Hindawi
Complexity
Volume 2019, Article ID 5370961, 15 pages
https://doi.org/10.1155/2019/5370961

Research Article

Evaluation of Residential Housing Prices on the Internet: Data Pitfalls

Ming Li,[1] **Guojun Zhang**,[2] **Yunliang Chen**,[3] **and Chunshan Zhou**[1]

[1]*School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China*
[2]*School of Public Policy and Management, Guangdong University of Finance and Economics, Guangzhou 510275, China*
[3]*School of Computer Science, China University of Geosciences, Wuhan 430074, China*

Correspondence should be addressed to Guojun Zhang; guojunz@gdufe.edu.cn and Yunliang Chen; Cyl_king@hotmail.com

Received 29 November 2018; Accepted 27 January 2019; Published 19 February 2019

Guest Editor: Ke Deng

Copyright © 2019 Ming Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Many studies have used housing prices on the Internet real estate information platforms as data sources, but platforms differ in the nature and quality of the data they release. However, few studies have analysed these differences or their effect on research. In this study, second-hand neighbourhood housing prices and information on five online real estate information platforms in Guangzhou, China, were comparatively analysed and the performance of neighbourhoods' raw information from four for-profit online real estate information platforms was evaluated by applying the same housing price model. The comparison results show that the official second-hand residential housing prices at city and district level are generally lower than those issued on four for-profit real estate websites. The same second-hand neighbourhood housing prices are similar across each of the four for-profit real estate websites due to cross-referencing among real estate websites. The differences of housing prices in the central city area are significantly fewer than those in the periphery. The variation of each neighbourhood's housing prices on each website decreases gradually from the city centre to the periphery, but the relative variation stays stable. The results of the four hedonic models have some inconsistencies with other studies' findings, demonstrating that errors exist in raw information on neighbourhoods taken from Internet platforms. These results remind researchers to choose housing price data sources cautiously and that raw information on neighbourhoods from Internet platforms should be appropriately cleaned.

1. Introduction

Housing sale price statistics for 70 large and medium-sized cities released in December 2016 by the National Bureau of Statistics of the People's Republic of China revealed that, in December, prices of newly constructed housing in megacities had not changed from the previous month. Prices of newly constructed houses in provincial capitals and other large cities rose by 0.2% compared with the previous month, and prices in medium-sized cities increased by 0.4%. According to public opinion, these price levels are underestimated, and this triggers media and public discussion of the accuracy of housing statistics.

Property agents emerged after housing market reforms were implemented in China [1]. With the development of the Internet economy, real estate agency websites have been established. These websites provide masses of information and neighbourhood descriptive information for the renting and selling of residential property, constituting a location-aware form of big data [2]. Data from real estate agency websites has served as a valuable data source for scholars. Research content and results are diverse. Researchers use online housing price data to investigate the determinants of housing prices, relevant policy, and macroeconomic and social situations, such as tax policy, stamp duty [3], housing purchase restriction policy [4], institutional mediation [5], and disease [6]. Structural attributes, such as gross floor area, storey level [7], age of properties [8], and differentials between large-scale estates and single-block buildings [9] and location attributes, such as metro services [10], green space [11, 12], neighbouring and environmental effects [13], and the effects of theme parks on local areas [14], were all investigated by using online housing price data. Moreover, online housing price data are employed to explain various

phenomena in the housing market, such as the spatiotemporal trends concerning housing price fluctuations [15], the spatial pattern of rent prices [16], the transmission of house price changes across quality tiers [17], the *housing ladder effect* [18], buyers' preferences for high-end residential property [19], and corruption in China's land market auctions [20]. Moreover, housing prices and neighbourhood information on real estate broker websites can be used as input variables for other models. Housing prices have been considered as influential factors when simulating urban growth [21], and neighbourhood data obtained from the Lianjia website have been used to create the Urban Form Index [22]. Housing prices on Internet information platforms have been extensively used in housing market research.

Internet real estate data have several advantages. First, users share housing information in a timely manner according to their own interests and they are willing to update this information. Internet real estate agency platforms can either hire their own agents or rent out some interfaces to other real estate agencies who can share their own property information. Furthermore, leasers can register accounts and list their own properties on these websites. The second advantage of Internet real estate data is that the cost of data acquisition is comparatively low. Most of the cost is paid by traditional real estate agencies. They gather and organize the data. The third advantage is that Internet real estate data are detailed. Users provide the location, type, structure, construction time, renovation pictures, and videos of an apartment or house. Some websites even provide information about local facilities, such as bus stops, supermarkets, hospitals, kindergartens, and subway stations. In addition, real estate agency websites document housing prices on different scales, at the city, district, subdistrict, and neighbourhood levels, as well as for individual houses and apartments.

However, as a type of big data, data on the real estate agency websites share the same defects. Sampling errors, measurement errors, aggregation errors, and errors associated with the systematic exclusion of information also exists [23]. The first reason for this is that the sampling process is biased. Due to the commercialization of real estate agencies, they do not tend to invest resources in areas where the market is small and the profit margins are low, such as suburban areas. Information density in developing areas is low, which may even lead to data blind zones. Furthermore, Internet housing data lack systematic validation. Some real estate agents may falsify lower housing prices to attract renters. Website operators and relevant government departments may have difficulty supervising such behaviour. Another reason for the inaccuracy of Internet housing data is the duplication of housing information. Different real estate agents may issue the same housing on the website. Each website calculates housing prices using their own property databases, and such problems certainly introduce errors to housing prices.

Although the accuracy of residential property prices is an essential foundation of real estate research and significant gaps exist between official housing prices and housing prices issued by each real estate agency website, the differences in various housing price data sources are likely to be overlooked. Much research uses housing prices without checking data reliability. Research about the quality of real estate price data products has mainly focused on various house price indexes [24–27]. Although house price indexes are crucial for academic research to more thoroughly understand the housing market, house price indexes are not intuitive for a public that lacks relevant background knowledge. Moreover, much research uses housing prices rather than price indexes as a data source [28]. How many differences exist among various property price data sources and to what extent these differences affect research are not yet known.

Accurate house prices are of theoretical importance and are crucial to understanding the operation of the housing market. Therefore, the primary objective of this study is to analyse differences among housing prices on mainstream online real estate information platforms and to evaluate the performance of neighbourhoods' raw information from for-profit online real estate information platforms by applying the same housing price model. Housing price data at the city and district levels from five Internet real estate information platforms were collected and compared. Then, second-hand neighbourhoods' housing prices from four for-profit Internet real estate information platforms were compared. Finally, raw information on neighbourhoods from the four platforms, including housing prices and the construction year, was put into the same hedonic housing price model to evaluate the performance of data on each platform (Figure 1). If results from the model contradicted other studies, the raw input housing information data were assumed to be unreliable.

2. China's Internet Real Estate Information Platforms

China's online real estate information platforms can be divided into four categories [29].

(1) Internet platforms for traditional bricks-and-mortar real estate agency firms: these websites are established by traditional real estate agency firms to promote their housing resources online. These websites serve mainly as a property database where agents and renters can search for housing information. Then, renter contact agents directly and continue the transaction offline. Typical platforms include Centaline Property, Lianjia, and Q Fang. Centaline Property (http://www.centanet.com), which has approximately 2,000 branches and over 60,000 employees in China, was selected. Centaline Property enjoys the largest portion of the Guangzhou real estate market. Lianjia.com (http://www.lianjia.com), the online platform of Beijing Lianjia Real Estate Brokerage Co., Ltd., was selected as a data source. Lianjia has approximately 8,000 stores and more than 1.3 million agents. Lianjia acquired My Top Home to enter the market in the Pearl River Delta in 2015.

(2) Internet real estate information platforms: these websites do not hire their own agents nor do they open stores. They serve as housing advertisement platforms. Traditional real estate agency firms can pay for their agents to release housing information on them, and individual users can share housing information for free. Moreover, such websites release real estate news and analysis reports. Anjuke, Sohu Focus, Sina Leju, and 58 Tongcheng are representative of such sites.

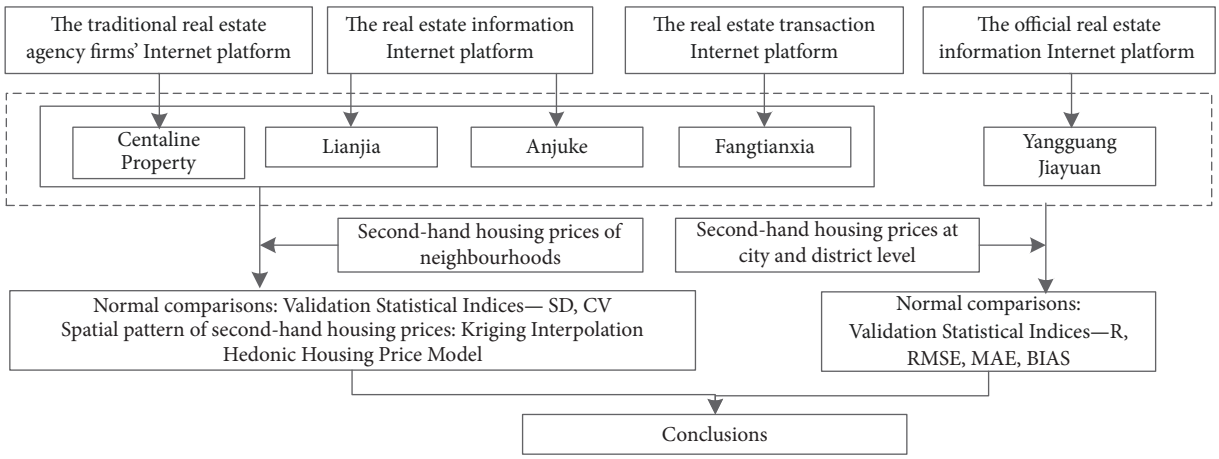


Figure 1: Flowchart of the housing price evaluation process.

Anjuke Inc. (http://www.anjuke.com), whose app has been installed 170 million times, was selected for this study. Its coverage of the market reaches 88%, encompassing 500 cities throughout the country.

(3) Real estate transaction Internet platforms: as for traditional real estate agency firms, such companies open stores and hire agents but not to the extent that Internet-based firms do. In contrast to traditional real estate agency firms' offline-to-online pattern, real estate transaction platforms started as online businesses and expanded offline. They also lend their interfaces to traditional real estate agency firms. Their stores are mainly for providing experiences and advertisements. Fangtianxia (formerly Soufang) is an example of this type of site. Fangtianxia, with more than 42 million registered users in March, 2015, hires approximately 3.7 million agents and covers more than 500 cities in China. It was the leading Internet portal for real estate in China, as measured by the numbers of page views and visitors to its websites in 2014, according to DCCI (http://www.dcci.com.cn).

(4) Official real estate information Internet platforms: relevant government departments, such as the Housing and Urban–Rural Development Bureau, create real estate information websites to release policies, property prices, property resources, and other information. Yangguang Jiayuan (http://www.gzcc.gov.cn/data/) is an official real estate information platform created by the Guangzhou Housing and Urban-Rural Construction Committee. Official statistical data, policies, and housing resources are released on it but its data volume is less than for-profit Internet real estate information platforms'.

3. Study Area and Data

3.1. Study Area. With a resident population of over 14 million in 2016, Guangzhou is the provincial capital of Guangdong Province, the economic and political centre of the Pearl River Delta region, and one of China's megacities. It was selected for the study because it has been at the forefront of reform since the 1980s and was the first provincial capital city to implement comprehensive housing system reform from 1978 [30]. Following many years of administrative division adjustments, Guangzhou now has 11 districts (Figure 2). According to the *Guangzhou Master Plan (2011–2020)*, the central area of Guangzhou contains Yuexiu District, Liwan District, Haizhu District, Tianhe District, the southern part

Figure 2: District Boundary of Guangzhou in 2016.

Table 1: List of the statistical indices for validation used to compare for-profit real estate information website housing prices with the official data.

Validation Statistical Index	Equation	Perfect Value
Pearson correlation coefficient (R)	$R = \frac{\sum_{i=1}^{n}(P_i - \overline{P})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}}$	1
Root Mean Squared Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2}$	0
Mean Absolute Error (MAE)	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert P_i - O_i\right\rvert$	0
Relative Bias (BIAS)	$\mathrm{BIAS} = \frac{\sum_{i=1}^{n}(P_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100\%$	0

Note: number of months is represented by n; P and O are the for-profit real estate information website housing prices and the official data, respectively.

of Baiyun District, the southern part of Huangpu District, and the northern part of Panyu District; other areas belong to the periphery of Guangzhou. The urban spatial structure of Guangzhou has developed to become polycentric. The traditional city centre is Renmin Park, near the municipal government in Yuexiu District, and the new city centre is Zhujiang New Town in Tianhe District [31–33].

3.2. Data. Data used in this study can be divided into real estate data and point of interest (POI) data. The business scope of real estate agency websites involves various types of commercial real estate, including residences, offices, shops, parking lots, and factory buildings. This research focuses on second-hand residential houses, which are closely related to people's livelihoods. Five representative real estate agency websites were selected: Centaline Property, Lianjia, Anjuke, Fangtianxia, and Yangguang Jiayuan. Second-hand residential housing prices at city and district level were collected between May 2015 and May 2016. Yangguang Jiayuan does not release second-hand neighbourhood housing prices. Each neighbourhood's housing prices and construction time on the four for-profit real estate information websites were collected on June 7, 2016, including 9,941 records from Centaline Property, 6,500 records from Lianjia, 8,198 records from Anjuke, and 6,119 records from Fangtianxia.

A point of interest (POI) is a specific point location that may be useful or of interest. It is a type of point datum representing a real geographic entity, including spatial information, such as latitude and longitude, and address; attribute information, such as names and categories, restaurants, stores, cinemas, and theatres. For this study, the locations of subway stations, parks, and key schools in Guangzhou were obtained from a Chinese map website, Gaode Map (http://www.amap.com) in June 2016. The location of each neighbourhood was obtained through the Gaode API (http://lbs.amap.com/console/show/picker).

4. Methodology

4.1. Normal Comparisons. Housing prices released on the official real estate information platform, Yangguang Jiayuan, are used as fiducial market prices in this study. To quantitatively compare for-profit real estate information website housing prices and official data, two types of statistical measure, namely, degree of agreement and error and bias, are used. Degree of agreement is represented by the Pearson correlation coefficient (R), which reflects the degree of linear correlation between the for-profit real estate information website housing prices and official data. In terms of error and bias, three statistical indices for validation were considered: the mean absolute error (MAE) represents the average magnitude of the error. Although the root mean square error ($RMSE$) also measures average error magnitude, it gives greater weight to the larger errors relative to the MAE. Relative bias ($BIAS$) describes the systematic bias of the for-profit real estate information website housing prices. Equations of each index are presented in Table 1.

The statistical indices for validation used to detect the variation in a neighbourhood's housing prices between websites were standard deviation and coefficient of variation. Standard deviation (SD) quantifies the variation or dispersion of a set of data values. The coefficient of variation (CV), also known as relative standard deviation (RSD), demonstrates the extent of relative variability. It expresses the precision and repeatability of a data set. The equations of each index are listed in Table 2.

4.2. Kriging Interpolation. Kriging is one of the optimal linear predictors based on spatial autocorrelation. The Kriging method predicts values on a continuous surface based on observed sampled data [34]. Housing price predictions at unobserved locations require geostatistical approaches, particularly Kriging interpolation. Kriging compares favourably to the ordinary least squares method (OLS) for predicting house prices [35].

4.3. Hedonic Housing Price Model. The hedonic model has been extensively employed in numerous empirical housing market studies and has proven to be effective [36, 37]. This model is therefore used to evaluate the performance of neighbourhoods' raw information from the for-profit online real

Table 2: List of the statistical indices for validation used to detect variations in a neighbourhood's housing prices between websites.

Validation Statistical Index	Equation	Perfect Value
Standard Deviation (SD)	$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{m}(p_i - \overline{p})^2}{m}}$	0
Coefficient of Variation (CV)	$\mathrm{CV} = \frac{\sqrt{(\sum_{i=1}^{m}(p_i - \overline{p})^2)/m}}{\overline{p}}$	0

Note: the number of websites is represented by m; p_i is the second-hand neighbourhood housing prices on website i.

Table 3: Descriptions of variables.

Variable		Description
Dependent variable	Price	Log of the second-hand neighbourhood's housing prices (¥/Chinese Yuan)
Independent variables	Year	The year of neighbourhood construction (a)
	Centre	Log of distance from the neighbourhood to the nearest city centre (m)
	Subway	Log of distance from the neighbourhood to the nearest subway station (m)
	School	Log of distance from the neighbourhood to the nearest key school (m)
	Park	Log of distance from the neighbourhood to the nearest park (m)

estate information platforms instead of other less common and more complicated models.

The hedonic model is based on Lancaster's [38] consumption theory. Goods are assumed to possess multiple characteristics in fixed proportions, and these characteristics—not the goods themselves—are assumed to dictate consumers' preferences. Rosen [39] developed market equilibrium theory. The aim of the hedonic pricing model is to assess the relationship between the market value of a composite good and each single attribute by generating a set of implicit prices for all these attributes. In general, housing price can be classified as [40]

$$P = f(S, L, N) \tag{1}$$

where P is the market price of a neighbourhood; S is structural attributes, such as construction time, building materials, and ratio of green space; L is location attributes, such as the distance to the city centre, shopping centre, and the nearest subway stations; and N is neighbourhood attributes, for example, school quality, environment quality, and natural scenery.

The three equation types most often used for hedonic price models are pure linear, semilog, and log–log. The two log forms are more appropriate than the linear form is, because the law of diminishing marginal utility applies to the situation. Coefficients of the log-log form are the percentage change in market price in response to a 1% change in each attribute's implicit prices. The log-log form was employed in this study. It can be defined as

$$\ln P = \alpha + \sum_{i=1}^{n} \beta_i \ln C_i + z \tag{2}$$

where P represents the market price of a neighbourhood; C represents the quantity of utilities or services that the house provides, concerning house characteristics and nearby infrastructure; for example, β is the regression coefficient of the study variables; α is the constant; z is the random error term; and n is the number of neighbourhoods [36].

Second-hand housing price data and neighbourhood information from the four for-profit real estate information websites are entered into the same hedonic house price model to test whether the use of different raw data sources would affect the modelling outcomes.

The descriptions of the explanatory variables that are used in the hedonic models are listed in Table 3. Neighbourhoods' property prices and the years in which they were constructed are obtained from Centaline Property, Lianjia, Anjuke, and Fangtianxia, respectively. *Price* is the log form of a neighbourhood's average second-hand housing price.

Housing prices are observed to have a negative relationship with age [8]. *Year* is the year when a neighbourhood was constructed. Location is widely recognized to be the primary determinant of housing price. The distance to the city centre accounts for a substantial proportion of variations in housing prices, which corresponds with the predictions of the bid-rent curve of renting prices [41]. Because Guangzhou is a polycentric city [31], the People's Government of Guangzhou Municipality and Zhujiang New Town are selected as the two city centres. *Centre* is the log form of distance from the neighbourhood to the nearest city centre. Most studies have concluded that the proximity of housing to subway stations positively affected value [10]. The subway station list used for this study was obtained from Guangzhou Metro (http://www.gzmtr.com/). *Subway* is the log form of the distance from the neighbourhood to the nearest subway station. The nearby enrolment policy and the school district system have been implemented since 1986 in China's compulsory education system. Key public schools have significant effects on housing prices [42]. The key school list used for this study was obtained from the Education Bureau of Guangzhou (http://www.gzedu.gov.cn/), and *School* is the log form of the distance from a neighbourhood to the nearest key school. A park is a major green space with ecological, entertainment,

Table 4: Validation indices for city-level second-hand housing price data.

Real Estate Agency Websites	R	RMSE	MAE	BIAS
Centaline Property	0.737	10222.360	10197.920	72.262
Lianjia	0.836	5772.276	5736.600	40.649
Anjuke	0.943	6192.564	6173.846	43.747
Fangtianxia	0.950	6384.281	6374.231	54.008

Notes: units for RMSE, MAE, and BIAS are CNY.

social, and cultural functionality. Therefore, other studies have concluded that house prices increase with increasing proximity to nearby parks [43]. For this study, parks in Guangzhou were found using Gaode Map. *Park* is the log form of the distance from a neighbourhood to the nearest park. The locations of neighbourhoods, subway stations, and key schools were obtained from the Gaode Map API. Distances were calculated based on their coordinates.

5. Comparison Results and Discussions

5.1. Second-Hand Housing Prices at City Level. The second-hand residential housing price trends from May 2015 to May 2016 are shown in Figure 3. The official second-hand residential housing prices released on Yangguang Jiayuan are significantly lower than those on the four for-profit real estate information websites. All the data present a steady upward trend in second-hand residential housing prices in Guangzhou. The price data released by Centaline Property are substantially higher than the other websites' data and they fluctuate dramatically. They also follow an upward trend in general. Prices from Lianjia, Anjuke, and Fangtianxia are similar. Their volatility is low and their rises are minimal.

The index values of each for-profit real estate information website's housing prices are listed in Table 4. Although housing prices at city level for Anjuke and Fangtianxia are considerably higher than those for Yangguang Jiayuan, their price data share an extremely similar trend, with correlation coefficients of 0.943 and 0.950, respectively. Second-hand residential housing prices in Guangzhou released by Lianjia and Yangguang Jiayuan also have a high correlation (the correlation coefficient reaches 0.836), whereas the correlation between the data of Centaline Property and Yangguang Jiayuan is relatively lower.

Several reasons could be suggested for why the official second-hand residential housing prices released on Yangguang Jiayuan are significantly lower than those on the four for-profit real estate information websites. (1) The final deal price is, in general, higher than the original offer price in second-hand residential housing transactions, because buyers usually bargain with landlords. Housing prices on Yangguang Jiayuan are based on the contract submitted to the relevant administrative housing department, namely, the final deal price; for-profit real estate information website housing prices are based on real estate agents' own databases. The quoted price and the final price are all included in the housing price model. Some transaction records may take the original offer price as the final deal price. (2) Real estate agency firms are for-profit. They do not tend to invest resources into lower priced or marginal regions, such as urban villages or suburban areas, such as Nansha District and Conghua District. Their agency branches are mostly distributed in district centres, where housing prices are higher than in other areas; however, the official data include all transaction records, meaning that many low-priced housing transactions are included. For example, Centaline Property has no branches in Nansha District or Conghua District. This situation magnifies the gap between official and for-profit real estate information websites' second-hand residential housing prices. (3) Tax evasion occurs. Some buyers sign twin contracts with the landlord, one of which is at a lower price and is submitted to the relevant administrative housing department to incur less tax [44].

Some variations also exist between price data of different for-profit real estate information websites because their housing resources are different in each district. For instance, Centaline Property has no housing resources in Nansha District or Conghua District, whereas the Anjuke website has 196 and 121 neighbourhoods in Nansha District and Conghua District, respectively.

5.2. Second-Hand Housing Prices at District Level. Guangzhou has undergone many administrative division adjustments. Therefore, variations exist in statistical division for different real estate agency websites. For this study, eight common districts were selected, namely Yuexiu District, Liwan District, Haizhu District, Tianhe District, Baiyun District, Panyu District, Huadu District, and Zengcheng District.

Figure 4 presents five websites' second-hand residential housing prices for eight districts in Guangzhou between May 2015 and May 2016. The official second-hand residential housing prices of each district are evidently still significantly lower than those from the for-profit real estate information websites, and the volatility of the official price data at district level is higher than that at city level. The overall trend is rising. The second-hand residential housing prices of Centaline Property in the eight respective districts are relatively high. Because housing resources in peripheral areas such as Huadu District and Zengcheng District are fewer than those in central areas, the second-hand residential housing prices of Centaline Property in these two districts are volatile. The second-hand residential housing prices of Lianjia in each district are the least volatile and maintain a gentle upward trend. Except in Yuexiu District, the gaps in second-hand residential housing prices among five websites are significant, both in suburban and urban districts. Although the housing prices of Lianjia, Anjuke, and Fangtianxia are at the same

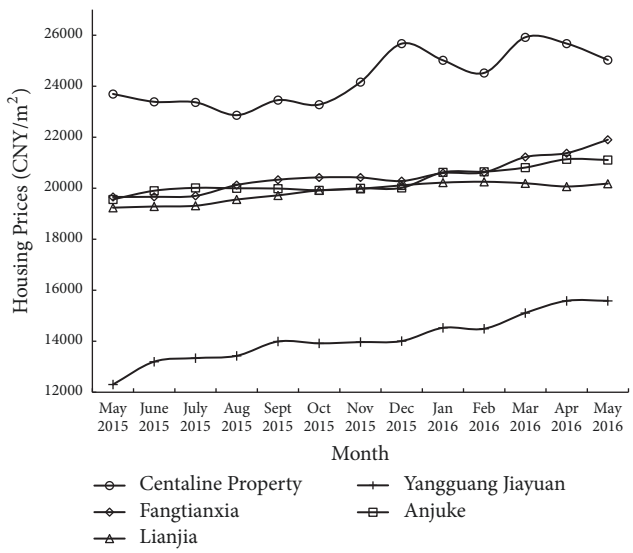


Figure 3: Second-hand residential housing prices in Guangzhou (May 2015-May 2016).

moderate level in Huadu District, Baiyun District, and Panyu District, those of Centaline Property are always higher and those of Yangguang Jiayuan are significant lower.

In general, the volatility of district-level prices is higher than that of city-level prices because of the smaller sample size at district level. In terms of prices, the official data of each district are still significantly lower, and Centaline Property's prices are at a comparatively high level for most districts. The gaps of second-hand residential housing prices among the five websites between prices do not vary significantly from suburban to urban districts.

5.3. Second-Hand Housing Prices of Neighbourhoods. Because Yangguang Jiayuan does not release the housing prices of each neighbourhood, only second-hand neighbourhoods' housing prices from four for-profit real estate information websites'—namely Centaline Property, Lianjia, Anjuke, and Fangtianxia—were collected. Then, 1,897 common neighbourhoods were selected. If each neighbourhood's housing prices were equal on the four websites, scatter plot points would be distributed on the $y = x$ line. Each neighbourhood's second-hand housing prices were compared pairwise on each of the four agency websites, and the results are given in Figure 5.

A pairwise comparison illustrates, as in Figure 5, that each neighbourhood's housing prices are similar on different for-profit real estate information websites. Moreover, each neighbourhood's housing prices demonstrate consistency between Centaline Property and Fangtianxia data. Consultations with real estate agency staff revealed that, in the real estate agency industry, second-hand neighbourhoods' housing prices are not only calculated with their own databases but can also be artificially modified using other real estate agency websites' price data. Therefore, one neighbourhood's housing prices can be approximately similar on different websites.

The same neighbourhood's housing prices on the four for-profit real estate information websites can be treated as one group. The *SD* and *CV* of each group were calculated. *SD* was used to quantify the variation of second-hand neighbourhoods' housing prices on different websites, and *CV* was used to express the relative variation. Kriging interpolation was used to detect the spatial distribution features of *SD* and *CV* for second-hand neighbourhoods' housing prices on different websites, and the interpolation results are provided in Figure 6.

The variation of each neighbourhood's housing prices on each website decreases gradually from the city centre to the periphery (Figure 6(a)), because the second-hand residential housing prices in the central area are higher than those in the suburban areas. However, the relative variation of prices is stable across Guangzhou (Figure 6(b)). The two peaks in the north of Tianhe District and in the south-east of Panyu District are caused by errors in websites' prices after verification, which also proves that errors occur in housing prices on the Internet.

5.4. Spatial Pattern of Second-Hand Residential Housing Prices. For this study, neighbourhoods with housing prices were selected and located using the Gaode Map API. Finally, 9,941 records from Centaline Property, 6,500 records from Lianjia, 8,198 records from Anjuke, and 6,119 records from Fangtianxia were entered into the Kriging interpolation, which was used to analyse the spatial pattern of second-hand residential housing prices in Guangzhou (Figure 7).

Spatial patterns of second-hand residential housing prices obtained using the four websites' data are seen to be similar in the central areas, namely Yuexiu District, Liwan District, Haizhu District, Tianhe District, southern Huangpu District, southern Baiyun District, and northern Panyu District. Second-hand neighbourhoods' housing prices in Zhujiang New Town, Ersha Island, Huijing New Town, Pazhou, and Baiyun Fortress Villa have the highest prices. The differences between the four spatial patterns of second-hand residential housing prices in the peripheral areas, such as Conghua

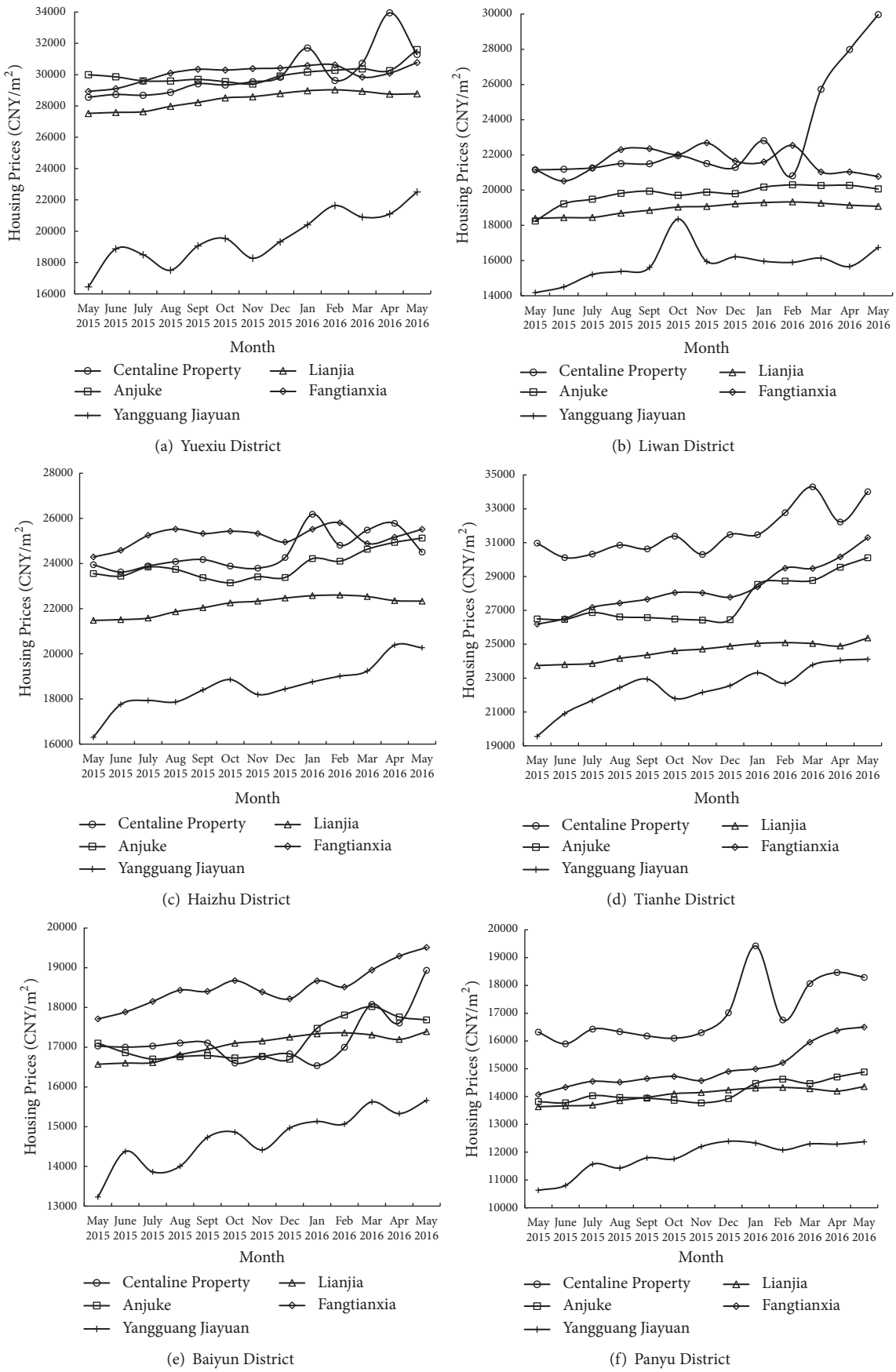
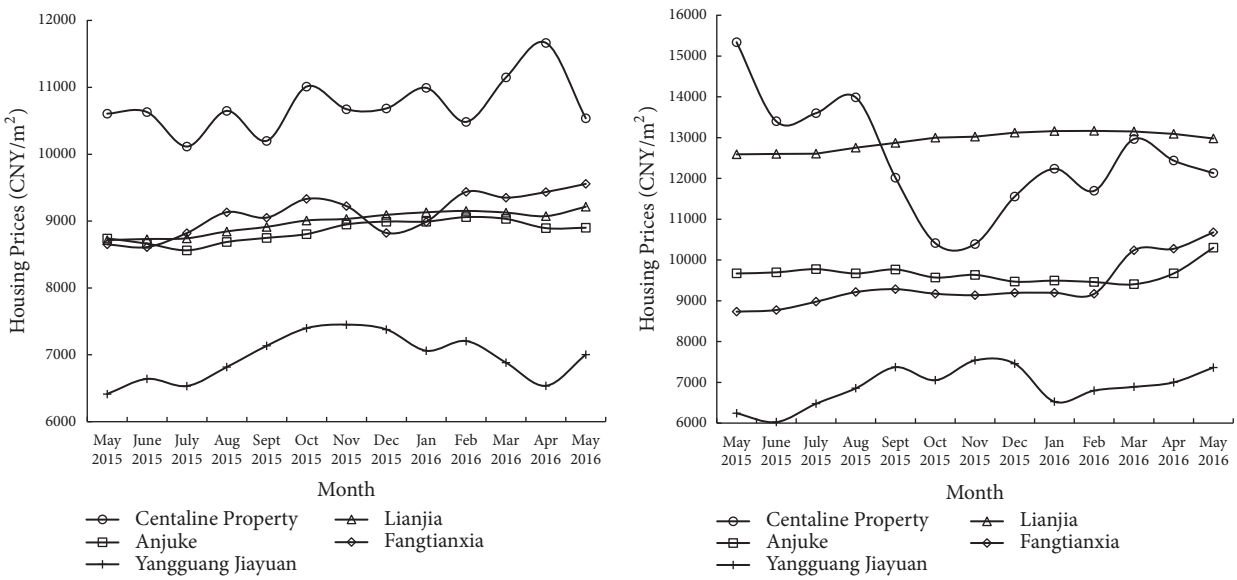


Figure 4: Continued.



(g) Huadu District (h) Zengcheng District

Figure 4: Second-hand residential housing prices of each district (May 2015-May 2016).

District, Zengcheng District, Huadu District, Nansha District, northern Panyu District, northern Baiyun District, and northern Huangpu District are significant. This is because fewer agency branches exist in suburban areas. Real estate agency firms tend to invest more resources in city centres, which leads to fewer peripheral second-hand neighbourhoods being included in the databases. The differences in second-hand neighbourhoods' housing prices on the four for-profit real estate information websites are therefore amplified after spatial interpolation.

5.5. Results of Hedonic Housing Price Models. Because we focused on evaluating the performance of raw data on neighbourhoods from Internet real estate information platforms in housing market research, a classic, reliable, and widely used model—the hedonic housing price model—was selected. If results of the model contradicted those of other studies, the raw input housing information data were assumed to be unreliable.

To maintain data consistency for the four for-profit real estate information websites in question, complex data cleaning was not applied. Neighbourhoods with relatively complete information were selected and used in the hedonic house price model. Statistical variations in second-hand neighbourhoods' housing prices on the four websites are provided in Figure 8. Anjuke has the largest dispersion of second-hand neighbourhoods' housing prices; Lianjia has the smallest. Centaline Property has the largest proportion of second-hand neighbourhood housing prices that qualify as being at the lower level.

The results of the hedonic housing price models using Centaline Property, Lianjia, Anjuke, and Fangtianxia neighbourhood data are listed in Table 5. All the four models' P values are less than 0.001 and F-statistic values are larger than 800. Therefore, all four models are effective. The Fangtianxia model's adjusted R^2 is the highest of the models' values, at 0.530, and the Centaline Property model's adjusted R^2 is the lowest of the models' values at 0.369.

The distance to the nearest city centre [45] and subway station [46] exhibits a significantly negative relationship to second-hand neighbourhoods' housing prices, as in other studies. According to the standardized coefficients of the distance from the neighbourhood to the nearest city centre (-0.516 in Centaline Property, -0.611 in Lianjia, -0.588 in Anjuke and -0.678 in Fangtianxia) and subway station (-0.150 in Centaline Property, -0.102 in Lianjia, -0.138 in Anjuke and -0.119 in Fangtianxia), the distance to the nearest city centre has a more substantial effect on second-hand neighbourhoods' housing prices than the distance to the nearest subway station does. This matches the findings from Shanghai [47] and Hangzhou [48], China. For all models except that of Centaline Property, the year of construction has a significant positive effect on second-hand neighbourhoods' housing prices, which agrees with the results of other studies [49]. The standardized coefficient of the construction year in Anjuke's model (0.119) is much lower than that in Lianjia (0.253) and Fangtianxia (0.231) models, which also reflects the effects, on one model, of using the data of different agencies. The distance to the nearest park has a significant positive effect on second-hand neighbourhoods' housing prices in all four models. This is inconsistent with findings from other research [43]. The distance to the nearest key school is established to have a significant positive effect in the Centaline Property and Lianjia models, which also contradicts other studies [50]. In Fangtianxia model, the distance to the nearest key school had a negative relationship with housing price, but this was not significant.

By building four hedonic housing price models and comparing their results with findings from other research, errors

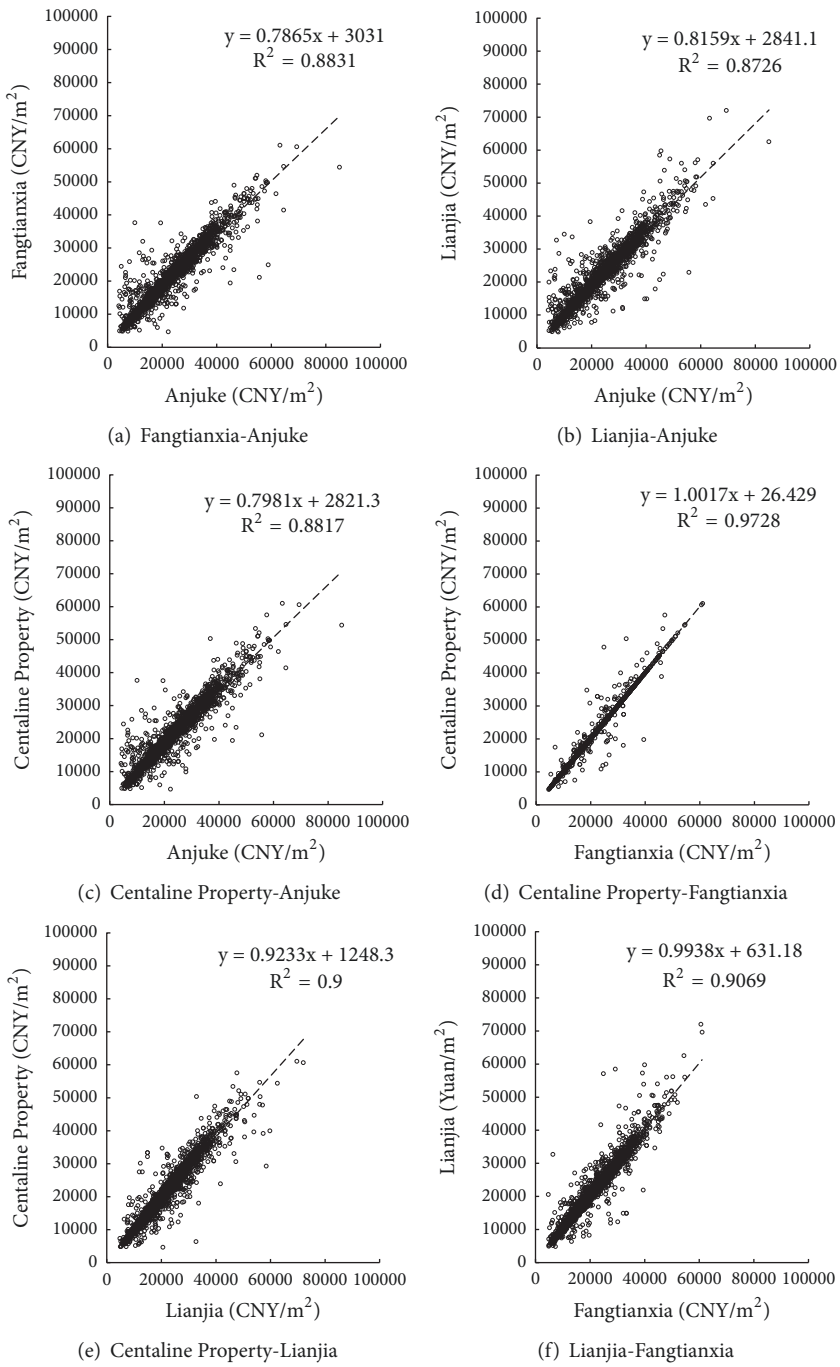
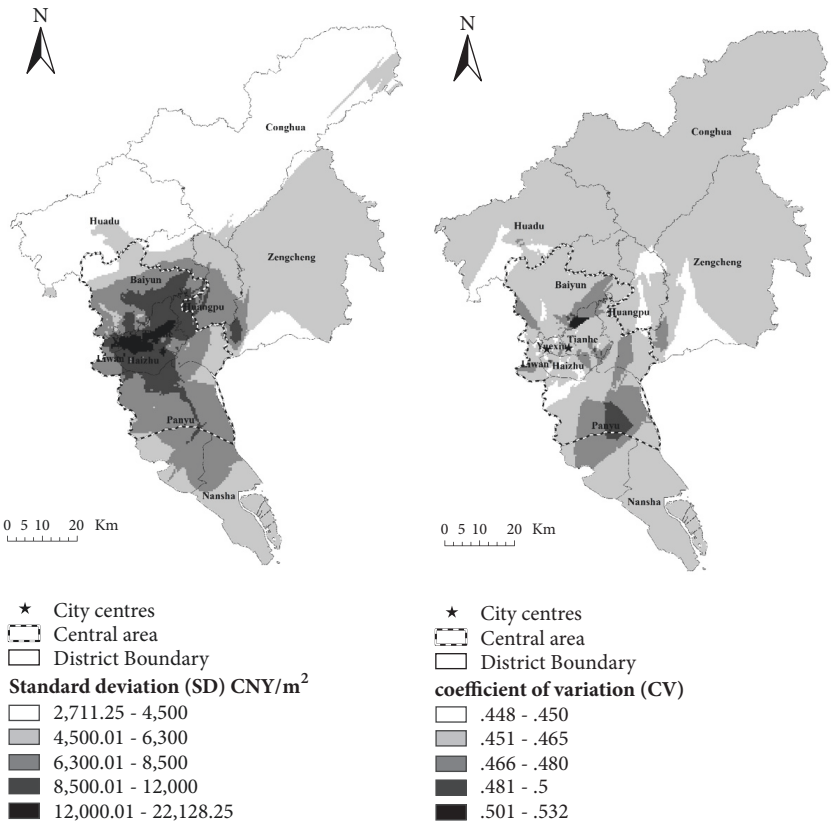


Figure 5: Scatter plots of neighbourhoods' second-hand housing prices (June 7, 2016).

are highlighted in the raw information on neighbourhoods from real estate agency websites, which somewhat affects the accuracy of the housing price models. Some flaws are revealed in the data on construction years for Centaline Property, because property age is proven to negatively correlate with property price [49] which contradicts results from the Centaline Property model. The distance to key school and park are supposed to demonstrate a negative correlation with housing price [43, 50], but this was not the case for results from the four models in the study. Moreover, the differences in the standardized coefficient for the same variables reflect the effects of using different firms' data in a single model. The performance of Fangtianxia model is better than that of other models, because its results match studies more closely. Thus, raw information on neighbourhoods from Fangtianxia are more reliable. But a process of appropriate data cleaning is still essential before we use raw information on neighbourhoods from real estate agency websites.



(a) Variation (b) The relative variation

Figure 6: Spatial interpolation of variation for corresponding second-hand neighbourhoods' housing prices on each website.

Table 5: Results and evaluation of hedonic housing price models.

		Year	Center	Subway	School	Park	Constant
Centaline Property	Coefficient	0.001	-0.253	-0.061	0.015	0.044	10.592
	Standardized Coefficient	0.010	-0.516	-0.150	0.034	0.064	
	Probability	0.236	<0.001	<0.001	0.002	<0.001	<0.001
	R^2: 0.369; adjusted R^2: 0.369; F-statistic: 1163.589; P-value<0.001						
Lianjia	Coefficient	0.016	-0.267	-0.045	0.009	0.043	-20.785
	Standardized Coefficient	0.253	-0.611	-0.102	0.023	0.068	
	Probability	<0.001	<0.001	<0.001	0.061	<0.001	<0.001
	R^2: 0. 451; adjusted R^2: 0. 451; F-statistic: 880.570; P-value<0.001						
Anjuke	Coefficient	0.007	-0.298	-0.059	0.010	0.050	-1.742
	Standardized Coefficient	0.119	-0.588	-0.138	0.022	0.068	
	Probability	<0.001	<0.001	<0.001	0.114	<0.001	0.165
	R^2: 0.449; adjusted R^2: 0. 449; F-statistic: 861.807; P-value<0.001						
Fangtianxia	Coefficient	0.020	-0.321	-0.045	-0.008	0.025	-12.403
	Standardized Coefficient	0.231	-0.678	-0.119	-0.019	0.039	
	Probability	<0.001	<0.001	<0.001	0.259	0.002	<0.001
	R^2: 0.530; adjusted R^2: 0.529; F-statistic: 832.682; P-value<0.001						

6. Conclusions

Housing prices on the Internet are not only a valuable data source for studies but are also commonly used by the public to track real estate market trends. Differences exist among housing prices released on various real estate agency websites, but few studies have compared such data or investigated how much effect differences will exert on relative housing price models. By comparing housing prices in Guangzhou released on official real estate information

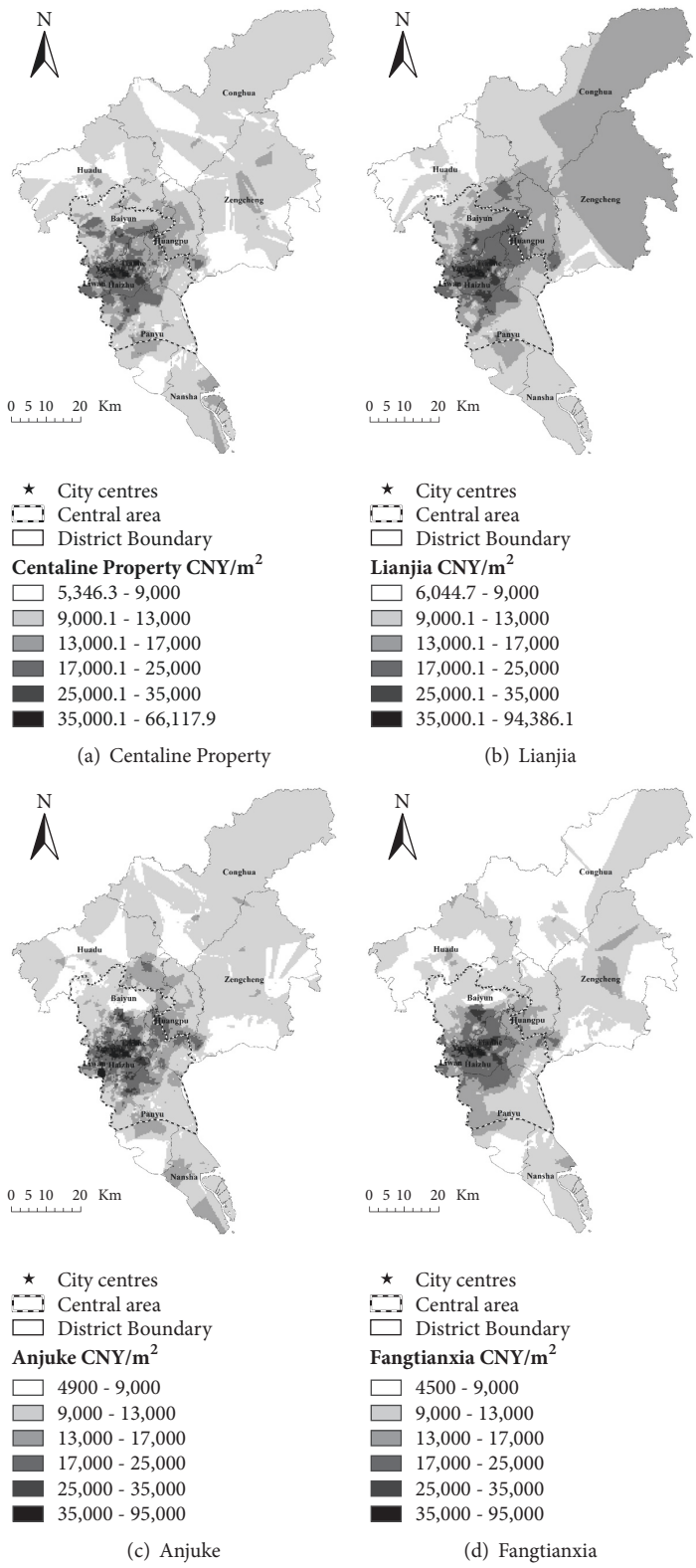


Figure 7: Spatial interpolation of second-hand residential housing prices in Guangzhou.

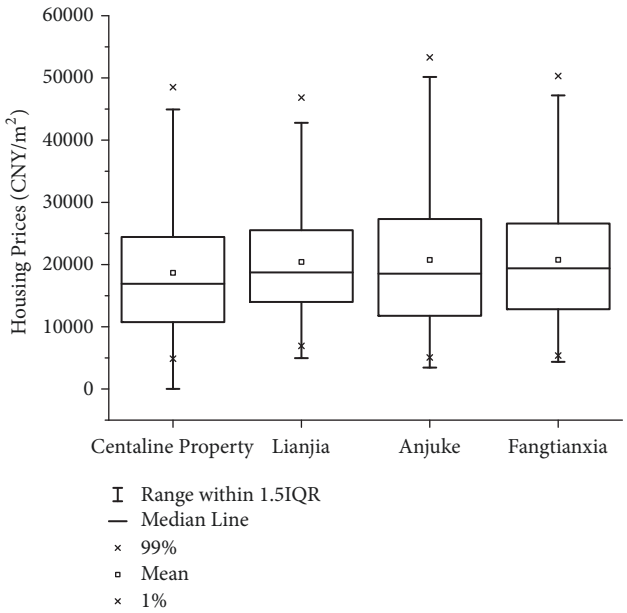


Figure 8: Statistical variations of second-hand neighbourhoods' housing prices on for-profit real estate information platforms.

platform Yangguang Jiayuan and on four for-profit real estate information platforms—namely, Centaline Property, Lianjia, Anjuke, and Fangtianxia—the key results from the analysis are as follows.

(1) The official second-hand residential housing prices in Guangzhou at city level and district level are generally lower than the housing prices issued on for-profit real estate information websites, whereas the overall trend for all types of housing price data is a rising one. Moreover, differences exist among the housing price data of various for-profit real estate information websites. In general, the volatility of district-level prices is higher than that of city-level prices. Data sources should be carefully selected in the study of city-level and district-level housing prices.

The city-level second-hand residential housing prices of Anjuke, Fangtianxia, and Lianjia display a high correlation with the official data. However, Centaline Property's second-hand residential housing prices at city and district level are relatively higher and more volatile than those in official data.

(2) Property prices for corresponding neighbourhoods are similar across the four for-profit real estate information websites as confirmed by cross-referencing. Spatial patterns of second-hand residential housing prices using Kriging interpolation of the four websites' data in Guangzhou are similar in the city centre area, but the differences in the peripheral areas are significant. Housing price variation decreases gradually from the city centre to the periphery, but the relative variation is stable.

(3) The results of the four hedonic models using neighbourhoods' raw information on Centaline Property, Lianjia, Anjuke, and Fangtianxia somewhat contradict other findings. In this study, the distance to the nearest city centre and subway station exhibited a negative relationship with second-hand neighbourhood housing prices, whereas the construction year exhibited a significantly positive relationship, consistent with other studies' findings. However, the distance to the nearest key school and park exerted a positive influence in most models, which was inconsistent with other research findings. Moreover, the differences in standardized coefficient for the same variable demonstrate the effects of different data resources. Fangtianxia model outperformed others. These contradictions and differences demonstrate that errors exist in raw information on neighbourhoods from the Internet, producing incorrect results.

Research shows that differences exist among second-hand residential housing prices of different Internet real estate information platforms at city, district, and neighbourhood levels. Raw information on neighbourhoods from the Internet may be erroneous. Thus, researchers should choose housing price data sources cautiously. Only with appropriate data cleaning can Internet-based information about neighbourhoods be used effectively in studies.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study is funded by the National Science Foundation of China [41601161].

References

[1] B.-S. Tang, S.-W. Wong, and S.-C. Lui, "Property agents, housing markets and housing services in transitional urban China," *Housing Studies*, vol. 21, no. 6, pp. 799–823, 2006.

[2] Y. Bar-Yam, "From big data to important information," *Complexity*, vol. 21, no. S2, pp. 73–98, 2016.

[3] E. C. M. Hui and C. Liang, "The spatial clustering investment behavior in housing markets," *Land Use Policy*, vol. 42, pp. 7–16, 2015.

[4] Y. Wu and Y. Li, "Impact of government intervention in the housing market: evidence from the housing purchase restriction policy in China," *Applied Economics*, vol. 50, no. 6, pp. 691–705, 2018.

[5] K. K. Fung and R. Forrest, "Institutional mediation, the Hong Kong residential housing market and the Asian Financial Crisis," *Housing Studies*, vol. 17, no. 2, pp. 189–207, 2002.

[6] G. Wong, "Has SARS infected the property market? Evidence from Hong Kong," *Journal of Urban Economics*, vol. 63, no. 1, pp. 74–95, 2008.

[7] E. C. M. Hui, J. W. Zhong, and K. H. Yu, "The impact of landscape views and storey levels on property prices," *Landscape and Urban Planning*, vol. 105, no. 1-2, pp. 86–93, 2012.

[8] H.-G. Kim, K.-C. Hung, and S. Y. Park, "Determinants of housing prices in Hong Kong: a box-cox quantile regression

approach," *The Journal of Real Estate Finance and Economics*, vol. 50, no. 2, pp. 270–287, 2015.

[9] M. H. C. Ho, "Liquidity and price differentials: evidence in the Hong Kong residential re-sale market," *Housing Studies*, vol. 18, no. 5, pp. 745–763, 2003.

[10] S. Li, L. Chen, and P. Zhao, "The impact of metro services on housing prices: a case study from Beijing," *Transportation*, vol. 5, pp. 1–27, 2017.

[11] Y. Xiao, Z. Li, and C. Webster, "Estimating the mediating effect of privately-supplied green space on the relationship between urban public green space and property value: Evidence from Shanghai, China," *Land Use Policy*, vol. 54, pp. 439–447, 2016.

[12] Y. Xiao, Y. Lu, Y. Guo, and Y. Yuan, "Estimating the willingness to pay for green space services in Shanghai: Implications for social equity in urban China," *Urban Forestry & Urban Greening*, vol. 26, pp. 95–103, 2017.

[13] E. C. M. Hui, C. K. Chau, L. Pun, and M. Y. Law, "Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix," *Building and Environment*, vol. 42, no. 6, pp. 2333–2343, 2007.

[14] H. Tsai, W.-J. Huang, and Y. Li, "The impact of tourism resources on tourism real estate value," *Asia Pacific Journal of Tourism Research*, vol. 21, no. 10, pp. 1114–1125, 2016.

[15] S. Li, X. Ye, J. Lee, J. Gong, and C. Qin, "Spatiotemporal analysis of housing prices in china: a big data perspective," *Applied Spatial Analysis and Policy*, vol. 10, no. 3, pp. 421–433, 2017.

[16] Y. Chen, X. Liu, X. Li, Y. Liu, and X. Xu, "Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning," *Applied Geography*, vol. 75, pp. 200–212, 2016.

[17] L. S. Ho, Y. Ma, and D. R. Haurin, "Domino effects within a housing market: The transmission of house price changes across quality tiers," *The Journal of Real Estate Finance and Economics*, vol. 37, no. 4, pp. 299–316, 2008.

[18] L. S. Ho and G. W.-C. Wong, "The first step on the housing ladder: A natural experiment in Hong Kong," *Journal of Housing Economics*, vol. 18, no. 1, pp. 59–67, 2009.

[19] W. M. Jayantha and J. M. Lau, "Buyers' property asset purchase decisions: an empirical study on the high-end residential property market in Hong Kong," *International Journal of Strategic Property Management*, vol. 20, no. 1, pp. 1–16, 2016.

[20] H. Cai, J. V. Henderson, and Q. Zhang, "China's land market auctions: Evidence of corruption?" *The RAND Journal of Economics*, vol. 44, no. 3, pp. 488–521, 2013.

[21] R. Tan, Y. Liu, K. Zhou, L. Jiao, and W. Tang, "A game-theory based agent-cellular model for use in urban growth simulation: A case study of the rapidly urbanizing Wuhan area of central China," *Computers, Environment and Urban Systems*, vol. 49, pp. 15–29, 2015.

[22] J. Wu, N. Ta, Y. Song, J. Lin, and Y. Chai, "Urban form breeds neighborhood vibrancy: a case study using a GPS-based activity survey in suburban Beijing," *Cities*, vol. 74, pp. 100–108, 2018.

[23] R. M. Kaplan, D. A. Chambers, and R. E. Glasgow, "Big data and large sample size: a cautionary note on the potential for bias," *Clinical and Translational Science*, vol. 7, no. 4, pp. 342–346, 2014.

[24] J. H. Mark and M. A. Goldberg, "Alternative housing price indices: an evaluation," *Real Estate Economics*, vol. 12, no. 1, pp. 30–49, 1984.

[25] H. O. Pollakowski, "Data sources for measuring house price," *Journal of Housing Research*, vol. 6, no. 3, pp. 377–387, 1995.

[26] S. C. Bourassa, M. Hoesli, and J. Sun, "A simple alternative house price index method," *Journal of Housing Economics*, vol. 15, no. 1, pp. 80–97, 2006.

[27] Y. M. Goh, G. Costello, and G. Schwann, "Accuracy and robustness of house price index methods," *Housing Studies*, vol. 27, no. 5, pp. 643–666, 2012.

[28] G. Meen, "Regional house prices and the ripple effect: A new interpretation," *Housing Studies*, vol. 14, no. 6, pp. 733–753, 1999.

[29] H. Zhang, "Discussion on present situation and developing tendency of real estate websites in China," *Sci-Tech Information Development and Economy*, vol. 6, pp. 69-70, 2001 (Chinese).

[30] S.-M. Li and Z. Yi, "Financing home purchase in China, with special reference to Guangzhou," *Housing Studies*, vol. 22, no. 3, pp. 409–425, 2007.

[31] F. Wu, "Polycentric urban development and land-use change in a transitional economy: the case of Guangzhou," *Environment and Planning A*, vol. 30, no. 6, pp. 1077–1100, 1998.

[32] L. Jiang and F. Wu, "A study on Guangzhou's employment subcentres and polycentricity," *Urban Planning Forum*, 2009 (Chinese).

[33] J. Liang, "The analysis of the user of the real estate website and suggestion," *China High-Tech Enterprises*, vol. 72, pp. 172–174, 2012 (Chinese).

[34] T. K. Koramaz and V. Dokmeci, "Spatial determinants of housing price values in istanbul," *European Planning Studies*, vol. 20, no. 7, pp. 1221–1237, 2012.

[35] R. A. Dubin, "Predicting house prices using multiple listings data," *The Journal of Real Estate Finance and Economics*, vol. 17, no. 1, pp. 35–59, 1998.

[36] L. Zhang, T. Leonard, and J. C. Murdoch, "Time and distance heterogeneity in the neighborhood spillover effects of foreclosed properties," *Housing Studies*, vol. 31, no. 2, pp. 133–148, 2016.

[37] P. Nilsson, "Are valuations of place-based amenities driven by scale?" *Housing Studies*, vol. 32, no. 4, pp. 449–469, 2017.

[38] K. J. Lancaster, "A new approach to consumer theory," *Journal of Political Economy*, vol. 74, no. 2, pp. 132–157, 1966.

[39] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[40] H. M. K. Mok, P. P. K. Chan, and Y.-S. Cho, "A hedonic price model for private properties in Hong Kong," *The Journal of Real Estate Finance and Economics*, vol. 10, no. 1, pp. 37–48, 1995.

[41] J. Chen and Q. Hao, "The impacts of distance to cbd on housing prices in shanghai: a hedonic analysis," *Journal of Chinese Economic and Business Studies*, vol. 6, no. 3, pp. 291–302, 2008.

[42] H. Wen, Y. Xiao, and L. Zhang, "School district, education quality, and housing price: Evidence from a natural experiment in Hangzhou, China," *Cities*, vol. 66, pp. 72–80, 2017.

[43] C. Wu, X. Ye, Q. Du, and P. Luo, "Spatial effects of accessibility to parks on housing prices in Shenzhen, China," *Habitat International*, vol. 63, pp. 45–54, 2017.

[44] J. A. Cao and R. Keivani, "The limits and potentials of the housing market enabling paradigm: an evaluation of China's housing policies from 1998 to 2011," *Housing Studies*, vol. 29, no. 1, pp. 44–68, 2014.

[45] J. Wu, Y. Deng, and H. Liu, "House price index construction in the nascent housing market: the case of china," *The Journal of Real Estate Finance and Economics*, vol. 48, no. 3, pp. 522–545, 2014.

[46] W. Sun, S. Zheng, D. M. Geltner, and R. Wang, "The housing market effects of local home purchase restrictions: evidence from Beijing," *The Journal of Real Estate Finance and Economics*, vol. 55, no. 3, pp. 288–312, 2017.

[47] L. Tan, W. Li, and W. Yu, "A GWR- based study on spatial pattern and structural determinants of Shanghai's housing price," *Economic Geography*, vol. 32, no. 2, pp. 52–58, 2012 (Chinese).

[48] H. Wen and Y. Tao, "Polycentric urban structure and housing price in the transitional China: Evidence from Hangzhou," *Habitat International*, vol. 46, pp. 138–146, 2015.

[49] J. Zietz, E. N. Zietz, and G. S. Sirmans, "Determinants of house prices: a quantile regression approach," *The Journal of Real Estate Finance and Economics*, vol. 37, no. 4, pp. 317–333, 2008.

[50] V. Sah, S. J. Conroy, and A. Narwold, "Estimating school proximity effects on housing prices: the importance of robust spatial controls in hedonic estimations," *The Journal of Real Estate Finance and Economics*, vol. 53, no. 1, pp. 50–76, 2016.

Advances in Operations Research
Advances in Decision Sciences
Journal of Applied Mathematics
The Scientific World Journal
Journal of Probability and Statistics
International Journal of Mathematics and Mathematical Sciences
Hindawi
Submit your manuscripts at
www.hindawi.com
Journal of Optimization
International Journal of Engineering Mathematics
International Journal of Analysis
Journal of Complex Analysis
Advances in Numerical Analysis
Mathematical Problems in Engineering
International Journal of Differential Equations
Discrete Dynamics in Nature and Society
International Journal of Stochastic Analysis
Journal of Mathematics
Journal of Function Spaces
Abstract and Applied Analysis
Advances in Mathematical Physics