutgers.edu
[◇]Tsinghua University, China, liugn.10@sem.tsinghua.edu.cn
[+]Big Data Lab, Baidu Research, China, zhuhengshu@baidu.com
[§]UNC Charlotte, USA, yong.ge@uncc.edu
[†]University of Science and Technology of China, China, zhuzhen@mail.ustc.edu.cn

ABSTRACT

Mixed land use refers to the effort of putting residential, commercial and recreational uses in close proximity to one another. This can contribute economic benefits, support viable public transit, and enhance the perceived security of an area. It is naturally promising to investigate how to rank real estate from the viewpoint of diverse mixed land use, which can be reflected by the portfolio of community functions in the observed area. To that end, in this paper, we develop a geographical function ranking method, named FuncDivRank, by incorporating the functional diversity of communities into real estate appraisal. Specifically, we first design a geographic function learning model to jointly capture the correlations among estate neighborhoods, urban functions, temporal effects, and user mobility patterns. In this way we can learn latent community functions and the corresponding portfolios of estates from human mobility data and Point of Interest (POI) data. Then, we learn the estate ranking indicator by simultaneously maximizing ranking consistency and functional diversity, in a unified probabilistic optimization framework. Finally, we conduct a comprehensive evaluation with real-world data. The experimental results demonstrate the enhanced performance of the proposed method for real estate appraisal.

1. INTRODUCTION

Mixed land use is increasingly popular in the real estate development of big cities. Mixed land use is the combination of multiple compatible land uses, including residential, commercial, and recreational uses within a certain area [30]. Mixed land use can: (i) contribute economic benefits, e.g., commercial areas in close proximity to residential areas can increase property values; (ii) support viable public transit; and (iii) enhance the perceived security, e.g., by helping increase activity and hence the presence of people on the street. More importantly, a balanced mix of land uses leads to the co-location of socio-economic functions, and thus yields livable, sustainable, and viable neighborhoods.

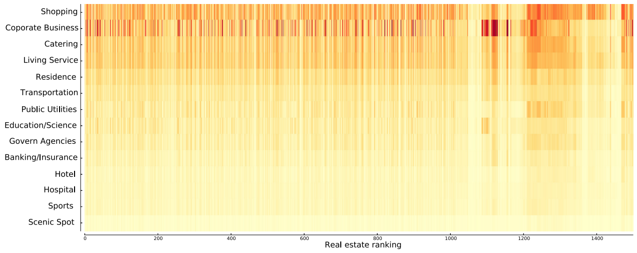


Figure 1: The POI density spectrum of estates over multiple POI categories.

Research literature has developed empirical evidence for the value of mixed land use. Many studies have shown that, in big cities, people value a balanced mix of land uses more than other key indicators of real estate value [30, 19, 22]. A recent study reported that people are willing to pay almost 25% more for a house in an area with appropriate mixed land use, and one standard deviation increase in diversity increases real estate prices by 1.00%–4.25% [19]. Indeed, Figure 1 shows the point of interest (POI) density spectrum of real estate over multiple POI categories. As can be seen, the spectrum of high-ranked estates (left) is more evenly balanced than that of low-ranked estates (right). The evidence illustrates that investment value of real estate with a balanced mix of neighborhood functions is usually higher than otherwise comparable real estate in mono-functional areas.

All the above evidence suggests it is highly appealing to investigate how to rank real estate values based on the functional diversity of land uses. Two unique challenges arise in achieving this goal. First, the community functions and the corresponding portfolios that affect value need to be effectively identified. Second, the relationship between these portfolios and real estate value ranking needs to be modeled. We outline how we tackle these two main challenges next.

First, the impact of mixed use on property values largely depends on the *specific* composition of land uses. Some functions can increase real estate values, while others may not have significant impact. For instance, manufacturing usually degrades property values. In contrast, more commercial land use, such as entertainment and retail stores, can lead to higher property values. People are generally willing to pay more for uses that are compatible with residential values and less for uses that negatively impact house prices. Therefore, compatible functions should be carefully selected for mixed land use. However, identifying these functions is a nontrivial task. For example, some studies [22] revealed that, within a certain range, proximity to commercial uses

KDD '15, August 10-13, 2015, Sydney, NSW, Australia.
© 2015 ACM. ISBN 978-1-4503-3664-2/15/08…$15.00.
http://dx.doi.org/10.1145/2783258.2783383.

has a negative effect on real estate value. Therefore, the first question arises: how to identify functions that are compatible with real estate values and learn the portfolio of these identified functions in the target community? Traditionally, real estate professionals use regression analysis to determine the significance and the direction of the relationship between real estate value and functions.

Unlike traditional approaches, we treat these unknown functions as latent factors and learn the portfolio of functions from human mobility data. During different time periods, there are different perceived functions in a community, and thus different patterns can be observed in the human mobility data of the community, which include taxi GPS traces, bus GPS traces, and user check-in data. The human mobility patterns in a community jointly reflect the diverse mixtures of neighborhood functions [34]. For example, on workdays, people generally leave a residential area in the morning and return in the evening. Also, people usually check into entertainment places on workday evenings or during the entire day over weekends. Therefore, in this paper, we exploit human mobility patterns for identifying the latent compatible functions and for learning the portfolio of community functions.

Second, after we learn the portfolio of community functions, we naturally come up with another question: how to evaluate the impact of the distribution of community functions on real estate value? Traditionally, real estate professionals use a two-step paradigm, which first defines entropy-like indexes, such as the Hirschmann-Herfindahl index, to measure the diversity of community functions, and then includes these indexes into regression models as independent variables [19, 12]. However, this paradigm may not be optimal for ranking, because these two steps are independently modeled. Instead, we treat the learned portfolio as the functional spectrum of the estate ranked list over K functions in a listwise manner. For each function k, we calculate the relevance score of the whole estate ranked list conditioned on k. Then, we aggregate the weighted sum of K relevance scores as a measure of functional diversity. Finally, we can jointly model both functional diversity and ranking consistency as a unified estate ranking objective for optimization.

Specifically, in this paper, we first develop a geographic functional learning model to jointly model the interrelationship among estate neighborhoods, urban functions, temporal effects, and mobility patterns for learning the portfolio of functions for each estate's neighborhood. In particular, we assume there are K latent functions and treat them as a latent categorical variable. At different time periods, an estate neighborhood exhibits different functions due to its particular mix of land uses. Given a specific function and a time period, an estate neighborhood has specific mobility patterns of taxi rides, bus trips, and check-ins. Here, we treat these patterns as three types of words in three different vocabularies (i.e., three different latent spaces). Hence, given a time period, a neighborhood has three clusters of words. We treat each word cluster as a mobility document. By fitting our geographic functional learning model to mobility data, we derive the portfolio of K neighborhood functions for each estate. Next, we incorporate functional diversity to learn an estate ranking indicator. In particular, we extract raw features from urban geography data and human mobility data, learn meta features by decision trees, and linearly regress these features to predict estate investment values. Moreover, we design a weighted sum function to capture the diversity of neighborhood functions in an estate ranked list. Along these lines, we train an estate ranking indicator by simultaneously maximizing ranking consistency and functional diversity in a unified probabilistic framework. Finally, we have conducted a comprehensive performance evaluation on real world data. The experimental results demonstrate the enhanced performance of the proposed method for real estate evaluation.

2. THE GEOGRAPHIC FUNCTIONAL RANKING FRAMEWORK

In this section, we first formally introduce the problem of geographic functional ranking, and then provide an overview of our ranking framework.

2.1 Problem Statement

Real estate investment value, different from market value (i.e., price), reflects the growth potential of resale value that can be higher or lower than market value to a particular investor. The unique characteristic of investment value motivates investors to enter the real estate marketplace, seek estates with high investment value, and maximize their investment returns. Therefore, the capability to rank estates based on investment ranking is necessary. Essentially, ranking estates is similar to ranking documents with a defined relevance, where an estate is analogized as a document and its investment value is considered as the relevance.

Formally, given a set of of M estates $E = \{e_1, e_2, ..., e_M\}$, the goal of our problem is to rank them in a descending order according to their investment values $Y = \{y_1, y_2, ..., y_M\}$. In this study, we assume each estate m has a location (i.e., latitude and longitude) and a neighborhood area (e.g., a circle with radius of 1 km), which we call an *estate community* in this paper. According to the theory of mixed land use, in urban areas of super cities, an estate's investment value largely depends on the functional portfolio of its community. In other words, a diverse mixture of community functions usually leads to high investment value of an estate. Indeed, the rankings of estates according to their investment value could be estimated by incorporating functional diversity of estate communities, using urban geography and human mobility data. Essentially, there are two major tasks: (1) learning the functional portfolios of estate communities from heterogeneous human mobility, and (2) predicting estate ranking by incorporating the impact of functional diversity.

2.2 Framework Overview

Figure 2 shows the framework of our geographic functional ranking. This framework consists of two major stages.

(1) Functional Portfolio Learning. As shown in Figure 2, we propose to learn the functional portfolio by mining three types of mobility patterns (i.e., mobile checkins, taxi trajectories, and bus trajectories), defined next.

Definition 1. (Checkin Pattern): Given a checkin event, the checkin pattern is a triple including information about (1) checkin day, (2) checkin hour, and (3) POI category of the checkin place.

Definition 2. (Taxi Mobility Pattern): Given a taxi trajectory, we extract the leaving (i.e., pick-up) and arriving

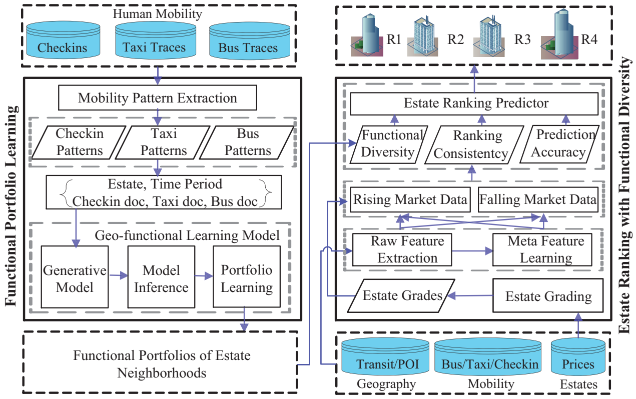


Figure 2: The framework overview of geographical functional ranking for estates.

(i.e., drop-off) patterns as two tuples, each of which contains information about (1) weekday or weekend, (2) hour, and (3) leaving or arriving.

Definition 3. (Bus Mobility Pattern): Given a bus trajectory, we extract the leaving (i.e., pick-up) and arriving (i.e., drop-off) patterns as two tuples, each of which contains information about (1) weekday or weekend, (2) hour, and (3) leaving or arriving.

We then associate all these mobility patterns to a nearby estate community once their checkin, pickup or dropoff points are located within the circle area of the estate with a radius of 1 km. Besides, we argue that the heterogeneous mobility patterns around an estate collectively reflect the mixed functions of its community. To this end, we assume there are multiple latent functions within the community of an estate. Moreover, an estate community shows different functions during different time periods. Therefore, given an estate and a time period, we can identify a unique mobility segmentation, which is defined as follows.

Definition 4. (Mobility Segment): A mobility segment is a six-item tuple including an estate, a time period, a latent function of the estate community in this time period, checkin pattern cluster, taxi pattern cluster, and bus pattern cluster.

According to the above definition, in each mobility segment, the estate has three clusters of mobility patterns generated by the functional portfolio of its community. To learn the functional portfolio of each estate community, here we adapt the idea of topic modeling and develop a novel generative model, where the mobility patterns and clusters are analogized as words and documents, respectively.

(2) Estate ranking with functional diversity. After learning the functional portfolios of estate communities, we extract the raw features from urban geography and human mobility. Furthermore, the raw features are then fed into ensemble decision trees (in our experiments, random forests) for generating meta features, and the output of each individual tree is treated as a meta feature. Here, we treat the investment value of an estate as a linear combination of both raw and meta features. Based on the above, we can learn an estate ranking predictor by jointly maximizing prediction accuracy, ranking consistency, and functional diversity. Finally, we infer the rankings of estates with the learned parameters. Next, Section 3 addresses the first problem of portfolio learning, and Section 4 of estate ranking.

3. LEARNING THE PORTFOLIO OF COMMUNITY FUNCTIONALITIES

Here we propose a topic modeling approach for learning the functional portfolios of estate communities with a collection of heterogeneous mobility patterns.

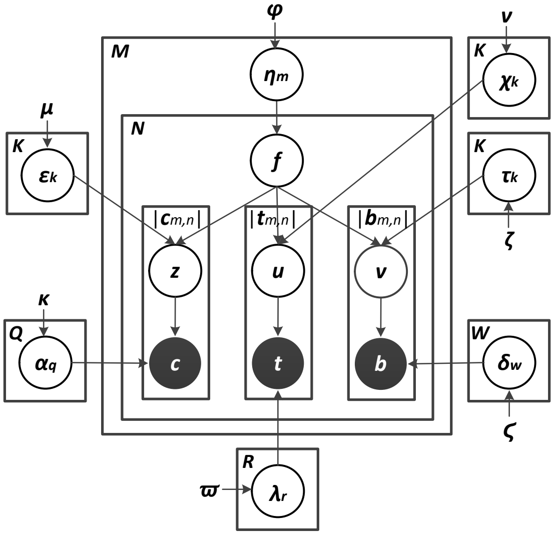


Figure 3: The graphical representation of the proposed geographic functional learning model.

3.1 Model Intuition

There are correlations among estate communities, urban functions, temporal effects, and mobility patterns. Therefore, in our approach, we model the generative process of checkin, taxi, and bus mobility for each estate community, based on the following intuition.

Intuition 1: A mixed estate community is represented as a mixture of urban functions in terms of its mixed land uses, and thus forms a portfolio of a fixed set of functions.

Intuition 2: The urban functions of a mixed community change over time. For example, people may visit an area for work on workday mornings, but visit the same area for entertainment during nights and weekends.

Intuition 3: Mobility patterns reflect the functions of a community. For example, the residential function of a place can be indicated by massive leaving patterns in the early morning (e.g., people take public transit to work) and massive arriving patterns around 6PM (e.g., people go home after work). Therefore, over a certain time period, a community shows specific mobility patterns which reflect a particular urban function.

Intuition 4: Given a time period, an estate community has three clusters of mobility patterns. By treating mobility patterns and clusters as words and documents, respectively, we can model the corresponding generative processes and uncover the latent urban function through topic modeling.

3.2 Model Specification

Figure 3 shows the graphical representation of our geographic functional learning model. Specifically, we use a multinomial distribution η_m over K latent functions to model

Table 1: The generative process of the geographic functional learning model.

```
For each function $f = k \in \{1, ..., K\}$:
    Draw a multinomial distribution $\epsilon_k \sim P(\epsilon_k|\mu)$
    Draw a multinomial distribution $\chi_k \sim P(\chi_k|\nu)$
    Draw a multinomial distribution $\tau_k \sim P(\tau_k|\zeta)$
For checkin latent topic $z = q \in \{1, ..., Q\}$:
    Draw a multinomial distribution $\alpha_q \sim P(\alpha_q|\kappa)$
For taxi latent topic $u = r \in \{1, ..., R\}$:
    Draw a multinomial distribution $\lambda_r \sim P(\lambda_r|\varpi)$
For bus latent topic $v = w \in \{1, ..., W\}$:
    Draw a multinomial distribution $\delta_w \sim P(\delta_w|\varsigma)$
For each estate $m \in \{1, ..., M\}$:
    Draw a multinomial distribution $\eta_m \sim P(\eta_m|\rho)$;
    For each time period $n \in \{1, ..., N\}$:
        Draw a community function $f \sim P(f|\eta_m)$;
        For each checkin mobility pattern $c \in \boldsymbol{c}_{m,n}$:
            Draw a latent topic of checkin document $z \sim P(z|\epsilon_f)$;
            Draw a checkin mobility pattern $c \sim P(c|\alpha_z)$.
        For each taxi mobility pattern $t \in \boldsymbol{t}_{m,n}$:
            Draw a latent topic of taxi document $u \sim P(u|\chi_f)$;
            Draw a taxi mobility pattern $t \sim P(t|\lambda_u)$.
        For each bus mobility pattern $b \in \boldsymbol{b}_{m,n}$:
            Draw a latent topic of taxi document $v \sim P(v|\tau_f)$;
            Draw a bus mobility pattern $b \sim P(b|\delta_v)$.
```

the functional portfolio of the estate m (**Intuition 1**). Based on **Intuition 2**, the functions of an estate community may vary over time. We thus segment historical mobility patterns of checkin, taxi, and bus into multiple segments in terms of N defined time periods. For example, if we define seven time periods (i.e., Monday to Sunday), we first segment mobility patterns day by day, and then group these segments into seven clusters, each of which corresponds to a day of the week. We denote a mobility segment by a tuple $\{m, n, f, \boldsymbol{c}_{m,n}, \boldsymbol{t}_{m,n}, \boldsymbol{b}_{m,n}\}$ introduced in Definition 4, which is generated as follows. For each time period n, an estate m shows a specific urban function f drawn from η_m. Note that each function f has: (1) a multinomial distribution ϵ_f over checkin latent topics, which represents the relevance of checkin latent topics to the urban function f; (2) a multinomial distribution χ_f over taxi latent topics, which represents the relevance of taxi latent topics to the urban function f; and (3) a multinomial distribution τ_f over bus latent topics, which represents the relevance of bus latent topics to the urban function f (**Intuition 3**). We iteratively draw: (1) a checkin latent topic z for each checkin pattern $c \in \boldsymbol{c}_{m,n}$ in checkin mobility document $\boldsymbol{c}_{m,n}$; (2) a taxi latent topic u for each taxi pattern $t \in \boldsymbol{t}_{m,n}$ in taxi mobility document $\boldsymbol{t}_{m,n}$; and (3) a bus latent topic v for each bus pattern $b \in \boldsymbol{b}_{m,n}$ in bus mobility document $\boldsymbol{b}_{m,n}$ (**Intuition 4**). In summary, Table 1 shows the generative process.

3.3 Model Inference

Let us denote all parameters by $\Psi = \{\boldsymbol{\eta}, \boldsymbol{\epsilon}, \boldsymbol{\chi}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\delta}\}$ where $\boldsymbol{\eta} = \{\eta_m\}_{m=1}^{M}$, $\boldsymbol{\epsilon} = \{\epsilon_k\}_{k=1}^{K}$, $\boldsymbol{\chi} = \{\chi_k\}_{k=1}^{K}$, $\boldsymbol{\tau} = \{\tau_k\}_{k=1}^{K}$, $\boldsymbol{\alpha} = \{\alpha_q\}_{q=1}^{Q}$, $\boldsymbol{\lambda} = \{\lambda_r\}_{r=1}^{R}$, $\boldsymbol{\delta} = \{\delta_w\}_{w=1}^{W}$, the hyperparameters $\Omega = \{\rho, \mu, \nu, \zeta, \kappa, \varpi, \varsigma\}$, the latent assignments of functions and topics $\Upsilon = \{\boldsymbol{F}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{V}\}$, and the observed mobility collection $D = \{\boldsymbol{C}, \boldsymbol{T}, \boldsymbol{B}\}$ where $\boldsymbol{C} = \{\boldsymbol{c}_{m,n}\}_{m=1,n=1}^{M,N}$, $\boldsymbol{T} = \{\boldsymbol{t}_{m,n}\}_{m=1,n=1}^{M,N}$, and $\boldsymbol{B} = \{\boldsymbol{b}_{m,n}\}_{m=1,n=1}^{M,N}$ are the checkin, taxi, and bus mobility documents of M estates for N time periods, respectively. Also, we use $\boldsymbol{P}_c$, $\boldsymbol{P}_t$, $\boldsymbol{P}_b$ to denote the vocabularies of checkin, taxi, and bus mobility patterns, respectively.

Following the generative process in Table 1, the joint distribution can be factored as

$$\begin{aligned} &P(D, \Upsilon, \Psi|\Omega) = P(D, \Upsilon|\Psi)P(\Psi|\Omega) \\ &= P(\boldsymbol{C}|\boldsymbol{\alpha})P(\boldsymbol{\alpha}|\kappa)P(\boldsymbol{T}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|\varpi)P(\boldsymbol{B}|\delta)P(\boldsymbol{\delta}|\varsigma)P(\boldsymbol{Z}|\boldsymbol{\epsilon})\times \\ &P(\boldsymbol{\epsilon}|\mu)P(\boldsymbol{U}|\boldsymbol{\chi})P(\boldsymbol{\chi}|\nu)P(\boldsymbol{V}|\boldsymbol{\tau})P(\boldsymbol{\tau}|\zeta)P(\boldsymbol{F}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\rho). \end{aligned} \tag{1}$$

We use Collapsed Gibbs sampling for training the model. Specifically, we derive the full conditional posteriors and obtain the update rules of both the latent assignments and the parameters. Let $\mathbb{C}_{z,*} = \{\mathbb{C}_{z,c}\}_{c=1}^{|\boldsymbol{P}_c|}$ where $\mathbb{C}_{z,c}$ denotes the number of checkin pattern c generated by checkin latent topic z; $\mathbb{T}_{u,*} = \{\mathbb{T}_{u,t}\}_{t=1}^{|\boldsymbol{P}_t|}$ where $\mathbb{T}_{u,t}$ denotes the number of taxi pattern t generated by latent topic u; $\mathbb{B}_{v,*} = \{\mathbb{B}_{v,b}\}_{b=1}^{|\boldsymbol{P}_b|}$ where $\mathbb{B}_{v,b}$ denotes the number of bus pattern b generated by latent topic v; $\mathbb{Z}_{f,*} = \{\mathbb{Z}_{f,z}\}_{z=1}^{Q}$ where $\mathbb{Z}_{f,z}$ denotes the number of checkin latent topic z generated by function f; $\mathbb{U}_{f,*} = \{\mathbb{U}_{f,u}\}_{u=1}^{R}$ where $\mathbb{U}_{f,u}$ denotes the number of taxi latent topic u generated by function f; $\mathbb{V}_{f,*} = \{\mathbb{V}_{f,v}\}_{v=1}^{W}$ where $\mathbb{V}_{f,v}$ denotes the number of bus latent topic v generated by function f; $\mathbb{F}_{m,*} = \{\mathbb{F}_{m,f}\}_{f=1}^{K}$ where $\mathbb{F}_{m,f}$ denotes the number of mobility segments whose urban function is f in an estate community m; $\mathbb{X}^{-(*)}$ represent the count of $\mathbb{X}$ excluding the component $(*)$ (e.g., $\mathbb{F}_{m,k}^{-(m,n)}$ represents the count of $\mathbb{F}_{m,k}$ excluding mobility segment (m,n)); Γ denote the gamma function.

For the n-th mobility segment in estate m , the conditional posterior probability for its latent function assignment f is computed by

$$\begin{aligned} P(f_{m,n} = k|\mathcal{D}, \Upsilon - f_{m,n}) &= \frac{\mathbb{F}_{m,k}^{-(m,n)} + \rho_k}{\sum_{f=1}^{K} \mathbb{F}_{m,f}^{-(m,n)} + \rho_f} \\ &\times \frac{\prod_{z=1}^{Q} \Gamma(\mathbb{Z}_{k,z} + \mu_z)\Gamma(\sum_{z=1}^{Q} \mathbb{Z}_{k,z}^{-(m,n)} + \mu_z)}{\prod_{z=1}^{Q} \Gamma(\mathbb{Z}_{k,z}^{-(m,n)} + \mu_z)\Gamma(\sum_{z=1}^{Q} \mathbb{Z}_{k,z} + \mu_z)} \\ &\times \frac{\prod_{u=1}^{R} \Gamma(\mathbb{U}_{k,u} + \nu_u)\Gamma(\sum_{u=1}^{R} \mathbb{U}_{k,u}^{-(m,n)} + \nu_u)}{\prod_{u=1}^{R} \Gamma(\mathbb{U}_{k,u}^{-(m,n)} + \nu_u)\Gamma(\sum_{u=1}^{R} \mathbb{U}_{k,u} + \nu_u)} \\ &\times \frac{\prod_{v=1}^{W} \Gamma(\mathbb{V}_{k,v} + \zeta_v)\Gamma(\sum_{v=1}^{W} \mathbb{V}_{k,v}^{-(m,n)} + \zeta_v)}{\prod_{v=1}^{W} \Gamma(\mathbb{V}_{k,v}^{-(m,n)} + \zeta_v)\Gamma(\sum_{v=1}^{W} \mathbb{V}_{k,v} + \zeta_v)}. \end{aligned} \tag{2}$$

For the i-th checkin pattern $c_{m,n,i} \in \boldsymbol{c}_{m,n}$, the conditional posterior for its latent checkin topic is computed by

$$\begin{aligned} &P(z_{m,n,i} = q|D, \Upsilon - z_{m,n,i}) \\ &= \frac{\mathbb{C}_{q,c_{m,n,i}}^{-(m,n,i)} + \kappa_{c_{m,n,i}}}{\sum_{c=1}^{|\boldsymbol{P}_c|} \mathbb{C}_{q,c}^{-(m,n,i)} + \kappa_c} \frac{\mathbb{Z}_{f_{m,n},q}^{-(m,n,i)} + \mu_q}{\sum_{z=1}^{Q} \mathbb{Z}_{f_{m,n},z}^{-(m,n,i)} + \mu_z)}. \end{aligned} \tag{3}$$

For the i-th taxi pattern $t_{m,n,i} \in \boldsymbol{t}_{m,n}$, the conditional posterior for its latent taxi topic is computed by

$$\begin{aligned} &P(u_{m,n,i} = r|D, \Upsilon - u_{m,n,i}) \\ &= \frac{\mathbb{T}_{r,t_{m,n,i}}^{-(m,n,i)} + \varpi_{t_{m,n,i}}}{\sum_{t=1}^{|\boldsymbol{P}_t|} \mathbb{T}_{r,t}^{-(m,n,i)} + \varpi_t} \frac{\mathbb{U}_{f_{m,n},r}^{-(m,n,i)} + \nu_r}{\sum_{u=1}^{R} \mathbb{U}_{f_{m,n},u}^{-(m,n,i)} + \nu_u}. \end{aligned} \tag{4}$$

For the i-th bus pattern $b_{m,n,i} \in \boldsymbol{b}_{m,n}$, the conditional posterior for its latent bus topic is computed by

$$\begin{aligned} &P(v_{m,n,i} = w|D, \Upsilon - v_{m,n,i}) \\ &= \frac{\mathbb{B}_{w,b_{m,n,i}}^{-(m,n,i)} + \varsigma_{b_{m,n,i}}}{\sum_{b=1}^{|\boldsymbol{P}_b|} \mathbb{B}_{w,b}^{-(m,n,i)} + \varsigma_b} \frac{\mathbb{V}_{f_{m,n},w}^{-(m,n,i)} + \zeta_w}{\sum_{v=1}^{W} \mathbb{V}_{f_{m,n},v}^{-(m,n,i)} + \zeta_v}. \end{aligned} \tag{5}$$

After all the latent assignments are learned, we obtain the update rules of the model parameters as $\eta_{m,f} = \frac{\mathbb{F}_{m,f}+\rho_f}{\sum_{k=1}^{K} \mathbb{F}_{m,k}+\rho_k}$, $\epsilon_{f,z} = \frac{\mathbb{Z}_{f,z}+\mu_z}{\sum_{q=1}^{Q} \mathbb{Z}_{f,q}+\mu_q}$, $\chi_{f,u} = \frac{\mathbb{U}_{f,u}+\nu_u}{\sum_{r=1}^{R} \mathbb{U}_{f,r}+\nu_r}$, $\tau_{f,v} = \frac{\mathbb{V}_{f,v}+\zeta_v}{\sum_{w=1}^{W} \mathbb{V}_{f,w}+\zeta_w}$,

$\alpha_{z,c} = \frac{\mathbb{C}_{z,c}+\kappa_c}{\sum_{p=1}^{|\boldsymbol{P}_c|}\mathbb{C}_{z,p}+\kappa_p}$, $\lambda_{u,t} = \frac{\mathbb{T}_{u,t}+\varpi_t}{\sum_{p=1}^{|\boldsymbol{P}_t|}\mathbb{T}_{u,p}+\varpi_p}$, $\delta_{v,b} = \frac{\mathbb{B}_{v,b}+\varsigma_b}{\sum_{p=1}^{|\boldsymbol{P}_b|}\mathbb{B}_{v,p}+\varsigma_p}$.

So far, we have learned the portfolios of M estate communities over K functions, i.e., $\boldsymbol{\eta} \in \mathbb{R}^{M\times K}$. Also, we can obtain the global portfolio of the entire city over K functions denoted by $\theta = \{\theta_f\}_{f=1}^{K}$ where $\theta_f = \frac{\sum_{m=1}^{M}\eta_{m,f}}{M}$.

4. ENHANCING ESTATE RANKING WITH FUNCTIONAL DIVERSITY

Next, we introduce the proposed estate ranker by incorporating the impact of functional diversity.

4.1 Modeling Estate Investment Value

Before introducing the overall objective function, let us first introduce how to model the investment value of estates.

Raw Features. Table 2 shows the raw features we have extracted from urban geography (e.g., bus stops, subway stations, road networks, POIs, etc.), human mobility (e.g., taxi trajectories, bus smart card transactions, checkins, etc.) and social media (e.g., online business reviews, etc.).

Table 2: The raw features extracted by neighborhood profiling.

Category	Source	Feature Design
Urban Geography	Transportation	Number of bus stop Distance to nearest bus stop
		Number of subway station Distance to nearest subway station
		Number of road network entries Distance to nearest road network entry
	POIs	Number of POIs of different POI categories
Human Mobility	Taxi	Taxi Arriving Volume Taxi Leaving Volume Taxi Transition Volume Taxi Driving Velocity Taxi Commute Distance
	Bus	Bus Arriving Volume Bus Leaving Volume Bus Transition Volume Bus Stop Density
	Checkin	Checkin Count Topical Profile
Social Media	Online User Reviews	Overall Rating Service Rating Environment Rating Consumption Cost

Meta Features. We exploit a random forest based method to learn meta features via supervised non-linear transformation. Indeed, the work in [14] proved that decision trees can help improve the accuracy of predicting clicks on online advertisements . Therefore, we feed raw features and ground-truth real estate investment values into random forest, and learn a set of decision trees (weak classifiers). We then treat each individual tree as a categorical feature which is represented by a binary-valued vector. The elements of vectors correspond to tree leaves and the values indicate whether an estate falls into the corresponding leaf. For example, [1,0,0] indicates the tree has three leaves and the estate falls into the first leaf.

Finally, we linearly combine both raw and meta features to formulate estate investment value. Formally, let $\boldsymbol{x_m}$ denote the I-size vector representation of estate m with the above extracted features, $\boldsymbol{w}$ denote the weights of features, g_m denote the predicted estate value of estate m, y_m denote the ground-truth investment value of estate m, and $\mathcal{N}$ represent the normal distribution. The generative process of our linear model is

- Draw feature weights $w_i \sim \mathcal{N}(w_i; 0, \sigma_w^2)$.
- For each estate m, generate estate value $y_m \sim \mathcal{N}(y_m; g_m, \sigma^2)$ where $g_m = \boldsymbol{w}^\top \boldsymbol{x_m} = \sum_{i=1}^{I} w_i x_{mi}$.

4.2 Incorporating Functional Diversity

Here, we introduce how to jointly model prediction accuracy, ranking consistency, and functional diversity in a unified objective function of posterior probability. Let us denote all the parameters by $\Phi = \{\boldsymbol{w}\}$, the hyperparameters $\Lambda = \{\sigma_w^2, \sigma_f^2\}$. Indeed, the estate ranked list contains three-component information of its ranking structure, denoted by $\Delta = \{Y, \Pi, \Xi\}$ where Y, Π, Ξ are the investment values, rankings, and functional diversity of M estates respectively. Let $\overline{\Pi}$ represent the inverse of Π and $\overline{\pi}_m$ be the index of the m-th ranked estate. For simplicity, we first assume that $m = \pi_m = \overline{\pi}_m$. In other words, the estates in Δ are sorted and indexed in a descending order in terms of their investment values (which coincides with descending rating rank). Therefore, the objective is to learn the parameters Φ that maximize the posterior probability $P(\Phi;\Delta\,,\Lambda)$ given the observed data and hyperparameters. By Bayesian inference, the posterior probability is

$$P(\Phi;\Delta\,,\Lambda) = P\left(\Delta|\Phi,\Lambda\right) P\left(\Phi|\Lambda\right). \tag{6}$$

We follow the commonly-used "bag of words" assumption [2], which in our setting corresponds to conditional independence of the investment value, ranking, and functional diversity of an estate, given parameters Φ and Λ. Then, the term $P\left(\Delta|\Phi,\Lambda\right)$ is the likelihood of the observed data collection Δas

$$\begin{aligned} &P\left(\Delta|\Phi,\Lambda\right) = P\left(\{Y,\Pi,\Xi\}\,|\Phi,\Lambda\right) \\ &= \underbrace{P\left(Y|\Phi,\Lambda\right)}_{\text{Prediction Accuracy}} \times \underbrace{P\left(\Pi|\Phi,\Lambda\right)}_{\text{Ranking Consistency}} \times \underbrace{P\left(\Xi|\Phi,\Lambda\right)}_{\text{Functional Diversity}}, \end{aligned} \tag{7}$$

where $P\left(Y|\Phi,\Lambda\right)$ denotes the likelihood of the observed investment values of estates given the parameters, which corresponds to prediction accuracy. $P\left(\Pi|\Phi,\Lambda\right)$ denotes the likelihood of the rankings of estates given the parameters, which captures ranking consistency. $P\left(\Xi|\Phi,\Lambda\right)$ denotes the likelihood of the functional diversity of the estate ranking list. Next, we introduce the modeling of prediction accuracy, ranking consistency, and functional diversity in detail.

Prediction Accuracy. The smaller loss, the higher prediction accuracy for estate investment value.

$$P\left(Y|\Phi,\Lambda\right) = \prod_{m=1}^{M}\mathcal{N}(y_m|g_m,\sigma) = \prod_{m=1}^{M}\frac{1}{\sigma}\exp\left(-\frac{(y_m-g_m)^2}{2\sigma^2}\right). \tag{8}$$

Ranking Consistency. The ranked list of estates indeed can be encoded into a directed acyclic graph (DAG), $G = \{V, E\}$, with the node set V as estates and the edge set E as pairwise ranking orders. For instance, edge $m \to h$ represents that estate m is ranked higher than estate h. From a generative modeling angle, edge $m \to h$ is generated by our model through a likelihood function $P(m \to h)$. The more valuable an estate m is compared to estate h, the larger $P(m \to h)$ should be.

$$P\left(\Pi|\Phi,\Lambda\right) = \prod_{m=1}^{M-1}\prod_{h=m+1}^{M} P(m \to h|\Phi,\Lambda), \tag{9}$$

where the generative likelihood of each edge $m \to h$ is defined as Sigmoid$(g_m - g_h)$: $P(m \to h) = \frac{1}{1+exp(-(g_m-g_h))}$.

Functional Diversity. So far, each estate is associated with a vector of K-dimensional distribution of functions. An estate with diverse functions is likely to have higher investment value and appears earlier in the estate ranked list. Therefore, one goal of our estate ranker is to find a list of estate such that high-ranked estates maximally cover the K functions. Specifically, for each function k, we calculate the relevance score of the entire estate ranked list conditioned on the function k. We then aggregate the weighted sum of K relevance scores as a measurement of functional diversity.

$$
\begin{aligned}
&P\left(\Xi|\Phi,\Lambda\right) = \sum_{f=1}^{K} P(f)P(\Xi|f,\Phi,\Lambda) \\
&= \sum_{f=1}^{K} \frac{\theta_f}{1 + exp(-(\sum_{m=1}^{M} g_m \frac{\sum_{h=1}^{m}\eta_{h,f}}{m} - \sum_{m=1}^{M} g_m \eta_{m,f}))}.
\end{aligned}
\tag{10}
$$

Second, the term $P(\Phi|\Lambda)$ is the prior of the parameters Φ. Since we have extracted many features, we impose a zero-mean Gaussian distribution with variance σ^2 for each weight. This is known to enforce weak sparse representations during learning, by setting some feature weights to zero for automatic feature selection, $P(\Phi|\Lambda) = \prod_{i=1}^{I} \mathcal{N}(w_i|0,\sigma_w^2)$.

4.3 Parameter Estimation

With the formulated posterior probability, the learning objective is to find the optimal estimate of the parametersΦ that maximizes the posterior. Hence, by inferring Equation 6, we can obtain the log of the posterior for the proposed model.

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{w}|Y,\Pi,\Xi,\sigma^2,\sigma_w^2) = \sum_{m=1}^{M}\left[-\frac{1}{2}\ln\sigma^2 - \frac{(y_m - f_m)^2}{2\sigma^2}\right] \\
&+ \sum_{m=1}^{M-1}\sum_{h=m+1}^{M} ln\frac{1}{1+exp(-(g_m - g_h))} + \sum_{i=1}^{I}\left[-\frac{1}{2}\ln\sigma_w^2 - \frac{w_i^2}{2\sigma_w^2}\right] \\
&+ ln\sum_{f=1}^{K}\theta_f \frac{1}{1 + exp(-(\sum_{m=1}^{M} g_m \frac{\sum_{h=1}^{m}\eta_{h,f}}{m} - \sum_{m=1}^{M} g_m \eta_{m,f}))}
\end{aligned}
\tag{11}
$$

We apply a gradient descent method to maximize the posterior, by updating w_i through $w_i^{(t+1)} = w_i^{(t)} - \epsilon\frac{\partial(-\mathcal{L})}{\partial w_i}$, where ϵ is the learning rate.

4.4 Ranking Inference

After obtaining the parameters, we can construct the ranking function for predicting the investment value of estates, i.e., $\mathbb{E}(y_m|\Phi) = \boldsymbol{x_m}^{\top}\boldsymbol{w}$. For a new estate k (lacking historical transaction information), we may predict its investment value accordingly. The larger the $\mathbb{E}(y_k|\Phi)$ is, the higher investment value it has.

5. EXPERIMENTAL RESULTS

This section details our empirical evaluation of the proposed method on real-world data.

5.1 Data Description

Table 3 shows the detailed statistics of our real-world data sets. The transportation data covers the bus system, the subway system, and the road networks of Beijing. We also extracted POI features from the Beijing POI data set. The taxi GPS traces were collected from a Beijing taxi company. Each trajectory contains trip ID, distance (m), travel time (s), average speed d(km/h), pick-up time, drop-off time, pick-up point, and drop-off point. In addition, we crawled the smart card transactions from the official website of Beijing Public Transportation Group. Each bus trip has card ID, time, expense, balance, route name, pick-up and drop-off stop information (name, longitude, and latitude). Moreover, the Beijing check-in data were crawled from www.jiepang.com, which is a Chinese version of Foursquare. Each check-in event includes checkin time, POI name, POI category, address, longitude, latitude, and comments. Furthermore, we crawled Beijing online business reviews from www.dianping.com, which is a business review site in China. Each review contains shop ID, name, address, latitude, longitude, consumption cost, star rating (1–5), POI category, environment, service, and overall rating. Finally, we crawled Beijing second-hand real estate data from www.soufun.com, which is the largest online real-estate system in China.

Table 3: Statistics of the experiment data.

Data Sources	Properties	Statistics
Bus stop(2011)	Number of bus stop	9,810
Subway(2011)	Number of subway station	215
Road networks (2011)	Number of road segments	162,246
	Total length(km)	20,022
	Percentage of major roads	7.5%
POIs	Number 0f POIs	300,811
	Number of categories	13
Taxi Traces	Number of taxis	13,597
	Effective days	92
	Time period	Apr. - Aug. 2012
	Number of trips	8,202,012
	Number of GPS points	111,602
	Total distance(km)	61,269,029
Smart Card Transactions	Number of bus stops	9,810
	Time Period	Aug 2012 to May 2013.
	Number of car holders	300,250
	Number of trips	1,730,000
Check-Ins	Number of check-in POIs	5,874
	Number of check-in events	2,762,128
	Number of POI categories	9
	Time Period	01/2012-12/2012
Business Review	Number of reviews	470846
	Number of users	159820
Real Estates	Number of estates	2,851
	Size of bounding box (km)	40*40
	Time period of transactions	04/2011 - 09/2012

In the real estate industry, investment value of a property is measured by return rate. This is the ratio of the price increase relative to the starting price of a market period , i.e., $r = \frac{P_f - P_i}{P_i}$, where P_f and P_i denote the final and initial prices, respectively. To prepare the benchmark investment values of estates (Y) for training data, we first calculated the return rate of each estate during a given market period. We then sorted the return rates of all the estates in descending order. Finally, we partition them into five clusters using variance-based top-down hierarchical clustering [12]. In this way, we segmented the estates into five ordered value categories (i.e., 4>3>2>1>0, the higher the better). Estate grading is a way to evaluate the investment potential and reduce the impact of fluctuations in return rates that do not

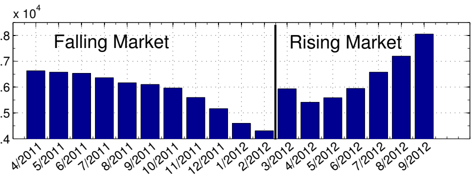


Figure 4: The rising and falling market periods in Beijing.

provide meaningful information about differences in real estate value.

Finally, a list of estates, together with the extracted features and investment value of each, were split into two data sets, corresponding to the falling market period (from July 2011 to February 2012) and the rising market period (from February 2012 to September 2012), as shown in Figure 4. Here we follow the norms of real estate research, which typically studies rising and falling markets separately [24, 6].

5.2 Baseline Algorithms

Since our work is related to Learning-To-Rank (LTR), we compared our method against the following algorithms. (1) **Coordinate Ascent [23]:** uses domination loss and coordinate descent optimization. (2) **LambdaMART [3]:** the boosted tree version of LambdaRank. (3) **FenchelRank [21]:** designed for solving sparse learning-to-rank with an L1 constraint. (4) **ListNet [5]:** a listwise ranking model with permutation top-k ranking likelihood as the objective function.

Beyond traditional ranking models, we further compare with two methods specifically designed for real estate ranking. (5) **SEK [12]:** exploits regression modeling, pairwise ranking objective, and sparsity regularization, to solve the real estate ranking problem. Also, its feature design includes the entropy of POI distribution, which is an summary index of functional diversity. (6) **ClusRanking [13]:** solves the estate ranking problem by capturing individual, peer, and zone dependencies.

In our experiments, we used RTree to index geographic items (e.g., POIs, trajectories, checkins, etc.) and extracted the defined features. For traditional LTR algorithms, we used RankLib [1]. For Coordinate Ascent, we set step base = 0.05, step scale = 2.0, tolerance = 0.001, and slack = 0.001. For LambdaMART, we set number of trees = 100, number of leaves = 10, number of threshold candidates = 256, learning rate = 0.1. For FenchelRank, we use the source code[2] provided by the author. For SEK, we set $a = 0.01$, $b = 0.01$, and $\sigma^2 = 1000$. For ClusRanking, we set β_1=0.8, β_2=25m, latent business areas $K = 10$, $\eta = \frac{1}{K}$, $\mu_q = \mu_w = 0$, $\sigma_q = \sigma_w = \sigma = 35$ and $M = 3$ for hyperparameters. For our method,we implemented the geo-functional learning model in C and DivFuncRanking model in Python with the Scipy optimization package. We used a KNN-based method to impute the values of missing features. To learn the meta features, we leveraged the scikit-learn random forest package, where the number of trees is set to 100. We randomly divided the data into 70% for training and 30% for testing, and used Matlab for result visualization.

5.3 Evaluation Metrics

Normalized Discounted Cumulative Gain. The discounted cumulative gain (DCG) metric is evaluated over top N estates on the ranked estate list by assuming that high-value estates should appear earlier in the ranked list. $DCG[n] = \begin{cases} rel_1 \quad if\, n = 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n}, \quad if\, n >= 2 \end{cases}$ Later, given the ideal discounted cumulative gain $DCG^{'}$, $NDCG$ at the n-th position can be computed as $NDCG[n] = \frac{DCG[n]}{DCG^{'}[n]}$, where ref_n refers to the investment rating of estate n.

Table 4: Examples of temporal topics and their patterns of check-in mobility.

Weekday Topics			Weekend Topics		
Topic 7	Topic 6	Topic 5	Topic 7	Topic 6	Topic 5
R@6PM	E@9PM	S@4PM	R@6PM	E@9PM	S@4PM
R@7PM	E@6PM	S@7PM	R@8PM	E@10PM	S@4PM
R@8PM	E@10PM	S@4PM	R@7PM	S@10PM	S@7PM
R@12	E@10PM	S@12	R@1PM	E@6PM	S@11PM
R@1PM	E@8PM	S@11PM	R@12	E@8PM	S@12

Note: R, E, and S denote restaurant, entertainment, and shopping.

Precision. We binarize our five-level rating system ($4 > 3 > 2 > 1 > 0$) by treating the ratings ≥ 3 as "high-value" and ratings < 3 as "low-value". Given a top-N estate list E_N sorted in descending order of prediction values, the precision is defined as Precision@$N = \frac{|E_N \bigcap E_{\geq 3}|}{N}$, where $E_{\geq 3}$ are the estates whose ratings are greater or equal to 3.

Kendall's Tau Coefficient. Kendall's Tau Coefficient (or Tau for short) measures the overall ranking accuracy. Let us assume that each estate i is associated with a benchmark score y_i and a predicted score f_i. Then, an estate pair $\langle i, j \rangle$ is said to be concordant, if both $y_i > y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i < f_j$. Conversely, $\langle i, j \rangle$ is said to be discordant, if both $y_i < y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i > f_j$. Tau is given by Tau $= \frac{\#_{conc} - \#_{disc}}{\#_{conc} + \#_{disc}}$. Diversity is defined as $\sum_{f=1}^{K} \theta_f \frac{\sum_{m=1}^{M} y_m \frac{\sum_{h=1}^{m} \eta_{h,f}}{m}}{\sum_{m=1}^{M} y_m \eta_{m,f}}$. The larger diversity, the better.

Perplexity and Diversity. Perplexity and diversity are used to study parameter sensitivity, defined by $Perplexity = exp\left\{ - \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \log P(\boldsymbol{c}_{m,n}, \boldsymbol{t}_{m,n}, \boldsymbol{b}_{m,n})}{\sum_{m=1}^{M} \sum_{n=1}^{N} (|\boldsymbol{c}_{m,n}| + |\boldsymbol{t}_{m,n}| + |\boldsymbol{b}_{m,n}|)} \right\}$, and $Diversity = \sum_{f=1}^{K} \theta_f \frac{\sum_{m=1}^{M} y_m \frac{\sum_{h=1}^{m} \eta_{h,f}}{m}}{\sum_{m=1}^{M} y_m \eta_{m,f}}$.

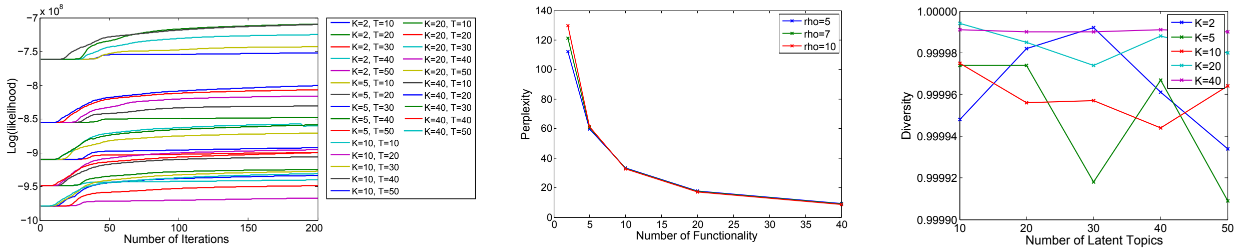
5.4 Evaluation of Geographical Functional Portfolio Learning

Next, we study our geographic functional learning model in terms of parameter sensitivity, temporal popular topics and patterns, and community functional portfolios.

(1) Study of Parameter Sensitivity.
Here, we investigate the sensitivity of different parameter settings in terms of three metrics: likelihood, perplexity, and diversity. Figure 5(a) plots the likelihood against the number of iterations. The likelihoods in all settings converge after 100 iterations. To ensure convergence, we retrieve all the results after 200 iterations. Figure 5(b) shows that the perplexity decreases as the number of functions decreases, in terms of different prior (ρ) settings. Since the trends of perplexity for different numbers of latent topics are similar, we only show the plots where $Q = R = W = 10$. Meanwhile, we notice that a smaller ρ results in a larger perplexity when K is small, and the perplexity gaps between different settings become small with the increase of K. Hence, we make a trade-off and set ρ to 7 in the following experiments. In addition, when K increases from 5 to 20, the perplexity decreases smoothly. Figure 5(c) shows that the differences among the diversities in all settings are not significant, and the number of latent topics is less related with diversity. Therefore, to avoid overfitting, we set $K = 5$, $Q = R = W = 7$, because the number of time periods for mobility segments is small (i.e., $N = 7$, one day per segment), and the sizes of vocabularies of checkin, taxi, and bus patterns are also small.

[1] `http://sourceforge.net/p/lemur/wiki/RankLib/`

[2] `http://ss.sysu.edu.cn/~py/fenchelcode.rar`

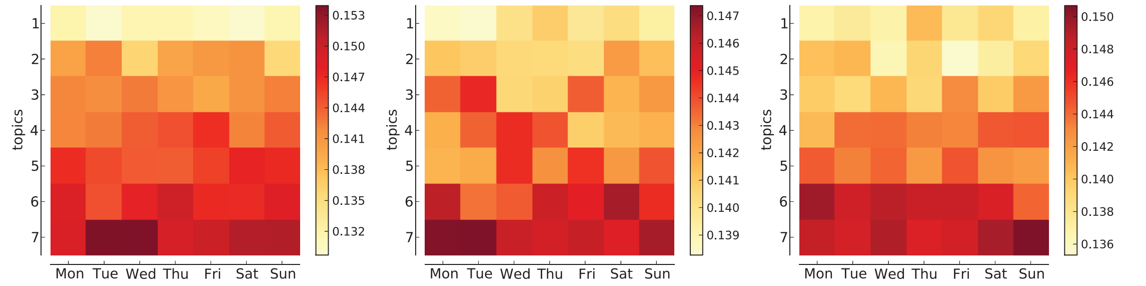


(a) No. of iterations vs. likelihood ($T = Q = R = W$)

(b) No. of functions vs. perplexity

(c) No. of latent topics vs. diversity

Figure 5: Sensitivity analysis of parameters.



(a) Checkin latent topics.

(b) Taxi latent topics.

(c) Bus latent topics.

Figure 6: Heatmaps of temporal popularity of checkin, taxi and bus latent topics during weekdays.

Table 5: Examples of temporal topics and their patterns of taxi mobility.

Weekday Topics				Weekend Topics	
Topic 6	Topic 7	Topic 3	Topic 4	Topic 6	Topic 7
L@6PM	A@6PM	L@5PM	A@8AM	L@6PM	A@6PM
A@8AM	A@8AM	A@8AM	L@5PM	A@8AM	A@8AM
A@5PM	L@8AM	L@7AM	L@6PM	A@5PM	L@8AM
A@6PM	L@5PM	L@6PM	L@8AM	A@6PM	L@5PM

Note: L and A denote leaving and arriving patterns respectively.

Table 6: Examples of temporal topics and their patterns of bus mobility.

Weekday Topics			Weekend Topics		
Topic 7	Topic 6	Topic 5	Topic 7	Topic 6	Topic 4
L@6PM	A@8AM	A@8AM	L@6PM	A@8AM	L@10PM
A@8AM	L@6PM	L@5PM	A@8AM	L@6PM	A@5PM
A@5PM	A@5PM	A@6PM	A@5PM	A@5PM	A@7PM
A@6PM	L@7AM	A@2PM	A@6PM	L@7AM	A@6PM
A@5PM	A@6PM	A@7AM	A@5PM	A@6PM	L@9PM

Note: L and A denote leaving and arriving patterns respectively.

(2) Study of temporal popularity of checkin, taxi, and bus latent topics.
We compute the topic distributions of checkin, taxi, and bus with respect to different week days. Figure 6 presents the topic distributions over seven days, with values represented by color darkness. We also list the representative words for these popular topics in Tables 4, 5, and 6, respectively. Figure 6 validates that the topic distribution of mobility has a temporal pattern. First, Figure 6(a) shows that checkin latent topics 1, 3, and 4 are popular during both weekdays and weekends. This is because topics 5, 6, 7 respectively represent shopping, entertainment, and catering activities at noon or at night, as shown in Table 4. Next, Figure 6(b) shows that taxi latent topics 3 and 4 are popular only during weekdays, while topics 4 and 6 are popular during both weekdays and weekends. From Table 5, we can see topics 3 and 4 generally include arriving patterns in the morning (i.e., go to work) and leaving patterns at night (i.e., leave after work), and thus mainly happen in weekdays. Topics 6 and 7 are combinations of both working activities (i.e., arriving early in the morning and leaving after 5PM) and catering, entertainment, and commercial activities (i.e., arriving after 5PM and leaving at night), and thus are popular during both weekdays and weekends. In addition, Table 6 shows that bus latent topics 6 and 7 include both working activities as well as catering, entertainment, and commercial activities, and thus cover both weekdays and weekends. On the other hand, bus latent topic 5 with only working activities is popular on weekdays. Bus latent topic 4 is mostly about recreation activities at night and is thus popular on weekends. The above analysis demonstrates that the geographic functional learning model can capture temporal patterns of checkin, taxi, and bus mobility.

(3) Study of functional distribution of high-ranked and low-ranked estates.
Here, we visualize the functional distribution of high-ranked and low-ranked estates, and study the correlation between real estate value and functional diversity. Figure 7 compares the functional distributions of high-ranked (i.e., top 1–25) and low-ranked (i.e., top 2505–2530) estates. High ranked estates generally show diverse and balanced distributions among different functions, Whereas low ranked estates show unbalanced distributions with low heterogeneity. This observation validates the assumption that a good functional portfolio can increase investment value.

5.5 Evaluation on Real Estate Ranking

Here, we report the evaluation results of our method, compared to baseline algorithms, on the rising and falling markets, in terms of NDCG, Precision, and Tau.

Rising Market. Figure 8 shows our method performs better than the baselines over top-k ranking in rising market.

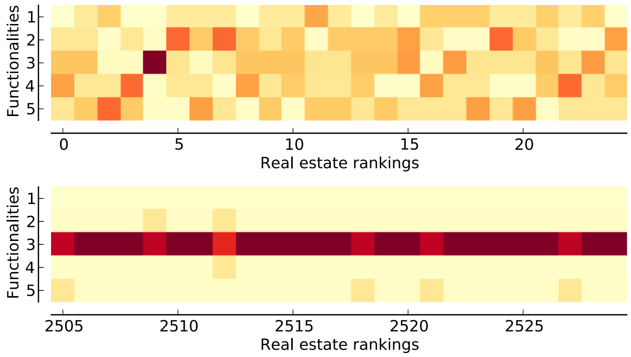


Figure 7: Comparison of functional distributions of high-ranked and low-ranked estates.

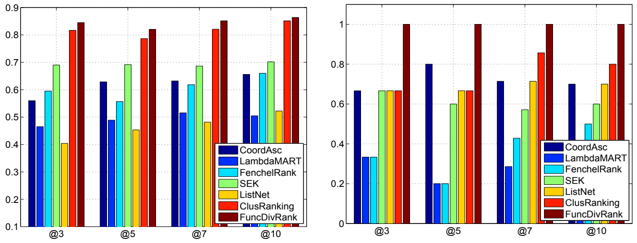
For example, our method offers 21%, 32.4%, 47.2% improvement in terms of NDCG@3 compared to SEK, FenchelRank, and RankBoost, respectively. Figure 8(b) shows that the top-K results ($K = 3, 5, 7, 10$) of our method consist almost exclusively of estates with rating ≥ 3. For example, *all* our top-10 results are high-value, compared to just 2 for random or CoordAsc ranking, and 7–8 for the best competitor.

Falling Market. As can be seen in Figure 9, our method outperforms the baselines over top-K ranking by a significant margin in falling market. Specifically, our method achieves 27.5%, 17.3%, and 99% improvement in terms of NDCG@3 compared to SEK, RankBoost, and FenchelRank, respectively. Unfortunately, we observe the overall ranking accuracy of our method decreases and is lower than ClusRanking and SEK. Finally, although our goal is to identify *top* investment opportunities, for completeness we also evaluate the total ranking of all estates, showing Tau scores in Table 7.

Next, we discuss how our work differs from previous work on real estate ranking. First, while ClusRanking [13] considers proximity and zone dependencies to capture pairwise ranking consistency, our method takes into account not only prediction accuracy and ranking consistency, but also the impact of mixed land use (i.e., functional diversity). As a result, we can better capture the ranking of the list of estates. Indeed, we observe a significant improvement in top-K ranking over classic LTR methods. Second, we exploit random forests to generate meta features from raw features. Third, although SEK [12] includes the entropy of POI distribution as one of the features, its predictive power may be diluted by the large number of other extracted features. In contrast, our method can emphasize the functional diversity directly in the ranking objective.

6. RELATED WORK

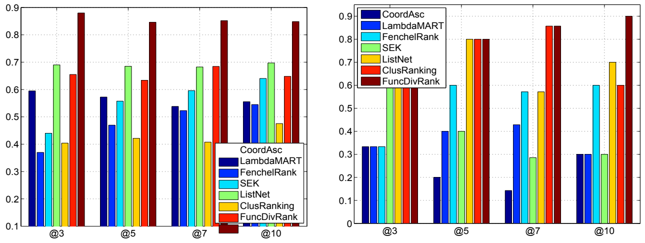
Real Estate Appraisal and Ranking. Traditional research on estate appraisal is based on financial real estate theory, typically constructing an explicit index of estate value [20], for example, price to income ratio. Some studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of estate prices [7, 10]. More classic works are based on repeat sales methods and hedonic methods [1, 29, 17, 32]. The work in [10] studies the automated valuation models which aggregate and analyze physical characteristics and sales prices of comparable properties, to provide property valuations. The work in [12] extracts features from user reviews and mobility behaviors and integrates sparsity regularization into pairwise estate ranking. The work in [13] jointly models the geographical individual, peer, and zone dependencies for enhancing prediction of estate investment value. More recent works [18] apply general additive mode, support vector machine regression, and multilayer perceptron ensembles for computational estate appraisal.



(a) NDCG@N

(b) Prec@N

Figure 8: Performance comparison, rising market.



(a) NDCG@N

(b) Prec@N

Figure 9: Performance comparison, falling market.

Learning To Rank with Diversity. Also, our work is related to LTR. The pair-wise methods, such as RankNet [4], RankBoost [11], RankSVM [15], and LambdaRank [26], reduce the LTR task to a classification problem. The goal of the pairwise ranking is to learn a binary classifier to identify the better document in a given document pair by minimizing the average number of rank inversions. Works [33, 27] provide full Bayesian explanations and optimize the posterior of point-wise and pair-wise ranking models, respectively. Study [28] unifies both rating error and ranking error as objective function to enhance Top-K recommendation. More recent works [35, 31, 25] study diversified learning to rank. For example, [35] ranks items by random walks in an absorbing Markov chain and achieves both diversity and centrality. The work in [31] proposes a diversified ranking objective by incorporating subtopics into MAP (Mean Average Precision) for expert finding.

Urban Computing and Site Selection. Our work also has a connection with mining of mobile, geographical, and mobility data, to tackle issues in the urban space. Yuan et al. discover regional functions of a city using POIs and taxi traces [34] . Work [16] selects the optimal sites for retail stores by mining Foursquare data. Also, our work is related to measuring similarity for ranking [9, 8].

7. CONCLUDING REMARKS

Summary. In this paper, we investigated how to rank real estate investment values by considering the impact of

Table 7: The Tau values of different algorithms in rising and falling markets.

Period	CoordAsc	LambdaMART	FenchelRank	SEK	ListNet	ClusRanking	FuncDivRank
Rising Market	-0.1370415	0.07150473	0.1224318	0.3493753	0.1722723	0.3428617	0.350517
Falling Market	0.223312	0.2311301	-0.124769	0.3347548	0.0538088	0.2363498	-0.09250678

mixed land use, which can be reflected by diverse community functions. Since human mobility patterns provide a reasonable estimation of diverse functions present in the community of an estate, we developed a latent factor model to learn the portfolio of community functions for real estate from human mobility data. Then, we designed a unified probabilistic framework which allows simultaneous maximization of ranking consistency and of functional diversity for real estate ranking. Finally, we conducted extensive experiments on real-world human mobility data, urban geographical data, and user check-in data collected from location based social networks. As revealed in the experimental results, a diverse view of mixed land use can help to better capture real estate values and the performance improvement of our proposed method is substantial compared to benchmark methods.

Discussion. This paper focused on assessing the investment ratings of residential complexes in urban areas of big cities, whose developing strategy is mixed land using, for business site selection. In different cities, buyers may have personalized expectations on functional diversity, the method of incorporating functional diversity can be further enhanced for personalized real estate recommendation.

Acknowledgment. This research was partially supported by National Science Foundation (NSF) via the grant number CCF-1018151. Also, it was supported in part by Natural Science Foundation of China (71028002). In addition, this research was supported in part by National Institutes of Health under Grant 1R21AA023975-01.

8. REFERENCES

[1] M. Bailey, R. Muth, and H. Nourse. A regression method for real estate price index construction. *J. Am. Stat. Assoc.*, 58:933–942, 1963.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] C. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 2010.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML'05*, 2005.

[5] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML'07*, 2007.

[6] K. E. Case and R. J. Shiller. The behavior of home buyers in boom and post-boom markets, 1988.

[7] L. D. B. Chaitra H. Nagaraja and L. H. Zhao. An autoregressive approach to house price modeling, 2009.

[8] S. Chang, C. C. Aggarwal, and T. S. Huang. Learning local semantic distances with limited supervision. In *2014 IEEE International Conference on Data Mining*, pages 70–79, 2014.

[9] S. Chang, G. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang. Factorized similarity learning in networks. In *2014 IEEE International Conference on Data Mining*, 2014.

[10] M. L. Downie and G. Robson. Automated valuation models: an international perspective. 2007.

[11] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 2003.

[12] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, and N. J. Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors. In *the 14th IEEE International Conference on Data Mining (ICDM 2014)*, 2014.

[13] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of clustering and ranking. In *KDD'14*, 2014.

[14] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. n. Candela. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, ADKDD'14, 2014.

[15] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*, 1999.

[16] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *KDD '13*, 2013.

[17] R. H. Knight, J.R. and C. Sirmans. Biased prediction of housing values. *Journal of the American Real Estate and Urban Economics Association*, 20:427–456, 1992.

[18] V. Kontrimas and A. Verikas. The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 11:443 – 448, 2011.

[19] H. R. Koster and J. Rouwendal. The impact of mixed land use on residential property values*. *Journal of Regional Science*, 52(5):733–761, 2012.

[20] J. Krainer and C. Wei. House prices and fundamental value. *FRBSF Economic Letter*, 2004.

[21] H. Lai, Y. Pan, C. Liu, L. Lin, and J. Wu. Sparse learning-to-rank via an efficient primal-dual algorithm. *Computers, IEEE Transactions on*, 2013.

[22] S. Loehr. Mixed-use, mixed impact: Re-examining the relationship between non-residential land uses and residential property values. 2013.

[23] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 2007.

[24] R. K. Pace. Appraisal using generalized additive models. *Journal of Real Estate Research*, 15:77–100, 1998.

[25] L. Qin and X. Zhu. Promoting diversity in recommendation by entropy regularizer. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2698–2704. AAAI Press, 2013.

[26] C. Quoc and V. Le. Learning to rank with nonsmooth cost functions. *NIPS'07*, 2007.

[27] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI '09*, 2009.

[28] Y. Shi, M. Larson, and A. Hanjalic. Unifying rating-oriented and ranking-oriented collaborative filtering for improved recommendation. *Information Sciences*, 2012.

[29] R. J. Shiller. Arithmetic repeat sales price estimators. Technical report, Cowles Foundation for Research in Economics, Yale University, 1991.

[30] Y. Song and G.-J. Knaap. Measuring the effects of mixed land uses on housing values. *Regional Science and Urban Economics*, 34(6):663–680, 2004.

[31] H. Su, J. Tang, and W. Hong. Learning to diversify expert finding with subtopics. In *Advances in Knowledge Discovery and Data Mining*, pages 330–341. Springer, 2012.

[32] L. O. Taylor. The hedonic method. In *A primer on nonmarket valuation*. Springer, 2003.

[33] R. C. Weng and C.-J. Lin. A bayesian approximation method for online ranking. *The Journal of Machine Learning Research*, 2011.

[34] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *KDD'12*, 2012.

[35] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, pages 97–104. Citeseer, 2007.