Exploiting Geographic Dependencies for Real Estate Appraisal: A Mutual Perspective of Ranking and Clustering

Yanjie Fu[‡], Hui Xiong[‡*], Yong Ge[+], Zijun Yao[‡], Yu Zheng[◇], Zhi-Hua Zhou[†]
[‡]Rutgers University, {yanjie.fu, hxiong, zijun.yao}@rutgers.edu
[+]University of North Carolina at Charlotte, USA, yong.ge@uncc.edu
[◇]Microsoft Research Asia, yuzheng@microsoft.com
[†]Nanjing University, zhouzh@nju.edu.cn

ABSTRACT

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 paper, 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.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

*Contact author.

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2623675.

General Terms

Algorithms, Design, Experimentation

Keywords

Real Estate Appraisal, Geographic Dependencies, ClusRanking

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 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.

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

Specifically, in this paper, 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 paper, 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.

2. REAL ESTATE RANKING

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

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 paper, 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 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 paper, 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.

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 1: Mathematical Notations

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 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 1 shows the framework of our method.

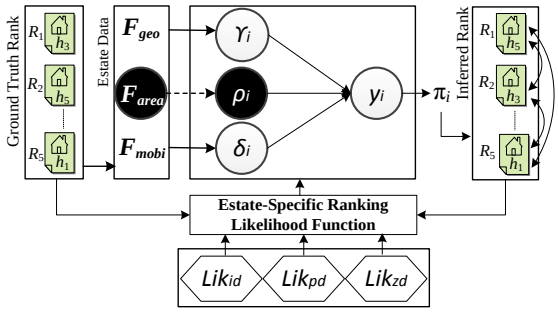


Figure 1: The framework of ClusRanking. (The black plates represent the latent effects.)

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.

2.3.1 Geographic Utility: γ

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

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 Table 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 [16], 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.

2.3.2 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.

2.3.3 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 $\kappa(p) = \sum_{d \in D} P(dist(d,p))$.

1 For each estate i:
1.1 Draw a business area $r \sim$ Multinomial(η).
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: The generative process of ClusRanking

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. 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 trajectory 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.

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 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

$$Pr(\Psi; \mathcal{D}, \Omega) = P\left(\mathcal{D}|\Psi, \Omega\right) P\left(\Psi|\Omega\right) \tag{1}$$

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

$$\begin{aligned} P\left(\mathcal{D}|\Psi, \Omega\right) &= P\left(\{Y, \Pi, L\}\,|\Psi, \Omega\right) \\ &= \underline{P\left(\{Y, L\}\,|\Psi, \Omega\right)} \times P\left(\Pi|\Psi, \Omega\right), \end{aligned} \tag{2}$$

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}. $P\left(\Pi|\Psi, \Omega\right)$ 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, 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

$$\begin{aligned} Lik_{id} &= \prod_{i}^{I} P(\{y_i, l_i\}|\Psi, \Omega) = \prod_{i}^{I} P(\{y_i, l_i, r_i\}|\Psi, \Omega) \\ &= \prod_{i=1}^{I} \mathcal{N}(y_i|f_i, \sigma) \prod_{i=1}^{I} \mathcal{N}(l_i|\mu_{r_i}, \Sigma_{r_i}) \prod_{i=1}^{I} Mult(r_i|\boldsymbol{\eta}) \\ &= \prod_{i=1}^{I} \frac{1}{\sigma} \exp\left(-\frac{(y_i - f_i)^2}{2\sigma^2}\right) \prod_{i=1}^{I} \frac{1}{\Sigma_{r_i}} \exp\left(-\frac{(l_i - \mu_{r_i})^2}{2\Sigma_{r_i}^2}\right) \prod_{i=1}^{I} Mult(r_i|\boldsymbol{\eta}) \end{aligned} \tag{3}$$

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

$$\begin{aligned} Lik_{pd} &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(i \to h|\Psi, \Omega)^{\mathbb{I}(r_i = r_h)} \\ &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left(\frac{1}{1 + exp(-(f_i - f_h))}\right)^{\mathbb{I}(r_i = r_h)} \end{aligned} \tag{4}$$

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 2.3.3). In this way, we capture the spatial dependency between estate and business area. Lik_{zd} is then given by

$$\begin{aligned} Lik_{zd} &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(r_i \to r_h|\Psi, \Omega)^{\mathbb{I}(r_i \neq r_h)} \\ &= \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left(\frac{1}{1 + exp(-(\eta_{r_i} - \eta_{r_h}))} \right)^{\mathbb{I}(r_i \neq r_h)}, \end{aligned} \tag{5}$$

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

$$\begin{aligned} &P(\Psi|\Omega) = P(q|\mu_q, \sigma_q^2)P(W|\mu_w, \sigma_w^2) \\ &= \prod_{m=1}^{M} \mathcal{N}(q_m|\mu_q, \sigma_q^2) \times \prod_{m=1}^{M} \prod_{n=1}^{N} \mathcal{N}(w_{mn}|\mu_w, \sigma_w^2) \\ &= \prod_{m=1}^{M} \frac{1}{\sigma_q} \exp\left(-\frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right) \prod_{m=1}^{M} \prod_{n=1}^{N} \frac{1}{\sigma_w} \exp\left(-\frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right) \end{aligned} \tag{6}$$

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,

$$r \sim P\left(l_i|r, \Psi^{(t)}\right) P\left(\{Y, \Pi\}|r, \Psi^{(t)}\right) P\left(r|\boldsymbol{\eta}^{(t)}\right) \tag{7}$$

where

$$P\left(l_i|r, \Psi^{(t)}\right) = \mathcal{N}\left(l_i|\mu_r^{(t)}, \Sigma_r^{(t)}\right) \tag{8}$$

$$\begin{aligned} P\left(\{Y, \Pi\}|r, \Psi^{(t)}\right) = P(y_i|f_i, \sigma^2) \prod_{h=i+1}^{I} P(i \to h|r, \Psi^{(t)})^{\mathbb{I}(r_i = r_h)} \\ \prod_{h=i+1}^{I} P(r_i \to r_h|r, \Psi^{(t)})^{\mathbb{I}(r_i \neq r_h)} \end{aligned} \tag{9}$$

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.

$$\begin{aligned} \boldsymbol{\mu}_r &= \frac{1}{\#(i,r)} \sum_{i=1}^{I} \mathbb{I}(r_i = r) l_i \\ \boldsymbol{\Sigma}_r &= \frac{1}{\#(i,r) - 1} \sum_{i=1}^{I} \mathbb{I}(r_i = r) \left((l_i - \mu_r)^\top (l_i - \mu_r) \right) \end{aligned} \tag{10}$$

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

$$\boldsymbol{\eta}_r^{(t+1)} = \frac{C_r^{(t+1)} + \gamma}{C^{(t+1)} + |R|\gamma} \tag{11}$$

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

$$\begin{aligned} &\mathcal{L}(q, W|R^{(t+1)}, \mathcal{D}) = \\ &\sum_{i=1}^{I} \left[-\frac{1}{2} ln\sigma^2 - \frac{(y_i - f_i)^2}{2\delta^2} \right] + \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} ln \frac{1}{1 + exp(-(f_i - f_h))} \mathbb{I}(r_i = r_h) \\ &+ \sum_{m=1}^{M} \left[-\frac{1}{2} ln\sigma_q^2 - \frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right] + \sum_{m=1}^{M} \sum_{n=1}^{N} \left[-\frac{1}{2} ln\sigma_w^2 - \frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right] \end{aligned} \tag{12}$$

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 $w_{mn}^{t+1} = w_{mn}^t - \epsilon \frac{\partial(-\mathcal{L})}{\partial w_{mn}}$

$$\begin{aligned} \frac{\partial(\mathcal{L})}{\partial q_m} &= \sum_{i=1}^{I} \frac{(y_i - f_i) w_{m \cdot} e_i}{\sigma^2} + \sum_{m=1}^{M} -\frac{q_m - \mu_q}{\sigma_q^2} + \\ &\sum_{i=1}^{I-1} \sum_{h=i+1}^{I} \frac{exp(f_h - f_i) w_{m \cdot} (e_i - e_h)}{1 + exp(f_h - f_i)} \mathbb{I}(r_i = r_h) \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{\partial(\mathcal{L})}{\partial w_{mn}} &= \sum_{i=1}^{I} \frac{(y_i - f_i) q_m e_{in}}{\sigma^2} + \sum_{m=1}^{M} -\frac{w_{mn} - \mu_w}{\sigma_w^2} + \\ &\sum_{i=1}^{I-1} \sum_{h=i+1}^{I} \frac{exp(f_h - f_i) q_m (e_{in} - e_{hn})}{1 + exp(f_h - f_i)} \mathbb{I}(r_i = r_h) \end{aligned} \tag{14}$$

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. EXPERIMENTAL RESULTS

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

3.1 Experimental Data

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 4: Statistics of the experimental data.

Table 4 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 [30]. Finally, we crawl the Beijing estate 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 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.

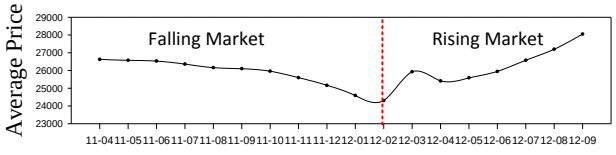


Figure 2: The rising market period and the falling market period in Beijing.

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 2.

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

$$DCG[n] = \begin{cases} rel_1 & if\ n = 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n}, & if\ n >= 2 \end{cases} \quad (15)$$

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 = $\frac{\#_{conc} - \#_{disc}}{\#_{conc} + \#_{disc}}$.

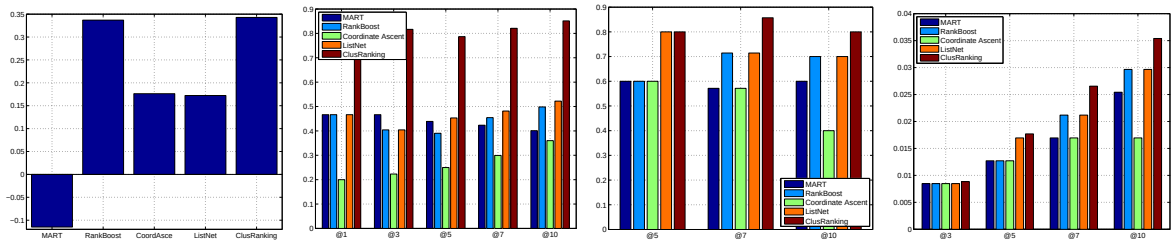
3.3 Baseline Algorithms

To show the effectiveness of the proposed method, we compare the ranking accuracy of our methods against following baseline algorithms. (1) **MART [10]:** it is a boosted tree model, specifically, a linear combination of the outputs of a set of regression trees. (2) **RankBoost [9]:** it is a boosted pairwise ranking method, which trains multiple weak rankers and combines their outputs as final ranking. (3) **Coordinate Ascent [20]:** it uses domination loss and applies coordinate descent for optimization. (4) **ListNet [4]:** 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 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.

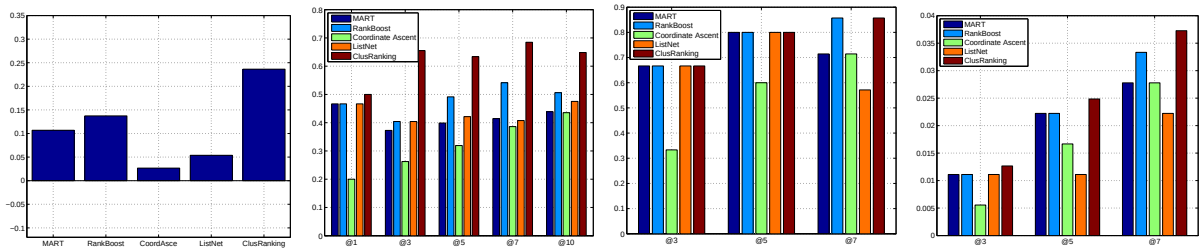
The codes are implemented in R (modeling), Python (preprocessing), and Matlab (visualization). The experiments were performed on a x64 machine with Intel i5 2.60GHz

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



(a) Tau (b) NDCG@N (c) Precision@N (d) Recall@N

Figure 3: The overall performances on the rising market dataset.



(a) Tau (b) NDCG@N (c) Precision@N (d) Recall@N

Figure 4: The overall performances on the falling market dataset.

dual-core CPU and 16GB RAM. The operation system is Microsoft Windows 7 Professional.

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(a) shows the Kendall's Tau Coefficient. Our method achieves 0.3428617 and outperforms the baselines. Figure 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(c) and Figure 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 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.

3.5 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 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 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, since the peer and zone dependencies better capture the ranking consistency of estates than the individual dependency, as individual dependency indeed models the prediction accuracy of the observed data collection $\{Y, L\}$.

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 5: Performance comparison of different geographic dependencies on the rising market data.

Falling Market Data. Table 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**

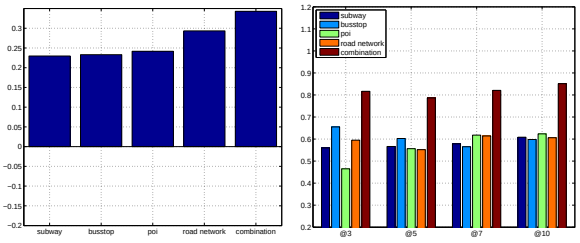
Table 6: Performance comparison of different geographic dependencies on the falling market data.

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 .

3.6 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 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 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.

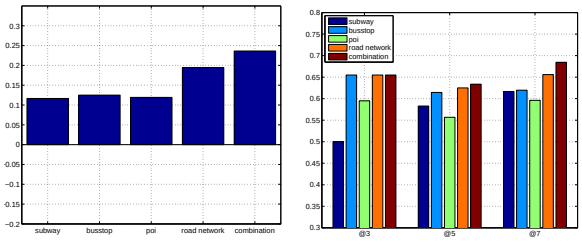


(a) Tau (b) NDCG@N

Figure 5: Performance comparison of different geographic features on rising market data.

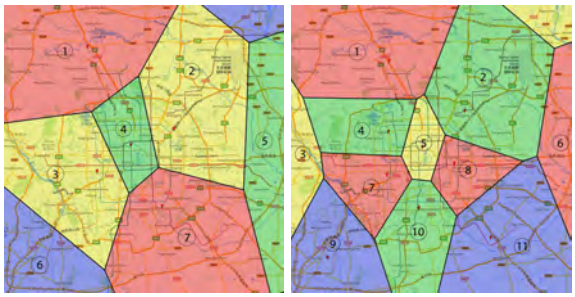
Falling Market Data. Figure 6(a) shows a comparison of the five feature sets on Tau: combination > road network > bus stop, subway and poi. This result is consistent with that of rising market data. Regarding top-k ranking, Figure 6(b) shows the NDCG@N (N=3, 5, 7 respectively) of different feature sets 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 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) Tau (b) NDCG@N

Figure 6: Performance comparison of different geographic features on falling market data.



(a) Kmeans (b) ClusRanking

Figure 7: A comparison of the learned business areas within the Beijing Fifth Ring (K=10).

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 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 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 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 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 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.

3.8 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 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 8. First, the upper half are darker than the lower half, which indicates

POI categories in the upper half is more important than those in the lower half. In other words, people prefer their homes near schools, malls, office, restaurants, transportation. 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-rank estates is evenly distributed for smooth whereas the POI density spectrum of the right-side low-rank 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 using the three inflection points. Finally, we present a triangle structure of needs of Beijing citizens as shown in Figure 9. The higher, the more fundamental and urgent in human needs.

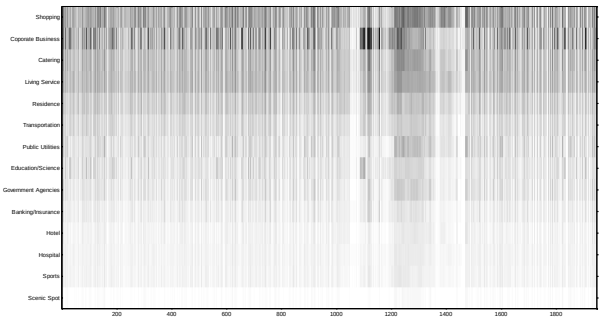


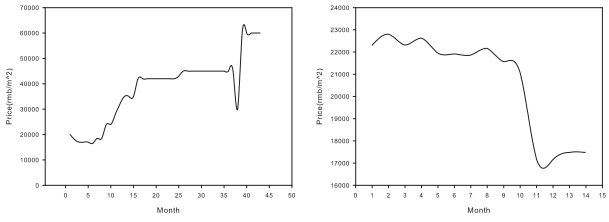
Figure 8: The POI density spectral of estates over multiple poi categories

Figure 9: The triangle need hierarchy of Beijing

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 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 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 price to RHF than JR11. This validate the individual dependency. Besides, RHF is located in the prosperous area of MuXiDi (inside No. 7 area in Figure 7(b)) near the 2nd ring road whereas JR11 is located in the area of DongFengXiang (inside No.2 area in Figure 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.



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

Figure 10: Price Trend Comparison.

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 7: A comparison of transportation, POI and mobility of RHF and JR11

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 [16]. More studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of estate prices. Work [8] checks the volatility of estate price and concludes that low investment-valued estate values relatively volatile. Work [5] 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 sales methods. The hedonic methods [27, 1] assume the price of a property depends on its characteristics and location. The repeat sales methods [1, 2, 26] construct a predefined price index based on properties sold more than once during the given period. Recent works [8, 21] study the automated valuation models, which aggregate and analyze physical characteristics and sales prices of comparable properties to provide property valuations. More recent studies [22, 15, 17, 2] 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 [12, 7] reduce the LTR task to a regression problem: given a single query-document pair, predict its score. The pair-wise methods, such as RankBoost [9], RankSVM [14] and LambdaRank [23], 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 [29], LambdaMART [3] and ListNet [4], optimize a ranking loss metric over lists instead of document pairs. Works [28, 24, 13] provide full Bayesian explanations and optimize the posterior of point-wise, pair-wise and list-wise ranking models. Study [25] 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 [32] or combining supervised and unsupervised ranking models [18].

Furthermore, our work has a connection with recent studies of exploring the geographic influence for POI recommendation. Works [6, 11] 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 [19] 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 [31].

5. CONCLUSION

In this paper, 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.e., 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.

Acknowledgement

This research was supported in part by National Science Foundation Grant IIS-1256016, the National Science Foundation of China under Grant number 61333014 and the 111 Project under Grant Number B14020.

6. REFERENCES

[1] E. M. Assil. Constructing a real estate price index: the moroccan experience. 2012.

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

[3] C. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11:23–581, 2010.

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

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

[6] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI'12*, 2012.

[7] W. S. Cooper, F. C. Gey, and D. P. Dabney. Probabilistic retrieval based on staged logistic regression. In *SIGIR'92*, 1992.

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

[9] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[11] Y. Fu, B. Liu, Y. Ge, Z. Yao, and H. Xiong. User preference learning with multiple information fusion for restaurant recommendation. In *SDM'14*, 2014.

[12] N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems (TOIS)*, 7(3):183–204, 1989.

[13] Z. Gantner, L. Drumond, C. Freudenthaler, and L. Schmidt-Thieme. Personalized ranking for non-uniformly sampled items. *Journal of Machine Learning Research-Proceedings Track*, 18:231–247, 2012.

[14] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*, pages 115–132, 1999.

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

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

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

[18] M. Li, H. Li, and Z.-H. Zhou. Semi-supervised document retrieval. *Information Processing and Management*, 2009.

[19] B. Liu, Y. Fu, Z. Yao, and H. Xiong. Learning geographical preferences for point-of-interest recommendation. In *KDD'13*, 2013.

[20] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10:257–274, 2007.

[21] A. Mitropoulos, W. Wu, and G. Kohansky. Criteria for automated valuation models in the uk. *Fitch Ratings*, 2007.

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

[23] C. Quoc and V. Le. Learning to rank with nonsmooth cost functions. *Proceedings of the Advances in Neural Information Processing Systems*, 19:193–200, 2007.

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

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

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

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

[28] R. C. Weng and C.-J. Lin. A bayesian approximation method for online ranking. *The Journal of Machine Learning Research*, 12:267–300, 2011.

[29] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *SIGIR '07*, 2007.

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

[31] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM TIST*, 2014.

[32] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 2006.