REAL ESTATE RANKING: FROM BLACK MAGIC TO DATA SCIENCE

by

YANJIE FU

A Dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

written under the direction of

Dr. Hui Xiong

and approved by

_____

_____

_____

_____

Newark, New Jersey

October 2016

ABSTRACT OF THE DISSERTATION

REAL ESTATE RANKING: FROM BLACK MAGIC TO DATA SCIENCE

By YANJIE FU

Dissertation Director: Dr. Hui Xiong

With the advent of mobile, Internet, and sensing technologies, large-scale urban and mobile data are available and are linked with locations near real properties. These data can be a source of rich intelligence for classifying high-rated residential locations, developing livable communities, and enhancing urban planning in big cities. In this dissertation, we aim to address the unique challenges of real estate ranking, especially (i) how to build an effective ranking system by exploiting heterogeneous mobile data and modeling geographic dependencies; (ii) what are the underlying drivers for livable and sustainable communities.

Along these lines, I first introduced a method for ranking residential complexes based on invest- ment ratings by mining users opinions about residential complexes from online user reviews and offline moving behaviors (e.g., taxi traces, smart card transactions, check-ins). While a variety of features could be extracted from these data, these features are intercorrelated and redundant. Thus, selecting good features and integrating the feature selection into the fitting of a ranking model are essential. To this end, I first strategically mined the fine-grained discriminative features from user reviews and moving behaviors. Then, I proposed a Sparse Pairwise Ranking method by combining a pairwise ranking objective and a sparsity regularization in a unified probabilistic framework.

In addition, with the development of new ways to collect estate-related mobile data, there is a potential to leverage geographic dependencies of residential complexes for enhancing real estate evaluation. Indeed, the geographic dependencies of the value of a residential complex can be from the characteristics of its own neighborhood (individual), the values of its nearby residential complexes (peer), and the prosperity of the affiliated latent business area (zone). To this end, I proposed an enhanced method, named ClusRanking, for real estate evaluation by leveraging the mutual enforcement of ranking and clustering power. In ClusRanking, three influential factors (i.e., geographic utility, neighborhood popularity, and influence of business areas) are constructed and extracted for predicting real estate investment ratings. An estate-specific ranking objective is also proposed to jointly model individual, peer and zone dependencies.

Moreover, 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 residential complexes 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, I further developed a geographical function ranking method, named FuncDivRank, by incorporating the functional diversity of communities into real estate evaluation. In FunDivRank, a mix-land use latent model is developed to learn latent community functions and the corresponding portfolios. Also, a real estate ranking indicator is learned by simultaneously maximizing ranking consistency and functional diversity.

Finally, we present experimental results to demonstrate the effectiveness of our methods.

ACKNOWLEDGEMENTS

I would like to express my great appreciation to all the people who provided me tremendous support and help during my Ph.D. study.

First, I would like to express my deep gratitude to my advisor, Prof. Hui Xiong, for his continuous support, guidance and encouragement, which are necessary to survive and thrive the graduate school and the beyond. I thank him for generously giving me motivation, support, time, assistance, opportunities and friendship; for teaching me how to identify key problems with impact, present and evaluate the ideas. He helped making me a better writer, speaker and scholar.

I also sincerely thank my other committee members: Prof. Rui Kuang, Prof. Jian Yang, and Prof. Spiros Papadimitriou. All of them not only provide constructive suggestions and comments on my work and this thesis, but also offer numerous support and help in my career choice, and I am very grateful for them. Prof. Spiros Papadimitriou has been a great professor to me over the past three years. His experience in machine learning and data mining has inspired me a lot to solve the challenging problems in my research, and I have learned a great deal from the collaboration with him on many exciting projects.

I wrote my first paper during my internship in Microsoft Research Asia under the supervision of Dr. Yu Zheng and Dr. Xing Xie, who open a window for me toward

urban computing. My Ph.D. study was financially supported by Futurewei Technology, and thereby special thanks go to Dr. Jin Yang, Dr. Nandu Gopalakrishnan, and Dr. Yan Xin. They inspired me with another new area: wireless big data analytics. Finally, I particularly enjoyed my internship at IBM Thomas J. Watson Research Center. At this great research center, I met numerous outstanding researchers, and therefore I sincerely thank Dr. Deepak S. Turaga, Dr. Srinivasan Parthasarathy, and Dr. Charu Aggarwal for their generous guidance.

Special thanks are due to Prof. Wenjun Zhou at University of Tennessee, Prof. Ge Yong at University of Arizona, Prof. Zhi-hua Zhou at Nanjing University. It was a great pleasure working with all of them. I also owe a hefty amount of thanks to my colleagues and friends Bin Liu, Zijun Yao, Meng Qu, Konstantin Patev, Jingyuan Yang, Can Chen, Hao Zhang, Farid Razzak, Qingxin Meng, Xinjiang Lu, Yayao Zuo, Yanhong Guo, Leilei Sun, Huang Xu, Jie Liu, Bowen Du, Chuanren Liu, Chu Guan, Jiadi Du, Xiaolin Li, Ling Yin, Hongting Niu, Jing Sun, Zhongmou Li, Guannan Liu, Tong Xu, Bo Jin, Yuanchun Zhou, Keli Xiao, Qi Liu, Liyang Tang, Hengshu Zhu, Xue Bai, Chang Tan, Aimin Feng, Yanchi Liu, Chunyu Luo, Qi Liu for their help, friendship and valuable suggestion.

I would like to acknowledge the Department of Management Science and Information Systems (MSIS) and Center for Information Management, Integration and Connectivity (CIMIC) for supplying me with the best imaginable equipment and facilities that helped me to accomplish much of this work.

Finally, I would like to thank my family for their love, support and understanding. Without their encouragement and help, this thesis would be impossible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Preliminaries

Today, things are more connected. There are new emerging trends in urban areas: (i) more sensors are installed to sense the pulses of our cities and residents; (ii) Internet and cloud technologies meet and rejuvenate traditional industries (e.g., logistics, agriculture, finance) via the Internet Plus strategy; (iii) more transactions and events happen on mobile devices. With the advent of mobile, Internet, and sensing technologies, large-scale urban and mobile data are available and are linked with locations near real properties. These data can be a source of rich intelligence for classifying high-rated residential locations, developing livable communities, and enhancing urban planning in big cities. In this dissertation, we aim to address the unique challenges of real estate ranking, especially (i) how to build an effective ranking system by exploiting heterogeneous mobile data and modeling geographic dependencies; (ii) what are the underlying drivers for livable and sustainable communities by exploring heterogeneous human mobility.

1.2 Research Challenges and Contributions

However, it is not easy to achieve this goal. There are two major challenges. First, prior literature in housing price appraisal regards each residential location as a product. These methods mainly consider price information, coarse-grained location information, and basic building information. However, they might not consider fine-grained urban geography data such as Point of Interests, public transportations, and dynamic human mobility patterns. Therefore, we are the first to bring these fine-grained urban geography data and dynamic human mobility patterns. Second, once we bring in these heterogeneous urban and mobile data, these data make the modeling of ranking difficult. In particular, we need to address three modeling questions: (i) how to fuse heterogeneous information for ranking; (ii) how to model geographic dependencies for ranking; (iii) how to explore mobility patterns for ranking. More importantly, through the modeling of ranking, we uncover the underlying driver of livable and sustainable communities: a balance mix of land uses.

1.3 Overview

Chapter 2 presents a sparse ranking method for fusing heterogeneous urban and mobile data into a pairwise ranking indicator.

Chapter 3 presents a geographic ranking method by jointly modeling geographic individual, peer, and zone dependencies via the mutual enhancement of ranking and clustering.

Chapter 4 presents a mobility ranking method by exploring the impact of mixed land use via learning optimal portfolios of community functions from heterogeneous

mobility patterns.

Chapter 5 presents conclusion remarks and future work.

CHAPTER 2

EXPLOITING HETEROGENEOUS INFORMATION FUSION FOR REAL ESTATE RANKING

Ranking residential real estates based on investment values can provide decision making support for home buyers and thus plays an important role in estate marketplace. In this chapter, we aim to develop methods for ranking estates based on investment values by mining users opinions about estates from online user reviews and offline moving behaviors (e.g., taxi traces, smart card transactions, check-ins). While a variety of features could be extracted from these data, these features are intercorrelated and redundant. Thus, selecting good features and integrating the feature selection into the fitting of a ranking model are essential. To this end, in this chapter, we first strategically mine the fine-grained discriminative features from user reviews and moving behaviors, and then propose a probabilistic sparse pairwise ranking method for estates. Specifically, we first extract the explicit features from online user reviews which express users opinions about point of interests (POIs) near an estate. We also mine the implicit features from offline moving behaviors from multiple perspectives (e.g., direction, volume, velocity, heterogeneity, topic, popularity, etc.). Then we learn an estate ranking predictor by combining a pairwise ranking objective and a sparsity regularization in a unified probabilistic framework. And we develop an effective solution for the optimization problem. Finally, we conduct a comprehensive

performance evaluation with real world estate related data, and the experimental results demonstrate the competitive performance of both features and the proposed model.

2.1 Introduction

There are several definitions of estate value according to International Valuation Standards [1] . For instance, market value is defined as the price at which an estate would trade in a competitive Walrasian auction setting. Another example is investment value, which is the value of an estate to one particular investor and may or may not be higher than the market value of the estate. Difference between the investment value and the market value for a particular estate provides the motivation for buyers or sellers to enter the estate marketplace. Thus, providing a ranking of estates based on investment values will greatly help buyers make their purchase decisions.

Which estates have high investment values? While estate industry professionals have used different housing indexes (e.g., price-rent ratio) to approximate the fundamental value of estates, researchers have also used financial time series analysis to investigate the trend, periodicity and volatility of estate prices and assess estate investment potentials (Downie & Robson, 2007; Chaitra H. Nagaraja & Zhao, 2009). Recent studies have tried to correlate the estate value to the static statistics of urban infrastructure (e.g., the numbers of POIs, the distances to bus stops), because they explicitly reflect the physical facilities of a neighborhood (Taylor, 2003; Fu, Xiong, et al., 2014). However, infrastructure statistics is not sufficient for evaluating invest-

[1] http://www.ivsc.org/

ment values of estates. Considering the distance to public transit, while an estate near public transit usually leads to high rent and sale price in many cities, there is also possible negative effect when living nearby public transit. For example, the noise and pollution associated with train/bus systems can lower the value of an estate as reported in (Landis, Guhathakurta, Huang, Zhang, & Fukuji, 1995; Bowes & Ihlanfeldt, 2001; Lewis-Workman & Brod, 1997). Thus, there is some limitation for using these infrastructure statistics. Moreover, these statistics are often lack of dynamics and hardly reflect the changing pulses of a city.

On the contrary, there are more estate-related dynamic and information-rich data which has been accumulated with the development of mobile, internet and sensor technologies. For example, people may post comments and ratings for POIs (e.g., schools, restaurants and shopping centers, etc.) via mobile apps after their consumptions. Also, the mobility data, such as smart card transactions and taxi GPS traces, comprise both trajectories and consumption records of residents' daily commutes. People's check-ins may reflect the popularity of POIs. If properly analyzed, these data (e.g., user reviews, location traces, smart card transactions, check-ins, etc.) can be a rich source of intelligence for discovering estates of high investment-value.

Indeed, these estate-related dynamic data generated by users could better reflect investment values of estates than urban infrastructure statistics. Generally speaking, if people have better opinions for an estate, the demand for this estate is higher and its investment value will be higher. The challenge is how to uncover people's opinions for an estate. In fact, the opinions of users for an estate can be mined from (1) online user reviews and (2) offline moving behaviors. Specifically, the online reviews (e.g.,

Zagat/Yelp ratings) contain the explicit opinions for places surrounding an estate. For example, the quality of neighborhood can be partially approximated by the ratings of business venues, such as overall rating, service rating, environment rating, etc. Meanwhile, the offline moving behaviors near an estate not only encode the static statistics of urban infrastructure, but also reflect the implicit "opinions" of residents for a neighborhood. For example, the arriving, transition, and leaving volumes of taxies and buses imply the mobility density of a neighborhood; the average velocity of taxies and buses indicates the degree of traffic congestion or accessibility; the daily frequency of check-ins shows regional popularity and prosperity; the heterogeneity of distributions of check-ins over categories reflects if the facility planning is balanced or not. All these indications by the estate-related dynamic user-generated data comprise the important facets of an estate that home buyers care very much and convey the implicit "opinions" of users for a neighborhood. Therefore, we consider and mine both the explicit opinions from user reviews and the implicit opinions from moving behaviors to enhance the evaluation of estate investment value.

Although we may extract a lot of features from the variety of data sources, these extracted estate-related features usually are correlated and redundant. The feature redundancy results in poor generalization performance. In reality, a small number of good features can determine the ranking of estates based on investment values. Therefore, we explore the sparse learning technique for the ranking of estates. However, classic sparse learning methods use a two-step paradigm, which is basically to first select a feature subset and then learn a ranking model based on the selected features. But the selected feature subset may not be optimal for ranking because the

two steps are modelled separately. In contrast, combining sparsity and ranking in a unified model can help to identify the optimal feature subset for better learning an estate ranker, and also have less computational cost in prediction.

Along this line, in this chapter, we propose to mine opinions of mobile users and explore the learning-to-rank with sparsity for the investment value based estate ranking. We consider and explore both explicit and implicit opinions that reflect estate investment value by mining online user reviews and offline moving behaviors. Specifically, to capture the opinions of mobile users toward estates, we extract the explicit features from user reviews to reveal user satisfaction of estate neighborhoods. Besides, we measure the traffic volumes with respect to different directions, traffic velocity, functionality heterogeneity, neighborhood popularity, topical profile of estate neighborhoods by mining multi-type mobility data including taxi traces, smart card transactions and check-ins. Moreover, we learn a linear ranking predictor by combining pairwise ranking objective and sparsity regularization in a unified probabilistic framework, which is greatly enhanced by simultaneously conducting feature selection and maximizing estate ranking accuracy. Finally, we conduct comprehensive performance evaluations for the feature sets and models with large-scale real world data and the experimental results demonstrate the competitive performance of our method with respect to different validation metrics.

2.2 Sparse Estate Ranking

In this section, we present the proposed system of sparse estate ranking, namely SEK.

2.2.1 The Overview of Sparse Estate Ranking

As shown in Figure 2.1, our estate ranking system consists of two major components: (1) estate feature extraction and (2) sparse estate ranking.

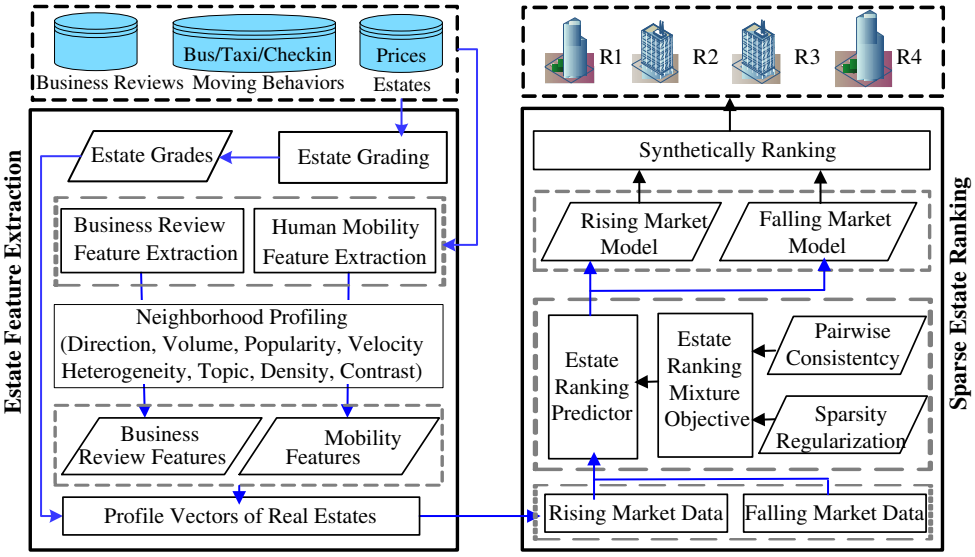


Figure 2.1. The framework of the proposed system.

Estate Feature Extractions: As shown in Figure 2.1, we first collect historical prices of each estate, compute the return rates [2] of estates and grade estates into five bins/levels in terms of investment returns to prepare labels for training data. The discretization of the estate returns is important because the small difference between estate values in the same value category might be noisy for the ranking model.

Specifically, we first calculate the average estate price of a city for each month. For instance, Figure 4.4 shows the trend of the average estate prices in Beijing. We can see an inflection point in the curve. The point is used to split the time period into two phases, i.e., the rising phrase (from Feb. 2012 to Sept. 2012) and the falling

[2] http://financial-dictionary.thefreedictionary.com/rate+of+return

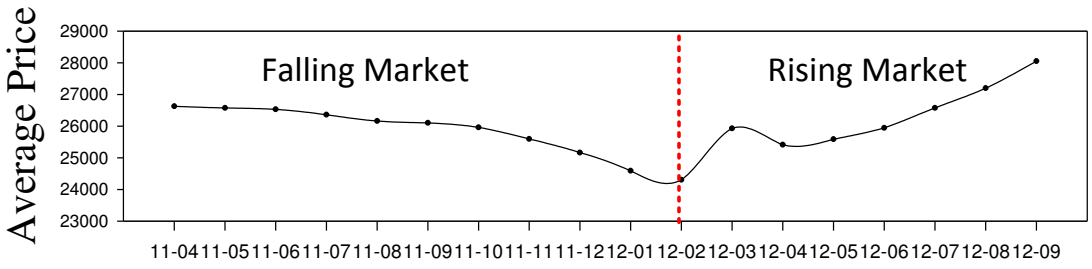


Figure 2.2. The rising market period and the falling market period in Beijing.

phrase (from Apr. 2011 to Feb. 2012). We then sort estates in rising phase and falling phrase according to their investment returns in the decreasing order as shown in Figures 2.3 (a) and (d), where the horizontal axis is the order of an estate in the sorted list and the vertical axis represents return rates. As can be seen, the prices of a small number of estates significantly increase or decrease whereas many estates' prices remain stable. In fact, these distributions indicate the power law distribution for estate investment returns. After computing the second order derivatives of these two curves, we find out four inflection points, which show the significant change of return rates as shown in Figures 2.3 (b) and (e). As a result, we obtain five rating levels for the rising and falling phrases as shown in Figures 2.3 (c) and (f).

Next we aim at extracting the features from online user reviews and offline moving behaviors such as taxi traces, smart card transactions, check-ins as shown in Table 4.2. The features from user reviews are summarized by spatial statistics and the features from moving behaviors are derived from multiple angles (e.g., direction, volume, velocity, heterogeneity, topic, contrast, popularity).

Sparse Estate Ranking: We learn a linear ranking predictor by combining a pairwise ranking objective and a sparsity regularization together. By optimizing the

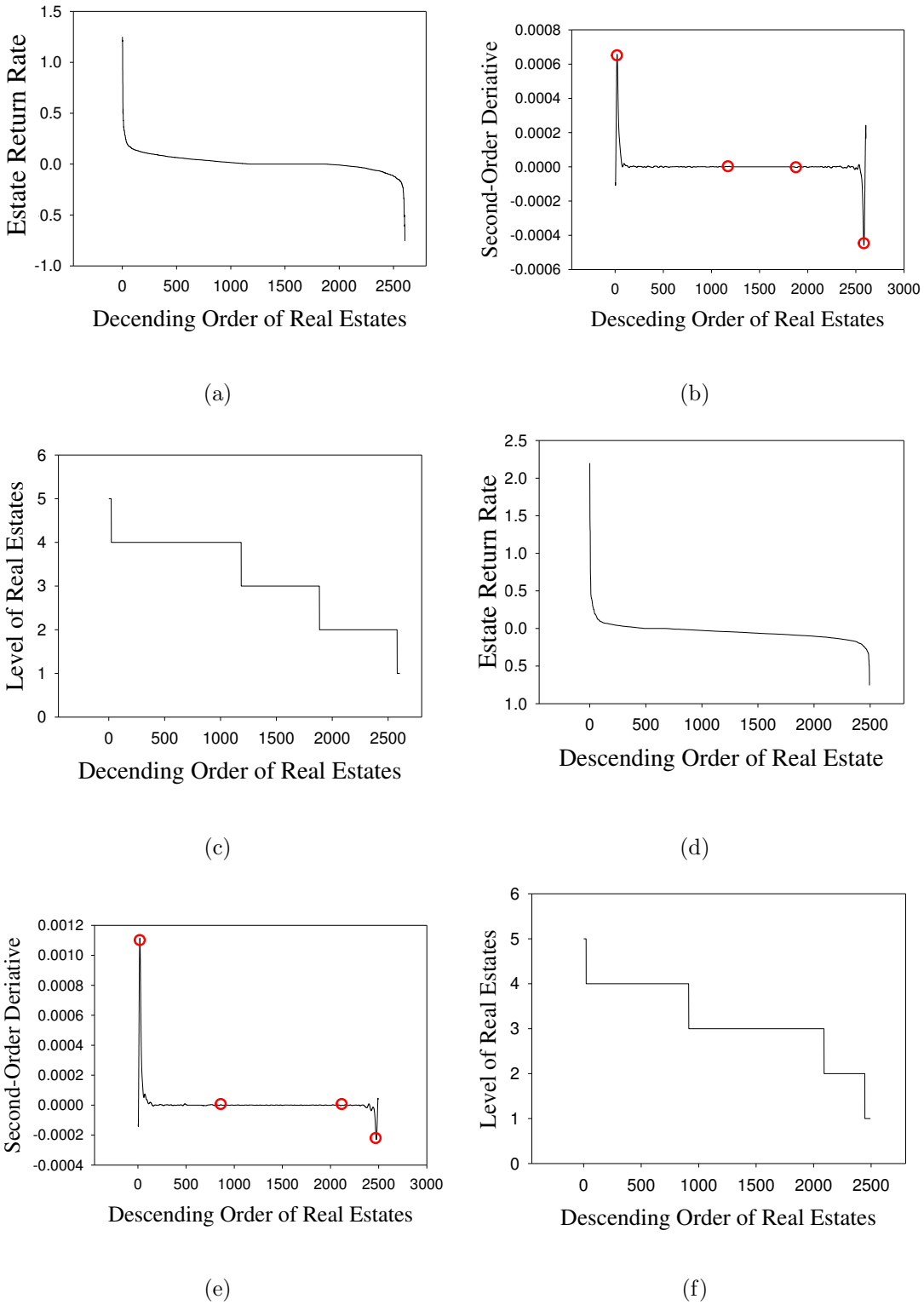


Figure 2.3. The grading process of estates.

Table 2.1. The extracted features.

Online User Reviews	Offline Moving Behaviors		
User Reviews	**Taxi**	**Bus**	**Check-in**
Overall Salification	Arriving Volume	Arriving Volume	Popularity
Service Quality	Leaving Volume	Leaving Volume	Topic
Environment Class	Transition Volume	Transition Volume	
Consumption Cost	Driving Velocity	Bus Stop Density	
Functionality Planning	Commute Distance	Smart Card Balance	

overall objective function, we learn the estate ranker by simultaneously conducting feature selection and maximizing ranking accuracy. Two separated models are then built to infer the value-adding and value-protecting ability of an estate in a rising and a falling market respectively. Given a set of estates specified by a user, we extract the features in the same way as we show in Figure 2.1. Since we do not know whether the market will go up or down, the extracted features are fed into two ranking models respectively to produce the potential ranks of these estates at the current time. Finally, we generate a final score for an estate by aggregating the ranking outputs of these two models.

2.2.2 Estate Feature Extraction

Rather than simply considering the static statistics of urban infrastructure (e.g., the numbers of POIs, the distances to bus stops), we introduce the fine-grained features we have extracted from online users reviews and offline moving behaviors for estate

ranking.

Explicit Features from Online User Reviews

Both prosperity and users' opinion of neighborhood are two important factors determining property investment value. Recent study (Wardrip, 2011) shows that a strong regional economy usually indicates high housing demand. (b. Hj. Mar Iman al Murshid, 2008) further points out the word-of-mouth reflects the satisfaction of people toward the quality of a neighborhood. We thus consider to mine the online user reviews of Beijing collected from www.dianping.com. More specifically, for each estate e_i, we measure (1) overall satisfaction, (2) service quality, (3) environment class, (4) consumption level, and (5) functionality planning of the neighborhood r_i by mining the reviews of business venues located in r_i, $\{p : p \in P \& p \in r_i\}$ in which P is the set of business venues in Beijing.

Overall Satisfaction: For each estate e_i, we access the overall satisfaction of users over the neighborhood r_i. Since the overall rating of a business venue p represents the satisfaction of users, we extract the average of overall ratings of all business venues located in r_i as a numeric score of overall satisfaction. Formally we have:

$$f_i^{OS} = \frac{\sum_{p \in P \& p \in r_i} OverallRating_p}{|\{p : p \in P \& p \in r_i\}|}. \tag{2.1}$$

Service Quality: Similarly, we compute the average of service rating of business venues in r_i and represent the service quality of the neighborhood of e_i by

$$f_i^{SQ} = \frac{\sum_{p \in P \& p \in r_i} ServiceRating_p}{|\{p : p \in P \& p \in r_i\}|} \tag{2.2}$$

Environment Class: The environment class of business venues could reflect whether the neighborhood is high-class or not. Therefore, we extract the average environment

ratings as

$$f_i^{EC} = \frac{\sum_{p \in P \& p \in r_i} EnvironmentRating_p}{|\{p : p \in P \& p \in r_i\}|} \tag{2.3}$$

Consumption Cost: Average costs of consumption behaviors in business venues can partially reflect the salary income and neighborhood class. We calculate the average consumption cost of business venues of a targeted neighborhood as a feature.

$$f_i^{CC} = \frac{\sum_{p \in P \& p \in r_i} AverageCost_p}{|\{p : p \in P \& p \in r_i\}|} \tag{2.4}$$

Functionality Planning: A competitive neighborhood usually provides convenient access to diverse facilities, such as living demands (e.g., restaurants, supermarkets, and hospitals), education demands (e.g., schools and libraries), safety demands (e.g., police and fire department) and entertainment demands (e.g., theaters and parks), so that it meets various demands of residents. Shortage of diverse facility would reduce estate investment value. High facility diversity of a neighborhood helps to enhance the attractiveness of its estates. This effect is called mixed/diverse land use which plays an important role in metropolitan realty market. We therefore investigate the distribution of POIs over categories in each neighborhood. A high-class neighborhood is expected to provide balanced and heterogeneous categories of facilities. Hence, we apply an entropy to measure the functionality heterogeneity of a neighborhood. Let $\#(i, c)$ denotes the number of business venues of category $c \in C$ located in r_i, $\#(i)$ be the total number of business venues of all categories located in r_i. The entropy is defined as

$$f_i^{FP} = -\sum_{c \in C} \frac{\#(i, c)}{\#(i)} \times log \frac{\#(i, c)}{\#(i)} \tag{2.5}$$

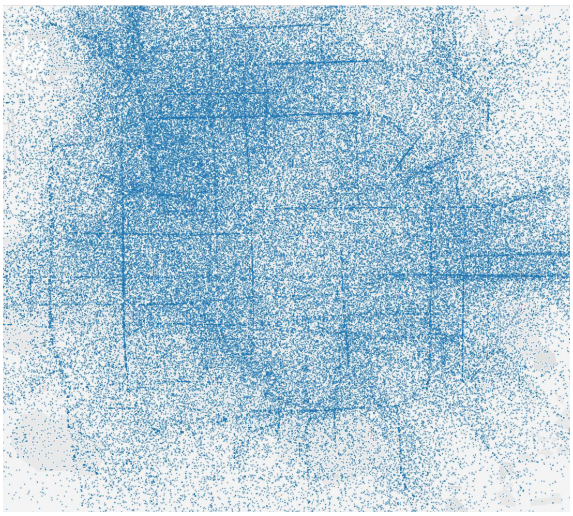
Implicit Features from Offline Moving Behaviors

Recent study (Wardrip, 2011) reports different types of transit systems (e.g., taxi, bus) have different impacts on estate values due to their different fares, frequencies, speeds, and scopes of service. Figure 2.4 (a), (b) and (c) show the density distribution of three types of moving behaviors respectively (i.e., taxi, bus and check-in) in Beijing. Taxi transits are fast, expensive and mainly distributed in central business district (CBD) and financial areas. Bus transits are slow, cheap and mainly distributed in information technology (IT) and education areas. Check-ins reflect a broad range of mobility and are mainly distributed in areas full of attractions, entertainments, and POIs. Since different moving behaviors reflect different geographic preferences and social classes of mobile users, we exploit these three types of moving behaviors to uncover the implicit preference of mobile users toward a neighborhood.

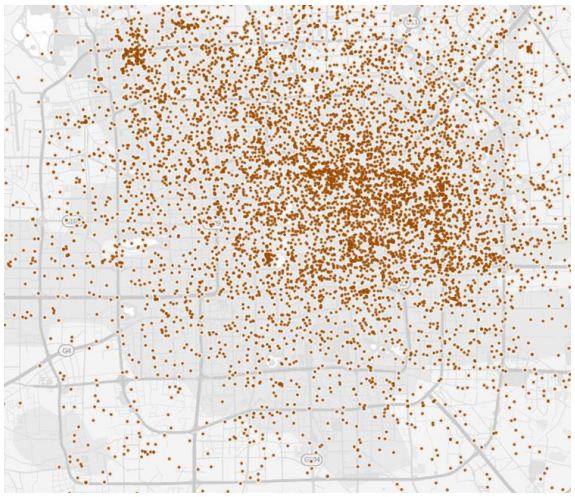
Taxi-Related Features. Recent study (Wardrip, 2011) suggests that the ability to travel within a large metropolitan area in a short time, for example, by taxi, is highly valued by residents. To extract the taxi related features, we measure the arriving volume, leaving volume, transition volume, driving velocity and commute distance of a neighborhood using taxi GPS traces. Let TT denote the set of all taxi trajectories of Beijing, each of which represents a taxi trajectory, denoted by a tuple $< p, d >$ where p is a pickup point and d is a drop-off point.

Taxi Arriving, Leaving and Transition Volume: According to (Wardrip, 2011), most affluent homeowners expect time-saving commute to white-collar jobs downtown and value faster taxies access.Therefore, the arriving, leaving, and transition volumes of

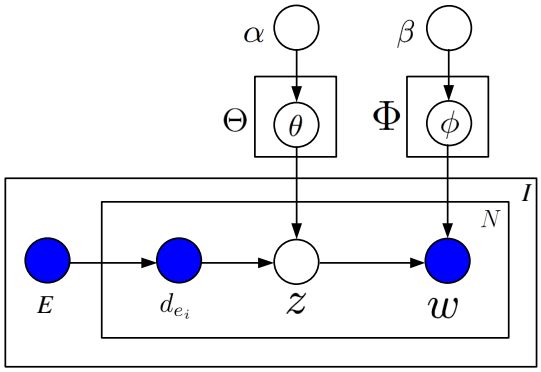
(a) Taxi drop-off points



(b) Bus drop-off points



(c) Check-ins



(d) Estate topic profiling

Figure 2.4. (a), (b), and (c) respectively show spatial distribution of taxi drop-offs, bus drop-offs and check-ins; (d) illustrates the process of estate topic profiling using the associated word-of-mouth from check-ins.

taxi mobility reflect the income and social class of residents of the targeted neighborhood. We define a feature as the counted taxi arriving volume of external passengers toward the targeted neighborhood. Formally, the taxi arriving volume is given by

$$f_i^{TAV} = |\{< p, d >\in TT : p \notin r_i \& d \in r_i\}| \tag{2.6}$$

Similarly, we define a feature as the counted taxi leaving volume from the targeted neighborhood to external venues. Formally, the taxi leaving volume is defined as

$$f_i^{TLV} = |\{< p, d >\in TT : p \in r_i \& d \notin r_i\}| \tag{2.7}$$

We also define a feature as the taxi transition volume between different venues inside the targeted neighborhood. Formally,

$$f_i^{TTV} = |\{< p, d >\in TT : p \in r_i \& d \in r_i\}| \tag{2.8}$$

Taxi Driving Velocity: According to (Wardrip, 2011), the value of increased travel velocity and reduced traffic congestion should be reflected in home values. We investigate the average taxi velocity of the neighborhood of each estate, namely f_i^{TDV}. Usually, the taxi speed of a neighborhood indicates the accessibility of road network and transportation efficiency. Formally, f_i^{TDV} is given by

$$f_i^{TDV} = \frac{\sum_{p \in r_i \& d \in r_i} dist(p, d)/time(p, d)}{|\{< p, d >\in TT : p \in r_i \& d \in r_i\}|} \tag{2.9}$$

Taxi Commute Distance: Taxi is a kind of expensive but fast transit. Normally, passengers take taxi to the important places (e.g., work place, theater, hotel, etc.) for business or urgent purposes. The shorter distance an estate neighbor is from important places, the more prosperous the neighborhood is, and the higher commute convenience the neighborhood has. A huge part of motivations of trading an estate comes from the incentive of convenient living environment. Formally, the taxi

commute distance is defined by

$$f_i^{TCD} = \frac{\sum_{p \in r_i || d \in r_i} dist(p, d)}{|\{< p, d > \in TT : p \in r_i || d \in r_i\}|} \tag{2.10}$$

Bus-Related Features. Most of moderate-income residents choose buses which are cheaper with acceptable speed rather than taxies which are expensive with faster speed (Wardrip, 2011). Since most of the residents in a city are middle-class, bus traffic represents the majority of urban mobility. Besides, according to (Montanari & Staniscia, 2012), there is a connection between a drop in estate prices and a decreased flow of bus mobility. We thus measure the arriving, leaving and transition volumes of buses in the neighborhood of each estate. Let BT denote the set of all the bus trajectories of Beijing, each of which represents a bus trajectory, denoted by a tuple $< p, d >$ where p is a pickup bus stop and d is a drop-off bus stop.

Bus Arriving, Leaving and Transition Volume: Similar to taxi mobility volume, we also extract the arriving volume, leaving volume and transition volume of buses from smart card transactions. Formally,

$$\begin{aligned} f_i^{BAV} &= |\{< p, d > \in BT : p \notin r_i \& d \in r_i\}| \\ f_i^{BLV} &= |\{< p, d > \in BT : p \in r_i \& d \notin r_i\}| \\ f_i^{BTV} &= |\{< p, d > \in BT : p \in r_i \& d \in r_i\}| \end{aligned} \tag{2.11}$$

Bus Stop Density: Recent work (Robert Cervero, 2011) reports that price premiums of up to ten percents are estimated for estates within 300m of more bus stops. In other words, the bus stop density is positively correlated to estate prices. Here, we propose an alternative approach and strategically estimate bus stop density using smart card transactions. In smart card transactions, the ticket fare of a trajectory

indeed reflects the number of bus stops in this trajectory. This is because the Beijing Public Transportation Group charges passengers according to the number of stops of each trip. Given the pick-up stop p and the drop-off stop d, the trip distance between p and d is fixed in a designed bus route. Then, the ratio of trip distance to bus stop number implicitly suggests in average distance between every two consecutive bus stops. Since the bus stop number of a trip can be approximated by the fare, we compute the ratio of distance to fare for estimating the density of bus stop in a neighborhood. The smaller the distance-fare ratio is, the higher the bus stop density is.

$$f_i^{BSD} = \frac{\sum_{p \in r_i || d \in r_i} dist(p,d)/fare(p,d)}{|\{< p,d > \in BT : p \in r_i || d \in r_i\}|} \tag{2.12}$$

Smart Card Balance: The smart card balances imply the patterns of the consumption and recharge behaviors. If residences always maintain a higher balance in their smart card, this suggests the card holders spend more money on bus travel. The large expense of bus travel implies: (1) residences depend on buses more than other transportation (e.g., subway, taxi), which may indicate that the affiliated neighborhood is lack of subways and taxies; (2) residences travel a longer distance to work, shop and pick up children, and thus need to maintain a high balance. In other words, this place is remote and inconvenient. We thus consider to extract the smart card balance as a feature. Formally,

$$f_i^{SCB} = \frac{\sum_{p \in r_i || d \in r_i} balance(p,d)}{|\{< p,d > \in BT : p \in r_i || d \in r_i\}|} \tag{2.13}$$

Check-in Related Features. Mobile users check in at online location-aware social networks when they walk in an important place. These check-ins are a significant

portion of urban mobility. Estate price is likely high in communities where there are convenient transit stations with good access to retail stores and services (Wardrip, 2011). Therefore, check-in behaviors could partially reflect the access convenience to these locations. In our data set, each check-in event can be denoted by a tuple, $< p, t, c > \in CI$, where p, t, c and CI represent the POI of the check-in, the check-in time stamp, the category of POI, and the set of check-in events, respectively.

Neighborhood Popularity: We count the total number of check-ins reported in the neighborhood of each estate as popularity measurement. Formally,

$$f_i^{NP} = |\{< p, t, c > \in CI : p \in r_i\}| \tag{2.14}$$

Topic Profile: The goal of topic distillation is to learn the topic distribution of a neighborhood based on the textual information of check-ins via a two-step approach. *STEP1: Propagating word-of-mouth from poi to neighborhood.* In check-in data, each POI is associated with textual reviews posted by users. This textual information reflects opinion of users toward this POI. Since each neighborhood is associated with a cluster of POIs, we therefore propose to propagate the word-of-mouth of mobile users from poi to neighborhoods by spatio-textual aggregation using check-in data. We get a cluster of textual posts denoted as d_{e_i} for the neighborhood of each estate e_i. We then segment these sentences into words and extract the semantically significant tags for each neighborhood. One reason for propagating word-of-mouth from poi to neighborhood is that the terms associated with a single POI are usually short, incomplete and ambiguous. Moreover, LDA is proven non-effective for short texts. The aggregation process can better learn thousands of mobile users' opinions toward

estates in terms of latent topic distributions.

STEP2: Textual profiling from words to topics. Next we exploit the LDA model for estate topic profiling by treating each estate neighborhood as a document. In LDA, each document is represented as a probability distribution over topics (document-topic distribution) and each topic is represented as a probability distribution over a number of words (topic-word distribution).In this way, we build an aggregated LDA model as shown in Figure 2.4(d). Here, the topic distribution of each document $\Pr(z \mid d_{e_i})$ is treated as topical features of estate, where z and d_{e_i} are topic and document respectively. The topic profiling process of the estates is as following:

1. For each topic $z \in \{1, ..., K\}$, draw a multinomial distribution over terms, $\phi_z \sim Dir(\beta)$.

2. For the document d_{e_i} given an estate e_i

   (a) Draw a multinomial distribution over topics, $\theta_{d_{e_i}} \sim Dir(\alpha)$

   (b) For each word $w_{d,n}$ in document d_{e_i}:

      i. Draw a topic $z_{d,n} \sim Mult(\theta_{d_{e_i}})$

      ii. Draw a word $w_{d,n} \sim Mult(\phi_{z_{d,n}})$

So far, we have extracted two categories of estate features as shown in Table 4.2. We emphasize that the above features are defined in terms of the neighborhood (r_i) of each estate, which is parameterized by its radius d. Hence, we can extract multiple groups of estate features with respect to different neighborhood radius (e.g., d=0.25,0.5,0.75,1,1.25,...,3km).

2.2.3 Sparse Pairwise Ranking for Estate Appraisal

Here we present the sparse pairwise estate ranker.

Model Description: Since many existing learning-to-rank algorithms use linear rankers, we learn a linear ranking predictor. Let $\boldsymbol{x_i}$ denote the M-size vector representation of estate e_i with the above extracted features, f_i denote the predicted estate value, and y_i denote the ground truth estate value, then we have $f_i(\boldsymbol{x_i}; \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{x_i} + \epsilon_i = \sum_{m=1}^{M} w_m x_{im} + \epsilon_i$, where ϵ_i is a zero-mean Gaussian bias with variance σ^2, and $\boldsymbol{w}$ is the weights of features. In other words, $P(y_i|\boldsymbol{x_i}) = \mathcal{N}(y_i|f_i, \sigma^2) = \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x_i}, \sigma^2)$ where $\mathcal{N}$ represents normal distribution.

Objective Function: While these features indeed capture residents' opinions about estates to be ranked, they usually are inter-correlated and redundant. Thus possible confounders lead to poor generalization performance. To address this issue, we adopt a strategy which simultaneously conducts feature selection while maximizing estate ranking accuracy. Since pairwise ranking strategy is effective with lower complexity comparing with listwise ranking strategy, we combine a pairwise ranking objective and a sparsity regularization term in a unified probabilistic modeling framework.

Next we introduce how to derive the mixture objective of sparse pairwise estate ranking. Let us denote all parameters by $\Psi = \{\boldsymbol{w}, \boldsymbol{\beta}^2\}$ which are the parameters of estate ranker (we will introduce $\boldsymbol{\beta}^2$ in the following), the hyperparamters by $\Omega = \{a, b, \sigma^2\}$ which are the parameters of sparsity regularization, and the observed data by $\mathcal{D} = \{Y, \Pi\}$ where Y and Π are the investment values and ranks of I estates respectively. For simplicity, we assume the real estates in $\mathcal{D}$ are sorted and indexed

in a descending order in terms of their investment values, which compiles a descending ranks as well. In other words, i is both the index and the ranking order of the given estate x_i By Bayesian inference, we have the posterior probability as

$$Pr(\Psi; \mathcal{D}, \Omega) = P\left(\mathcal{D}|\Psi, \Omega\right) P\left(\Psi|\Omega\right) \tag{2.15}$$

First, the term $P\left(\mathcal{D}|\Psi, \Omega\right)$ is the likelihood of the observed data collection $\mathcal{D}$, which can be explained as a joint probability of both estate investment values, $P\left(Y|\Psi, \Omega\right)$, and estate ranking consistency, $P\left(\Pi|\Psi, \Omega\right)$. Here we treat the ranked list of estates as a directed graph, $G =< V, E >$, with nodes as estates and edges as pairwise ranking orders. For instance, edge $i \to h$ represents an estate i is ranked higher than estate h. From a generative modeling angle, edge $i \to h$ is generated by our model through a likelihood function $P(i \to h)$. The more valuable estate i is than estate h, the larger $P(i \to h)$ should be. On the contrary, the case, in which $i \to h$ but $f_i < f_h$, will punish $P(i \to h)$. Therefore,

$$\begin{aligned} P\left(\mathcal{D}|\Psi, \Omega\right) &= P\left(Y|\Psi, \Omega\right) P\left(\Pi|\Psi, \Omega\right) \\ &= \prod_{i=1}^{I} \mathcal{N}(y_i|f_i, \sigma^2) \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(i \to h|\Psi, \Omega) \end{aligned} \tag{2.16}$$

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

Second, the term $P\left(\Psi|\Omega\right)$ is the prior of the parameters Ψ. Here, we introduce a sparse weight prior distribution by modifying the commonly used Gaussian prior, such that a different and separate variance parameter β_m^2 is assigned for each weight. Thus, $P\left(\boldsymbol{w}|\boldsymbol{\alpha}\right) = \prod_{m=1}^{M} \mathcal{N}(w_m|0, \beta_m^2)$, where β_m^2 represents the variance of corresponding parameter w_m and $\boldsymbol{\beta}^2 = (\beta_1^2, ..., \beta_M^2)^\top$, each of which is treated as a random variable.

Later, an Inverse Gamma prior distribution is further assigned on these hyperparameters, $P(\boldsymbol{\beta}^2|a,b) = \prod_{m=1}^{M}\text{Inverse-Gamma}(\beta_m^2;a,b)$, where a and b are constants and are usually set close to zero. By integrating over the hyperparameters, we can obtain a student-t prior for each weight, which is known to enforce sparse representations during learning by setting some feature weights to zero and avoiding overfitting.

$$\begin{aligned} P(\Psi|\Omega) &= P(\boldsymbol{w}|0,\boldsymbol{\beta}^2)P(\boldsymbol{\beta}^2|a,b) \\ &= \prod_{m=1}^{M}\mathcal{N}(w_m|0,\beta_m^2)\prod_{m=1}^{M} Inverse-Gamma(\beta_m^2|a,b) \end{aligned} \tag{2.17}$$

Parameter Estimation: With the formulated posterior probability, the learning objective is to find the optimal estimation of the parameters Ψ that maximize the posterior. Hence, by inferring Equation 4.6, we can have the log of the posterior for the proposed model.

$$\begin{aligned} &\mathcal{L}(\boldsymbol{w},\boldsymbol{\beta}^2|Y,\Pi,a,b,\sigma^2) = \\ &\sum_{i=1}^{I}\left[-\frac{1}{2}\ln\sigma^2 - \frac{(y_i-f_i)^2}{2\sigma^2}\right] + \sum_{i=1}^{I-1}\sum_{h=i+1}^{I} ln\frac{1}{1+exp(-(f_i-f_h))} \\ &+\sum_{m=1}^{M}\left[-\frac{1}{2}\ln\beta_m^2 - \frac{w_m^2}{2\beta_m^2}\right] + \sum_{m=1}^{M}\left[-(a+1)\ln\beta_m^2 - \frac{b}{\beta_m^2}\right] \end{aligned} \tag{2.18}$$

We apply a gradient descent method to maximize the posterior by updating w_m, β_m^2 through $w_m^{(t+1)} = w_m^{(t)} - \epsilon\frac{\partial(-\mathcal{L})}{\partial w_m}$ and $\beta_m^{2(t+1)} = \beta_m^{2(t)} - \epsilon\frac{\partial(-\mathcal{L})}{\partial \beta_m^2}$ where

$$\begin{aligned} \frac{\partial(\mathcal{L})}{\partial w_m} &= \sum_{i=1}^{I}\frac{1}{\sigma^2}(y_i - \sum_{m=1}^{M} w_m\cdot x_{im})x_{im} + \\ &\sum_{i=1}^{I-1}\sum_{h=i+1}^{I}\frac{exp(-(f_i-f_h))}{1+exp(-(f_i-f_h))}(x_{im}-x_{hm}) + \frac{-w_m}{\beta_m^2} \end{aligned} \tag{2.19}$$

$$\frac{\partial(\mathcal{L})}{\partial \beta_m^2} = \frac{w_m^2+b}{\beta_m^4} - \frac{3+2a}{2\beta_m^2} \tag{2.20}$$

2.2.4 Ranking Inference

After parameters Ψ are estimated via maximizing the posterior probability, we will obtain the learned model for investment value of estate, i.e., $\mathbb{E}(y_i|\boldsymbol{w}, \boldsymbol{\beta}) = \boldsymbol{x_i w}$ given a rising or falling market period. For a new coming estate k, we may predict its investment value accordingly. The larger the $\mathbb{E}(y_k|\boldsymbol{w}, \boldsymbol{\beta})$ is, the higher investment value it has.

For practical usage, we train two ranking models, $g(x)$ and $g^{'}(x)$, for the rising and falling markets respectively. Since we do not predict whether a market will go up or go down, we feed the features of a real estate into two models respectively and generate two value levels, which denote its value-adding and value-protecting abilities in rising and falling markets. To provide a unified ranking to users, the output of these two models can be aggregated as $R = \alpha \cdot g(x) + (1 - \alpha) \cdot g^{'}(x)$.

2.3 Experimental Results

We provide an empirical evaluation of the performances of the proposed method on real-world estate related data.

2.3.1 Experimental Data

Table 4.3 shows five data sources. The taxi GPS traces are collected from a Beijing taxi company. Each trajectory contains trip id, distance(m), travel time(s), average speed(km/h), pick-up time and drop-off time, pick-up point and drop-off point. Also, we extract features from the Beijing smart card transactions. Each bus trip has

card id, time, expense, balance, route name, pick-up and drop-off stops information (names, longitudes and latitudes). Moreover, the check-in data of Beijing is crawled from www.jiepang.com which is a Chinese version of Fourquare. Each check-in event includes poi name, poi category, address, longitude and latitude, comments. Furthermore, we crawl the online business reviews of Beijing from www.dianping.com which is a business review site in China. Each review contains shop ID, name, address, latitude and longitude, consumption cost, star (from 1 to 5), poi category, city, environment, service, and overall ratings. Finally, we crawl the Beijing estate data from www.soufun.com which is the largest real-estate online system in China.

2.3.2 Baseline Algorithms

To show the effectiveness of our method, we compare our method against the following algorithms. (1) **MART (Friedman, 2001):** it is a boosted tree model, specifically, a linear combination of the outputs of a set of regression trees. (2) **RankBoost (Freund, Iyer, Schapire, & Singer, 2003):** it is a boosted pairwise ranking method, which trains multiple weak rankers and combines their outputs as final ranking. (3) **Coordinate Ascent (Metzler & Croft, 2007):** it uses domination loss and applies coordinate descent for optimization. (4) **LambdaMART (Burges, 2010):** it is the boosted tree version of LambdaRank, which is based on RankNet. LambdaMART combines MART and LambdaRank. (5) **FenchelRank (Lai, Pan, Liu, Lin, & Wu, 2013)** beyond traditional ranking methods, we further compare with FenchelRank which is designed for solving the sparse learning-to-rank (LTR) problem with a L1 constraint.

Table 2.2. Statistics of the experimental data.

Data Sources	**Properties**	**Statistics**
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 business POIs	1472
	Number of reviews	470846
	Number of users	159820
Real Estates	Number of real estates	2,851
	Size of bounding box (km)	40*40
	Time period of transactions	04/2011 - 09/2012

We utilize RTree [3] to index geographic items (i.e., taxi and bus trajectories, checkins, etc.) and extract the defined features. We use Jieba [4] which is a Chinese/English text segmentation module to segment words and extract tags. For traditional LTR algorithms, we use RankLib [5] . We set the number of trees = 1000, the number of leaves = 10, the number of threshold candidates = 256, and the learning rate = 0.1 for MART. We set the number of iteration = 300, the number of threshold candidates = 10 for RankBoost. We set step base = 0.05, step scale = 2.0, tolerance = 0.001, and slack = 0.001 for Coordinate Ascent. We set number of trees = 100, number of leaves = 10, number of threshold candidates = 256, learning rate = 0.1 for LambdaMART. For FenchelRank, we use the source code[6] provided by the author. We set a=0.01, b=0.01, and $\sigma^2 = 1000$ for our model.

All the codes are implemented in R (modeling), Python (feature extraction) and Matlab (visualization). And all the evaluations are performed on a x64 machine with i7 3.40GHz Intel CPU (with 4 cores) and 24GB RAM. The operation system is Microsoft Windows 7.

2.3.3 Evaluation Metrics

Normalized Discounted Cumulative Gain. The discounted cumulative gain (DCG@N) is given by $DCG[n] = \begin{cases} rel_1 \quad if\, n = 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n}, \quad if\, n >= 2 \end{cases}$ Later, given the ideal discounted cumulative gain DCG', NDCG at the n-th position can be com-

[3] https://pypi.python.org/pypi/Rtree/

[4] https://github.com/fxsjy/jieba

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

[6] http://ss.sysu.edu.cn/ py/fenchelcode.rar

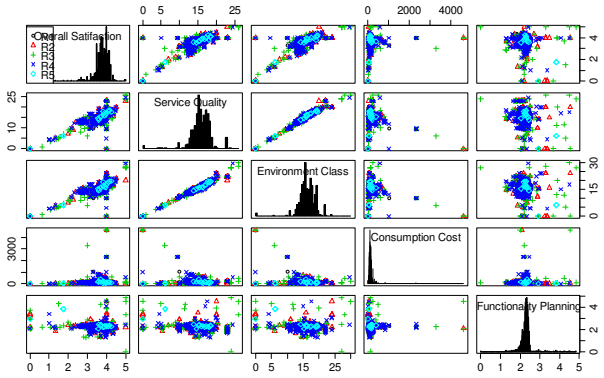
puted as NDCG[n]= $\frac{DCG[n]}{DCG'[n]}$. The larger NDCG@N is, the higher top-N ranking accuracy is.

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

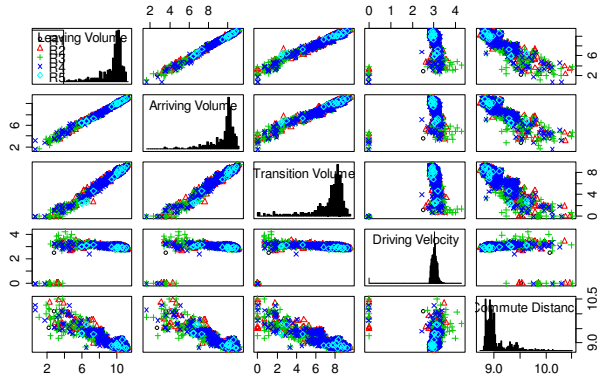
Kendall's Tau Coefficient. Kendall's Tau Coefficient (or Tau for short) measures the overall ranking accuracy. Let us assume that each estate i is associated with a benchmark score y_i and a predicted score f_i. Then, for an estate pair $< i, j >$, $< i, j >$ is said to be concordant, if both $y_i > y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i < f_j$. Also, $< i, j >$ 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 $= \frac{\#_{conc} - \#_{disc}}{\#_{conc} + \#_{disc}}$.

2.3.4 Correlation Analysis

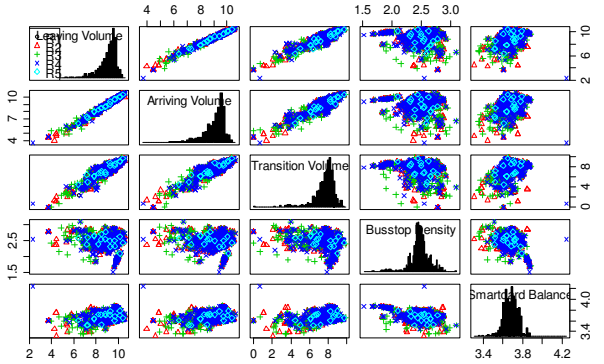
We provide a visualization analysis to validate the correlation between the extracted features and estate investment values. We use scatter-plot matrix for correlation analysis. Each non-diagonal chart in a scatter plot matrix shows the correlation between a pair of features whose feature names are listed in the corresponding diagonal charts. Given a set of N features, there are N-choose-2 pairs of features, and thus the same numbers of scatter plots. The dots represent the estates and their colors represent the grades of investment value. For readability, we use $R5 > R4 > R3 >$



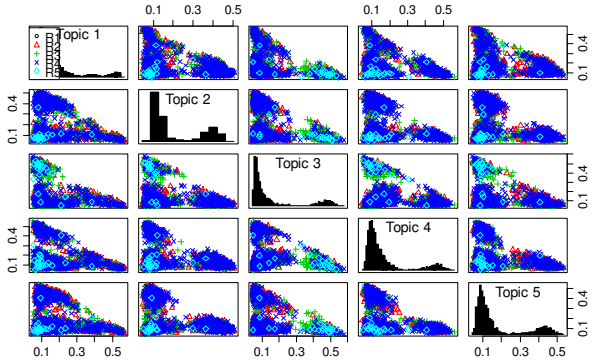
(a) Features of business review



(b) Features of taxi traces



(c) Features of bus traces



(d) Topics of mobile check-ins

$R2 > R1$ (symbol) to represent $4 > 3 > 2 > 1 > 0$ (number) in Figure 2.5.

In Figure 2.5(a), we present the correlation between business review features (overall satisfaction, service quality, environment class, consumption cost) and estate investment value. As can be seen, the R5 estates tend to appear at the top right corner of all the non-diagonal charts. This implies that if mobile users have higher ratings for estate neighborhoods, estate investment values are the higher. Remind that we mean the heterogenesis of poi planning by the entropy of frequency of categorized POIs. Interestingly, we observe if the heterogenesis of functionality planning is too high or too low, these estates are usually low-value. This can be intuitively explained by the fact that people are willing to live in a community that can meet and balance the needs of their life.

In Figure 2.5(b), we show the positive correlation between the taxi leaving, arriving and transition volumes of estate neighborhoods and estate investment value. However, the commute distance of taxies has negative correlation with estate investment value. In other words, the shorter the commute distance of taxies is, the higher is the estate investment value. A potential interpretation of this observation is that since taxies are valued by white-collar and business people, the destinations of taxi trajectories usually are important places (e.g., conference centers, business hotels, companies and government organizations, etc). If the commute distance of taxies is short, the targeted neighborhood is close to these important places.

In Figure 2.5(c), we show the positive correlation between estate investment value and bus related features, such as the leaving, arriving, and transition volumes of buses, bus stop density. Figure 2.5(d) illustrates that Topic 4 has positive correlation

with estate investment value whereas Topic 1,2,3,5 have negative correlation. This validates topic profiling of checkin posts can help discriminate estate values.

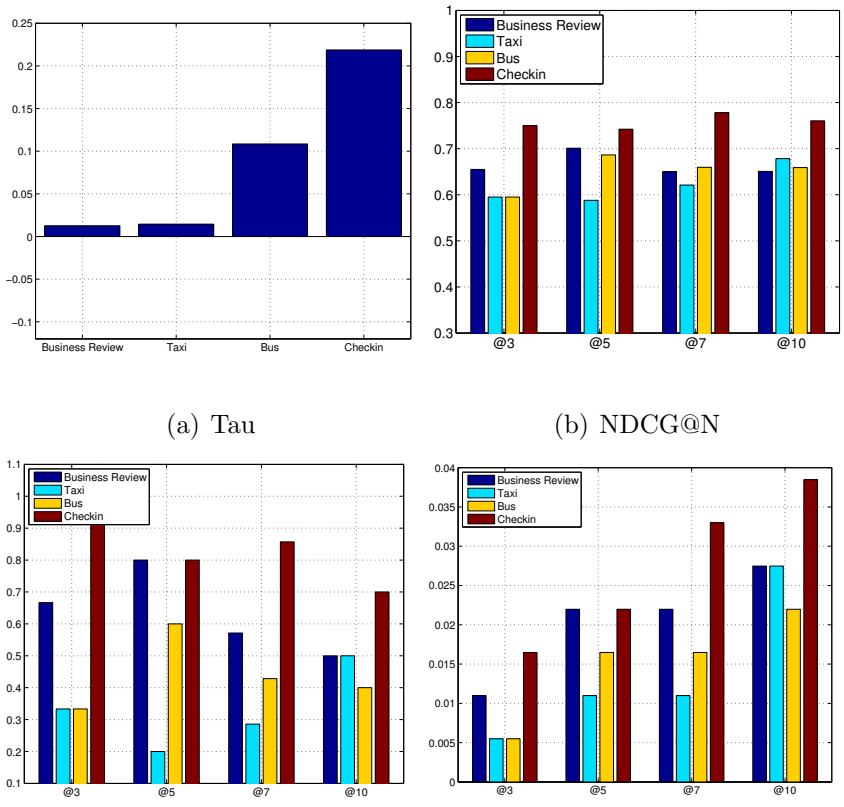
The visualization results show the collectiveness of our intuitions for defining and extracting discriminative features

2.3.5 Feature Evaluation

We evaluate the performances of different features segmented from two perspectives.

Evaluation on features of different data sources. We segment the extracted features in terms of different data sources and investigate which source is more effective for ranking estates. Figure 2.6 and Figure 2.7 shows the Tau, NDCG, Precision, and Recall of four feature sets (business reviews, taxi traces, smart card transactions and check-ins) in rising market and falling market respectively. In all cases, we observe the extracted features achieve good performances, yet there are features which are substantially better than others.

Specifically, the check-in features perform best with Tau 0.1046198, NDCGs $>$ 0.75, Precisions $>$ 0.85, and Recalls $>$ 0.24 in rising market, and consistently achieve the best ranking results in falling market. The features of business reviews hold the second place of overall and top-k rankings in rising and falling markets. In sum, business reviews and check-ins performs better than taxi and bus traces. One possible reason is that people's outdoor activities consist (1) moving phrase and (2) attending phrase. Although moving phrase (taxi and bus trajectories) help realize activity attending (check-ins and business reviews), the drop-off points of taxi and bus trajectories are not always the destinations of outdoor activities. Whereas, the

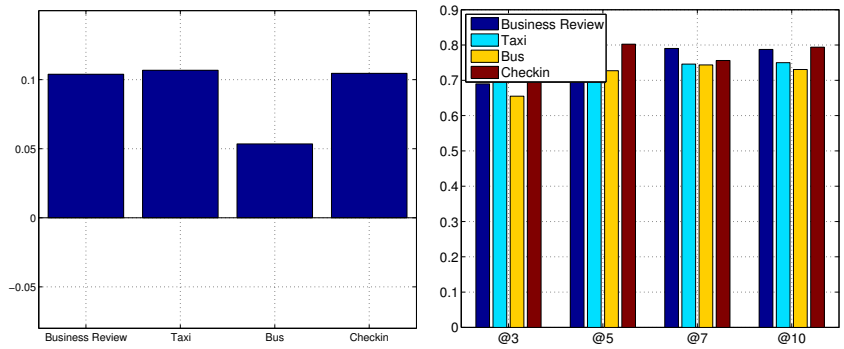


(a) Tau (b) NDCG@N

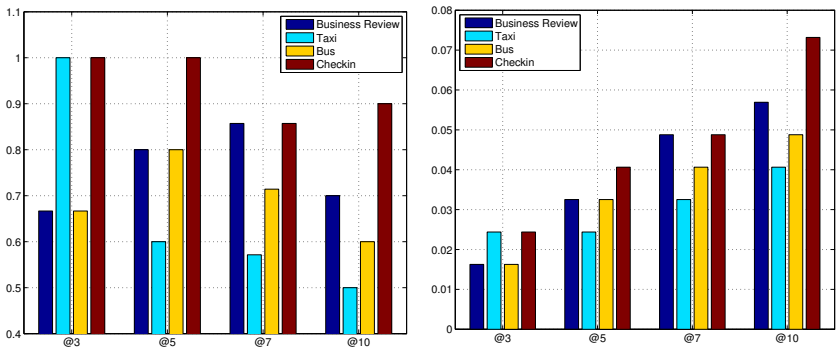
(c) Precision@N (d) Recall@N

Figure 2.6. Feature performances of different sources on the rising market dataset.

locations of check-ins and business reviews usually are the final destinations of people's visits. They reflect direct interaction between users and activities via locations, and thus have semantically richer information than public transits. Besides, a comparison between Figure 2.6 and Figure 2.7 shows that bus features perform better than taxi features in rising market, whereas taxi features perform better than bus ones in falling market. We note that bus traces stand for the mobility of mediate classes while taxi traces stand for the mobility of white-collar and business people. This observation implies that in falling market, despite economic recession, since the high-income groups still have strong purchasing power of estates, their preferences have more influence on estate prices than middle class.



(a) Tau (b) NDCG@N



(c) Precision@N (d) Recall@N

Figure 2.7. Feature performances of different sources on the falling market dataset.

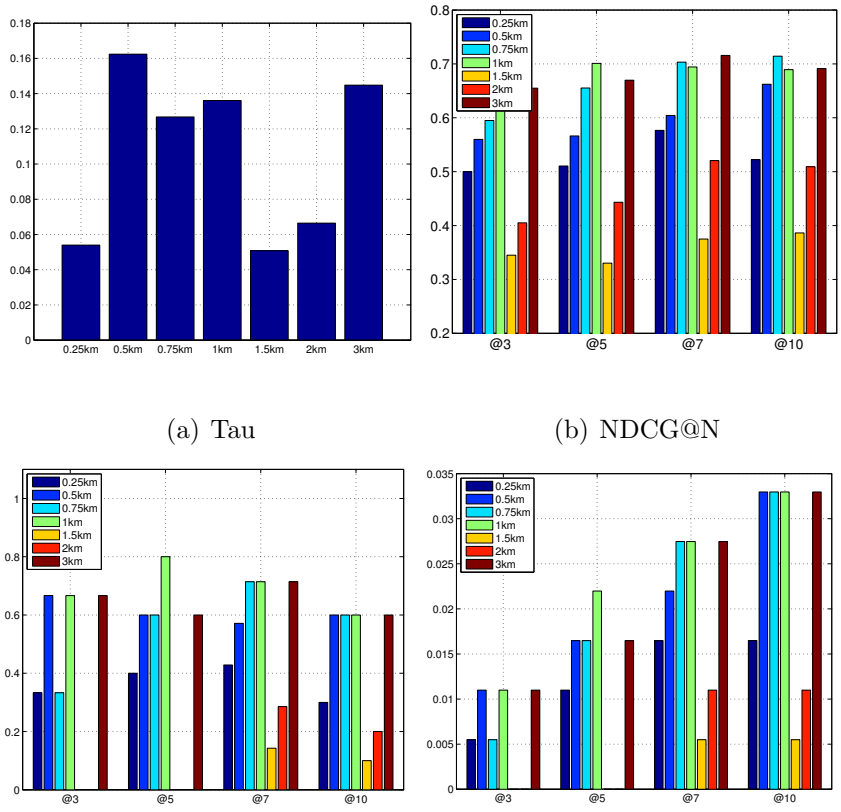
Evaluation on features of different radius distances. We segment the features in terms of different neighborhood radiuses and investigate the proper radiuses of neighborhoods for estate ranking. In Figure 2.8 and Figure 2.9, we report the performance comparison of feature sets of different radius distances (i.e., 0.25, 0.5, 1, 1.5, and 3km) in both rising and falling markets. We observe that the radius distance of neighborhood can affect the ranking performance. Some radius distances substantially outperform others. Figure 2.8 illustrates the radiuses of 0.5km, 0.75km, 1km outperform other radiuses with a significant margin with respect to both overall and top-k ranking in rising market. The setting of 0.25, 1.5 or 2km in rising market lead to lower ranking accuracy. In falling market, 0.75km performs best. Besides, 0.25, 1.5 and 2km consistently perform worst as they do in rising market. Therefore, we recommend to set the radius of neighborhood to 0.75 ± 0.25km, rather than too short($<$0.25km) or too long($>$2km). This might be because 0.75km is not only a comfortable walking distance for bus and taxi stops, but also sufficient to capture the outdoor activities of estate neighborhoods.

The results justify the mining and fusion methods of feature extraction (e.g., direction, volume, velocity, heterogeneity, density, popularity, etc.).

2.3.6 Model Evaluation

We report the performance comparison of our method comparing to five baseline algorithms on rising market and falling market in terms of Tau and NDCG.

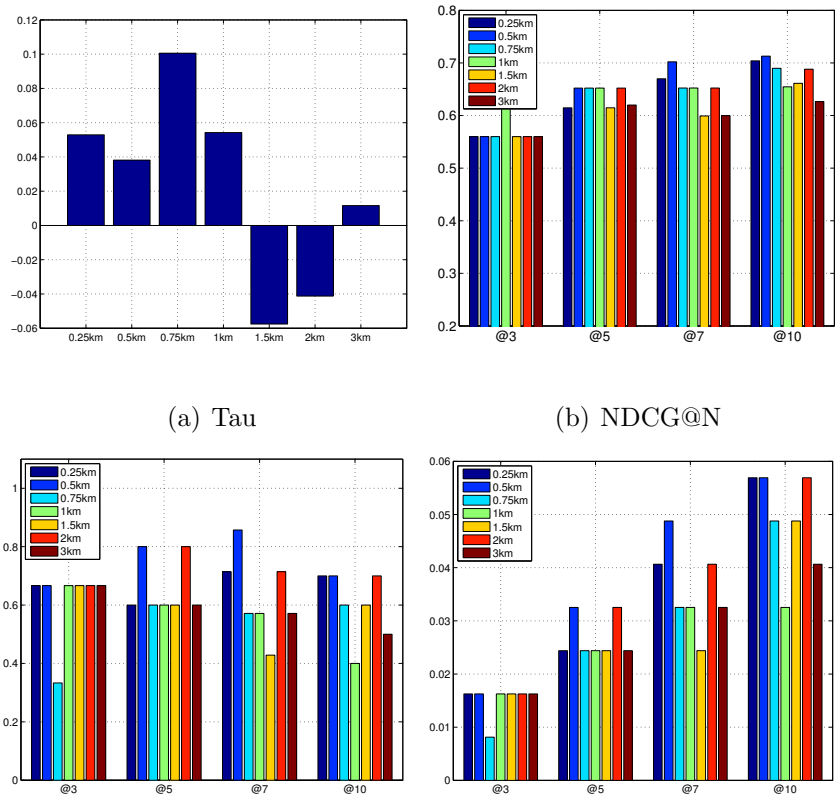
Rising Market Data. In Table 2.3, we present the performance comparison of NDCG and Tau in rising market. Our method achieves 0.75 NDCG@1, 0.6900469



(a) Tau (b) NDCG@N

(c) Precision@N (d) Recall@N

Figure 2.8. Feature performances of different radius on the rising market dataset.



(a) Tau (b) NDCG@N

(c) Precision@N (d) Recall@N

Figure 2.9. Feature performances of different radius on the falling market dataset.

NDCG@3, 0.6915216 NDCG@5, 0.6861585 NDCG@7, and 0.7016248 NDCG@10, which obviously outperform the baseline algorithms with a significant margin. Our method fuses sparsity regularization and pairwise ranking objective, and offers an increase in comparison to FenchelRank (Lai et al., 2013) which is a newly proposed sparse ranking algorithm. Specifically, our method achieves 15.9% increase in NDCG@3 and 24.2% increase in NDCG@5 comparing to FechelRank. This observation validates the superiority of our method when considering many intercorrelative features with confounders. Meanwhile, we also observe FenchelRank achieve the second best ranking accuracy comparing traditional ranking algorithms. This justifies the benefits of considering both sparsity regularization and ranking accuracy. With respect to overall ranking, our method achieves the highest Tau (0.3493753). In the comparison between tau and NDCG, an observation stands out is that although FenchelRank holds the second place in top-k ranking, it surprisingly achieve the lowest tau value. However, our method achieves a balance performance in both top-k and overall ranking.

Table 2.3. performance comparison of our approach and baselines in rising market.

Metric	MART	RankBst	CoordAsc	LamMART	FenRank	SEK
NDCG@3	0.50089	0.46493	0.55995	0.46493	0.59502	**0.69005**
NDCG@5	0.58295	0.52506	0.628623	0.48887	0.55679	**0.69152**
NDCG@7	0.59649	0.59105	0.63199	0.51548	0.61837	**0.68616**
NDCG@10	0.62377	0.56735	0.65563	0.50471	0.65999	**0.70162**
Tau	-0.01755	0.08892	-0.13704	0.07150	0.12243	**0.34938**

Falling Market Data. Table 2.4 shows the performance comparison of NDCG and Tau in falling market. We first compare all the six methods in terms of NDCG. Our method and RankBoost outperform the other algorithms with a significant margin. Regarding Tau, our method achieves the highest accuracy with 0.3347548. Although RankBoost obtains impressive NDCGs in falling market, it fails to consistently achieve good NDCGs in rising market. Whereas, our method consistently reports high and balanced performances in both rising and falling markets.

Table 2.4. performance comparison of our approach and baselines in falling market.

Metric	MART	RankBst	CoordAsc	LamMART	FenRank	SEK
NDCG@3	0.46493	0.75	0.59502	0.36991	0.44005	**0.69005**
NDCG@5	0.57712	0.77008	0.5725	0.46968	0.55746	**0.68514**
NDCG@7	0.61288	0.80305	0.53820	0.52281	0.59603	**0.68249**
NDCG@10	0.65570	0.81719	0.55537	0.54510	0.64049	**0.69719**
Tau	0.09481	0.12978	0.22331	0.23113	-0.12477	**0.33475**

The results validate the advantages of considering both ranking accuracy and sparsity regularization with the extracted intercorrealative features from heterogenous sources.

2.4 Related Work

Related work can be grouped into two categories. The first one includes the work on estate appraisal. In the second category, we present the ranking related methods.

Real estate appraisal is the process of valuing the property's market value. Tra-

ditional research on estate appraisal is based on financial real estate theory, typically constructing an explicit index of estate value (Krainer & Wei, 2004), such as price-rent ratio, price to income ratio. More studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of estate prices (Chaitra H. Nagaraja & Zhao, 2009). Work (Downie & Robson, 2007) checks the volatility of estate price and concludes that low investment-valued estate values relatively volatile. More classic works are based on repeat sales methods and hedonic methods. The repeat sales methods (Shiller, 1991a) construct a predefined price index based on properties sold more than once during the given period. The hedonic methods (Taylor, 2003) assume the price of a property depends on its characteristics and location. Work (Downie & Robson, 2007) studies the automated valuation models which aggregate and analyze physical characteristics and sales prices of comparable properties to provide property valuations. More recent works (Kontrimas & Verikas, 2011; Fu, Xiong, et al., 2014) apply general additive mode, support vector machine regression, multilayer perceptron, ranking and clustering ensemble method to computational estate appraisal. In our earlier work (Fu, Xiong, et al., 2014), we focus on exploiting the mutual enhancement between ranking and clustering to model geographic utility, popularity and influence of latent business area for estimating estate value. Besides, in (Fu, Xiong, et al., 2014), we identify and jointly capture the geographical individual, peer, and zone dependencies as an estate-specific ranking objective for enhancing prediction of estate value. However, in this chapter, we details comprehensive feature designs that cover most of aspects that have an impact on estate value. Also, we integrate sparsity regularization into pairwise ranking strategy because the extracted features

are usually correlated and redundant.

Also, our work can be categorized into Learning-To-Rank (LTR) which includes pointwise, pairwise, and listwise approaches (Hang, 2011). The point-wise methods (Hang, 2011) reduce the LTR task to a regression problem: given a single query-document pair, predict its score. The pair-wise methods approximate 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 minimize average number of inversions in ranking (Burges et al., 2005; Freund et al., 2003; Quoc & Le, 2007; Fürnkranz & Hüllermeier, 2003). The list-wise methods, optimize a ranking loss metric over lists instead of document pairs (Xia, Liu, Wang, Zhang, & Li, 2008). For instance, H. Li et al. propose AdaRank (Xu & Li, 2007) and ListNet (Cao, Qin, Liu, Tsai, & Li, 2007) and Burges et al. propose LambdaMART (Burges, 2010). More recent work (Lai et al., 2013) further learn the ranking model which is constrained to be with only a few nonzero coefficients using L1 constraint and propose a learning algorithm from the primal dual perspective.

Urban computing (Zheng, Capra, Wolfson, & Yang, 2014) is a process of acquisition, integration, and analysis of urban data (e.g., sensors, devices, vehicles, buildings, human) to tackle the major issues that cities face. Our work also has a connection with mining mobile, geography and mobility data to tackle issues in urban space. Work (Ceci, Appice, & Malerba, 2007) identifies emerging patterns with multirelational approach from spatial data. Liu et al. detects spatio-temporal causality of outliers in traffic data (W. Liu, Zheng, Chawla, Yuan, & Xing, 2011). Yuan et al. discovers regional functions of a city using POIs and taxi traces (Yuan, Zheng, & Xie,

2012a) . Heierman et al. mines the device usage patterns of homeowners for smart houses (Heierman III & Cook, 2003) . Paper (Karamshuk, Noulas, Scellato, Nicosia, & Mascolo, 2013) selects the optimal sites for retail stores by mining Foursquare data. (Zheng et al., 2014) mines the driving route for end users by considering physical feature of a route, traffic flow, and driving behavior.

2.5 Conclusions

In this chapter, we aimed to assess estate investment value by mining a variety of user-generated data. We collected a large scale of online user reviews and offline moving behaviors (taxi traces, smart card transactions, and checkins) of mobile users. We index, filter, propagate, distill, aggregate mobile data, and extract the fine-grained features from multiple perspectives (e.g., direction, volume, velocity, heterogeneity, popularity, topic, etc.) for evaluating estate values. However, since the extracted estate features usually are intercorrelated and redundant, we proposed to learn a sparse pairwise ranker, which is mutually enhanced by simultaneously conducting feature selection and maximizing estate ranking accuracy. Finally, the experimental results with real world estate-related data demonstrates the competitive effectiveness of both extracted features and learning models.

CHAPTER 3

MODELING GEOGRAPHIC DEPENDENCIES FOR REAL ESTATE RANKING

It is traditionally a challenge for home buyers to understand, compare and contrast the investment values of real estates. While a number of estate appraisal methods have been developed to value real property, the performances of these methods have been limited by the traditional data sources for estate appraisal. However, with the development of new ways of collecting estate-related mobile data, there is a potential to leverage geographic dependencies of estates for enhancing estate appraisal. Indeed, the geographic dependencies of the value of an estate can be from the characteristics of its own neighborhood (individual), the values of its nearby estates (peer), and the prosperity of the affiliated latent business area (zone). To this end, in this chapter, we propose a geographic method, named ClusRanking, for estate appraisal by leveraging the mutual enforcement of ranking and clustering power. ClusRanking is able to exploit geographic individual, peer, and zone dependencies in a probabilistic ranking model. Specifically, we first extract the geographic utility of estates from geography data, estimate the neighborhood popularity of estates by mining taxicab trajectory data, and model the influence of latent business areas via ClusRanking. Also, we use a linear model to fuse these three influential factors and predict estate investment values. Moreover, we simultaneously consider individual, peer and zone dependencies, and derive an estate-specific ranking likelihood as the objective function. Finally,

we conduct a comprehensive evaluation with real-world estate related data, and the experimental results demonstrate the effectiveness of our method.

3.1 Introduction

There are a number of online estate information systems, such as Yahoo! Homes, Zillow.com, and Realtor.com, which provide functions to help people to search estate-related information. In these systems, home buyers can also rank estates based on some criteria, such as prices, the number of bedrooms, and the home size. However, the decision process of buying a house is different from that of buying a regular product. Home buyers not only aim to gain utility from a house, but also seek resale values and long-term capital growth. Therefore, home buyers often need the tool to rank estates based on their investment values. Indeed, the investment value is more related to the potential capital growth in the future. The **return rate**[1] is often used to quantify the investment values of estates instead of using the price. In fact, a high price does not necessarily mean a high investment value, and vice versa.

Traditionally, estate appraisal methods can help for the estimation of the values of estates, but the performances of these methods have been limited by the traditional data sources for estate appraisal. For instance, traditional estate price modeling methods exploit the trend, periodicity and volatility of price time series. However, both rigid and speculative demands have a big impact on the prices of estates. It is difficult to identify the true estate values only with the current prices. Also, the comparative estate analysis, e.g. automated valuation models (AVMs), typically aggregates

[1] http://en.wikipedia.org/wiki/Rate_of_return

and analyzes the physical characteristics and sales prices of comparable properties to provide property evaluations. However, AVMs could fail to appraise new or planned estates due to the lack of comparable property data.

Indeed, with the development of new ways of collecting estate-related mobile data, there is a potential to exploit geographic dependencies of estates for enhancing estate appraisal. In fact, a large amount of estate-related mobile data, such as urban geographic data and human mobility information near estates, have been accumulated. If properly analyzed, these data could be a source of rich intelligence for finding estates with high investment values.

Specifically, in this chapter, we study three types of geographic dependencies, which categorize estate values from three perspectives: (1) the geographic characteristics of its own neighborhood (individual), (2) the values of its nearby estates (estate-estate peer), and (3) the values of its affiliated latent business area (estate-business zone). First, the investment value of an estate is largely determined by the geographic characteristics of its own neighborhood. This is called **individual dependency**. For example, people are usually willing to pay higher prices for estates close to the best public schools. The individual dependency can be captured by correlating the estate investment values with urban geography (e.g. bus stops, subway stations, road network entries, and point of interests (POIs)) as well as human mobility patterns. Second, the estate investment value can be reflected by its nearby estates. This is called **peer dependency**. The peer dependency can be captured by the comparative estate analysis which is a popular method in estate appraisal and evaluates estates based on peer estate comparison. An intuitive understanding along

this line is, if the surrounding estates are of high investment values, the targeted estate will usually have a high value as well.

Third, the estate value can also be influenced by the values of its affiliated latent business area. This is called **zone dependency**. A business area is a self-organized region with many estates. The formation of business areas are driven by the long-term commercial activities under two mutually-enhanced effects: (1) estates tend to co-locate in multiple centers, and thus bring human activities to those business areas; (2) prosperous business areas in return lead to more estate constructions. Hence, a prosperous business area represents a high density cluster of human activities, commercial activities, and estates. Here, we assume that each estate is affiliated with a latent business area and each business area is endowed with a value function of estate investment preferences, which measures the prosperity of the estate industry in this business area. The more prosperous the business area is, the easier we can identify a high investment-value estate from this business area.

In summary, the individual dependency shows that the estate investment value can be reflected by urban geography information and human mobility data. This allows us to value real property when we lack of comparable estates. Also, the peer dependency allows to exploit spatial autocorrelation of investment values through the comparison between the targeted estate and its peer estates. Moreover, the zone dependency allows to explore the influence of the associated latent business area of an estate. Based on the above, in this chapter, we propose a geographic method, named ClusRanking, for estate appraisal by leveraging the mutual enforcement of ranking and clustering power. ClusRanking is able to exploit geographic individual, peer and

zone dependencies into a unified probabilistic ranking model.

Specifically, we first extract the geographic utility from urban geography data. Then, we estimate the neighborhood popularity through spatial propagation and aggregation of passenger visit probabilities by mining taxicab trajectory data. Moreover, we model the influence of latent business areas via ClusRanking. In particular, since we assume there are multiple latent business areas in a city, we embed a dynamic spatial-clustering approach into the ranking process. Here, each business area is treated as a spatial hidden state. A business area not only shows the locations of its estates, but also reflects the influence on estate investment values in terms of geographic proximity between estate and the centroids of the business area. Our method is iteratively updated by mutual enhancement between spatial-clustering and ranking until the boundaries of latent business areas are learned. After this, we fuse the three factors and learn estate investment values for estate ranking. In addition, we derive a mixture likelihood objective, which simultaneously considers the geographic individual, peer and zone dependencies. Here, individual dependency describes the prediction accuracy of estate investment values and locations. Peer dependency captures the ranking consistency of intra-business-area estate pairs. Zone dependency models the ranking consistency of inter-business-area estate pairs. Finally, we conduct a comprehensive performance evaluation on real world estate related data and the experimental results demonstrate the effectiveness of our method.

3.2 Real Estate Ranking

In this section, we introduce a geographic ClusRanking method for estate appraisal.

3.2.1 Problem Statement

In estate industry, two concepts are often used for an estate: value-adding capability and value-protecting capability, which are quantified by the investment value of estates in rising and falling markets respectively. In this chapter, we focus on estimating the investment value of estates and ranking all estates accordingly during these two markets. Ranking estates is very similar to the traditional information retrieval problem, where documents are ranked according to a defined relevance. Here, each estate is treated as a document and the value-adding capability or the value-protecting capability is considered as the relevance.

Formally, let $E = \{e_1, e_2, ..., e_I\}$ be a set of I estates, each of which is represented by all associated geographic features denoted as e_i as shown in Table 3.1, where more notation are listed. Our goal is to rank the estates in descending order according to the investment value in two markets. In fact, the essential task of this problem is how to estimate the investment value (denoted as y_i) of each estate i by modeling all associated relevant information of estates in a unified way. In this chapter, we consider a group of heterogenous information associated with estates, which include the public transportation information (e.g., bus stop, subway, road network), point of interest (e.g., restaurant and shopping mall), neighborhood popularity, and the influence among estate geographic zone.

3.2.2 The Overview of ClusRanking

Assume that each estate i is endowed with an investment value function y_i. We first build a model to predict y_i with the geographic information. Specifically, the estate

Symbol	Size	Description
$\boldsymbol{E}$	$I \times N$	estate geographic feature vector, e_i is the i^{th} estate
$\boldsymbol{Y}$	$1 \times I$	benchmark values, y_i is the benchmark value of e_i
$\boldsymbol{F}$	$1 \times I$	predicted values, f_i is the predicted value of e_i
$\boldsymbol{\Pi}$	$1 \times I$	ranks, π_i is the rank of e_i, smaller is better
$\overline{\boldsymbol{\Pi}}$	$1 \times I$	indexes, $\overline{\pi}_i$ is the index of i-th ranked estate, inverse of Π
$\boldsymbol{\gamma}$	$1 \times I$	geographic utility
$\boldsymbol{\delta}$	$1 \times I$	neighborhood popularity
$\boldsymbol{\rho}$	$1 \times I$	influence of business area
N	I	neighborhood set, n_i is the neighborhood of the i-th estate
D	-	drop-off point set
C	J	POI category set
R	$1 \times I$	business area assignments I estates
$\mathcal{R}$	K	latent business area set
$\boldsymbol{\eta}$	$1 \times K$	business area level prosperity distribution

Table 3.1. Mathematical Notations

value is affected by three factors: $y_i \propto \gamma_i + \rho_i + \delta_i$, in which (1) γ_i: the geographic utility extracted from urban geography data F_{geo}; (2) ρ_i: the influence of latent business area F_{area}; (3) δ_i: the neighborhood popularity estimated from human mobility data F_{mobi}. Then, we will be able to get a ranked list of estates based on their predicted investment values, and thus each estate i is associated with an inferred rank π_i. With the ranked list of estates, we formulate a likelihood function, which simultaneously captures the geographic individual (Lik_{id}), peer (Lik_{pd}) and zone (Lik_{zd}) dependencies. This likelihood function unifies both the prediction accuracy based on geographic data of estates and the ranking consistency of the estate ranked list. By maximizing this likelihood function, we could optimize the prediction accuracy of estate investment value and the ranking list of estates at the same time. Finally, we solve the optimization problem using a Expectation Maximization (EM) method. Figure 3.1 shows the framework of our method.

3.2.3 Modeling Estate Investment Value

Before introducing the overall objective function which captures the three dependencies altogether, let us first introduce how to model the investment value of estates with geographic information. Specifically, we will first introduce the modellings of γ_i, ρ_i and δ_i separately, and then state how they are combined together.

Geographic Utility: γ

Estate values are largely determined by its geographic location. Therefore, we naturally relate the geographic utility of estate to its location characteristics. More specifically, we first extract geographic features from estate neighborhoods (refer to

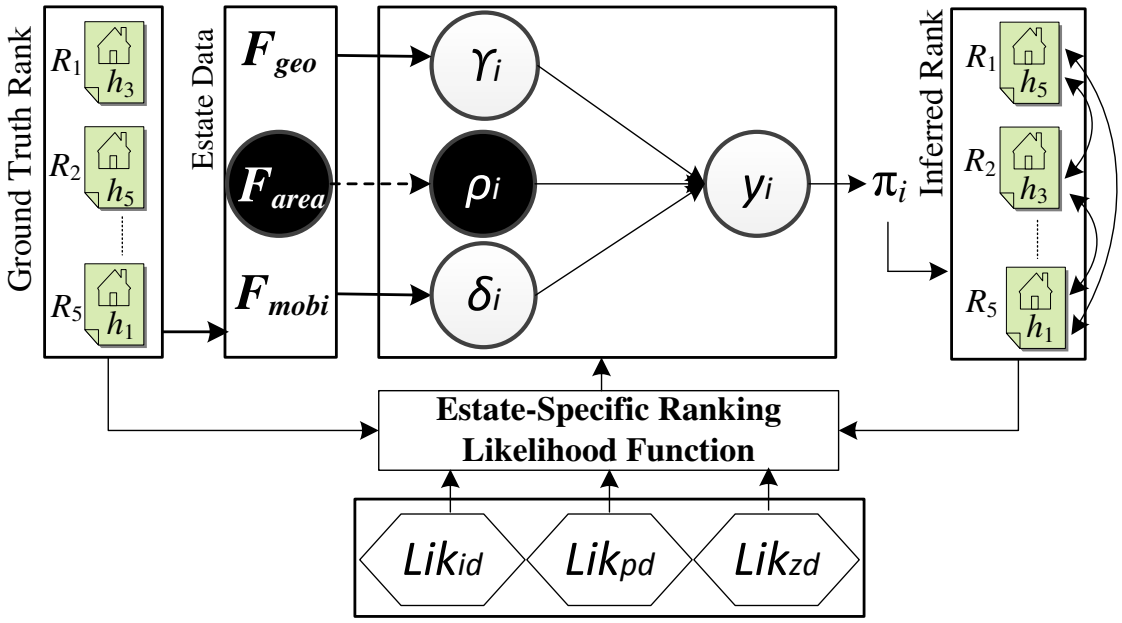


Figure 3.1. The framework of ClusRanking. (The black plates represent the latent effects.)

Data	Feature Design
Transportation	Number of bus stop Distance to bus stop
	Number of subway station Distance to subway station
	Number of road network entries Distance to road network entries
Point of interest	Number of POIs of different POI categories (Shopping, Sports, Education, etc.)

Table 3.2. Neighbourhood Profiling (a neighborhood is defined as a cell area with a radius of 1km.)

Table 4.2) and treat the raw representations of estates as a vector E. The raw representations of estates E are then learned and transformed to the meta representations WE using a single-layer perceptron, where $W \in M \times N$ is indeed a coefficient matrix. Finally, we parameterize geographic utility by a linear aggregation over transferred features in meta representation: $\gamma = qWE^{\top}$, where $q \in 1 \times M$ are the weights of the transferred features.

According to estate financial theory (Krainer & Wei, 2004), the estate investment value can be partially approximated by rent-interest ratio from market performances explicitly. We incorporate the rent-interest ratio into $\gamma = \frac{rent}{interest} + qWE^{\top}$ as side information to strengthen the robustness of our method.

Influence of Latent Business Area: ρ

Since we assume each estate is associated with a latent business area, the estate investment value also depends on the value of the associated business area. Suppose there are K latent business areas, we first choose the business area for each estate. We apply a multinominal distribution over latent business area $r \sim p(r|\boldsymbol{\eta})$, where $\boldsymbol{\eta} \in 1 \times K$ denotes the values (prosperity of estate industry or estate investment preference) of K business areas respectively. Later, each estate location l_i is drawn from a multivariate normal distribution: $l_i \sim \mathcal{N}(\mu_r, \Sigma_r)$, where $\mu_r \in 1 \times 2$ and $\Sigma_r \in 2 \times 2$ is the center and covariance of business area r, respectively. Finally, to model the influence of business area, we treat all the K business areas as K latent spatial states. The K latent spatial states together show the influence on each estate. Assume the influence is inversely proportional to the distance between the estate

location and the business area center: $d(i,r) = \sqrt{\|\mu_r - l_i\|_2}$, the influence of K business areas over estate i is defined by an aggregate power-law weighted parametric term $\rho_i = \sum_{k=1}^{K} \left(\frac{d_0}{d_0+d(i,r_k)}\right)^e \frac{\eta_k}{\sum_{k=1}^{K}\eta_k}$ where d_0 as a parameter and e is a mathematical constant.

Neighborhood Popularity: δ

Neighborhood popularity can affect the investment value of an estate to a certain extent. In general, people are willing to live in a popular neighborhood. A popular neighborhood usually has lots of notable POIs, which can be measured from two perspectives: (1) POI numbers, representing the quantitative measurement; (2) POI visit probability, representing the quality of those POIs. We propose to estimate the neighborhood popularity of a targeted estate by strategically combining POI numbers and POI visit probabilities using the taxicab GPS traces via a three-stage algorithm.

Propagating visit probability. In the first stage, given the drop-off point of a taxi trace d , we model the probability of a POI p visited by the passenger as a parametric function, whose input x is the road network distance between d and p: $P(x) = \frac{\beta_1}{\beta_2} \cdot x \cdot exp(1 - \frac{x}{\beta_2})$, where $\beta_1 = \max_x(P(x))$ and $\beta_2 = \operatorname{argmax}_x(P(x))$. The reasons why we adopt this function are as follows. First, when $x = 0$, $P(x) = 0$. Since a taxi could not send passengers into a POI directly, the drop-off point usually is not the same with the destination. A passenger often walks a short distance to reach the destination. Second, the drop-off point usually is close to the destination. Hence, when the distance exceeds a threshold β_2, the probability keeps decreasing with an exponential heavy tail. With this function, we can propagate the visit probability of

a passenger from the drop-off point to its surrounding POIs.

Aggregating POI-level visit probability. Given a POI p, the visit probability of p is measured by summarizing all the visit probabilities propagated from all the drop-off points in taxicab trace data via $\kappa(p) = \sum_{d \in D} P(dist(d, p))$.

Aggregating POI-category-level visit probability. In the third stage, we first identify the POIs located in the neighborhood n_i of the i-th estate. Then, we summarize the visit probability of those POIs per category c_j and obtain the category-level aggregated visit probability as $\phi_{ij} = \sum_{p \in c_j \wedge p \in n_i} \kappa(p)$. In this way, we reconstruct the representation of neighborhood popularity as an aggregated visit probability vector $\boldsymbol{\phi_i} =< \phi_{i1}, \cdots, \phi_{iJ} >$ over different POI categories for the i-th estate. Finally, we aggregate and normalize the popularity score as $\delta_i = \frac{1}{J} \sum_{j=1}^{J} \frac{\phi_{ij}}{max_{i \in r}\{\phi_{ij}\}}$.

Finally, we combine all modellings of γ_i, ρ_i and δ_i together and get the overall generative process of estate investment value as shown in Table 3.3. Specifically, we first assume there are K latent business areas in a city. Each business area is a cluster of estates. We treat K latent business areas as K spatial hidden states, each of which is endowed with a latent value η_k, which represents estate investment preference (or prosperity of estate industry) in the k-th business area. For each estate i, we draw a business area r from all K business areas following a multinomial distribution: Multi($\boldsymbol{\eta}$). The location of estate l_i is drawn from the sampled business area r. Later, given the estate location l_i is drawn, we are able to identify the neighborhood area and represent estate by a geographic feature vector e_i via neighborhood profiling. We then extract geographic utility γ_i from e_i. Moreover, we estimate the neighborhood popularity δ_i by strategically mining the taxicab trajec-

1 For each estate i:

1.1 Draw a business area $r \sim \text{Multinomial}(\eta)$.

1.2 Draw a location $l_i \sim \mathcal{N}(l_i; \mu, \sigma^2)$

1.3 Generate geographic utility

1.3.1 Draw coefficient matrix of meta representation

$$w_{mn} \sim \mathcal{N}(w_{mn}|\mu_w, \sigma_w^2)$$

1.3.2 Draw coefficient vector of geography utility

$$q_m \sim \mathcal{N}(q_m|\mu_q, \sigma_q^2)$$

1.3.3 Estate geographic utility $\gamma_i = \frac{rent_i}{interest} + qWe_i^\top$

1.4 Compute influence given by latent business areas

$$\rho_i = \sum_{k=1}^{K} \left(\frac{d_0}{d_0 + d(i, r_k)}\right)^e \frac{\eta_k}{\sum_{k=1}^{K} \eta_k}$$

1.5 Compute neighborhood popularity $\delta_i = \frac{1}{J} \sum_{j=1}^{J} \frac{\phi_{ij}}{max_{i \in r}\{\phi_{ij}\}}$

1.6 Generate the estate investment value $y_i \sim \mathcal{N}(y_i|f_i, \sigma^2)$ where

$$f_i = \gamma_i + \delta_i + \rho_i$$

2 Compile the ranked list Π of estates in terms of all y_i

Table 3.3. The generative process of ClusRanking

tory traces. Since the estate investment value depends on the value of the associated latent business area, the K business areas together show the value influence on the estate: $\rho_i = \sum_{k=1}^{K} \left(\frac{d_0}{d_0 + d(i, r_k)} \right)^e \frac{\eta_k}{\sum_{k=1}^{K} \eta_k}$, which is penalized by the distance between area centroid and estate location. After incorporating the three factors, we generate the investment value y_i of real estate i. With all the estate investment values, we compile a ranked list of estates denoted as Π.

3.2.4 Modeling Three Dependencies

Here, we introduce how to model the geographic individual, peer and zone dependencies of estates together in a unified objective function, as shown in Figure 3.1. Let us denote all parameters by $\Psi = \{q, W, \eta, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, the hyperparamters $\Omega = \{\mu_q, \sigma_q^2, \mu_w, \sigma_w^2, \sigma^2\}$, and the observed data collection $\mathcal{D} = \{Y, \Pi, L\}$ where Y, Π and L are the investment value, ranks and locations of I estates respectively. For simplicity, we first assume that $i = \pi_i = \overline{\pi}_i$. In other words, the real estates in $\mathcal{D}$ are sorted and indexed in a descending order in terms of their investment values, which compiles a descending ranks as well.

By Bayesian inference, we have the posterior probability as

$$Pr(\Psi; \mathcal{D}, \Omega) = P\left(\mathcal{D}|\Psi, \Omega\right) P\left(\Psi|\Omega\right) \tag{3.1}$$

The term $P\left(\mathcal{D}|\Psi, \Omega\right)$ is the likelihood of the observed data collection $\mathcal{D}$ as

$$\begin{aligned} P\left(\mathcal{D}|\Psi, \Omega\right) &= P\left(\{Y, \Pi, L\}|\Psi, \Omega\right) \\ &= \underline{P\left(\{Y, L\}|\Psi, \Omega\right)} \times \underline{P\left(\Pi|\Psi, \Omega\right)}, \end{aligned} \tag{3.2}$$

where $P\left(\{Y, L\}|\Psi, \Omega\right)$ denotes the likelihood of the observed investment values and locations of estates given the parameters. $P\left(\{Y, L\}|\Psi, \Omega\right)$ can be explained as to be

proportional to the individual dependency Lik_{id}. $P(\Pi|\Psi,\Omega)$ denotes the likelihood of the ranking of estates given the parameter, which we argue is proportional to the product of peer dependency Lik_{pd} and zone dependency Lik_{zd}. Next, we introduce the modeling of each dependency in detail.

Individual Dependency. The smaller loss, the higher Lik_{id}. Specifically we model Lik_{id} as a joint probability of the estate investment values, the estate locations, and the business areas to learn the geographic interinfluence between estate investment values and locations. As shown in Table 3.3, we assume each location of estate is drawn from a business area and all business areas are drawn from a Multinomial distribution. Along this line, Lik_{id} is formulated by

$$\begin{aligned} Lik_{id} &= \prod_{i}^{I} P(\{y_i, l_i\}|\Psi,\Omega) = \prod_{i}^{I} P(\{y_i, l_i, r_i\}|\Psi,\Omega) \\ &= \prod_{i=1}^{I} \mathcal{N}(y_i|f_i,\sigma) \prod_{i=1}^{I} \mathcal{N}(l_i|\mu_{r_i},\Sigma_{r_i}) \prod_{i=1}^{I} Mult(r_i|\boldsymbol{\eta}) \\ &= \prod_{i=1}^{I} \frac{1}{\sigma} \exp\left(-\frac{(y_i - f_i)^2}{2\sigma^2}\right) \prod_{i=1}^{I} \frac{1}{\Sigma_{r_i}} \exp\left(-\frac{(l_i - \mu_{r_i})^2}{2\Sigma_{r_i}^2}\right) \prod_{i=1}^{I} Mult(r_i|\boldsymbol{\eta}) \end{aligned} \tag{3.3}$$

where we introduce a latent variable $R \in 1 \times I$, each of which r_i represents the latent business area assignment of estate i.

Peer and Zone Dependencies.

While directly modeling likelihood of the ranking list of estates cannot comprehensively capture the spatial correlation of estate-estate and estate-business area, we model the ranking consistency by Lik_{pd} and Lik_{zd} instead. In fact, the ranked list of all the estates indeed can be encoded into a directed graph, $G = \{V, E\}$, with the node set V as estates and the edge set E as pairwise ranking orders. For instance, edge $i \to h$ represents an estate i is ranked higher than estate h. From a generative

modeling angle, edge $i \to h$ is generated by our model through a likelihood function $P(i \to h)$. The more valuable estate i is than estate h, the larger $P(i \to h)$ should be. Since an estate pair $< i, h >$ can be located inside one business area or cross two different business areas, the edges of G then can be categorized into two sets: (1) edges intra business area which corresponds to peer dependency and (2) edges inter business area which corresponds to zone dependency.

Specifically, Lik_{pd} is defined as the ranking consistencies of estate pairs within the same business area. In other words, peer dependency captures the likelihood of the edges intra business area. Here the generative likelihood of each edge $i \to h$ is defined as Sigmoid$(f_i - f_h)$: $P(i \to h) = \frac{1}{1+exp(-(f_i - f_h))}$. Therefore, Lik_{pd} is defined by

$$\begin{aligned} Lik_{pd} &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(i \to h|\Psi, \Omega)^{\mathbb{I}(r_i = r_h)} \\ &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left(\frac{1}{1 + exp(-(f_i - f_h))} \right)^{\mathbb{I}(r_i = r_h)} \end{aligned} \tag{3.4}$$

where $\mathbb{I}(r_i = r_h)$ is the indicator function with $\mathbb{I}(r_i = r_h) = 1$ when estate i and estate h are in the same business area (or $r_i = r_h$), and $\mathbb{I}(r_i = r_h) = 0$ otherwise.

While the peer dependency considers the estate pairs which are within the same business area, zone dependency yet targets the estate pairs, each of which are within two different business areas. We use the generative likelihood of edges inter business area as the zone dependency. There is investment value conformity between estate and business area. That is, the higher prosperity of estate industry in the associated business area, the higher possibility we can draw a high-value estate from it. Thus, when the estate pair $< i, h >$ is drawn from two different business areas $< r_i, r_h >$, we compare the values of the two associated business areas $(r_i \to r_h)$ instead of the values

of estates ($i \to h$). Therefore, the generative likelihood of an inter-business-area edge is define as Sigmoid($\eta_{r_i} - \eta_{r_h}$): $P(i \to h) = \frac{1}{1+exp(-(\eta_{r_i} - \eta_{r_h}))}$, where the values of r_i and r_h are represented by η_{r_i} and η_{r_h} respectively (refer to Section 3.2.3). In this way, we capture the spatial dependency between estate and business area. Lik_{zd} is then given by

$$\begin{aligned} Lik_{zd} &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(r_i \to r_h | \Psi, \Omega)^{\mathbb{I}(r_i \neq r_h)} \\ &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left(\frac{1}{1 + exp(-(\eta_{r_i} - \eta_{r_h}))} \right)^{\mathbb{I}(r_i \neq r_h)}, \end{aligned} \tag{3.5}$$

Second, term $P(\Psi|\Omega)$ is the prior of the parameters Ψ

$$\begin{aligned} &P(\Psi|\Omega) = P(q|\mu_q, \sigma_q^2) P(W|\mu_w, \sigma_w^2) \\ &= \prod_{m=1}^{M} \mathcal{N}(q_m|\mu_q, \sigma_q^2) \times \prod_{m=1}^{M} \prod_{n=1}^{N} \mathcal{N}(w_{mn}|\mu_w, \sigma_w^2) \\ &= \prod_{m=1}^{M} \frac{1}{\sigma_q} \exp\left(-\frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right) \prod_{m=1}^{M} \prod_{n=1}^{N} \frac{1}{\sigma_w} \exp\left(-\frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right) \end{aligned} \tag{3.6}$$

3.2.5 Parameter Estimation

With the formulated posterior probability, the learning objective is to find the optimal estimation of the parameters Ψ that maximize the posterior. Specifically, we use EM mixed with a sampling algorithm. The algorithm iteratively updates the parameters by mutually enhancement between Geo-clustering and estate ranking. The Geo-clustering updates the latent business areas based on locations and the three geographic dependencies; estate ranking learns the estate scores and generate a ranked list.

E-Step. In the E-step, we iteratively draw latent business area assignments for all real estates. For each estate i, we treat its latent business area r as a latent

variable, which is drawn from the posterior of r in terms of the complete likelihood: $r \sim P\left(r|\mathcal{D}, R^{(t)}, \Psi^{(t)}\right)$. More specifically,

$$r \sim P\left(l_i|r, \Psi^{(t)}\right) P\left(\{Y, \Pi\}|r, \Psi^{(t)}\right) P\left(r|\boldsymbol{\eta}^{(t)}\right) \tag{3.7}$$

where

$$P\left(l_i|r, \Psi^{(t)}\right) = \mathcal{N}\left(l_i|\mu_r^{(t)}, \Sigma_r^{(t)}\right) \tag{3.8}$$

$$\begin{aligned} P\left(\{Y, \Pi\}|r, \Psi^{(t)}\right) = P(y_i|f_i, \sigma^2) \prod_{h=i+1}^{I} P(i \to h|r, \Psi^{(t)})^{\mathbb{I}(r_i = r_h)} \\ \prod_{h=i+1}^{I} P(r_i \to r_h|r, \Psi^{(t)})^{\mathbb{I}(r_i \neq r_h)} \end{aligned} \tag{3.9}$$

Here the latent business area assignment of real estate e_i is updated by three effects: (1) $P(r|\boldsymbol{\eta}^{(t)})$ updates business area assignment in terms of the prosperity distribution of multiple business areas ; (2) $P\left(l_i|r, \Psi^{(t)}\right)$ is the location emission probability given the latent business area as a hidden spatial state. (3) $P\left(\{Y, \Pi\}|r, \Psi^{(t)}\right)$ updates business area assignment by both prediction accuracy and ranking consistency.

When the latent business area assignment of each estate is updated, we further update the neighborhood popularity $\delta_i = \frac{1}{J}\sum_{j=1}^{J} \frac{\phi_{ij}}{max_{i \in r}\{\phi_{ij}\}}$, because the normalization term is conditional on the updated business area r_i.

M-Step. In the M-step, we maximize the log likelihood of the model given the business area assignments R are fixed in the E-step. Since business area assignments

are known, we can update $\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r, \boldsymbol{\eta}$ directly from the samples.

$$\begin{aligned}\boldsymbol{\mu}_r &= \frac{1}{\#(i,r)} \sum_{i=1}^{I} \mathbb{I}(r_i = r) l_i \\ \boldsymbol{\Sigma}_r &= \frac{1}{\#(i,r) - 1} \sum_{i=1}^{I} \mathbb{I}(r_i = r) \left((l_i - \mu_r)^\top (l_i - \mu_r) \right)\end{aligned} \tag{3.10}$$

where $\#(i,r)$ is the number of real states assigned to region r. Through imposing a conjugate Dirichlet prior $\text{Dir}(\boldsymbol{\gamma})$, we update $\boldsymbol{\eta}^{(t+1)}$ by

$$\boldsymbol{\eta}_r^{(t+1)} = \frac{C_r^{(t+1)} + \gamma}{C^{(t+1)} + |R|\gamma} \tag{3.11}$$

where $C_r = \sum_{i \in r} y_i$, $C = \sum y_i$ and $\gamma = \frac{1}{K}$.

Note that the centers ($\boldsymbol{\mu}$) and estate investment values ($\boldsymbol{\eta}$) of latent business areas are updated, so updated is the influence of latent business areas $\rho_i = \sum_{k=1}^{K} \left(\frac{d_0}{d_0 + d(i,r_k)} \right)^e \frac{\eta_k}{\sum_{k=1}^{K} \eta_k}$.

After updating the parameters $\{\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and latent business area assignments R, we update $\Psi^{(t+1)}$ that maximizes the log of posterior

$$\begin{aligned}&\mathcal{L}(q, W | R^{(t+1)}, \mathcal{D}) = \\ &\sum_{i=1}^{I} \left[-\frac{1}{2} ln\sigma^2 - \frac{(y_i - f_i)^2}{2\delta^2} \right] + \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} ln \frac{1}{1 + exp(-(f_i - f_h))} \mathbb{I}(r_i = r_h) \\ &+ \sum_{m=1}^{M} \left[-\frac{1}{2} ln\sigma_q^2 - \frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right] + \sum_{m=1}^{M} \sum_{n=1}^{N} \left[-\frac{1}{2} ln\sigma_w^2 - \frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right]\end{aligned} \tag{3.12}$$

We apply a gradient descent method to update q, W through $q_m^{t+1} = q_m^t - \epsilon \frac{\partial(-\mathcal{L})}{\partial q_m}$ and $w_{mn}^{t+1} = w_{mn}^t - \epsilon \frac{\partial(-\mathcal{L})}{\partial w_{mn}}$

$$\begin{aligned}\frac{\partial(\mathcal{L})}{\partial q_m} &= \sum_{i=1}^{I} \frac{(y_i - f_i) w_{m\cdot} e_i}{\sigma^2} + \sum_{m=1}^{M} -\frac{q_m - \mu_q}{\sigma_q^2} + \\ &\sum_{i=1}^{I-1} \sum_{h=i+1}^{I} \frac{exp(f_h - f_i) w_{m\cdot}(e_i - e_h)}{1 + exp(f_h - f_i)} \mathbb{I}(r_i = r_h)\end{aligned} \tag{3.13}$$

$$\frac{\partial(\mathcal{L})}{\partial w_{mn}} = \sum_{i=1}^{I} \frac{(y_i - f_i) q_m e_{in}}{\sigma^2} + \sum_{m=1}^{M} -\frac{w_{mn} - \mu_w}{\sigma_w^2} + \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} \frac{exp(f_h - f_i) q_m (e_{in} - e_{hn})}{1 + exp(f_h - f_i)} \mathbb{I}(r_i = r_h) \tag{3.14}$$

3.2.6 Ranking Inference

After parameters Ψ are estimated via maximizing the posterior probability, which essentially captures both prediction accuracy of estate investment value and the ranking consistence of estates, we will obtain the learned model for investment value of estate, i.e., $\mathbb{E}(y_i|q, e_i) = \gamma_i + \delta_i + \rho_i$ given a rising or falling market period. For a new coming estate k, we may predict its investment value accordingly. The larger the $\mathbb{E}(y_k|q, e_k)$ is, the higher investment value it has. With the predicted investment values for all new estates, we are able to compile a ranking list of those estate.

3.3 Experimental Results

In this section, we provide an empirical evaluation of the performances of the proposed ClusRanking method on real-world estate data.

3.3.1 Experimental Data

Table 4.3 shows four data sources. The transportation data set includes the data about the bus system, the subway system, and the road network in Beijing, China. Also, we extract POI features from the Beijing POI dataset. Moreover, mobility patterns are extracted from the taxi GPS traces. In Beijing, taxi traffic contributes more than 12 percent of the total traffic, and thus reflects a significant portion of human mobility (Yuan, Zheng, & Xie, 2012b). Finally, we crawl the Beijing estate

Data Sources	**Properties**	**Statistics**
Real estates	Number of real estates	2,851
	Size of bounding box (km)	40*40
	Time period of transactions	04/2011 - 09/2012
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 Trajectories	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

Table 3.4. Statistics of the experimental data.

data from www.soufun.com, which is the largest real-estate online system in China.

In estate industry, the estate return rate is used to measure the investment value of an estate. The estate return rate is the ratio of the price increase relative to the start price of a market period as $r = \frac{P_f - P_i}{P_i}$, where P_f and P_i denote the final price

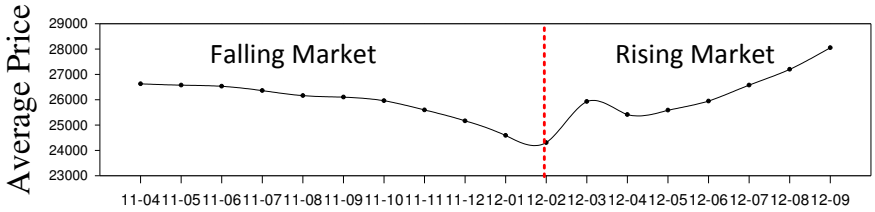


Figure 3.2. The rising market period and the falling market period in Beijing.

and the initial price, respectively.

To prepare the benchmark investment values of estates (Y) for training data, we first calculate the return rate of each estate during a given market period. We then sort the return rates of all the estates in a descending order. Finally, we cluster them into five clusters using variance based top-down hierarchical clustering. In this way, we segment the estates into five ordered value categories (i.e., $4 > 3 > 2 > 1 > 0$, the higher the better).

By discretizing estate return rates into five categories, we can understand estate investment potentials and reduce the noise led by the small fluctuations in return rates.

Finally, a list of estates, each of which with the extracted features and investment values, are split into two data sets in terms of the falling market period (from Jul. 2011 to Feb. 2012) and the rising market period (from Feb. 2012 to Sep. 2012) as shown in Figure 4.4.

3.3.2 Evaluation Metrics

To show the effectiveness of the proposed model, we use the following metrics for evaluation.

Normalized Discounted Cumulative Gain. The discounted cumulative gain

(DCG@N) is given by

$$DCG[n] = \begin{cases} rel_1 \quad if\, n = 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n}, \quad if\, n >= 2 \end{cases} \tag{3.15}$$

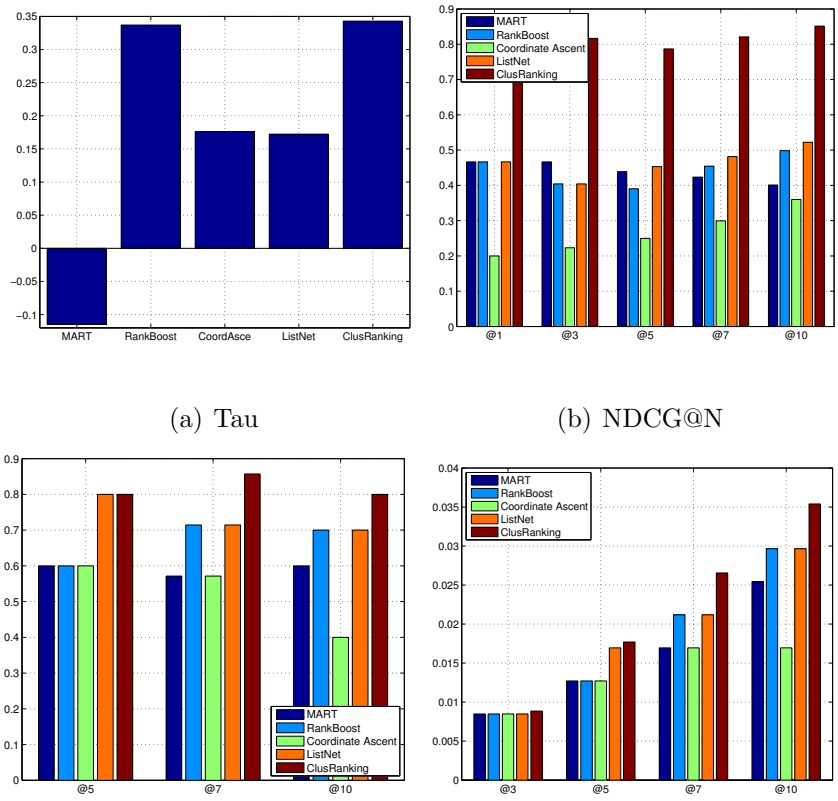
Later, given the ideal discounted cumulative gain $DCG^{'}$, NDCG at the n-th position can be computed as NDCG[n]= $\frac{DCG[n]}{DCG^{'}[n]}$. The larger NDCG@N is, the higher top-N ranking accuracy is.

Precision and Recall. Since we use a five-level rating system ($4 > 3 > 2 > 1 > 0$) instead of binary rating, we treat the rating ≥ 3 as "high-value" and the rating $<$ 3 as "low-value". Given a top-N estate list E_N sorted in a descending order of the prediction values, precision and recall are defined as Precision@$N = \frac{|E_N \bigcap E_{\geq 3}|}{N}$ and Recall@$N = \frac{|E_N \bigcap E_{\geq 3}|}{|E_{\geq 3}|}$, 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, for an estate pair $< i, j >$, $< i, j >$ is said to be concordant, if both $y_i > y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i < f_j$. Also, $< i, j >$ 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 = $\frac{\#_{conc} - \#_{disc}}{\#_{conc} + \#_{disc}}$.

3.3.3 Baseline Algorithms

To show the effectiveness of the proposed method, we compare the ranking accuracy of our methods against following baseline algorithms. (1) **MART (Friedman, 2001):** it is a boosted tree model, specifically, a linear combination of the outputs of a set of regression trees. (2) **RankBoost (Freund et al., 2003):** it is a boosted



(a) Tau (b) NDCG@N

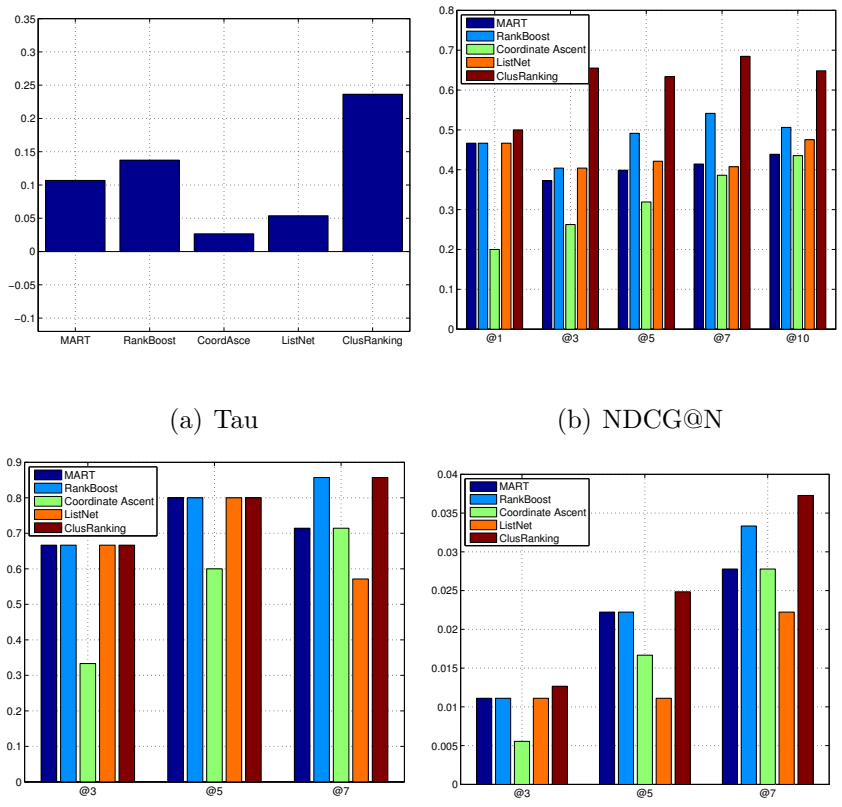
(c) Precision@N (d) Recall@N

Figure 3.3. The overall performances on the rising market dataset.

pairwise ranking method, which trains multiple weak rankers and combines their outputs as final ranking. (3) **Coordinate Ascent (Metzler & Croft, 2007):** it uses domination loss and applies coordinate descent for optimization. (4) **ListNet (Cao et al., 2007):** it is a listwise ranking model with permutation top-k ranking likelihood as the objective function.

For the baseline algorithms, we use RankLib[2] . We set the number of trees = 1000, the number of leaves = 10, the number of threshold candidates = 256, and the learning rate = 0.1 for MART. For RankBoost, we set the number of iteration = 300, the number of threshold candidates = 10. Regarding Coordinate Ascent, we

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



(a) Tau (b) NDCG@N

(c) Precision@N (d) Recall@N

Figure 3.4. The overall performances on the falling market dataset.

set step base = 0.05, step scale = 2.0, tolerance = 0.001, and slack = 0.001. For our model, we set β_1=0.8 and β_2=25m. We set $d_0 = 1$ and $d(i, r_k)$ is computed based on degree (°) instead of mile or km for simplicity. We set latent business areas K=10 and initialize the mean and covariance of the locations of each business area by Kmeans clustering. Finally, we set $\eta = \frac{1}{K}$, $\mu_q = \mu_w = 0$, $\sigma_q = \sigma_w = \sigma = 35$ and M=3 for hyperparameters.

The codes are implemented in R (modeling), Python (preprocessing), and Matlab (visualization). The experiments were performed on a x64 machine with Intel i5 2.60GHz dual-core CPU and 16GB RAM. The operation system is Microsoft Windows 7 Professional.

3.3.4 Overall Performances

We provide the performance comparison on the rising market dataset and the falling market dataset in terms of Tau, NDCG, Precision and Recall.

Rising Market Data. Figure 3.3(a) shows the comparison of Kendall's Tau Coefficient. Our method achieves 0.3428617 and outperforms the baselines. Figure 3.3(b) shows the NDCG comparison. Our method achieves 0.75 NDCG@1, 0.81 NDCG@3, 0.78 NDCG@5, 0.82 NDCG@7, and 0.85 NDCG@10 whereas the NDCGs of the four baselines only range from 0.2 to 0.61. Figure 3.3(c) and Figure 3.3(d) respectively show the precision@N and recall@N. In Precision, ClusRanking ¿ ListNet ¿ MART, RankBoost, Coordinate Ascent. In Recall, ClusRanking achieves 0.0088 recall@3, 0.017 recall@5, 0.026 recall@7, and 0.035 recall@10, which in overall outperforms ListNet, MART, RankBoost, Coordinate Ascent with a significant margin.

Falling Market Data. Figure 3.4 shows the comparison in terms of Kendall's Tau. Our method achieves a higher accuracy at 0.2363498 than four baselines. We also compare all the five methods in terms of NDCG, Precision and Recall. Our method achieves around 0.65 NDCG@3, 0.63 NDCG@5, 0.68 NDCG@7, and 0.64 NDCG@10 whereas the NDCGs of the four baselines are lower than 0.6111. Moreover, the Precision@3,5,7 of our method are relatively higher than the baselines in overall. Finally, our method achieves 0.012 recall@3, 0.024 recall@5, and 0.037 recall@7, which are generally better than RankBoost but significantly outperforms MART, Coordinate Ascent and ListNet.

The above overall performances validate the effectiveness of our ClusRanking

method.

3.3.5 The Study on Geographic Dependencies

Here, we study the impact of three geographic dependencies. Specifically, we designed three internal competing methods in terms of variants of posterior likelihood $Pr(\Psi; \mathcal{D}, \Omega) = P(\mathcal{D}|\Psi, \Omega) P(\Psi|\Omega)$: (1) **Individual Dependency (ID)**, in which we only consider the individual dependency as the objective function. In other words, $P(\mathcal{D}|\Psi, \Omega) = Lik_{id}$. (2) **Peer Dependency (PD)**, in which we only consider the peer dependency as the objective function. (3) **Peer Dependency + Zone Dependency (PD+ZD)**, in which we consider the combination of peer and zone dependencies as the objective function. (4) **Combination (ClusRanking)**, in which we consider individual, peer, and zone dependencies simultaneously. This is exactly our method: $P(\mathcal{D}|\Psi, \Omega) = Lik_{id} \times Lik_{pd} \times Lik_{zd}$.

Rising Market Data. Table 3.5 shows the performance comparison on the rising market data in terms of Tau and NDCG. It is clear that our method achieves around 0.81 NDCG@3, 0.78 NDCG@5, 0.82 NDCG@7 and 0.85@10 on the rising market data, which outperforms PD+ZD, PD, and ID. In the Tau comparison, the results lead to: ClusRanking ¿ PD ¿ ID ¿ PD+ZD. From Table 3.5, we conclude that (1) the strategy of capturing three dependencies helps ClusRanking to achieve the highest Tau and NDCG; (2) considering both peer and zone dependencies enhances the top-k accuracy but degrades the overall ranking comparing to individual dependency only. This might be because the peer and zone dependencies better capture the ranking consistency of estates than the individual dependency, as individual dependency

Metric	@N	ID	PD	PD+ZD	ClusRanking
NDCG	3	0.5599531	0.6549766	0.6900469	**0.8166009**
	5	0.5771226	0.6024622	0.6101556	**0.7867076**
	7	0.587992	0.6048394	0.641282	**0.8208795**
	10	0.6518163	0.6723095	0.694175	**0.8513267**
Tau	-	0.2494531	0.2535907	0.2203712	**0.3428617**

Table 3.5. Performance comparison of different geographic dependencies on the rising market data.

indeed models the prediction accuracy of the observed data collection $\{Y, L\}$.

Falling Market Data. Table 3.6 shows the performance comparison of different geographic dependencies on the falling market data. It is clear that our method outperforms ID, PD and PD+ZD. PD+ZD achieves the second highest NDCG. Moreover, ClusRanking > PD+ZD > PD > ID in terms of Kendall's Tau.

Metric	@N	ID	PD	PD+ZD	ClusRanking
NDCG	3	0.570193	0.5950234	0.6250234	**0.6549766**
	5	0.6144799	0.6004235	0.6144799	**0.633635**
	7	0.6196808	0.654487	0.6196808	**0.6845354**
	10	0.6415102	0.6252658	0.6307051	**0.6482665**
Tau	-	0.1186736	0.1313437	0.1433408	**0.2363498**

Table 3.6. Performance comparison of different geographic dependencies on the falling market data.

This experiment not only justifies the spatial autocorrelation of estate investment

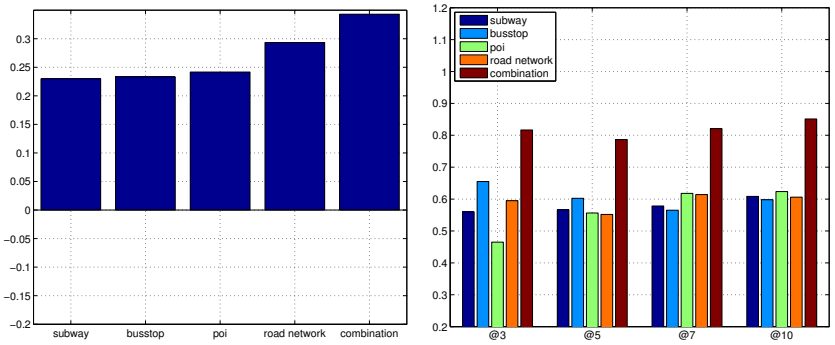
values (e.g., individual, estate-estate peer, estate-business area), but also shows the advantages of considering three geographical dependencies .

3.3.6 The Study on Geographic Features

We compare the performances of ClusRanking with different geographic feature sets (i.e., subway, bus stop, POI, and road network) over rising and falling markets.

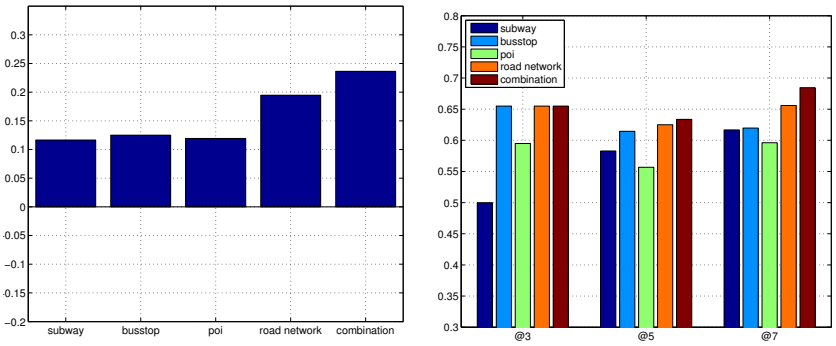
Rising Market Data. First, Figure 3.5(a) shows the performance comparison of the five feature sets in terms of Tau: combination ¿ road network ¿ bus stop, subway and poi. Next, Figure 3.5(b) shows the NDCG@N of different feature sets (N=3, 5, 7, 10 respectively). As can be seen, the combination of all the four feature sets achieves 0.81 NDCG@3, 0.78 NDCG@5, 0.82 NDCG@7, 0.85 NDCG@10, and outperforms the other four individual feature sets. Moreover, the NDCGs of the bus stop and road network feature sets are lower than combination but higher than the POI and subway feature sets. Finally, we can conclude that, in rising market, the combination of all geographic information is the best. Road network outperforms bus stop, subway and POI. Bus stop is more suitable for top-k ranking than road network whereas road network performs better than bus stop in overall ranking.

Falling Market Data. Figure 3.6(a) shows a comparison of the five feature sets on Tau: combination ¿ road network ¿ bus stop, subway and poi. This result is consistent with that of rising market data. Regarding top-k ranking, Figure 3.6(b) shows the NDCG@N (N=3, 5, 7 respectively) of different feature sets in terms of ClusRanking. First, the POI feature set achieves the worst performance in NDCG@5,7. Second, the road network feature set achieves the second highest NDCGs@3,5,7. Finally, the



(a) Tau (b) NDCG@N

Figure 3.5. Performance comparison of different geographic features on rising market data.

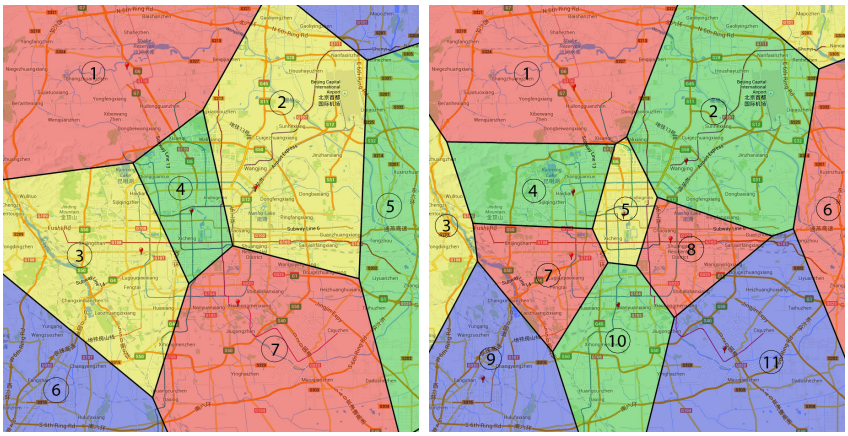


(a) Tau (b) NDCG@N

Figure 3.6. Performance comparison of different geographic features on falling market data.

combination of all the four feature sets outperforms all the individual feature sets. In summary, in falling market, combination ¿ bus stop ¿ subway, road network, and POI.

The results validate the effectiveness of using multiple information fusion (subway, bus stop, POI and road network).



(a) Kmeans (b) ClusRanking

Figure 3.7. A comparison of the learned business areas within the Beijing Fifth Ring (K=10).

3.3.7 Implication of Latent Business Areas

Our model also provides a unique understanding of the latent business areas of Beijing from an estate perspective. Figure 3.7 clearly shows our method, learned from geography, mobility and estate data, is more reasonable than K-means, which simply cluster the estates by location information. For instance, in Figure 3.7(b), NO.4 area, named Zhongguancun, is the Chinese Silicon Valley and is famous for high-tech companies. This area is a high density cluster of human mobility, estates and POIs. However, in Figure 3.7(a), the Zhongguancun area is improperly separated into NO.3 and NO.4 area by K-means. Another example is the NO.2 and NO.8 areas, namely Wangjing and CBD respectively, in Figure 3.7(b). Wangjing is a quick-growing residential sub-center with easy-access transportation and luxury apartments. Currently, about 203,000 young people, including company executives, white-collar workers, expatriates and returnees, are living in Wangjing. CBD is the Center Business District

with numerous financial business offices, culture media companies and high-end enterprise information services. However, in Figure 3.7(a), Wangjing and CBD are improperly united into NO.2 area by K-means. The visualization results show the effectiveness of ClusRanking learned from multi-source estate related data and the effectiveness of capturing the three geographic dependencies as the objective function.

3.3.8 Hierarchy of Needs for Human Life

We show how our ranking results can be used to understand the hierarchy of human needs from a POI aspect. Figure 3.8 shows the estate-POI density spectrum. From left to right, x-axis represents the estate rankings in the descending order. From up to down, y-axis represents POI categories in the descending order in terms of POI numbers. Several interesting findings can be drawn from Figure 3.8. First, the upper half are darker than the lower half. This indicates POI categories in the upper half are more important than those in the lower half. In other words, people prefer their homes near schools, malls, offices, restaurants, and transits. Whereas, hotels, hospitals, sports and scene spots are not must-have POIs to be located close to living places. Second, along x-axis, the POI density spectrum of the left-side high-ranked estates is evenly distributed for smooth whereas the POI density spectrum of the right-side low-ranked estates are non-smooth. This illustrates high-value estates usually balance the needs of human beings. Third, we calculate the average POI density of each POI category based on the top 2000 estates. We then sort all POI categories in terms of POI densities, show the smoothed POI density curve and find three inflection points. Later, we segment those POI categories into four clusters

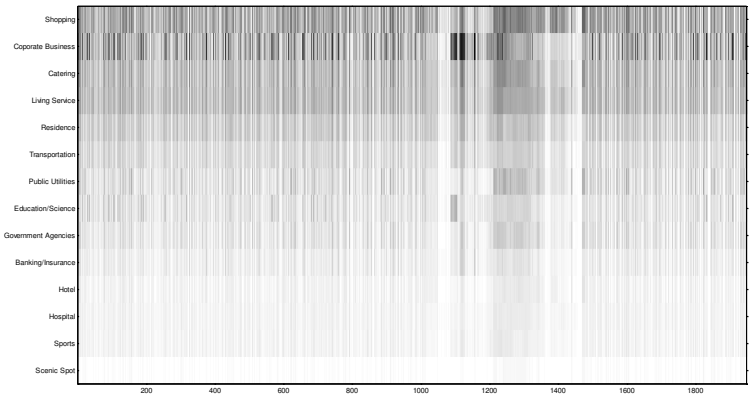


Figure 3.8. The POI density spectral of estates over multiple poi categories

using the three inflection points. Finally, we present a triangle structure of needs of Beijing citizens as shown in Figure 3.9. The higher, the more fundamental and urgent in human needs.

Figure 3.9. The triangle need hierarchy of Beijing

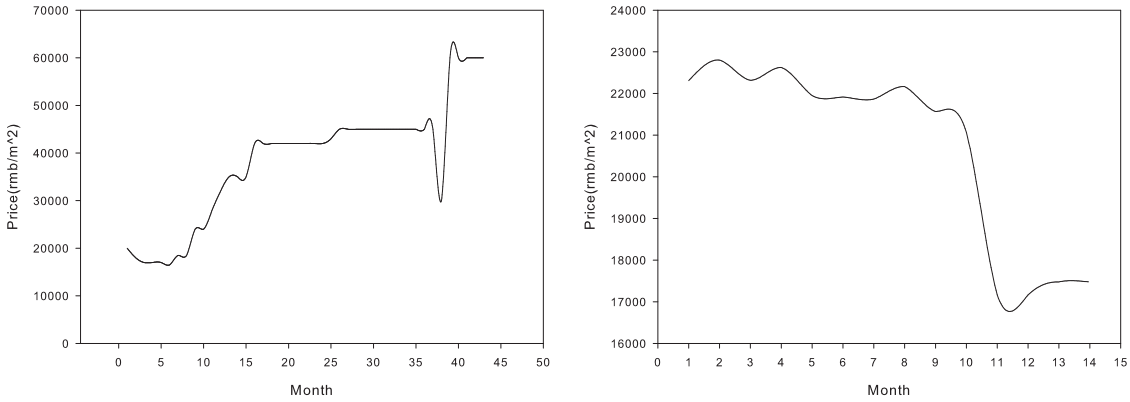
3.3.9 A Case Study

Here, we present a case study. First, we select one high-ranked estate called "Red Hill Family" (RHF) and one low-ranked estate called "Jiuxianqiao Road No. 11"

(JR11) from our ranking results. Then, we compare RHF with JR11 from historical transaction prices. As can be seen in Figure 3.10, during the past 43 months, the prices of RHF increase in both rising and falling markets. However, for the past 15 months, the overall prices of JR11 continuously fall even in the rising market.

To show why, we first check the neighborhood profiles (individual dependency) of two estates. Specifically, we extract geographic and mobility features of the neighborhoods of RHF and JR11, respectively. Table 3.7 shows RHF has higher road network density, larger amount of POIs (especially schools), bus stops and subway stations, and higher neighborhood popularity than JR11. It thus is reasonable that people are willing to afford higher prices to RHF than JR11. This validates the individual dependency. Besides, RHF is located in the prosperous area of MuXiDi (inside No. 7 area in Figure 3.7(b)) near the 2nd ring road whereas JR11 is located in the area of DongFengXiang (inside No.2 area in Figure 3.7(b)) outside the fifth ring road. The average rating of estates in MuXiDi is round to 3, which is higher than that (round to 1) of estates in DongFengXiang. This justifies the zone dependency.

Traditional learning to ranking (LTR) methods feed document feature vectors into predictive models (such as regression, tree based models, neural network) and optimize the model over objective functions, which describe the ranking accuracy in a point-wise, pair-wise or list-wise manner. In real estate ranking, LTR simply represents estates as feature vectors, optimizes a general ranking accuracy metric, and thus fails to achieve higher performance. However, our method extracts the geographic utility and neighborhood popularity by strategically mining geography and mobility data. Besides, our method model the implicit influence of latent business



(a) Red Hill Family (b) Jiuxianqiao Road No. 11

Figure 3.10. Price Trend Comparison.

area via ClusRanking. Moreover, ClusRanking simultaneously captures geographic individual, peer and zone dependencies as objective function. Hence, we can observe significant improvements against baselines.

3.4 Related Work

Related work can be grouped into two categories. The first one includes the work on estate appraisal. In the second category, we present the ranking related methods.

Traditional research on estate appraisal are based on financial estate theory, typically constructing an explicit index of estate value (Krainer & Wei, 2004). More studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of estate prices. Work (Downie & Robson, 2007) checks the volatility of estate price and concludes that low investment-valued estate values relatively volatile. Work (Chaitra H. Nagaraja & Zhao, 2009) applies an autoregression method to learn the trend and periodicity of price and predicts estate value. More studies are conducted from an econometric angle, for example, hedonic methods and repeat

Type	Name	RHF	JR11
transportation	bus stop(1km)	12	3
	subway(3km)	9	0
	shortest distance to subway	1061	3597
	road network level-2 entry(3km)	102	46
POI number (1km)	catering	146	17
	shopping	127	18
	living	201	16
	sports	27	3
	healthcare	44	2
	education	67	13
	finance	55	1
	public facility	79	10
popularity	average accumulated visit probability	1.64e+7	1.36e+6

Table 3.7. A comparison of transportation, POI and mobility of RHF and JR11

sales methods. The hedonic methods (Taylor, 2003; Assil, 2012) assume the price of a property depends on its characteristics and location. The repeat sales methods (Assil, 2012; Bailey, Muth, & Nourse, 1963; Shiller, 1991b) construct a predefined price index based on properties sold more than once during the given period. Recent works (Downie & Robson, 2007; Mitropoulos, Wu, & Kohansky, 2007) study the automated valuation models, which aggregate and analyze physical characteristics and

sales prices of comparable properties to provide property valuations. More recent studies (Pace, 1998; Kontrimas & Verikas, 2011; Lam, 1996; Bailey et al., 1963) shift to computational estate appraisal and apply general additive mode, support vector machine regression, multilayer perceptron and ensemble method to evaluate estate value.

Also, our work can be categorized into Learning-To-Rank (LTR). The LTR methods are threefold: point-wise, pair-wise and list-wise. The point-wise methods (Fuhr, 1989; Cooper, Gey, & Dabney, 1992) reduce the LTR task to a regression problem: given a single query-document pair, predict its score. The pair-wise methods, such as RankBoost (Freund et al., 2003), RankSVM (Herbrich, Graepel, & Obermayer, 1999) and LambdaRank (Quoc & Le, 2007), approximate the LTR task as a classification problem and learn a binary classifier that can tell which document is better in a given document pair. The list-wise methods, such as AdaRank (Xu & Li, 2007), LambdaMART (Burges, 2010) and ListNet (Cao et al., 2007), optimize a ranking loss metric over lists instead of document pairs. Works (Weng & Lin, 2011; Rendle, Freudenthaler, Gantner, & Schmidt-Thieme, 2009; Gantner, Drumond, Freudenthaler, & Schmidt-Thieme, 2012) provide full Bayesian explanations and optimize the posterior of point-wise, pair-wise and list-wise ranking models. Study (Shi, Larson, & Hanjalic, 2012) further unifies both rating error and ranking error as objective function to enhance Top-K recommendation. There are also studies that improve ranking performance by semi-supervised learning through exploiting the disagreement between two learners (Zhou, Chen, & Dai, 2006) or combining supervised and unsupervised ranking models (Li, Li, & Zhou, 2009).

Furthermore, our work has a connection with recent studies of exploring the geographic influence for POI recommendation. Works (Cheng, Yang, King, & Lyu, 2012; Fu, Liu, Ge, Yao, & Xiong, 2014) consider the multi-center of user check-in patterns and apply a static pre-clustering method to extract the influence of geographic proximity in choosing a POI. Work (B. Liu, Fu, Yao, & Xiong, 2013) exploits multi-center user mobility and embeds a POI clustering method into matrix factorization. Finally, our work is related to studies of city region function via geographic topic modeling using POI and mobility (Zheng et al., 2014).

3.5 Conclusion

In this chapter, we proposed a ClusRanking method for ranking estates based on their investment values. Specifically, this method has the ability in capturing the geographic individual, peer, and zone dependencies via ClusRanking by exploiting various estate related data. Also, our method has two advantages. First, for predictive modeling, we establish a hierarchical generative structure to capture both explicit factors (i.g., geographic utility and neighborhood popularity) and latent influences (e.g., the influence of latent business area) based on the estate data. This generative structure profiles, filters, aggregates and fuses multi-source information to predict estate investment values. It helps to take advantage of rich estate-related data sources. Second, in the learning framework, we leverage the mutual enforcement of ranking and clustering power. In addition, we simultaneously consider three dependencies and construct an estate-specific ranking likelihood as the objective function for enhancing model learning. Finally, the experimental study demonstrates the effectiveness of our

method on real-world estate-related data over several alternative methods.

CHAPTER 4

EXPLORING MIXED LAND USE FOR REAL ESTATE RANKING

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 chapter, 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.

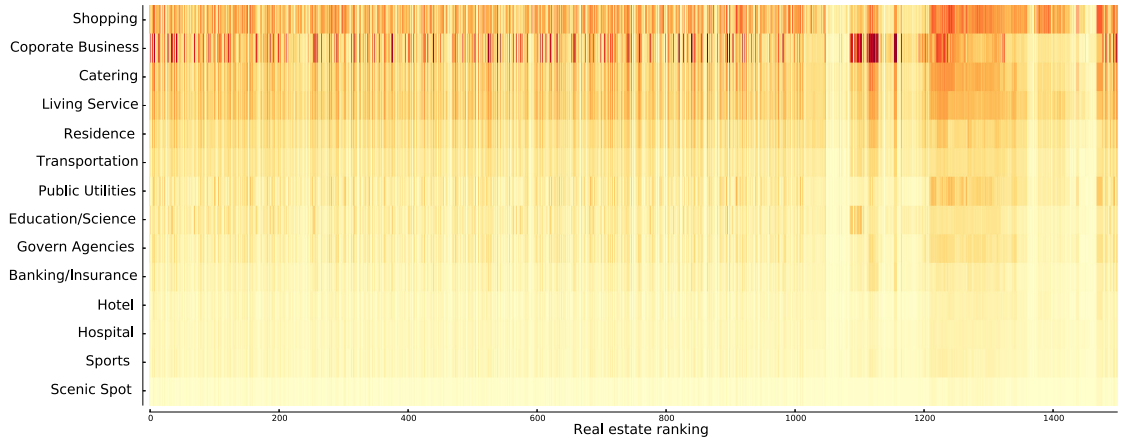


Figure 4.1. The POI density spectrum of estates over multiple POI categories.

4.1 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 (Song & Knaap, 2004). 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.

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 (Song & Knaap, 2004; Koster & Rouwendal, 2012; Loehr, 2013). 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% (Koster & Rouwendal, 2012). Indeed, Figure 4.1 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 (Loehr, 2013) revealed that, within a certain range, proximity to commercial uses 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 (Yuan et al., 2012a). 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, 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 (Koster & Rouwendal, 2012;

Fu, Ge, et al., 2014). 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, 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.

4.2 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.

4.2.1 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 chapter. 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.

4.2.2 Framework Overview

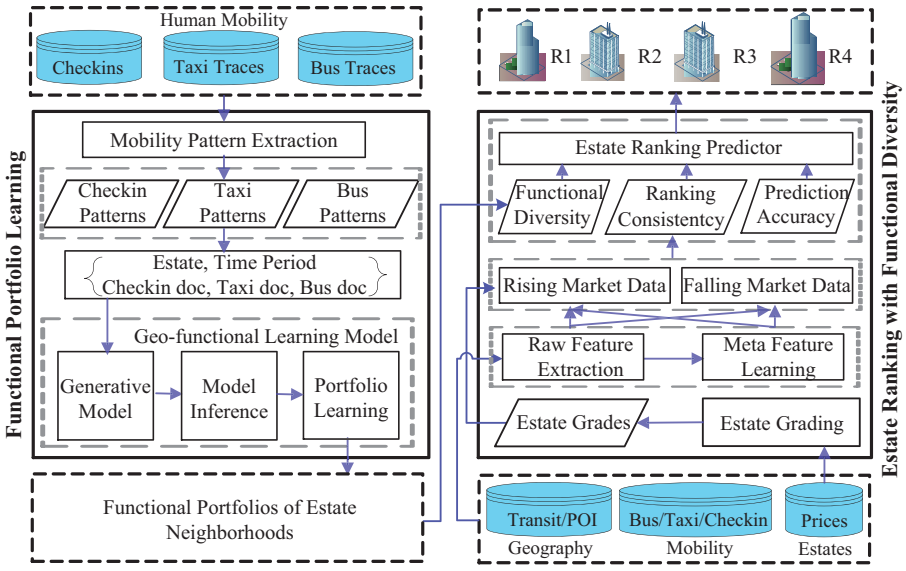


Figure 4.2. The framework overview of geographical functional ranking for estates.

Figure 4.2 shows the framework of our geographic functional ranking. This framework consists of two major stages.

(1) Functional Portfolio Learning. As shown in Figure 4.2, 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 (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 4.3 addresses the first problem of portfolio learning, and Section 4.4 of estate ranking.

4.3 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.

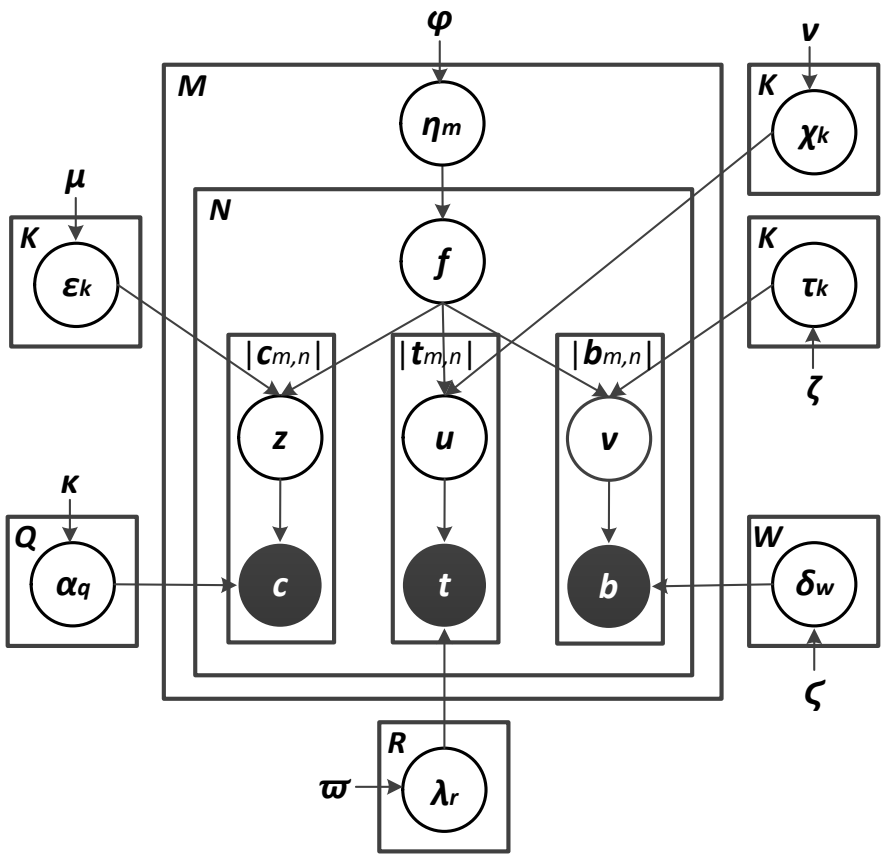


Figure 4.3. The graphical representation of the proposed geographic functional learning model.

4.3.1 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.

4.3.2 Model Specification

Figure 4.3 shows the graphical representation of our geographic functional learning model. Specifically, we use a multinomial distribution η_m over K latent functions to model 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

Table 4.1. 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 $b \sim P(b|\delta_v)$.

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 4.1 shows the generative process.

4.3.3 Model Inference

Let us denote all parameters by $\Psi = \{\boldsymbol{\eta}, \boldsymbol{\epsilon}, \boldsymbol{\chi}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\delta}\}$ 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 4.1, the joint distribution can be factored

as

$$
\begin{aligned}
&P(D, \Upsilon, \Psi|\Omega) = P(D, \Upsilon|\Psi)P(\Psi|\Omega) \\
&= P(\boldsymbol{C}|\boldsymbol{\alpha})P(\boldsymbol{\alpha}|\kappa)P(\boldsymbol{T}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|\varpi)P(\boldsymbol{B}|\delta)P(\boldsymbol{\delta}|\varsigma)P(\boldsymbol{Z}|\boldsymbol{\epsilon})\times \\
&P(\boldsymbol{\epsilon}|\mu)P(\boldsymbol{U}|\boldsymbol{\chi})P(\boldsymbol{\chi}|\nu)P(\boldsymbol{V}|\boldsymbol{\tau})P(\boldsymbol{\tau}|\zeta)P(\boldsymbol{F}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\rho).
\end{aligned} \tag{4.1}
$$

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

$$\begin{aligned} P(f_{m,n}=k|\mathcal{D},\Upsilon-f_{m,n}) &= \frac{\mathbb{F}_{m,k}^{-(m,n)}+\rho_k}{\sum_{f=1}^{K}\mathbb{F}_{m,f}^{-(m,n)}+\rho_f} \\ &\times \frac{\prod_{z=1}^{Q}\Gamma(\mathbb{Z}_{k,z}+\mu_z)\Gamma(\sum_{z=1}^{Q}\mathbb{Z}_{k,z}^{-(m,n)}+\mu_z)}{\prod_{z=1}^{Q}\Gamma(\mathbb{Z}_{k,z}^{-(m,n)}+\mu_z)\Gamma(\sum_{z=1}^{Q}\mathbb{Z}_{k,z}+\mu_z)} \\ &\times \frac{\prod_{u=1}^{R}\Gamma(\mathbb{U}_{k,u}+\nu_u)\Gamma(\sum_{u=1}^{R}\mathbb{U}_{k,u}^{-(m,n)}+\nu_u)}{\prod_{u=1}^{R}\Gamma(\mathbb{U}_{k,u}^{-(m,n)}+\nu_u)\Gamma(\sum_{u=1}^{R}\mathbb{U}_{k,u}+\nu_u)} \\ &\times \frac{\prod_{v=1}^{W}\Gamma(\mathbb{V}_{k,v}+\zeta_v)\Gamma(\sum_{v=1}^{W}\mathbb{V}_{k,v}^{-(m,n)}+\zeta_v)}{\prod_{v=1}^{W}\Gamma(\mathbb{V}_{k,v}^{-(m,n)}+\zeta_v)\Gamma(\sum_{v=1}^{W}\mathbb{V}_{k,v}+\zeta_v)}. \end{aligned} \tag{4.2}$$

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

$$\begin{aligned} &P(z_{m,n,i}=q|D,\Upsilon-z_{m,n,i}) \\ &= \frac{\mathbb{C}_{q,c_{m,n,i}}^{-(m,n,i)}+\kappa_{c_{m,n,i}}}{\sum_{c=1}^{|\boldsymbol{P}_c|}\mathbb{C}_{q,c}^{-(m,n,i)}+\kappa_c}\frac{\mathbb{Z}_{f_{m,n},q}^{-(m,n,i)}+\mu_q}{\sum_{z=1}^{Q}\mathbb{Z}_{f_{m,n},z}^{-(m,n,i)}+\mu_z)}. \end{aligned} \tag{4.3}$$

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

$$\begin{aligned} &P(u_{m,n,i}=r|D,\Upsilon-u_{m,n,i}) \\ &= \frac{\mathbb{T}_{r,t_{m,n,i}}^{-(m,n,i)}+\varpi_{t_{m,n,i}}}{\sum_{t=1}^{|\boldsymbol{P}_t|}\mathbb{T}_{r,t}^{-(m,n,i)}+\varpi_t}\frac{\mathbb{U}_{f_{m,n},r}^{-(m,n,i)}+\nu_r}{\sum_{u=1}^{R}\mathbb{U}_{f_{m,n},u}^{-(m,n,i)}+\nu_u}. \end{aligned} \tag{4.4}$$

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

$$\begin{aligned} &P(v_{m,n,i}=w|D,\Upsilon-v_{m,n,i}) \\ &= \frac{\mathbb{B}_{w,b_{m,n,i}}^{-(m,n,i)}+\varsigma_{b_{m,n,i}}}{\sum_{b=1}^{|\boldsymbol{P}_b|}\mathbb{B}_{w,b}^{-(m,n,i)}+\varsigma_b}\frac{\mathbb{V}_{f_{m,n},w}^{-(m,n,i)}+\zeta_w}{\sum_{v=1}^{W}\mathbb{V}_{f_{m,n},v}^{-(m,n,i)}+\zeta_v}. \end{aligned} \tag{4.5}$$

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 $\theta_f = \frac{\sum_{m=1}^{M}\eta_{m,f}}{M}$.

4.4 Enhancing Estate Ranking with Functional Diversity

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

4.4.1 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 4.2 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.).

Meta Features. We exploit a random forest based method to learn meta features via supervised non-linear transformation. Indeed, the work in (He et al., 2014) 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 val-

ues 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 $g_m = \boldsymbol{w}^\top \boldsymbol{x_m} = \sum_{i=1}^{I} w_i x_{mi}$.

4.4.2 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

Table 4.2. 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

Bayesian inference, the posterior probability is

$$P(\Phi; \Delta, \Lambda) = P\left(\Delta|\Phi, \Lambda\right) P\left(\Phi|\Lambda\right). \tag{4.6}$$

We follow the commonly-used "bag of words" assumption (Blei, Ng, & Jordan, 2003), 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 Δ as

$$\begin{aligned} P\left(\Delta|\Phi, \Lambda\right) &= P\left(\{Y, \Pi, \Xi\}\,|\Phi, \Lambda\right) \\ &= \underbrace{P\left(Y|\Phi, \Lambda\right)}_{\text{Prediction Accuracy}} \times \underbrace{P\left(\Pi|\Phi, \Lambda\right)}_{\text{Ranking Consistency}} \times \underbrace{P\left(\Xi|\Phi, \Lambda\right)}_{\text{Functional Diversity}}, \end{aligned} \tag{4.7}$$

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.

$$P\left(Y|\Phi, \Lambda\right) = \prod_{m=1}^{M} \mathcal{N}(y_m|g_m, \sigma) = \prod_{m=1}^{M} \frac{1}{\sigma} \exp\left(-\frac{\left(y_m - g_m\right)^2}{2\sigma^2}\right). \tag{4.8}$$

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.

$$P\left(\Pi|\Phi,\Lambda\right) = \prod_{m=1}^{M-1} \prod_{h=m+1}^{M} P(m \to h|\Phi,\Lambda), \tag{4.9}$$

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.

$$\begin{aligned} P\left(\Xi|\Phi,\Lambda\right) &= \sum_{f=1}^{K} P(f) P(\Xi|f,\Phi,\Lambda) \\ &= \sum_{f=1}^{K} \frac{\theta_f}{1 + exp(-(\sum_{m=1}^{M} g_m \frac{\sum_{h=1}^{m} \eta_{h,f}}{m} - \sum_{m=1}^{M} g_m \eta_{m,f}))}. \end{aligned} \tag{4.10}$$

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, $P(\Phi|\Lambda) = \prod_{i=1}^{I} \mathcal{N}(w_i|0,\sigma_w^2)$.

4.4.3 Parameter Estimation

With the formulated posterior probability, the learning objective is to find the optimal estimate of the parameters Φ that maximizes the posterior. Hence, by inferring Equation 4.6, we can obtain the log of the posterior for the proposed model.

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}|Y,\Pi,\Xi,\sigma^2,\sigma_w^2) &= \sum_{m=1}^{M}\left[-\frac{1}{2}\ln\sigma^2 - \frac{(y_m-f_m)^2}{2\sigma^2}\right] \\ &+ \sum_{m=1}^{M-1}\sum_{h=m+1}^{M} ln\frac{1}{1+exp(-(g_m-g_h))} + \sum_{i=1}^{I}\left[-\frac{1}{2}\ln\sigma_w^2 - \frac{w_i^2}{2\sigma_w^2}\right] \\ &+ ln\sum_{f=1}^{K}\theta_f\frac{1}{1+exp(-(\sum_{m=1}^{M} g_m\frac{\sum_{h=1}^{m}\eta_{h,f}}{m} - \sum_{m=1}^{M} g_m\eta_{m,f}))}\end{aligned} \tag{4.11}$$

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.

4.4.4 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.

4.5 Experimental Results

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

4.5.1 Data Description

Table 4.3 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.

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,

Table 4.3. 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

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

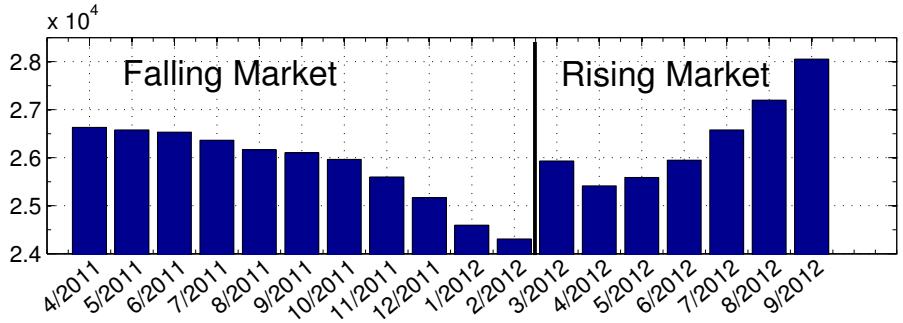


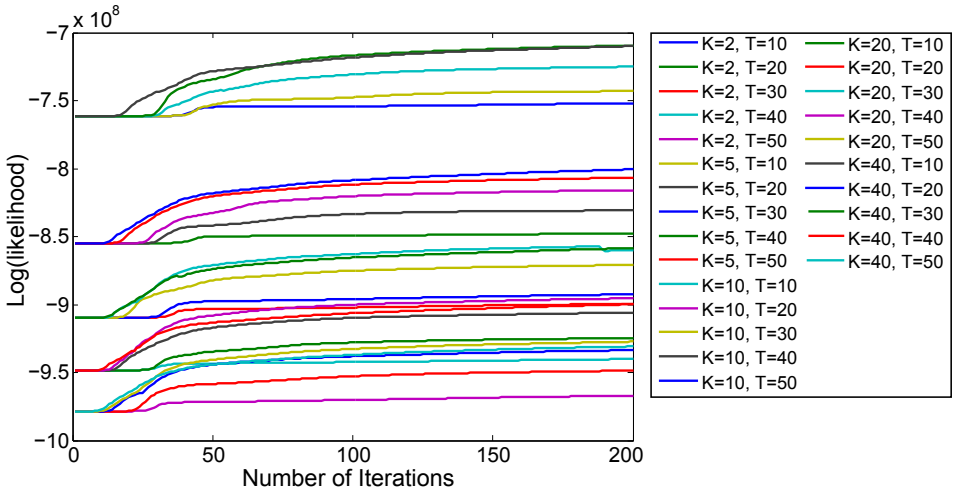
Figure 4.4. The rising and falling market periods in Beijing.

(Fu, Ge, et al., 2014). 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 provide meaningful information about differences in real estate value.

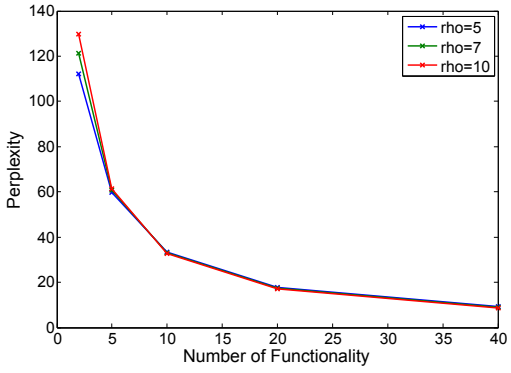
Finally, a list of estates, together with the extracted features and investment value of each, were split into two data sets, corresponding to the falling market period (from July 2011 to February 2012) and the rising market period (from February 2012 to September 2012), as shown in Figure 4.4. Here we follow the norms of real estate research, which typically studies rising and falling markets separately (Pace, 1998; Case & Shiller, 1988).

4.5.2 Baseline Algorithms

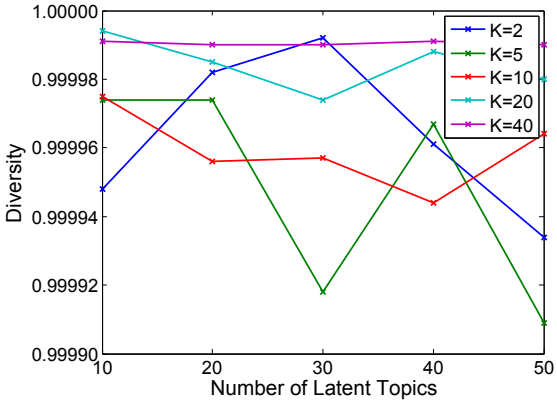
Since our work is related to Learning-To-Rank (LTR), we compared our method against the following algorithms. (1) **Coordinate Ascent (Metzler & Croft, 2007):** uses domination loss and coordinate descent optimization. (2) **LambdaMART (Burges, 2010):** the boosted tree version of LambdaRank. (3) **FenchelRank (Lai et al., 2013):** designed for solving sparse learning-to-rank with an L1 constraint.



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



(b) No. of functions vs. perplexity



(c) No. of latent topics vs. diversity

Figure 4.5. Sensitivity analysis of parameters.

(4) **ListNet (Cao et al., 2007):** 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 (Fu, Ge, et al., 2014):** 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 (Fu, Xiong, et al., 2014):** 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

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

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

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.

4.5.3 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.

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 param-

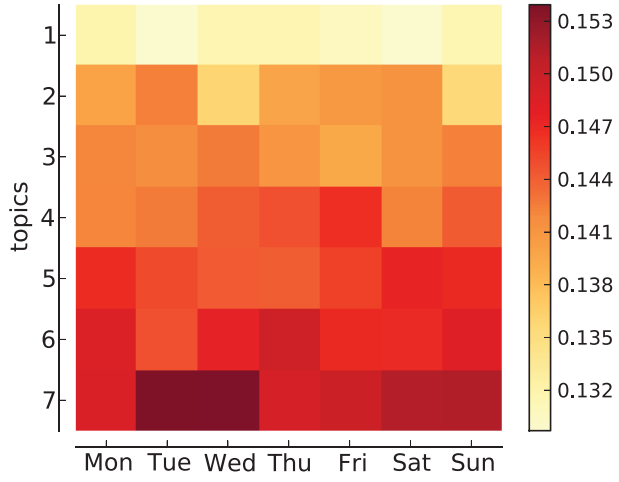
eter sensitivity, defined by $Perplexity = exp\left\{-\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\log P(\boldsymbol{c}_{m,n},\boldsymbol{t}_{m,n},\boldsymbol{b}_{m,n})}{\sum_{m=1}^{M}\sum_{n=1}^{N}(|\boldsymbol{c}_{m,n}|+|\boldsymbol{t}_{m,n}|+|\boldsymbol{b}_{m,n}|)}\right\}$, and $Diversity = \sum_{f=1}^{K}\theta_f\frac{\sum_{m=1}^{M}y_m\frac{\sum_{h=1}^{m}\eta_{h,f}}{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

4.5.4 Evaluation of Geographical Functional Portfolio Learning

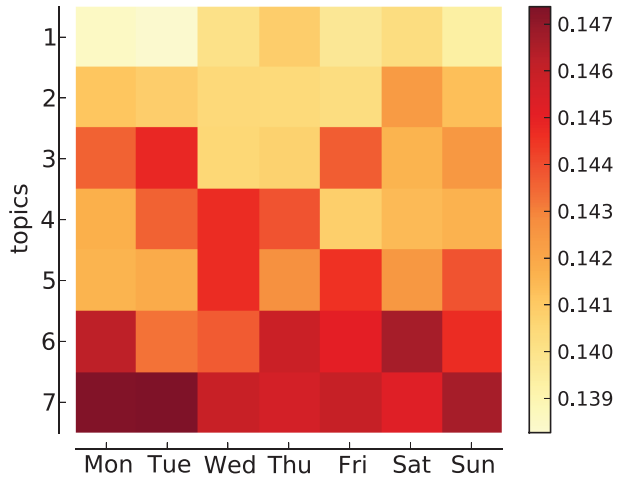
Next, we study our geographic functional learning model in terms of parameter sensitivity, temporal popular topics and patterns, and community functional portfolios.

(1) Study of Parameter Sensitivity.

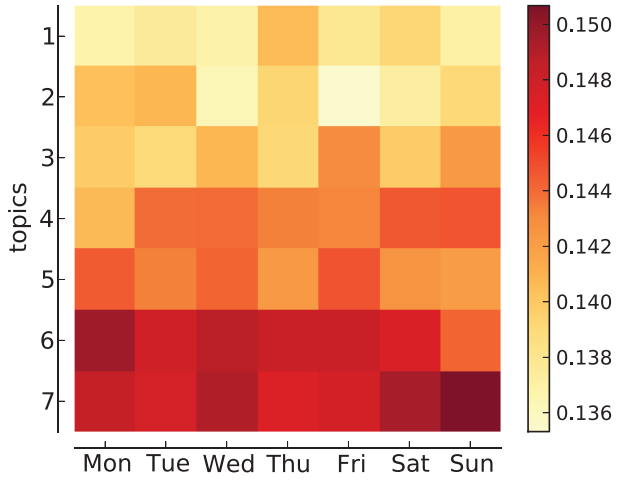
Here, we investigate the sensitivity of different parameter settings in terms of three metrics: likelihood, perplexity, and diversity. Figure 4.5(a) plots the likelihood against the number of iterations. The likelihoods in all settings converge after 100 iterations. To ensure convergence, we retrieve all the results after 200 iterations. Figure 4.5(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 4.5(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.



(a) Checkin latent topics.



(b) Taxi latent topics.



(c) Bus latent topics.

Figure 4.6. Heatmaps of temporal popularity of checkin, taxi and bus latent topics during weekdays.

Table 4.4. 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.

Table 4.5. 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 4.6. 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 4.6 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 4.4, 4.5, and 4.6, respectively. Figure 4.6 validates that the topic distribution of mobility has a temporal pattern. First, Figure 4.6(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 4.4. Next, Figure 4.6(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 4.5, 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 4.6 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 4.7 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

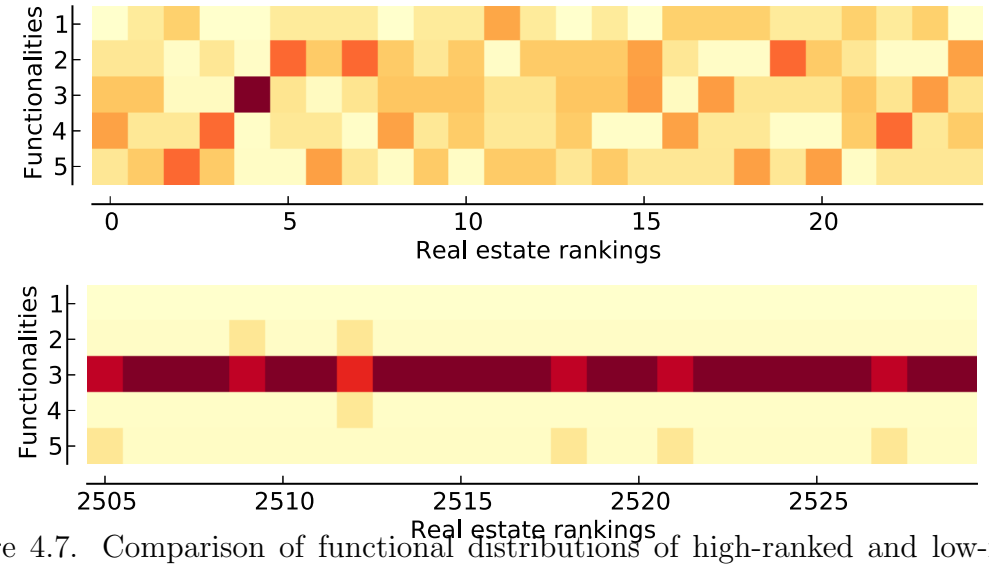


Figure 4.7. Comparison of functional distributions of high-ranked and low-ranked estates.

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.

4.5.5 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.

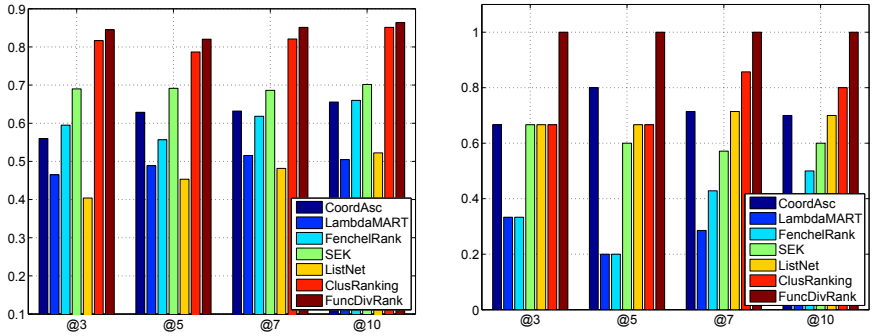
Rising Market. Figure 4.8 shows our method performs better than the baselines over top-k ranking in rising market. For example, our method offers 21%, 32.4%, 47.2% improvement in terms of NDCG@3 compared to SEK, FenchelRank, and RankBoost, respectively. Figure 4.8(b) shows that the top-K results ($K = 3, 5, 7, 10$) 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 4.9, 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 4.7.

Next, we discuss how our work differs from previous work on real estate ranking. First, while ClusRanking (Fu, Xiong, et al., 2014) 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 (Fu, Ge, et al., 2014) 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.

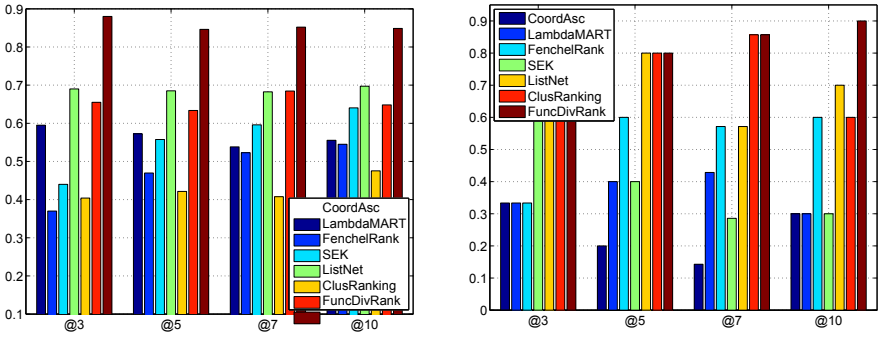
Table 4.7. The Tau values of different algorithms in rising and falling markets.

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



(a) NDCG@N (b) Prec@N

Figure 4.8. Performance comparison, rising market.



(a) NDCG@N (b) Prec@N

Figure 4.9. Performance comparison, falling market.

4.6 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 (Krainer & Wei, 2004), for example, price to income ratio. Some studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of estate prices (Chaitra H. Nagaraja & Zhao, 2009; Downie & Robson, 2007). More classic works are based on repeat sales methods and hedonic methods (Bailey et al., 1963; Shiller, 1991a; Knight & Sirmans, 1992; Taylor, 2003). The work in (Downie & Robson, 2007) studies the automated valuation models which aggregate and analyze

physical characteristics and sales prices of comparable properties, to provide property valuations. The work in (Fu, Ge, et al., 2014) extracts features from user reviews and mobility behaviors and integrates sparsity regularization into pairwise estate ranking. The work in (Fu, Xiong, et al., 2014) jointly models the geographical individual, peer, and zone dependencies for enhancing prediction of estate investment value. More recent works (Kontrimas & Verikas, 2011) apply general additive mode, support vector machine regression, and multilayer perceptron ensembles for computational estate appraisal.

Learning To Rank with Diversity. Also, our work is related to LTR. The pair-wise methods, such as RankNet (Burges et al., 2005), RankBoost (Freund et al., 2003), RankSVM (Herbrich et al., 1999), and LambdaRank (Quoc & Le, 2007), 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 (Weng & Lin, 2011; Rendle et al., 2009) provide full Bayesian explanations and optimize the posterior of point-wise and pair-wise ranking models, respectively. Study (Shi et al., 2012) unifies both rating error and ranking error as objective function to enhance Top-K recommendation. More recent works (Zhu, Goldberg, Van Gael, & Andrzejewski, 2007; Su, Tang, & Hong, 2012; Qin & Zhu, 2013) study diversified learning to rank. For example, (Zhu et al., 2007) ranks items by random walks in an absorbing Markov chain and achieves both diversity and centrality. The work in (Su et al., 2012) 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 (Yuan et al., 2012a) . Work (Karamshuk et al., 2013) selects the optimal sites for retail stores by mining Foursquare data. Also, our work is related to measuring similarity for ranking (Chang, Qi, et al., 2014; Chang, Aggarwal, & Huang, 2014).

4.7 Concluding Remarks

Summary. We investigated how to rank real estate investment values by considering the impact of 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.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this dissertation, we address the unique challenges of building a geographic ranking system by effectively modeling and efficiently computing with various mobile data.

Along these lines, I first introduced a method for ranking residential complexes based on investment ratings by mining users opinions about residential complexes from online user reviews and offline moving behaviors (e.g., taxi traces, smart card transactions, check-ins). While a variety of features could be extracted from these data, these features are intercorrelated and redundant. Thus, selecting good features and integrating the feature selection into the fitting of a ranking model are essential. To this end, I first strategically mined the fine-grained discriminative features from user reviews and moving behaviors. Then, I proposed a Sparse Pairwise Ranking method by combining a pairwise ranking objective and a sparsity regularization in a unified probabilistic framework.

Also, with the development of new ways to collect estate-related mobile data, there is a potential to leverage geographic dependencies of residential complexes for enhancing real estate evaluation. Indeed, the geographic dependencies of the value of a residential complex can be from the characteristics of its own neighborhood (individual), the values of its nearby residential complexes (peer), and the prosperity of the affiliated latent business area (zone). To this end, I proposed an enhanced

method, named ClusRanking, for real estate evaluation by leveraging the mutual enforcement of ranking and clustering power. In ClusRanking, three influential factors (i.e., geographic utility, neighborhood popularity, and influence of business areas) are constructed and extracted for predicting real estate investment ratings. An estate-specific ranking objective is also proposed to jointly model individual, peer and zone dependencies.

Finally, 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 residential complexes 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, I further developed a geographical function ranking method, named FuncDivRank, by incorporating the functional diversity of communities into real estate evaluation. In FunDivRank, a mix-land use latent model is developed to learn latent community functions and the corresponding portfolios. Also, a real estate ranking indicator is learned by simultaneously maximizing ranking consistency and functional diversity.

BIBLIOGRAPHY

Assil, E. M. (2012). Constructing a real estate price index: the moroccan experience.

Bailey, M., Muth, R., & Nourse, H. (1963). A regression method for real estate price index construction. *J. Am. Stat. Assoc.*, *58*, 933–942.

b. Hj. Mar Iman al Murshid, A. H. (2008). Modelling locational factors using geographic information system generated value response surface techniques to explain and predict residential property values. In *Naprec conference.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bowes, D. R., & Ihlanfeldt, K. R. (2001). Identifying the impacts of rail transit stations on residential property values. *Journal of Urban Economics*, *50*(1), 1–25.

Burges, C. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *Icml'05.*

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Icml'07.*

Case, K. E., & Shiller, R. J. (1988). *The behavior of home buyers in boom and post-boom markets.* National Bureau of Economic Research Cambridge, Mass., USA.

Ceci, M., Appice, A., & Malerba, D. (2007). Discovering emerging patterns in spatial databases: A multi-relational approach. In *Pkdd'07.* Springer Berlin Heidelberg.

Chaitra H. Nagaraja, L. D. B., & Zhao, L. H. (2009). *An autoregressive approach to house price modeling.*

Chang, S., Aggarwal, C. C., & Huang, T. S. (2014). Learning local semantic distances with limited supervision. In *2014 ieee international conference on data mining* (pp. 70–79).

Chang, S., Qi, G., Aggarwal, C. C., Zhou, J., Wang, M., & Huang, T. S. (2014). Factorized similarity learning in networks. In *2014 ieee international conference on data mining.*

Cheng, C., Yang, H., King, I., & Lyu, M. R. (2012). Fused matrix factorization with geographical and social influence in location-based social networks. In *Aaai'12.*

Cooper, W. S., Gey, F. C., & Dabney, D. P. (1992). Probabilistic retrieval based on staged logistic regression. In *Sigir'92.*

Downie, M. L., & Robson, G. (2007). Automated valuation models: an international perspective.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of machine learning research.*

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics.*

Fu, Y., Ge, Y., Zheng, Y., Yao, Z., Liu, Y., Xiong, H., & Yuan, N. J. (2014). Sparse real estate ranking with online user reviews and offline moving behaviors. In *the 14th ieee international conference on data mining (icdm 2014).*

Fu, Y., Liu, B., Ge, Y., Yao, Z., & Xiong, H. (2014). User preference learning with multiple information fusion for restaurant recommendation. In *Sdm'14.*

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

Fuhr, N. (1989). Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems.*

Fürnkranz, J., & Hüllermeier, E. (2003). Pairwise preference learning and ranking. In *Machine learning: Ecml 2003.* Springer.

Gantner, Z., Drumond, L., Freudenthaler, C., & Schmidt-Thieme, L. (2012). Personalized ranking for non-uniformly sampled items. *Journal of Machine Learning Research*.

Hang, L. (2011). A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., ... Candela, J. Q. n. (2014). Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the eighth international workshop on data mining for online advertising.* Retrieved from `http://doi.acm.org/10.1145/2648584.2648589` doi: 10.1145/2648584.2648589

Heierman III, E. O., & Cook, D. J. (2003). Improving home automation by discovering regularly occurring device usage patterns. In *Icdm'03.*

Herbrich, R., Graepel, T., & Obermayer, K. (1999). Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013). Geo-spotting: Mining online location-based services for optimal retail store placement. In *Kdd '13.*

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

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, *11*, 443 - 448.

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

Krainer, J., & Wei, C. (2004). House prices and fundamental value. *FRBSF Economic Letter*.

Lai, H., Pan, Y., Liu, C., Lin, L., & Wu, J. (2013). Sparse learning-to-rank via an efficient primal-dual algorithm. *Computers, IEEE Transactions on*, *62*(6), 1221–1233.

Lam, E.-K. (1996). Modern regression models and neural networks for residential property valuation. *RICS Research-The Cutting Edge*.

Landis, J., Guhathakurta, S., Huang, W., Zhang, M., & Fukuji, B. (1995). Rail transit investments, real estate values, and land use change: a comparative analysis of five california rail transit systems.

Lewis-Workman, S., & Brod, D. (1997). Measuring the neighborhood benefits of rail transit accessibility. *Transportation Research Record: Journal of the Transportation Research Board*, *1576*(1), 147–153.

Li, M., Li, H., & Zhou, Z.-H. (2009). Semi-supervised document retrieval. *Information Processing and Management*.

Liu, B., Fu, Y., Yao, Z., & Xiong, H. (2013). Learning geographical preferences for point-of-interest recommendation. In *Kdd'13.*

Liu, W., Zheng, Y., Chawla, S., Yuan, J., & Xing, X. (2011). Discovering spatio-temporal causal interactions in traffic data streams. In *Kdd '11.*

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

Metzler, D., & Croft, W. B. (2007). Linear feature-based models for information retrieval. *Information Retrieval*.

Mitropoulos, A., Wu, W., & Kohansky, G. (2007). Criteria for automated valuation models in the uk. *Fitch Ratings*.

Montanari, A., & Staniscia, B. (2012). From global to local: Human mobility in the rome coastal area in the context of the global economic crisis*. *Belgeo. Revue belge de géographie*(3-4), 187–200.

Pace, R. K. (1998). Appraisal using generalized additive models. *Journal of Real Estate Research*, *15*, 77-100. Retrieved from `http://ideas.repec.org/a/jre/issued/v15n11998p77-100.html`

Qin, L., & Zhu, X. (2013). Promoting diversity in recommendation by entropy regularizer. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2698–2704).

Quoc, C., & Le, V. (2007). Learning to rank with nonsmooth cost functions. *NIPS'07*.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Uai '09.*

Robert Cervero, C. D. K. (2011). Bus rapid transit impacts on land uses and land values in seoul, korea. *Transport Policy*, *18*, 102-116.

Shi, Y., Larson, M., & Hanjalic, A. (2012). Unifying rating-oriented and ranking-oriented collaborative filtering for improved recommendation. *Information Sciences*.

Shiller, R. J. (1991a). *Arithmetic repeat sales price estimators* (Tech. Rep.). Cowles Foundation for Research in Economics, Yale University.

Shiller, R. J. (1991b). *Arithmetic repeat sales price estimators* (Tech. Rep.). Cowles Foundation for Research in Economics, Yale University.

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

Su, H., Tang, J., & Hong, W. (2012). Learning to diversify expert finding with subtopics. In *Advances in knowledge discovery and data mining* (pp. 330–341). Springer.

Taylor, L. O. (2003). The hedonic method. In *A primer on nonmarket valuation.* Springer.

Wardrip, K. (2011). Public transits impact on housing costs: a review of the literature.

Weng, R. C., & Lin, C.-J. (2011). A bayesian approximation method for online ranking. *The Journal of Machine Learning Research*.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., & Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. In *Icml'08.*

Xu, J., & Li, H. (2007). Adarank: a boosting algorithm for information retrieval. In *Sigir '07.*

Yuan, J., Zheng, Y., & Xie, X. (2012a). Discovering regions of different functions in a city using human mobility and pois. In *Kdd'12.*

Yuan, J., Zheng, Y., & Xie, X. (2012b). Discovering regions of different functions in a city using human mobility and pois. In *Kdd'12.*

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM TIST*.

Zhou, Z.-H., Chen, K.-J., & Dai, H.-B. (2006). Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*.

Zhu, X., Goldberg, A. B., Van Gael, J., & Andrzejewski, D. (2007). Improving diversity in ranking using absorbing random walks. In *Hlt-naacl* (pp. 97–104).