# Price Recommendation on Vacation Rental Websites

Yang Li* Suhang Wang† Tao Yang* Quan Pan* Jiliang Tang‡

**Abstract**

Vacation rental websites such as Airbnb have become increasingly popular where rentals are typically short-term and travels or vacations related. Reasonable rental prices play a crucial role in improving user experiences and engagements in these websites. However, the unique properties of their rentals challenge traditional house rentals that are often long-term and study or work related. Therefore, in this paper we investigate the novel problem of price recommendation in vacation rental websites. We identify some important factors that affect the rental prices and propose a framework that consists of Multi-Scale Affinity Propagation (MSAP) to cluster houses, Nash Equilibrium filter to remove unreasonable price and Linear Regression model with Normal Noise (LRNN) to predict the reasonable prices. Experimental results demonstrate the effectiveness of the proposed framework. We conduct further experiments to understand the important factors in rental price recommendation.

## 1 Introduction

As an increasingly popular application of online booking service, online vacation rental websites such as Airbnb [1] and FlipKey [2] have attracted millions of travelers and hosts. Generally, these websites allow people to list, discover, and book accommodations around the world online, which not only benefits travelers but also increases the income of the hosts. For example, Airbnb has helped people experience the unique travel in more than 34,000 cities and 191 countries with over 60 million guests; and it also gives more than 20 million hosts the opportunities to earn extra money from the extra accommodations[3]. However, due to various reasons such as the lack of knowledge about how to set a reasonable price, some prices listed in these websites are unreasonable. Figure 1 is a snapshot from Airbnb. It illustrates an example that a host finds that her listing price is unreasonably high[4]. As online vacation

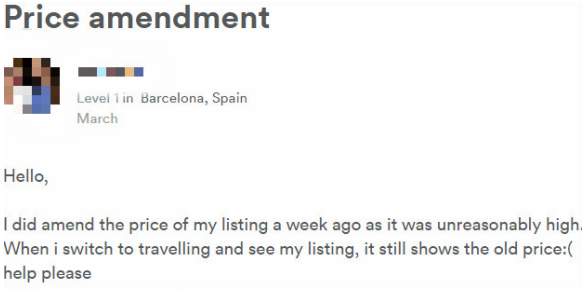


Figure 1: An Example of An Unreasonable Price

rental services are becoming more and more popular and there could be many unreasonable prices, the task of predicting and recommending reasonable rental prices of the accommodations becomes very important and necessary for both travelers and hosts. For travelers, a reasonable price can save their money and help them make better decision about which accommodations to rent. While for hosts, recommending a reasonable price can save their efforts in figuring out the price to list and reduce the risk that few travelers want to rent due to the unreasonable price.

The problem of reasonable price prediction on vacation rental websites is a novel and challenging problem. It is inherently different from traditional house rental price prediction. For traditional house rentals, tenants tend to have long-term rentals such as one year that are usually for work or study. Unlike traditional house rentals, studies reveal that vacation rental websites such as Airbnb are largely used for group travels or vacations that are often short term[5]. Thus, users in these websites are likely to value more on famous landmarks and transportation around the room. For instance, the closer the room to landmarks, the more travelers may want to pay. Meanwhile, for rooms within similar landmark ranges, traveler then care more about room facilities such as wifi since they need such facilities that are not likely to install by themselves for short-term stays. These observations suggest that it has great potentials to predict reasonable prices by investigating landmarks around and the facilities in the room.

In this paper, we study the novel problem of reasonable price recommendation for vacation rental websites

*Northwestern Polytechnical University, liyangnpu@mail.nwpu.edu.cn,{yangtao107,quanpan}@nwpu.edu.cn
†Arizona State University, swang187@asu.edu
‡Michigan State University, tangjili@msu.edu

[1] https://www.airbnb.com/

[2] https://www.flipkey.com/

[3] https://www.airbnb.com/about/about-us?locale=en

[4] https://community.airbnb.com/t5/Hosts/Price-amendment/td-p/39943

[5] https://www.jumpshot.com/airbnb-infographic-who-uses-airbnb-and-why/

such as Airbnb by investigating nearby landmarks and room facilities. In particular, we investigate the following two challenges - (1) How to utilize the landmarks and room facilities in a mathematical way for reasonable price recommendation; and (2) How to eliminate unreasonable prices (or outliers) so that we can learn a better model for price recommendation. In an attempt to solve these two challenges, we propose a novel framework, which is composed of three components – (a) Multi-Scale Affinity Propagation (MSAP) to cluster houses appropriately by landmarks and house facilities, (b) Nash equilibrium to remove unreasonable prices and (c) Linear Regression with Normal Noise (LRNN) to predict reasonable prices. The main contributions of the paper are listed as follows:

- Identify important factors related to the house price in terms of its distance to the landmarks, landmark popularity, and its facilities;
- Propose a method MSAP that groups the houses based on the landmarks and house facilities;
- Provide a filter of Nash equilibrium to remove the unreasonable prices in the clustered houses; and
- Conduct extensive experiments to demonstrate the effectiveness of the proposed methods in price recommendation in vacation rental websites.

The rest of the paper is organized as follows. In Section 2, we conduct preliminary data analysis, which lays the foundation of our work. In Section 3.1, we introduce the details of the multi-scale clustering using MSAP. In section 3.2 the Nash equilibrium filter is introduced, Section 3.3 is the Linear regression model with normal noise, then followed the experiments with discussions in Section 4. In Section 5, we review the related work about rental price prediction. In Section 6, we conclude our work with possible future research directions.

## 2 Preliminary Data Analysis

In vacation rental websites, landmarks and room facilities should affect rental prices. Thus, in this section, we initially analyze the correlation among rental prices, popularity of nearby landmarks and the room facilities, which lays the ground work for our price prediction framework. For better explanation, the mathematical notations used in the paper are listed in Table 1.

**2.1 Datasets** To conduct the preliminary data analysis, we collect a dataset from Airbnb and TripAdvisor[6]. For Airbnb, we collect the house information which includes house location (latitude, longitude), facilities status, rental price, comment numbers, and review grades from three cities, i.e., Los Angeles, London and Tokyo. Since there is no landmark information available in Airbnb, we collect the landmark information from TripAdvisor for the three cities. The landmark information includes the location (latitude, longitude) and comment numbers. The statistics of house and landmark information are listed in Table 2 and 3, respectively.

There are thirty dimensions in facility status, which are listed as follows: *Carbon Monoxide Detector, Meeting Facilities, Washing Machine, Buzzer/Wireless Intercom, TV, Gym, Indoor Fireplace, Elevator, Daily Necessities, Fire Extinguishers, Swimming Pool, Security Card, Wireless Network, Smoking Allowed, Kitchen, Cable TV, Shampoo, Guard, Heater, Dryer, Family/Children Friendly, Free Parking, Breakfast, Network, Air Conditioning, Pets Allowed, Smoke Detectors, Events Friendly, Hot Tub, and First Aid Kit.*

Table 2: Statistics for Accommodations

| | Los Angeles | London | Tokyo |
|---|---|---|---|
| # Accommodations | 1,008 | 100,8 | 2,496 |
| # Comments | 176,954 | 163,225 | 388,378 |
| Average Price ($) | 165 | 134 | 78 |

Table 3: Statistics for Landmarks

| | Los Angeles | London | Tokyo |
|---|---|---|---|
| # Landmarks | 160 | 668 | 337 |
| # Comments | 59,769 | 340,502 | 51,463 |

**2.2 Data Analysis** City landmarks such as tour sights, museums and theaters are the city symbols which attract tourists from all over the world. Thus, the distance to a popular landmark is a factor that affects the house rental price. A traveler seeking for famous landmarks or vacations would like to pay more to rent a house near landmarks. This intuition can be stated as the following assumption:

ASSUMPTION 1. *The closer the room to famous landmarks, the higher the house rental is.*

To verify this assumption, we show the distance of the house to nearest landmark and the rental price of the house in Figure 2. A house and landmark can be regarded as data points $n^h$, $n^m$ in the map, respectively. The nearest landmark $n_j^m$ to the house $n_j^h$ is defined as follow:

$$\forall n_k^m \in M \quad |n_j^h - n_j^m| \leq |n_j^h - n_k^m| \tag{2.1}$$

where $|n_j^h - n_j^m|$ denotes the Euler distance. From the figure, we observe that the house rental price is inversely proportional to its distance to its nearby landmark, which validates the reasonability of Assumption

[6] www.tripadvisor.com

Table 1: The mathematical notations expressions

| Notation | Means | Notation | Means |
|---|---|---|---|
| $n^h$, $n^m$ | sample point of house and landmark | $st$ | strategy |
| $M$ | landmark points set | $g_l^t$,$g_l^h$ | income at action $l$ |
| $e$ | facility vector | $S$ | instinct income of tenant |
| $E$ | Hoffman coding | $p$,$p^t$,$p^h$ | probability of strategy |
| $H$ | house facility variety | $a$ | satisfaction threshold |
| $s_m$ | similarity between two nodes | $b$ | upper bound of review grade |
| $q$ | viscosity coefficient | $\epsilon$ | value error |
| $R$ | # of comments | $EP^t$,$EP^h$ | expected profit |
| $\alpha$ | nondimensionalize distance factor | $\Delta P$ | price gap distribution |
| $r_m$,$r_h$ | responsibility | $i, k, j, l$ | index number |
| $a_m$,$a_h$ | availability | $Sil$ | measurement of silhouette |
| $V$ | house value | $D$ | dissimilarity |
| $P$ | house price | $A$ | inner dissimilarity in a cluster |
| $W$ | parameters | $F$ | dissimilarity set between clusters |
| $d$ | distance | $B$ | minimum dissimilarity between clusters |
| $\xi$ | celebration activities | $C$ | other clusters set except current cluster |
| $ST^t$,$ST^h$ | strategy set | $w$ | the deviation of price gap distribution |
| $c$ | cluster | | |

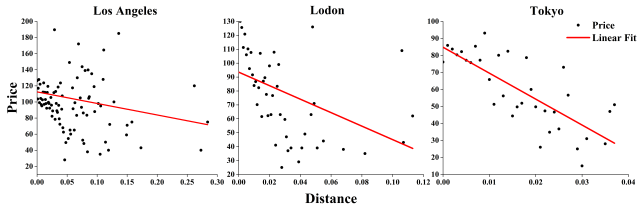


Figure 2: The relation between the house rental prices and their distances to the nearest landmarks.

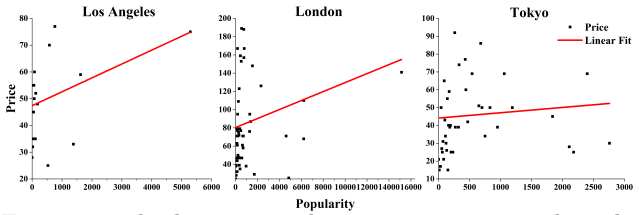


Figure 3: The house rental price is proportional to the landmark popularity.

1. Likewise, study from [12] suggests that there is a strong relationship between house price and its location.

Meanwhile if a place is very popular, such as London Tower which is one of the most popular landmarks in London, intuitively the rental prices of the house around it will be more expensive than those around non-popular landmarks. This intuition is summarized in Assumption 2 as:

ASSUMPTION 2. *The popularity of the landmark positively affects the house rental prices nearby.*

To validate this assumption, we use the numbers of the comments to indicate the landmark popularity since a more popular landmark will generally receive more comments. Note that there are other ways to assess the landmark popularity and we would like to leave it as one future work. We then analyze the relation between the popularity of a landmark and the average rental prices of houses near it. The result is shown in Figure 3. From the figure, in general we observe – the more popular the landmark is, the higher the rental price of house near it.

In addition, the house facilities should affect the rental price. For example, a house with a swimming pool is more expensive, and furthermore such houses are rare. Apart from the swimming pool, facilities like elevator, gym, and guard could make the rental price higher. We first define the coverage of facility as the portion of houses with the facility in a community. For example, if every house in a community is equipped with *TV*, then its coverage in the community is 100%; while if there are only two houses in ten with *gym*, then the coverage of *gym* is 20%. This phenomenon is summarized in Assumption 3.

ASSUMPTION 3. *In a community, the coverage of facility is inversely proportional to house rental prices.*

To verify the assumption, we plot correlation between facility coverage of houses and the average prices in Figure 4. From the figure, we note that when a facility is covered by the majority of the houses, the contribution of the facility to the house rental price is less. The work of [2] also suggests that the house price is positively

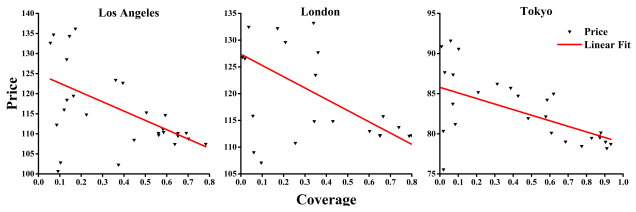


Figure 4: The house rental price is inversely proportional to the facility coverage.

affected by the house facilities.

In summary, preliminary analysis on three cities shows that rental price is closely related to the distance to landmark, the popularity of landmark nearby and the house facility coverage. These observations imply that houses within similar district of similar house facilities should have similar values to travelers, which lay the ground work of our price recommendation framework.

## 3 The Proposed Framework

In this section, we introduce the proposed framework for price recommendation. The framework is composed of three components: (i) Multi-scale clustering to divide houses into groups such that houses in each group are likely to have similar rental prices; (ii) Nash equilibrium to remove unreasonable prices that helps clean the data; and (iii) Linear regression with normal noise by integrating the landmark and house facility information for prediction. Next, we introduce each component in detail.

**3.1 Multi-Scale Clustering** We are informed by the previous three assumptions that houses in the same district with similar facilities are likely to have similar prices. We refer the cluster after Landscape Clustering as a district that will be explained next. Thus if we can first cluster the houses based on landmarks into districts and then further group houses in each cluster based on facilities, we can have two advantages: (i) we may enable to give better price prediction as in each cluster the prices are desired to be similar; and (ii) a house with very high or low price compared to others in the same cluster indicates that the price of the house could be unreasonable. Therefore, with clustering, we may enable to eliminate certain outliers.

We propose Multi-Scale Affinity Propagation (MSAP) to perform the two stage clustering, i.e., Landmark Clustering (LC) and House Clustering (HC). We use the similarity $s_m(i,k) = -\|n_i^m - n_k^m\|^2$ to indicate how well the landmark with index $k$ suitable to be the exemplar for landmark $i$. According to Assumption 2, if the landmark is popular, it has a large viscosity that will attract more people renting the house in this district. Thus, we add the viscosity coefficient $q_j = \frac{R_j}{\sum_{j=1}^{|M|} R_j + 1}$ to the negative Euclidean distance, where $R$ is the number of comments, $\alpha$ the distance threshold to nondimensionalize distance factor. Then the similarity between two nodes is defined as follows:

$$s_m(i,k) = -\frac{\|n_i^m - n_k^m\|^2}{\alpha(1+e^{2q_k})} \tag{3.2}$$

The responsibility $r_m(i,k)$ means the likelihood of landmark $k$ attracting $i$ into its cluster, and the availability $a_m(i,k)$ measures the probability of the landmark $i$ choosing $k$ as the exemplar. Below are the rules for the responsibility $r_m$ and availability $a_m$:

$$r_m(i,k) = s_m(i,k) - \max\{a_m(i,j) + s_m(i,j)\} \quad (j \in \{1,2,..,k-1,k+1,...,N\}) \tag{3.3}$$

$$a_m(i,k) = min\{0, r_m(k,k) + \sum_j \{\max(0, r_m(j,k))\}\} \quad (j \in \{1,2,..,k-1,k+1,...,N\}) \tag{3.4}$$

With responsibility $r_m(i,k)$ and the availability $a_m(i,k)$ defined above, we perform Affinity Propagation [6] on the landmarks. The general idea of affinity propagation is that all data points are simultaneously considered as exemplars, but exchange deterministic messages as defined above until a good set of exemplars gradually emerges. These exemplars (or landmarks) are considered as the most attractive landmarks. Then for each house, the nearest exemplar landmark is treated as its cluster center and thus all houses are grouped into different clusters. We name these clusters as districts and the center of the district is the exemplar landmark.

After houses are assigned to each district, we further perform clustering on houses in each district, separately. The procedure is called HC. Instead of defining the similarity on locations, HC defines the similarity on house facilities. Specifically, let a vector $e$ which only contains 0 and 1 to represent a house facility status. The order of these thirty facilities in the vector is decided by its coverage. According to Assumption 3, the lower the facility coverage is, the more luxurious the house is. So the facility position in the vector $e$ is in ascending order based on its coverage. We mainly use two housing properties including the location and facilities. Therefore, to make these two variables unified, we need to map the vector $e$ into two dimensions as the latitude and longitude in geographic coordinates. Thus the vector $e$ is divided into two vectors $\{e_1, e_2\}$ staggered. For example, if the vector $e = [1,0,1,0,1,1]$,

then $e_1 = [1, 1, 1]$ and $e_2 = [0, 0, 1]$. Then the house facility variety can be defined as follows:

$$H_j = \log(E(e_j)) \quad j \in \{1, 2\} \tag{3.5}$$

where $E$ is the Hoffman Coding on $e_j$. Finally, facility variety is quantified. Then the similarity $s_h$ is measured by $H$ as $s_h(i, k) = -\|H_i - H_k\|^2$. The responsibility $r_h$ and availability $a_h$ are the same as that in LC. With responsibility $r_h$ and availability $a_h$, we perform affinity propagation and cluster houses in each district into groups.

**3.2 Nash Equilibrium Filter** Our observations suggest that houses with similar facilities in one district should have similar prices. However, the prices of some houses are astonishingly high, while some are extremely low compared to houses with similar facilities in the same district. This suggests the existence of unreasonable prices that should be eliminated for accurate price prediction. In order to clean the data, for each district, we use a filter of Nash equilibrium to remove unreasonable prices, where the district is found by LC described above.

**3.2.1 Price and Value** From Assumption 3, the facility can affect the house value $V$ expressed by (3.6). However, the house rental price $P$ not only relays on the facilities, but also the location (e.g., the distance to the landmark and landmark's popularity), celebration activities and so on, which shows in (3.7). The gap between the value and the price is the income to the host.

$$V_k = WH_k + \epsilon \tag{3.6}$$

where $W$ is the parameters, and $\epsilon$ the value error.

$$P_k \propto V_k + W\{d_k, q_k, \xi_k, c_k\} \tag{3.7}$$

where $W$ is the parameters, $d_k$ the distance between the nearest landmark and the house $k$, $q_k$ the viscosity coefficient of nearest landmark, $c_k$ the cluster that the house belongs to, and $\xi_k$ the celebration activities, such as festivals, events and so on.

**3.2.2 Mixed Strategy Nash Equilibrium** Before introducing the game theory, two definitions are necessary to fully understand the proposed model.

DEFINITION 3.1. *A strategy $st_j$ of a player $j$ corresponds to a complete plan of actions, selected from a set of possible actions $ST^j$ that determines his/her behavior in any stage of the game. The player may, instead of using a fixed action $st_j$, define a probability distribution $p_j$ for the set $ST^j$ to determine his/her actions where $p_j$ is a mixed strategy.*

DEFINITION 3.2. *Let $p_j$ be the strategy of the $j$-th player in a set of $J$ players under a given game. A Nash Equilibrium is a vector of probabilities $p^* = (p_1^*, ..., p_J^*)$ containing the strategies of the players such that no player has incentives to change his/her particular strategy. If $ST^i(p)$ is the payout for the player $j$, then a Nash Equilibrium is*

$$ST^j(p^*) = \max_{p_j^*} S(p_1^*, ..., p_j^*, ..., p_J^*) \quad \forall j \in \{1, ..., J\}$$

DEFINITION 3.3. *If the Nash Equilibrium is an equilibrium to the game, and also is the equilibrium to all the sub-games, then that equilibrium will be a Perfect Sub-Game Equilibrium.*

The price filter is mainly formed by the strategy game theory. In the game, there are two players and each player has a finite strategy set. The whole process can be divided into three stages – the host pricing, the tenant deciding whether to rent or not, and if choosing to rent, the tenant giving a review for his or her living experience after leaving the house. In the whole process, each player has a possible strategy set. As stated in the Definition 3.1, we can use mixed strategy Nash equilibrium model to build the filter.

In the first stage, the host action set $ST^h$ contains three pricing levels $ST^h \in \{high, medium, low\}$ in the same cluster. However, the tenant action set is $ST^t \in \{rent, not\text{-}rent\}$ at first. If the tenant decides to rent, he will grade his living experience to evaluate whether the price, living environment, traffic, etc. are good or not. So the action set of the tenant expands to $ST^t \in \{rent, not\text{-}rent\} \cup l = \{0, 1, 2, ..., b\}$. If the tenant does not rent the house, $st^t = 0$. If the grade is bigger than the satisfaction threshold $a$, the tenant will gain the income for the satisfactory accommodation. The higher the satisfaction, the greater the income. However, if the grade is smaller than the satisfaction threshold, he will lose the house value. To quantitatively analyze these factors, below is the income function of the tenant $t$ when he takes the action $st^t$:

$$g_l^t = \begin{cases} 0 & st^t = 0 \text{ or } st^t = l = a \\ \frac{P_k}{b-l+1} & st^t = l \ \& \ l > a \\ -\frac{V}{l} & st^t = l \ \& \ l < a \end{cases} \tag{3.8}$$

where $l \in \{1, 2, .., b\}$ is the rating score from the tenant to the house. If the tenant grades a higher score than $a$ (satisfaction threshold), which means the house worth the price, his income is house price. On the contrary, he does not satisfy the value of the house, hence his income is the house value.

As we have described before, the host income is the gap between the price and the value. But if the tenant

gives a low grade in the review, it is going to have bad influence on the host income. Conversely, it will have positive effects. So the income of the host $h$ when a tenant takes the action $st^t$ is defined as:

$$g_l^h = \begin{cases} -V & st^t = 0 \\ P_k - V + \frac{(l-a)P}{2} & st^t = l \;\&\; l \geq a \\ P_k - V + \frac{(l-a)V}{2} & st^t = l \;\&\; l < a \end{cases} \tag{3.9}$$

$P_k$ is the rental price when the host takes action $k$, and $V$ denotes the house value. To evaluate the goodness of the high grades, the term of $\frac{(l-a)P}{2}$ is used, while the bad influence about the low grades is defined by $\frac{(l-a)V}{2}$.

In the process, the probability of the host choosing the action $k$ is $p_k^h$ (where $\sum_{k=1}^{|ST^h|} p_k^h = 1$), while the probability of the tenant reviewing action $l$ is $p_l^t$ (where $\sum_{l=1}^{b} p_l^t = 1$) after he rents the house.

In the strategy game model, the expected income of the host is $EP^h$:

$$EP^h = \sum_k \sum_l g_l^h p_k^h p_l^t$$

And the expected income of the tenant is $EP^t$:

$$EP^t = \sum_l \sum_k g_l^t p_l^t p_k^h$$

Based on the Kakutani's fix point theorem [11], we can prove that for every finite player, finite strategy game has at least one Nash equilibrium if we admit mixed strategy equilibrium as well as pure. In our paper, the strategy sets of the two players are all limited, hence there is at least one equilibrium point in the mixed strategy game theoretic model.

In order to find out the mixed strategy Nash equilibrium, we have:

$$EP^h = EP^t \tag{3.10}$$

Based on this equation, we can form the filter of Nash equilibrium to remove unreasonable prices and its effectiveness will be described in detail in Section 4.

**3.3 Linear Regression with Normal Noise** After Nash Equilibrium filtering, we get a clean data. As rental price is closely related to the distance of landmark, the popularity of landmark and the house facility coverage, we use these factors to indicate the house price which could be modeled via a multivariable linear regression model. However, due to the uncertainties in reality, it is better to capture the unsureness as the noise $\Delta P$ that will be further explained in Section 4.

$$P = W \begin{bmatrix} H \\ d \\ q \end{bmatrix} + distribution(\Delta P) \tag{3.11}$$

Here, $H$ is the facility variety, $d$ the distance to the nearest landmark that is calculated by the Euler Distance, $q$ the viscosity coefficient of the landmark, and $W$ the parameters. In addition, since houses within similar district with similar house facilities are likely to have similar rental prices, we train a LRNN in each cluster separately. In this way, each LRNN has more stable and dedicated data and can train better regression models.

## 4 Experiments

In this section, we conduct experiments to verify the effectiveness of the proposed framework. Via the experiments, we aim to answer two questions:

- Can MSAP generate reasonable clusters by aggregating the landmark and house facility information into affinity propagation?
- Is the proposed framework effective in predicting reasonable price?

To answer these two questions, we use the same datasets used in Section 2.

**4.1 Quality of MSAP in Clustering** To answer the first question, we first perform MSAP on Los Angeles, London and Tokyo datasets. In the stage of LC, we aggregate the landmarks which are crawled from TripAdvisor into several districts based on the negative Euler distance with viscosity coefficient in (3.2). In the stage of HC, the houses in the same district are grouped into several clusters based on the facility status. Because of using the nearest distance, if there is more than one district nearest, we select the one which has more landmarks. We visualize the clustering results in Figure 5.

To evaluate the quality of clusters, we use Silhouette measurement [4] as the evaluation metric, which is a popular method for evaluating the clustering performance. The definition of Silhouette measurement is:

$$Sil_i = \frac{B_i - A_i}{max(A_i, B_i)} \tag{4.12}$$

where $A_i$ denotes the average dissimilarity $D$ between house $i$ and all the other houses in the same cluster. Let $C_i$ be the set of all the clusters except the cluster where the house $i$ is in. Then $F(i, C)$ is the set of dissimilarities from house $i$ to the houses in the cluster $C$. And $B_i = min(F(i, C))$. The dissimilarity is defined as the summation of the geo-distance difference and facility variety difference.

We compare the performance of MSAP with K-Means and Affinity Propagation [7]. For K-Means and Affinity Propagation (AP), we use both geo-location and facilities as features to perform clustering. The

Table 4: Values of silhouette

| Methods | Los Angeles | London | Tokyo |
|---|---|---|---|
| MSAP | 0.27 | 0.075 | 0.17 |
| Kmeans | -0.13 | -0.059 | 0.070 |
| AP | -0.66 | -0.46 | -0.76 |

results are showed in Table 4. From the table, we make the following observations:

- MSAP outperforms both Kmeans and AP, which demonstrates the effectiveness of MSAP by using both landmark and house facilities to perform two stage clustering.
- Kmeans and AP don't perform well. Possible reasons leading to the low Silhouette performance include – (1) two houses have almost the same facilities but are far away from each other, and (2) two houses are in the neighborhood with completely different facilities.

Since the focus of this paper is to make reasonable price recommendation. By performing MSAP, we expect the rental prices and house facilities within the same cluster to be similar. Thus, we also calculate the distortion of prices and the distortion of dissimilarities in each cluster. The smaller the distortion is, the more similar the prices or house facilities are. The average distortion of dissimilarities and prices are reported in Table 5 with the standard deviations. From the table, we can see that MSAP, which fuses house location and the facilities information appropriately, has the best performance in price and facility clustering.

In summary, MSAP gives better clusters in terms of both Silhouette and distortion, which demonstrates the effectiveness of MSAP in clustering and achieves the expected goal that houses in the same cluster have similar prices and facilities.

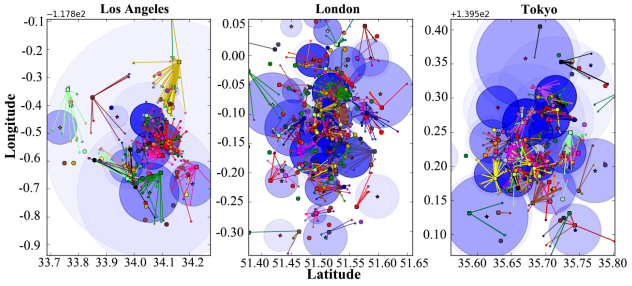


Figure 5: A visualization of clusters identified by MSAP. The stars are the center landmarks, the dots are the houses, blue circles are the districts from the LC, and the dots with the same colored edges are the clusters from the HC.

**4.2 Quality of Nash Equilibrium and LRNN in Price Predicting** To answer the second question, we perform price prediction using LRNN based on the cluster information getting from MSAP. The data is first divided into 80% and 20% where 80% are used as training. Since the training data contains outliers, i.e., unreasonable prices, we perform Nash equilibrium to clean the training data. We empirically set $a = 3$, $b = 5$, and assume that tenant actions of giving the review ratings follow the uniform distribution. So the expected incomes of the tenant are $EP^t$:

$$EP^t = \sum_l \sum_k g_l^t p_l^t p_k^h = -\frac{3}{10}V + \frac{3}{10}(p_1^h P_1 + p_2^h P_2 + p_3^h P_3)$$
$$= -\frac{3}{10}V + \frac{3}{10}\bar{P}$$

And the host expected income is below,

$$EP^h = \sum_k \sum_l g_l^h p_k^h p_l^t = -\frac{13}{10}V + \frac{13}{10}(p_1^h P_1 + p_2^h P_2 + p_3^h P_3)$$
$$= -\frac{13}{10}V + \frac{13}{10}\bar{P}$$

Then we can find that:

$$EP^t > 0, EP^h > 0 \rightarrow \bar{P} > V \tag{4.13}$$

According to (3.10), the equilibrium point is $V = \bar{P}$. Based on (4.13), a higher price $P$ will bring in more income for both hosts and tenants. But before the host increasing the house price, he needs to enrich the house facilities first to enhance the value of the house, which will bring the tenant much satisfaction in return. So during the filter building, we remove the price which is under $\bar{P}$, and the remaining house prices can be used as the training data. After the filtering, we analyze the price gap $\Delta P$ distribution between reasonable prices and the average prices in each city. The results are shown in Figure 6. From the figure, we can see that price gaps follow the normal distribution, but have different shapes in different cities. Tokyo are sharper than other two cities, which suggests that Los Angeles and London have wide price ranges, while the house price in Tokyo is more concentrated.

Let $p_{pred}$ be the predicted price of a house and $p_{true}$ be the true price. We use RMSE (Root-Mean-Square Error) as a metric to assess the prediction performance. Meanwhile the predicted price around the true price could be also acceptable to users. Therefore, we define the precision as the percentage of predictions that are within [-K, K] range of the true price. The prediction performance is demonstrated in Table 6. Note that for the precision performance, we choose $K = 10$. The baseline methods in the table are defined as – (1) IMean: it makes predictions as the average price

Table 5: The distortion of the dissimilarity and price in each city using those three methods.

| | | Kmeans | AP | MSAP |
|---|---|---|---|---|
| Los Angeles | Dissimilarity | 0.074(±0.007) | 0.041(±0.009) | **0.032(±0.007)** |
| | Price | 38.880(±1.925) | 43.729(±2.963) | **33.890(±2.112)** |
| London | Dissimilarity | 0.039(±0.001) | 0.029(±0.008) | **0.014(±0.003)** |
| | Price | 36.056(±0.897) | 44.481(±2.355) | **31.479(±1.386)** |
| Tokyo | Dissimilarity | 0.016(±0.002) | 0.010(±0.003) | **0.008(±0.001)** |
| | Price | 13.449(±0.809) | 16.695(±1.424) | **11.765(±0.860)** |

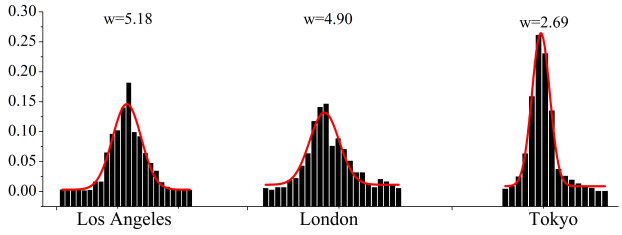


Figure 6: Price gap distributions are different in different cities, and all of them follow the normal distributions where $w$ is the deviation.

of similar houses that is a variant of item-based CF method; (2) LRNN: it only performs linear regression with normal noise that is a variant of the proposed framework without MSAP and Nash Equilibrium filtering; (3) Filter+LRNN: it is LRNN with Nash Equilibrium filtering; (4) LC+LRNN : it performs landmark clustering first and then uses LRNN to predict prices; (5) LC+Filter+LRNN: it is Filter+LRNN with landmark clustering; (6) MSAP+LRNN: it performs landmark and house clustering and then adopts LRNN to predict prices; and (7)MSAP+Filter+LRNN: it is the proposed framework. From the table, it can be observed that the proposed framework obtains much better performance than IMean that supports the necessity to investigate the unique properties to recommend prices in vacation rental websites. We also note that the performance of the LRNN prediction is improved consistently with clustering components. Finally, the results with Nash Equilibrium filter are enhanced in Los Angeles and London; Especially in Tokyo where there is a huge increase in prediction accuracy. This observation can be explained by Figure 6 – house prices are concentrated in Tokyo; while those in Los Angeles and London are more diverse. The x-axis in the figure is the gap, and the y-axis stands for the gap proportion.

## 5 Related Works

Predicting rental price is a popular economic research topic that has been well studied by economist [8, 9, 1, 14]. For example, Gallin et al. [8] investigate the relationship between house prices and rents. Ayuso et al. [1] study the impact of discount in rental price. These works focus on traditional house rentals, i.e., long term offline house rentals. However, rentals in online vacation rental websites such as Airbnb and FlipKey are different from these as the majority of rentals in such websites are for group travels or vacations and are short-term. Thus, price prediction for traditional house rental is not applicable to these sites.

As these websites are becoming more and more popular, they have attracted increasing attention from researchers [15, 10]. Lee [13] finds that the features of house price have close relations with sales, which always bother the hosts. So an attractive reasonable price is the key to both the hosts and tenants. Ikkala et al. [10] investigate the host rental experience qualitatively and suggest on how to gain the reputation and trust from the guests. Edelman [5] reports that non-black hosts in New York City charged approximately 12% more than black hosts for equivalent rentals. Choi [3] uses the panel regression model to investigate the impacts of Airbnb on the hotel revenue, and Lee et al. [13] analyze the features that are significantly associated with house sales. However, the work on predicting reasonable price is very limited. Given the importance of reasonable price prediction, in this paper, we investigate the novel and challenging problem of predicting reasonable price on websites such as Airbnb. Specifically, given the fact that the majority of the tenants on Airbnb are travelers which care more about landmarks near rooms and the room facilities, we tackle the problem from the landmark and room facility perspectives.

## 6 Conclusions

In this paper, we study the novel and important problem of the reasonable price prediction in vacation rental websites. We propose a framework which consists of three components, i.e., MSAP for clustering the houses into groups based on landmark and house facility information, Nash equilibrium for eliminating unreasonable prices and LRNN for predicting price. Experimental results on three real-world datasets demonstrate that, MSAP can cluster the house accurately and aggregate

Table 6: The price recommendation performance.

| | Los Angeles | | London | | Tokyo | |
|---|---|---|---|---|---|---|
| | Precision | RMSE | Precision | RMSE | Precision | RMSE |
| IMean | 18.73% | 79.40 | 12.95% | 61.88 | 33.26% | 36.75 |
| LRNN | 11.24% | 89.29 | 28.91% | 81.69 | 31.28% | 44.01 |
| Filter+LRNN | 17.93% | 38.46 | 16.08% | 49.90 | 28.22% | 26.11 |
| LC+LRNN | 30.36% | 45.31 | 27.42% | 51.87 | 41.94% | 25.58 |
| LC+Filter+LRNN | 33.55% | 26.28 | 41.93% | 26.95 | 43.65% | 20.07 |
| MSAP+LRNN | 37.45% | 39.51 | 25.52% | 49.06 | 74.27% | 13.14 |
| MSAP+Filter+LRNN | 42.84% | 30.82 | 53.01% | 40.07 | 77.51% | 21.47 |

the houses into different districts. The gap distribution between the house price and the city mean price reflects the diverse price in each city. It helps LRNN provide an accessible way for the reasonable price prediction that can be further enhanced by the filter of Nash Equilibrium and the clustering components.

There are several interesting directions needing further investigation. First, currently we only use the information about the distance from the landmarks, and the popularity of the nearest landmarks; while semantic information such as user comments can be used to further improve the precision of the price recommendation. Second, the price of house will fluctuate seasonally and we would like to consider temporal information for price prediction.

## 7 Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61402373).

## References

[1] Juan Ayuso and Fernando Restoy. House prices and rents in spain: does the discount factor matter? *Journal of housing economics*, 16(3):291–308, 2007.

[2] Richard J. Cebula, Reese Goldman, and Michael Toma. Housing prices in the savannah historic landmark district: Preliminary analysis. *Journal of Global Business Issues*, 2, 2008.

[3] K Hong Choi, J Hyun Jung, S Yeol Ryu, S Do Kim, and S Min Yoon. The relationship between airbnb and the hotel revenue: In the case of korea. *Indian Journal of Science and Technology*, 8(26), 2015.

[4] Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):1–21, 2002.

[5] Benjamin G. Edelman and Michael Luca. Digital discrimination: The case of airbnb.com. *Social Science Electronic Publishing*, 2014.

[6] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[7] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[8] Joshua Gallin. The long-run relationship between house prices and rents. *Real Estate Economics*, 36(4):635–658, 2008.

[9] Eden Hatzvi and Glenn Otto. Prices, rents and rational speculative bubbles in the sydney housing market. *Economic Record*, 84(267):405–420, 2008.

[10] Tapio Ikkala and Airi Lampinen. Defining the price of hospitality: Networked hospitality exchange via airbnb. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing*, CSCW Companion '14, pages 173–176, 2014.

[11] S. Kakutani. A generalization of brouwer's fixed point theorem. *Duke Mathematical Journal*, 8(3):457–459, 1941.

[12] Katherine A. Kiel and Jeffrey E. Zabel. Location, location, location: The 3l approach to house price determination . *Journal of Housing Economics*, 17(2):175–190, 2008.

[13] Donghun Lee, Woochang Hyun, Jeongwoo Ryu, Woo Jung Lee, Wonjong Rhee, and Bongwon Suh. An analysis of social features associated with room sales of airbnb. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 219–222. ACM, 2015.

[14] Kamila Sommer, Paul Sullivan, and Randal Verbrugge. The equilibrium effect of fundamentals on house prices and rents. *Journal of Monetary Economics*, 60(7):854–870, 2013.

[15] L. Zekanovic-Korona and J. Grzunov. Evaluation of shared digital economy adoption: Case of airbnb. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1574–1579, 2014.