Ann. Data. Sci. (2015) 2(4):453–469
DOI 10.1007/s40745-015-0057-5

Segmentation of Chinese Urban Real Estate Market: A Demand-Supply Distribution Perspective

Jichang Dong[1] · **Xiuting Li**[1] · **Wencong Li**[1] · **Zhi Dong**[1]

Received: 2 June 2015 / Revised: 3 December 2015 / Accepted: 6 December 2015 /
Published online: 17 December 2015

Abstract This study proposed a new perspective on the analysis of the regional features of real estate market and explored a more reliable segmentation method for Chinese urban real estate market based on the optimization of supply-demand resource distribution. A two-stage clustering procedure is proposed based on supply and demand elements and market performance respectively. And six clustering algorithms were used to divide 283 Chinese cities at the prefecture level or above into three clusters and 13 sub-clusters, which are identified as key regulatory region, stable development region and region that needs policy support. Differentiated regulatory policy suggestions are accordingly provided for each cluster.

Keywords Real estate market · Segmentation · Clustering analysis · Gini coefficient

1 Introduction

Real estate market segmentation is essential for property assessment [1,2] and real estate investment portfolio diversification [3]. It should also serve as the basis for the differentiated real estate market regulation that has been the major effort made by Chinese government to control high housing price in recent years. However, there has been no commonly-agreed segmentation method for the Chinese real estate market so far. The so-called "differentiated real estate market regulation" are either based on administrative division (e.g. provinces or municipalities), or on geo-economic division (e.g. Eastern, Western and the middle-area of China). Are these segmentation methods

✉ Xiuting Li
lixiuting@ucas.ac.cn

[1] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

Table 1 The IDI and SDI of Chinese cities in different regions (average from 2009 to 2011)

	Chinese mainland	Eastern China	Middle area of China	Western China	Beijing-Tianjin-Hebei	Pearl River Delta	Yangtze River Delta	Bohai Bay
IDI	0.6331	0.5691	0.5575	0.6973	0.6711	0.4197	0.4914	0.6165
SDI	0.6682	0.4793	0.6146	0.6630	0.6795	0.2884	0.3591	0.6321

appropriate? In order to answer this question, we proposed a new perspective on the analysis of the regional features of real estate market and the assessment of the rationality of using administrative division or geo-economic division as the basis for differentiated regulation. And two indicators, i.e. real estate investment distribution index (IDI) and commercial property sales distribution index (SDI), were constructed to measure distribution of the demand and supply of urban real estate market, by referring to the methodology of the Gini coefficient that has been widely used to evaluate the distribution of personal income. The Gini efficient can be expressed and calculated in many ways. And one of the most popular and simplest is the formula proposed by [4], referring to which we got the formulae to calculate IDI and SDI.

$$IDI = \frac{2\mathrm{cov}ar(inve, rank_{inve})}{N\overline{inve}} \tag{1}$$

$$SDI = \frac{2\mathrm{cov}ar(sale, rank_{sale})}{N\overline{sale}} \tag{2}$$

where $rank_{inve}, rank_{sale} \in [1, N]$ are the rankings of Chinese cities according to the real estate investment volume and sales volume, N is the number of cities, $\overline{inve}$ and $\overline{sale}$ are the average investment volume and average sales volume of N cities, $\mathrm{cov}ar(inve, rank_{inve})$ is the covariance of investment $inve$ and investment ranking $rank_{inve}$, $\mathrm{cov}ar(sale, rank_{sale})$ is the covariance of sales volume $sale$ and sales ranking $rank_{sale}$.

By using the two formulae, we calculated the IDI and SDI of Chinese mainland and cities in main economic zones. It can be found that except cities in the Pearl River Delta Economic Zone and the Yangtze River Delta Economic Zone, the IDI and SDI of cities in other economic zones and the Chinese mainland as a whole are above the warning level of 0.4[1], among which the IDI and SDI of Beijing-Tianjin-Hebei economic zone, the Bohai Bay economic zone and Chinese mainland as a whole exceed 0.6 (See Table 1). The results show that a polarization phenomenon widely exists in the real estate markets of China and many economic zones, and especially severe at the demand level. Besides, the regional imbalance also indicates that the current real estate market segmentation method based on geo-economic division is not appropriate, and a unified regulation policy for the regionally unbalanced real estate market cannot achieve the expected targets. More importantly, inappropriate regulation policy may aggravate the imbalance of the market, which would be harm

[1] The warming level was set by the United Nations for the Gini coefficient.

to people's well being and the sustainable development of the real estate market. Therefore, it is very important to explore a sound segmentation method to aggregate cities with similar abilities of capturing demand-side and supply-side resources, and thus provide a theoretical basis for the differentiated government regulatory policy.

There are mainly two theoretical perspectives on the real estate market segmentation in previous literature, which are segmentation based on factors influencing consumers' housing preferences and segmentation based on real estate market performance [5,6]. The first is a long-term perspective that emphasizes on classifying the real estate market into submarkets according to key factors that influence the market. And the second is a short-term perspective which divides the real estate market into groups according to indicators of the actual market performance. A majority of the available literature took the first perspective and most of them focus on the segmentation at the intra-city level, i.e. delineating a city's local market [7–11]. In contrast, only a few studies have focused on the national and/or regional level housing market segmentation, i.e. segmenting the 30 metropolitan US housing markets [12] and 71 Turkish metropolitan residential markets [2].

There are even fewer Chinese researches on real estate market segmentation. The report *Top 10 Most Attractive Prefecture-level Chinese Cities for Real Estate Investment* issued by China Index Academy since 2010 adopted a 7-cluster partition for prefecture-level Chinese cities based on three indicators: commercial housing sales income, GDP and permanent resident population [13]. Though the partition method was oversimple without theoretical basis, the result can be a useful reference for our present study because it was the first study on housing market segmentation for Chinese prefecture-level cities. Another study proposed a new time series clustering method that integrated both wavelet analysis and DBScan's algorithm, and divided 70 Chinese cities into six groups [14]. The perspective of this study was close to the second type, i.e. a market performance-based perspective, but the authors did not discuss the reasons behind different market performance.

Our present study combined both perspectives and tried to get a more reliable segmentation of the real estate markets of prefecture-level Chinese cities based on the optimization of supply-demand resource distribution by a two-stage clustering analysis. The remainder of this paper is arranged as follows. In Sect. 2, we describe the two-step clustering procedures, including defining the variables, algorithm selection and validity evaluation. In Sect. 3, we present the clustering results of real estate markets in 283 Chinese cities. The last section offers a summary of this paper and future research directions.

2 Model and Methodology

2.1 Analytical Procedures of Two-Stage Clustering

The performance of urban real estate market is determined by basic elements of supply and demand. If two cities possess similar supply and demand resources, the development level of their real estate market tends to converge in the long run. Therefore, segmentation based on supply and demand elements aggregates cities with converging

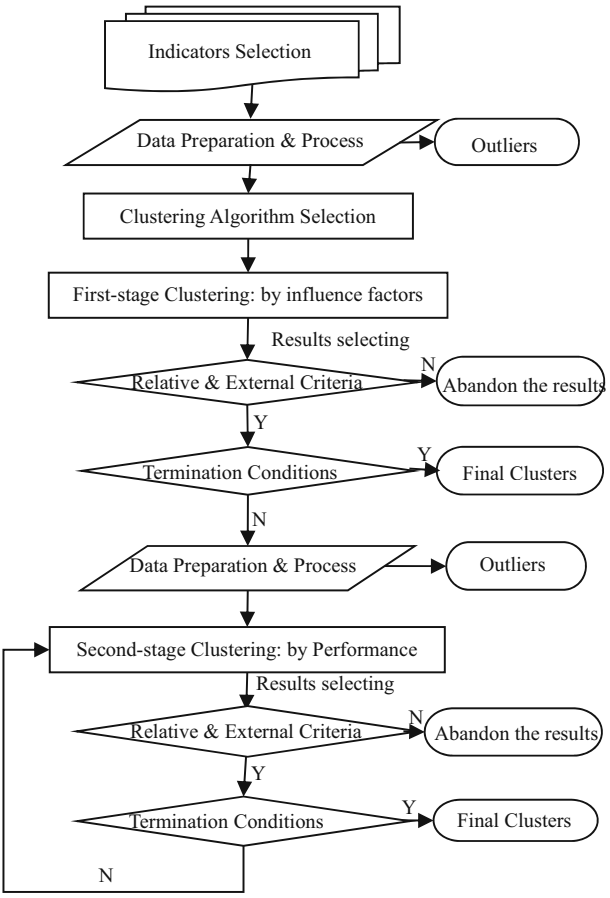

Fig. 1 Two-stage clustering procedure

long-term real estate market development into one cluster. However, market performance is also affected by other factors such as speculation [15] and government policy [16]. The performances of urban real estate markets at the same level of supply and demand elements may vary because of these factors. In segmentation based on market performance, cities with similar short-term performances can be aggregated into one cluster considering the fluctuation and variance of urban real estate market caused by factors other than the fundamental ones. The two perspectives respectively focus on the long-term and short-term development of urban real estate market.

This study adopted both perspectives and conducted clustering analysis for Chinese urban real estate markets according to both market performance and supply-demand elements. And a two-stage clustering procedure was constructed as shown in Fig. 1. The first stage was the clustering based on supply-demand elements. Cities in the sample were aggregated according to their potential supply and demand elements and economic fundamentals. The second stage was the clustering based on urban real estate market performance, in which cities with similar short-term market performances

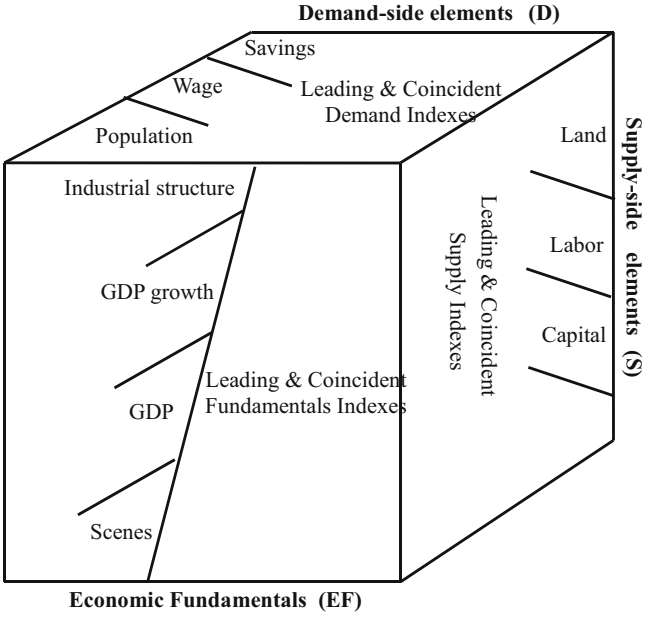

Fig. 2 Segmentation cube for real estate market

were aggregated into one cluster. Both stages involve key procedures such as handling outliers, selection of clustering criteria, and selection of clustering algorithm, as well as clustering result evaluation and screening.

2.2 Segmentation Cube and Indicators Selection

For the Indicators Selection in stage one, this study constructed a segmentation cube for real estate market (Fig. 2) with reference to the concept of "The Economic Cycle Cube" of the Economic Cycle Research Institute[2]. Under this framework, we divided leading and coincident factors influencing the real estate market into three dimensions including supply-side elements, demand-side elements and economic fundamentals. We further summarized key indicators of the 3 dimensions based on classic literatures on real estate economics. Then 12 indicators were selected for stage-one clustering through a review of the literature and with the consideration of data availability. For the dimension of economic fundamentals, GDP, GDP growth rate, percentage of secondary industry, percentage of tertiary industry and scenes were selected to measure economic aggregate, growth and structure [15,17,18]. For the dimension of demand-side elements, population at year-end, natural population growth rate, saving deposit of urban and rural households at year-end and average wage of employed persons were selected to measure a city's population, population growth and the purchasing power of residents [19–22]. For the supply-side elements, total amount of foreign investment

[2] Available:https://www.businesscycle.com/pdf/ECRI_Putting_it_All_Together.pdf.

actually utilized, number of employed persons, land area and deposits of financial institutions at year-end were selected to measure the stock amount and incremental amount of real estate supply [23–25]. For stage-two clustering, investment in real estate, commercial housing sale volume and commercial housing price were used as the clustering indicators, according to which the results of stage one were aggregated.

2.3 Selection of Clustering Algorithm and Relative Evaluation Criteria

Clustering algorithms for real estate market segmentation include partitional algorithms, hierarchical algorithms, density-based algorithms, graph-based algorithms, model-based algorithms and grid-based algorithms [26]. Each of the six most commonly-used clustering algorithms has specific advantages and disadvantage as shown in Table 2. This study selected several classic ones from them for the clustering of urban real estate markets, including K-means [27], hierarchical clustering [28], DBSCAN [29], MST [30], SOM [31] and WaveCluster [32]. These algorithms complement each other and thus ensure the validity of clustering results.

Generally, three approaches can be used to evaluate the validity of clustering results. The first is external evaluation that compares clustering results with predetermined benchmark classes of the dataset, and the validity is measured by the closeness of clustering results to predetermined classes. The second is internal evaluation which deals with datasets with unknown structures and compares clustering results with the dataset's inherent characteristics. The validity is usually defined by compactness and separation and measured by intra-cluster variances and correlation coefficients. The third approach is relative evaluation, in which the best clustering algorithm is determined by comparing different algorithms or the clustering results of one algorithm under different parameter settings.

Classic indices in relative evaluation include DUNN index [33] and SD index [34]. Both external and internal evaluations are based on statistical analysis and have certain limitations. For example, external evaluation requires classification of datasets in advance, but correct predetermined classification is difficult to obtain in practice.

Table 2 Performance analysis of clustering algorithm

Algorithm	Discovering arbitrary-shaped dataset	Discovering dataset with uneven density distribution	Unaffected by noise	Identifying neighboring dataset	Less reliance on prior knowledge	Time complexity
K-means	×	✓	×	✓	×	O(n)
Hierarchical	×	×	×	×	✓	$O(n^2)$
DBSCAN	✓	×	✓	×	×	O(nlog(n))
MST	✓	×	✓	✓	×	$O(n^2)$
SOM	×	×	×	✓	×	O(kmn)
WaveCluster	✓	×	✓	×	×	O(n)

Notes: ✓ = yes; × = no; k is the number of clusters for SOM, m is the number of neurons for SOM

The research of [8–11] used the out-of-sample forecasting accuracy of hedonic price model to test the validity of clustering, which in essence is one type of external evaluation. Relative evaluation does not require statistical test and is the most widely-used clustering evaluation method. We took relative evaluation as the basis and developed an external criteria based on IDI and SDI according to the idea of supply-demand optimization.

After preliminary clustering by the six algorithms, SD validity index was used to fine-tune the clustering results. The SD validity index is defined based on the average scattering for clusters and total separation between clusters, as given by the equation:

$$SD(c) = \alpha Scat(c) + Dis(c) \tag{3}$$

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c} \|\sigma(v_i)\| \,/\, \|\sigma(X)\| \tag{4}$$

$$Dis(c) = \frac{D_{\max}}{D_{\min}} \sum_{i=1}^{c} \left(\sum_{j=1}^{c} \left\| v_i - v_j \right\| \right)^{-1} \tag{5}$$

where c is the number of clusters; $\alpha = Dis(c_{\max})$ is the weighting factor; $c_{\max}$ is the maximum number of input clusters. $\sigma(X) = \frac{\sum_{k=1}^{N} (x_k - \overline{x})^2}{N}$ is the variances of dataset X; $\overline{x}$ is the center of dataset. $\sigma(v_i) = \frac{\sum_{k=1}^{N_i} (x_k - v_i)^2}{N_i}$ is the variances of cluster i; v_i and v_j, $\forall i, j \in \{1, 2, ..., c\}$ are the respective centers of cluster i and cluster j; $D_{\max} = \max(\left\| v_i - v_j \right\|)$ and $D_{\min} = \min(\left\| v_i - v_j \right\|)$ are the maximum and minimum distances between cluster centers respectively.

As argued in [34], the SD index can be used to identify the best number of clusters. When a local optimum is found by SD and defined by the Eq. (6). And c_b is the best number of clusters.

$$|SD(c_b) - SD(c_b - 1)| < 1/3, c_b \in [2, c_{\max}] \tag{6}$$

2.4 Optimization of Demand and Supply Distribution and External Criteria

Cities aggregated into one cluster have similar ability in capturing supply and demand resources, i.e. cities of the same cluster have similar supply and purchasing power in the real estate market. Therefore a reasonable segmentation of the real estate market should ensure a balanced distribution of the real estate investments and commercial housing sales volumes of cities in each cluster. Based on the principle of supply and demand optimization, we used IDI and SDI as the external criteria and selected the result with the smallest IDI and SDI from the six clustering results that met the relative criteria (given by Eqs. (7), (8), and (9)). Smaller IDI and SDI indicated more balanced distribution of supply and demand resources of cities in the cluster and thus reflected better clustering result.

$$\underset{1\le i\le a}{Min}\left(mean^{i}(IDI_j)\right), \forall j \in \{1, 2, ..., c_i\} \tag{7}$$

$$\underset{1\le i\le a}{Min}\left(mean^{i}(SDI_j)\right), \forall j \in \{1, 2, ..., c_i\} \tag{8}$$

$$IDI_j \le 0.3, SDI_j \le 0.3 \tag{9}$$

where, $a = 6$ is the six algorithms; c_i is the best number of clusters that meets the relative criteria by algorithm i; IDI_j and SDI_j are the real estate investment distribution index and commercial property sales distribution index of cluster j respectively; $mean^i(IDI_j)$ and $mean^i(SDI_j)$ are the average investment distribution index and average sales distribution index of clusters by algorithm i.

We further set a threshold value of external criteria as the termination condition of clustering procedure. In General, a Gini coefficient below 0.2 indicates an absolutely equal wealth distribution, and that between 0.2 and 0.3 indicates a relatively mean distribution. The wealth distribution is relatively reasonable when the Gini coefficient is between 0.3 and 0.4. The coefficient between 0.4 and 0.5 means a medium level of wealth inequality, and that higher than 0.5 indicates a serious level of wealth inequality [35]. This study used 0.3 as the threshold value of external criteria. The clustering process stopped when the IDI and SDI of clusters were lower than 0.3 which is given by Eq. (9) and the result was considered as the final. The external criteria based on IDI and SDI has both statistical significance and economic implication, and thus complements the relative criteria. Compared to the external criteria based on out-of-sample forecasting accuracy of hedonic price model proposed by [8–11], IDI and SDI are easy to calculate and the requirement on data quality is far lower than hedonic price model. It is a more economical evaluation method for real estate market segmentation.

3 Empirical Analysis

3.1 Data Collection and Preprocessing

Our sample included 283 cities at the prefecture level or above from 30 provinces, autonomous regions and directly-controlled municipalities of the Chinese mainland except the Tibet Autonomous Region. Data of the 283 cities from 2009 to 2011 were collected from China Economic Information Network, CEIC database, the statistical yearbook of provinces, autonomous regions and municipalities, China City Statistical Yearbook and China Index Academy Database. Average values were calculated and the data were normalized.

The indicator of "scenes" requires coding to the scenes. In the theory of scenes, scenes are used to measure the amenity of certain cities and mainly refer to educational and cultural facilities, entertainment facilities and medical facilities.

Indicators used in this study included the number of higher education institutions, the number of secondary schools, the number of primary schools, the number of college and university faculty, the number of secondary school faculty, the number of primary school faculty, the number of theaters, the number of public library collections per hundred persons, the number of hospitals and heal centers per hundred persons, the

number of hospital beds, the number of doctors, the number of public transportation vehicles per ten thousand population, per capita area of paved roads, coverage rate of afforestation in developed area. The index scores were calculated by the following factor weighting method:

$$Scene_i = \sum_{j}^{m} \frac{F_{ij} \times \sigma_j}{\sum_{j}^{m} \sigma_j} \tag{10}$$

where $Scene_i$ is the index score of city i; m is the number of extracted common factors; F_{ij} is the score of factor j of city i; σ_j is the proportion of variance explained of factor j.

Considering the possible interference of outliers, data were preprocessed according to Pauta criterion (3σ criteria) before each clustering analysis. Outlier cities were sorted out for independent analysis and the remaining cities were aggregated into clusters by the two-stage clustering procedures.

3.2 Clustering of Outlier Cities

Outliers were defined on the basis of the average value of each indicator from 2009 to 2011. Twenty-five cities were sorted out from the sample. Each had certain indicator values far higher or lower than the average level. These cities had an IDI of 0.6110 and a SDI of 0.6225, both much higher than the threshold level of 0.3. Segmentation of the real estate markets of these cities required an independent round of clustering.

Firstly, the 25 outlier cities were respectively clustered by six algorithms. Then the six clustering results were compared according to the SD index. It was found that the optimal number of clusters by K-means, SOM, Hierarchical, MST and DBSCAN was two, but WaveCluster algorithm got only one cluster. Besides, Hierarchical clustering and MST clustering got identical results. We further calculated the IDI and SDI of each clustering result of the six algorithms. The results of Hierarchical clustering and MST clustering had the smallest average IDI and SDI and were thus considered as the optimal result as shown in Table 3.

Table 3 Clustering results of outlier cities

Clustering results	K-means	SOM	Hierarchical	MST	DBSCAN	WaveCluster
Number of clusters	2	2	**2**	**2**	2	1
IDI	0.4968	0.4968	**0.3945**	**0.3945**	0.5179	0.6110
SDI	0.5215	0.5215	**0.3337**	**0.3337**	0.5237	0.6225
Number of clusters	**3**	2	**3**	3	**3**	1
IDI	**0.3494**	0.4163	**0.3494**	0.4464	**0.3494**	0.5974
SDI	**0.4306**	0.4267	**0.4306**	0.4092	**0.4306**	0.5981

The result of the algorithm bolded was optimal measured by relative and external criteria

The 25 cities were preliminarily divided into two clusters. Cluster 1 included Beijing and Shanghai. Both its SDI and IDI were lower than the threshold level of 0.3. Since the termination condition was met, the result was considered as final. Cluster 2 included the remaining 23 cities. Its IDI and SDI were 0.5974 and 0.5981 respectively and should go through the second clustering procedure.

Table 3 also shows the results of the second clustering, which was based on market performance. K-means, Hierarchical and DBSCAN got identical results, and according to external criteria and relative criteria, the result was considered as optimal. The 23 cities were further divided into three sub-clusters. One of them met the termination condition and become one cluster of the final result. The other two sub-clusters went through another round of outlier processing and re-clustering and were further divided into three groups.

In the end, the 25 outlier cities were divided into five clusters, each having IDI and SDI lower than 0.3 as the termination condition was fulfilled. The first cluster included Beijing and Shanghai. Both have much bigger markets and higher housing prices than other cities. The second cluster included ten cities represented by Guangzhou, Shenzhen and Tianjin. The housing supply and demand volumes and prices of these cities are second only to cities in the first cluster. The third cluster included 7 cites represented by Hezhou and Heihe. These cities have relatively undeveloped real estate markets and low housing prices. The fourth cluster included Fangchenggang and Yulin. Their real estate markets have good supply and demand basis but the housing prices are low, indicating bigger potential of price increase.

3.3 Clustering of Non-outlier Cities

After removing outliers, the remaining 258 cites were clustered by the above-mentioned six algorithms. According to relative criteria and external criteria, the result of K-means clustering was optimal as shown in Table 4, in which the 258 cities were clustered into two clusters. The IDI and SDI of the 23 cities in the first cluster were 0.2440 and 0.2430 respectively, both lower than the threshold of 0.3. Clustering stopped and the cluster was put in the final result. The IDI and SDI of the 235 cities in another cluster were 0.4337 and 0.4730 respectively, both higher than 0.3, which means the termination condition was not fulfilled and further clustering based on real estate market performance was needed. Eleven outlier cities in terms of market performance were removed after outlier detection for real estate market performance of cities in the second cluster. The supply and demand volumes and housing prices of

Table 4 Results of the first clustering of the non-outlier cities

The first clustering	K-means	SOM	Hierarchical	MST	DBSCAN	WaveCluster
Number of clusters	**2**	2	3	2	2	1
IDI	**0.3388**	0.3863	0.5426	0.3739	0.5425	0.5441
SDI	**0.3580**	0.3795	0.5547	0.3877	0.5552	0.5584

The result of the algorithm bolded was optimal measured by relative and external criteria

Table 5 Results of the second clustering of the non-outlier cities

The second clustering	K-means	SOM	Hierarchical	MST	DBSCAN	WaveCluster
Number of clusters	**3**	2	2	5	2	1
IDI	**0.2279**	0.2900	0.2173	0.3160	0.2963	0.4050
SDI	**0.2102**	0.2936	0.2121	0.3408	0.2952	0.4367
Number of clusters	3	**4**	3	3	3	1
IDI	0.2244	**0.1571**	0.3084	0.1971	0.2096	0.3207
SDI	0.2112	**0.1570**	0.3154	0.2977	0.2893	0.3158

The result of the algorithm bolded was optimal measured by relative and external criteria

these cities exceeded the remaining 224 cities by far. The IDI and SDI of the 11 cities were 0.2049 and 0.1756 respectively, both lower than 0.3. Therefore the 11 cities were grouped into one cluster in the final result. The IDI and SDI of remaining 224 cities did not meet the termination condition and need further clustering.

After removing outliers, the 224 cities were clustered by the six algorithms. According to relative criteria and external criteria, the result of K-means clustering was optimal as shown in Table 5, in which the 224 cities were clustered into three clusters. Cluster 2–1 included 19 cities represented by Anshan, Baotou and Haikou. The termination condition was fulfilled and the cluster was considered as part of the final result. Cluster 2–2 included 69 cities represented by Guilin, Linyi and Beihai and was also considered as part of the final result. The remaining 136 cities constituted cluster 2–3. The IDI and SDI of this cluster were 0.3207 and 0.3158 respectively, indicating that further clustering was needed. The 136 cities in Group 2–3 went through another round of two-stage clustering. The result of SOM was optimal measured by relative and external criteria. The 132 cities were divided into four sub-clusters. The first sub-cluster included 28 cities represented by Baoji and Bozhou; the second sub-cluster included 22 cities represented by Anyang and Binzhou; the third sub-clustser included 21 cities represented by Hanzhou and Guang'an; the fourth sub-cluster included 65 cities represented by Ankang and Baiyin. IDIs and SDIs of all the four sub-groups were lower than the threshold level of 0.3 and no further clustering was needed.

3.4 Results and Discussion

In the final result, the real estate markets of 283 Chinese cities were divided into three clusters and 13 sub-clusters by the two-stage clustering procedures (as shown in Table 6; Figs. 3, 4). And the 52 cities in sub-cluster 1–6 constituted the key regulatory region. Cities in sub-cluster 1 and 2 have much higher level of supply and demand elements such as GDP, capital, scene index and purchasing power than other cities. For example, in sub-cluster 1, Beijing is the political and cultural center of China and Shanghai is the economic and financial center of China. Both have well-developed real estate markets. The serious imbalance between supply and demand has lifted housing prices out of reach. Sub-cluster 2 included big cities with robust economy such as Guangzhou, Shenzhen, Hangzhou and Tianjin. The real estate market size of Guangzhou and Shenzhen are just next to that of Beijing and Shanghai. The housing

Table 6 Clustering results for Chinese urban real estate markets

Regions	Sub-clusters	Cities	Investment	Sales	Price
Key regulatory region	Cluster1	Beijing, Shanghai (2)	231,505	308,016	14,764
	Cluster2	Guangzhou, Shenzhen, Tianjin, Hangzhou, Suzhou, Chongqing, Chengdu, Dalian, Shenyang, Qingdao (10)	101,091	122,212	8568
	Cluster3	Changchun, Changsha, Changzhou, Dongguan, Foshan, Fuzhou, Harbin, Hefei, Jinan, Kunming, Nanchang, Nanjing, Nanning, Ningbo, Shijiazhuang, Taiyuan, Tangshan, Urumqi, Wuxi, Wuhan, Xi'an, Xiamen, Zhengzhou (23)	52,325	48,899	6059
	Cluster4	Ordos, Guiyang, Jinhua, Langfang, Nantong, Shaoxing, Taizhou[1], Weifang, Yantai, Zhoushan, Zhuhai (11)	26,454	31,014	6161
	Cluster5	Sanya, Wenzhou, Suihua, Yingkou (4)	20,254	20,204	8553
	Cluster6	Fangchenggang, Yulin (2)	8620	5443	2770
Stable development region	Cluster7	Anshan, Baotou, Haikou, Hohhot, Huzhou, Huai'an, Huizhou, Jiaxing, Quanzhou, Taizhou[2], Weihai, Wuhu, Xuzhou, Yancheng, Yangzhou, Yinchuan, Zhenjiang, Zhongshan, Zibo (19)	20,494	23,601	4714
	Cluster8	Anqing, Bengbu, Baoding, Beihai, Benxi, Cangzhou, Chengde, Chizhou, Chifeng, Chuzhou, Daqing, Dandong, Dezhou, Dongying, Fushun, Fuyang, Ganzhou, Guilin, Handan, Heze, Hulun Buir, Huainan, Huangshan, Jilin, Jining, Jiangmen, Jinzhou, Jiujiang, Lanzhou, Leshan, Lishui, Lianyungang, Liaoyang, Linyi, Liuzhou, Liu'an, Longyan, Luoyang, Ma'anshan, Mianyang, Nanchong, Nanping, Ningde, Panjin, Putian, Qinhuangdao, Qingyuan, Qujing, Quzhou, Rizhao, Sanming, Shantou, Suqian, Tai'an, Tieling, Tongling, Xining, Xianyang, Xinxiang, Xinyang, Xuancheng, Yibin, Yichang, Zaozhuang, Zhanjiang, Zhangjiakou, Zhangzhou, Zhaoqing, Zhuzhou (69)	10,455	10,013	3640
	Cluster9	Baoji, Bozhou, Dazhou, Datong, Deyang, Fuzhou, Guigang, Huludao, Huaibei, Jiamusi, Jingzhou, Kaifeng, Liaocheng, Luzhou, Maoming, Meishan, Mudanjiang, Neijiang, Qiqihar, Qinzhou, Shangrao, Shaoguan, Songyuan, Suzhou, Xiangtan, Yangjiang, Yuxi, Zigong (28)	5857	5816	3017

Table 6 continued

Regions		Sub-clusters	Cities	Investment	Sales	Price
Policy Region	Support	Cluster10	Anyang, Binzhou, Changde, Chaoyang, Chenzhou, Hengshui, Hengyang, Jiaozuo, Nanyang, Shangqiu, Shiyan, Suining, Tonghua, Xingtai, Xuchang, Yichun, Yiyang, Yueyang, Ziyang, Zunyi, Zhumadian, Zhoukou(22)	7527	5962	2425
		Cluster11	Bayan Nur, Baise, Guang'an, Hanzhong, Huaihua, Huanggang, Ji'an, Jingmen, Loudi, Pingdingshan, Puyang, Shaoyang, Siping, Tongliao, Weinan, Ulanqab, Xianning, Xiaogan, Yongzhou, Yuncheng, Wuzhou (21)	4706	3876	2263
		Cluster12	Ankang, Baiyin, Baoshan, Chaozhou, Ezhou, Fuxin, Guangyuan, Heyuan, Huangshi, Jixi, Jincheng, Jingdezhen, Laiwu, Lijiang, Linfen, Panzhihua, Pingxiang, Qitaihe, Qingyang, Shanwei, Shuangyashan, Tianshui, Wuhai, Wuzhong, Wuwei, Ya'an, Yan'an, Yingtan, Yulin, Yunfu, Zhangye, Anshun, Bazhong, Baicheng, Baishan, Changye, Chongzuo, Dingxi, Guyuan, Hechi, Hebi, Hegang, Jiayuguan, Jieyang, Jinchang, Laibin, Liaoyuan, Liupanshui, Longnan, Lvliang, Luohe, Meizhou, Pingliang, Sanmenxia, Shangluo, Shizuishan, Shuozhou, Suizhou, Tongchuan, Xinzhou, Xinyu, Yangquan, Zhangjiajie, Zhongwei, Zhaotong (65)	2221	2031	2645
		Cluster13	Hezhou, Heihe, Jinzhong, Jiuquan, Karamay, Lincang, Yichun (7)	1815	1514	2333

Note: Taizhou[1] and Taizhou[2] are respectively "台州" and "泰州" in Chinese

prices of these cities are also very high. Hangzhou has smaller economic aggregate and lower levels of supply and demand elements than Guangzhou and Shenzhen, but it has higher housing price and larger market size. It should be attributed to Hangzhou's outstanding economic performance in southeastern China and its key location in the Yangtze-River Delta. Both Tianjin and Chongqing are directly-controlled municipal cities and have similar economic aggregates compared to Guangzhou and Shenzhen, but their real estate market sizes are smaller. The housing price of Chongqing ranks at the end of sub-cluster 2 cities. Other cities like Chengdu, Dalian, Qingdao and Shenyang are characterized by good supply and demand, large market size and big growth potentials. Sub-cluster 3 was constituted by provincial capitals and large cities represented by Wuhan, Tangshan, Foshan, Jinan and Changchun. Common features of these cities are high levels of economic development, large real estate market size and high housing price. Sub-cluster 4 contained cities with high levels of real estate market

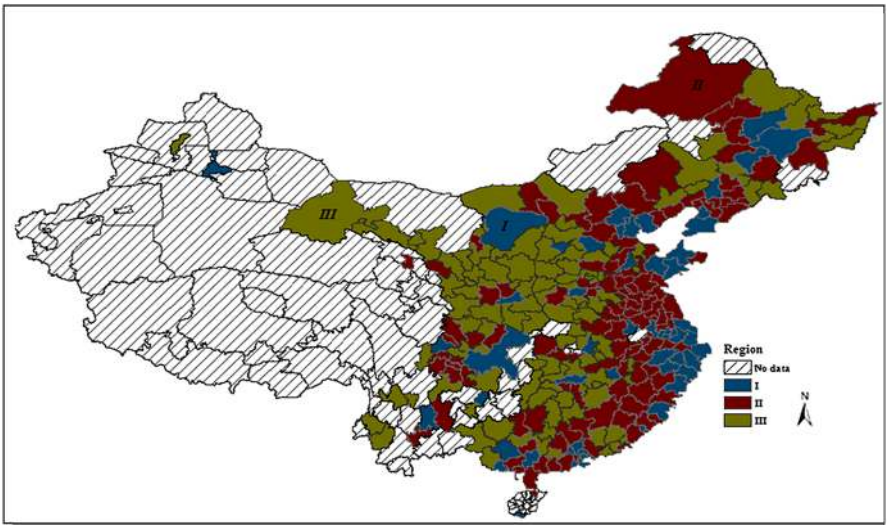

Fig. 3 Geographical distribution of the three regions of China urban real estate market

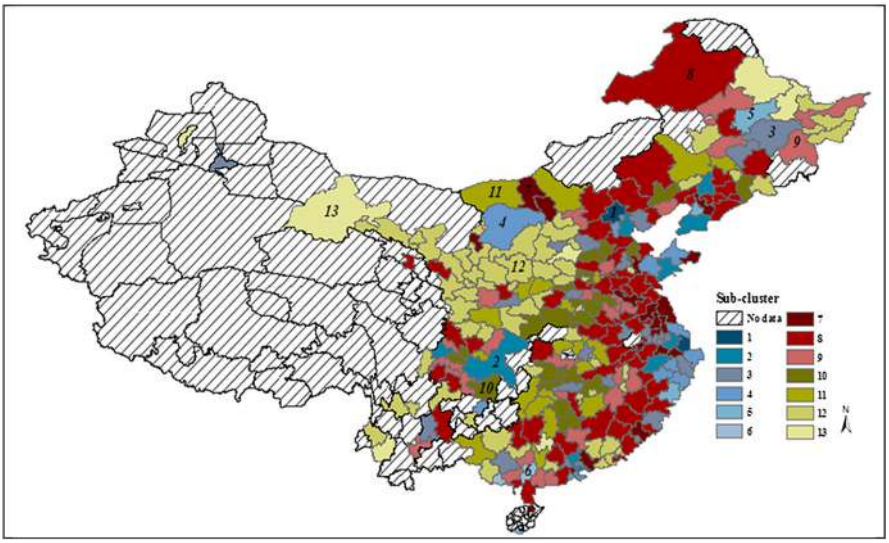

Fig. 4 Geographical distribution of the 13 sub-clusters of China urban real estate market

development and low levels of economic development as represented by Ordos, Shaoxing and Taizhou. Among cities in sub-cluster 5, Sanya is an international tourism city and has high level of trade openness. Its real estate market develops faster than local economy and the housing price is even higher than that of Shanghai. Wenzhou has an untamed shadow banking system and abundant capital for speculation and investment. Its real estate market develops with astonishing speed and the housing price is higher than that of Shanghai. Imbalance between real estate market performance and

economic development also exists in Suihua and Yingkou. The two cities have bigger market sizes and higher housing prices among cities with similar levels of supply and demand elements. Both Fangchenggang and Yulin in sub-cluster 6 have uncoordinatedly high housing prices compared to their low levels of economic development. The 116 cities in sub-cluster 7–9 constituted the stable development region. All of these cities have moderate sizes of real estate markets and moderate levels of housing prices. The third cluster was the region that needs policy support and contained 115 cities with underdeveloped real estate markets in sub-cluster 10–13.

It should be pointed out that the clustering results depend on the set threshold value of IDI and SDI to certain extent. A threshold value higher than the current 0.3 would make the termination condition easier to reach and less sub-clusters would got through less rounds clustering procedures. For example, if the threshold value was set as 0.35, the sub-cluster 1 and sub-cluster 2 would be aggregated into one cluster, and the sub-cluster 10, sub-cluster 11 and sub-cluster12 would also be aggregated into one cluster. Then ten sub-clusters would be finally got. However, the result wouldn't affect the division of three clusters of the market and main conclusions of this study.

3.5 Policy Implication

Our present study has practical implications for government policy. Specific regulatory objectives and policies can be made for each sub-cluster based on their supply and demand level and market performance. For cities in the key regulatory region, the government can implement a strict regulation. The goal is to keep a balanced supply and demand and controlling the growth speed of housing price from rising too fast. For Beijing and Shanghai in sub-cluster 1 and Guangzhou and Shenzhen in sub-cluster 2, government policy should focus on the structural adjustment of supply and demand and give priority to the optimization of supply structure and scale. Cities in sub-cluster 5 have uncoordinatedly booming real estate market and high housing price compared with their economy development. Strict regulation should be implemented to prevent speculation from driving up prices. For cities in other sub-clusters, the government can make a moderate regulation. For cities in stable development region, the key is to keep a real time monitoring of the real estate market and stabilize the housing price. For the region that needs policy support. Government policy should shift from control to stimulation, and provide support to local residents' demand and the development of local real estate industry and market. Adjustments should be made from time to time according to the development of real estate markets. For example, Wenzhou and Sanya have posted decline in housing price. Local governments can turn to a moderate regulatory policy to avoid risks of a hard landing of the real estate market.

4 Conclusion

This study proposed a new perspective on the analysis of the regional features of real estate market and explored a more reliable segmentation method for Chinese urban real estate market based on the optimization of supply-demand resource distribution. With reference to the methodology of the Gini coefficient, this study proposed two indices

to measure the supply-demand distribution of urban real estate market—Investment distribution Index (IDI) and sales distribution Index (SDI)—and discussed the importance of market segmentation for Chinese real estate market. This study designed a two-stage clustering procedure and made a clustering analysis for Chinese cities of prefecture-level and above by six different algorithms. The 283 cities in the sample were clustered into 3 clusters and 13 sub-clusters. This study further proposed differentiated regulatory policies for the real estate market of each sub-cluster. Innovations of this study are as follows: firstly, we proposed a clustering method based on optimization of supply and demand and designed a two-stage clustering procedure based on supply-demand elements and market performance respectively. Secondly, we constructed the segmentation cube to facilitate index selection and used IDI and SDI as the external evaluation criteria which had both statistical basis and economic implications. Thirdly, at the policy level, the clustering result of this study can be a framework and useful reference for differentiated regulation by the government.

We are aware that this study was a static analysis of the real estate market and did not reflect the short-term market dynamics. The effect of the change of some major factors is not considered. In future researches, we will incorporate time dimension into clustering, for example the time series clustering of housing price fluctuation, and consider the dynamic changes of clustering result to achieve a real time monitoring of real estate markets in each sub-cluster.

Acknowledgments We would like to thank for the financial support of the Project of National Natural Science Foundation of China (No.71173213, No. 71203217, No. 71403260), and the Project of China Postdoctoral Science Foundation (No. 2013M540129).

References

1. Bourassa SC, Hamelink F, Hoesli M, MacGregord BD (1999) Defining housing submarket. J Hous Econ 8:160–183
2. Bourassa SC, Hoesli M, Pema VS (2003) Do housing submarkets really matter? J Hous Econ 12:12–28
3. Hepsen A, Vatanseer M (2012) Using hierarchical clustering algorithms for Turkish residential market. Int J Econ Financ 4:38–50
4. Pyatt G, Chen CN, Fei J (1980) The distribution of income by factor components. Q J Econ 95:451–473
5. Kauko T, Hooimeijer P, Hakfoort J (2002) Capturing housing market segmentation: an alternative approach based on neural network modeling. Hous Stud 17:875–894
6. Isiam KS, Asami Y (2009) Housing market segmentation: a review. Rev Urban Reg Dev Stud 21:93–109
7. Watkins CA (2001) The definition and identification of housing submarkets. Environ Plan A 33:2235–2253
8. Goodman AC, Thibodeau TG (1998) Housing market segmentation. J Hous Econ 12:121–143
9. Goodman AC, Thibodeau TG (2003) Housing market segmentation and hedonic prediction accuracy. J Hous Econ 12:12–28
10. Goodman AC, Thibodeau TG (2007) The spatial proximity of metropolitan area housing submarkets. Real Estate Econ 35:209–232
11. Helbich M, Brunauer W, Hagenauer J (2013) Data-driven regionalization of housing markets. Ann Assoc Am Geogr 103:871–889
12. Abraham JM, Goetzmann WN (1994) Homogeneous grouping of metropolitan housing market. J Hous Econ 3:186–206
13. China Index Academy (2013) Top 10 most attractive prefecture-level Chinese cities for real estate investment in 2013. http://fdc.soufun.com/news/zt/201305/2013zgdj.html

14. Guo K, Wang J, Shi G, Gao X (2012) Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. Procedia Comput Sci 9:1299–1305
15. Kuang WD (2010) Expectation, speculation and urban housing price volatility in China. Econ Res J 9:67–78
16. Yuan ZG, Fan YY (2003) Financial intermediation and relationship banking. Econ Res J 3:34–43
17. Quigley JM (1999) Real estate prices and economic cycles. Int Real Estate Rev 2:1–20
18. Shen Y, Liu HY (2004) Housing prices and economic fundamentals: a cross city analysis of China for 1995 to 2002. Econ Res J 6:78–86
19. Mankiw NG, Weil DN (1989) The baby boom, the baby bust, and the housing market. Reg Sci Urban Econ 19:235–258
20. Hort K (1998) The determinants of urban house price fluctuations in Sweden 1968–1994. J Hous Econ 7:93–120
21. Li JY, Sun G, Li G (2011) Effects of changes in the number of intergenerational population on stock market and real estate market in US and China. Manag World 8:171–172
22. Chen BK, Xu F, Tan L (2012) Demographic change and housing demand in China 1999–2025: a micro-empirical analysis based on the census data. J Financ Res 1:129–140
23. Peng RJ, Wheaton WC (1994) Effects of restrictive land supply on housing in Hong Kong: an econometric analysis. J Hous Res 5:263–291
24. Zhang JQ, Wu YH (2009) On the issue of foreign capital flow's effect on the real estate price in China. J Syst Eng 5:568–573
25. Wang ST (2009) Urban openness and real estate prices: empirical evidence from thirty-five large scale Chinese cities. Nankai Econ Stud 2:91–102
26. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. Intell Inf Syst 17:107–145
27. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkley symposium on mathematical statistical and probability, vol 1. University of California Press, Berkley, pp 281–297
28. Theodoridis S, Koutroubas K (1999) Pattern recognition. Published by Academic Press, San Diego
29. Ester M, Kriegel HP, Sander J et al (1996) A density-based algorithm for discovering clusters in large spatial database with noise. In: Proceedings of 2rd international conference on knowledge discovery and data mining, pp 226–231
30. Zahn CT (1971) Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans Comput C–20:68–86
31. Kohonen T (1995) Self-organizing maps., Springer Series in Information SciencesPublished by Springer, New York
32. Sheikholeslami G, Chatterjee S, Zhang A (1998) WaveCluster: a multi-resolution clustering approach for very large spatial database. In: Proceedings of 24th VLDB conference, pp 428–439
33. Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. J Cybern 4:95–104
34. Halkidi M, Vazirgiannis M, Batistakis Y (2000) Quality scheme assessment in the clustering process. Princ Data Mining Knowl Discov 1910:265–276
35. Liu HY, Yang F, Xu YJ (2013) Analysis of China's urban housing conditions based on 2010 census data analysis. J Tsinghua Univ (Philos Soc Sci) 28(6):138–147