Exploiting Geographic Dependencies for Real Estate Appraisal

Yanjie Fu
Joint work with Hui Xiong, Yu Zheng, Yong Ge, Zhihua Zhou, Zijun Yao
Rutgers, the State University of New Jersey
Microsoft Research Asia, UNC Charlotte, Nanjing University
Agenda

- Background and Motivation
- Problem Statement
- Methodology
- Evaluation
- Conclusions
Housing Matters

Win your lover

Settle down your family

Make extra money as investment

Zillow

Realtors

Yahoo Homes

housing consultant services

Call for an intelligent system to rank estates based on estate investment value
We **don’t** predict future price
- We predict **growth potential of resale value** (long term thing)

- **Return Rate:**

\[ r = \frac{P_f - P_i}{P_i} \]

- Prepare the benchmark investment values of estates for training data
- **Value-adding capability** (in Rising Market)
- **Value-protecting capability** (in Falling Market)
What Special In Estate

- **Location! Location! Location!**
  - Urban geography (poi, road networks) features the geographical utility of houses

- **Human Mobility** reflects neighborhood popularity of houses

- **Prosperity of Business Area**
Individual dependency

- the investment value of an estate is determined by the geographic characteristics of its own neighborhood
Peer dependency

- Inside a business area, value can be reflected by its nearby estates.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to Level2 road network</td>
<td>156 meters</td>
<td>143 meters</td>
</tr>
<tr>
<td>Distance to subway station</td>
<td>1385 meters</td>
<td>1585 meters</td>
</tr>
<tr>
<td>#Restaurants</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>#Transportation facilities</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>
Zone dependency

- The estate value can also be influenced by the values of its associated latent business area.
Problem Definition

- **Given**
  - Urban geography (poi, road networks)
  - Human mobility (taxi GPS traces)
  - Estates with locations and historical prices

- **Objective**
  - Rank estates based on their investment values

- **Core tasks**
  - Predict estate investment value using urban geography and human mobility
  - Jointly model three geographic dependencies as objective function to learn estate ranking predictor
Methodology Overview

Estate Investment Value

Location

Geographic Utility
- Urban Geography
  - Road Networks
  - Subway Stations
  - Bus Stops
  - Places of interests

Neighborhood Popularity
- Human Mobility
  - Cell Tower Traces
  - Taxicab GPS Traces
  - Bus Traces
  - Check-ins

Influence of Business Area’s Prosperity
- Business Area

Prediction Model

Objective Function
- Individual Dependency
- Peer Dependency
- Zone Dependency
Figure 1: The framework of ClusRanking. (The black plates represent the latent effects.)
Geographic Utility

- Feature extraction by spatial indexing
- linearly regress geographic utility from geographic features of the neighborhood of each estate

<table>
<thead>
<tr>
<th>Data</th>
<th>Feature Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transportation</td>
<td>Number of bus stop</td>
</tr>
<tr>
<td></td>
<td>Distance to bus stop</td>
</tr>
<tr>
<td></td>
<td>Number of subway station</td>
</tr>
<tr>
<td></td>
<td>Distance to subway station</td>
</tr>
<tr>
<td></td>
<td>Number of road network entries</td>
</tr>
<tr>
<td></td>
<td>Distance to road network entries</td>
</tr>
<tr>
<td>Point of interest</td>
<td>Number of POIs of different POI categories</td>
</tr>
<tr>
<td></td>
<td>(Shopping, Sports, Education, etc.)</td>
</tr>
</tbody>
</table>

Neighbourhood Profiling (a neighborhood is defined as a cell area with a radius of 1km.)
The prosperous, the more easier we identify an estate from business area. The inverse influence of geo-distance between estate and business area center. Each business area is a cluster of estates, estates tend to co-locate along multiple business areas. There are K business areas in a city. The prosperities of K business areas (K spatial hidden states) show influence on estates. The more prosperous, the more easier we identify an estate from business area. The inverse influence of geo-distance between estate and business area center. Each business area is a cluster of estates, estates tend to co-locate along multiple business areas. There are K business areas in a city. The prosperities of K business areas (K spatial hidden states) show influence on estates. The more prosperous, the more easier we identify an estate from business area. The inverse influence of geo-distance between estate and business area center. Each business area is a cluster of estates, estates tend to co-locate along multiple business areas. There are K business areas in a city. The prosperities of K business areas (K spatial hidden states) show influence on estates.

Gaussian Mixture Model + Learning To Rank

Prosperities of K business areas

\[ \rho_i = \sum_{k=1}^{K} \frac{d_0}{d_0 + d(i, k)} \left( \frac{d_0}{d_0 + d(i, k)} \right)^{\frac{1}{n_k}} \sum_{k=1}^{K} \frac{n_k}{n_k} \]

\[ l_i \sim N(\mu_i, \Sigma_i) \]
Modeling Estate Investment Value (3)

- Neighborhood Popularity (A propagation view)
  - Propagate visit probability to POIs per drop-off point
  - Aggregate visit probability per POI
  - Aggregate visit probability per POI category
  - Compute popularity score

- Spatial propagation and aggregation from taxi to house
Generative Processes

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
   \[ w_{mn} \sim \mathcal{N}(w_{mn}|\mu_w, \sigma^2_w) \]
1.3.2 Draw coefficient vector of geography utility
   \[ q_m \sim \mathcal{N}(q_m|\mu_q, \sigma^2_q) \]
1.3.3 Estate geographic utility
   \[ \gamma_i = \frac{rent_i}{interest} + q^T W e_i \]
1.4 Compute influence given by latent business areas
   \[ \rho_i = \sum_{k=1}^{K} \left( \frac{d_0}{d_0 + d(i, r_k)} \right)^c \frac{\eta_k}{\sum_{k=1}^{K} \eta_k} \]
1.5 Compute neighborhood popularity
   \[ \delta_i = \frac{1}{J} \sum_{j=1}^{J} \max_{i \in r} \phi_{ij} \]
1.6 Generate the estate investment value
   \[ y_i \sim \mathcal{N}(y_i|f_i, \sigma^2) \text{ where} \]
   \[ f_i = \gamma_i + \delta_i + \rho_i \]
2 Compile the ranked list $\Pi$ of estates in terms of all $y_i$

Table 3: The generative process of ClusRanking
Modeling Three Dependencies

- **Individual Dependency**
  - Capture prediction accuracy of investment values, locations, and business area assignment

\[ \text{Lik}_i = \prod_{i} P(\{y_i, l_i, r_i\} | \Psi, \Omega) \]

The more accurate, the higher likelihood

- **Peer Dependency (pairwise analysis on estate level)**
  - Capture ranking consistency of predicted investment value of each estate pair

\[ \text{Lik}_{pd} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(r_i \rightarrow r_h | \Psi, \Omega) \text{Sigmoid}(f_i - f_h) \]

- **Zone Dependency (pairwise analysis on business area level)**
  - Capture ranking consistency of learned prosperities of the corresponding business area pair of each estate pair

\[ \text{Lik}_{zd} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(r_i \rightarrow r_h | \Psi, \Omega) \text{Sigmoid}(\eta_{r_i} - \eta_{r_h}) \]

- **Overall likelihood**
  \[ P(D | \Psi, \Omega) = \text{Lik}_i \times \text{Lik}_{pd} \times \text{Lik}_{zd} \]
Parameter Estimations

- Given $\mathcal{D} = \{Y, \Pi, L\}$ where $Y$, $\Pi$, and $L$ are the investment value, ranks and locations of I estates respectively.

- To maximize the posterior $Pr(\Psi; \mathcal{D}, \Omega) = P(\mathcal{D}|\Psi, \Omega) P(\Psi|\Omega)$

Parameters $\Psi = \{q, W, \eta, \mu, \Sigma\}$  Priors $\Omega = \{\mu_q, \sigma_q^2, \mu_w, \sigma_w^2, \sigma^2\}$

- Expectation Maximization (EM) to learn the parameters by treating latent business area of each estate as a latent variable
  - Geo-clustering updates the latent business area by maximizing the posterior of latent business area
  - After business area assignments are updated, maximizing the posterior of model parameters
## Experimental Data

- **Beijing real-world Data**
  - **Beijing estate data**
    - 2851 estates with transaction records from 04/2011 to 09/2012
    - Falling market (04/2011 to 02/2012) and Rising market (02/2012 to 09/2012)
  - **Beijing transportation facility data** including bus stop, subway, road networks
  - **Beijing POI data**
  - **Beijing taxi GPS traces**

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Properties</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real estates</td>
<td>Number of real estates</td>
<td>2,851</td>
</tr>
<tr>
<td></td>
<td>Size of bounding box (km)</td>
<td>40*40</td>
</tr>
<tr>
<td></td>
<td>Time period of transactions</td>
<td>04/2011 - 09/2012</td>
</tr>
<tr>
<td>Bus stop (2011)</td>
<td>Number of bus stop</td>
<td>9,810</td>
</tr>
<tr>
<td>Subway (2011)</td>
<td>Number of subway station</td>
<td>215</td>
</tr>
<tr>
<td>Road networks (2011)</td>
<td>Number of road segments</td>
<td>162,246</td>
</tr>
<tr>
<td></td>
<td>Total length (km)</td>
<td>20,022</td>
</tr>
<tr>
<td></td>
<td>Percentage of major roads</td>
<td>7.5%</td>
</tr>
<tr>
<td>POIs</td>
<td>Number of POIs</td>
<td>300,811</td>
</tr>
<tr>
<td></td>
<td>Number of categories</td>
<td>13</td>
</tr>
<tr>
<td>Taxi Trajectories</td>
<td>Number of taxis</td>
<td>13,597</td>
</tr>
<tr>
<td></td>
<td>Effective days</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Number of trips</td>
<td>8,202,012</td>
</tr>
<tr>
<td></td>
<td>Number of GPS points</td>
<td>111,602</td>
</tr>
<tr>
<td></td>
<td>Total distance (km)</td>
<td>61,269,029</td>
</tr>
</tbody>
</table>

Table 4: Statistics of the experimental data.
Evaluation Methods and Metrics

- **Baseline algorithms**
  - **MART**: it is a boosted tree model, specifically, a linear combination of the outputs of a set of regression trees
  - **RankBoost**: it is a boosted pairwise ranking method, which trains multiple weak rankers and combines their outputs as final ranking
  - **Coordinate Ascent**: it uses domination loss and applies coordinate descent for optimization
  - **ListNet**: it is a listwise ranking model with permutation top-k ranking likelihood as the objective function

- **Evaluation metrics**
  - **Normalized Discounted Cumulative Gain (NDCG)**
  - **Precision** \[ \text{Precision}_{@N} = \frac{|E_N \cap E_{\geq 3}|}{N} \]
  - **Recall** \[ \text{Recall}_{@N} = \frac{|E_N \cap E_{\geq 3}|}{|E_{\geq 3}|} \]
  - **Kendall’s Tau Coefficient** \[ \text{Tau} = \frac{\#\text{conc}-\#\text{disc}}{\#\text{conc}+\#\text{disc}} \]
We investigate the ranking performances by comparing to baseline algorithms in terms of Tau, NDCG, Precision and Recall.

Figure 3: The overall performances on the rising market dataset.
We investigate the ranking performances by comparing to baseline algorithms in terms of Tau, NDCG, Precision and Recall.

Figure 4: The overall performances on the falling market dataset.
We study the impact of three geographic dependencies by designing different variants of objective function (posterior likelihoods).

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>@N</th>
<th>ID</th>
<th>PD</th>
<th>PD+ZD</th>
<th>ClusRanking</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG</td>
<td>3</td>
<td>0.5599531</td>
<td>0.6549766</td>
<td>0.6900469</td>
<td>0.8166009</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.5771226</td>
<td>0.6024622</td>
<td>0.6101556</td>
<td>0.7867076</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.587992</td>
<td>0.6048394</td>
<td>0.641282</td>
<td>0.8208795</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.6518163</td>
<td>0.6723095</td>
<td>0.694175</td>
<td>0.8513267</td>
</tr>
<tr>
<td>Tau</td>
<td>-</td>
<td>0.2494531</td>
<td>0.2535007</td>
<td>0.2203712</td>
<td>0.3428617</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>@N</th>
<th>ID</th>
<th>PD</th>
<th>PD+ZD</th>
<th>ClusRanking</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG</td>
<td>3</td>
<td>0.570193</td>
<td>0.5950234</td>
<td>0.6250234</td>
<td>0.6549766</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.6144799</td>
<td>0.6004235</td>
<td>0.6144799</td>
<td>0.633635</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.6196808</td>
<td>0.654487</td>
<td>0.6196808</td>
<td>0.6845354</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.6415102</td>
<td>0.6252658</td>
<td>0.6307051</td>
<td>0.6482665</td>
</tr>
<tr>
<td>Tau</td>
<td>-</td>
<td>0.1186736</td>
<td>0.1313437</td>
<td>0.1433408</td>
<td>0.2363498</td>
</tr>
</tbody>
</table>

Individual dependency can achieve good overall ranking performance;

Peer and zone dependency can achieve good top-k ranking performance;

We recommend combination usage of three dependencies
We hope we go shopping, work, eat, access transportation quickly and easily, children go to schools near our home.

We DONOT need hotels/hospitals/sports/spots located near our home.

Good house is always balance people's need.
The triangle need structure of human life in Beijing
Conclusions

- Housing analysis is funny
- Use urban geography and human mobility to model estate investment value
- Capture geographic individual, peer, and zone dependencies to better learn estate ranking predictor

Business applications
- Decision making support for homebuyers
- Improve price structure for housing agent/broker/consultant
- Optimize site selection for housing developer
Thank You!

Email: yanjie.fu@rutgers.edu
Homepage: http://pegasus.rutgers.edu/~yf99/