

Complexity

# Computational Methods Applied to Data Analysis for Modeling Complex Real Estate Systems

Lead Guest Editor: Marco Locurcio

Guest Editors: Francesco Tajani and Pierluigi Morano





---

**Computational Methods Applied to Data  
Analysis for Modeling Complex Real Estate  
Systems**

Complexity

---

# **Computational Methods Applied to Data Analysis for Modeling Complex Real Estate Systems**

Lead Guest Editor: Marco Locurcio

Guest Editors: Francesco Tajani and Pierluigi  
Morano



---

Copyright © 2020 Hindawi Limited. All rights reserved.

This is a special issue published in "Complexity." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

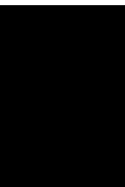
# Chief Editor

Hiroki Sayama, USA

## Editorial Board

Oveis Abedinia, Kazakhstan  
José Ángel Acosta, Spain  
Carlos Aguilar-Ibanez, Mexico  
Mojtaba Ahmadiéh Khanesar, United Kingdom  
Tarek Ahmed-Ali, France  
Alex Alexandridis, Greece  
Basil M. Al-Hadithi, Spain  
Juan A. Almendral, Spain  
Diego R. Amancio, Brazil  
David Arroyo, Spain  
Mohamed Boutayeb, France  
Átila Bueno, Brazil  
Arturo Buscarino, Italy  
Ning Cai, China  
Eric Campos, Mexico  
Émile J. L. Chappin, The Netherlands  
Yu-Wang Chen, United Kingdom  
Diyi Chen, China  
Giulio Cimini, Italy  
Danilo Comminiello, Italy  
Sergey Dashkovskiy, Germany  
Manlio De Domenico, Italy  
Pietro De Lellis, Italy  
Albert Diaz-Guilera, Spain  
Thach Ngoc Dinh, France  
Jordi Duch, Spain  
Marcio Eisenkraft, Brazil  
Mondher Farza, France  
Thierry Floquet, France  
José Manuel Galán, Spain  
Lucia Valentina Gambuzza, Italy  
Harish Garg, India  
Carlos Gershenson, Mexico  
Peter Giesl, United Kingdom  
Sergio Gómez, Spain  
Lingzhong Guo, United Kingdom  
Xianggui Guo, China  
Sigurdur F. Hafstein, Iceland  
Chittaranjan Hens, India  
Giacomo Innocenti, Italy  
Sarangapani Jagannathan, USA  
Mahdi Jalili, Australia  
Peng Ji, China

Jeffrey H. Johnson, United Kingdom  
Mohammad Hassan Khooban, Denmark  
Toshikazu Kuniya, Japan  
Vincent Labatut, France  
Lucas Lacasa, United Kingdom  
Guang Li, United Kingdom  
Qingdu Li, China  
Xinzhi Liu, Canada  
Chongyang Liu, China  
Xiaoping Liu, Canada  
Rosa M. Lopez Gutierrez, Mexico  
Vittorio Loreto, Italy  
Eulalia Martínez, Spain  
Marcelo Messias, Brazil  
Ana Meštrović, Croatia  
Ludovico Minati, Japan  
Saleh Mobayen, Iran  
Christopher P. Monterola, Philippines  
Marcin Mrugalski, Poland  
Roberto Natella, Italy  
Sing Kiong Nguang, New Zealand  
Irene Otero-Muras, Spain  
Yongping Pan, Singapore  
Daniela Paolotti, Italy  
Cornelio Posadas-Castillo, Mexico  
Mahardhika Pratama, Singapore  
Matilde Santos, Spain  
Michele Scarpiniti, Italy  
Enzo Pasquale Scilingo, Italy  
Dan Selișteanu, Romania  
Dehua Shen, China  
Dimitrios Stamovlasis, Greece  
Shahadat Uddin, Australia  
Gaetano Valenza, Italy  
Jose C. Valverde, Spain  
Alejandro F. Villaverde, Spain  
Dimitri Volchenkov, USA  
Christos Volos, Greece  
Zidong Wang, United Kingdom  
Qingling Wang, China  
Wenqin Wang, China  
Yan-Ling Wei, Singapore  
Honglei Xu, Australia  
Yong Xu, China



---

Xingang Yan, United Kingdom

Zhile Yang, China

Baris Yuce, United Kingdom

Massimiliano Zanin, Spain

Hassan Zargarzadeh, USA

Rongqing Zhang, China



Xianming Zhang, Australia

Xiaopeng Zhao, USA

Quanmin Zhu, United Kingdom


# Contents

## **Computational Methods Applied to Data Analysis for Modeling Complex Real Estate Systems**

Marco Locurcio , Francesco Tajani , and Pierluigi Morano



Editorial (3 pages), Article ID 8519060, Volume 2020 (2020)

## **Real Estate Asset Management Companies' Economies of Scale: Is It a Dream or Reality? The Italian Case**

Evita Allodi , Claudio Cacciamani, Michele Caliolo, Pier Paolo De Santis, Fabio Della Marra, and Simona Sanfelici

Research Article (9 pages), Article ID 8752865, Volume 2020 (2020)

## **GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China**

Shaopei Chen , Dachang Zhuang , and Huixia Zhang


Research Article (16 pages), Article ID 1079024, Volume 2020 (2020)

## **A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems**

José-Luis Alfaro-Navarro , Emilio L. Cano , Esteban Alfaro-Cortés , Noelia García , Matías Gámez , and Beatriz Larraz 


Research Article (12 pages), Article ID 5287263, Volume 2020 (2020)

## **Predicting Days on Market to Optimize Real Estate Sales Strategy**

Mauro Castelli , Maria Dobreva, Roberto Henriques, and Leonardo Vanneschi

Research Article (22 pages), Article ID 4603190, Volume 2020 (2020)

## **Efficiency of Chinese Real Estate Market Based on Complexity-Entropy Binary Causal Plane Method**

Yan Chen, Ya Cai, and Chengli Zheng 

Research Article (15 pages), Article ID 2791352, Volume 2020 (2020)

## **Grey Spectrum Analysis of Air Quality Index and Housing Price in Handan**

Kai Zhang, Yan Chen , and Lifeng Wu 

Research Article (6 pages), Article ID 8710138, Volume 2019 (2019)

## **Developing Statistical Optimization Models for Urban Competitiveness Index: Under the Boundaries of Econophysics Approach**

Cem Çağrı Dönmez  and Abdulkadir Atalan 

Research Article (11 pages), Article ID 4053970, Volume 2019 (2019)

## Editorial

# Computational Methods Applied to Data Analysis for Modeling Complex Real Estate Systems

**Marco Locurcio** <sup>1</sup>, **Francesco Tajani** <sup>2</sup>, and **Pierluigi Morano**<sup>1</sup>

<sup>1</sup>Department of Civil Engineering Sciences and Architecture, Polytechnic University of Bari, Via Orabona 4, Bari 70125, Italy

<sup>2</sup>Department of Architecture and Design, Sapienza University of Rome, Via Flaminia 359, Rome 00196, Italy

Correspondence should be addressed to Marco Locurcio; marco.locurcio@uniroma1.it

Received 17 June 2020; Accepted 18 June 2020; Published 9 July 2020

Copyright © 2020 Marco Locurcio 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.

Over the last few decades, as a result of the serious economic and financial crisis that has affected the USA and most European countries, there has been an increasing need for tools that provide reliable mass appraisals. The crisis was triggered by the drastic price reduction of properties as securities for credit exposures, characterized by values that, at the time of the sale due to the debtor's default, had revealed a real market price that made it impossible for the banks to recover the loaned capital.

This contingency has outlined the global connections of the real estate markets and has highlighted, on the one hand, the complex relationship between the real economy and property finance and, on the other hand, the need for multidisciplinary models—e.g., automated valuation ones—that are able to appropriately interpret the available data, identify space-time interactions, and forecast real estate cycles. The complexity of real estate systems concerns the numerous social, economic, and environmental implications that are related to property valuations and regional economic growth, as well as the reciprocal interdependencies between the territorial transformations and their socioeconomic factors.

In this framework, the aim of this Special Issue has been to collate both original research and review articles that contribute to the development of new tools for modeling, optimizing, and simulating complex real estate systems and are related to the applications of data analysis models that take into account the continuous changes of the economic boundary conditions and are able to automatically capture the causal relationships among the variables involved as well as predict property values in the short term.

The main topics of the Special Issue include the following: (i) mass appraisal methods applied to the interpretation of the real estate markets; (ii) multicriteria decision systems as support for valuations in uncertain contexts; (iii) big data analysis for modeling and control approaches; (iv) econometric analysis for the forecasting of real estate trends; (v) GIS-based systems for the identification of spatial correlations among real estate factors; (vi) artificial intelligence implemented for the automated valuation model; (vii) genetic algorithms for the investigation of complex real estate systems.

A total of twenty-one papers were submitted. Following a rigorous procedure of peer review, only seven papers were accepted and published. The different countries of the authors' affiliation (China, Turkey, Portugal, Spain, Italy, and so on) had given the Special Issue a strongly international character.

The paper by S. Chen et al. entitled “GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China” describes the housing markets of the city of Guangzhou (China). Through an empirical analysis with the using of traditional regressive models and GIS-based spatial autocorrelation tools, the authors characterize the spatial homogeneity and nonstationarity of the housing prices in the period 2009–2015 with reference to the context of the neighborhoods in Guangzhou. The average annual housing price (AAHP) is the dependent variable, as a function of twelve explanatory variables related to the geographical location



condition, the transportation accessibility, the commercial service intensity, and the public service intensity. The results outline that (1) the temporal and spatial evolution of the AAHP in Guangzhou shows the circle characteristic with the center of the urban core; (2) there are spatial differences in the growth of AAHP in Guangzhou, which is closely related to the urban planning and the spatial pattern of the urban functional area; (3) the global spatial autocorrelation analysis reveals that the housing price has significant spatial aggregation; (4) the analysis of the linear regression model illustrates the role of the urban infrastructure; (5) the factor of geographical location presents the extreme significant impact on the housing price; (6) the analysis based on the geographically weighted regression model points out the specific effect of each factor on the spatial heterogeneity and nonstationarity of the housing price. The outputs of this study could be a valid support for governing policies and spatial controlling mechanisms of housing markets.

The paper by Y. Chen et al. entitled “Efficiency of Chinese Real Estate Market Based on Complexity-Entropy Binary Causal Plane Method” analyzes the mechanism of house price formation, assuming that the price is a function of the rental income and the average annual household income. The authors implement the complexity-entropy binary causal plane method to detect the hidden structure of real estate market price and then measure its efficiency and complexity. By comparing the applications to Chinese and American contexts, some considerations are derived to define possible guidelines for reducing the information asymmetry in the market and elaborating more detailed rules to help information disclosure in the process of real estate transactions, so as to guarantee the liquidity, accuracy, and timeliness of the information data.

The paper by C. Ç. Dönmez and A. Atalan entitled “Developing Statistical Optimization Models for Urban Competitiveness Index: Under the Boundaries of Econophysics Approach” aims to elaborate a specific urban competitiveness index (UCI) by using the statistical optimization method for the econophysics approach. UCI is defined as a combination of the gross domestic product of urban and the gross domestic product per capita of urban with other factors (education, health, and training, labor and transport, technology and industry, market size, product efficiency, and financial service). The complexity of the developed model is attested by the number of indicators that represent the considered factors: totally, thirty-eight indicators have been specified, useful to characterize the complex structure of the urban areas. Borrowing the goal programming logic, the proposed model is applied to thirty cities located in fifteen countries worldwide. The outputs allow identifying the most attractive cities for institutional investors, by considering possible variations of the different factors.

The paper by M. Castelli et al. entitled “Predicting Days on Market to Optimize Real Estate Sales Strategy” develops a model for limiting the problem of irregularities and frauds that are frequent in the real estate market. Starting from the real estate market in Bulgaria, the authors outline that agencies frequently advertise unreal or unavailable

apartment listings for a cheap price, as a method to attract unaware potential new customers. According to the authors, the absence of rigorous laws that regulate the real estate market and the duplication of the real estate listings on different advertising portals have increased the need of appropriate transparency. Therefore, the authors present a systematical approach based on data analysis techniques and machine learning methods aimed at identifying frauds in real estate advertisements and improving the transparency of the property listings.

The paper by J.-L. Alfaro-Navarro et al. entitled “A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems” proposes an Automated Valuation Model (AVM) to carry out the assessment of real estate prices for an entire country. The authors point out the growing importance of real estate valuations, also highlighted by the close relationship among the collateral value and the bank stability. Starting from a database of about 800,000 properties located in 433 cities in Spain, the AVM is differentiated (based on bagging, boosting, and random forest) for each municipality. The outputs outline that the ensemble methods usually provide good prediction results, although they tend to sacrifice the possible interpretation of the relationships between the predictor variables and the target.

The paper by K. Zhang et al. entitled “Grey Spectrum Analysis of Air Quality Index and Housing Price in Handan” assesses the relationship between an air quality index (AQI) and the housing prices in order to define appropriate guidelines for government programs aimed at reducing the air pollution and regulating the property market. The temporal analysis, based on the grey system theory, is referred to the city of Handan (China) in southern Hebei Province. The application shows that there is a negative correlation between AQI and the housing price; furthermore, due to the specific Chinese policies, the housing prices are gradually stabilizing and the air quality tends to improve.

The paper by E. Allodi et al. entitled “Real Estate Asset Management Companies’ Economies of Scale: Is It a Dream or Reality? The Italian Case” aims to verify the presence (or absence) of economies of scale of Italian real estate management companies. The authors highlight that operations of mergers and acquisitions have consolidated the real estate asset management companies, with a consequent increase in the company size. Through a multivariate analysis, the study concerns twenty-six asset management companies, characterized by a total asset under management (AUM) equal to seventy billion euros, i.e., corresponding to 85% of the total AUM managed by all real estate asset management companies operating in Italy. The implementation of a series of multivariate regressions outlines the absence of relationships that would suggest economies of scale.

All the papers submitted in this Special Issue have demonstrated the current, widespread, and increasing interest of the scientific and professional operators for the implementation of innovative mass appraisal models. In particular, the Special Issue is part of the broader international debate on the contribution that data mining has on the global economy, in accordance with the guidelines of the

European Coordinated Plan on Artificial Intelligence, and how this typology of models can support to interpret the complexity of real estate dynamics.

### **Conflicts of Interest**

The lead guest editor and the guest editors declare that there are no conflicts of interest or agreements with private companies, which will prevent them working impartially in the editorial process.

### **Acknowledgments**

The editors thank all the authors for their contributions.

*Marco Locurcio  
Francesco Tajani  
Pierluigi Morano*

## Research Article

# Real Estate Asset Management Companies' Economies of Scale: Is It a Dream or Reality? The Italian Case

**Evita Allodi** <sup>1</sup>, **Claudio Cacciamani**<sup>1</sup>, **Michele Caliolo**<sup>2</sup>, **Pier Paolo De Santis**<sup>2</sup>,  
**Fabio Della Marra**<sup>1</sup>, and **Simona Sanfelici**<sup>1</sup>

<sup>1</sup>Department of Economic and Business Sciences, University of Parma, Parma, 43125, Italy

<sup>2</sup>Mazars, Milan, 20154, Italy

Correspondence should be addressed to Evita Allodi; [evita.allodi@unipr.it](mailto:evita.allodi@unipr.it)

Received 2 January 2020; Revised 5 April 2020; Accepted 27 April 2020; Published 27 May 2020

Guest Editor: Marco Locurcio

Copyright © 2020 Evita Allodi 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.

The research focuses on a sample of 26 Italian real estate asset management companies (Società di Gestione del Risparmio “SGR”)—whose asset management is totally linked to real estate funds—that considers a period of six years (2013–2018). Using some variables extrapolated from the internal accountability of each SGR, the analysis investigates possible relationships between them to verify the presence or absence of economies of scale of Italian real estate management companies by multivariate regressions. The results show that there is no single model for profit maximization and cost minimization, but all depends on the business model that each SGR decides to adopt.

## 1. Introduction

Both researchers and professionals agree that numerous and significant changes affected the real estate industry over the last twenty years [1–7]. The uncertainty about the trend in the real estate industry in the near future is high and is expected to increase [8]. Anyway, with quantitative easing and negative interest rates in Europe, attention in real estate asset class is more and more increasing.

Finance is gaining progressively more importance in every macroeconomic industry, in general, and in the real estate market, in particular [1, 9–11].

The main role of finance is to raise funds at the lowest cost. Real estate is a “capital intensive” industry and needs a lot of capital, from the smallest development operations to the largest investments. Finance is also useful in investment analysis; real estate represents an important asset class in both institutional and private portfolios. Analyzing the real estate investment with a financial approach is therefore fundamental. In fact, more and more space has been given to the evaluation of all economic, income, equity, and financial factors affecting the management of real estate portfolios

with a consequent impact on the portfolio ability to produce profit. These quantitative factors became as significant as the qualitative (material and architectural) characteristics of real estate [12].

On the other side, the number of companies offering services in the real estate market increased, resulting in higher competition both nationally and internationally [1]. All this led to greater exposure of national operators to threats and opportunities with consequent competitive challenges [13–15]. In this scenario, the real estate industry faced several changes, increasingly dealing with new logics and stakeholders, with ever-changing organizational and structural dynamics. This continuous transformation implied more and more interest in the organizational models of the companies providing asset management services (in Italy, the so-called “Società di Gestione del Risparmio”—SGR).

This interest is also due, on the one hand, to the need to identify more efficient production assets (in a context of reinvigorated requests for the protection of investors in financial instruments exposed to stock market turbulence) and, on the other hand, to the increase of demand for improving the quality of the offered services [16].

These structural and organizational changes led over the years to mergers, acquisitions, and consolidation of real estate asset management companies, with a consequent increase in company size. Economies of scale are the cost advantages that companies gain from their scale of operation (typically measured by the amount of output produced), with cost per unit of output decreasing with increasing scale.

By transferring this notion to the real estate asset management industry, it can be assumed that as real estate portfolios increase in size, the incremental cost of managing additional properties should fall [17, 18]. So real estate asset management companies with larger property portfolios should be more efficient than those with smaller portfolios [19].

While most previous studies focused on the economies of scale in US Real Estate Investment Trusts (REITs), this analysis aims to show, starting from the financial statement data, that there are no specific economies of scale in Italian real estate management companies.

This study briefly summarizes the literature on the subject and the issues arising on the matter. Then, considering some empirical research variables, it tries to find some relationships between them to verify the presence or absence of economies of scale of Italian real estate management companies using multivariate regressions. The results show that there is no single model for profit maximization and cost minimization, but all depends on the business model that each SGR decides to adopt.

## 2. Related Literature

The concept “economies of scale” means that efficiency in production and operations increases with size. Historically, firms in various industries often expand not only to consolidate their power, but also to capture these efficiency gains [17].

In the real estate market, this type of study started to be developed mainly from American REITs.

The first studies, dating back to the 70s and 80s, tried to find economies of scale. Anyway, it resulted difficult to identify and measure economies of scale due to the different technologies available and the lack of data [17]. In fact, early studies suggested a “small firm effect” in American REITs: smaller companies earned higher average returns than larger companies [20].

At the beginning of the 90s, researchers found some lower costs coming from the increase in the size of REITs.

Linneman [21] notes that large REITs can achieve greater shareholder value via economies of scale with respect to costs.

Bers and Springer [22] show that economies of scale exist for REITs and also that economies of scale differ chronologically. The same researchers [22–24] and others [25], using the standard approach of estimating the cost function without allowing for the possibility of inefficient production, find evidence of economies of scale for REITs. Capozza and Seguin [26] find that only general and administrative costs exhibit substantial economies of scale. Latzko [27] find economies of scale in managing mutual funds.

Ambrose et al. [25] report that large REITs have higher net operating income (NOI) growth, but they also show that this ratio is weak. In the same year, Capozza and Seguin [28], contradicting what was reported only two years earlier, find evidence of diseconomies of scale. In their study, Anderson et al. [29] measure technical efficiency and economies of scale for REITs by employing data envelopment analysis (DEA), a linear programming technique. They find that REITs are technically inefficient. In particular, inefficiencies are a result of both poor input use and failure to operate at constant returns to scale. The dynamics emerged for the American REITs have recently been tested in the European real estate market by Ambrose et al. [19], investigating the effect of firm size on expense, revenue, return, and capital cost for European real estate companies and comparing the size effects of REITs and non-REIT real estate companies. They found that larger real estate companies are able to generate higher revenue per unit of company size, incur lower costs, and produce higher returns. Moreover, NOI ratios and return ratios increase while selling, general and administrative cost ratios decrease with the size of a company.

The question of the existence of economies of scale in real estate companies still remains important [19]. Even if there is not yet an answer for this question, disagreements about the concept of scale economies in real estate asset management continue to exist.

The uncertainty on the possible existence of economies of scale in the real estate industry is due to multiple factors. First factor being the market, which is characterized by cycles of expansion and of contraction influencing its trend. Secondly, it is difficult to quantify economies of scale: it is easy to identify the presence or absence of economies of scale in industries where production is quantified and measured. In most manufacturing industries, the final output is always an equal product generated by clear and defined input processes. In the analysis of real estate management companies, not only the final output is not easy to identify, but it is not even homogeneous. Third, the assumption underlying in the studies on economies of scale is that all the companies of the sample under investigation can benefit of the same frontier of costs and use the same technology [17, 30]: reality is quite different. Fourth, young, growing, and expanding businesses need time to achieve the right size to ensure economies of scale. No research directly addresses the existence of scale economies in Italian real estate asset management companies: SGR.

There are very few studies focused on Italian asset management companies in general [16, 31, 32] in which an analysis is carried out considering balance sheet multiples [31] or company X-efficiencies measured to estimate a possible efficient frontier [16, 32]. Almost absent are the research studies referring to the Italian real estate management companies: only Abate [33] and Giannotti and Mattarocci [34]. Both these papers show that the Italian real estate companies represent a sector that, after a few years of constant and high growth, did not reach the typical maturity of the asset management industry.

The possible reason is, on the one hand, the difficulty in identifying a univocal and truthful representation of the

production process of the asset management industry in the real estate industry and, on the other, the total lack of data. In fact, SGRs (companies that mainly manage real estate investment funds—closed funds) are not obliged to be listed on the stock exchange.

This analysis aims to demonstrate, starting from the financial statement data of SGR, that there are no economies of scale for the Italian real estate management companies.

### 3. The Sample

The research focuses on a sample of 26 asset management companies (25 (the number of SGRs decreases by one unit because two SGRs have merged) in the years 2017 and 2018) whose asset management is totally based on real estate funds. With reference to the number of SGRs active in Italy, the sample constitutes the 96.30% (it was not possible to analyse 100% of the SGRs operating in Italy because one of them is in extraordinary administration) of the population; consequently, the database is almost representative of reality. The assets managed by the companies included in the sample, as of 31 December 2018, amounted to 70 billion euros, corresponding to 85% of the total AUM managed by all real estate asset management companies operating in Italy, calculated by the Bank of Italy at 82 billion euros [35].

The variables taken into account were extrapolated from the SGRs' confidential internal accountability which they provided to the research team.

The sample, considering a period of six years (2013–2018), is divided into three different clusters, identified on the basis of the assets under management (AUM) managed by each company included in the sample. The three clusters were identified by setting thresholds based on the AUMs of each SGR as follows:

- (i) Cluster 1: average AUM > 5 billion euros
- (ii) Cluster 2: 2 billion euros ≤ average AUM ≤ 5 billion euros
- (iii) Cluster 3: average AUM < 2 billion euros

Obviously, from one year to the next, SGRs may change clusters. The variables describing each of the 26 SGRs are the following five:

- (1) Net fees: they represent the amount of the budget line no. 30 of the SGRs' income statement scheme, according to the Instruction of Bank of Italy named "Il bilancio degli intermediari IFRS diversi dagli intermediari bancari"
- (2) Asset under management: this variable is equal to the sum of the total assets of the all real estate alternative investment funds managed by every SGR of the sample, at the end of the year
- (3) Administrative costs: they are the amount of the line no. 120 of the SGRs' income statement scheme, according to the Instruction of Bank of Italy named "Il bilancio degli intermediari IFRS diversi dagli intermediari bancari"

- (4) No. of real estate funds managed: this variable is equal to the number of the real estate alternative investment funds managed at the end of the year, as reported in the explanatory note of the SGRs' annual financial statements
- (5) Average number of employees: this variable is equal to the average number of employees of the year, as reported in the explanatory note of the SGRs' annual financial statements

In order to make the variables comparable, they have been standardised, i.e., the following transformation has been applied for each variable  $x$ :

$$x = \frac{x - E(x)}{\sigma_x}, \quad (1)$$

where  $E(\cdot)$  identifies the operator mean value and  $\sigma$  standard deviation. The purpose of this transformation is to reduce all variables to the same order of size.

As reported in Table 1, all variables are highly correlated. In particular, the following pairs of variables show a correlation of more than 90%: (net commissions, assets under management), (administrative costs, net commissions), (average number of employees, net commissions), and (administrative costs, average number of employees).

### 4. Empirical Results

Two multivariate regressions are performed, as other dependences did not prove to be significant at the 10% level. The first one, as reported in Table 2, considers the asset under management as a dependent variable and the number of real estate funds managed and the average number of employees for each SGR as independent variables.

The second multivariate regression, according to Table 3, considers net commissions as a dependent variable and the number of real estate funds managed, personnel expenses, and AUM as independent variables.

Each regression is performed with reference to two scenarios. The first scenario is represented by the entire data sample, without any distinction between different SGRs, while the second considers the splitting of the sample into clusters, depending on the AUM. Dummy variables are added to perform the latter analysis.

Before analysing the first regression, it should be considered the path followed by SGRs in terms of AUM, number of funds, and percentage of administrative costs in terms of AUM. As reported below, Table 4 represents the amount of asset under management for each SGR, Table 5 represents the number of real estate funds for each SGR, and Table 6 represents the percentage of administrative costs in terms of AUM. Each table refers to the period from 2013 to 2018.

Over the six-year period considered, an increase in investments made in Italy through alternative real estate investment funds is observed. Indeed, the AUM analysed in the sample shows a compound annual growth rate (CAGR) of 7.21%. This increase can be expressed in absolute value as +6.6 billion euro in the last year (2018). Considering the

TABLE 1: Correlation matrix for each year.

	Net fees	AUM	Administrative costs	No. of real estate funds	No. of employees
2013					
Net fees	1.0000000	0.9366408	0.9203043	0.7812385	0.7872560
AUM	0.9366408	1.0000000	0.8572549	0.7694823	0.8768293
Administrative costs	0.9203043	0.8572549	1.0000000	0.8867641	0.7720600
No. of real estate funds	0.7812385	0.7694823	0.8867641	1.0000000	0.6830251
No. of employees	0.7872560	0.8768293	0.7720600	0.6830251	1.0000000
2014					
Net fees	1.0000000	0.9176423	0.9083548	0.7871646	0.8014534
AUM	0.9366408	1.0000000	0.8781287	0.8051264	0.8657633
Administrative costs	0.9203043	0.8781287	1.0000000	0.9056471	0.7427939
No. of real estate funds	0.7812385	0.8051264	0.9056471	1.0000000	0.6758336
No. of employees	0.7872560	0.8657633	0.7427939	0.6758336	1.0000000
2015					
Net fees	1.0000000	0.9477420	0.9496365	0.7399840	0.8903021
AUM	0.9477420	1.0000000	0.9168684	0.7787273	0.8643887
Administrative costs	0.9496365	0.9168684	1.0000000	0.7687266	0.9463153
No. of real estate funds	0.7399840	0.7787273	0.7687266	1.0000000	0.6741669
No. of employees	0.8903021	0.8643887	0.9463153	0.6741669	1.0000000
2016					
Net fees	1.0000000	0.9465023	0.9509846	0.6359986	0.9216262
AUM	0.9465023	1.0000000	0.8819332	0.7457024	0.8260694
Administrative costs	0.9509846	0.8819332	1.0000000	0.5626679	0.9609642
No. of real estate funds	0.6359986	0.7457024	0.5626679	1.0000000	0.5183680
No. of employees	0.9216262	0.8260694	0.9609642	0.5183680	1.0000000
2017					
Net fees	1.0000000	0.8641211	0.9738102	0.5489571	0.9383796
AUM	0.8641211	1.0000000	0.8544921	0.7638472	0.7778732
Administrative costs	0.9738102	0.8544921	1.0000000	0.5696372	0.9550878
No. of real estate funds	0.5489571	0.7638472	0.5696372	1.0000000	0.5026150
No. of employees	0.9383796	0.7778732	0.9550878	0.5026150	1.0000000
2018					
Net fees	1.0000000	0.8886413	0.9628419	0.5543713	0.9310833
AUM	0.8886413	1.0000000	0.9130547	0.7538074	0.8420773
Administrative costs	0.9628419	0.9130547	1.0000000	0.6393171	0.9541886
No. of real estate funds	0.5543713	0.7538074	0.6393171	1.0000000	0.5685592
No. of employees	0.9310833	0.8420773	0.9541886	0.5685592	1.0000000

TABLE 2: First multivariate regression.

Response	Y	Asset under management
Regressors	$X_1$	Number of real estate funds managed
Regressors	$X_2$	Average number of employees

TABLE 3: Second multivariate regression.

Response	Y	Net fees
Regressors	$X_1$	Number of real estate funds managed
Regressors	$X_2$	Administrative costs
Regressors	$X_3$	Asset under management

entire reference period, this deviation amounts to approximately +24 billion euro (+51.8%).

At the same time, there is an increase in the number of funds (+181 alternative real estate investment funds between 2013 and 2018; +73%). Nevertheless, this growth stopped over the past year. This is illustrated by the smaller increase in the number of funds in 2018: the lowest in the last 6 years

with an absolute value of +17 and a growth rate of +4.11%. On the contrary, the AUM increased by +10.4% in 2018.

The dynamic of administrative costs is related to the amount of AUM and to the number of real estate funds. More considerations about this variable are made later.

Launching the first regression, in all six years considered, we found that the intercept is not significant at the 10% level. By repeating the regression, we obtain the results reproduced in Table 7, where only significant regressors at 0.1% and 1% level are reported.

In the analysed period (2013–2018), the AUM always depends positively on the average number of employees and the number of funds managed. This empirical evidence shows that by increasing the number of employees by one unit compared to their average number during the year, a higher marginal effect is achieved by a unit increase in the number of funds. This can be interpreted as follows: by increasing the number of employees, the asset management company has more resources to employ in setting up and subsequently managing a new real estate fund and, consequently, new assets from which having fees of management.

TABLE 4: Amount of asset under management for each SGR from 2013 to 2018.

SGR	AUM (bn €)					
	2018	2017	2016	2015	2014	2013
1	9.396	9.488	8.593	7.867	8.983	9.179
2	7.321	7.521	6.890	6.769	7.074	3.882
3	4.880	5.060	4.592	5.567	5.505	4.515
4	7.040	5.070	5.351	5.508	5.421	5.498
5	5.453	<b>4.870</b>	<b>3.829</b>	<b>3.270</b>	<b>2.935</b>	<b>2.725</b>
6	<b>4.400</b>	<b>3.551</b>	<b>3.796</b>	<b>3.583</b>	<b>3.243</b>	<b>3.349</b>
7	<b>3.770</b>	<b>3.361</b>	<b>3.220</b>	<b>3.562</b>	<b>3.404</b>	<b>2.272</b>
8	<b>3.955</b>	<b>2.810</b>	<b>1.780</b>	<b>500</b>	<b>307</b>	<b>40</b>
9	<b>3.142</b>	<b>2.300</b>	<b>1.663</b>	<b>1.326</b>	<b>840</b>	<b>520</b>
10	<b>2.503</b>	<b>2.198</b>	<b>2.065</b>	<b>1.523</b>	<b>1.014</b>	<b>541</b>
11	<b>2.352</b>	<b>2.127</b>	<b>1.967</b>	<b>1.587</b>	<b>1.410</b>	<b>1.360</b>
12	<b>1.394</b>	<b>1.644</b>	<b>1.494</b>	<b>1.591</b>	<b>1.380</b>	<b>1.231</b>
13	<b>933</b>	<b>1.266</b>	<b>1.392</b>	<b>1.411</b>	<b>1.469</b>	<b>1.482</b>
14	<b>1.053</b>	<b>1.033</b>	<b>1.120</b>	<b>1.151</b>	<b>843</b>	<b>872</b>
15	—	—	<b>1.135</b>	<b>1.139</b>	<b>850</b>	<b>435</b>
16	<b>1.518</b>	<b>1.179</b>	<b>1.036</b>	<b>931</b>	<b>820</b>	<b>1.036</b>
17	<b>1.280</b>	<b>1.173</b>	<b>1.120</b>	<b>919</b>	<b>1.007</b>	<b>1.003</b>
18	<b>1.958</b>	<b>1.411</b>	<b>1.411</b>	<b>773</b>	<b>709</b>	<b>497</b>
19	<b>1.235</b>	<b>1.215</b>	<b>804</b>	<b>738</b>	<b>659</b>	<b>620</b>
20	<b>1.152</b>	<b>938</b>	<b>653</b>	<b>592</b>	<b>540</b>	<b>515</b>
21	<b>1.137</b>	<b>985</b>	<b>943</b>	<b>490</b>	<b>501</b>	<b>509</b>
22	<b>1.018</b>	<b>826</b>	<b>814</b>	<b>910</b>	<b>756</b>	<b>682</b>
23	<b>27</b>	<b>29</b>	<b>259</b>	<b>299</b>	<b>276</b>	<b>356</b>
24	<b>48</b>	<b>108</b>	<b>171</b>	<b>268</b>	<b>334</b>	<b>390</b>
25	<b>92</b>	<b>100</b>	<b>107</b>	<b>229</b>	<b>236</b>	<b>236</b>
26	<b>473</b>	<b>273</b>	<b>117</b>	<b>147</b>	<b>258</b>	<b>311</b>
Tot.	67.530	60.536	56.322	52.650	50.774	44.056

Values in italics represent cluster 1, values in bold represent cluster 2, and values in bold italics represent cluster 3.

On the other side, increasing the number of funds by one unit does not necessarily mean a substantial increase in the SGR's AUM, as there is no minimum quantum leap for setting up a fund. Such a fund could be small in terms of size, with a marginal impact on the SGR's overall AUM. However, such an increase in assets would inevitably involve the use of resources and therefore costs.

The absence of the intercept implies that if the regressors are equal to their average, as a consequence, the AUM is not significantly different from its average.

Multivariate regression is later repeated on clustered data. The dummy explanatory variables *ind\_cl1*, *ind\_cl2* and *ind\_cl3* indicate membership in the different AUM-based clusters.

More precisely, the dummy variables split the sample into three clusters, as shown below:

- (i) *Ind\_cl1* relates to the first cluster, which contains SGRs whose AUM is greater than 5 billion euros
- (ii) *Ind\_cl2* relates to the second cluster, which contains SGRs whose AUM is bounded between 2 and 5 billion euros
- (iii) *Ind\_cl3* relates to the third cluster, which contains SGRs whose AUM is less than 2 billion euros

TABLE 5: Number of real estate funds for each SGR from 2013 to 2018.

SGR	Number of real estate funds managed					
	2018	2017	2016	2015	2014	2013
1	47	43	41	37	36	32
2	44	42	35	34	33	29
3	27	27	27	27	26	21
4	15	8	7	7	7	7
5	22	22	20	17	12	11
6	31	32	33	30	28	25
7	14	12	12	14	13	12
8	34	23	14	5	2	2
9	7	5	5	4	2	1
10	27	25	18	11	7	4
11	36	32	30	17	13	12
12	16	17	17	9	9	8
13	12	14	14	12	10	9
14	6	5	5	3	3	3
15	—	—	6	6	5	3
16	19	19	19	19	18	15
17	12	12	9	8	8	8
18	6	6	3	3	3	3
19	22	21	20	18	18	15
20	5	5	4	4	4	3
21	4	3	2	2	2	2
22	12	10	8	9	9	4
23	2	7	7	10	9	9
24	1	2	2	2	3	3
25	5	4	3	4	4	4
26	4	4	3	2	2	4
Tot.	430	400	364	314	286	249

In this study, to avoid multicollinearity, the dummy variable *Ind\_cl3* is not explicitly considered in the regressions.

We now want to test whether membership to different clusters is one of the qualitative variables relevant to the regression. The intercept would be the constant term for the base group with lowest AUM, while for members of the first and second AUM-based groups the constant term would be the intercept plus the coefficient of the membership dummy. Again, the results obtained for each regression are reported after removing any insignificant regressors in Table 8.

For the time interval considered, all variables are at least 5% significant. In addition, all variables, except the intercept, show positive coefficients. As all the variables are standardised, the negative intercept implies that when all the regressors are zero (i.e., for the lowest AUM group), the expected AUM is obviously lower than the average value.

The regression shows that with obvious differences among clusters, while in 2013 and 2018 the number of employees is the only explanatory variable that influences the AUM, for the period 2014–2017 the number of funds managed is the only one having effects on the AUM. Nevertheless, it is noted that the regressor that identifies the number of funds impacts more modestly than the case in which the data are not clustered.

TABLE 6: Percentage of administrative costs in terms of AUM.

SGR	Administrative costs in % of AUM					
	2018	2017	2016	2015	2014	2013
1	0.000271%	0.000273%	0.000288%	0.000331%	0.000319%	0.000292%
2	0.000277%	0.000269%	0.000267%	0.000269%	0.000323%	0.000464%
3	0.000305%	0.000287%	0.000309%	0.000219%	0.000216%	0.000252%
4	0.000503%	0.000657%	0.000573%	0.000364%	0.000109%	0.000129%
5	0.000311%	0.000327%	0.000374%	0.000355%	0.000202%	0.000157%
6	0.000268%	0.000315%	0.000294%	0.000343%	0.000372%	0.000358%
7	0.000182%	0.000195%	0.000203%	0.000196%	0.000203%	0.000340%
8	0.000184%	0.000186%	0.000220%	0.000336%	0.000437%	0.004489%
9	0.000194%	0.000253%	0.000242%	0.000184%	0.000290%	0.000260%
10	0.000255%	0.000267%	0.000253%	0.000291%	0.000378%	0.000782%
11	0.000347%	0.000358%	0.000342%	0.000337%	0.000380%	0.000414%
12	0.000421%	0.000413%	0.000401%	0.000283%	0.000336%	0.000315%
13	0.000500%	0.000590%	0.000393%	0.000378%	0.000345%	0.000279%
14	0.000311%	0.000287%	0.000270%	0.000258%	0.000256%	0.000173%
15	—	—	0.000192%	0.000252%	0.000260%	0.000390%
16	0.000513%	0.000596%	0.000612%	0.000595%	0.000580%	0.000418%
17	0.000306%	0.000322%	0.000296%	0.000366%	0.000305%	0.000346%
18	0.000443%	0.000612%	0.000563%	0.000890%	0.000873%	0.001051%
19	0.000555%	0.000564%	0.000795%	0.000805%	0.000834%	0.000683%
20	0.000191%	0.000256%	0.000352%	0.000384%	0.000365%	0.000388%
21	0.000203%	0.000213%	0.000214%	0.000417%	0.000350%	0.000323%
22	0.000401%	0.000434%	0.000602%	0.000505%	0.000276%	0.000304%
23	0.003491%	0.004599%	0.000511%	0.000515%	0.000484%	0.000397%
24	0.006215%	0.002744%	0.001554%	0.001071%	0.000966%	0.001381%
25	0.001138%	0.001181%	0.001081%	0.000463%	0.000638%	0.000726%
26	0.000270%	0.000286%	0.000688%	0.000669%	0.000427%	0.000340%

TABLE 7: Coefficients resulting from the first regression.

Y = asset under management						
Year	Regressor	Estimate	Std. error	t value	Pr (>  t )	Adjusted R <sup>2</sup>
2013	No. of employees	0.87683	0.09429	9.299	$9.42e - 10^{***}$	0.7599
2014	No. of employees	0.5921	0.1090	5.432	$1.22e - 05^{***}$	0.8257
	No. of r.e. funds	0.4050	0.1090	3.716	0.00102**	
2015	No. of employees	0.6222	0.1157	5.380	$1.4e - 05^{***}$	0.803
	No. of r.e. funds	0.3593	0.1157	3.106	0.00467**	
2016	No. of employees	0.60102	0.09916	6.061	$2.47e - 06^{***}$	0.8059
	No. of r.e. funds	0.43415	0.09916	4.378	0.000187***	
2017	No. of employees	0.5271	0.1102	4.782	$8.02e - 05^{***}$	0.773
	No. of r.e. funds	0.4989	0.1102	4.526	0.000152***	
2018	No. of employees	0.6110	0.1097	5.570	$1.34e - 05^{***}$	0.8046
	No. of r.e. funds	0.4064	0.1097	3.705	0.00123**	

Signif. codes: \*\*\*0.001; \*\*0.01; \*0.05. All variables exhibit positive coefficients.

TABLE 8: Coefficients resulting from the first regression on clustered data.

Y = asset under management						
Year	Regressor	Estimate	Std. error	t value	Pr (>  t )	Adjusted R <sup>2</sup>
2013	Intercept	-0.3292	0.1237	-2.661	0.01395*	0.8288
	No. of employees	0.3859	0.1706	2.261	0.03349*	
	ind_cl1	1.5280	0.4766	3.206	0.00392**	
	ind_cl2	0.6940	0.2558	2.713	0.01240*	
2014	Intercept	-0.42872	0.06451	-6.646	$8.84e - 07^{***}$	0.9375
	No. of r.e. funds	0.25319	0.07031	3.601	0.00151**	
	ind_cl1	2.14006	0.19282	11.099	$1.03e - 10^{***}$	
	ind_cl2	0.75379	0.15181	4.965	$5.08e - 05^{***}$	



TABLE 8: Continued.

Y = asset under management						
Year	Regressor	Estimate	Std. error	<i>t</i> value	Pr ( $>  t $ )	Adjusted $R^2$
2015	Intercept	-0.48072	0.05868	-8.192	$2.85e-08^{***}$	0.9479
	No. of r.e. funds	0.20154	0.06283	3.208	$0.00391^{**}$	
	ind_cl1	2.22211	0.17092	13.001	$4.40e-12^{***}$	
	ind_cl2	1.02272	0.14032	7.288	$2.04e-07^{***}$	
2016	Intercept	-0.42832	0.07138	-6.000	$4.05e-06^{***}$	0.9211
	No. of r.e. funds	0.31868	0.06846	4.655	$0.00011^{***}$	
	ind_cl1	2.29186	0.20763	11.038	$1.15e-10^{***}$	
	ind_cl2	0.78153	0.14725	5.307	$2.18e-05^{***}$	
2017	Intercept	-0.4801	0.1142	-4.205	$0.000398^{***}$	0.8633
	No. of r.e. funds	0.3098	0.1028	3.015	$0.006599^{**}$	
	ind_cl1	1.9595	0.2731	7.174	$4.51e-07^{***}$	
	ind_cl2	0.5950	0.2003	2.970	$0.007307^{**}$	
2018	Intercept	-0.5594	0.1007	-5.556	$1.94e-05^{***}$	0.9052
	No. of employees	0.3236	0.1072	3.018	$0.0068^{**}$	
	ind_cl1	1.7988	0.2939	6.121	$5.56e-06^{***}$	
	ind_cl2	0.8902	0.1523	5.846	$1.02e-05^{***}$	

Signif. codes: \*\*\*0.001; \*\*0.01; \*0.05.

TABLE 9: Coefficients resulting from the second regression.

Y = net fees						
Year	Regressor	Estimate	Std. error	<i>t</i> value	Pr ( $>  t $ )	Adjusted $R^2$
2013	AUM	0.5571	0.1033	5.393	$1.35e-05^{***}$	0.9236
	Administrative costs	0.4427	0.1033	4.285	$0.000238^{***}$	
	Number of r.e. funds managed	—	—	—	—	
2014	AUM	0.5242	0.1399	3.747	$0.000945^{***}$	0.8791
	Administrative costs	0.4480	0.1399	3.203	$0.003693^{**}$	
	Number of r.e. funds managed	—	—	—	—	
2015	AUM	0.4835	0.1237	3.910	$0.000625^{***}$	0.9342
	Administrative costs	0.5063	0.1237	4.094	$0.000389^{***}$	
	Number of r.e. funds managed	—	—	—	—	
2016	AUM	0.48515	0.08832	5.493	$1.05e-05^{***}$	0.9532
	Administrative costs	0.52311	0.08832	5.923	$3.50e-06^{***}$	
	Number of r.e. funds managed	—	—	—	—	
2017	AUM	—	—	—	—	0.9483
	Administrative costs	0.97381	0.04641	20.983	$6.0142e-17$	
	Number of r.e. funds managed	—	—	—	—	
2018	AUM	—	—	—	—	0.9239
	Administrative costs	0.96284	0.05631	17.1	$1.43e-14^{***}$	
	Number of r.e. funds managed	—	—	—	—	

Signif. codes: \*\*\*0.001; \*\*0.01; \*0.05.

The second multivariate regression is carried out to verify the dependence of net commissions. Again, the results obtained for each regression are reported in Table 9 after removing any insignificant regressors.

All regression coefficients show positive coefficients. In the range considered, commissions depend positively on assets under management and personnel expenses. Exceptions are the last two years during which net commissions depend only on personnel expenses.

This empirical evidence means that by increasing assets (for all years except 2017–2018) or expenses by one unit

compared to their average, a similar marginal effect on net commissions is achieved. The absence of the intercept means that if the regressors are equal to the average, commissions are not significantly different from the average.

The analyses resulting from repeating the multivariate regression on clustered data are exactly the same as those carried out on nonclustered data because clusterization turns out to be not statistically significant. There is no evidence of economies of scale: increasing assets under management or employee expenses increases net commission costs.

## 5. Conclusions

The analyses carried out so far have not shown any relationship that would suggest economies of scale. The first regression clearly showed that the assets under management, i.e., those capable of generating income, depend on the number of funds managed by each SGR and the number of employees. As the number of funds increases, there is logically an increase in assets. Also by increasing the number of employees, there are more resources able to manage the possible creation of new funds. This last aspect translates in practical terms into higher net commissions to be paid as more resources to be employed, alias costs, and lower profitability.

Since empirical evidence does not provide a *modus operandi* to be followed to minimize costs and maximize assets under management, it may be thought that the issue is still little known and of great interest to all market participants. This uncertainty suggests policy considerations that deserve to be further investigated.

The absence of economies of scale could be linked to the particular characteristics of each fund and its business model. This consideration deserves to be studied in greater depth. In fact, when a “traditional” fund is established, typically based on big asset located to primary tenants, the management company cannot charge high unit fees. In this case, high volumes managed do not imply high level of cost, but also high revenues to earn. When an innovative fund with high added value to be performed in the management of the assets is established (this implies the presence of unique and not easily replicable assets), the management company may charge higher fees, due to the specificity of the assets, with the consequent risk that management costs increase more than proportionally.

In addition, anyway minimum management costs are imposed by national law and supervisory authorities to protect and safeguard the rights of all stakeholders and of the market. These kinds of costs, for the part directly related to the funds managed, do not allow economy of scale.

In summary, there is no single model for profit maximization and cost minimization, but all depends on the business model that each real estate asset management company decides to adopt and, consequently, on the type of assets that are managed.

## Data Availability

The data used were obtained through the balance sheets that each SGR provided to the Department of Business Economics of the University of Parma.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] G. Abatecola, A. Caputo, M. Mari, and S. Poggesi, “Real estate management: past, present, and future research directions,” *International Journal of Globalisation and Small Business*, vol. 5, no. 1, pp. 98–113, 2013.

- [2] J. Dombrow and G. K. Turnbull, “Trends in real estate research, 1988–2001: what’s hot and what’s not,” *The Journal of Real Estate Finance and Economics*, vol. 29, no. 1, pp. 47–70, 2004.
- [3] L. L. Johnson and T. Keasler, “An industry profile of corporate real estate,” *Journal of Real Estate Research*, vol. 8, pp. 455–473, 1993.
- [4] M. A. O’Mara, *Strategy and Place; Managing Corporate Real Estate for Competitive Advantage*, Free Press, New York, NY, USA, 1999.
- [5] A. Schmitz and D. L. Brett, *Real Estate Market Analysis. A Case Study Approach*, The Urban Land Institute, Washington, DC, USA, 2001.
- [6] B. P. Singer, B. A. G. Bossink, and H. J. M. Vande Putte, “Corporate real estate and competitive strategy,” *Journal of Corporate Real Estate*, vol. 9, no. 1, pp. 25–38, 2007.
- [7] F. Tajani, P. Morano, M. P. Saez-Perez et al., “Multivariate dynamic analysis and forecasting models of future property bubbles: empirical applications to the housing markets of Spanish metropolitan cities,” *Sustainability*, vol. 11, no. 13, 2019.
- [8] P. Morano, F. Tajani, and M. Locurcio, “Multicriteria analysis and genetic algorithms for mass appraisals in the Italian property market,” *International Journal of Housing Markets and Analysis*, vol. 11, no. 2, 2018.
- [9] N. Barlow and E. Lawson, “Real estate’s vital role in corporate finance,” *Journal of Corporate Accounting & Finance*, vol. 1, no. 4, pp. 361–365, 1990.
- [10] J. D. Benjamin, P. Chinloy, and W. G. Hardin, “Local presence, scale and vertical integration: brands as signals,” *The Journal of Real Estate Finance and Economics*, vol. 33, no. 4, pp. 389–403, 2006.
- [11] M. B. Trundle, “Capturing hidden value for your shareholders,” *Journal of Corporate Real Estate*, vol. 7, no. 1, pp. 55–71, 2005.
- [12] C. Giannotti and G. Mattarocci, “Risk diversification in a real estate portfolio: evidence from the Italian market,” *Journal of European Real Estate Research*, vol. 1, no. 3, pp. 214–234, 2008.
- [13] F. J. Acoba and S. P. Foster, “Aligning corporate real estate with evolving corporate missions: process—based management models,” *Journal of Corporate Real Estate*, vol. 5, no. 2, pp. 143–164, 2003.
- [14] S. Duckworth, “Realizing the strategic dimension of corporate real property through improved planning and control systems,” *Journal of Real Estate Research*, vol. 8, pp. 495–509, 1993.
- [15] A. Lindholm, K. M. Gibler, and K. I. Leväinen, “Modeling the value-adding attributes of real estate to the wealth,” *Journal of Real Estate Research*, vol. 28, pp. 445–475, 2006.
- [16] A. Banfi, G. Borello, and F. Pampurini, *Una Stima del Livello di Efficienza Delle Società di Gestione del Risparmio Operanti in Italia*, Banca d’Italia, Rome, Italy, 2011.
- [17] B. W. Ambrose, M. J. Highfield, and P. D. Linneman, “Real estate and economies of scale: the case of REITs,” *Real Estate Economics*, vol. 33, no. 2, pp. 323–350, 2005.
- [18] H. Y. Kim, “Economies of scale and economies of scope in multiproduct financial institutions: further evidence from credit unions,” *Journal of Money, Credit and Banking*, vol. 18, no. 2, pp. 220–226, 1986.
- [19] B. W. Ambrose, F. Fuerst, N. Mansley, and Z. Wang, “Size effects and economies of scale in European real estate companies,” *Global Finance Journal*, vol. 42, 2019.

- [20] W. McIntosh, Y. Liang, and D. L. Thompkins, "An examination of the small-firm effect within the REIT industry," *Journal of Real Estate Research*, vol. 6, pp. 9–18, 1991.
- [21] P. Linneman, "Forces changing the real estate industry forever," *Wharton Real Estate Review*, vol. 1, pp. 1–12, 1997.
- [22] M. Bers and T. M. Springer, "Economies-of-scale for real estate investment trusts," *Journal of Real Estate Research*, vol. 14, pp. 275–290, 1997.
- [23] M. Bers and T. M. Springer, "Sources of scale economies for REITs," *Real Estate Finance*, vol. 14, pp. 47–56, 1998.
- [24] M. Bers and T. M. Springer, "Differences in scale economies among real estate investment trusts: more evidence," *Real Estate Finance*, vol. 15, pp. 37–44, 1998.
- [25] B. W. Ambrose, S. R. Ehrlich, W. T. Hughes, and S. M. Wachter, "REIT economies of scale: fact of fiction," *The Journal of Real Estate Finance and Economics*, vol. 20, no. 2, pp. 211–224, 2000.
- [26] D. R. Capozza and P. J. Seguin, "Managerial style and firm value," *Real Estate Economics*, vol. 26, no. 1, pp. 131–150, 1998.
- [27] D. A. Latzko, "Economies of scale in mutual fund administration," *Journal of Financial Research*, vol. 22, no. 3, pp. 331–339, 1999.
- [28] D. R. Capozza and P. J. Seguin, "Debt, agency and management contracts in REITs: the external advisor puzzle," *The Journal of Real Estate Finance and Economics*, vol. 20, no. 2, pp. 91–116, 2000.
- [29] R. I. Anderson, R. Fok, T. Springer, and J. Webb, "Technical efficiency and economies of scale: a non-parametric analysis of REIT operating efficiency," *European Journal of Operational Research*, vol. 139, no. 3, pp. 598–612, 2002.
- [30] L. J. Mester, "A study of bank efficiency taking into account risk-preferences," *Journal of Banking & Finance*, vol. 20, no. 6, pp. 1025–1045, 1996.
- [31] M. L. Bianchi and M. G. Miele, *I Fondi Comuni Aperti in Italia: Performance Delle Società di Gestione del Risparmio*, Banca d'Italia, Rome, Italy, 2011.
- [32] E. Geretto and R. Morassut, "La valutazione delle performance economico—operative delle società di gestione del risparmio," *Banche e Banchieri*, vol. 6, pp. 452–470, 2010.
- [33] G. Abate, "Real estate finance e sgr immobiliari: caratteristiche strutturali e dinamiche reddituali," *Bancaria*, vol. 3, 2011.
- [34] C. Giannotti and G. Mattarocci, "La redditività e l'efficienza delle sgr immobiliari," *Bancaria*, vol. 69, no. 10, pp. 32–36, 2014.
- [35] Banca d'Italia, *Rapporto Sulla Stabilità Finanziaria*, Banca d'Italia, Rome, Italy, 2019.

## Research Article

# GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China

Shaopei Chen , Dachang Zhuang , and Huixia Zhang

*School of Public Administration, Guangdong University of Finance and Economics, Guangzhou, China*

Correspondence should be addressed to Dachang Zhuang; [Zhuang-dc@163.com](mailto:Zhuang-dc@163.com)

Received 28 December 2019; Revised 14 February 2020; Accepted 2 March 2020; Published 23 April 2020

Guest Editor: Francesco Tajani

Copyright © 2020 Shaopei Chen 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.

In the past decades, the booming growth of housing markets in China triggers the urgent need to explore how the rapid urban spatial expansion, large-scale urban infrastructural development, and fast-changing urban planning determine the housing price changes and spatial differentiation. It is of great significance to promote the existing governing policy and mechanism of housing market and the reform of real-estate system. At the level of city, an empirical analysis is implemented with the traditional econometric models of regressive analysis and GIS-based spatial autocorrelation models, focusing in examining and characterizing the spatial homogeneity and nonstationarity of housing prices in Guangzhou, China. There are 141 neighborhoods in Guangzhou identified as the independent individuals (named as area units), and their values of the average annual housing prices (AAHP) in (2009–2015) are clarified as the dependent variables in regressing analysis models used in this paper. Simultaneously, the factors including geographical location, transportation accessibility, commercial service intensity, and public service intensity are identified as independent variables in the context of urban development and planning. The integration and comparative analysis of multiple linear regression models, spatial autocorrelation models, and geographically weighted regressing (GWR) models are implemented, focusing on exploring the influencing factors of house prices, especially characterizing the spatial heterogeneity and nonstationarity of housing prices oriented towards the spatial differences of urban spatial development, infrastructure layout, land use, and planning. This has the potential to enrich the current approaches to the complex quantitative analysis modelling of housing prices. Particularly, it is favorable to examine and characterize what and how to determine the spatial homogeneity and nonstationarity of housing prices oriented towards a microscale geospatial perspective. Therefore, this study should be significant to drive essential changes to develop a more efficient, sustainable, and competitive real-estate system at the level of city, especially for the emerging and dynamic housing markets in the megacities in China.

## 1. Introduction

Since the 1990s, the reform of real-estate system and land-use system in China have been implemented, leading to a continuous booming growth of the real-estate market, which has grown into one of the pillar industries in the national economic development. After decades of the dramatic development of housing market, the housing prices in China has been driven to consistent increasing, especially in the mega cities, such as Beijing, Shanghai, Shenzhen, and Guangzhou. Currently, the tremendous changes of housing prices in China are increasingly being watched by relevant research

organizations, scholars, and other interest groups, as it is not only a crucial economic issue related to the benign development of the housing market and the optimization and adjustment of the national economic structure but also a paramount social issue related to the urban residents' livelihood [1, 2]. Especially, the long-run urban spatial expansion, large-scale urban infrastructural construction, and fast-changing urban planning have been deriving the rapid changes of housing prices in the megacities in China. This triggers the significant need to examine the specific affecting factors in the context of urban spatial development, infrastructure layout, and planning for promoting the governing

policy and spatial controlling mechanism of housing market, and the reform of real-estate system and land-use system in the level of city.

Scholars have conducted research on the effects of housing prices from two perspectives, i.e., the national level and the city level. At the national level, the factors influencing on the changes in housing prices generally include monetary, fiscal, and housing policies [3–7], while the specific factors affecting the housing prices at the level of city mainly involve with urban development, land use, population, housing supply and demand, and urban planning [8–12]. At the level of city, the study on the fluctuations of housing prices and the correlations to the influencing factors has been a hot topic in the fields of governing policy of housing market and the reform of real-estate and related urban systems, e.g., land-use system, as well as the hotspots in urban spatial layout, urban planning, and sustainability development [10]. Due to the commodity attribute of the house, since the late 1970s, the traditional regressive models for asset pricing, i.e., hedonic price models [13], have been widely used in the evaluation of house values [14–17]. Hedonic price models define that the price of a specific commodity should be constituted by several different elements in which their number and combination are discrepant, leading to unequal prices for different commodities. The application of hedonic price models in housing market aims at decomposing the key components of housing prices and applying a regression analysis to quantitatively measure the impact of each element [13]. Currently, hedonic price models have been adopted and extended widely to investigate the complex effects on housing prices in the context of urban neighborhoods and submarkets. As an example, Basu and Thibodeau examine the spatial autocorrelation in transaction prices of single-family properties in Dallas, Texas, by an empirical analysis which applies a semilog hedonic house price equation and a spherical autocorrelation function with the data for over 5000 transactions of homes sold. This study reveals a strong evidence of spatial autocorrelation in transaction prices within submarkets [10]. Can utilizes the traditional econometric models based on a hedonic price regressive analysis extended to incorporate spatial neighborhood dynamics to model the housing price determination process from an explicit geographic perspective, leading to characterize the spatial variation with respect to the influence of housing attributes on housing prices [11]. To more precisely clarify the boundaries of the housing submarkets, Leishman introduces an application of the multilevel hedonic model as a tool to identify housing submarkets, and a method for identifying temporal changes within the submarket system [9]. Importantly, the complex environmental factors affecting housing prices have been increasingly kept getting attention by some professionals and researchers, who are actively looking for a new paradigm in exploring the influencing factors and regional differences of housing prices for effectively regulating housing markets and constructing the scientific governing policy for the real-estate system and other related urban systems, such as the land-use system. For instance, Keskin et al. develop a novel extension of the standard use of hedonic price models in

event studies to investigate the impact of natural disasters on real-estate values, aiming to explore the impact of a recent earthquake activity on housing prices and their spatial distribution in the Istanbul housing market on introducing a multilevel approach with the extension of hedonic models. Such an approach allows the isolation of the effects of earthquake risk and explores the differential impact in different submarkets, resulting in better capturing the granularity of the spatial effects of environmental events than the standard approach [18]. Furthermore, to examine the existence of the ripple effect from the change of housing prices between different regional housing markets, Larm et al. (2017) uses the autoregressive distributed lags (ARDL) cointegration and causality techniques [19] to characterize the ripple effect as a lead-lag effect and a long-run convergence between the regional and Amsterdam housing prices in the Netherlands [20]. The abovementioned studies are of significance to promote the traditional econometric models that utilize the hedonic price regression to reveal the direction of causality between housing prices and the complex affecting factors and further detect the spatial differentiation and spatial nonlinear characteristics of housing prices. With the rapid development of spatial econometric models, the spatial properties of housing prices have received increasing attention. For instance, Barreca et al. propose a new perspective on the spatial relationship between urban vibrancy and neighborhood services and the real-estate market. In this study, a Neighborhood Services Index (NeSI) is provided to identify the most and least vibrant urban areas in a city, and spatial autoregressive models are used to manage spatial effects and to identify the variables that significantly influence the process of housing price determination.

Indeed, it is complex and difficult to explore the influencing factors that trigger the housing price changes and spatial differentiation, particularly for the emerging and dynamic housing markets in China [21]. In the 1990s, most of the research studies oriented to evaluate the influencing factors of housing prices at the level of city conducted by Chinese scholars focused on the qualitative analysis on the correlations between the housing supply and demand, the composition of housing prices, the reform of housing system, and the effects of real-estate policy. Since the early 2000s, some scholars in China have gradually used hedonic price models to quantitatively characterize the influencing factors of housing prices [22, 23]. However, the spatiality of a house makes it different from the ordinary commodities, mainly reflected in the fact that the housing value usually shows spatial homogeneity and nonstationarity. For example, the houses located in the developed urban area usually possess the favorable geographical location and the well-developed urban infrastructure, leading to much higher prices than those of the houses located in the newly developing urban area even with better building structure design and construction quality. This implies that the traditional regression of hedonic price models without the consideration of spatial differentiation and spatial nonlinear characteristics is difficult to identify the determinations of housing prices. In recent years, spatial econometric models,

such as spatial interpolation, spatial autocorrelation analysis model, ESDA (Exploratory Spatial Data Analysis), and geographically weighted regression (GWR) model, have been widely applied to characterize the spatiality of housing prices and the spatial autocorrelation among the influencing factors [24–28]. Among the spatial econometric models, the local spatial autocorrelation analysis models, e.g., GWR model, have obvious advantages in characterizing the changes and spatial differentiation of housing prices. The GWR model extends the traditional regression models by introducing the local location factors of housing prices, leading to better deal with the issue of quantifying the spatial heterogeneity of housing prices under the affecting factors which are identified with specific spatial correlations [29]. Furthermore, the GWR model has been an effective way to achieve the spatial nonstationarity analysis of housing price changes [30].

Currently, many studies on the determinations of housing prices and the correlations to the influencing factors have been implemented by introducing the locations of houses and the spatial adjacency relations into hedonic price models [31–35]. Moreover, various theories and methods of spatial econometric models, such as ESDA and spatial autocorrelation analysis, are applied to examine the spatiality of housing prices [21, 36–42]. Nevertheless, under the environment of long-run urban spatial development, large-scale construction of urban infrastructure, and rapid transformation of urban planning in the megacities in China, the connections which are not limited to spatial links between adjacent areas within the city have become effective reflections of the spatial correlation among the housing prices in the context of neighborhoods. Although the traditional hedonic price models and spatial autocorrelation analysis models have been extended to use local spatial weights as the parameters of regression analysis, they only consider the spatial correspondence between the given dependent variables and the independent variables. Relevant research studies in recent years have focused on the identification of urban areas characterized by urban development and housing prices, which can effectively support the revision of the urban development plan and its regulatory acts, as well as the strategic urban policies and actions [12]. Nevertheless, at the level of a city, how the urban development differences, particularly with a microscale geospatial perspective (i.e., in the context of neighborhoods), triggers varying degrees of impact on housing prices, especially what and how to determine the spatial heterogeneity and nonstationarity of housing prices are still lacking attention. This presents the urgent need to explore how the intense urban spatial expansion, large-scale urban infrastructure development, and fast-changing urban planning determine and characterize the housing prices changes and spatial differentiation, which is of great significance to promote the governing policy and spatial controlling mechanism of housing markets and the reform of the real-estate system and land-use system.

Therefore, at the level of city, this paper implements an empirical analysis with the integration and comparative discussion of the traditional econometric models of

regressive analysis and GIS-based spatial autocorrelation analysis tools, focusing in exploring and characterizing the spatial homogeneity and nonstationarity of the housing prices in 2009–2015 in the context of the neighborhoods in Guangzhou, China. There are total 141 neighborhoods in Guangzhou identified as area units, and their average annual housing prices (AAHP) at different points in time (2009–2015) are defined as dependent variables. Simultaneously, the factors including geographical location condition, transportation accessibility, commercial service intensity, and public service intensity are identified as independent variables in the context of urban spatial expansion, infrastructural layout, land use, and urban planning, leading to examine the spatial correlations to the AAHP in neighborhoods. This aims to examine and characterize what and how to determine and characterize the spatial homogeneity and nonstationarity of housing prices, resulting in promoting the governing policy and spatial controlling mechanism of the housing markets and the reform of real-estate and land-use systems in Guangzhou, China. This would support a deep knowledge of the spatial heterogeneity and nonstationarity of housing prices identified and characterized by the regional differences in urban spatial expansion, infrastructural layout, land use, and urban planning. It is of great significance to trigger significant changes to develop a more efficient, sustainable, and competitive governing policy and spatial controlling mechanism for the real-estate system and land-use system at the level of city, especially in the megacities in China.

## 2. Study Area and Methodology

*2.1. Study Area.* The main building area of Guangzhou is composed of 8 administrative districts with 141 neighborhoods, which have a considerable population of about ten million and a large number of residential areas. Among the districts, Yuexiu District is the historical urban center of the city and forms the urban core with Liwan, Haizhu, and Tianhe Districts (see Figure 1). The Central Business District (CBD), which includes the neighborhoods of Liede, Yuncun, and Licun, is the business center of the city and located in Tianhe District. Other administrative districts surround the urban core, i.e., Baiyun and Huadu Districts in the north, Huangpu District in the east, and Panyu District in the south. As shown in Figure 1, the primary urban functional areas are illustrated, which include the CBD, Baiyun New Town, and University Town.

*2.2. Data Sources and Preprocessing.* The data adopted in this study is collected and preprocessed from different data sources. The average annual housing prices (AAHP) mainly comes from the “Report on the Operation of Guangzhou Real Estate,” which is issued by the Guangzhou Municipal Housing and Urban-Rural Development Bureau at each month through the government website of “Sunshine Family” ([http://zfcj.gz.gov.cn/data/Category\\_623/Index.aspx](http://zfcj.gz.gov.cn/data/Category_623/Index.aspx)). The report provides the specific data of the monthly average housing prices of all neighborhoods in

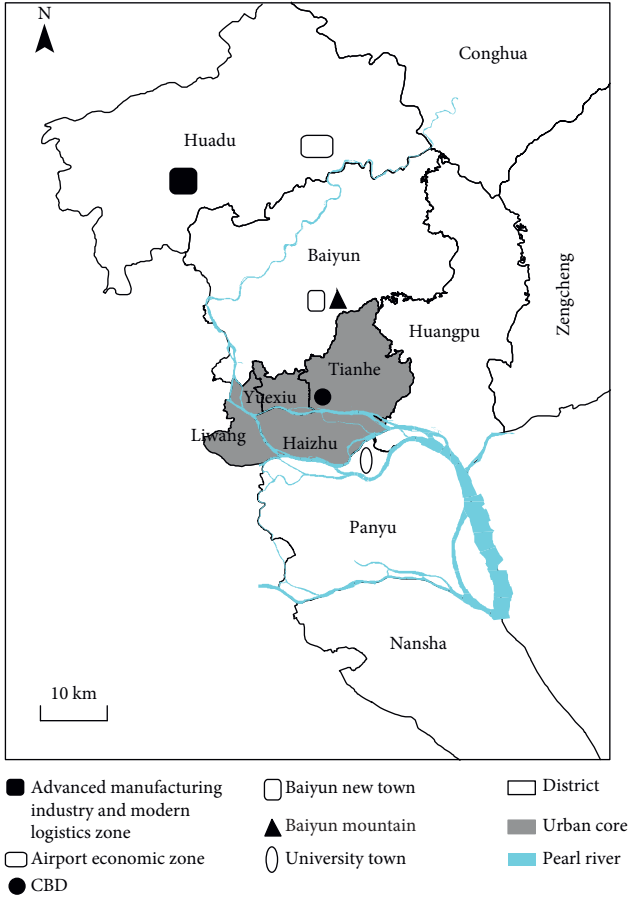


FIGURE 1: Study area.

Guangzhou, including the monthly average transaction prices of new housing, stock housing, and second-hand housing. The data of the monthly average prices in the neighborhoods can better present the spatial and temporal fluctuation of housing prices and accurately characterize the overall evolution of the housing prices in Guangzhou. In addition, the geographical spatial data of Guangzhou are extracted through vector digital processing, remote image nesting, and coordinate registration based on the OpenStreetMap database, and the medium and high resolution remote sensing images of the city of Guangzhou, which include the neighborhoods, urban transportation data (involving roads, bus stops, and metrostations), residential areas, buildings, and other urban infrastructure, and land-use data, e.g., the financial institutions (such as banks), restaurants, retail stores, supermarkets, educational institutions (such as primary and secondary schools), medical institutions (such as clinics and hospitals), government agencies (such as the administrative service centers at all administrative grades), parks, and squares. Particularly, in the environment of ArcGIS 10.2, the administrative area of each neighborhood is built as a polygon with its geometric center point, which stores all the information of the neighborhood, including the AAHP, number of bus stops and metrostations, and road density. Moreover, the road network is based on ArcGIS for topology modelling to fit the

network analysis for searching the shortest path between two given points. Other data, i.e., bus stops and metrostations, residential areas, buildings, and other urban infrastructure data, are represented as points in GIS, leading to calculate their density indexes in the context of neighborhoods.

### 2.3. Methodology

**2.3.1. Regression Model.** The regression model is a mathematical tool for the quantitative description of the statistical relations between given observation values (i.e., variables). Especially, it can provide the calculation method and theory for regression analysis to study the specific dependence of one explained variable (dependent variable) on another (independent variable). Regression linear analysis including multiple regression variables is identified as a multiple linear regression model, and its general equation is presented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (1)$$

where  $Y$  is a random variable,  $K$  is the number of independent variables,  $\beta_0$  is a constant term, and  $\beta_j$  ( $j = 1, 2, \dots, k$ ) is the regression coefficient.

Hedonic price model is a multiple regression model, which regresses on a specific price through quantifying the different properties of the price by multiple explanatory variables (independent variables). Hedonic price models mainly have three functional forms: linear model, log-linear model, and semilog model. In a specific application, therefore, the effective variables and function equations need to be selected properly according to different functional forms [34].

**2.3.2. GWR Model and Factor Selection.** The GWR model assumes that the regression coefficient is a function of the geographic location of the observation point and incorporates the spatial characteristics of the data into the regression model, leading to realizing the analysis of the spatial differences of the explained variables (dependent variables) [43]. The general formula of the model is listed as follows:

$$y_i = \beta_0(\mu_i, v_i) + \sum_{k=1}^p \beta_k(\mu_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (2)$$

where  $(\mu_i, v_i)$  is the coordinate of observation point  $i$  and  $\beta_k(\mu_i, v_i)$  is its  $k$ th regression parameter, which is a function describing the geographic location of observation point  $i$ .

Indeed, the changes of housing prices are affected by the rapid operation of urban systems, especially by the intense urban spatial expansion, large-scale urban infrastructural development, and fast-changing urban planning in megacities in China. To ensure the quantitative analysis under the framework of the geographically weighted regression model, in the context of urban spatial development, land use, infrastructure layout, and planning, four factors including geographical location condition, transportation accessibility, commercial service intensity, and public intensity are selected. Particularly, the specific variables which measure the intensity of each factor are shown in Table 1.

TABLE 1: Selection of the influencing factors and their variables.

Factor	Variable
Geographical location condition ( $D$ )	Distance away from the CBD ( $d_1$ )
Transportation accessibility ( $T$ )	Bus stop density ( $t_1$ )
	Number of metro station ( $t_2$ )
	Road weighted density ( $t_3$ )
Commercial service intensity ( $B$ )	Commercial building density ( $b_1$ )
	Financial institution density ( $b_2$ )
	Restaurant density ( $b_3$ )
	Retail store/supermarket density ( $b_4$ )
Public service intensity ( $p$ )	Educational institution density ( $p_1$ )
	Medical institution density ( $p_2$ )
	Government agency density ( $p_3$ )
	Park/square density ( $p_4$ )

In Table 1, the distance away from the CBD ( $d_1$ ) is identified as the length of the minimum travel-time path between the neighborhood and the CBD. Such a path is attained by the algorithm of Dijkstra with an average speed of 60 km/hour of cars through the road network topology model of Guangzhou in the environment of GIS. The application of Dijkstra shortest path algorithm and road network modelling can be found in our previous works [44–46]. In the factor of transportation accessibility, the value of bus stop density ( $t_1$ ) is calculated by the total number of bus stops located in the neighborhood divided by its area ( $\text{km}^2$ ). Under a microscale geospatial perspective, the neighborhood usually has only one or two metrostations; therefore, the number of metro stations ( $t_2$ ) is used as one of its transportation accessibility variables. The road weighted density ( $t_3$ ) in each neighborhood in the city of Guangzhou is calculated as follows:

$$M_j^w = \frac{\sum_{i=1}^n l_i^\alpha w_i}{S_j}, \quad i = 1, 2, \dots, n, \quad (3)$$

where  $M_j^w$  is the road weighted density of neighborhood  $j$  ( $\text{km}/\text{km}^2$ ),  $l_i^\alpha$  presents the length of road  $i$  with a grade of  $\alpha$  and a weight of  $w_i$  in neighborhood  $j$ ,  $n$  is the total number of roads, and  $S_j$  is defined as the area of neighborhood  $j$ . In this paper, the urban roads are classified into four grades, Expressway (EX), Main Road (MR), Secondary Road (SR), and Internal Road (IR), according to “Highway Engineering Technical Standards” (JTG B01-2014), which are issued by the Ministry of Transport of China. Furthermore, the weight of each grade of road is implemented by the Analytic Hierarchy Process (AHP) [47] and the expert scoring method, as shown in Table 2. The AHP is given by Thomas Saaty (1980) and is usually referred to as the Saaty method. As a systematic method, the analytic hierarchy process (AHP) takes a complex multiobjective decision-making problem as a system and then decomposes the objective into multiple objectives or criteria. Furthermore, the criterium are resolved into several levels of multiobjective (or criteria constraints), and the single rank (weight) and the total rank can be calculated by the fuzzy quantitative method of qualitative index, to be used as the system method of objective (multi index) and multischeme optimization decision [48, 49]. Therefore, this study adopts the analytical

TABLE 2: Grades and weights of the urban roads in the study area.

Road	EX	MR	SR	IR
Grade	1	2	3	4
Weight	0.5	0.25	0.2	0.1

hierarchical process (AHP) to evaluate the significance of the urban road for promoting the result of the calculation of road weighted density in neighborhood.

For the factor of commercial service intensity, commercial buildings, financial institutions, restaurants, retail stores, supermarkets, educational institutions, medical institutions, government agencies, parks, and square areas, respectively, are represented as point objects. For instance, the commercial building density index ( $b_1$ ) can be attained through the number of points (which represent the commercial buildings) in the neighborhood divided by the neighborhood’s area ( $\text{km}^2$ ). Following this method, therefore, other variables, including financial institution density ( $b_2$ ), restaurant density ( $b_3$ ), retail store/supermarket density ( $b_4$ ), educational institution density ( $p_1$ ), medical institution density ( $p_2$ ), government agency density ( $p_3$ ), and park/square density ( $p_4$ ), can be processed and calculated.

To access the relative development intensity of each neighborhood, the variables measuring the intensity of each factor are further processed, respectively, and the general formulation can be presented as follows:

$$A_i = \frac{x_i}{\left(\sum_{i=1}^N x_i\right)/n}, \quad i = 1, 2, \dots, n. \quad (4)$$

$A_i$  identifies the relative development level in neighborhood  $i$  based on  $x_i$ , which is defined as a specific variable of the factors, such as the road weighted density ( $t_3$ ), and  $n$  is the total number of neighborhoods. Therefore, for each factor, the indicator measuring its intensity in neighborhood  $i$  can be characterized as the arithmetic mean of  $A_i$ . For example, the indicator measuring the intensity of the factor of transport accessibility in neighborhood  $i$  can be calculated by

$$K_i^T = \frac{A_i^{t_1} + A_i^{t_2} + A_i^{t_3}}{3}. \quad (5)$$



In terms of formula (5), all indicators measuring the intensity of the factors of location condition, transport accessibility, commercial service intensity, and public service intensity are represented by  $K_i^D, K_i^T, K_i^B$ , and  $K_i^P$ , respectively.

**2.3.3. Spatial Autocorrelation Analysis Model.** The spatial autocorrelation analysis can be divided into global spatial autocorrelation and local spatial autocorrelation. The key index of the global spatial autocorrelation analysis is Moran's I index, which can be calculated by [50]

$$I = \frac{n}{\sum_{i=1}^n (y_i - \bar{y})^2} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y}) w_{ij} (y_j - \bar{y})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}, \quad (6)$$

where  $y_i$  is the attribute value of observation point  $i$  and  $w_{ij}$  is the adjacency matrix. The value of Moran's I ranges from  $-1$  to  $1$ , when  $I > 0$ , it means that a given area has a positive correlation with its surrounding areas, and they tend to spatial agglomeration, otherwise, it implies that the areas tend to spatial dispersion. When  $I = 0$ , it illustrates that there is no spatial correction between the given area and its surrounding areas. However, the global spatial autocorrelation analysis cannot detect whether the observation values are high aggregation or low aggregation in space [31]. It is necessary to further explore the local spatial homogeneity or heterogeneity between a given area and its surrounding areas through the local spatial autocorrelation analysis, so as to clarify how the spatial dependence among the areas in space.

Anselin proposed the local indicators of spatial association (LISA) in 1995 to detect the aggregation in local space and analyze its nonstationarity [51]:

$$I_i = \frac{(y_i - \bar{Y})}{S_i^2} \sum_{j=1, j \neq i}^n w_{ij} (y_j - \bar{Y}), S_i^2 = \frac{\sum_{j=1, j \neq i}^n (y_j - \bar{Y})^2}{n-1} - \bar{Y}^2, \quad (7)$$

where  $y_i$  is the value of observation point  $i$ ,  $\bar{Y}$  is the average value of the associated value, and  $w_{ij}$  represents the spatial weight of observation points  $i$  and  $j$ . The LISA provides four spatial association patterns, in which high-high (HL) level and low-low (LL) level associations define the positive spatial autocorrection and high-low (HL) level and low-high (LH) level associations are identified as the negative spatial autocorrection.

Spatial autocorrelation analysis is the basis for constructing the geographically weighted regression model. When the spatial correlation of observation points is not significant, it demonstrates that the distance between the points has a weak influence on their correlation, resulting in being unnecessary to use the geographically weighted regression.

### 3. Results and Analysis

**3.1. Spatial and Temporal Changes of Housing Prices in Guangzhou.** In 2008, the global financial crisis slowed down the economic growth of China. As a result, the State Council of China launched the "Four Trillion" economic stimulus

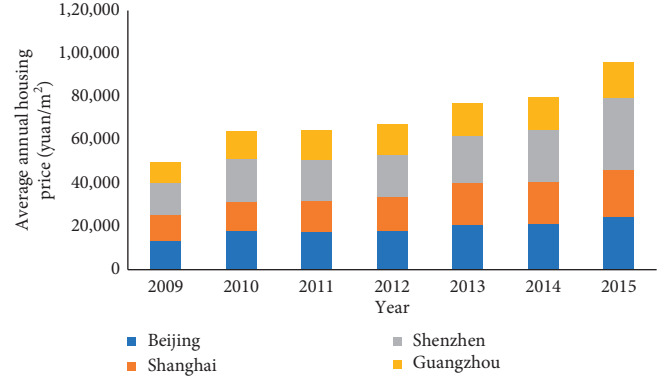


FIGURE 2: Increasing of housing price in 2009–2015.

plan in 2009, which prompted the real-estate market to recover after the financial crisis and the housing prices to return on the upward trend. Figure 2 shows the trend of AAHP in China's major metropolians, including Beijing, Shanghai, Guangzhou, and Shenzhen in 2009–2015. As illustrated in Figure 2, since 2009, the AAHP in the four cities has been rising up. Simultaneously, it can be found that the AAHP in Guangzhou has the slowest growth among these cities. Particularly, it is significantly different from Shenzhen, which is also located in the Pearl River Delta and nearby Guangzhou. During this period, the AAHP in Shenzhen presents a fluctuating increasing and shows the change range is the largest among the four cities.

Through using the Kriging spatial interpolation model of ArcGIS 10.2, the spatial pattern of the average annual housing price (AAHP) in Guangzhou is illustrated in Figure 3. The results show that, in 2009–2015, the area with the higher AAHP in Guangzhou was concentrated in the urban core, i.e., Yuexiu, Liwang, Haizhu, and Tianhe Districts. Furthermore, the CBD in Tianhe District is highlighted with the highest AAHP in Guangzhou, involving with the neighborhoods of Liede, Licun, and Yuancun, which are nearby the Pearl River. As shown in Figure 3, the favorable location of the neighborhoods of Tangxia, Xintang, and Changxing, which are located in the east of Tianhe District and adjacent to the CBD, drives a sharp increasing of the AAHP. According to the perspective of spatiotemporal evolution, the spatial pattern of AAHP in Guangzhou evolved into a ring structure with the center of CBD in 2009–2015.

Figure 3 further demonstrates the existence of spatial differences in the rising of AAHP in 2009–2015. That is, the AAHP of the neighborhoods located in Baiyun, Tianhe, and Huangpu District increased rapidly in 2009–2015, and the area with the slowest growing is concentrated in Panyu District, which has an obvious advantage of location and is an important hub connecting the cities of Shenzhen and Hong Kong. Therefore, as early as 2000, the housing market in Panyu District has been developed and grew rapidly. After nearly ten years of booming growth, the AAHP in the district of Panyu in 2009–2015 is in a relatively stable increasing.

Furthermore, the AAHP in Tonghe and its surrounding neighborhoods of Yongping, Jiahe, and Huangshi in Baiyun

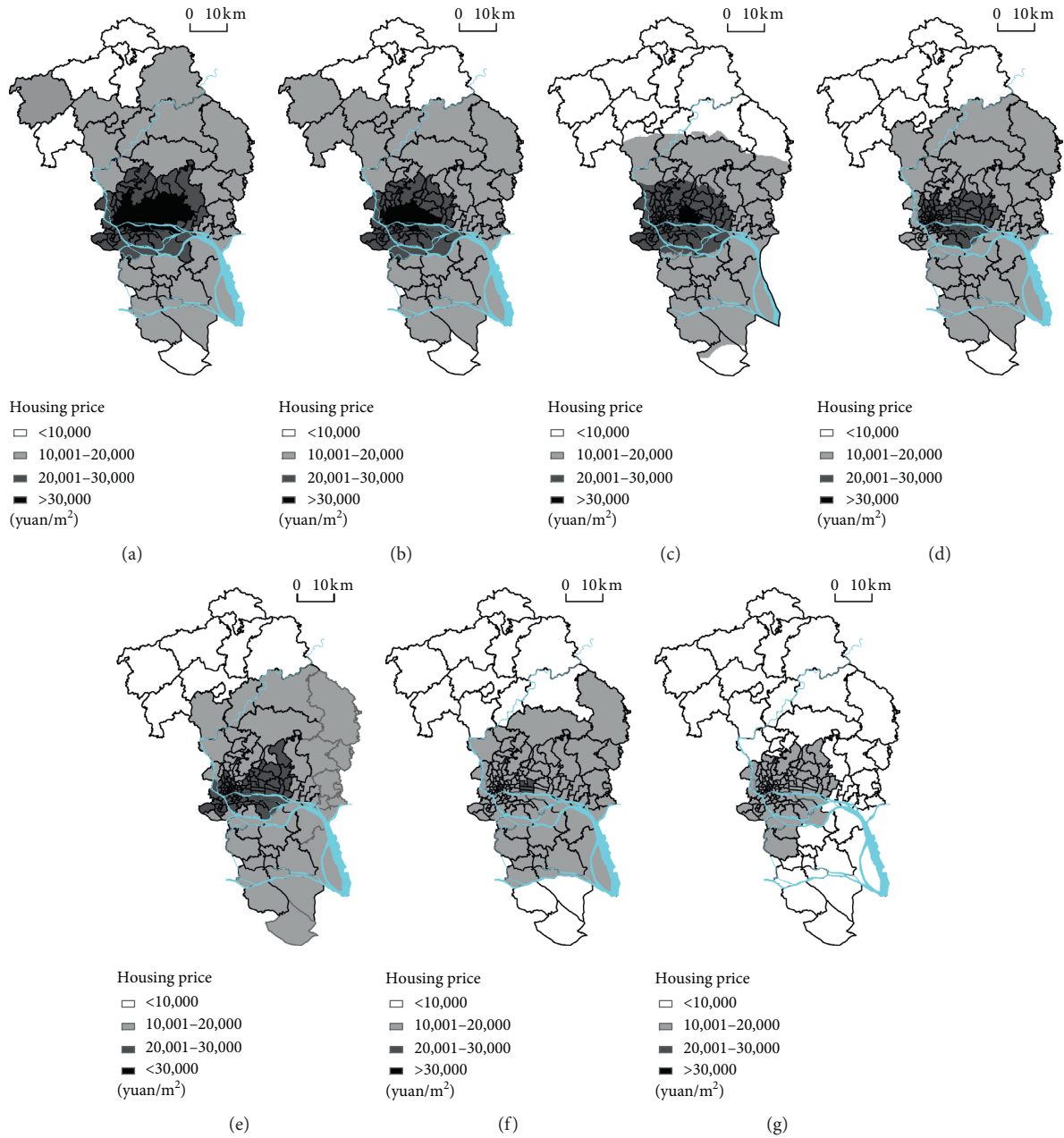


FIGURE 3: Spatial and temporal evolution of housing price in 2009–2015. (a) 2015. (b) 2014. (c) 2013. (d) 2012. (e) 2011. (f) 2010. (g) 2009.

District rapidly increased in 2009–2015, as the central area of the neighborhood of Tonghe in BaiYun District is the Baiyun mountain, which has outstanding natural environment and tourism resources. Especially, the neighborhoods of Yongping, Jiahe, and Huangshi are planned as another business center in the north of the city, i.e., Baiyun New Town (BNT). According to the 2015–2020 Urban Development Plan of Guangzhou, the BNT has been designated as the hug of spatiality, transportation, logistics, and business of the Central Comprehensive Service Function Zone (i.e., the urban core), the Northern Airport Economic Zone (i.e., the New Baiyun International Airport), the Western Advanced Manufacturing Industry, and the Modern Logistics Area (Huadu). Clearly, the transportation infrastructure

inside the BNT, especially the rapid development of metronetwork and urban public service facilities stimulates the booming growth of its housing markets.

As shown in Figure 4, the northwestern neighborhoods of Baiyun District also have a high increasing trend of AAHP in 2009–2015, mainly including Jinshazhou and Luochongwei, which are an important connection point between Guangzhou and Foshan City. This area once became the main region for the construction of resettlement houses (i.e., noncommercial houses), to arrange for residents who have been relocated because of the renewal of the urban core. However, with the promotion of the strategic goals of the integration of Guangzhou-Foshan City, the internal transportation networks in this region have been continuously

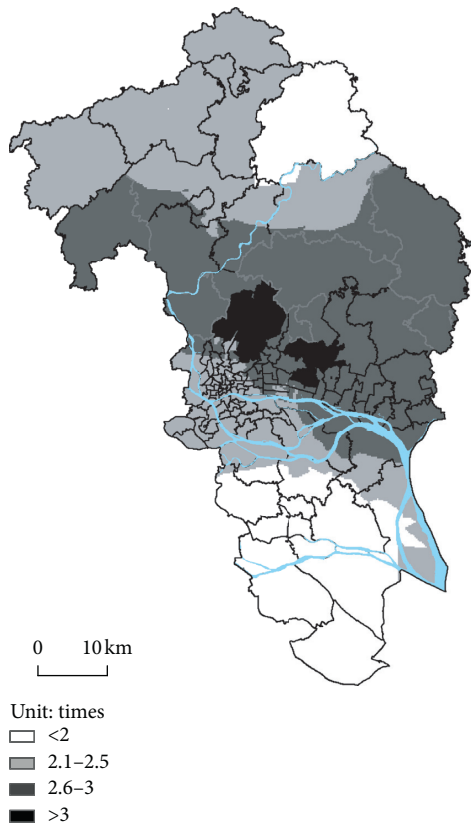


FIGURE 4: Spatial differences in housing prices in Guangzhou (2009–2015).

improved, and the region has gradually grown into a transportation hub for a primary link between Guangzhou and Foshan City. In Figure 4, nevertheless, it is interesting to find that the AAHP growth in 2009–2015 in the urban core including Yuexiu, Liwan, and Haizhou Districts has been relatively flat, but remains at a high level. The excellent geographical location, outstanding urban infrastructure, and favorable public service resources have kept the housing prices at a high level in Yuexiu, Liwan, and Haizhou Districts. Therefore, the increasing space of the AAHP is in line with the expectation and smaller than newly developed areas, such as Baiyun District.

Comparing the spatial differences in the rising of the AAHP in the city of Guangzhou, the AAHP increasing spatial patterns in Beijing, Shanghai, and Shenzhen (see Figure 5) in 2009–2015 are also demonstrated in Figure 6. In these metropolitans, the administrative districts are identified as unit areas, respectively.

As shown in Figure 6(a), the highest growth of AAHP in Shenzhen is highlighted in Nanshan District, which is the innovational center for science and technology in the city. Correspondingly, the AAHP in its adjacent districts, i.e., urban historical center (i.e., Futian District) and the emerging urban center (i.e., Longgang District) also grow rapidly. Compared with the highest increase of housing prices in the CBD in Tianhe District and the BNT in Baiyun District in Guangzhou, which is driven by the robust high-end service industry clusters, mainly including finance and

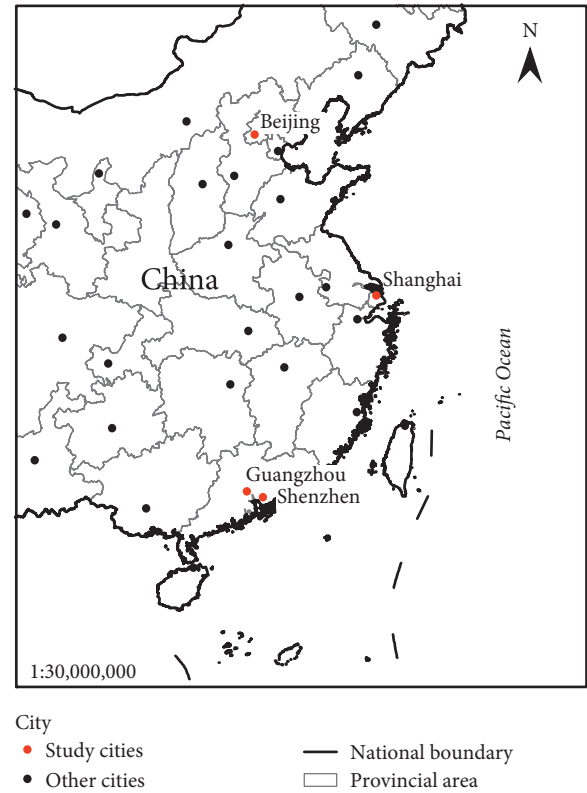


FIGURE 5: Locations of the study cities.

trade, Nanshan District in the city of Shenzhen is known as a scientific and technological manufacture and innovation center in Shenzhen and even across the country. It gathers a large number of high-tech industrial groups such as IT (Information Technology), communications, new materials, new energy, biomedicine, instruments, medical equipment, and mechatronics, resulting in being referred as China's most "Silicon Valley temperament" region. The city of Shenzhen is a young city which was established in 1979; nevertheless, its GDP has approached \$ 400 billion after the booming development of 40 years based on high-tech manufacturing industries. Triggered by the urban development and planning strategy in Shenzhen, Nanshan District has gradually shifted from an industrially oriented urban area to a fully functioning urban central area consisting of the economic center, science-technology center, cultural center, and international communication center, and particularly constitutes the dual main center of the city with Futian District. The area of Shenzhen's administrative district (1997 km<sup>2</sup>) is much smaller than the area of Guangzhou's (the total administrative area 7,436 km<sup>2</sup> and the area of the study area in this paper is 3,061 km<sup>2</sup>), leading to the spatial distribution pattern of the AAHP in the inner area is concentrated at two levels, i.e., Nanshan, Futian, and Luohu Districts in the urban core area around 50,000 yuan/m<sup>2</sup> (about \$ 7,200), and Longgang, Baoan, and Lantian Districts are all around 30,000 yuan/m<sup>2</sup> (about \$ 4,300). Therefore, the influence of the locations on the AAHP in the city of Shenzhen is not obvious. Aside from the factors, such as housing policies and the structure and materials of the

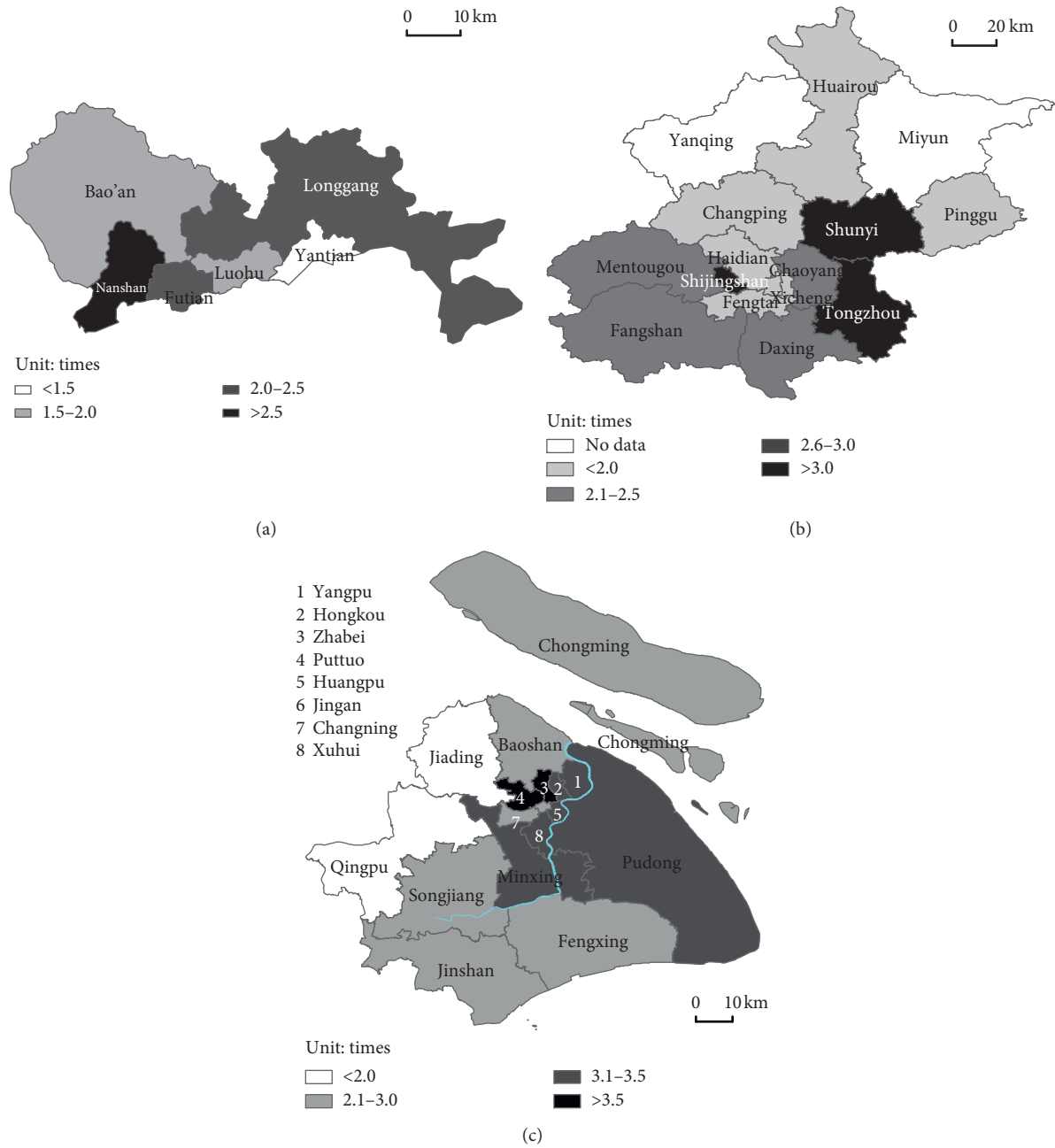


FIGURE 6: Spatial differences in the rise of housing prices in Beijing, Shanghai, and Shenzhen in 2009–2015. (a) Shenzhen. (b) Beijing. (c) Shanghai.

house itself, the affecting impacts are mainly derived from the urban strategic planning and functional area layout in the urban development of Shenzhen.

Figure 6(b) reveals that the areas with the fastest-growing housing prices in Beijing are Tongzhou and Shunyi Districts, which are identified as the subcenters in urban planning, as well as Shijingshan District, which has the well-developed urban infrastructure, favorable geographical location (the urban core in Beijing), and excellent landscape resources (the urban leisure and entertainment zone). In 2009–2015, the increase of AAHP in Shanghai was the fastest among the four cities. Compared with the area where the

rapidest growth of AAHP is the emerging urban center in Guangzhou, Shenzhen, and Beijing, the highest increase of the AAHP in Shanghai is highlighted in Putuo and Zhabei Districts, which are the urban historical center (see Figure 6(c)). Furthermore, it is similar to Guangzhou that the river crossing the center of the city; this provides an excellent landscape and plays the significant influencing on the rise of housing prices.

According to the comparative analysis of the spatial patterns of housing prices increasing in the four major cities of China, it can be found that the urban spatial development, urban infrastructure layout, urban planning, and land use

should play a significant role in the changes of housing prices. From this perspective, this paper further takes 141 neighborhoods in Guangzhou as area units to further explore the spatial correlation of housing prices within the city, and then quantify the impacting factors in the context of urban spatial development, infrastructural layout, planning, and land use, to characterize the spatially nonstationarity and heterogeneous characteristics of the housing prices in the city of Guangzhou.

### 3.2. Spatial Autocorrelation Analysis of Housing Prices in Guangzhou

**3.2.1. Global Spatial Autocorrelation Analysis.** Using the global Moran's I index analysis tool of ArcGIS 10.2, the global spatial autocorrelation analysis for the average annual housing price (AAHP) in 141 neighborhoods of Guangzhou in 2009–2015 is achieved, in which the first-order adjacency of polygon is adopted as the spatial relationship criterion.

As shown in Table 3, in 2009–2015, Moran's I index is more than 0. This illustrates that the AAHP in Guangzhou presents the positive spatial autocorrelation characteristic and demonstrates a significant statistical phenomenon of spatial clustering ( $Z > 1.96$ ,  $P < 0.001$ ), that is, when the AAHP of a neighborhood is high, that of its surrounding neighborhoods is correspondingly high, and vice versa.

**3.2.2. Local Spatial Autocorrelation Analysis.** Although the global spatial autocorrelation analysis can better detect the spatial clustering characteristics of the AAHP in Guangzhou with the spatial aggregation patterns of similar values of observation points (positive correlation) or nonsimilarity values (negative correlation), the spatial homogeneity, i.e., whether the AAHP is a high-value cluster or a low-value cluster cannot be explored clearly. Therefore, this paper further applies the local spatial correlation analysis tool, i.e., the local indicator of spatial association (LISA), to provide a better understanding of the spatial homogeneity of the AAHP in Guangzhou, as shown in Figure 7. In Figure 7(a), the high-high (HH) clustering area is mainly concentrated in Yuexiu District (i.e., the historical urban center), the Central Business District (i.e., the urban economic center), the Pearl River coast in Haizhu District, and the Baiyun New Town (BNT). Furthermore, the HH clustering area is surrounded by the low-high (LH) clustering area, but no high-low (HL) clustering area is found (see Figure 7(b)). This implies that the high housing prices have a certain effect on spatial spillover. The low-low (LL) clustering area is concentrated in the northern mountainous area and the southern region. There is no significant clustering characteristic in the East, indicating that the housing prices of the neighborhoods are more homogeneous in this region.

### 3.3. Geographically Weighted Analysis of the Housing Prices in Guangzhou

**3.3.1. Linear Regression Analysis of the Influencing Factors of Housing Prices.** To examine the global (average) influence of

the four factors of “location condition,” “transport accessibility,” “commercial service intensity,” and “public service intensity” on the formation of housing price, a linear regression model is used. In the regression model, the quantitative indexes of the factors are set as the explanatory variables (independent variables), i.e.,  $K_i^D$ ,  $K_i^T$ ,  $K_i^B$ , and  $K_i^P$ , and the AAHP of each neighborhood in 2015 is the explanatory variable (dependent variable), and the results are illustrated in Table 4.

As shown in Table 4, the value of Adjusted  $R^2$  is 0.60, which indicates how well the regression line fits the observations. Furthermore, it can be found that the values of P in the observations of  $K_i^T$  and  $K_i^P$  are much higher than 0.05, which implies the poor linear fitness, and especially reveals nonsignificant impacts of the transportation accessibility ( $K_i^T$ ) and public service intensity ( $K_i^P$ ) on the housing prices in the context of linear regression analysis. As shown in Figures 8(a) and 8(b), the Line Fit Plots of  $K_i^T$  and  $K_i^P$  to AAHP in neighborhoods show a strong log-fitting trend, respectively. For the factor of commercial service intensity, the value of P in the observations of  $K_i^B$  is lower than 0.05, but close to 0.01. Although it reflects the significant impact of the commercial service intensity on the house prices, Figure 8(c) shows that the Line Fit Plot of  $K_i^B$  to AAHP in neighborhoods presents the extremely significant log curve fitting instead of linear fitting. Figure 8 reveals that when the urban infrastructure development level of the neighborhood is lower, the influence of the factors oriented towards the perspective of urban infrastructure layout (i.e., transportation accessibility, public service intensity, and commercial service intensity) on the housing prices is more significant, that is, with the urban infrastructure development, the influence is getting weaker. This demonstrates that regional urban infrastructure development differences have varying degrees of impact on housing prices, further revealing that housing prices are spatially heterogeneous and nonstationarity. It is a fact that, with the intense urban spatial expansion, a large number of residential areas are bound to extend outward from the urban core into the city of Guangzhou, leading to the development of urban infrastructure lagging the spread of urban areas. Furthermore, the scarcity and uneven distribution of high-quality education resources, which are highly concentrated in the center of the city of Guangzhou, causes strong impacts on the AAHP in the developing neighborhoods with poor educational infrastructure, but causes weak impacts on the developed neighborhoods with complete educational infrastructure. Especially, in the Chinese traditional concept, the proximity of a house to a medical institution such as the hospital is not a favorable condition, and it even has a negative impact on the housing prices.

In Table 4, the factor of geographical location condition (i.e., the distance away from the CBD) presents the extreme significant impact on the housing prices, in terms of the value of P in cis much lower than 0.01, as well as the significant linear fitness on  $K_i^D$  to the AAHP of neighborhoods (see Figure 9). Specifically, every 1 km decrease in the distance between the neighborhood and the Central Business District (CBD), leading to its housing price to rise by about

TABLE 3: Results of the global Moran's I index analysis.

Value	Year						
	2009	2010	2011	2012	2013	2014	2015
Moran's I index	0.195145	0.365298	0.268577	0.288585	0.272338	0.346694	0.549534
Exp. index	-0.009901	-0.009901	-0.009901	-0.009901	-0.007143	-0.007143	-0.007143
Variance	0.004628	0.004643	0.004591	0.004643	0.002473	0.002483	0.002465
Z-score	3.014119	5.506398	4.109836	4.380566	5.620302	7.101614	11.211547
P value	0.002577	0.000000	0.000040	0.000012	0.000000	0.000000	0.000000

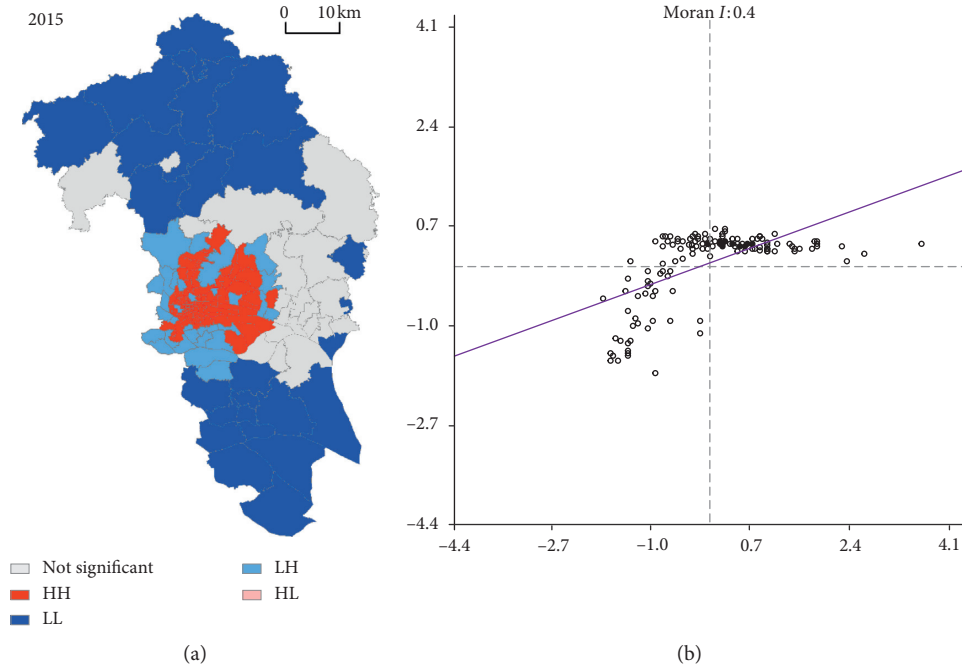


FIGURE 7: Local spatial autocorrelation analysis of housing price (2015). (a) LISA aggregation chart. (b) Moran scatter chart.

TABLE 4: Results of regression model analysis.

	Coefficients	Standard error	t-Stat	P value
Intercept	30388.85677	1543.939504	19.68267325*	1.25352E-41
$K_i^D$	-636.54367	75.18006459	-8.466921031*	3.59862E-14
$K_i^T$	219.1213257	157.9274463	1.387480966*	0.167564468
$K_i^P$	6.55650564	125.9266689	0.052066061*	0.958552453
$K_i^B$	346.8583018	135.4364537	2.561040932*	0.01152663
Adj. $R^2$			<b>0.60</b>	
Significance F			<b>2.03E-26</b>	

\*Significance at the level of 0.05.

636 yuan (about 90 dollars) per square meter in the neighborhood.

According to the linear regression analysis, it can be found that housing prices have obvious spatial differentiation and spatial nonlinear characteristics. Therefore, in the following content in this paper, the geographically weighted regression (GWR) model is further applied to more clearly and accurately interpret and characterize the spatial heterogeneity and nonstationarity of housing prices.

**3.3.2. Geographically Weighted Analysis of the Influencing Factors of Housing Prices.** In the environment of ArcGIS

10.2, the geographically weighted regression model is utilized to characterize the spatial homogeneity and nonstationarity of the housing prices with the AIC identified as the criterion for bandwidth optimization. In the GWR mode, the AAHPs of neighborhoods in Guangzhou in 2015 are identified as the dependent variables, and the independent variables being the observations in  $K_i^D$ ,  $K_i^T$ ,  $K_i^B$ , and  $K_i^P$ . The result shows that the value of  $R^2$  is 0.63, which is higher than the fitness of the traditional linear regression analysis (0.60). The residuals of the results derived from the global spatial autocorrelation analysis in the GWR model illustrate that Moran's I

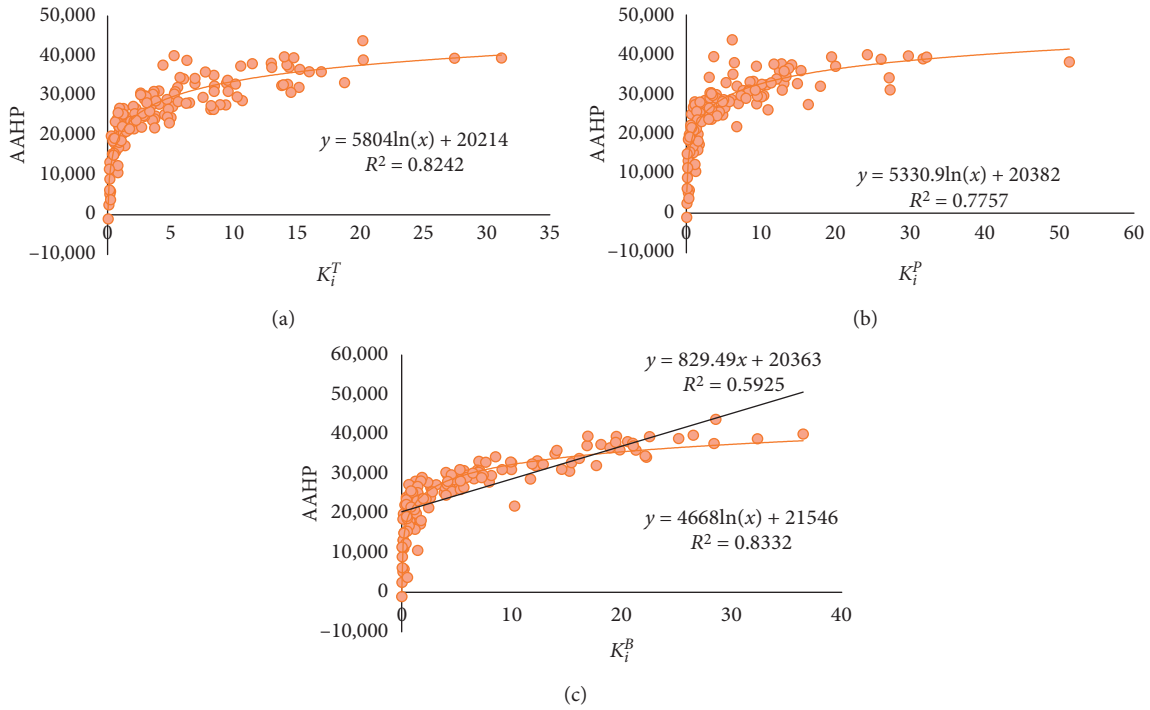


FIGURE 8: The Line Fit Plots of  $K_i^T$ ,  $K_i^P$ , and  $K_i^B$ . (a) Transportation accessibility. (b) Public service intensity. (c) Commercial service intensity.

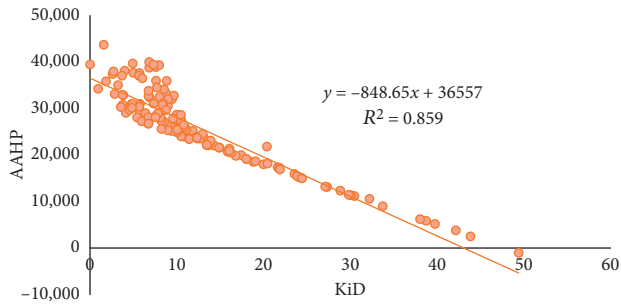


FIGURE 9: The line fit plots of  $K_i^D$ .

index is 0.03, the z-score is 0.5, and the  $P$  value is 0.4, indicating that the residuals exhibit spatial random distribution characteristics.

To clarify the impact of  $K_i^D$ ,  $K_i^T$ ,  $K_i^B$ , and  $K_i^P$  on the spatial nonstationarity of the housing prices in Guangzhou, Figure 10 illustrates, respectively, the equivalent partition based on the geographically weighted regression coefficient. In Figure 10(a), the distance away from the CBD has a significant effect on the housing prices in the eastern and southern neighborhoods, while the weakest impact on the housing prices is in Huadu District because of the longest distance between its neighborhoods and the CBD. Although the districts of Yuexiu, Liwan, and Haizhu in the west are adjacent to the CBD, their AAHP is rather less affected than that in Huangpu District and Panyu District. The reason is that the districts of Yuexiu, Liwan, and Haizhu are urban developed areas, in which the housing price is more affected by their own factors, such as the favorable location and well-developed urban infrastructure. As shown in Figure 10(b), it

is revealed that the transportation infrastructural development in Yuexiu, Liwan, and Haizhu Districts is flawless, leading to the highest level of transportation accessibility which plays the significant effect on the housing prices. Furthermore, the housing prices in the outstanding urban functional areas with the well-developed transportation infrastructure, such as the CBD, the BNT, and the University Town in Guangzhou, are highlighted by the significant influence of the factor on transport accessibility. Figure 10(c) demonstrates that the spatial distribution characteristic of the geographically weighted regression coefficient of  $K_i^B$  is similar to  $K_i^T$ . Moreover, the influence of the factor on commercial service intensity radiates outward from the urban core and shows the attenuation law based on the increasing of distance. As shown in Figure 10(d), the influence of the public service intensity on housing prices presents an interesting spatial heterogeneity and non-stationarity characteristic, that is, the east is strong and the west is weak. This phenomenon further characterizes the imbalance in the public service resources, that is, more resources are concentrated in the west (i.e., urban core) and less in the east. As the public service resources in the east are scarcer, its housing prices is more significantly affected by the public service intensity.

#### 4. Conclusion and Discussion

The consistent booming growth of the housing market in China highlights the urgent need to examine and detect how the intense urban spatial expansion, large-scale urban infrastructural development, and fast-changing urban planning determine and characterize the changes and spatial

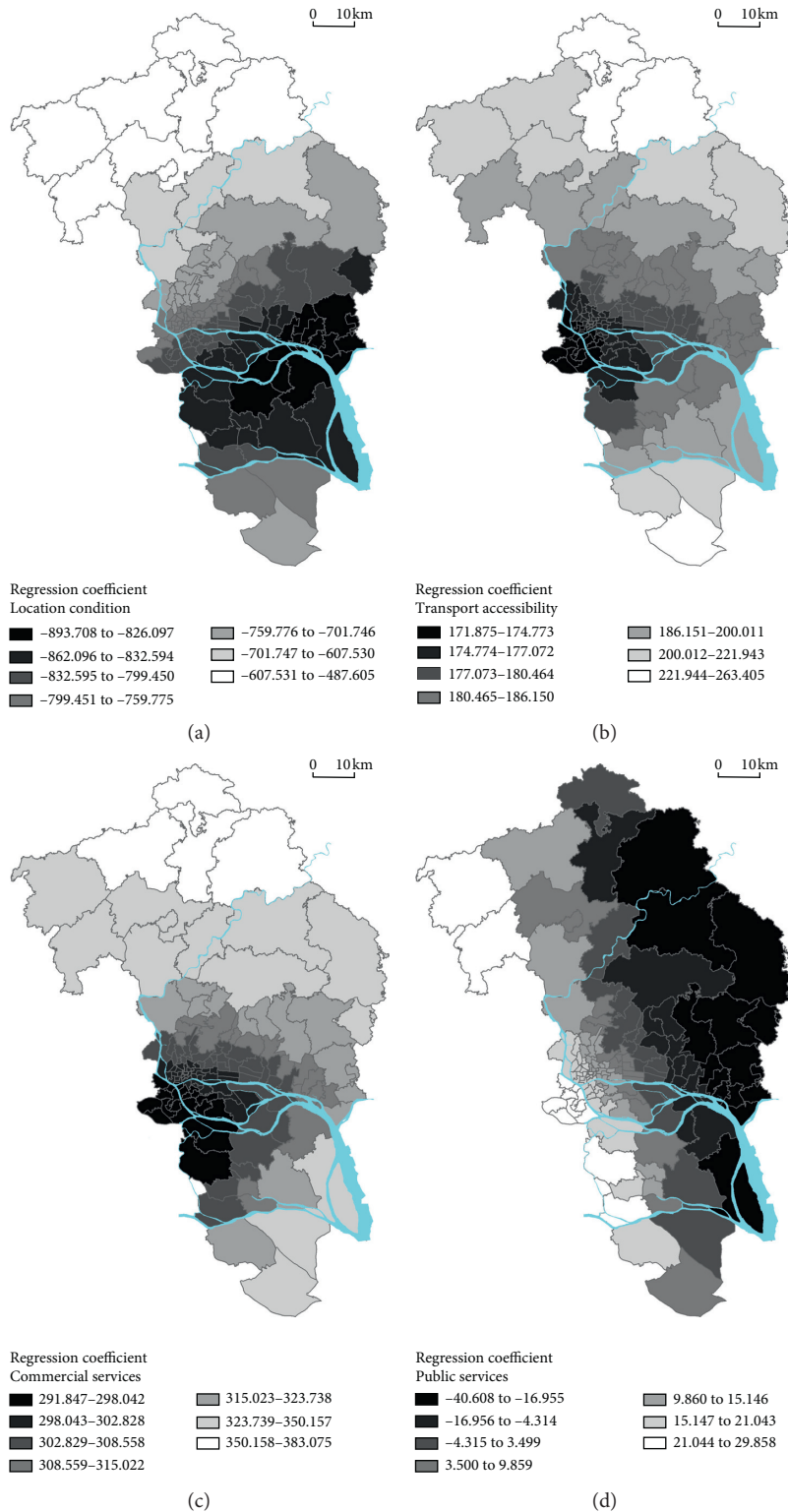


FIGURE 10: Spatial differences of geographically weighted regression coefficients. (a)  $K_i^D$ . (b)  $K_i^T$ . (c)  $K_i^B$ . (d)  $K_i^P$ .

differentiation of housing prices. This should be of great significance to promote the governing policy and spatial controlling mechanism of housing markets and the reform of the real-estate system and related urban systems, such as the land-use system. At the level of city, therefore, this paper

implements an empirical analysis with the using of the traditional econometric models of regressive analysis and GIS-based spatial autocorrelation analysis tools, focusing in exploring and characterizing the spatial homogeneity and nonstationarity of the housing prices in 2009–2015 in the



context of the neighborhoods in Guangzhou, China. In Guangzhou, there are total 141 neighborhoods identified as area units, and their average annual housing prices (AAHP) in 2009–2015 are represented as dependent variables. Simultaneously, the factors including geographical location condition, transportation accessibility, business service intensity, and public service intensity, which are identified in the context of urban development and planning, are defined as independent variables, leading to explore the spatial correlation between the AAHP in neighborhoods (area units). Furthermore, the quantitative analysis models, including multiple linear regression models, spatial autocorrelation analysis models, and geographically weighted regression (GWR) models in the environment of ArcGIS 10.2 are integrated and applied.

Firstly, in the environment of GIS (ArcGIS), the Kriging spatial interpolation method is used to reveal the spatio-temporal evolution of the average annual housing prices (AAHP) in 2009–2015, especially with the comparative analysis of the spatial patterns of housing prices increasing in the major cities of China, including Beijing, Shanghai, Guangzhou, and Shenzhen. It aims to reveal that the urban spatial expansion, urban infrastructure development, urban planning, geographical location, transportation accessibility, and land use have significant roles in the changes of housing prices. Furthermore, the global and local spatial autocorrelation models are used to explore the spatial clustering characteristics of the AAHP in the city of Guangzhou in 2015, integrating with the traditional linear regression model and geographically weighted regression model (GWR model). Finally, the in-depth investigation and discussion of the influencing factors on the housing prices in Guangzhou is achieved by characterizing the spatial heterogeneity and nonstationarity of the housing prices caused by these factors.

The specific results derived from this study show that (1) the temporal and spatial evolution of the AAHP in Guangzhou shows the circle characteristic with the center of the urban core; (2) there are obvious spatial differences in the growth of AAHP in Guangzhou, which is closely related to the urban planning and the spatial pattern of urban functional area; (3) the global spatial autocorrelation analysis reveals that the housing prices has significant spatial aggregation, and the local spatial autocorrelation analysis further characterizes the spatial homogeneity in the aggregation which highlights the critical characteristic of the high aggregation in the urban core; however, no area with a high-low aggregation is found, indicating that the housing price has a spatial spillover effect; (4) the analysis of the traditional linear regression model illustrates that when the urban infrastructure development level of neighborhood is lower, the influence of the factors oriented towards a perspective of urban infrastructure layout (i.e., transportation accessibility, public service intensity, and commercial service intensity) on the housing prices is more significant, that is, with the urban infrastructure development, the influence is getting weaker; (5) the factor of geographical location (i.e., the distance away from the CBD) presents the extreme significant impact on the housing prices; (6) the analysis based on the geographically weighted regression model

further illustrates the specific effect of each factor on the spatial heterogeneity and nonstationarity of the housing prices, that is, the spatial pattern of the regression coefficients of  $K_i^D$  and  $K_i^P$  shows “the east is strong and the west is weak,” while that of the regression coefficients of  $K_i^T$  and  $K_i^B$  is “the west is strong and the east is weak;” moreover, the spatial heterogeneity and nonstationarity of the housing prices demonstrates a ring structure with the center of urban core and the decreasing law with the increasing of distance.

All the abovementioned results can better reflect the elementary spatial characteristics and influencing factors of the housing prices within Guangzhou. The contribution of our study is to examine and characterize what and how to determine the spatial homogeneity and nonstationarity of housing prices oriented towards a microscale geographical perspective, i.e., in the context of neighborhoods, aiming to promote the governing policy and spatial controlling mechanism of the housing markets and the reform of real-estate and land-use systems in Guangzhou, China. The empirical analysis supports a deep knowledge of the spatial heterogeneity and nonstationarity of housing prices determined under the spatial differences in urban spatial development, infrastructural layout, land use, and urban planning. This is significant to drive significant changes to develop a more efficient, sustainable, and competitive governing policy and spatial controlling mechanism for the real-estate system and land-use system at the level of city, especially in the megacities in China.

However, the selection and processing of quantitative indicators are still in exploratory, and the existing research literature does not have a sound basis and solution. Therefore, the fitness of the regression models is not very satisfactory (around 6.0), which may cause deviations in the analysis of influencing factors. Therefore, the future work of this paper needs to do more in-depth analysis and investigation on the selection and quantification of related indicators and compare and analyze more cities to improve the current study.

## Data Availability

The data used to support the findings of this study are available from the first 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 was supported by the Guangdong Natural Science Fund (2015A030313626).

## References

- [1] M. P. Larkin, Z. Askarov, H. Doucouliagos et al., “Do house prices ride the wave of immigration?” *Journal of Housing Economics*, vol. 46, pp. 101–630, 2019.

- [2] C. Day, "House price post-GFC: more household debt for longer," Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University, Canberra, Australia, CAMA Working Papers 2019-52, 2019.
- [3] M. Luciani, "Monetary policy and the housing market: a structural factor analysis," *Journal of Applied Econometrics*, vol. 30, no. 2, pp. 199–218, 2013.
- [4] G. Favara and J. Imbs, "Credit supply and the price of housing," *American Economic Review*, vol. 105, no. 3, pp. 958–992, 2015.
- [5] J. Vandebussche, U. Vogel, and E. Detragiache, "Macroeconomic policies and housing prices: a new database and empirical evidence for central, eastern, and southeastern Europe," *Journal of Money, Credit and Banking*, vol. 47, no. 1, pp. 333–375, 2015.
- [6] C. A. L. Hilber and W. Vermeulen, "The impact of supply constraints on house prices in England," *The Economic Journal*, vol. 126, no. 591, pp. 358–405, 2014.
- [7] Z. Du and L. Zhang, "Home-purchase restriction, property tax and housing price in China: a counterfactual analysis," *Journal of Econometrics*, vol. 188, no. 2, pp. 558–568, 2015.
- [8] S. C. Bourassa, E. Cantoni, and M. Hoesli, "Spatial dependence, housing submarkets, and house price prediction," *Journal Real Estate Finance and Economics*, vol. 35, no. 2, pp. 143–160, 2007.
- [9] C. Leishman, "Spatial change and the structure of urban housing sub-markets," *House Study*, vol. 24, no. 5, pp. 563–585, 2009.
- [10] S. Basu and T. G. Thibodeau, "Analysis of spatial autocorrelation in house prices," *Journal of Real Estate Finance and Economics*, vol. 17, no. 1, pp. 61–85, 1998.
- [11] A. Can, "The measurement of neighborhood dynamics in urban house prices," *Economic Geography*, vol. 66, no. 3, pp. 254–272, 1990.
- [12] A. Barreca, R. Curto, and D. Rolando, "Urban vibrancy: an emerging factor that spatially influences the real estate market," *Sustainability*, vol. 12, no. 1, p. 346, 2020.
- [13] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.
- [14] S. Stevenson, "New empirical evidence on heteroscedasticity in hedonic housing models," *Journal of Housing Economics*, vol. 13, no. 2, pp. 136–153, 2004.
- [15] G. S. Sirmans, D. A. Macpherson, and E. N. Zietz, "The composition of hedonic pricing models," *Journal of Real Estate Literature*, vol. 13, no. 1, pp. 3–43, 2005.
- [16] R. Y. C. Tse, "Estimating neighborhood effects in house prices: towards a new hedonic model approach," *Urban Studies*, vol. 39, no. 7, pp. 1165–1180, 2002.
- [17] R. G. Michaels and V. K. Smith, "Market segmentation and valuing amenities with hedonic models: the case of hazardous waste sites," *Journal of Urban Economics*, vol. 28, no. 2, pp. 223–242, 1990.
- [18] B. Keskin, R. Dunning, and C. Watkins, "Modelling the impact of earthquake activity on real estate values: a multi-level approach," *Journal of European Real Estate Research*, vol. 10, no. 1, pp. 73–90, 2017.
- [19] M. H. Pesaran, Y. Shin, and R. Smith, "Bounds testing approaches to the analysis of level relationships," *Journal of Applied Econometrics*, vol. 16, no. 3, pp. 289–326, 2001.
- [20] A. Larm, M. Knoppel, J. Haan et al., "Amsterdam house price ripple effects in the Netherlands," *Journal of European Real Estate Research*, vol. 10, no. 3, pp. 331–345, 2017.
- [21] L. Yao, G. Gu, and J. K. Wang, "The spatial effect of building new housing in Zhengzhou City based on the spatial econometric model," *Chinese Journal of Economic Geography*, vol. 34, no. 1, pp. 69–74, 2014.
- [22] J. M. Le, "Influence factors analysis of housing price in Nanchang based on hedonic model," *Chinese Journal of East China Jiaotong University*, vol. 32, no. 5, pp. 128–135, 1990.
- [23] S. X. Ma and A. Li, "House price and its determinations in Beijing based on hedonic model," *Chinese Journal of Civil Engineering*, vol. 36, no. 9, pp. 59–64, 2003.
- [24] B. Lu, M. Charlton, P. Harris, and A. S. Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 660–681, 2014.
- [25] G. T. Smerth and M. T. Smith, "Accessibility changes and urban house price appreciation: a constrained optimization approach to determining distance effects," *Journal of Housing Economics*, vol. 9, no. 3, pp. 187–196, 2000.
- [26] G. Lee, D. Cho, and K. Kim, "The modifiable areal unit problem in hedonic house-price models," *Urban Geography*, vol. 37, no. 2, pp. 223–245, 2016.
- [27] R. Crespo, A. S. Fotheringham, and M. Charlton, "Application of geographically weighted regression to a 19-year set of house price data in London to calibrate local hedonic price models," in *Proceedings of the 9th International Conference on Geocomputation*, National University of Ireland, Maynooth, Ireland, September 2007.
- [28] M. Helbich, W. Brunauer, E. Vaz, and P. Nijkamp, "Spatial heterogeneity in hedonic house price models: the case of Austria," *Urban Studies*, vol. 51, no. 2, pp. 390–411, 2014.
- [29] C. Brunson, S. Fotheringham, and M. Charlton, "Geographically weighted regression," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 3, pp. 431–443, 1998.
- [30] J. Tu and Z. Xia, "Examining spatially varying relationships between land use and water quality using geographically weighted regression I: model design and evaluation," *Science of the Total Environment*, vol. 407, no. 1, pp. 358–378, 2008.
- [31] X. Yang, C. Webster, and S. Orford, "Identifying house price effects of changes in urban street configuration: an empirical study in Nanjing, China," *Urban Studies*, vol. 53, no. 1, pp. 112–131, 2014.
- [32] F. Wang, X. L. Gao, and B. Q. Yan, "Research on urban spatial structure in Beijing based on housing prices," *Chinese Journal of Progress in Geography*, vol. 33, no. 10, pp. 128–133, 2014.
- [33] L. F. Zhang, L. J. Pu, J. Zhang et al., "Analysis on the space frame of city land price based on the hedonic model: take Ludi city of Hunan province as example," *Chinese Journal of Economic Geography*, vol. 29, no. 9, pp. 1322–1331, 2009.
- [34] H. Y. Zhong, A. L. Zhang, and Y. Y. Cai, "Impacts of the Nanhu lake in Wuhan City on the price of peripheral houses: empirical research based on Hedonic model," *China Land Sciences*, vol. 23, no. 2, pp. 63–68, 2009.
- [35] J. Li, X. J. Yang, and T. Su, "Residential spatial variation characteristics and regional value: a case of 2011 opened commercial residential buildings in Xi'an City," *Chinese Journal of Arid Land Geography*, vol. 37, no. 1, pp. 170–178, 2014.
- [36] Q. Y. Tang, W. Xu, and F. L. Ai, "A GWR-based study on spatial pattern and structural determinants of Shanghai's housing price," *Chinese Journal of Economic Geography*, vol. 32, no. 2, pp. 52–58, 2012.

- [37] Z. Li, S. L. Zhou, H. F. Zhang et al., "Exploring the factors impacting on the residential land price and measuring their marginal effects based on geographically weighted regression model: a case study of Nanjing," *China Land Sciences*, vol. 23, no. 10, pp. 20–25, 2009.
- [38] P. Lv and H. Zhen, "Affecting factor research of Beijing residential land price based on GWR model," *Chinese Journal of Economic Geography*, vol. 30, no. 3, pp. 472–478, 2010.
- [39] Y. L. Gong, X. L. Zhang, and L. L. Zhang, "Spatial autocorrelation of urban land price: a case study of Suzhou," *Chinese Journal of Economic Geography*, vol. 31, no. 11, pp. 1906–1911, 2011.
- [40] 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.
- [41] Z. X. Mei and X. Li, "Spatial analysis of house's price in Dongguan based on ESDA and Kriging techniques," *Chinese Journal of Economic Geography*, vol. 28, no. 5, pp. 862–866, 2008.
- [42] H. Wang, "Real estate price impact factors analysis based on spatial econometrics," *Chinese Journal of Economic Review*, vol. 1, pp. 48–56, 2012.
- [43] K. Sun and Z. M. Xu, "The impacts of human driving factors on grey water footprint in China using a GWR model," *Chinese Journal of Geographical Research*, vol. 35, no. 1, pp. 37–48, 2016.
- [44] S. P. Chen, J. J. Tan, C. Claramunt, and C. Ray, "Multi-scale and multi-modal transport data model," *Journal of Transport Geography*, vol. 19, no. 1, pp. 147–161, 2011.
- [45] S. Chen, C. Claramunt, and C. Ray, "A spatio-temporal modelling approach for the study of the connectivity and accessibility of the Guangzhou metropolitan network," *Journal of Transport Geography*, vol. 36, pp. 12–23, 2014.
- [46] S. P. Chen, J. Yang, Y. Li et al., "Multiconstrained network intensive vehicle routing adaptive ant colony algorithm in the context of neural network analysis," *Complexity*, vol. 2017, Article ID 8594792, 9 pages, 2017.
- [47] T. L. Saaty, *The Analytic Hierarchy Process*, McGrawHill, New York, NY, USA, 1980.
- [48] F. J. Carmone, A. Kara, and S. H. Zanakis, "A Monte Carlo investigation of incomplete pairwise comparison matrices in AHP," *European Journal of Operational Research*, vol. 102, no. 3, pp. 538–553, 1997.
- [49] N. Hovanov, J. Kolari, and M. V. Sololov, "Deriving weights from general pairwise comparison matrices," *Mathematical Social Sciences*, vol. 55, no. 2, pp. 205–220, 2008.
- [50] J. Tian, F. Q. Xiong, X. P. Cheng et al., "Road density partition and its application in evaluation of road selection," *Geomatics and Information Science of Wuhan University*, vol. 41, no. 9, pp. 1225–1231, 2016.
- [51] L. Anselin, "The local indicators of spatial association-LISA," *Geographical Analysis*, vol. 27, pp. 93–115, 1995.

## Research Article

# A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems

José-Luis Alfaro-Navarro <sup>1</sup>, Emilio L. Cano <sup>2</sup>, Esteban Alfaro-Cortés <sup>2</sup>,  
Noelia García <sup>1</sup>, Matías Gámez <sup>2</sup> and Beatriz Larraz <sup>2</sup>

<sup>1</sup>Faculty of Economics and Business Administration, University of Castilla-La Mancha, Albacete, Spain

<sup>2</sup>Quantitative Methods and Socio-Economic Development Group, Institute for Regional Development (IDR), University of Castilla-La Mancha (UCLM), Albacete, Spain

Correspondence should be addressed to Beatriz Larraz; [beatriz.larraz@uclm.es](mailto:beatriz.larraz@uclm.es)

Received 12 November 2019; Accepted 19 March 2020; Published 14 April 2020

Guest Editor: Marco Locurcio

Copyright © 2020 José-Luis Alfaro-Navarro 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.

The close relationship between collateral value and bank stability has led to a considerable need to a rapid and economical appraisal of real estate. The greater availability of information related to housing stock has prompted to the use of so-called big data and machine learning in the estimation of property prices. Although this methodology has already been applied to the real estate market to identify which variables influence dwelling prices, its use for estimating the price of properties is not so frequent. The application of this methodology has become more sophisticated over time, from applying simple methods to using the so-called ensemble methods and, while the estimation capacity has improved, it has only been applied to specific geographical areas. The main contribution of this article lies in developing an application for the entire Spanish market that fully automatically provides the best model for each municipality. Real estate property prices in 433 municipalities are estimated from a sample of 790,631 dwellings, using different ensemble methods based on decision trees such as bagging, boosting, and random forest. The results for estimating the price of dwellings show a good performance of the techniques developed, in terms of the error measures, with the best results being achieved using the techniques of bagging and random forest.

## 1. Introduction

Since the year 2008, the global economic crisis caused a slowdown in the economy which resulted in a decrease in the price of real estate properties. The appraised valuation of a property is considered key for any transaction related to the property and particularly for its sale or for a mortgage application, so it is essential that the price is a true reflection of its value. Banks also need to periodically review the value of their real estate portfolio by updating their appraised valuation, see the Basel II International Banking Agreement [1, 2]. Normally, appraisals for the purpose of a mortgage are carried out by professional appraisers visiting the property. However, the appraisal procedure developed in this way is expensive, both in terms of time and money, and makes this procedure unsustainable for the valuation of large real estate

portfolios. Furthermore, although the physical presence of an appraiser may help to give a more accurate valuation of the property, there is also the possibility of bias from interested parties, such as buyers, sellers, or banks themselves, which may make the valuation more subjective. There is clearly a need for the development of a prediction model which can present unbiased, realistic valuations.

The International Association of Assessing Officers (IAAO) considers mass appraisal as the process of valuing a group of properties using common data, standardized methods, and statistical procedures [3]. These valuation methods have been implemented through models known as Automated Valuation Models (AVM) and have enabled the appraisal of large real estate portfolios without the direct intervention of an appraiser [4, 5]. The development of these estimation procedures has been enhanced by the growth in

the quantity and quality of information related to both real estate prices and property characteristics, accessible to researchers. This level of information allows the application of increasingly sophisticated statistical techniques for the development of estimation procedures of a higher quality and precision. Real estate AVMs allow the valuation of property prices en masse, without the need for the physical presence of an appraiser, by using computer-assisted task appraisal systems [6]. In many cases, the presence of an appraiser is only necessary for those valuations considered to be out of the ordinary [7].

Estimation techniques include parametric regression analysis [8] and nonparametric [9] or machine learning methods such as neural networks [10, 11], decision trees [12, 13], random forests [14, 15], fuzzy logic [16], or ensemble methods [17]. These techniques are used primarily with three objectives in mind: to estimate a real estate property price, to find out the influence of a characteristic of the house on its price, and to create a hedonic price index.

Over the recent decades, the most commonly used procedures has been based on hedonic-based regression [18, 19]. However, these models present certain fundamental problems related to the assumptions of the model: normality of the residuals, homoscedasticity, independence, and the absence of multicollinearity. This situation has led to greater use of pattern recognition techniques, often known as data mining techniques, which include machine learning. These techniques are more flexible about the assumptions related to the distribution of data, they are easier to interpret, and they allow linear and nonlinear relationships to be analysed. In addition, they enable both categorical and continuous variables to be managed [13]. Although these techniques were initially used more as classification methods, in recent years, their application has been used in determining the most influential variables on house pricing and in estimating dwelling prices. Pérez-Rave et al. [20] provide a two-stage methodology for the analysis of big data regression under a machine learning approach for both inferential and predictive purposes.

Accurate and efficient prediction of real estate prices has been and will continue to be an essential but controversial issue, with an impact on the various actors in the economy such as buyers, sellers, commission agents, governments, and banks [21, 22]. Nowadays, the big data paradigm offers exciting possibilities for more accurate predictions and one of the main approaches for dealing with big data is machine learning.

These machine learning methods have been applied to the estimation of real estate properties in very specific locations. Research has been carried out by Jaen [12] in Coral Gables (Florida, USA); Fan et al. [13] in Singapore (Republic of Singapore); Özsoy and Şahin [23] in Istanbul (Turkey); Del Cacho [14] in Madrid (Spain); Pow et al. [17] in Montreal (Canada); Ceh et al. [24] in Ljubljana, (Slovenia); Nguyen [25] in 5 counties of USA; and Dimopoulos et al. [26] in Nicosia (Cyprus). In contrast, Pérez-Rave et al. [20] deal with the estimation of dwelling prices for a whole country, Colombia, using independent variables identifying the city in which each property is located, thus proposing a

unique model for the entire country (with a sample of 61,826 properties).

The main new element of this article is that it proposes a new methodology to carry out the automated estimation of real estate prices for an entire country (Spain in this case study), specifying automatically a different model for each municipality, with a sample size of 790,631 real estate properties. The whole country can be considered as a complex real estate system due to the great differences that exist between rural and urban areas and even inside the urban areas. Each municipality is trained with the information available in its training set so that each one will have its own model adapted to its characteristics and needs. This study addresses a program capable of covering a different model for each of the 433 municipalities with more than 100 properties for sale, with a population ranging from 1,559 inhabitants in the smallest municipality to 3,223,334 in Madrid, the biggest one. We focus on the application of this automated valuation system based on machine learning methods in estimating real estate prices and analyse the accuracy of each of them using error measures widely recognised in economic literature. As a rule, to evaluate model quality, aggregated diagnostic indicators are used (coefficient of determination) although there are few contributions in the relevant literature where the quality of the procedure is analysed using a measurement of the estimation error [15]. In this article, we use four measures to analyse the validity of the proposed methods, namely, the mean ratio, the mean absolute percentage error (MAPE), the median absolute percentage error (MdAPE), and the coefficient of dispersion (COD).

In this article, machine learning techniques used in estimating the price of dwellings are based on the decision tree technique. However, in general, it is difficult to build a single tree to make predictions because of incorrect parameter settings, simplicity rules, and tree instability. To overcome these problems and obtain better behaviour when making predictions, techniques of ensemble of decision trees have been developed, such as bagging, boosting, and random methods [27]. In bagging, the models are fitted using random independent bootstrap replicates that are then combined by averaging the output for regression [28]. In boosting, the fitted model is a simple linear combination of many trees that are fitted iteratively and boosted to reweight poorly modelled observations [29]. The random forest model, however, is constructed in a random vector of the data feature space sampled independently [30]. Starting from this base, we automatically design the best ensemble method, including bagging, boosted regression tree, and random forest in each municipality and then make a comparison to analyse their behaviour under different circumstances. In addition, the results obtained with a single decision tree are included in order to analyse and compare the benefit of using an ensemble of decision tree techniques.

A further consideration is the particular emphasis that specialized literature places on the need to include spatial information in hedonic models, given the significant influence that the location of the property can have on its price and therefore on its valuation. Ceh et al. [24] highlight the

growing interest in recent years in applying spatial statistics to hedonic price modeling, besides coupling the geographic information system and machine learning techniques. In this article, we include geographical coordinates in the explanatory variables to draw attention to the importance of the property location when valuing a property.

The article is laid out as follows. After the introduction, Section 2 presents a literature review highlighting the main research to date on the valuation of real estate property prices from the use of regression trees to tree ensemble models. In Section 3, the methodology used is presented with a description of the main techniques as well as the valuation measures of behaviour of the different models. In Section 4, the empirical argument for an application to the whole of Spain is developed and advocates the need for using ensemble methods for Spain. Finally, Section 5 presents the main conclusions and further lines of research.

## 2. Literature Review

The application of machine learning methods in the field of estimating property prices has attracted interest for some years now. However, the application of decision trees is relatively recent, initially being used as a classification technique and for determining which variables had the greatest influence on the price of housing. The application of decision trees was then used as a prediction technique through the so-called regression tree to obtain dwelling price predictions. One of the first proposals for the application of a regression tree was made by Jaen [12] who used information from 15 variables for 1,229 transactions in the city of Coral Gables (Florida) taken from the multiple listing system (MLS). Jaen [12] tests the effectiveness of using stepwise regression, CART decision tree, and neural networks in estimating the price of housing and in determining the most important variables for this prediction. The best results are achieved from CART measuring the estimation capacity with the mean absolute error (MAE), using a smaller number of variables, specifically five versus the nine used in stepwise regression.

Following on from [12], Fan et al. [13] demonstrate the good behaviour of regression trees, using the CART algorithm to identify the main determinants and predict the price of housing. This application is developed for the Singapore resale public housing market. However, although the process used to identify the main variables that affect the price of housing is extensive, its estimation is based solely on the average value in a leaf node of the tree, this value being thought of as a forecasting value or regression value. Özsoy and Şahin [23] develop a CART application in Turkey to determine the most influential characteristics on the price of housing in Istanbul based on a database taken from the Internet in 2007. The results lead them to conclude that the size of the house and the existence of an elevator, security, central heating, and views are the most influential variables on house prices in Istanbul.

Kok et al. [31] show the main advantages of the application of regression trees to predict property prices, taking into account that these models help overcome the problem

of regression models in nonlinear relationships. The advantages highlighted by the authors are they are simple to understand and interpret, and their statistical significance is easy to calculate; they can handle categorical variables without creating dummy variables; and they consume little computing time even with large amounts of data. In addition, the authors propose the use of a procedure called stochastic boosting which allows an unlimited number of variables to be handled with good results, including economic and demographic variables and hyperlocal metrics in the prediction model. The limitations of regression trees are they can show unlimited growth vertically until the sample has an observation which may generate models with poor generalization capacity; they are not robust to changes in the training set; and they usually suffer an underfitting effect giving rise to models with little predictive capacity. To solve these limitations, the authors propose the use of tree ensembles such as random forest. Though it is true that these models have been used before in the pioneering works in [14, 15], Breiman [30] produced one of the first papers that highlights the need to improve prediction using ensemble methods.

Following on from these papers, there have been numerous proposals that compare the behaviour of ensemble techniques with classical regression models, concluding that models behave better with machine learning techniques. Likewise, in the work by Pow et al. [17], they use 25,000 web data on Montreal properties with 130 characteristics; 70 related to the housing itself and 60 sociodemographic. These authors use principal component analysis (PCA) to reduce the dimension and four regression techniques to predict property prices: linear regression, support vector machine, K-nearest neighbors (KNN), and random forest regression and an ensemble approach by combining KNN and random forest technique. From the results, the authors highlight the good behaviour of the ensemble approach with a mean absolute percentage difference for the asking price of 9.85. In addition, they show that applying PCA does not improve the prediction error.

Ceh et al. [24] analyse the behaviour of random forest compared to multiple regression to select the most important variables. In the case of multiple regression, an analysis of main components allows it to go from 36 variables to 10 principal components and in the case of random forest, a procedure is carried out to determine the 10 most important variables. Interestingly, for random forest, the date of sale is important but not for ordinary least squares (OLS). Although the behaviour in terms of COD and MAPE of random forest is better than that of OLS, it should be noted that both overestimate the lowest prices and underestimate the highest. Specifically, in the application developed for the price of apartments in Ljubljana (Slovenia) with 7,497 observations for the 6-year period 2008–2013, the results in terms of MAPE for the test set were 7.27% for RF and 17.48% for multiple OLS, while in terms of COD the values obtained were 7.28% and 17.12%, respectively. Although the authors state that their model does not take into account the potential price differences over the 6-year time period under consideration, this price change could

influence their results. In our study, we use a static database of 2018 to avoid this problem.

Nguyen [25] develops an application in five counties in the United States using Zillow group web data by comparing linear regression models, random forest, and support vector machine. The results lead the author to conclude that both random forest and support vector machine behave better than linear regression in terms of the percentage of houses whose estimated prices fall within a 5% range of their actual sold prices. In addition, the conclusion emphasizes that it is not necessary to change the variables used in each county and that the accuracy of the model is practically the same using a series of common attributes for all of them. Dimopoulos et al. [26] develop an application to compare the behaviour of random forest and linear regression in estimating the prices of residential apartments in Nicosia (Cyprus). The results verify that the best behaviour in predictive terms is that of random forest, with average MAPE values of 25.2%. Shinde and Gawande [32] use data based on 3,000 observations with 80 parameters of a database called KaggleInc to compare the behaviour of logistic regression, support vector regression, lasso regression, and decision tree and show that the best behaviour, both in terms of accuracy and of error, is achieved with the decision tree. The variables used to estimate the sale price are area in square metres, overall quality which includes the overall condition and finish of the dwelling, location, the year in which the house was built, number of bedrooms and bathrooms, garage area and number of cars that can fit in the garage, swimming pool area, year in which the house was sold, and price at which the house was sold.

In addition to the comparison in the literature of machine learning techniques with classical regression models, there is a wide range of literature that compares different machine learning methods, concluding that there is no one technique that shows better behaviour than the others but highlights the best behaviour of tree ensemble techniques. For example, Kagie and Wezel [33] use Friedman's LSBoost and LADBoost boosting algorithms designed for regression with three main objectives: to predict dwelling prices in six areas in Netherlands; to determine the most important characteristics; and to build a price index. To do this, they use transaction data from the year 2004 obtained from Nederlandse Vereniging van Makelaars (NVM, Dutch Association of Real Estate Brokers) for the cities of Groningen, Apeldoorn, Eindhoven, Amsterdam, Rotterdam, and Zeeland, with 83 variables and a number of observations ranging from 2,216 for Zeeland to 8,490 for Amsterdam, also including sociodemographic variables. The results show that both boosting models improve the behaviour of linear and nonlinear models in the six areas considered, with improvements in terms of the absolute error of around 25–30% and in relative error of around 33–39%. In addition, they show that the models present a better behaviour in the prediction of errors in terraced houses and apartments and a worse behaviour in predicting errors in detached houses, which is consistent considering that the most influential characteristic on dwelling price is the size of the house.

Del Cacho [14] compares different ensemble methods for housing valuation in Madrid, based on a sample of 25,415 observations taken from an online real estate portal. The results show a better behaviour of ensemble of M5 model trees with a better behaviour of bagging unpruned decision trees, with a mean relative error of 15.25%. Similar results with a median percentage error of 15.11% and 13.18% are obtained for the English private rental market using gradient boost [34] and Cubist [35], respectively, by Clark and Lomax [36]. Graczyk et al. [37] use six machine learning algorithms: multilayer perceptron (MLP); radial basis function neural network for regression problems (RBF); pruned model tree (M5P); M5Rules (M5R); linear regression model (LRM); and NU-support vector machine (SVM) for the three ensemble methods of additive regression (an implementation of boosting in WEKA), bagging, and stacking, in Waikato Environment for Knowledge Analysis (WEKA). The results show that there are differences between the simple and ensemble methods used although all of them with good behaviour in terms of MAPE had values ranging from 19.02% to 15.89%. Bagging results are the most stable, with better results using SVM. However, the best results are obtained using stacking and SVM. The general conclusion of the study is that there is no single algorithm that produces the best results and, therefore, it is necessary to investigate the behaviour of different alternatives.

Antipov and Pokryshevskaya [15] show the best behaviour of random forest when estimating prices per square metre rather than for the total price due to heteroscedasticity and other real estate data problems. They propose comparing the behaviour of 10 algorithms: multiple regression; CHAID; exhaustive CHAID; CART; k-nearest neighbors (2 modifications); multilayer perceptron neural network (MLP); radial basis function neural network (RBF); boosted trees; and random forest. In the evaluation of each method, habitual metrics are used in the validation of the predictive capacity of automated valuation models such as the average ratio sale (SR), the coefficient of dispersion (COD), and the mean average percentage error (MAPE). All the analysed techniques showed acceptable values for all the metrics, both in the training set and in the test set and with better results for random forest with a MAPE of 17.25 and a COD of 16.97 while, using a two-step procedure, these are 14.86 and 14.77, respectively. In addition, this study proposes a classification of variables according to their relevance, highlighting the importance of the type of house and the district in which it is located. It also recommends a segmentation-based diagnostic method that determines segments based on the total area and the district in which the house is located, with any overestimated or underestimated value highlighting the need for the intervention of an appraiser. However, the main drawback of this study is that the data are too limited, focusing on 2-bedroom apartments with an area of up to 160 m<sup>2</sup> and a price below 30 million rubles. Such a limited profile is an unrealistic reflection of most cities.

Lasota et al. [38] propose that instead of using a single expert machine learning system, a combination of these should be used. They argue that in this way, the risk of selecting a poor model would be reduced in some of the cases

and large volumes of data could be analysed efficiently by applying the procedure to small partitions of the data and combining the results. This proposal is compared with the individual methods with two ensemble machine learning methods: mixture of experts (MoE) and Adaboost.R2 (AR2), concluding that Adaboost and this mixture of machine learning procedures show better behaviour, with no significant differences between the methods. In the case of MoE, the algorithms, multilayer perceptron, general linear model, and support vector regression, are used, while for AR2, multilayer perceptron, general linear model, and a regression tree are used. It is the mixture of machine learning procedures with multilayer perceptron and general linear model that show a better behaviour, without significant differences between MoE and AR2. However, in the study by Lasota et al. [38], they used information for the period 1998–2011 with the problem, as highlighted by the authors, of comparability in the data. They also use only four characteristics as explanatory variables which can give rise to an oversimplified model and could be the reason for a good behaviour of ensemble techniques with simple basic techniques such as the general linear model.

Another comparison, in this case of random forest with other machine learning methods was developed by Yoo et al. [39]. Machine learning is used to determine the variables which have the greatest impact on the price of housing in Onondaga (New York) and to establish a way of estimating dwelling prices. Specifically, OLS regression methods are compared with Cubist and random forest. In terms of determining the most important variables, though OLS uses a stepwise selection based on the level of significance, RF or Cubist uses boosting or bagging techniques that permit the handling of nonlinear models as they are nonparametric procedures. For predictability, the behaviour of the two machine learning techniques is better, highlighting RF in terms of root mean squared error (RMSE) with values, in relative terms with respect to their average, of 25.04 considering a neighbourhood within a radius of 100 metres and 22.47 within a radius of 1 km, for the test set. In addition, the model also incorporates environmental variables which have not previously been included in these types of models. The authors highlight that the application of machine learning methods in the selection of variables allows key variables to be selected without being based on a level of significance. These methods also allow a sufficiently parsimonious set of important variables to be found for good prediction, which means it is not so important that the model contains all relevant variables, as long as the prediction works well. Park and Bae [40] compare C4.5, RIPPER, Naive Bayesian, and Adaboost in the residential market of the county of Fairfax, Virginia, concluding that the best behaviour is achieved with RIPPER. In addition, their study uses these techniques as classification techniques, not regression, when classifying properties based on the presence of a positive or negative value in the difference between what they call closing (sold) prices and listing (for sale) prices.

Shahhosseini et al. [41] compare the behaviour of several ensemble models for the prediction of dwelling prices using two databases, widely cited in the relevant literature, the

Boston metropolitan area dataset [42] and the sales database of residential homes in Ames (Iowa) presented in [43]. To demonstrate the validity of the ensemble models, they use the following algorithms: multiple learners including lasso regression, random forest, deep neural networks, extreme gradient boosting (XGBoost), and support vector machines with three kernels (polynomial, RBF, and sigmoid). Based on the results of the median price prediction error, for Boston, the best performance in terms of MAPE appears for XGBoost and random forest with MAPE values of 16.44% and 16.35%, respectively. In the case of Ames housing, lasso and random forest are the models with the best MAPE with values of 0.66% and 0.77%, respectively. Incredibly low errors are attributable to the quantity and quality of information related to the 80 available variables as well as the huge sample size (2,930) in relation to the population size of Ames (Iowa, USA) of 50,781 inhabitants. Therefore, these results lead us to conclude that there is no one model which performs better than the others.

Finally, Neloy et al. [44] develop a model for predicting the rental price of houses in Bangladesh through a website database of 3,505 homes with information on 19 characteristics. To develop the model, the following simple algorithms are selected for prediction: advance linear regression, neural network, random forest, support vector machine (SVM), and decision tree regressor. In addition, the ensemble learning is stacked with the following algorithms: ensemble AdaBoosting regressor, ensemble gradient boosting regressor, and ensemble XGBoost. Also, ridge regression, lasso regression, and elastic net regression are used to combine the advanced regression techniques. The best results, in terms of accuracy, are obtained by the ensemble gradient boosting with 88.75% and the worst by the ensemble AdaBoosting with 82.26%. In terms of root mean square error (RMSE), the behaviour is similar, with values of 0.1864 and 0.2340, respectively.

Other uses of the decision tree include the application of the CART algorithm to segment the observations and to improve the ability to estimate the model by applying different models by segments or even with the assistance of an appraiser if necessary [45]. To do this, a CART algorithm is applied, using the percentage error (estimated value less real value in absolute value divided by real value) as a dependent variable and the sales ratio (estimated value divided by real value) to determine segments of observations that allow them to go from a general MAPE in the simple training of 12,688 to a value in the best segment of 9,783 and in the simple test of 14,859 to 12,364. Pérez-Rave et al. [20] propose a methodology that incorporates a variable selection procedure called simple incremental with resampling (MIN-REM). This procedure is used in combination with a principal component analysis in two cases; 61,826 homes sold in Colombia, and the data used in [46] from the 2011 Metropolitan American Housing Survey with 58,888 observations. The results show a MAPE value of 27% without using interactions and 20.9% using the procedure proposed, in the case of housing in Colombia.

From all these studies, it follows that the analysis of the behaviour of different machine learning techniques to



analyse the price of housing has been widely covered in the literature. While the majority of the applications stress the importance of determining the most influential variables on the price of housing, there are few applications which focus on prediction and, above all, there are few studies that use measures such as MAPE or COD to help evaluate the predictive capacity of the models; the majority are based on measuring the predictive capacity using the coefficient of determination. In addition, the applications developed focus on specific areas or cities without trying to cover a wide geographical area (except in the case of Colombia where the study use the same model for the whole country). In this study, we cover a wider geographical area by developing, through an automated procedure for estimating models, a model to be applied to each of the Spanish municipalities where information is available. This gives us a total of 433 municipalities.

### 3. Methodology

As it has been stated, the aim of this article is to develop an automatic application that contains, for each municipality, a model capable of accurately estimating the price of housing. Several models are fitted in each municipality, among a range of competing machine learning techniques. Then, they will be analysed in order to check if there exists one best method that achieves optimal results in terms of the error measures explained at the end of this section.

The selected models are bagging, boosting, and random forest. All of them are ensemble algorithms, and we use regression trees as base learners. For this reason, the results of the single decision tree model will also be displayed as a reference alongside the results of the more complex models. The ensemble methods usually provide good prediction results although it is true that they sacrifice in some way the possibility of interpretation of the relationships between the predictor variables and the target. In our context, given the large number of models that will be estimated to completely cover the Spanish territory, accurate predictions are more important than easily interpret models.

The following briefly shows what each of these ensemble methods consists of. To begin with, bagging is an ensemble method proposed in [47] from the basis of bootstrapping and aggregating methods. The main advantage of this methods is the reduction of noise presence in the observations in the random samples obtained with replacement from the original set. Once the trees are fitted over the bootstrap samples, the outputs are averaged. The noise reduction coupled with the instability often shown by individual predictors lead bagging to improvements, especially for unstable procedures.

For its part, boosting [48] is an ensemble method capable of converting a weak learner into one with much higher accuracy. Boosting, similar to bagging, applies an iterative learning process. The differential characteristic of this method is that each iteration is not independent of the previous ones but uses a reweighting system to focus the attention of the learning process on the observations that in former steps have been estimated with higher errors. The

chosen algorithm to implement boosting in this article is gradient boosting [34] that consists in adding weak learning models, such as decision trees, by using a gradient descent procedure to minimize a loss function.

Random forest was also proposed in [30], and it could be seen as a variation of the bagging method, with a higher dose of randomness. This added randomness is given because when constructing the successive trees, the optimal division is not sought among all the available predictive variables, but only among a subset randomly chosen in each node. The main advantage of this method is that it incurs a lower risk of overfitting and, therefore, usually provides more accurate estimates. It should be noted that bagging is a special case of random forest when the subset of variable candidates contains the total number of predictors available.

All the models have been applied using the statistical environment R [49]. Specifically, the R packages `rpart` [50], `gbm` [51], and `randomForest` [52] have been used for fitting individual trees, boosting and bagging, and random forest, respectively.

Due to the large number of models to be fitted in this complex problem, the parameter tuning in random forest has been optimized for each model through the `caret` R library [53]. There are three main parameters to be set in random forest. The first two are the number and size of trees to be grown. The number should not be too small to ensure that every input row participates in the learning process at least a few times. The size of trees depends on the minimum size of terminal nodes. Setting this number larger leads to smaller trees and quicker learning procedure. Another important parameter in random forest is the number of predictors randomly sampled as candidates at each split. With regard to bagging, it has been treated as a particular case of random forest.

Regarding boosting, there are four main parameters to be set in the `gbm` model. The first one is the learning rate (shrinkage), with values 0.001, 0.01, and 0.1, which controls how large the changes are from one iteration to the next one, similarly to the learning rate in neural networks. Secondly, the complexity of the tree is controlled by two parameters, interaction depth (tested among 1, 3, 5, and 10) and the minimum number of observations per node, similar to random forest (taking 1, 5, 10, and 20). Finally, but also very important, the number of trees (iterations), an ensemble of 1,000 trees is generated and then pruned according to the minimum cross-validation error.

The loss function chosen for the optimization of each supervised method is the mean square error (MSE). In order to guarantee a good generalization ability avoiding overfitted models, 2/3 of the observations in the sample has been randomly assigned to the train set and the other 1/3 to the validation set. Once the best model of each technique (regression tree, bagging, boosting, and random forest) has been chosen in each municipality, the comparison of the final behaviour of four models has been analysed by the following error measures. They will be able to analyse the goodness of fit and validate the predictive capacity of the models.

- (a) Mean ratio (average sales ratio) is the average of the  $SR_i$ ,  $SR$  being the sales ratio defined as

$$SR_i = \frac{\hat{y}_i}{y_i}, \quad (1)$$

where  $y_i$  is the property value  $i$  and  $\hat{y}_i$  is the estimated value.

- (b) Mean absolute percentage error (MAPE) or relative mean error:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100. \quad (2)$$

The measure is in percentage terms, so it is comparable among different models.

- (c) Median absolute percentage error (MdAPE):

$$MdAPE = Me \left( \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) 100, \quad (3)$$

where  $Me()$  is the median, i.e., the value separating the higher half from the lower half of the absolute percentage errors.

- (d) Coefficient of dispersion (COD):

$$COD = \sum_{i=1}^n \frac{(1/n) \sum_{i=1}^n |SR_i - Me(SR)|}{Me(SR)} 100, \quad (4)$$

where  $Me(SR)$  is the median of the  $SR_i$ . Its interpretation does not depend on the assumption of normality.

In line with the study by Pérez-Rave et al. [20], we carry out the estimation and measurement of errors using monetary values since any transformation of the variable to be estimated (price), such as the logarithmic transformation, can lead to an improvement in results from a fictitious statistical point of view. In addition, the estimated value of the price is made in monetary terms and does not require any transformation for its interpretation and comparison between locations.

#### 4. Empirical Application

To develop the empirical application, a database is constructed based on the information obtained from freely available real estate websites. The data from advertisements on the Internet allow the development of the application of big data techniques for the analysis of dwelling prices with greater precision because the volume of accessible data is large and enriched daily, both ideal characteristics to be able to apply these techniques. In addition, the data are quite varied and the sale value on the web and the offline sale value are seen to be of similar magnitude. The Internet source also

offers information on a variety of property and neighbourhood characteristics that are difficult to find from other sources [20]. However, this source of information has been little used despite the existence of applications developed with great success in both the real estate sector as well as other sectors [54]. These same aspects are highlighted in [55, 56] in which the authors point out that web prices offer a valuable opportunity for statistical analysis due to the constant generation of information, their accessibility, and availability as well as there being little notable differences compared to offline prices. Within real estate, applications developed using web data are used by Özsoy and Şahin [23] in Istanbul; Del Cacho [14] in Madrid; Larraz and Larraz and Población [57, 58] in Spain; Pow et al. [17] in Montreal; Larraz and Población [59] in Czech Republic; Nguyen [25] in the United States; Clark and Lomax [36] in England; Pérez-Rave et al. [20] in Colombia; or Neloy et al. [44] in Bangladesh.

In our study, the database contains information related to the price of the property (flat nonsingle-family home) and its reliable geolocation as well as information that refers specifically to the characteristics of each property. We have access to information on properties for sale in all Spanish municipalities during 2018. The information includes the price of the property and the following 33 variables which represent the characteristics of each property: a text variable that shows the postal code in which the property is located; three numerical variables that include the constructed surface area, the number of bedrooms, and the number of bathrooms; and 29 attributes that have been categorized into different levels. Among these, the variables considered the most influential on dwelling prices by the implemented methods are location (longitude and latitude coordinates), constructed area, number of bedrooms, number of bathrooms, floor (basement, normal, or attic), the state of conservation (new, with important improvements, adequate for the age, or need for major improvements), and the presence of air conditioning, heating, lift, garage, terrace, green areas, swimming pool, and storage room. Therefore, these variables are used to estimate the price of each property.

In the first phase, data of adverts with possible errors are removed, for example, properties with zero unit price. Subsequently, a descriptive analysis is performed to decide what type of variables to work with. Finally, a multivariate analysis of outliers with the available variables is carried out, based on Mahalanobis distance. After this preliminary analysis, we obtain a substantial database whose elements are uniformly distributed throughout the territory. From this database, we work with those municipalities where there are at least 100 sample observations, 100 homes for sale that allow the procedure to choose the best model in each case. Therefore, our study presents an empirical application developed in 433 municipalities (out of the 8,125 in Spain) which have more than 100 dwellings on sale during the period of research. To be precise, this amounts to information on 790,631 real estate properties distributed in 48 Spanish provinces (out of the 52 in Spain) made up of 433 municipalities.

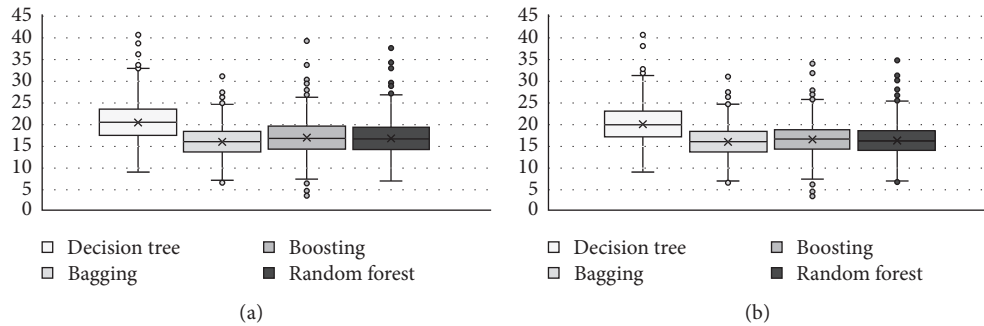


FIGURE 1: MAPE (a) and COD (b) main error measures of the four techniques (decision tree, bagging, boosting, and random forest) corresponding to the 433 municipalities in Spain.

As it has been said in Section 3, for the application of the different regression techniques, the data set is divided into two; a training set and a test set in a proportion of two-thirds and one-third, respectively. The data in the training set allow to fit the best combination in an automated way for each case, while assessments made on the properties of the test set are used to compare the suitability of the different techniques.

The errors obtained in the valuation of property prices in the 433 municipalities with information available show an average MAPE value of 20.49 with the decision tree technique, 16.54 with bagging, 16.98 with boosting, and 16.69 with random forest, while the average value of the COD was 20.03, 16.03, 16.53, and 16.23, respectively. Statistical dispersion is depicted through the box and whiskers plots in Figure 1. These results show a good performance of bagging, boosting, and random forest techniques given the heterogeneity of the sample and the wide geographic range analysed. However, the decision tree technique, overall, does not give satisfactory results.

A further line of research could be to what extent the error measures depend on the population size and even the sample size, or the number of properties available to be used as examples in the estimate. It is worth finding out whether there is a better or worse behaviour in the methods analysed in small, medium, or large cities, or if the maxim of “larger sample size, better estimates” is met. In fact, Table 1 shows how the results of the 4 techniques show a practically zero linear correlation between the different population and sample sizes with the different error measures. This may be due to both the quality of the starting information and the great arbitrariness present in the prices of housing in Spain. Since the quality of the starting information was controlled in the early stages of the analysis, the second option is considered more plausible.

Nor, are there any nonlinear correlation between error measures and population or sample sizes. Just as an example, Figure 2 shows the scatter plot for two main error measures, MAPE and COD versus population size for bagging results. Four techniques present very similar results. No regression structure can be deduced from the graphs. As observed in Figures 2(a) and 2(c), the biggest cities in Spain, Madrid, and Barcelona could be hiding the real correlation. But after having eliminated both cities (see Figures 2(b) and 2(d)), the

TABLE 1: Linear correlation coefficients between the MAPE and COD values obtained from the 433 municipalities where the valuations and population sizes (inhab.) and sample sizes (N) of said municipalities have been calculated.

Correlation coefficient	Technique			
	Decision tree	Bagging	Boosting	Random forest
MAPE vs inhab.	0.10	-0.07	-0.04	-0.07
MAPE vs N	0.26	-0.03	0.02	-0.04
COD vs inhab.	0.09	-0.07	-0.03	-0.06
COD vs N	0.24	-0.02	0.03	-0.03

Note. Own elaboration.

graphs do not show any relation between the errors and the population size. Coefficient of determination is stated in Table 2, having computed linear, exponential, potential, and logarithmic coefficients. Note that all of them are almost zero.

Because most of the assessments of real estate portfolios will be carried out in the largest cities, we decide to analyse the average results of the municipalities with a population greater than 100,000 inhabitants (see Table 3), 63 in all, and show the results of each of these in more detail. The results obtained for each of the 63 selected municipalities are presented in four tables in the appendix of this study (see Tables S1–S4 in the Supplementary Material for a more detailed analysis), one for each of the techniques used.

Table 3 reports the improvement in all cases when using tree ensemble methods (bagging, boosting, and random forest) compared to individual decision trees, with improvements in the estimation capacity measured through the MAPE of around six percentage points. The average MAPE values for the 63 municipalities show the best performance for bagging and random forest with an average value for the set of 63 municipalities with the largest population of 15.73 and 15.93, respectively. Both techniques achieve the same minimum MAPE value (8.58). However, in terms of a maximum value, although the value reached by random forest is a point higher than that reached by bagging, both achieve very acceptable values. In terms of COD, a similar situation is observed, with the results of bagging and random forest being the best, at around 15%, followed by those of the boosting technique that has an average COD of 16.3%. The

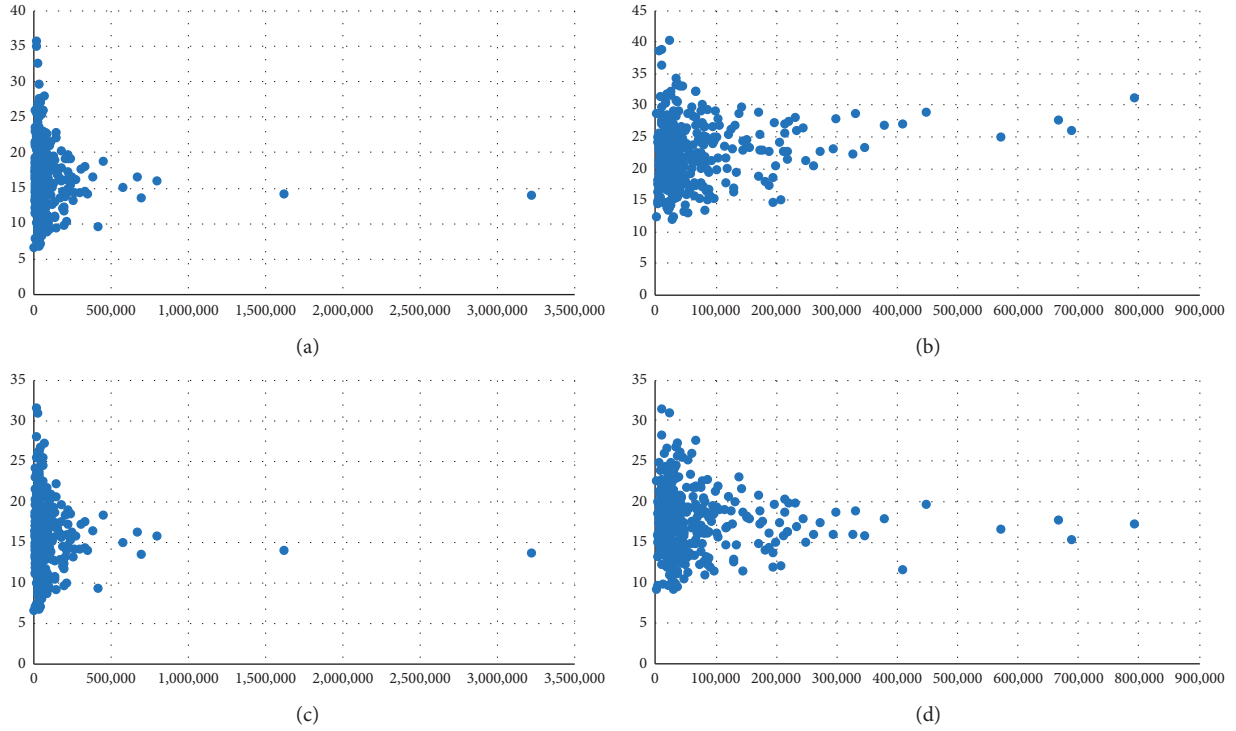


FIGURE 2: Bagging's MAPE and COD versus municipality population considering all the locations and without Madrid and Barcelona. (a) MAPE vs inhabitants, bagging. (b) MAPE vs inhabitants, bagging without Madrid and Barcelona. (c) COD vs inhabitants, bagging. (d) COD vs inhabitants, bagging without Madrid and Barcelona. Note. Own elaboration.

TABLE 2: Coefficients of determination of the regression between Bagging's MAPE and COD and municipality population considering all the locations and without Madrid and Barcelona.

Coefficient of determination	Technique: bagging			
	Linear	Exponential	Logarithmic	Potential
<i>Considering all locations</i>				
MAPE vs inhab.	0.0053	0.0034	0.0075	0.0040
COD vs inhab.	0.0044	0.0026	0.0049	0.0023
<i>Without Madrid and Barcelona</i>				
MAPE vs inhab.	0.0252	0.0246	0.0129	0.0148
COD vs inhab.	0.0053	0.0028	0.0038	0.0017

Note. Own elaboration.

TABLE 3: Average, minimum, and maximum results for MAPE and COD values obtained for the 63 municipalities with more than 100,000 inhabitants.

Technique	63 municipalities (>100,000 inhab.)					
	MAPE average	MAPE minimum	MAPE maximum	COD average	COD minimum	COD maximum
Decision tree	21.92	11.68	30.40	21.22	11.65	27.95
Bagging	15.73	8.58	22.93	15.41	8.48	22.19
Boosting	16.62	10.15	23.45	16.36	10.06	23.28
Random forest	15.93	8.58	23.92	15.56	8.48	23.08

COD of the decision tree technique was alone in achieving a value above the recommended 20%. Figure 3 graphically shows these results along with the dispersion of MAPE and COD measures. Note that all the outliers, municipalities with MAPE, and COD abnormally high or low have disappeared. They corresponded to municipalities with less than 100,000 inhabitants.

From the analysis of errors made in the valuation of properties of the test set for each of the 63 municipalities with more than 100,000 inhabitants of Spain, it is worth highlighting the good behaviour of the mean ratio that, in almost all cases, shows values between 0.98 and 1.1 (see Tables S1–S4 in the Supplementary Material). From the value of MdAPE, in the case of bagging, the smallest value

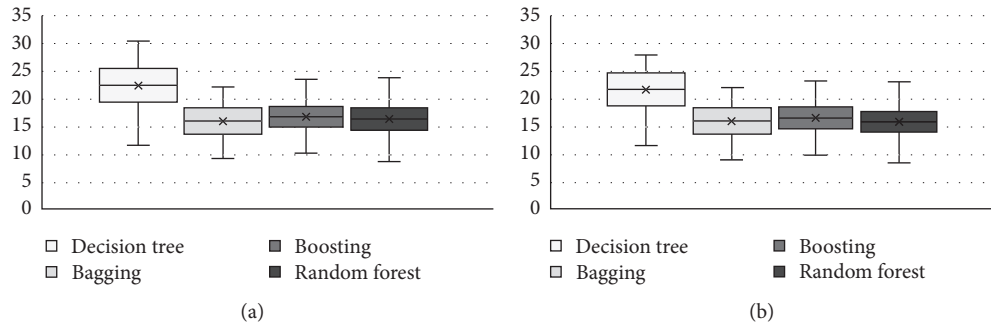


FIGURE 3: MAPE (a) and COD (b) main error measures of the four techniques (decision tree, bagging, boosting, and random forest) corresponding to the largest 63 municipalities in Spain.

for this measure is 6.59 and the maximum 18.91, which indicates that in the best-case municipalities, fifty percent of the valuations with the lowest error have an error of less than 6.59 and, in the worst-case municipalities, this error is no greater than 18.91. For random forest, these values are 6.15 and 19.64, respectively, reaching a higher value in terms of maximum and a lower value in terms of minimum compared to bagging. In addition, given that choosing between bagging and boosting is made more difficult by the similarity in the results obtained, it should be noted that bagging presents a better behaviour in terms of MdAPE since only 4 of the 63 municipalities analysed present values superior to 16% and 10 superior to 15%, while with random forest these values are 6 and 14, respectively.

## 5. Conclusions

The need for a rapid and economical appraisal of real estate and the greater availability of up-to-date information accessible through the Internet have led to the application of big data techniques and machine learning to carry out real estate valuation.

At the forefront of these machine learning techniques are tree ensemble methods, in particular, bagging, boosting, and random forest. So far, these techniques have been applied in many cases for purposes other than the estimation of property prices, and when they have been applied to real estate valuations, they have been done in a limited way to very specific geographical areas. In order to advance understanding of the value of the techniques of tree ensemble on an automated and massive scale, this study shows the results obtained from the application of different techniques for the whole of Spain with a total of 433 municipalities spread across 48 provinces. The article presents an automated algorithm which selects the best model for each technique in each municipality. Their behaviour in terms of estimation capacity is measured through error measures widely cited in the literature.

The results show that the behaviour of the tree ensemble clearly outperforms individual trees although of the three methods analysed (bagging, boosting, and random forest), none has a clear advantage over the others. Even so, looking more closely at the behaviour of the bagging and random forest methods, it seems that the slightly better results of bagging in terms of MAPE and COD together with the

results in terms of MdAPE would make us opt for the use of bagging in the case of Spain.

Reviewing the literature available so far, it can be concluded that the results obtained in terms of MAPE are better than those obtained in [26] with a value of 25.2% in Nicosia or in [46] for the US with 20.9%. The results are similar to those obtained in [14] with a value that ranges from 19.02% to 15.89% in Madrid and worse than those obtained for Ljubljana in [24] with an average MAPE of 7.28. However, it should be borne in mind that these applications focused on specific geographical areas while the application developed in this study covers the entire Spanish territory. The error measures provided are means of the MAPE and COD of each municipality, with municipalities of very different population, sample sizes, and socioeconomic characteristics.

From the global analysis of the 433 municipalities as a whole, it can be concluded that the error measures do not depend on the population size or the size of the sample set. This fact suggests the presence of a certain random component in the determination of sales prices since the greater the available sample information, the better the results should be.

Finally, it should be noted that this study has other active lines of research that are already being developed, such as the inclusion of a dynamic database that allows the handling of information with different temporal references or the inclusion of ensemble methods that allow machine learning techniques, not just simple trees, to be combined.

## Data Availability

The data base used to support the findings of this study were supplied by COHISPANIA, Consultoría y Valoración, under license and so cannot be made freely available. Requests for access to these data should be made to <http://www.cohispania.com/contacto>.

## Conflicts of Interest

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

## Acknowledgments

The authors want to thank the collaboration of Compañía Hispania de Tasaciones y Valoraciones, S. A. Emilio L. Cano

was partially funded by the Spanish “Agencia Estatal de Investigación” via the MTM2017-86875-C3-1-R AEI/FEDER, UE project. The present work was financed through the R&D contract between the University of Castilla-La Mancha and Cohispania with ref: UCTR180093.

## Supplementary Materials

Table S1: main results for decision trees. Table S2: main results for bagging. Table S3: main results for boosting. Table S4: main results for random forest. (*Supplementary Materials*)

## References

- [1] Bank for International Settlements, *International Convergence of Capital Measurement and Capital Standards*, Basel Committee on Banking Supervision, Basel, Switzerland, 2006.
- [2] European Council, “Directive 2006/48/EC of the European Parliament and of the Council of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions,” *Official Journal of the European Union*, vol. L177, pp. 1–200, 2006.
- [3] J. K. Eckert, *Property Appraisal and Assessment Administration*, International Association of Assessing Officers, Chicago, IL, USA, 1990.
- [4] V. Kontrimas and A. Verikas, “The mass appraisal of the real estate by computational intelligence,” *Applied Soft Computing*, vol. 11, no. 1, pp. 443–448, 2011.
- [5] R. Schulz, M. Wersing, and A. Werwatz, “Automated valuation modelling: a specification exercise,” *Journal of Property Research*, vol. 31, no. 2, pp. 131–153, 2014.
- [6] O. Kettani and M. Oral, “Designing and implementing a real estate appraisal system: the case of Québec Province, Canada,” *Socio-Economic Planning Sciences*, vol. 49, pp. 1–9, 2015.
- [7] M. Mooya, “Of mice and men,” *Urban Studies*, vol. 48, no. 11, pp. 2265–2281, 2011.
- [8] W. McCluskey and S. Anand, “The application of intelligent hybrid techniques for the mass appraisal of residential properties,” *Journal of Property Investment & Finance*, vol. 17, no. 3, pp. 218–239, 1999.
- [9] C. M. Filho and O. Bin, “Estimation of hedonic price functions via additive nonparametric regression,” *Empirical Economics*, vol. 30, no. 1, pp. 93–114, 2005.
- [10] H. Selim, “Determinants of house prices in Turkey: hedonic regression versus artificial neural network,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2843–2852, 2009.
- [11] N. García, M. Gámez, and E. Alfaro, “ANN+GIS: an automated system for property valuation,” *Neurocomputing*, vol. 71, no. 4–6, pp. 733–742, 2008.
- [12] R. D. Jaen, “Data mining: an empirical application in real estate valuation,” in *FLAIRS Conference*, S. M. Haller and G. Simmons, Eds., pp. 314–317, AAAI Press, Palo Alto, CA, USA, 2002.
- [13] G.-Z. Fan, S. E. Ong, and H. C. Koh, “Determinants of house price: a decision tree approach,” *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006.
- [14] C. Del Cacho, *A Comparison of Data Mining Methods for Mass Real Estate Appraisal*, University Library of Munich, Munich, Germany, 2010, <https://mpr.ub.uni-muenchen.de/id/eprint/27378MPRA Paper No. 27378>.
- [15] E. A. Antipov and E. B. Pokryshevskaya, “Mass appraisal of residential apartments: an application of Random forest for valuation and a CART-based approach for model diagnostics,” *Expert Systems with Applications*, vol. 39, no. 2, pp. 1772–1778, 2012.
- [16] M. Thériault, F. Des Rosiers, and F. Joerin, “Modelling accessibility to urban services using fuzzy logic,” *Journal of Property Investment & Finance*, vol. 23, no. 1, pp. 22–54, 2005.
- [17] N. Pow, E. Janulewicz, and L. Liu, “Applied machine learning project 4 prediction of real estate property prices in montreal,” 2014, [http://rl.cs.mcgill.ca/comp598/fall2014/comp598\\_submission\\_99.pdf](http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf).
- [18] O. Bin, “A prediction comparison of housing sales prices by parametric versus semi-parametric regressions,” *Journal of Housing Economics*, vol. 13, no. 1, pp. 68–84, 2004.
- [19] Shabana, G. Ali, M. K. Bashir, and H. Ali, “Housing valuation of different towns using the hedonic model: a case of Faisalabad city, Pakistan,” *Habitat International*, vol. 50, pp. 240–249, 2015.
- [20] J. I. Pérez-Rave, J. C. Correa-Morales, and F. González-Echavarría, “A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes,” *Journal of Property Research*, vol. 36, no. 1, pp. 59–96, 2019.
- [21] J. Bin, S. Tang, Y. Liu et al., “Regression model for appraisal of real estate using recurrent neural network and boosting tree,” in *Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI)*, IEEE, Beijing, China, pp. 209–213, September 2017.
- [22] 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.
- [23] O. Özsoy and H. Şahin, “Housing price determinants in Istanbul, Turkey,” *International Journal of Housing Markets and Analysis*, vol. 2, no. 2, pp. 167–178, 2009.
- [24] M. Ceh, M. Kilibarda, A. Lisec, and B. Bajat, “Estimating the performance of random forest versus multiple regression for predicting prices of apartments,” *International Journal of Geo-Information*, vol. 1, p. 168, 2018.
- [25] A. Nguyen, “Housing price prediction,” 2018, <https://pdfs.semanticscholar.org/782d/3fdf15f5ff99d5fb6acafb61ed8e1c60fab8.pdf>.
- [26] T. Dimopoulos, H. Tyrallis, N. P. Bakas, and D. Hadjimitsis, “Accuracy measurement of random forests and linear regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus,” *Advances in Geosciences*, vol. 45, pp. 377–382, 2018.
- [27] M. Skurichina and R. P. W. Duin, “Bagging, boosting and the random subspace method for linear classifiers,” *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121–135, 2002.
- [28] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, USA, 1993.
- [29] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.
- [30] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [31] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, “Big data in real estate? From manual appraisal to automated valuation,” *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, 2017.
- [32] N. Shinde and K. Gawande, “Valuation of house price using predictive techniques,” *International Journal of Advances in Electronics and Computer Science*, vol. 5, no. 6, pp. 34–40, 2018.
- [33] M. Kagie and M. V. Wezel, “Hedonic price models and indices based on boosting applied to the Dutch housing market,”

- Intelligent Systems in Accounting, Finance & Management*, vol. 15, no. 3-4, pp. 85–106, 2007.
- [34] J. H. Friedman, “Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [35] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, USA, 1984.
- [36] S. D. Clark and N. Lomax, “A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques,” *Journal of Big Data*, vol. 5, no. 1, p. 43, 2018.
- [37] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, “Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal,” in *Intelligent Information and Database Systems. ACIIDS 2010. Lecture Notes in Computer Science*, R. Goebel, J. Siekmann, and W. Wahlster, Eds., vol. 5991, pp. 340–350, Springer, Berlin, Germany, 2010.
- [38] T. Lasota, B. Londzin, Z. Telec, and B. Trawiński, “Comparison of ensemble approaches: mixture of experts and AdaBoost for a regression problem,” in *Intelligent Information and Database Systems. ACIIDS 2014, Lecture Notes in Computer Science*, N. T. Nguyen, B. Attachoo, B. Trawiński, and K. Somboonviwat, Eds., vol. 8398, Springer, Cham, Switzerland, 2014.
- [39] S. Yoo, J. Im, and J. E. Wagner, “Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County, NY,” *Landscape and Urban Planning*, vol. 107, no. 3, pp. 293–306, 2012.
- [40] B. Park and J. K. Bae, “Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data,” *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015.
- [41] M. Shahhosseini, G. Hu, and H. Pham, “Optimizing ensemble weights for machine learning models: a case study for housing price prediction,” *Smart Service Systems, Operations Management, and Analytics*, Springer, Berlin, Germany, 2019, [https://lib.dr.iastate.edu/imse\\_conf/185/](https://lib.dr.iastate.edu/imse_conf/185/).
- [42] D. Harrison and D. L. Rubinfeld, “Hedonic housing prices and the demand for clean air,” *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.
- [43] D. De Cock, “Ames, Iowa: alternative to the Boston housing data as an end of semester regression project,” *Journal of Statistics Education*, vol. 19, no. 3, p. 115, 2011.
- [44] A. Neloy, M. Sadman Haque, and M. Mahmud Ul Islam, “Ensemble learning based rental apartment price prediction model by categorical features factoring,” in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pp. 350–356, Zhuhai, China, 2019.
- [45] E. B. Pokryshevskaya and E. A. Antipov, “Applying a CART-based approach for the diagnostics of mass appraisal models,” *Economics Bulletin*, vol. 31, no. 3, pp. 2521–2528, 2011.
- [46] S. Mullainathan and J. Spiess, “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.
- [47] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [48] R. E. Schapire, “The strength of weak learnability,” *Machine Learning*, vol. 5, pp. 197–227, 1990.
- [49] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019, <https://www.R-project.org/>.
- [50] T. Therneau and B. Atkinson, *Rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-15*, 2019, <https://CRAN.R-project.org/package=rpart>.
- [51] B. Greenwell, B. Boehmke, and J. Cunningham, *Gbm: Generalized Boosted Regression Models. R package version 2.1.5*, 2019, <https://CRAN.R-project.org/package=gbm>.
- [52] A. Liaw and M. Wiener, “Classification and regression by random forest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [53] M. Kuhn, J. Wing, and S. Weston, *Caret: Classification and Regression Training. R Package Version 6.0-84*, 2019, <https://CRAN.R-project.org/package=caret>.
- [54] M. E. Beręsewicz, “On representativeness of Internet data sources for real estate market in Poland,” *Austrian Journal of Statistics*, vol. 44, no. 2, pp. 45–57, 2015.
- [55] A. Cavallo, *Scraped Data and Sticky Prices*, Social Science Research Network, Rochester, NY, USA, SSRN Scholarly Paper ID 1711999, 2012.
- [56] A. Cavallo, “Are online and offline prices similar? Evidence from large multi-channel retailers,” *American Economic Review*, vol. 107, no. 1, pp. 283–303, 2017.
- [57] B. Larraz, “An expert system for online residential properties valuation,” *Review of Economics & Finance*, vol. 2, pp. 69–82, 2011.
- [58] B. Larraz and J. Población, “An online real estate valuation model for control risk taking: a spatial approach,” *Investment Analysts Journal*, vol. 42, no. 78, pp. 83–96, 2013.
- [59] E. Hromada, “Mapping of real estate prices using data mining techniques,” *Procedia Engineering*, vol. 123, pp. 233–240, 2015.

## Research Article

# Predicting Days on Market to Optimize Real Estate Sales Strategy

Mauro Castelli <sup>1</sup>, Maria Dobрева,<sup>1</sup> Roberto Henriques,<sup>1</sup> and Leonardo Vanneschi<sup>1,2</sup>

<sup>1</sup>NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

<sup>2</sup>LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

Correspondence should be addressed to Mauro Castelli; [mcastelli@novaims.unl.pt](mailto:mcastelli@novaims.unl.pt)

Received 7 November 2019; Accepted 16 January 2020; Published 25 February 2020

Guest Editor: Francesco Tajani

Copyright © 2020 Mauro Castelli 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.

Irregularities and frauds are frequent in the real estate market in Bulgaria due to the substantial lack of rigorous legislation. For instance, agencies frequently publish unreal or unavailable apartment listings for a cheap price, as a method to attract the attention of unaware potential new customers. For this reason, systems able to identify unreal listings and improve the transparency of listings authenticity and availability are much on demand. Recent research has highlighted that the number of days a published listing remains online can have a strong correlation with the probability of a listing being unreal. For this reason, building an accurate predictive model for the number of days a published listing will be online can be very helpful to accomplish the task of identifying fake listings. In this paper, we investigate the use of four different machine learning algorithms for this task: Lasso, Ridge, Elastic Net, and Artificial Neural Networks. The results, obtained on a vast dataset made available by the Bulgarian company Homeheed, show the appropriateness of Lasso regression.

## 1. Introduction

The real estate market in Eastern Europe and former Soviet Union countries is emerging. In Bulgaria, the situation does not differ. Given the recent political and economic history of the country, the development of the Bulgarian property market can be presented in three main temporal stages: during socialism, the transition to a market economy, and the current internationally attractive market. The third stage is a period when the real estate market registered double-digit annual growth due to the international investment interest. Later, between 2003 and 2008, the sector was blooming which led to the creation of a price balloon formed by a 40% drop in the housing prices. After this crisis, property investments have registered again a gradual increase. Statistics show that the housing sales increased by 11.5% for the first quarter of 2018 and the interest rates remained at their low levels. Also, numerous new buildings were constructed, allowing for further housing sales growth of 6.3% [1]. Figure 1 reports the trend of interest rates and bank property loans from 2008 to 2018.

All these fluctuations in the market lead to the easy entrance and exit in the market of brokers, who compete for

customers. The market is not exclusive, and a single property can be offered on the market several times, in different sources and by a variety of brokers. Often brokers keep outdated or unreal, but attractive, listings online, to increase the chance of acquiring new customers. This usually creates wrong expectations and bad customer experience.

Homeheed is a Bulgarian startup, which tries to counteract this problem, by centralizing the redundant listings in one single platform. In technical terms, the company uses key points matching technique to identify duplicates of a listing, using several techniques including image recognition. Then, it summarizes the listings in one central unit. Currently, one apartment can be found online listed by different brokers and/or with changes in the description. This results in difficulties to extract a unique identification key for duplicated listings. Homeheed found out that images remain the only part of a listing offer by which one apartment can be tracked.

The value proposition of this process is to act as a single point of truth and to enable the customer to see all listings of a property, as well as to understand whether it is available or not. Homeheed entered the market recently with a first prototype to validate the idea and the demand. The team



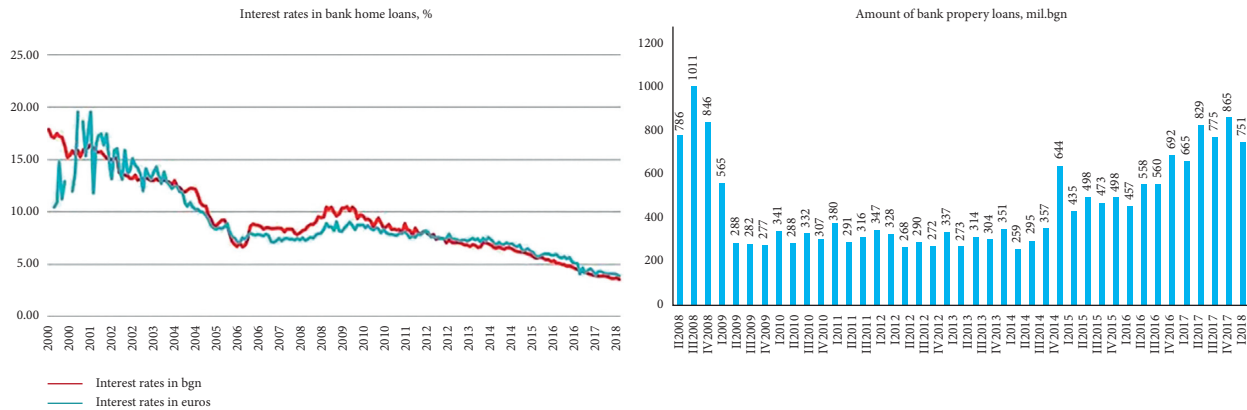


FIGURE 1: Loans interest rates Bulgaria [1] and amount of home loans [1].

provides the potential customers with a demo version of the platform, where the listings are filtered from fake offerings and only properties matching the individual preferences are received by email.

Homeheed collected information about the property market and listings from 2015 to 2018. The Startup aims to analyse these data to optimize its market entry program and to forecast return on investment (ROI). This work applies data mining techniques, based on this historical information, to forecast how many days a listed property with specific characteristics will be online. This will help Homeheed and several analogous organizations to provide customers firstly with the most attractive offers and so to optimize the revenue stream.

*1.1. Background and Problem Identification.* The topic of the irregularities and frauds in the real estate market has been raising heated debates in the Bulgarian media channels in the last few years. Generally speaking, the real estate market is not regulated by rigorous laws, which leads to the easy entrance of real estate agencies. Some agencies frequently publish unreal or unavailable apartment listings, often on a price below the average for the market, as a method to acquire customers looking for a new living property. These unaware customers either never see the desired place, or are even misled with fraud schemes for advanced payment before the deal. This not only creates a bad customer experience and dissatisfaction, but also makes the process of finding a living property challenging and time consuming. These instabilities and misappropriations in the property sector necessitate the development of a more transparent platform, like the one developed by Homeheed, and the establishment of better methods for assessing homes availability (Vasilev, n.d.) [2].

In the core of Homeheed's value proposition is the transparency of listings authenticity and availability. The startup goal is to provide a solution that can support the process of fixing the market irregularities, as well as should lead to better customer experience. Currently, the Homeheed team is trying to develop effective methods to identify unreal listings. Interestingly, it was observed that the number of days a published listing remains online can have a strong

correlation with the probability of a listing being unreal. More specifically, it has been observed that among all the available cases of ascertained fake listings, around 68% have stayed on the market for a number of days larger than the third quartile calculated over all the available data, while around 21% have stayed on the market for a number of days larger than the median calculated over all the available data. For this reason, building an accurate predictive model for the number of days a published listing stays online can be very helpful to accomplish the task of identifying fake listings. In this paper, we prefer to generate predictive models of "days-on-the-market," instead of directly predicting if a listing is fake or not because it is likely that the ascertained fake listings to which we were referring above are only a part of the fake listings contained in the Homeheed data. In other cases, the fraud is only suspected, but it was not ascertained. Last but not least, cases may exist in which deciding if the listing is real or fake may be a very hard, and subjective, task. For this reason, we believe that, in the specific case of our study, predicting "days-on-the-market" is more reliable and appropriate than "fraud."

*1.2. Study Objectives.* This paper aims to present a systematical approach based on data analysis techniques, in particular, predictive modelling, applied to the problem of identifying frauds in real estate advertisements. The core study objectives of this work are

- (1) Predicting days-on-the-market for housing
- (2) Identifying features which make a property more attractive

Concerning the first point, it should be pointed out that attaining a highly accurate model which can predict how long a given property will remain on the market is a compound task: first of all, data containing all required information are not currently available, and in general they are difficult to collect due to the high amount of not quantitatively measurable factors. Secondly, days-on-the-market is a variable that is highly influenced by a variety of dynamics, dependencies, and features such as location, price, and details regarding the condition of an apartment.

The second objective is closely related to the first one. In fact, different studies, focusing on predicting housing prices, identify and measure the effect of common housing attributes on the price. Here, the point of interest is to measure the effect of such features on days-on-the-market and identify what makes an apartment more attractive to a customer. The answer to this question will support the product development of Homeheed and will allow the team to provide customers with listings with a higher probability of being sold/rent.

*1.3. Study Relevance and Importance.* With respect to real estate market challenges in Bulgaria, this project will allow us to (i) explore historical market data and gain valuable insights, which will permit a more accurate estimation of the listings; (ii) streamline the market entry program significant for the revenue stream and ROI planning; and (iii) further support the design of the technology which can assess a property availability. The outcome of this work will help to determine important housing attributes and so will serve as a proposal for restructuring databases by introducing new features for future data mining projects.

Furthermore, the work aims to contribute to a platform that serves as a tool to achieve more fair competition on the Bulgarian unregulated real estate market. It is assumed that the findings can enhance the business model, the technology, and the market entry strategy. Data analysis techniques can influence positively the development of the system and enhance it by making it more sustainable, efficient, and transparent, as well as by improving customer satisfaction and general citizens' experience in the process of searching for a new home.

Several previous studies can be found about applying data science to housing price prediction. In different periods when the real estate market worldwide has recorded changes, bloom, or descent, questions regarding the accuracy of property value assessment have been raised. The instabilities made housing predictive models the subject of research among scholars. A literature review shows methods that can estimate the price of a property based on different features and in comparison to similar objects. However, the question of how long a listing will be on the market was not extensively studied yet. This work aims at filling this gap, by highlighting the importance of the concept of *day s\_on\_market*, as a significant feature in terms of investment and ROI planning.

*1.4. Manuscript Organization.* The paper is organized as follows: Section 2 contains a critical review of the literature. In Section 3, we describe the available data. Section 4 presents the data preprocessing phase that has allowed us to obtain a compact and informative dataset, to be used as an input for the machine learning algorithms. Section 5 discusses the obtained experimental results. Finally, Section 6 concludes the work and suggests ideas for future research.

Last but not least, Appendix A offers a presentation of the used machine learning algorithms.

## 2. Previous and Related Work

The application of data mining in the real estate has become widely popular in the last few years. Researchers and companies use a variety of prediction techniques to capture fluctuation periods and the factors influencing them to analyse the market trend through regression and machine learning algorithms, to describe property types by clustering heterogeneous housing data, including house attributes and geosocial information, and to find customer habits to determine sales strategies [3].

Several studies have appeared so far analysing the real estate prices. On the other hand, analysing the *day s\_on\_market* (DOM) and the popularity of a property is still an understudied area. DOM is an essential factor although challenging to measure for real estate listing since it is highly correlated with the popularity of a housing object. The literature review showed that some publications are focused on studying the relationship between DOM (or time on the market) and different factors, such as prices, brokers/broker agencies, marketing strategy, and others [4, 5]. The results show contradictory findings. For example, Belkin [6] suggest that DOM and sale price of housing have no relationship between each other, while Miller [7] uses DOM to explain sales prices and shows a positive correlation between these two variables. Other studies illustrate that DOM and sale price has an associated connection due to various factors such as quality, listing strategy, and real estate agency, which adds complexity to the relationship [8].

Hengshu Zhu [9] presents a study in which the authors measure the liquidity of the real estate market by developing an approach for predicting DOM. The authors use multitask learning-based regression to overcome the problem of location dependency and further compare the results by using baseline models such as linear regression (LR), Lasso, location-specific linear regression, decision trees (DTs), and others. Their results illustrate also the mutual importance of the different studied features. The performance of the method is assessed using real-world data and a designed prototype of a system showing the practical use of their analysis, which can be used as a reference for Homeheed software [9].

Ermolin [10] uses DTs to predict DOM within 7 days. The author makes the assumption that any accuracy for more than a week should be considered arbitrary due to the seasonality of the housing market. In Ermolin's work, it was concluded that geospatial features did not add value to the prediction [10].

Chao Mou [11] proposes a system to predict short DOM. This work provides a framework that can serve as a reference to estimate the market value of a housing property. The authors make the assumption that true market value can be approximated to the listing price when real estate agents

have similar offers because few brokers would be willing to sell a property at a much lower price. Further, housing with short DOM is detected by comparing their listing prices and estimated market values [11].

### 3. Data Description

The dataset provided by Homeheed consists of more than 550.000 observation points and 19 variables, describing apartments, houses, stores, restaurants, garages, lands, etc., for rent or sale in Sofia, Bulgaria. The data are collected from the main online property listing website and contains historical information for the listings published in the period from 01.07.2015 to 01.07.2018. Table 1 lists the features which characterize a listing from the dataset, with the respective description.

The dataset contains both qualitative and quantitative variables. The variables *date\_first/last\_seen* describe the dates when a listing has been online for the first time and in which it became not available anymore, respectively. These two variables are used for the creation of the dependent variable (the variable that the proposed system aims at predicting) that we call *days\_on\_market*. The variable *city* is constant for all observation points, namely, Sofia city, and so it will be removed from the dataset, as it does not add any useful information for the model. Also, the variable *broker\_name* will not be taken into consideration due to both poor quality (most of the names are in Cyrillic) and data privacy issues. Concerning the variable *lister\_username*, also some data privacy issues could exist, but they have been solved by encoding names, using unique numeric ids. The relevance of these ids will be examined for the model development since this might provide further insights for fraud detection. The rest of the variables describe a property in terms of location, value, and specific attributes.

The variables *specials* and *description* contain details about the listed property. The variable *description* provides full text about the property amenities, while *specials* contains only keywords characterizing the exterior or interior of a property. We decided to remove from the dataset the variable *description* since the content is in Cyrillic. However, the features provided by the variable *specials* summarize some of the main attributes of a property and will be further analysed with some text mining techniques, as explained in Section 4.

The variable *floor* mainly informs about the floor on which a property is, as well as the total number of floors in the building, e.g., “5 of 12”. However, it also contains misplaced values regarding the area of the garden in m<sup>2</sup> for houses and villas, or some other words which purpose for the dataset cannot be identified and are considered as mistakes. For explorative purposes, a new variable called *space\_m2\_garden* was created.

Finally, the variable *build\_type* contains several pieces of information concerning the building, namely, the type of bricks used to build it, beams, MICCS, type of concrete structure employed, sliding formwork (SF), panel, and under construction, together with the year when the building was constructed.

To provide the reader with a visual understanding of the frequency distribution of the selected property types, Figure 2 illustrates the total amount of listings of every property and their distribution by real estate owner types. As we can observe, most of the listings are provided by real estate agencies.

However, as discussed above, the collected data about DOM of listings made by real estate agencies may not be reliable and in some cases may even be not real. The missing piece of information here is a variable which states whether a listing was really available or not at the moment when it was published. Since this information is not available and hard to be collected, building a model that predicts DOM for listings made by real estate agents will be highly biased. To overcome this issue, we took the decision of removing from the dataset all listings made by agencies.

Generally speaking, different profiles of real estate owners/agents who publish listings are assumed to have different behavior. It is a point of interest to observe the distribution of DOM.

Figure 3 shows that listings published in July have the maximum DOM for most of the property types.

### 4. Data Preprocessing

In this section, we present the methods used to transform the data, to obtain a more compact and informative dataset. This new dataset will be given as the input to the computational methods that will generate a predictive model for the houses days on market.

*4.1. Univariate Analysis.* Different statistics and methods will be used in this section to understand the individual impact of continuous (or simply numerical, as they will be called in the continuation), textual, and categorical variables.

*4.1.1. Numeric Variables.* Figure 4 reports some basic statistics describing the numeric variables of our dataset, including measures for central tendency, variability, standard deviation, and several others. The study was performed for the numerical features available in the original dataset (marked with red) and also for some additional features created for the purpose of this work.

For normally distributed data, approximately 95% of the values lie within 2 standard deviations from the mean. For this reason, observing our data, we can state that only *year\_end* and *year\_start* can be assumed as normally distributed. The standard deviation is not the most suitable measure to study data distribution when the values in a variable are not normally distributed. On the other hand, histograms are one of the most common visual tools to quickly investigate data and make conclusions about central tendency, spread, modality, shape, and outliers. Furthermore, histograms support the illustration of the data distribution and serve as a method to envision skewness and kurtosis. Skewness measures the asymmetry, while kurtosis determines “peakedness” compared to the normal distribution. These measurements are useful for the

TABLE 1: Variable list and description.

Variable name	Description
<i>lid</i>	Listing ID
<i>date_first_seen</i>	The date on which the listing of a housing object first appeared online
<i>date_last_seen</i>	The date on which the listing of a housing object was last seen online
<i>rent_or_sell</i>	Variable which indicates whether a housing object is for renting or selling
<i>property_type</i>	Identifies the type of property being for sale or rent
<i>city</i>	The city in which a property is located
<i>neighborhood</i>	The neighborhood in which a property is located
<i>street</i>	The street on which a property is located
<i>space_m2</i>	The area of a property in m <sup>2</sup>
<i>price_in_bgn</i>	The price of a property in national currency
<i>price_in_currency</i>	The price of a property in different currency
<i>currency</i>	Specifies the currency
<i>build_type</i>	Specifies the building material type
<i>floor</i>	Names the floor on which is a property
<i>specials</i>	Gives details about the condition of a property
<i>description</i>	Text description of a property
<i>n_photos</i>	Number of photos which a property has included in the listing
<i>lister_type</i>	Specifies whether the listing was made by owner, agent, investor, etc.
<i>lister_username</i>	The name of the account from which the listing was made
<i>broker_name</i>	The name of the broker (company) which stays behind the listing

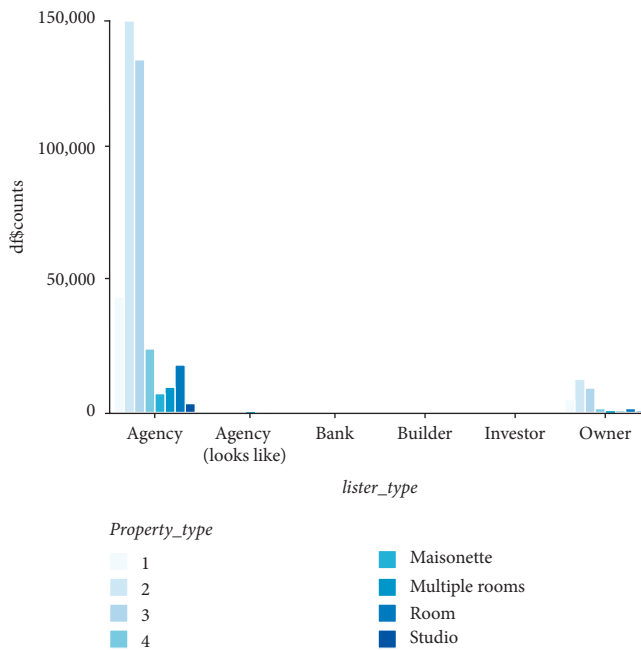


FIGURE 2: Property types provided by the real estate owner type.

differentiation of extreme values. In positively skewed (right-skewed) data values far from the mode are more regular and usually the mean is greater than the mode. If the skewness is negative, then the mean is less than the mode. Regarding kurtosis, a positive one allows the interpretation that values which are far from the central tendencies are more probable, as well as that the shape is more centrally peaked, but the tail is greater. When the kurtosis is negative, then the peak has wider “shoulders,” compared to the normal distribution [12]. Figure 5 shows the distribution of some of the variables in our dataset.

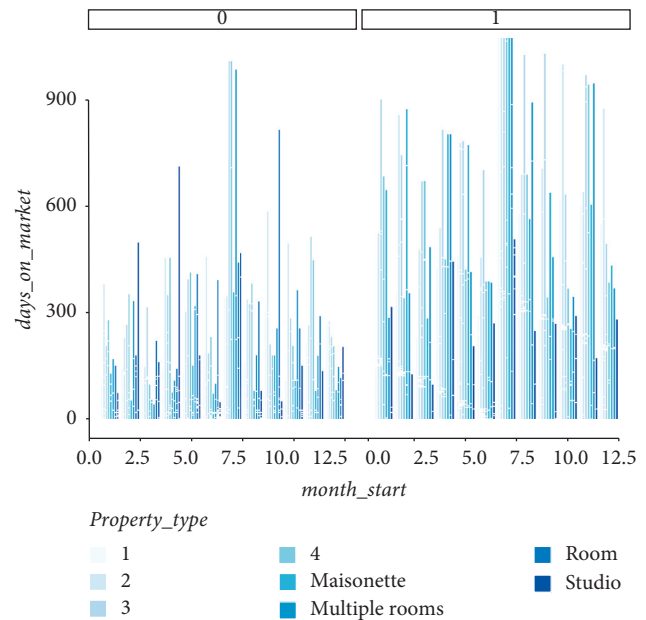


FIGURE 3: DOM based on month when a listing was published.

Additionally, Table 2 represents the values of the skewness and the kurtosis of the variables. The negative values imply that the distribution of the data is skewed to the left or negatively. The closer to zero, the slighter the skewness, and likewise if the number is more distant from zero. Oppositely, when the value is greater than zero, the distribution of the variable data is positive/skewed to the left. Concerning the kurtosis, a value smaller than 3 implies negative or flat and wide distribution, while a value larger than 3 should be interpreted as high and “slim” distribution [13].

	<i>space_m2</i>	<i>price_in_bgn</i>	<i>price_in_currency</i>	<i>n_photos</i>	<i>Year_start</i>	<i>Month_start</i>	<i>Day_start</i>	<i>Year_end</i>	<i>Month_end</i>	<i>Day_end</i>	<i>floor_new</i>	<i>total_floors</i>	<i>year_built</i>
nbr.val	33740	33398	33408	33740	33740	33740	33740	33740	33740	33740	32517	32517	10586
nbr.null	0	0	0	6104	0	0	0	0	0	0	1356	0	0
nbr.na	0	342	332	0	0	0	0	0	0	0	1223	1223	23154
min	1	39	20	0	2015	1	1	2015	1	1	0	1	1900
max	8065	5280741	3617400	17	2018	12	31	2018	12	31	24	26	2021
range	8064	5280702	3617380	17	3	11	30	3	11	30	24	25	121
sum	2684177	2.71E+09	1.42E+09	265439	68036243	217331	492121	68039886	220340	500516	127731	225508	21146836
median	70	1271	850	8	2017	7	14	2017	7	15	3	6	2006
mean	79.55474	81022.57	42467.9	7.86719	2016.486	6.441346	14.58568	2016.594	6.530528	14.8345	3.92813	6.93508	1997.623
SE.mean	0.584747	657.695	376.769	0.03091	0.005458	0.01761	0.049375	0.005454	0.017644	0.04971	0.015435	0.017281	0.190854
CI.mean.0.95	1.146124	1289.105	738.4804	0.060584	0.010699	0.034516	0.096777	0.01069	0.034583	0.097434	0.030254	0.033871	0.37411
var	11536.67	1.44E+10	4.74E+09	32.2351	1.005262	10.46293	82.25476	1.003647	10.50374	83.37514	7.747233	9.710418	385.5978
std.dev	107.4089	120194.5	68865.29	5.677596	1.002628	3.234646	9.069441	1.001822	3.240947	9.130999	2.783385	3.116154	19.63664
coef.var	1.350126	1.48347	1.621584	0.72168	0.000497	0.502169	0.621804	0.000497	0.496276	0.615525	0.708578	0.449332	0.00983

FIGURE 4: Numerical variables basic statistics.

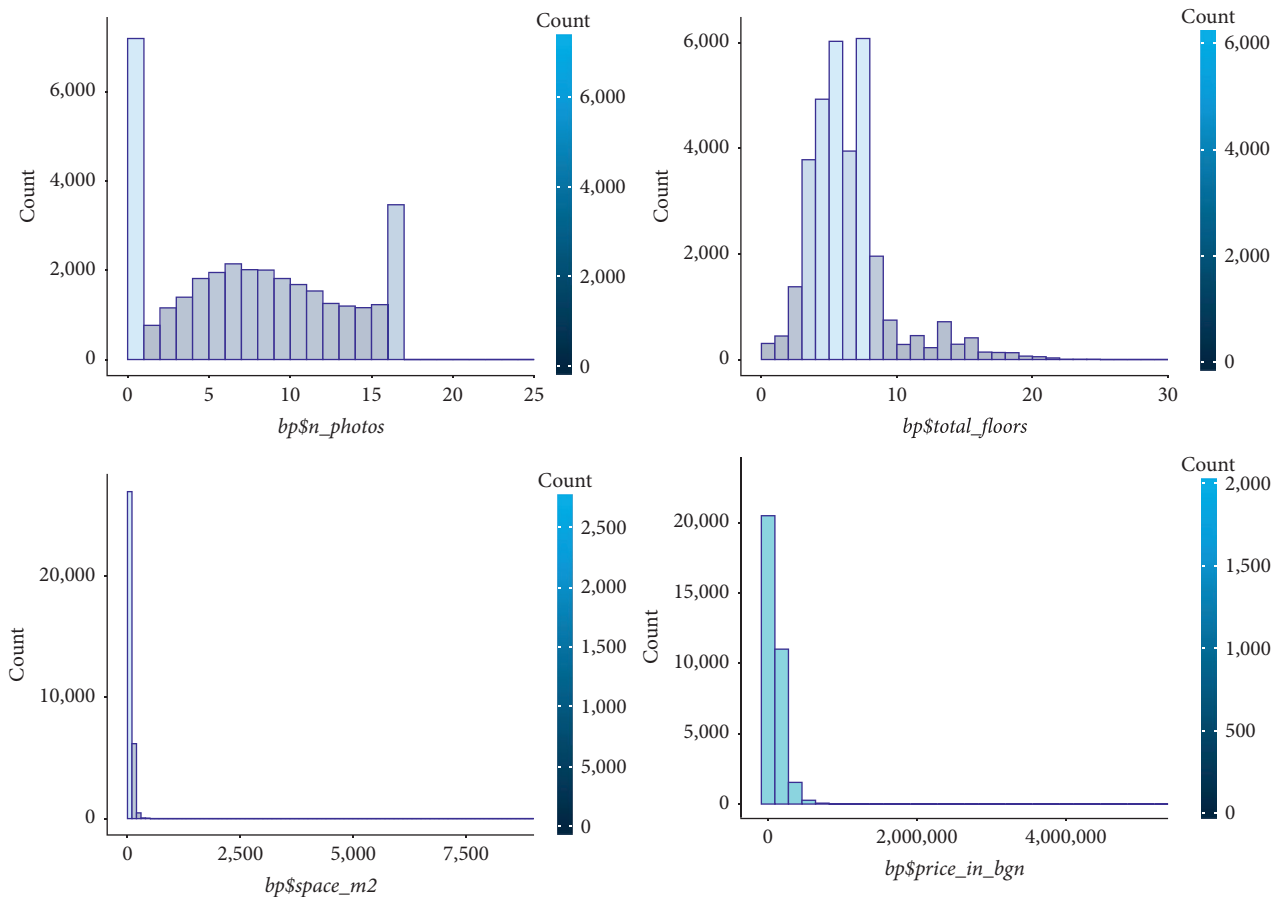


FIGURE 5: Histograms of some numerical variables.

**4.1.2. Textual Variables.** Textual features are extracted from the main description of the house and collected in the variable called *specials*. They include keywords that describe a property regarding its construction and/or amenities. To give the reader an overview of the features which are

generally used in a listing, a word cloud is created. Figure 6 shows that “elevator” and “furniture” are the most recurrent words in the description of houses, followed by “internet” and “brick.” These features have to be extracted through text modelling as an essential part of the data preparation.

TABLE 2: Skewness and kurtosis values of the variables.

Variable	Skewness_of_variable	Kurtosis_of_variable
<i>space_m2</i>	58.55332354	4076.35939
<i>price_in_bgn</i>	6.276439942	166.7471085
<i>price_in_currency</i>	13.76938472	556.4922087
<i>n_photos</i>	0.09699499	1.811480527
<i>year_start</i>	-0.019183275	1.928599009
<i>month_start</i>	-0.005586014	1.954408133
<i>day_start</i>	0.099231688	1.790969492
<i>year_end</i>	-0.085003855	1.933427314
<i>month_end</i>	-0.047972758	1.945445542
<i>day_end</i>	0.066149649	1.778833735
<i>floor_new</i>	1.464791477	6.643659231
<i>total_floors</i>	1.585019943	6.768881337
<i>year_built</i>	-1.06531323	3.845297287



FIGURE 6: Word cloud of the variable specials.

Organized text is usually represented by a table with one token per row. A token is an important component of the text, for instance, a word, which is noteworthy for analysis, and tokenization is the practice of separating the text into tokens. A token can also be a sequence of  $n$  words (called  $n$ -gram) or even a complete sentence. For instance, in our dataset, several combinations of words such as “entrance control” exist, and they are called bigrams. Also, the variable *specials* itself contains multiple words which define property features. Thus, it is interesting to examine the relationship and co-occurrence of words. Figure 7 shows the consecutive sequences of words which can be found in the *description* of a property.

Not only “furniture” and “elevator” are the words that appear most frequently, but also the combination between these two words occurs repeatedly. To examine the correlation between words, the so-called  $\phi$ -coefficient, which is a measure for the binary association of features, was used. This coefficient quantifies the correlation between the probability of two words appearing together and of the same two words appearing independently. Figure 8 illustrates the

four words which appear most often and the words which are most often associated with them. Here, it should be mentioned that, e.g., “under” and “construction” have the same  $\phi$ -coefficient related to “brick” since “under construction” is a predefined special bigram. The same is valid for several more word combinations. Interesting point was to study the correlation between the words “furniture” and “elevator” due to their common occurrence, but the analysis showed a  $\phi$ -coefficient of only 0.096.

**4.1.3. Categorical Variables.** The last type of variable that can be found in our dataset is the categorical variable. Table 3 shows that owners are the main listing publishers among the studied ones, and that flats with 2 or 3 rooms have the highest supply level for both rents or sell.

The variable *neighborhood* contains a large number of possible values. The center region offers slightly more listings, but still, none of the neighborhoods preponderates significantly.

Appealing fact, shown in Figure 9, is that the variable *type\_built* usually contains significant values only when a property is listed for selling. When a listing is marked for renting, then the construction type is often unknown. This should be considered during the management of missing values.

**4.2. Management of Missing Values.** Figure 10 provides an overview of the missing values in the original dataset. The variable *space\_m2\_garden* has the greatest amount of missing values since it makes sense only for houses and villas. Nevertheless, for the other types of dwellings, this variable can be informative, and so it was left in the dataset. On the other hand, Homeheed currently concentrates its interest and service to properties which can generically be clustered as “home.” Therefore, the focus of this work is only on properties listed for living purposes, mainly apartments. The type of apartment is stored in the variable *property\_type*, it can assume values such as 1, 2, 3, 4, or multiple rooms, studio, maisonette, and room, and it has no missing values. Our data contain more than 400.000 observations for this type of dwelling. Other types of listings will not be analysed and will be excluded from the dataset.

Other variables with missing values are *street*, *broker\_name*, and *build\_type*. Given their high percentage of missing values, variables *street* and *broker\_name* were removed from the dataset. Concerning the variable *build\_type*, as it will be discussed later in this document, a decision was taken to split the information contained in this variable, thus creating two new variables: *year\_built* and *type\_built*, containing information relative to the year of building and the building material, respectively. Interestingly, both these variables have a large number of missing values for houses that are for rent, while they present no missing values when houses are for sale. Nevertheless, we decided to remove *year\_built* and *type\_built* from the dataset. In fact, even though both variables contain information for properties that are for sale, the imputation or prediction for 50% of the

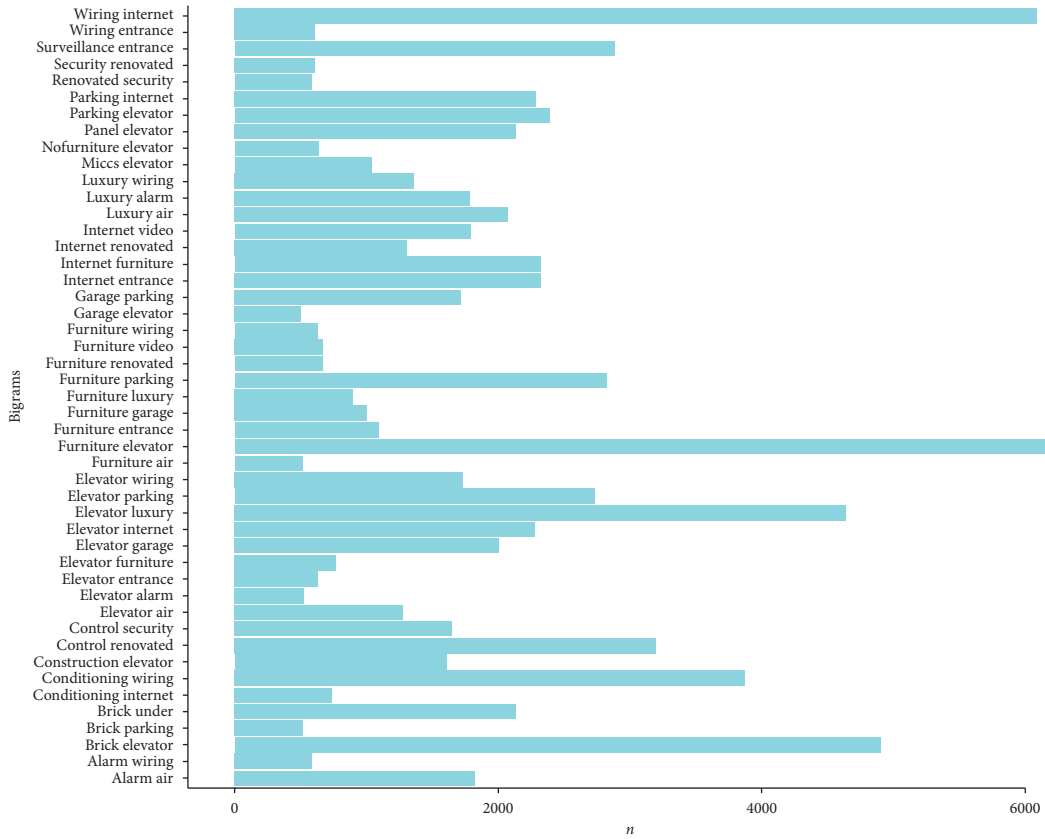


FIGURE 7: Relationship between the words in the description.

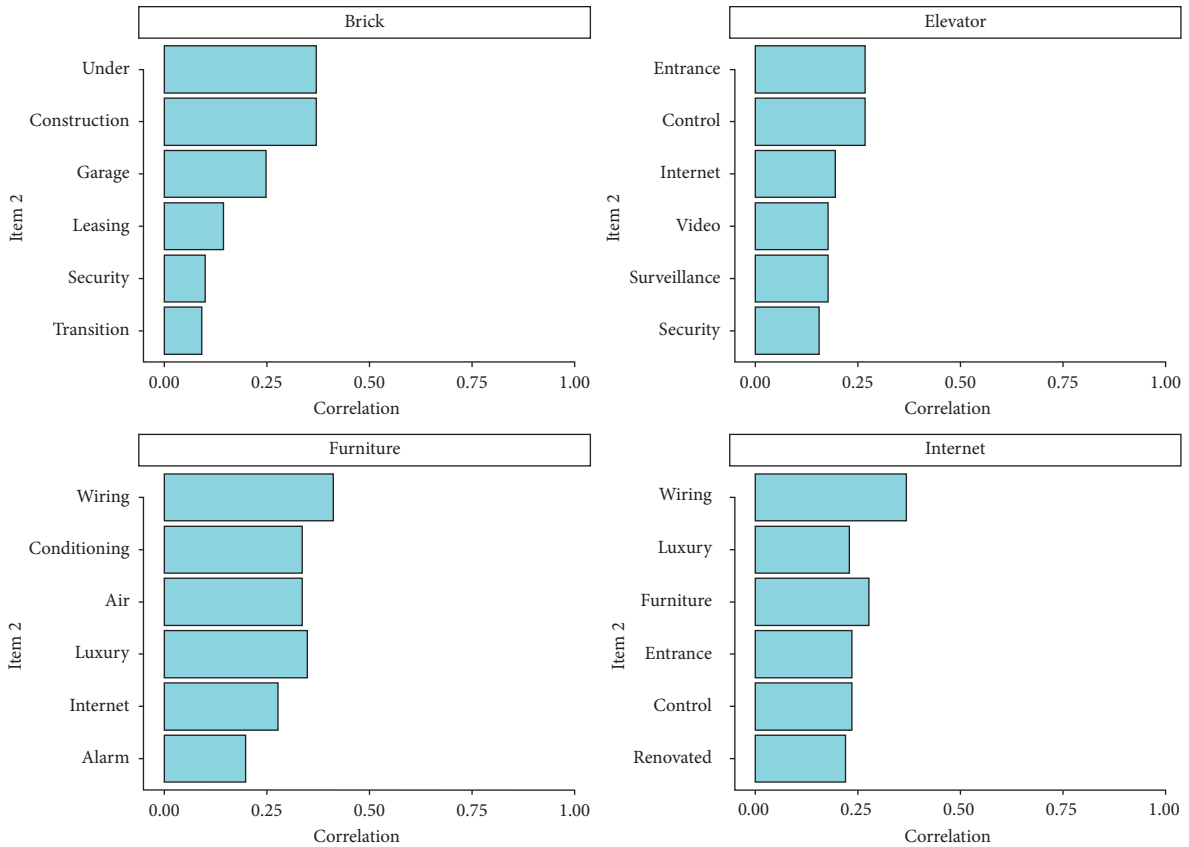


FIGURE 8: Words correlation.

TABLE 3: Cross table for property type by listing provider and by rent or sell.

	1	2	3	4	Maisonette	Multiple rooms	Room	Studio	Total
<b>1</b>									
Agency (looks like)	7	56	54	15	2	5	1	7	<b>147</b>
Bank	0	0	0	0	0	0	0	0	<b>0</b>
Builder	2	4	2	0	0	1	0	0	<b>9</b>
Investor	1	10	7	0	1	1	1	0	<b>21</b>
Owner	3251	7586	4027	455	174	191	1438	449	<b>17571</b>
<b>2</b>									
Agency (looks like)	93	615	695	159	63	89	0	44	<b>1758</b>
Bank	2	24	25	3	2	8	0	6	<b>70</b>
Builder	11	85	92	21	6	2	0	0	<b>217</b>
Investor	11	121	156	29	10	14	0	0	<b>341</b>
Owner	1571	5015	5165	970	335	393	0	157	<b>13606</b>
<b>Total</b>	<b>4949</b>	<b>13516</b>	<b>10223</b>	<b>1652</b>	<b>593</b>	<b>704</b>	<b>1440</b>	<b>6663</b>	<b>33740</b>

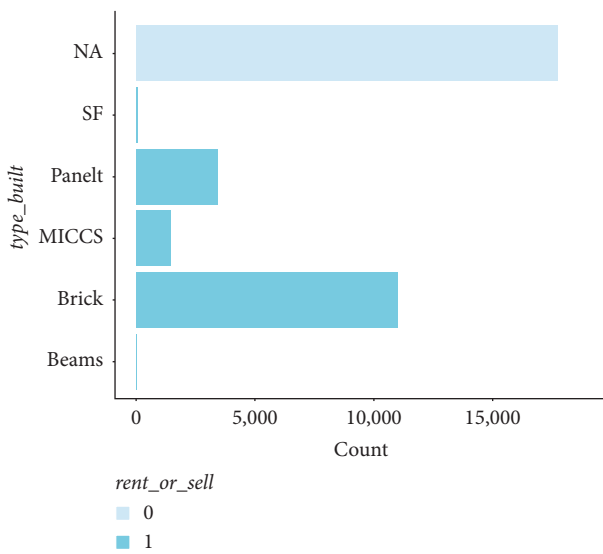


FIGURE 9: *type\_built* by rent or sell.

observation points would be either extremely time consuming or not be reliable.

Among the other variables with a significant amount of missing values, we also decided to remove the variable *lister\_username* from the dataset.

4.3. *Management of Outliers.* The next step before the transformation of the data is the detection and management of outliers. Outliers may have a significant impact on the data if no actions are taken. For instance, they can increase the error discrepancy and decrease the supremacy of numerical tests. Also, outliers can affect normality, as well as the fundamental hypothesis of some statistical models. In practice, an outlier can be interpreted as a value which is 1.5 times the IQR (interquartile range) more extreme than the quartiles of the distribution. The most applicable and useful way to detect an extreme value is by visualizing a boxplot. Figures 11–14 illustrate the boxplots of four features. Extreme values that require attention can be seen, as well as long tails in the distribution of the values.

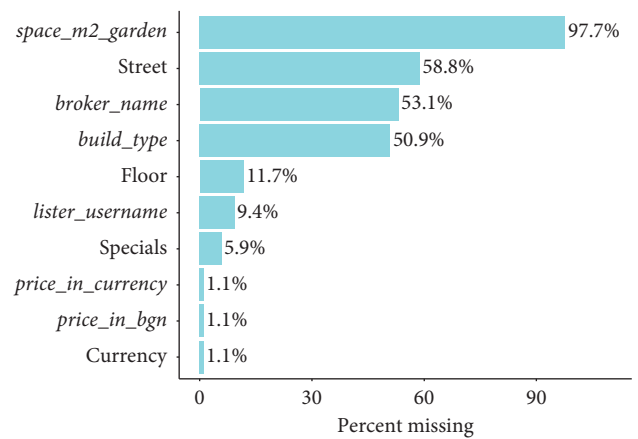


FIGURE 10: Missing values in the studied dataset.

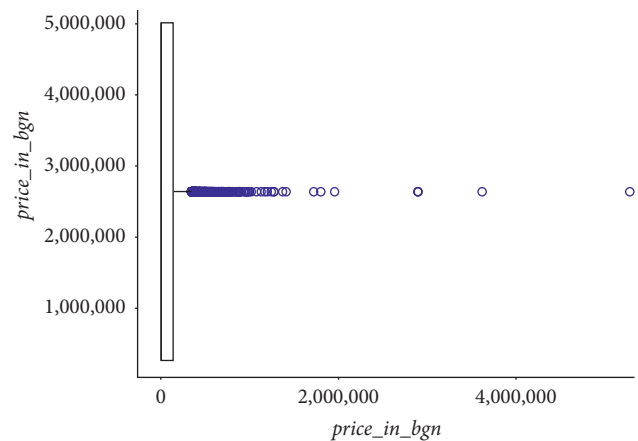


FIGURE 11: Boxplot for *price\_in\_bgn*.

One method which can support a better understanding of these outliers is the breakdown of the variable, which is observed based on the values in another feature. This is called multivariate analysis. Figures 15 and 16 show an example with separating *space\_m2* based on *property\_type* and by price in currency Bulgarian lev. The scatterplots show



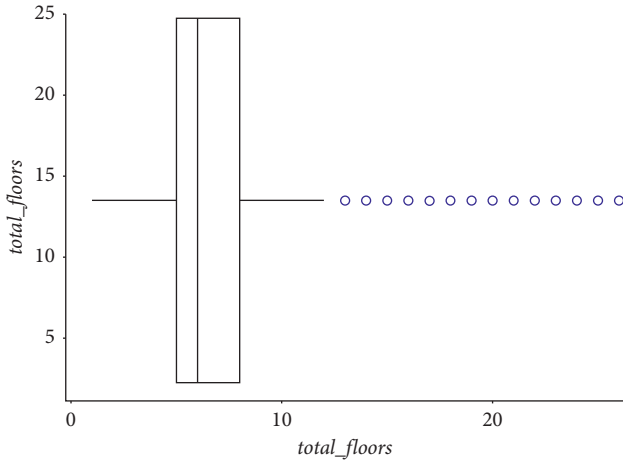


FIGURE 12: Boxplot for *total\_floors*.

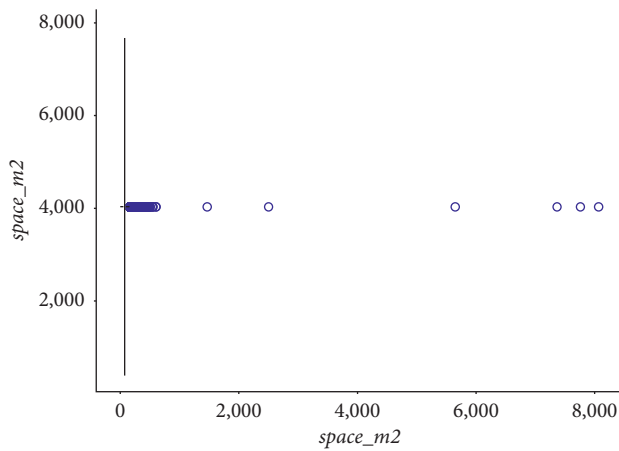


FIGURE 13: Boxplot for *space\_m2*.

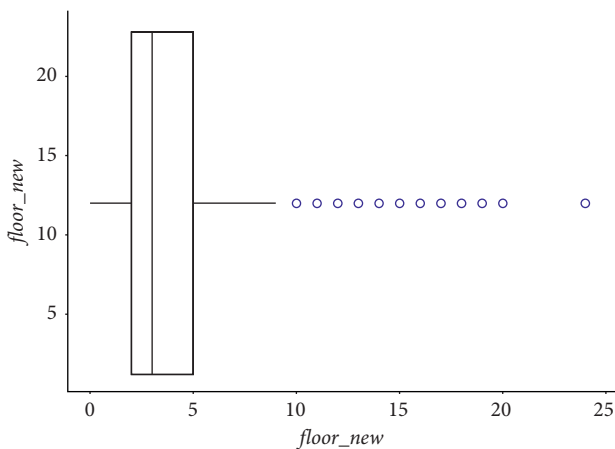


FIGURE 14: Boxplot for *floor\_new*.

that there are mainly outliers in 1-, 2-, or 3-room apartments and that places with extreme space have extreme prices.

Extreme points were detected only in two variables, *price\_in\_bgn* and *space\_m2*, and their total amount was

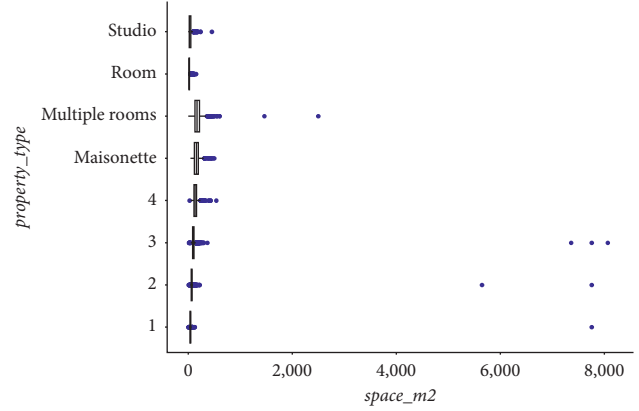


FIGURE 15: Outliers of *space\_m2* based on *property\_type*.

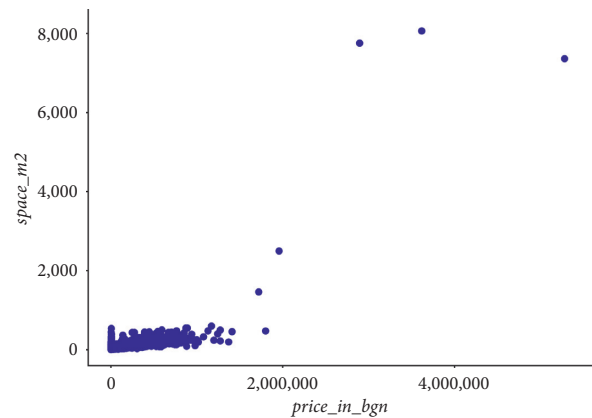


FIGURE 16: Outliers of *space\_m2* based on *price\_in\_bgn*.

small and inconsequential for the overall analysis, so these points were removed from the dataset.

**4.4. Data Transformation.** In our dataset, different variables have diverse ranges of possible values. Given that some algorithms base their functioning on the distance between observation points to make a prediction, a common scale is needed to assure that none of the features will be dominant. Further, as revealed previously, the distribution of the data in some variables shows skewness, which might represent a difficulty for some of the studied machine learning algorithms, and that can be alleviated by scaling the data. Common normalization methods are Min-Max, which scales the range between 0 and 1 and Z-score, which scales values between -1 and 1. In this work, the values have been normalized using Min-Max.

Table 4 reports a set of other modifications that were made to the variables, to obtain a more informative, and potentially more useful, dataset.

Among the other transformations reported in Table 4, it is worth discussing how we decided to transform the textual variable *specials*. The first task is the removal of punctuation because it has no added value to the information. Furthermore, all letters are converted to lower case. This prevents multiple extracted copies of the same word. Since the

TABLE 4: Recoding of original variables.

Variable	Original version	Recorded version
<i>date_first/last_seen</i>	Both variables are in format “yyyy-mm-dd”	6 new variables were created, namely <i>year_start/end</i> ; <i>month_start/end</i> ; and <i>day_start/end</i>
<i>rent_or_sell</i>	“rent” and “sell”	Recoded to binary on=0 rent, 1 for sale recoded completely to numbers-1, 2, 3, 4, 5, 6, 7, 8
<i>property_type</i>	1234 maisonette multiple rooms room studio	An id number was assigned for the different neighbourhoods
<i>lister_type</i>	Contains many character values with the name of the neighbourhood	Since no strict condition for the recognition between investor and builder was found, the value “agency (looks like)” was randomly replaced to be either builder or investor. New variables with codes from 1 to 4 were created
<i>build_type</i>	Originally the variable contains year and building material	The variable was split in two new variables- <i>year_built</i> and <i>type_built</i>
<i>specials</i>	Text variable in the format [\word1\,”\word2\,”\word3\,”. . .]	Binary variables for each word indicating the existence or lack of this feature
<i>floor</i>	Originally in the format for example “5 to 10”	Split in two new variables <i>floor_new</i> and <i>total_floors</i>

variable itself contains only keywords, some basic pre-processing procedures, such as stop words removal or stemming (removal of suffices), were not executed. However, the last step was the conversion of a single word into binary variables.

Finally, two new variables were additionally introduced-*price\_per\_m2* and *n\_features*. The first one is calculated based on *space\_m2* and *price\_in\_bgn*. The second one represents the total number of features, including as keywords in the *description*, available for a listing.

**4.5. Feature Selection.** The original dataset at our disposal included 19 variables. However, the transformations presented so far have increased the number of variables up to 54, so variable selection techniques have to be applied to choose the most valuable predictors for the model. Filter methods are usually employed as a data preparation step, to select features. First of all, a study of the correlation coefficients was performed, to have an idea of the relationship between the continuous variables. Figures 17 and 18 show both the heat matrixes of Pearson and Spearman correlation coefficients.

Both these figures show a significant correlation only between the expected *year\_start* and *year\_end* and between *price\_bgn* and *price\_per\_m2*.

Correlation alone can limit the detection of multicollinearity since it is only pairwise. One of the techniques which support the detection of more complex relationships is the usage of eigenvalues. A small magnitude shows that there is no multicollinearity, while a high range between the values is a signal for significant multicollinearity, which is the case here. The variance inflation factor (VIF), which indicates how much the variance of a regression coefficient is overestimated due to multicollinearity, can be calculated. The minimum possible VIF is equal to 1 and, as a rule of thumb, results between 5 and 10 are considered as indicators for the problem. In our dataset, *year\_start* and *year\_end* showed extreme results above 20 and *price\_bgn* has a result

around 9. To solve this issue, we have decided to remove those variables from the dataset. To examine the significance level between the categorical variables and the target, the Kruskal–Wallis test was performed. A *p* value which is less than 0.05 indicates a significance level between the groups. Only the variables extracted from the *description*, *telephone\_exchange*, and *elevator* had a *p* value higher than 0.05. All the others, having a smaller *p* value, cannot be excluded from the dataset.

Based on both these filter methods, i.e., correlation and Kruskal–Wallis test, not a significant amount of variables can be excluded. To select the proper variables for the model, we applied an embedded method: Lasso regression. Figure 19 illustrates the variables sorted based on their importance and based on the Lasso method.

Observing Figure 19, we can remark that, among the 8 variables that have importance larger than 0.05, two variables are highly correlated between each other: *price\_per\_m2* and *price\_bgn*. Given that these two variables, practically, contain the same type of information, it makes sense to choose only one of them and to remove the other from the dataset. The obvious choice is to keep in the dataset the variable which has the highest importance according to the Lasso algorithm and disregard the other. For this reason, *price\_per\_m2* was kept in the dataset, while *price\_bgn* was removed.

In conclusion, the resulting, final dataset, which was given as an input to the machine learning methods to build the predictive models, contains 7 variables. These variables are

- (i) *lister*
- (ii) *rent\_or\_sell*
- (iii) *under\_construction*
- (iv) *space\_m2*
- (v) *brick*
- (vi) *furniture*
- (vii) *price\_per\_m2*

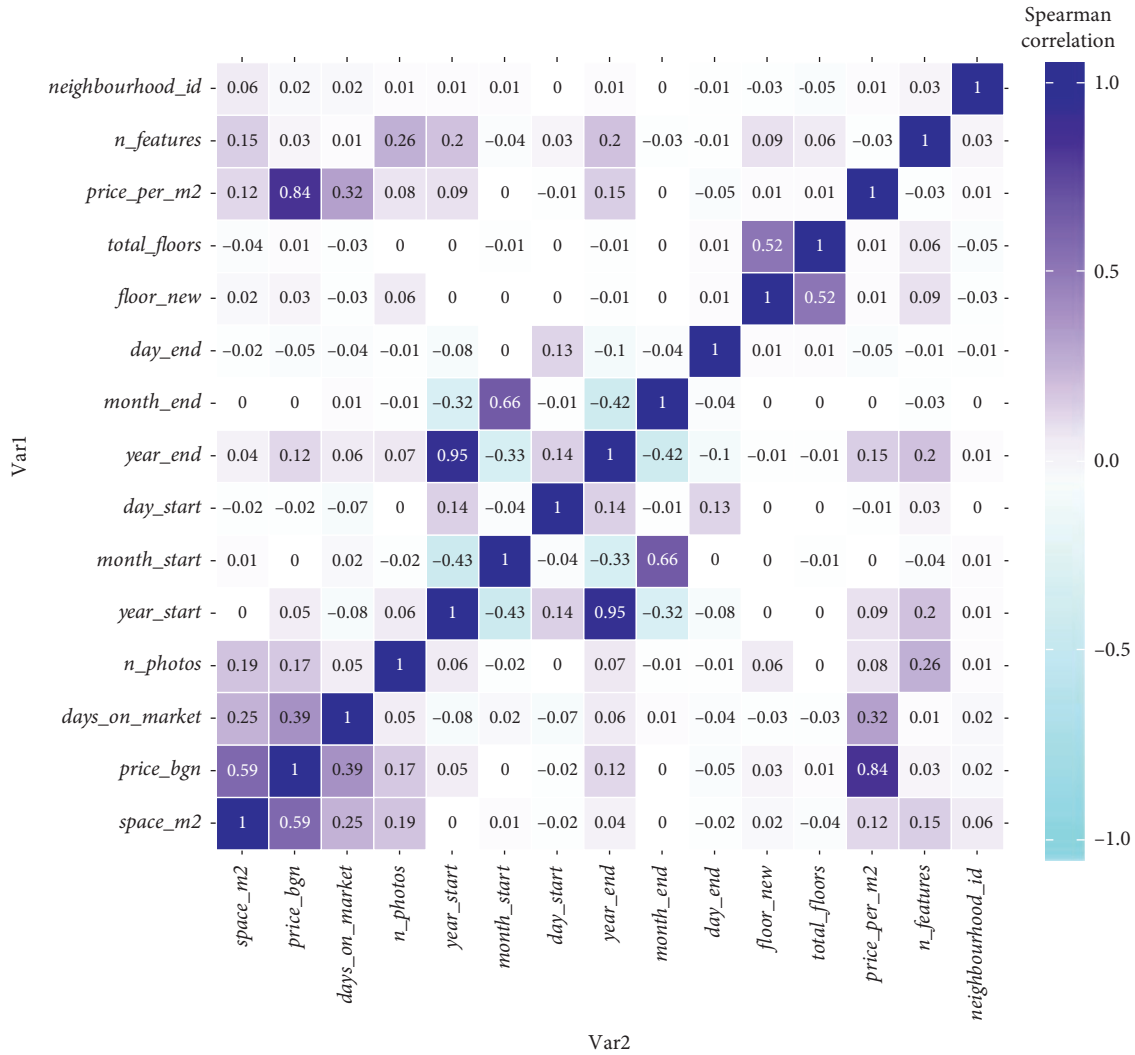


FIGURE 17: Spearman correlation heat map.

### 5. Experimental Results

All the results shown in this section have been obtained by performing 30 independent executions of each one of the studied machine learning algorithms. For each one of these executions, a different split of the available data into a learning set and a test set was considered. To obtain this split, 70% of the observations, selected at random with uniform distribution, were considered as the learning set, while the remaining 30% formed the test set. For each one of the studied machine learning methods, the training phase was executed on the learning set and the reported results are the results that have been obtained on the test set. When parameters needed to be set (it is the case, for instance, of the lambda parameter of Lasso, Ridge, and Elastic Net), only the learning set has been used to optimize the parameters' values, in the following way: the learning set was partitioned into 5 subsets and 5 different training phases were performed with different values of the parameters. In each one of these phases, 4 of these subsets were

used for training, while the other one was used for validation cyclically, so that each one of these 5 subsets was used once and only once for validation (5-fold cross-validation). The set of parameters that were used are the ones who allowed us to obtain the best median results on validation.

Let us begin the discussion of the experimental results by analysing the results obtained by Lasso, Ridge, and Elastic Net. Each of the three models was trained performing a grid search of predefined values of the parameter lambda. The value of lambda which minimizes the RMSE on validation was selected. The obtained values of lambda were 0.001 for Lasso, 0.0023 for Ridge, and 0.00014 for Elastic Net. With these values of lambda, the results shown in Table 5 were obtained:

As Table 5 shows, Lasso outperformed both Ridge and Elastic Net both in terms of minimum and median obtained RMSE.

Tables 6 and 7 show, for each one of the used features, the value of the coefficient that was obtained for each one of the studied algorithms.

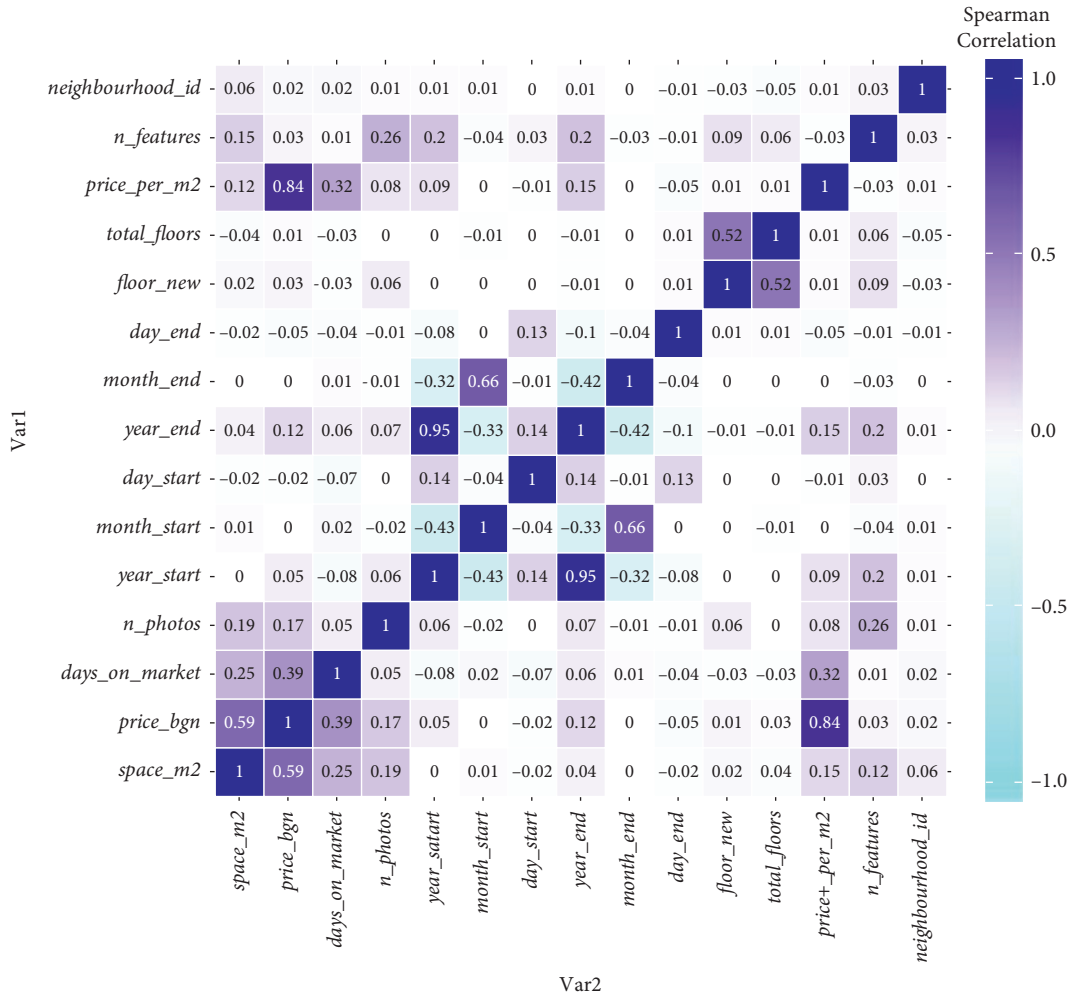


FIGURE 18: Pearson correlation heat map.

Tables 6 and 7 give an idea of the relative importance of the variables for each one of the studied algorithms. As we can see, none of the coefficients was equal to zero, except for the coefficient of variable *price\_per\_m2* for Lasso and Elastic Net. This confirms the appropriateness of the work that was done in the feature selection phase, corroborating that the 7 selected features are important for the prediction.

Let us now discuss the results obtained by Artificial Neural Networks. A grid search was performed to look for appropriate values of the number of hidden layers and the number of units per hidden layer. The results that returned the best median results on validation were 2 hidden layers, 3 units in the first hidden layer, and 2 units in the second hidden layer. Figure 20 illustrates the trained Neural Network that was possible to obtain with this configuration. The black lines give visibility on the connections and their weights, while the blue lines and values represent the bias term added on each step.

Figure 21 reports a comparison between the Neural Network and the Lasso regression, showing real vs predicted values. The closer the data points to the line, the better the model (theoretically, in the best-case scenario, the data

points should align perfectly with the line, when the RMSE is equal to 0).

The scatterplots show that the Neural Network has slightly more distant data points from the line than the Lasso. This gives a visual indication that Lasso may be a more accurate algorithm than Neural Networks for the studied problem. This qualitative result is also corroborated quantitatively: the RMSE obtained by the Neural Network is equal to 0.065, which means that Lasso performs slightly better.

Besides that one may also consider that Neural Networks are in general more complicated for interpretation and explanation.

Finally, to strengthen the robustness of the results obtained using Lasso, we perform a comparison against other well-known machine learning techniques commonly employed to address regression problems, namely, random forests (RFs), support vector regression (SVR), and k-nearest neighbors (K-NN). The reader is referred to the material in Appendix A for a brief overview of these techniques. To ensure a fair comparison, the values of the parameters characterizing the different techniques were chosen by performing a preliminary tuning phase. In particular, similar

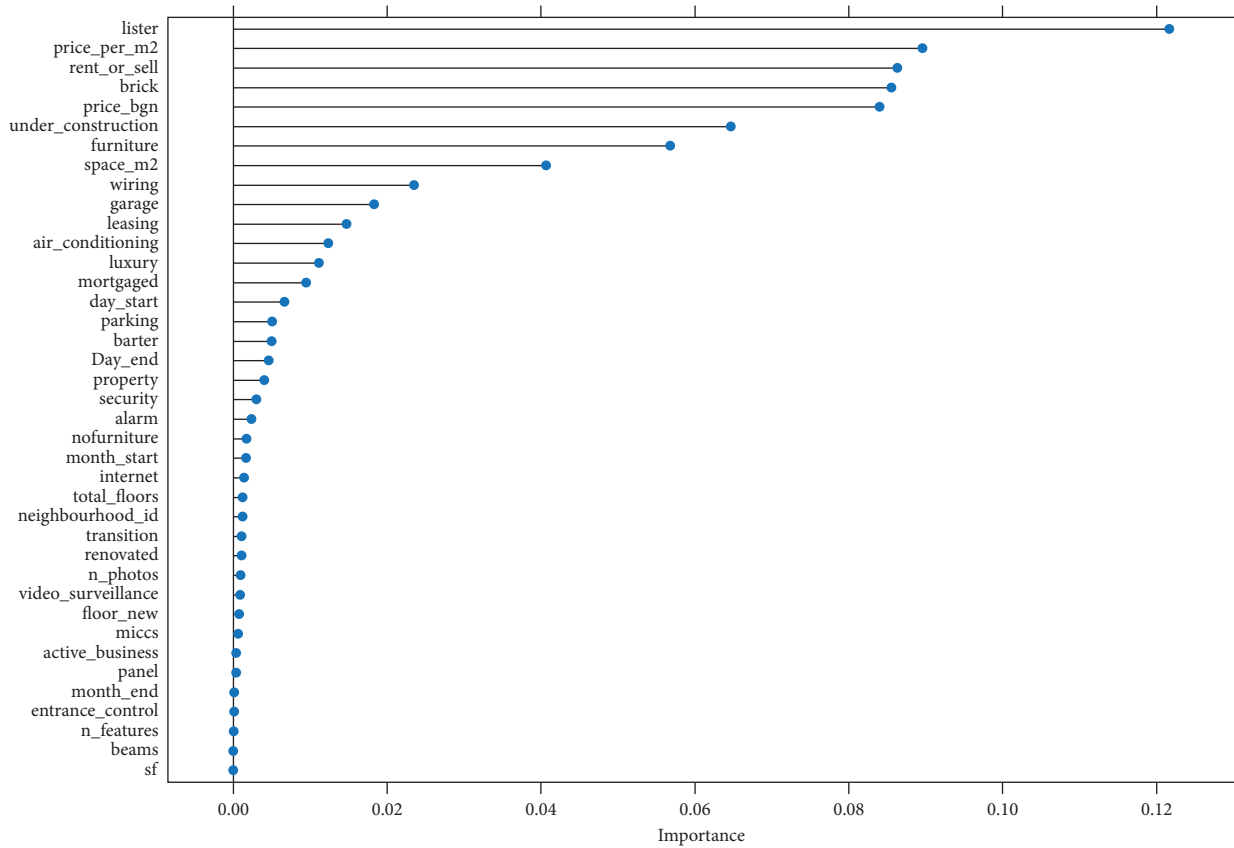


FIGURE 19: Lasso variables importance.

TABLE 5: RMSE obtained by Lasso, Ridge, and Elastic Net.

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NAs
Ridge	0.059	0.065	0.068	0.067	0.070	0.073	0
Lasso	0.056	0.064	0.066	0.067	0.070	0.081	0
Elastic	0.059	0.065	0.068	0.067	0.070	0.073	0

TABLE 6: Ridge regression coefficients.

Variable	Coef.
(Intercept)	0.007812742613
rent_or_sell	0.017990042310
space_m2	0.087954307932
brick	0.006890894870
furniture	-0.005648094327
under_construction	0.023199993261
price_per_m <sup>2</sup>	0.012193552241
lister	0.126633821677

to the experiments performed with Neural Networks and Lasso, we performed a grid search to determine the most suitable parameters for the considered machine learning techniques.

Focusing on RFs, the tuning phase returned a value of 70 for the maxnodes parameter (i.e., the parameter that limits the total number of nodes in each tree), 1000 for the number of trees in the random forest, and the function used to

TABLE 7: Coefficients of Lasso and Elastic Net.

Variable	Coef. Lasso	Coef. Elastic
(Intercept)	0.008262840044	0.006914658626
rent_or_sell	0.020773920773	0.021302049363
space_m2	0.081148777731	0.088121228126
brick	0.006277781636	0.006389871295
furniture	-0.004166719412	-0.004836976890
under_construction	0.020923002635	0.022619172277
price_per_m2	0	0
lister	0.127546013431	0.130465654935

measure the quality of a split in the trees was the Gini impurity. The RF with this configuration returned a median RMSE equal to 0.073.

Focusing on K-NN, it is important to highlight the importance of the parameter  $k$  (i.e., number of neighbors) on the performance of the model. In particular, the literature reports that a model with a very low value of  $k$  may tend to overfit the data, while higher  $k$  values can lead to underfitting. The grid search procedure returned a value of  $k$  equal to 15, leading to a final model with an RMSE of 0.064. Though this value is comparable to the one achieved with Lasso, K-NN has some weaknesses in the context of the problem studied here. In particular, K-NN requires an unbearable amount of time to return a prediction for unseen data since it has to compute the distance between each new

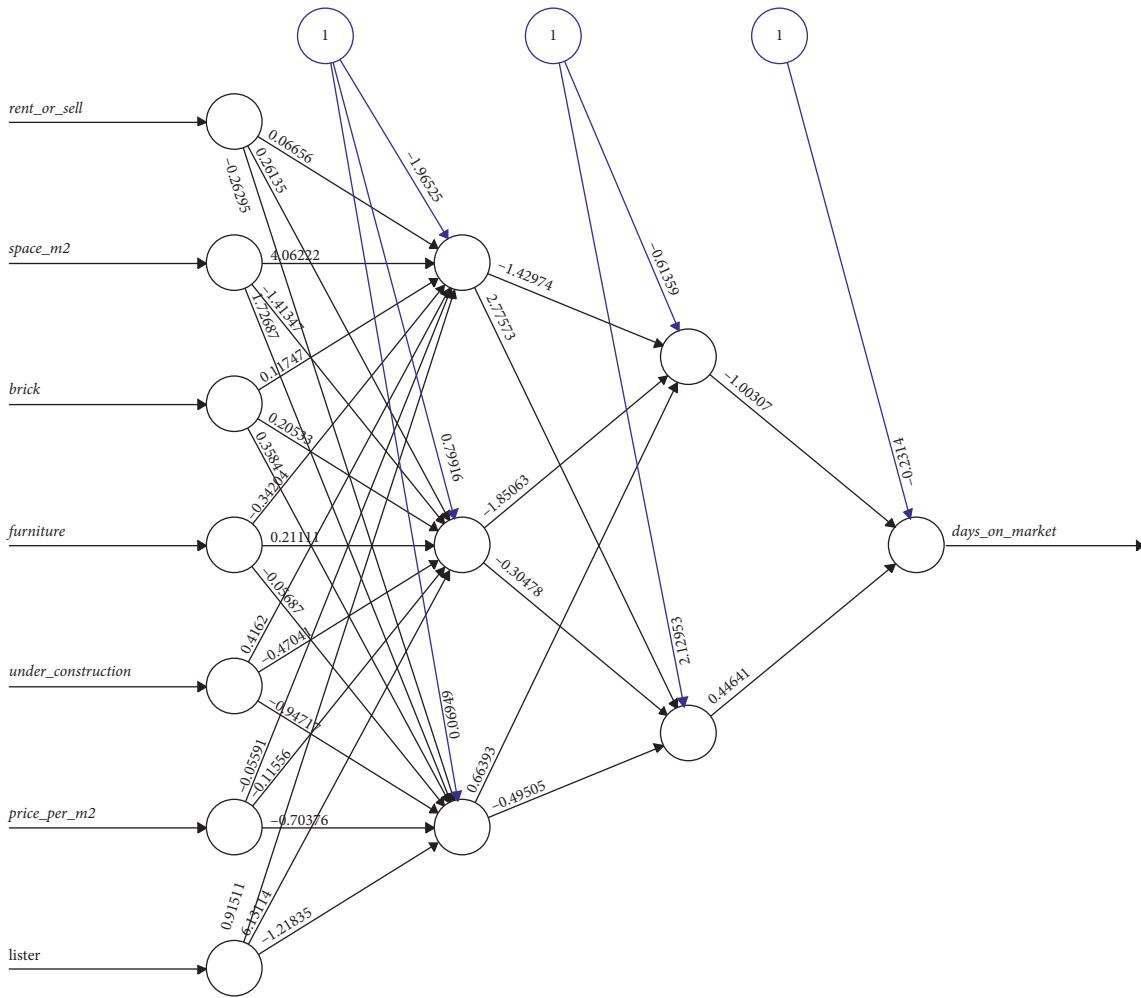


FIGURE 20: The best Neural Network that we were able to obtain in our experiments.

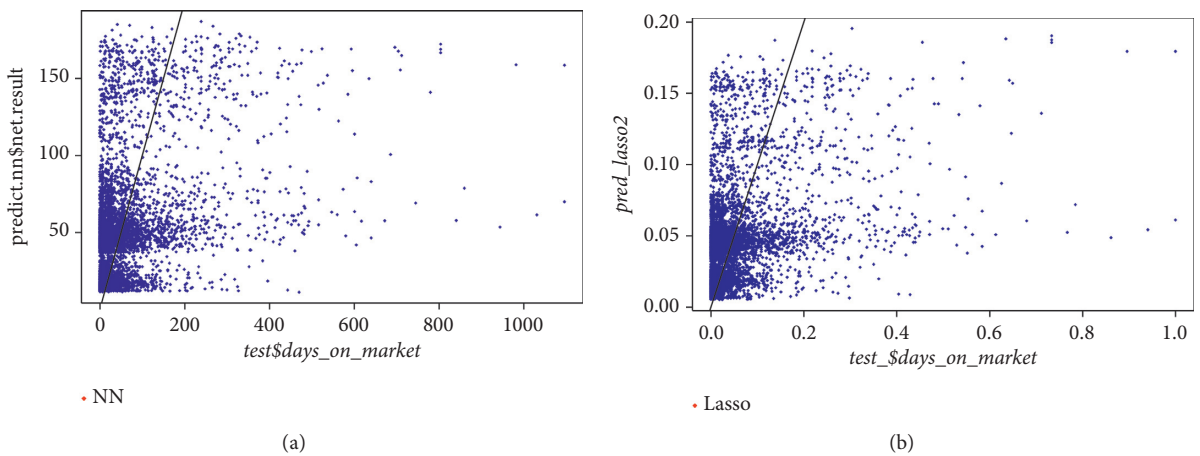


FIGURE 21: Continued.

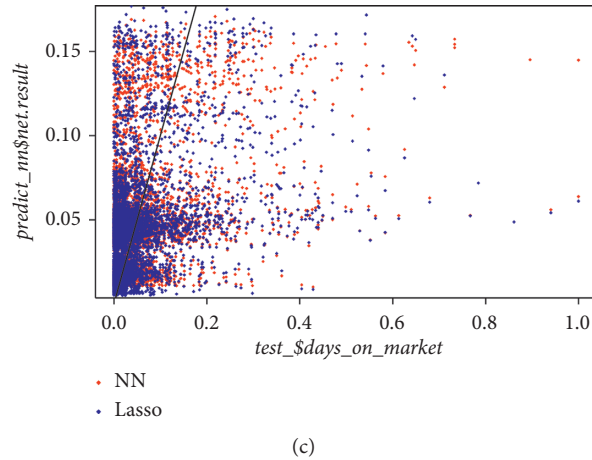


FIGURE 21: Real vs. predicted values for Neural Network and Lasso.

observation and the samples in the training set. Moreover, the interpretability of a model generated by means of a penalized regression is higher than the one of K-NN, since features' importance cannot be extracted from K-NN.

The performance of the last machine learning technique considered, SVR, generally depends on the choice of the kernel function. The kernel function defines the relationship/distance between the support vector and the target, by transforming the nonlinear input space into a linear space. The basic concept behind SVR is that the maximum admissible error for the prediction should be below a certain value defined as epsilon. To avoid overfitting, the regression is penalized by the usage of a cost parameter. In the experimental phase, we used the automated kernel function selection, but to define penalty cost and epsilon (maximum allowed error), we performed a grid search. Performing the experiments with epsilon = 0.5 and the cost parameter equal to 4.57, we obtained a median RMSE of 0.066.

Table 8 presents several performance measures to summarize and compare the models trained for the presented problem. MAE (mean absolute error) and MDAE (median absolute error) are both suitable measures as the data taken into account are characterized by some extreme values for DOM. The explained variance score takes into consideration the mean error, while  $R^2$  does not consider the mean error in the calculation and this makes the metric a bit more biased, which may lead to over- or underestimating the model in terms of how well the predictors explain the target.

All in all, it is possible to state that despite its simplicity, Lasso is the technique that we found most appropriate to address the problem at hand. In particular, it produced a competitive performance (i.e., low error) by also allowing us to analyse the most important features that characterize the problem. Section 5.1 is dedicated to this analysis.

**5.1. Feature Importance in the Model Found by Lasso.** One of the most known methods to measure the importance of features in a learned predictive model consists of measuring the increase in the error of the model, after modifying the values of the features, for instance, shuffling their values

along with the different observations. In other words, a given feature is considered less or not important if rearranging its values does not lead to any change in the model's error, and it is considered as important if it leads to a significant modification of the error. One of the interesting points of this method is that it takes into account not only the relationship of a feature with the output variable, but also with all the rest of the features. Additionally, the permutation importance does not require retraining of the model, but just a simple shuffling of the values of the features [14].

Figure 22 shows the features, sorted according to their importance (from the most important one that is reported at the top to the less important one that is reported at the bottom). For each feature, its importance is measured as a difference in the RMSE between the model executed with the original values of the feature and the model executed after shuffling. Table 9 gives detailed information on the results of the features importance test.

These results show that *lister* is considered as the most important feature by the Lasso model, followed, in the order by *rent\_or\_sell*, *under\_construction*, *space\_m2*, *brick*, and *furniture*. Finally, *price\_per\_m2* was considered as the less predictive feature.

## 6. Conclusions and Future Work

The objective of this paper was to develop a model to predict the *days\_on\_market* variable by applying several algorithms, in particular, Lasso, Ridge, and Elastic Net regressions and Neural Networks. The starting point of the work was the formulation of the following research questions, which will be answered in the upcoming paragraphs:

- (1) Can a machine learning algorithm predict the *days\_on\_market* variable for the housing units?
- (2) Which features effectively influence the property attractiveness for the customer target?

The various features were investigated and transformed to identify the key factors that affect the attractiveness of a property, which resulted in the reduction of features used in

TABLE 8: Model comparison—performance measures.

Model	RMSE	MAE	MDAE	$R^2$	Explained variance score
Random forest	0.073	0.0399	0.0202	0.3625	17.5037
Elastic Net	0.065	0.0341	0.0176	0.1983	9.5736
Lasso	0.064	0.0340	0.0177	0.1832	8.8431
Ridge	0.065	0.0341	0.0176	0.1942	9.3746
ANN	0.065	0.0339	0.0160	0.2	9.6594
K-NN	0.064	0.0331	0.0155	0.2394	11.5575
SVR	0.066	0.0396	0.0321	0.0761	3.6757

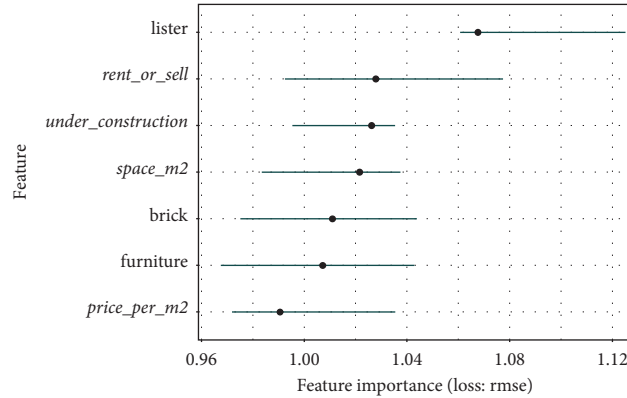


FIGURE 22: Feature importance in the model.

TABLE 9: Feature importance in the model.

Variable	Importance.05	Importance	Importance.95	Permutation error
<i>lister</i>	1.0611371	1.0676099	1.119868	0.07243278
<i>rent_or_sell</i>	0.9945100	1.0278727	1.072171	0.06973678
<i>under_construction</i>	0.9973978	1.0262011	1.030145	0.06962337
<i>space_m2</i>	0.9818909	1.0215203	1.032234	0.06930579
<i>brick</i>	0.9752772	1.0109712	1.038632	0.06859008
<i>furniture</i>	0.9664737	1.0071756	1.037656	0.06833257
<i>price_per_m2</i>	0.9698690	0.9905373	1.030250	0.06720373

the model to 7. Then, the studied algorithms were trained and Lasso regression outperformed the other studied algorithms. In conclusion, we were able to develop an accurate predictive model using Lasso regression, to predict the independent variable *days\_on\_market* with a selection of discriminators, which will be discussed in the second research question. The answer to the second question (2) was closely related to the findings of the first one: recognizing the features which make a property more interesting to the market. As many studies focus on measuring the effect of factors on houses' price, here the point of interest was measuring the effect of features on attractiveness. Based on this particular dataset, the features which have the most influence on the *days\_on\_market* are *lister*, *rent\_or\_sell*, *under\_construction*, *space\_m2*, *brick*, and *furniture*. One of the main limitations of this work is given by the available data. For instance, a significant amount of the variable characters was in Cyrillic, and while it was possible to translate some of them in English, others contained a major number of characters which made the automatic and correct

translation impossible. For example, analysing further the full description of the listings, or considering the names of the agencies/owners who published the listings could provide deeper insights.

To improve this work, several supplementary steps can be taken in the future. In the data collection phase, which was not part of the scope of this paper, additional data sources can be taken into account. For example, data for the neighborhood and residential profile (schools, supermarkets, transport, etc) can be collected and included in the research. The same is valid for other factors that influence the market. Additionally, as mentioned previously, the real days on market for a property were not available and known in this dataset. To assure the reliability of the outcome, information about properties' liquidity needs to be collected. This is not only time consuming but also a long-term task since such information would be available only if it provided directly by agencies and owners. Furthermore, the data used here were only for one city; a more complex dataset covering



various cities and regions with their own specifications would be more informative. In the long term, we plan to collect demographics, users profile, and in-app behavior data. Such information together with macroeconomic statistics for purchasing power, banking interest rates, employment level, wage rates, etc., can provide a broader picture not only about the market, but also about the factors which influence home preferences and attractiveness. Not to forget news and media data, which both can reveal interesting patterns for customer behavior and market fluctuations, as well as can provide some insights for the reputation of different agencies. Last but not least, another field of potential research involves the use of other machine learning algorithms, such as a k-nearest neighbor, support vector machines, and random forest.

## Appendix

Regression analysis is a statistical technique that models and approximates the relationship between a dependent variable and one or more independent variables. In the case of this study, the dependent variable is *days\_on\_market* (DOM), while the independent variables resulted from a complex phase of data preprocessing, described in Section 4. This Appendix describes the different techniques used in the paper to address the regression problem at hand.

*A.1. Lasso, Ridge, and Elastic Net.* Simple linear regression, also known as ordinary least squares (OLS) attempts to minimize the sum of error squared. The error, in this case, is the difference between the actual (observed) data point and its predicted value. The equation for this model is referred to as the cost function and is a way to find the optimal error by minimizing and measuring it:

$$\sum_{i=1}^M (y_i - \hat{y}_i)^2 = \sum_{i=1}^M \left( y_i - \sum_{j=0}^p w_j \times w_{ij} \right)^2. \quad (\text{A.1})$$

The gradient descent algorithm is used to find the optimal cost function by going over several iterations. But the data we need to define and analyse are not always so easy to characterize with the base OLS model. One situation is the data showing multicollinearity, this is when predictor variables are correlated to each other and the response variable. To produce a more accurate model of complex data, we can add a penalty term to the OLS equation. A penalty adds a bias towards certain values. These are known as L1 regularization (or Lasso regression) and L2 regularization (or Ridge regression).

Ridge regression adds the following penalty term, called L2 term, to the OLS equation:

$$+ \lambda \sum_{j=0}^p w_j^2. \quad (\text{A.2})$$

The L2 term is equal to the square of the magnitude of the coefficients. In this case, if lambda ( $\lambda$ ) is zero, then the

equation is the basic OLS. If lambda is greater than zero, then a constraint is added to the coefficients. This constraint has the objective of minimizing the coefficients (or, informally speaking, shrinking). The values of the coefficients tend towards zero as the values of lambda get larger. Shrinking the coefficients leads to lower variance and in turn a lower error value. Therefore Ridge regression decreases the complexity of a model. However, Ridge does not reduce the number of variables it rather just shrinks their effect.

Lasso (least absolute shrinkage and selection operator) regression uses the L1 penalty term, which is equal to the absolute value of the magnitude of the coefficients:

$$+ \lambda \sum_{j=0}^p |w_j|. \quad (\text{A.3})$$

Analogously to Ridge regression, also for Lasso, a lambda value equal to zero corresponds to the basic OLS equation. However, given an appropriate lambda value, Lasso can drive some coefficients to zero. The larger the value of lambda, the more features are shrunk to zero. This can eliminate some features and give us a subset of predictors that helps mitigate multicollinearity and model complexity. If a variable is not shrunk to zero, it means that the variable is important. In other words, L1 regularization allows for feature selection (sparse selection).

A third commonly used model of regression is the Elastic Net, which incorporates penalties from both L1 and L2 regularization:

$$\frac{\sum_{i=1}^n (y_i - x_i^j \hat{\beta})^2}{2n} + \lambda \left( \frac{1 - \alpha}{2} \sum_{j=1}^m \hat{\beta}_j^2 + \alpha \sum_{j=1}^m |\hat{\beta}_j| \right). \quad (\text{A.4})$$

In addition to choosing a value for the lambda parameter, Elastic Net also allows us to tune the alpha ( $\alpha$ ) parameter. A value of alpha equal to zero corresponds to Ridge; a value of alpha equal to one corresponds to Lasso. If we choose an alpha value between 0 and 1, we can incorporate penalties from both L1 and L2 regularization and alpha allows us to decide the relative importance of these two penalties. The interested reader is referred to Fonti [15] for deepening the functioning and properties of the Lasso, Ridge, and Elastic Net regression methods.

*A.2. Artificial Neural Networks.* An Artificial Neural Network (ANN) is a computational model based on the structure and functions of biological neural networks. It is composed of a set of elementary computational units, called neurons, strongly interconnected between each other by means of connections, or synapses, characterized by a weight. An ANN encodes a function (or model) that can produce outputs once inputs are presented to it. Supervised learning ANNs that are the ones studied in this paper have the objective of returning the expected outputs for each one of the input vectors contained in a given dataset. The learning phase, aimed at obtaining this expected input/output match, consists in a modification of the weights of the

connections in the network. Every single neuron can be represented as shown in Figure 23.

Once the values of the set of weights of the connections entering into a neuron have been established, the output of the neuron is calculated by

$$y = f\left(\sum_{i=1}^n w_i x_i + \theta\right). \quad (\text{A.5})$$

In an ANN, neurons are usually organized into layers. Supervised learning ANNs are formed by three different types of layers of artificial neurons:

- (i) Input layer
- (ii) Hidden layer
- (iii) Output layer

The input layer communicates with the external environment that presents data to the neural network. Its job is to deal with all the input values. These input values are transferred to the hidden layers, which are explained below. Every input neuron represents some independent variable that has an influence over the output of the neural network. The hidden layers are intermediate layers, found between the input layer and the output layer. The job of each hidden layer is to process the inputs obtained by its previous layer. Finally, the output layer contains the units that return the computed result to the outside world. The general structure of a feed-forward ANN, i.e., one of the most diffused types of supervised ANN and the one used in this work, is shown in Figure 24.

Several learning rules exist, aimed at looking for a configuration of the connection weights that allow a perfect input/output match. One of the most diffused ones and the one used in this paper is called backpropagation. The interested reader is referred to Gurney [16] to deepen the subject.

**A.3. Support Vector Regression.** Support vector machines (SVM) were introduced in [17], for classification problems. The objective is looking for the optimal separating hyperplane between classes. The points lying on classes' boundaries are called support vectors, and the in-between space, the hyperplane; when a linear separator is not able to find a solution, data points are projected into a higher-dimensional space, where the before nonlinearly separable points become linearly separable, using kernel functions. The whole task can be formulated as a quadratic optimization problem that can be solved with exact techniques. In Figure 25, an example of a linearly separable classification problem solved using SVM is presented. SVM aims at maximizing the margin between the support vectors and the hyperplane.

One year after the introduction of SVM, Smola [18] presented an alternative loss function, which allowed SVM to also be applied to regression problems. In SVR, the idea is to map the data events  $X$  into a  $k$ -dimensional feature space  $F$ , through a nonlinear mapping  $\phi_j(X)$ , so that it is possible to fit a linear regression model to the data points in this space. The obtained linear learner is then used to forecast in

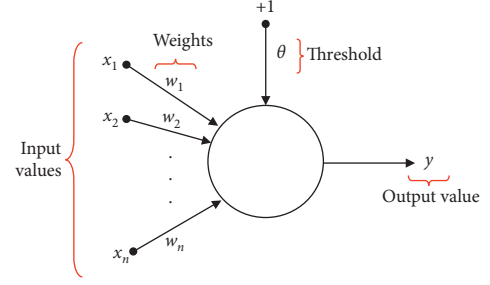


FIGURE 23: General structure of an artificial neuron.

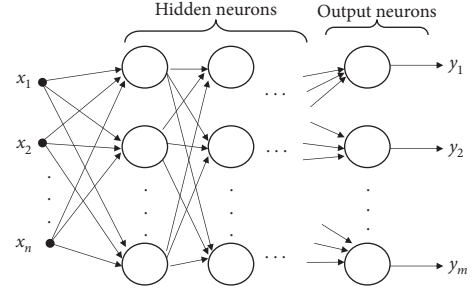


FIGURE 24: Architecture of a feed-forward ANN.

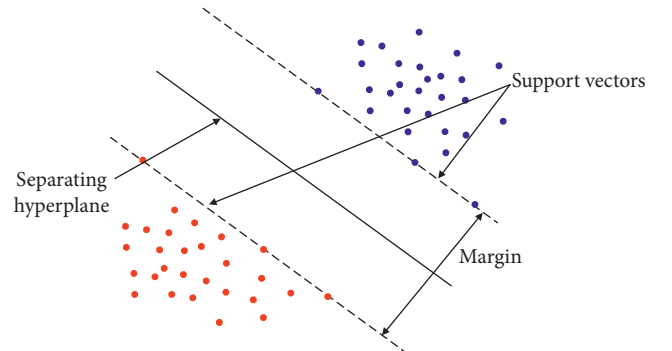


FIGURE 25: Linearly separable problem.

the new feature space. Once again, the mapping from the input space into the new feature space is defined by the kernel function. One of the most attractive characteristics of SVR is related with the model errors; instead of minimizing the observed training error, SVR minimizes a combination of the training error and a regularization term, aimed at improving the generalization ability of the model. Other attractive properties of SVR are related to the use of kernel functions, which make them applicable both to linear and nonlinear forecasting problems, and the absence of local minima in the error surface due to the convexity of the fitness function and its constraints. Given

- (i) Training dataset  $T$ , represented by

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}, \quad (\text{A.6})$$

where  $x \in X \subset \mathbb{R}^n$  are the training inputs and  $y \in Y \subset \mathbb{R}$  are the training expected outputs;

(ii) A nonlinear function:

$$f(x) = w^T \Phi(x_i) + b, \quad (\text{A.7})$$

where  $w$  is the weight vector,  $b$  is the bias, and  $\Phi(x_i)$  is the high-dimensional feature space, which is linearly mapped from the input space  $x$ .

The objective is to fit the training dataset  $T$ , by finding a function  $f(x)$  that has the smallest possible deviation  $\varepsilon$  from the targets  $y_i$ . Equation (A.7) can be rewritten into a constrained convex optimization problem as follows:

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} w^T w, \\ & \text{subject to} \quad \begin{cases} y_i - w^T \Phi(x_i) - b \leq \varepsilon, \\ w^T \Phi(x_i) + b - y_i \leq \varepsilon. \end{cases} \end{aligned} \quad (\text{A.8})$$

The aim of the objective function represented in equation (A.8) is to minimize  $w$ , while satisfying the other constraints. One assumption is that  $f(x)$  exists, i.e., the convex optimization problem is feasible. This assumption is

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \sum_{i,j=1}^m K(x_i, x_j) (\alpha_i^+ - \alpha_i^-) (\alpha_j^+ - \alpha_j^-) + \varepsilon \sum_{i=1}^m (\alpha_i^+ + \alpha_i^-) - \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-), \\ & \text{subject to} \quad \begin{cases} \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) = 0, \\ \alpha_i^+, \alpha_i^- \in [0, C], \end{cases} \end{aligned} \quad (\text{A.10})$$

where  $K(x_i, x_j)$  is the kernel function; the above formulation allows the extension of SVR to nonlinear functions, as the kernel function allows nonlinear function approximations while maintaining the simplicity and computational efficiency of linear SVR. The performance and good generalization of SVR depend on three training parameters:

- (i) Kernel function
- (ii)  $C$  (the regularization parameter)
- (iii)  $\varepsilon$  (the insensitive zone)

**A.4.  $K$ -Nearest Neighbors.**  $k$ -nearest neighbors (K-NN) [19] is one of the simplest existing machine learning algorithms and, despite its simplicity, is often capable of making accurate predictions on a large number of applications. The basic idea of K-NN is as follows: suppose we want to partition a dataset into classes and suppose we have a supervised training dataset, where some training observations are already categorized into the correct class. Suppose now that we have a new data  $x$  and we want to predict which class  $x$

not always true; therefore, one might want to trade off errors by the flatness of the estimate. Having this in mind, Vapnik reformulated equation (A.8) as

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} w^T w + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-), \\ & \text{subject to} \quad \begin{cases} y_i - w^T \Phi(x_i) - b \leq \varepsilon + \xi_i^+, \\ w^T \Phi(x_i) + b - y_i \leq \varepsilon + \xi_i^-, \\ \xi_i^+, \xi_i^- \geq 0, \end{cases} \end{aligned} \quad (\text{A.9})$$

where  $C < 0$  is a prespecified constant that is responsible for regularization and represents the weight of the loss function. The first term of the objective function  $w^T w$  is the regularized term, whereas the second term  $C \sum_{i=1}^m (\xi_i^+ + \xi_i^-)$  is called the empirical term and measures the  $\varepsilon$ -insensitive loss function. To solve equation (A.9), Lagrangian multipliers ( $\alpha_i^+, \alpha_i^-, \eta_i^+, \eta_i^-$ ) can be used to eliminate some of the primal variables. The final equation that translates the dual optimization problem of SVR is

belongs to. The idea is to consider the  $k$  training observations that are closest to or most similar to  $x$  (where similarity is quantified by a predefined distance measure) and return to the class to which most of these observations belong (majority vote). Following the same idea, for regression problems, the output on an unseen instance  $x$  is given by the average output of the  $k$  training observations most similar to  $x$ .

K-NN is a supervised, nonparametric, instance-based classification method. It is not parametric because before making the prediction, you do not have to make any assumptions about the distribution of the data, nor about the shape of the model. It is instance-based in the sense that there is no training phase: as long as we have the supervised data and the data we want to make predictions about, we can make the prediction. Although K-NN is nonparametric, we usually use two parameters to build the model:  $k$  (the number of neighbors) and the distance metric. There are no strict rules for selecting  $k$ . Indeed, this choice depends on the dataset and experience in choosing an optimal value. Generally, when  $k$  is small, the prediction would be easily impacted by noise and

when  $k$  is larger, while reducing the impact of outliers, it will show more bias (as a limit case, when we increase  $k$  up to the number of training data, the forecast will always be the majority class in the training set). The selection of the distance metric also varies in different cases. By default, the most commonly used metrics are Euclidean distance (L2 standard), Manhattan distance, and Minkowski distance.

There are several advantages of using K-NN: it is a simple method, very easy to implement and interpret, there is no model training phase, there are no previous assumptions about data distribution (this is especially useful when we have poor quality and unstructured data), and it generally has relatively high accuracy. Of course, there are also disadvantages: high memory requirements (we need to store all training data in memory to execute the method) and computationally expensive (we need to calculate the distance between the new data point and all existing data points to decide which  $k$  are closest), which is quite expensive in terms of computation and sensitive to noise (particularly if we choose a small  $k$ , the prediction results will probably be impacted by noise, if any).

**A.5. Random Forest.** Random forest [20] is a type of ensemble model, which uses bagging as an ensemble method and the decision tree as an individual model.

A decision tree is a predictive model, where each internal node represents a variable, an edge towards a child node represents a possible value for that property, and a leaf represents the predicted value for the target variable starting from the values of the other properties. A decision is represented by the path from the root node to a leaf node.

An ensemble method is a technique that combines predictions from multiple machine learning algorithms, to make predictions more accurate than any single model. Bagging represents a general procedure that can be used to reduce the variance of those algorithms that have a high variance, such as decision trees, in the case of random forests. Decision trees, in fact, are sensitive to the specific data on which they are formed. If the training data is changed (e.g., a tree is trained on a subset of the training data), the resulting decision tree can be quite different and, in turn, the forecasts can be quite different. Bagging is the application of the bootstrap procedure to a high-variance machine learning algorithm. A random forest combines many decision trees into one model. Individually, the predictions made by the decision trees may not be accurate, but combined together, the forecasts will on average be closer to the result. The final result returned by the random forest is nothing but the average of the numerical result returned by the different trees in the case of a regression problem, or the class returned by the largest number of trees for classification.

## Data Availability

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

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia) by the projects GADgET (DSAIPA/DS/0022/2018), BINDER (PTDC/CCI-INF/29168/2017), and AICE (DSAIPA/DS/0113/2019). Mauro Castelli acknowledges the financial support from the Slovenian Research Agency (research core funding no. P5-0410).

## References

- [1] P. Stoykova: Основни Показатели за жилищния Пазар в България През 2018: Bulgarian Properties, <https://www.bulgarianproperties.bg/novini-za-imoti/pokazateli-imoten-pazar-2018-7555.html>, 2018.
- [2] BTV, *Bulgaria. BTV Novinite*, <https://btvnovinite.bg/bulgaria/falshivi-brokeri-zalivat-pazara-na-imoti.html>, 2012.
- [3] Q. M. Xian Guang LI, *The Application of Data Mining Technology in Real Estate Market Prediction*, Fraunhofer Information Center for Space and Construction IRB, Stuttgart, Germany, 2006, <https://www.irbnet.de/daten/iconda/CIB5807.pdf>.
- [4] E. C. M. Hui, J. T. Y. Wong, and K. T. Wong, "Marketing time and pricing strategies," *Journal of Real Estate Research*, vol. 34, no. 3, pp. 375–398, 2012.
- [5] G. D. Jud, "Time on the market: the impact of residential brokerage," *Journal of Real Estate Research*, vol. 12, no. 3, pp. 447–458, 1996.
- [6] D. D. Belkin, "An empirical study of time on market using multidimensional segmentation of housing markets," *Real Estate Economics*, vol. 4, no. 2, pp. 57–75, 1976.
- [7] N. Miller, "Time on the market and selling price," *Real Estate Economics*, vol. 6, no. 2, pp. 164–174, 1978.
- [8] J. Z. Catherine-Tucker, J. Zhang, and T. Zhu, "Days on market and home sales," *Rand Journal of Economics*, vol. 44, no. 2, pp. 337–360, 2013.
- [9] H. X. Hengshu Zhu, "Days on market: measuring liquidity in real estate markets," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 393–402, Beijing, China, 2016.
- [10] S. V. Ermolin, *Predicting Days-on-Market for Residential Real Estate Sales*, Department of Computer Science Stanford University, Stanford, CA, USA, 2016, [http://cs229.stanford.edu/proj2016/report/ermolin\\_predicting\\_Days\\_on\\_market\\_for\\_Residential\\_Real\\_Estate\\_Sales\\_report.pdf](http://cs229.stanford.edu/proj2016/report/ermolin_predicting_Days_on_market_for_Residential_Real_Estate_Sales_report.pdf).
- [11] Q. Z. Chao Mou, "Recommending property with short days-on-market for estate agency," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 6, pp. 2077–2092, 2018.
- [12] H. Seltman, *Exploratory Data Analysis. Experimental Design and Analysis*, Carnegie Mellon University, Pittsburgh, PA, USA, 2015, <https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>.
- [13] A. A. Asaad, *Measures of Skewness and Kurtosis. R Bloggers*, <https://www.r-bloggers.com/measures-of-skewness-and-kurtosis/>, 2013.
- [14] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Christopher Molnar*, <https://christophm.github.io/interpretable-ml-book/>, 2019.

- [15] V. Fonti, *Research Paper in Business Analytics: Feature Selection with LASSO*, VU Amsterdam, Amsterdam, Netherlands, 2017.
- [16] K. Gurney, *An Introduction to Neural Networks*, University College London (UCL) Press, London, UK, 2004.
- [17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [18] A. J. Smola, "Regression estimation with support vector learning machines," Master Thesis, Technische Universität München, Munich, Germany, 1996.
- [19] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [20] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms," *Machine Learning*, vol. 36, no. 1/2, pp. 105–139, 1999.

## Research Article

# Efficiency of Chinese Real Estate Market Based on Complexity-Entropy Binary Causal Plane Method

Yan Chen,<sup>1</sup> Ya Cai,<sup>2</sup> and Chengli Zheng <sup>2</sup>

<sup>1</sup>School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China

<sup>2</sup>School of Economics and Business Administration, Financial Engineering Research Center, Central China Normal University, Wuhan 430079, China

Correspondence should be addressed to Chengli Zheng; zhengchengli168@163.com

Received 30 October 2019; Accepted 24 December 2019; Published 13 January 2020

Guest Editor: Marco Locurcio

Copyright © 2020 Yan Chen 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.

Real estate market is a complex system. A rational real estate market is not only helpful to people's living standards but also beneficial to countries' macroeconomic stability. Is Chinese real estate market rational? This paper attempts to study the efficiency of Chinese real estate market by using the complexity-entropy binary causal plane method. We firstly discuss the formation mechanism of real estate price, which provides a theoretical basis for testing the efficiency, and compute the real estate market efficiency of 70 main Chinese cities. The results show that neither the whole market nor the main cities have reached the weak efficiency, and the efficiency and complexity of each city are different, and the relationship between them is significantly negative. In addition, this paper also compares the efficiency and complexity of Chinese real estate market with American real estate market. Then, some suggestions for the healthy development of Chinese real estate market in the future are put forward.

## 1. Introduction

In July 1998, the notice of the State Council on further deepening the reform of the housing system and accelerating the housing construction was issued, which clearly put forward the reform goal of developing the housing trading market and accelerating the housing construction. Since then, the welfare housing system traversing nearly 40 years has been withdrawn from the historical stage. "Marketization" has become the new theme of the real estate system. Real estate began to appear as a commodity. From then on, real estate is not only an essential element to satisfy the basic residential attribute of residents but also an indispensable part of family property. Nowadays, real estate is not only a commodity but also an asset. With the development of financial industry, real estate has gradually become a major tool for investment and speculation. At the same time, the market size of real estate has gradually become larger and larger. By 2018, the added value of real estate industry has accounted for nearly 7% of GDP, while the contribution of real estate and its related industries to Chinese GDP has

reached about 1/3. It can be said that Chinese real estate has achieved great success in development size and speed after nearly 20 years of continuous reform and exploration. At the same time, the real estate has played a significant role in pushing the progress of other industries related to itself, dramatically promoting the rapid development of the national economy.

However, Chinese real estate price has kept increasing rapidly in recent years. Subsequently, the ratio of housing price to income is seriously unbalanced, and the house price bubble accumulates gradually. It can be seen that this abnormal high-speed rising trend has brought many potential problems. On the one hand, the imbalance of superficial demand and supply caused by the rapid rise of house prices will lead to an overheated economy and lead to housing price bubble, which has a significant crowding out effect on the real economy. Once this rapid growth cannot continue, the bubble will collapse and then result in the slump in the overall economy. More seriously, it will push the economy into depression, for instance, the subprime crisis caused by the breakdown of real estate market in America in 2007. That

crisis was a financial-market mess, as well as a housing one. On the other hand, the rapid rise of housing price or irrational housing price will lead to a dramatic decline in the housing consumption level of most ordinary people and affect people's living standards. Also, it does harm to the stability of economy and society.

In view of the importance of the reasonable and orderly development of the real estate market, it has become a matter of concern for many scholars to evaluate and quantify the efficiency of the market. Meese and Wallace [1] applied the efficient market theory to the real estate market in San Francisco and found that the real estate market in that region is efficient in the long term, but inefficient in the short term, owing to the deviation caused by the transaction cost between the housing price and fundamental value. However, other scholars believed that it was the existence of market bubbles and irrational expectations, both of which led to this bias and further led to the inefficient market, such as Clayton's [2] continuous rational housing price model. The analysis about housing price efficiency varies from the angle of study. Case and Shiller [3] proposed a new method to examine the applicability of "Efficient Market Hypothesis" in real estate market in different American cities by using excess return rate. They argued that the reason for housing inefficiency was because the interest rate was not correctly estimated. Abraham and Hendershott [4] used the economic fundamental information to measure the efficiency of real estate market in 30 cities of the United States and calculated the deviation degree between the fundamental price and the actual housing price. They finally thought this deviation as the bubble of housing price. Grenadier [5] studied this issue from the perspective of vacancy rate. He attributed the inefficiency of the real estate market to the high vacancy rate, which was caused by monopoly. Kunzel [6] made a more comprehensive analysis of the efficiency of real estate market and proposed that the inefficiency of housing price originated from several reasons: space restriction, long construction period, transaction cost, speculation, and so on.

Domestic scholars' research on this problem is relatively behind. Wang et al. [7] applied the sequence correlation method to test whether the housing price index of China 1998–2004 conformed to the random walk model and obtained a negative conclusion, that is, the real estate market of China had not yet reached weak form efficiency. In addition to the newly constructed houses, the second-hand real estate market is also worth the attention. Xie [8] conducted an empirical test on newly constructed real estate market and the second-hand real estate market in Shanghai by collecting the quarterly data of CFS Shanghai index from 2005 to 2013. The test methods included unit root test and sequence correlation test. The results showed that the efficiency of each market segment in Shanghai was not the same. Among them, the office market and shop market in the new real estate market and the rental real estate market in the second-hand real estate market had reached weak form efficiency, while the real estate market in the new real estate market and the sales real estate market in the second-hand real estate market had not. However, the results of the sequence correlation test showed that the efficiency of each market

segment and the overall real estate market in Shanghai failed to achieve weak form.

In order to study the efficiency of Chinese real estate market, this paper first puts forward the theory about the housing price mechanism, which provides a theoretical foundation for testing the efficiency of real estate market and then conducts empirical test with the complexity-entropy binary causal plane method. The results show that the real estate market of 70 main Chinese cities as well as the overall Chinese real estate market does not reach weak form efficiency. The degree of efficiency of each city shows difference and is significantly negatively related to the corresponding degree of complexity. In addition, in order to control housing prices and promote the orderly development of the real estate market more effectively, we further compare the efficiency of real estate market in China with that in America, whose real estate market has also experienced many boom years. Combined with the historical experience in the real estate industry development of two countries, this paper analyzes the reasons that why there are differences in the efficiency of the real estate market between two countries and then puts forward some valuable suggestions for the healthy development of Chinese real estate market in the future.

A novel feature of our study is that we propose a novel method called complexity-entropy binary causal plane method to detect the hidden structure in the housing price and get the efficiency and complexity of real estate market in 67 main cities and overall China. We collect the monthly data of 67 main cities of China during 2005 to 2017. With a wide range of samples and a long period of time, our result can be regarded as representative, so we think there is reference value for cities to implement policies to maintain the stability of housing prices. Moreover, we conduct the comparative study of the efficiency of the real estate market in China and in America. The results provide a direction for learning from foreign advanced management and regulation experience.

The rest of the paper is organized as follows. Section 2 clarifies the theory of housing price mechanism. Section 3 develops the efficiency and complexity measure method based on the complexity-entropy binary causal plane method. Section 4 conducts empirical analysis to demonstrate the efficiency of real estate markets in China and America. Finally, Section 5 concludes our paper.

## 2. Theory about Housing Price Mechanism

Analogous to the efficiency theory of financial market, the efficiency of real estate market can be described as follows: housing price can quickly respond to all kinds of relevant information so that actual housing price is consistent with its intrinsic fundamental value. Therefore, the research on the efficiency of housing price should first clarify the formation mechanism of housing price. The process of how the real estate market price is formed under the principle of no-arbitrage pricing is presented as follows.

Suppose that the participants in the real estate market (taking the buyer as an example) at time  $t$  have two choices:

(1) buying a house to live in, where the price is  $p_t$ ; (2) renting a house to live in, where the rent is  $d_t$  per month. Under the condition of short selling, above two cases should be the same, that is,

$$p_t = \sum_{i=1}^{\infty} \frac{E(d_{t+i-1} | \Pi_t)}{(1+r_f)^i}, \quad (1)$$

where  $r_f$  denotes the risk-free rate. If the right-hand side in equation (1) is not equal to the left-hand side, an arbitrary opportunity will appear. Specifically, if there is  $p_t < \sum_{i=1}^{\infty} (E[d_{t+i-1} | \Pi_t]) / (1+r_f)^i$ , we can make an arbitrage by constructing following portfolio: firstly, borrow money from a bank to buy a house and rent it out at a price of  $d_t$  per month and then repay the bank loan by installments. It is clear that the investor can make profit easily. We assume that the market participants are rational. They can find this arbitrary opportunity and then perform the same behavior that participants all buy houses to rent them out. In this case, the house price will rise and the rent will decline until equation (1) holds true. After this game, the arbitrary opportunity will disappear and the market reaches no-arbitrage equilibrium. On the contrary, if  $p_t > \sum_{i=1}^{\infty} (E[d_{t+i-1} | \Pi_t]) / (1+r_f)^i$ , the participants can conduct the opposite arbitrage until a new equilibrium appear. However, if the short selling is limited, the price cannot reach equilibrium by arbitrages. There will always exist  $p_t < \sum_{i=1}^{\infty} (E[d_{t+i-1} | \Pi_t]) / (1+r_f)^i$  or  $p_t > \sum_{i=1}^{\infty} (E[d_{t+i-1} | \Pi_t]) / (1+r_f)^i$ . At this time, the actual house price can be described as follows:

$$p_t = \sum_{i=1}^{\infty} \frac{E[d_{t+i-1} | \Pi_t]}{(1+r_f)^i} + h_t = p_t^* + h_t, \quad (2)$$

where  $p_t^*$  is the basic value of a house and  $h_t$  denotes the extra price that can contribute to arbitrage.

On the other hand, Pan and Wang [9] analyzed the formation of housing price from a rational perspective. They believed that the actual market price would deviate from the basic price if the market is irrational. Accordingly, there also exists bias between actual house price and its theoretical price in irrational real estate market. This situation yields the following price rules:

$$p_t = p_t^* + b_t, \quad (3)$$

where  $b_t$  denotes irrational price or house price bubbles, deviating from the basic value [10]. In most cases,  $b_t > 0$ . Generally,  $b_t < 0$  will not occur. But it also may appear due to the impact of demand in the actual transaction. In the statistical sense, it should be positive.

Comparing these two price formation mechanisms,  $b_t$  in formula (3) belongs to the irrational component, while  $h_t$  in formula (2) belongs to the reasonable component, or at least some of which are reasonable. For the simplification of presentation, we just think  $b_t$  as a reasonable component and think  $h_t$  as an irrational component. In this way, after considering the irrational component, we propose that the house price is formed by following components:

$$p_t = p_t^* + h_t + b_t. \quad (4)$$

The real estate market is efficient when market participants are fully rational. At this time, the irrational component  $b_t$  does not exist, and the housing price can be characterized by equation (2). From the perspective of market efficiency theory, this means

$$E[b_t] = 0 \text{ and } b_t \text{ follows random walk.} \quad (5)$$

Then,  $b_t$  can be expressed as

$$b_t = p_t - E[p_t^* + h_t]. \quad (6)$$

In this way, we can judge the efficiency of real estate market by testing whether series  $b_t$  has random walk characteristics.

However, it is not easy to get the basic value  $p_t^* + h_t$  of a house. Although  $p_t^* = \sum_{i=1}^{\infty} (E[d_{t+i-1} | \Pi_t]) / (1+r_f)^i$  in theory, future cash flow of yield  $d_t$  cannot be obtained. Moreover, the computation of  $h_t$  is more difficult, as a kind of right of resale or convenience yield. Such fact makes it difficult to get the irrational price  $b_t$  and test the efficiency of markets.

Then, we decide to start with the ratio of house price to income. As the main indicator to measure the purchasing power of housing in a certain period, the ratio of house price to income is also the basic index to evaluate whether the housing market is healthy [11]. In the early 1990s, Andrew Hamer, a World Bank expert, conducted a study on the reform of Chinese housing system and gave a ratio from 4 times to 6 times. This interval is considered as ideal by the World Bank. According to the relevant data published by the United Nations, the dispersion about the ratio of house price to income in different countries is quite large. According to the statistical results of 96 countries in 1998, the ratio of house price to income in these countries ranged from 0.8 to 30, with an average value of 8.4 and a median value of 6.4. Chen [12] believes that the ratio of house price to income is reflection of real interest rate and is the most accurate indicator of housing price. In general, the ratio  $k_t$  of average house price to income of a country or region is usually calculated by

$$k_t = \frac{p_t}{I_t}, \quad (7)$$

where  $p_t$  denotes the average price of a house and  $I_t$  denotes the average annual income of a family. According to the theory about the ratio of house price to income, the housing price and income ratio should be kept within a suitable interval in a reasonable real estate market. Otherwise, it is an unhealthy market, where the house price is always underestimated or overestimated (generally overestimated). In other words, there are irrational components or bubbles in the real estate market. If there is an irrational component, the market is considered to be inefficient.

From a general equilibrium perspective, all the borrowing and lending in a region would offset, only leaving income to support house prices. Therefore, a reasonable housing price should be in proportion to personal income. This viewpoint has been proved by Case and Shiller [13]. They think that there is a stable relationship between



personal income and housing price if a housing price bubble does not exist. In order to measure the efficiency of the real estate market, we just need to consider whether the rest of housing price that is not fully explained by income is efficient or not, that is, whether it has the characteristics of random walking. In addition, the reasonable ratio of house price to income is basically stable. In other words, it is a constant or  $k_t = k$ . To generalize the model, we add an intercept term, so we can get

$$p_t = a + kE[I_t] + \varepsilon_t. \quad (8)$$

Thus, we have  $\varepsilon_t = p_t - (a + kE[I_t])$ , and it is a random disturbance term and satisfies  $E[\varepsilon_t] = 0$ . At this time, we can test the efficiency of real estate market by calculating the relevant measures of  $\varepsilon_t$  with the permutation entropy method.

### 3. Efficiency Measure Based on Complexity Entropy Binary Causal Plane Method

In recent years, system complexity and nonlinear dynamic methods have been paid more attention. Zunino et al. [14, 15] pointed out that it can effectively detect the structural information hidden in system noise, even if the system is at the edge of chaos. In the process of empirical test, we use the complexity-entropy binary causal plane method based on  $b_t$  to detect the hidden structure of real estate market price and then measure its efficiency and complexity. This method, as proposed in the work of Rosso et al. [16], can not only distinguish Gaussian and non-Gaussian processes but also demonstrate their correlation degree. Therefore, it is considered to be a good method to test the market efficiency.

**3.1. Market Efficiency and Shannon Entropy.** Entropy can accurately test the uncertainty and confusion of time series without any additional restrictions on the distribution. If the price follows a pure random walk, then there will not exist correlation relationship between time series, where the entropy of the sequence is the largest and represents a completely disordered state. Otherwise, it is difficult to reach the maximum entropy. It was Gulko [17] who firstly applied entropy to study financial time series by showing that the maximum-entropy formalism, also called informational efficiency, made the efficient market hypothesis operational and testable. If the market is efficient, the time series will satisfy random walk and the normalized entropy is 1. The smaller the entropy is, the harder the market reaches random walk and the less efficient the market is. Therefore, the normalized entropy (relative maximum entropy) can be applied to measure the real estate market efficiency. There are many proposed concepts of entropy. Matesanz and Ortega [18] proposed Shannon entropy, Renyi entropy, Tsallis entropy, approximation entropy, and so on. Risso [19] pointed out that Shannon entropy is the most widely used entropy in financial market. For a given probability distribution  $P = \{p_i: i = 1, \dots, N\}$ , Shannon entropy is defined as

$$S[P] = - \sum_{i=1}^N p_i \ln p_i. \quad (9)$$

It is clear that when the probability distribution is uniformly distributed, that is,  $P_e = \{1/N, \dots, 1/N\}$ , the Shannon entropy can reach the maximum value, and  $S_{\max} = S[P_e] = \ln N$ . On the other hand, if there is a deterministic event, the corresponding distribution can be expressed as  $P_0 = \{0, \dots, 1, \dots, 0\}$ , and the value of Shannon entropy is 0.

**3.2. Permutation Entropy.** Before calculating the entropy value of a given time series, we generally need to deal with the series to get the corresponding distribution. Rosso et al. [16] found that if the basic probability distribution took the causal relationship between time series into account and then used the above information measure the entropy, the result would be excellent. More importantly, the difference between chaos and randomness can be clearly distinguished. Later, Bandt and Pompe [20] successfully proposed a symbol method based on the reconstruction of phase space, which could introduce such causal relationship into the basic probability distribution. It was suggested by Ridet et al. [21] that this method was the only popular method considering the intertemporal structure of time series so far. It has been used to study the efficiency of crude oil markets [22], foreign exchange markets [23], and stock markets [14]. This method can be summarized as follows:

- (1) Given a time series  $\{x_t: t = 1, \dots, M\}$ . Embedding dimension is  $D$ , and time delay is  $\tau$ . The connection between dimension  $D$  and new subsequence is represented as follows:

$$s \mapsto (x_{s-(D-1)\tau}, x_{s-(D-2)\tau}, \dots, x_{s-\tau}, x_s). \quad (10)$$

We call it the ordinal pattern of embedding dimension  $D$ . As for any time  $s$ , there is always a new subsequence mapping to it, which is formed by a vector with dimension  $D$  as shown in formula (10). It is easy to notice that the larger dimension can incorporate more historical information.

- (2) Each subsequence  $i$  is sorted in ascending order. We can obtain a new set related permutation, marking it as  $\pi_i = (r_0, r_1, \dots, r_{D-1})_i$ . The rearranged series is

$$x_{s-r_{D-1}\tau} \leq x_{s-r_{D-2}\tau} \leq \dots \leq x_{s-r_1\tau} \leq x_{s-r_0\tau}. \quad (11)$$

- (3) For each dimension, there is  $D!$  kinds of arrangement. The probability distribution of the permutation in these time series is given by

$$p(\pi_i) = \frac{\#\{s \mid 1 + (D-1)\tau \leq s \leq M\}}{M - (D-1)\tau}, \quad (12)$$

where  $s$  satisfies formula (10) and  $\#$  denotes the number of occurrences of the permutation  $\pi$  in the class.

According to above steps, the probability distribution of each ordinal pattern in the time series can be obtained. In

order to obtain a more accurate distribution, the length of time series  $M$  is required to be infinite. So, we can get the corresponding distribution when embedding dimension  $D$  and delay time  $\tau$  are given. For achieving more reliable statistical results, it is generally suggested  $M \gg D!$  [24]. Bandt and Pompe suggested  $3 \leq D \leq 7$  in practical applications. As for the time delay,  $\tau = 1$  is common. Of course, there are also different selections (for details, see the literature [25]). Some recent applications and development of permutation entropy also can be seen in related literature [26, 27].

**3.3. Complexity-Entropy Binary Causal Plane Method.** In addition to permutation entropy, Lamberti et al. [28] pointed out that statistical complexity measure (SCM) can also be used to measure the efficiency of markets. SCM can not only detect the dynamic details of the system but also distinguish the periodicity and the degree of chaos, which is superior to the entropy measure. Mathematically, SCM is defined as follows:

$$C_{JS}[P] = Q_J[P, P_e] \bullet H_S[P], \quad (13)$$

where  $H_S[P]$  is normalized Shannon entropy, and it is defined as

$$H_S[P] = \frac{S[P]}{S_{\max}}. \quad (14)$$

And  $Q_J[P, P_e]$  is the Jensen-Shannon divergence; it is defined as

$$Q_J[P, P_e] = Q_0 \bullet J[P, P_e], \quad (15)$$

where  $J[P, P_e] = \{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2\}$  and  $Q_0 = 1/J[P_0, P_e]$ .

SCM can describe the structural complexity of a system, which cannot be achieved by entropy just measuring randomness. Different entropy values mean different randomness, but they may have the same SCM value because they have the same complexity. The larger the SCM value is, the more complex the system is. Accordingly, the correlation degree of sequences in the system is higher. We take two simple examples to further illustrate the difference between SCM and entropy. The entropy of linear data is 0, and the entropy of absolute fair dice game is 1. However, both their SCMs are 0 because of simple systems.

Therefore, the binary causal plane method formed by the combination of permutation entropy and SCM can characterize the efficiency of markets better. This method analyzes the market function from different perspectives and then takes the structure of real estate market into consideration. There is no doubt that more available market information is incorporated in this method. Zunino et al. [15] adopted this method to measure the efficiency of global major stock markets and found that this method could not only accurately identify the markets of developed and developing countries but also divide the markets of many countries into more categories, fully reflecting the discrimination ability of the model. Theoretically, the higher the complexity of a financial market is, the stronger the

randomness and disorder of the market will be. At this time, the corresponding permutation entropy will be larger, that is, the permutation entropy is positively correlated with complexity. On the other hand, the higher the degree of disorder and chaos in the market is, the higher the probability that the market tends to be invalid is, that is, the permutation entropy is negatively correlated with the efficiency. Therefore, it can be considered that the complexity of the market is negatively correlated with its effectiveness through the transfer of permutation entropy (see Rosso et al. [29] and Shaobo et al. [30]).

## 4. Empirical Analysis

**4.1. Data and Statistical Description.** We initially take the 70 main cities defined by National Bureau of Statistics (NBS) as our research objects. Due to the lack of relevant data of Anqing, Dali, and Yangzhou, we finally select the remaining 67 cities as our research objects. To simplify the presentation, we number the cities from 1 to 67 according to the alphabetical order of cities. At the same time, the overall real estate market in China is also incorporated into our study and numbered as No. 68.

The data of each research unit used in the empirical test mainly involve three indicators: house price index (HPI), consumer price index (CPI), and per capita income index (PPI). We collect the data from Wind database and Chinese economic and social big data research platform. Considering the availability of the data, the time interval of the sample data spans from July 2005 to December 2017, and we have a total of 150 monthly samples. For convenience, we set the HPI in July 2015 as 100 points and treat it as the basis. Average monthly salary of residents is a proxy variable of PPI. As for the CPI, we set the CPI in July 2015 as 100 points and treat it as the basis, same as HPI. A few missing data are interpolated by linear interpolation.

Table 1 shows the simple descriptive statistics of HPI in Chinese overall real estate market. From Figure 1, we can see the specific trend of HPI from 2005 to 2017. Combined with Table 1 and Figure 1, the overall housing price in China presents an overall upward trend during the sample period. Chinese housing prices have risen about 1.7 times since July 2005, with a monthly average growth rate of 0.0036 and an annual average growth rate of 0.0576. Although the fluctuation of Chinese housing price is small on the whole, its fluctuation frequency is high. From July 2005 to the end of 2017, Chinese housing price experienced at least 7 fluctuations. Besides the role of market itself, most of these fluctuations are related to the government's regulation policies. There are several severe fluctuations that deserve attention. Affected by the global financial crisis in 2008, housing prices began to fall sharply in October 2008. However, with the issue of several stimulus policies of the government, housing prices began to recover and rise in April 2009, which led to the steady rise of housing prices in later three years. In 2011, many cities successively launched the control policies centering on the purchase restriction. In October 2011, the housing price dropped slightly. However, with the implementation of easy monetary policy in 2012 and the reduction of bank reserve ratio, the real estate

TABLE 1: The simple descriptive statistics of HPI in Chinese overall real estate market.

Type	Max	Min	Mean	Median	Std	Skewness	Kurtosis	Monthly average growth rate
China	117.61	68.3	93.45	96.13	12.77	-0.24	2.26	0.0036

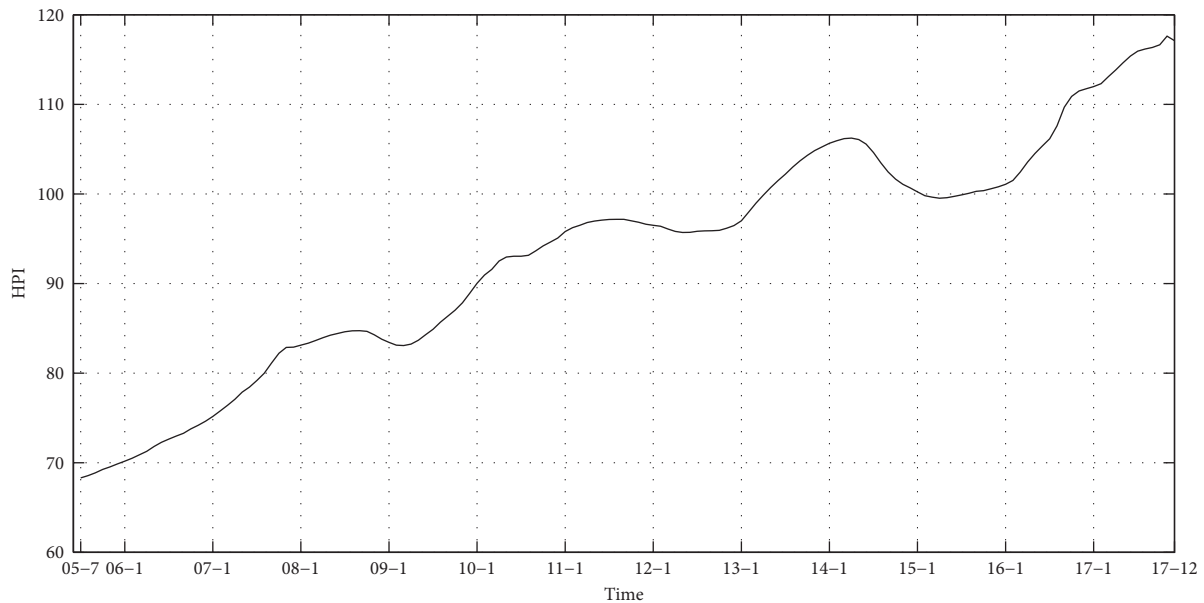


FIGURE 1: HPI trend of the overall real estate market.

market began a new growth period. In the latter half of 2014, the real estate market got into murky situation. This is due to the combined effect of regulatory policies and anticorruption efforts. But the new “330 policy” (a new real estate regulation and control policy on March 30, 2015) and the two interest-rate cuts in the latter half of 2015 initiated a wild rising tendency in a new round for housing prices. In 2016, due to the shortage of land, there was even a phenomenon that flour was more expensive than bread, which made the real estate market to be concerned again. In 2017, housing prices remained high, but the growth rate had been slower than that in 2016.

Considering the large amount of data, the basic descriptive statistics of HPI of 67 main cities cannot be fully presented due to space constraints. Here, we just demonstrate the HPI trend of each city, as shown in Figure 2.

As can be seen from Figure 2, housing prices in China had a rise tendency before 2015, but the tendency was moderate. It was after 2015 that severe fluctuations in housing prices occurred in many cities. From 2015 to 2016, housing prices soared rapidly, which was the fastest growth in the latest decade. Housing prices in many cities became relatively stable until 2017. In 67 cities, Wenzhou (No. 50) and Shenzhen (No. 44) are relatively special samples on studying the housing price. After the financial crisis in 2008, Chinese government launched “4 trillion stimulus plan” to stimulate the economic recovery. In the first half of 2009, housing prices in Wenzhou took the lead in response and started a new round of surge. At that time, housing prices in Wenzhou were even far higher than the first-tier cities, leading the country. However, bubbles would burst sooner or later. Its housing prices

declined steeply in the latter half of 2011. This price change lasted for more than 20 months. Then, a large number of investors and speculators withdrew from the property market. Although the government adjusted the purchase restriction policy, it achieved little success in saving the property market. In the next years, the property market was in a mild state until the appearance of property boom in 2016. Before 2015, Shenzhen real estate market compared with other first-tier cities seemed to be depressed, but after 2015, housing prices in Shenzhen rose rapidly and the increase was nearly 50% just in 2015. It is not only because of its strong ability to absorb population, resulting in huge housing demand, but also because of the government's deregulation policy, reducing the real estate market access threshold of residents.

**4.2. Econometric Model Setup.** We have mentioned that reasonable housing price in general equilibrium state is proportional to income. Therefore, if we take per capita income of residents as the independent variable and housing price index as the dependent variable, we can construct a linear econometric model as follows:

$$p_{it} = \alpha_i + \beta_i PPI_{it} + \varepsilon_{it}, \quad (16)$$

where  $i$  denotes the number of research object,  $t$  denotes the time,  $p_{it}$  is the HPI of research object  $i$  at given time  $t$ ,  $PPI_{it}$  is the corresponding average income, and  $\varepsilon_{it}$  is a random disturbance term. Based on the above theory, we just need to analyze the residual series  $\varepsilon_{it}$  to test the market efficiency.

Firstly, we carried out the regression analysis on collected data according to the established econometric model

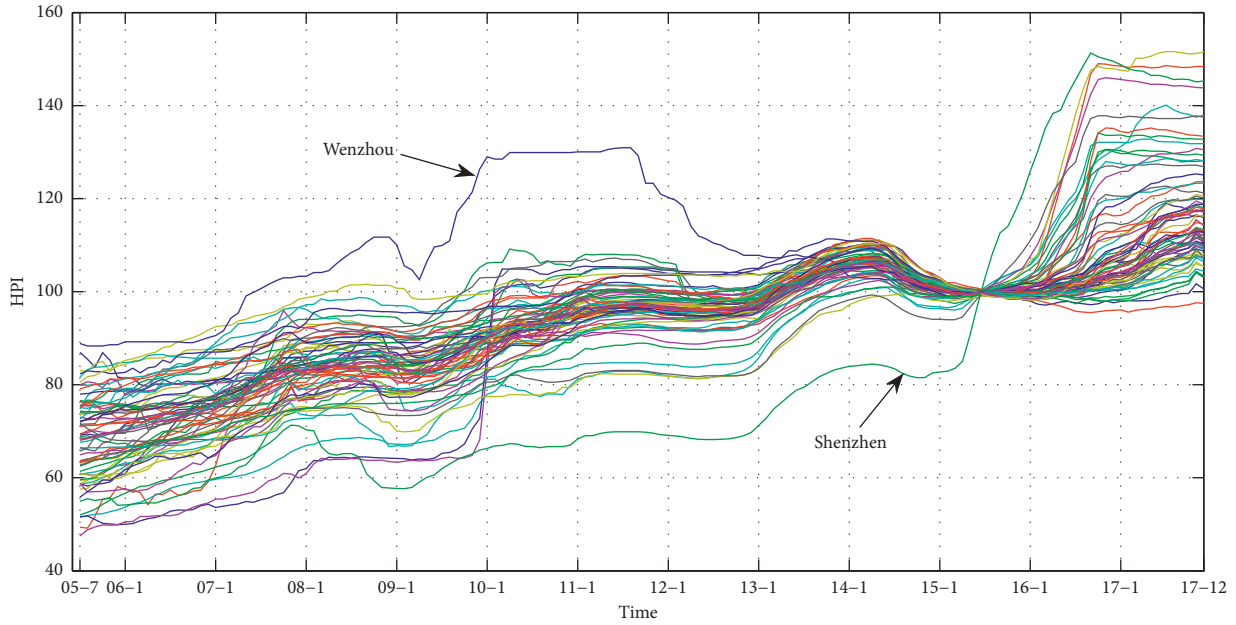


FIGURE 2: HPI trend of each city.

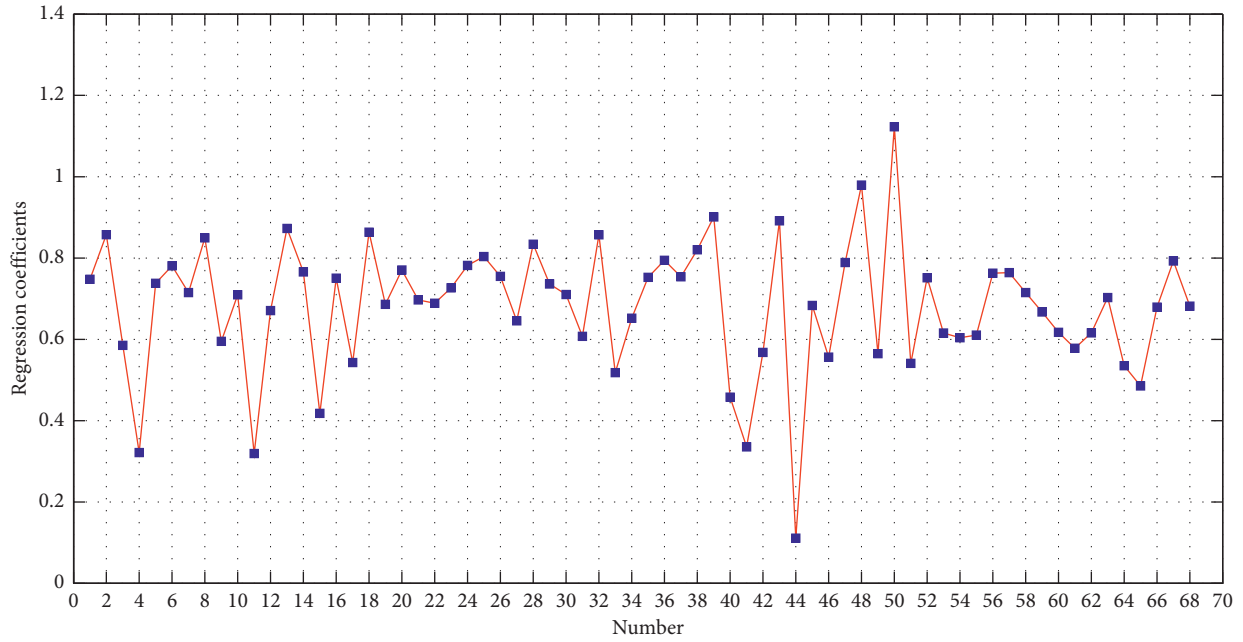
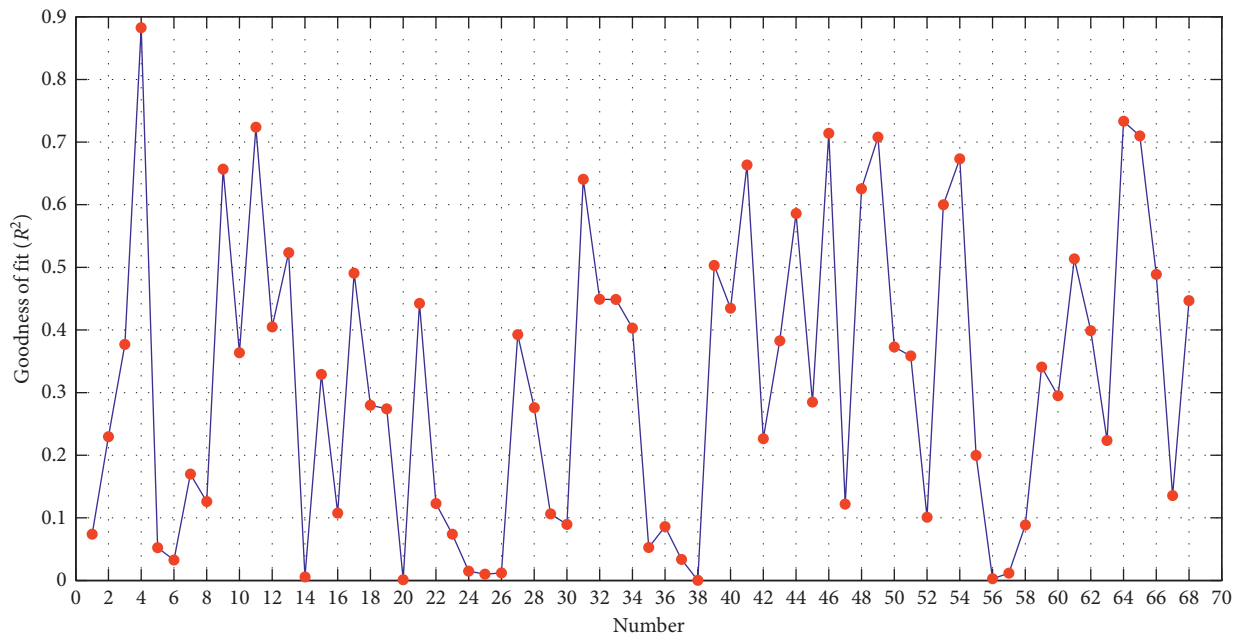
and then obtained regression coefficients and determination coefficients (also called goodness of fit), as shown in Figures 3 and 4. The signs of regression coefficients in Figure 3 are all positive. The results show that HPI has positive correlation relationship with the PPI to some extent, but the correlation degree varies with cities. In most cities, the results of  $t$ -test aimed at regression coefficients are significant at the 1% confidence level, and the  $p$  value approaches to 0. However, the  $t$ -test results of 8 cities, including Guilin, Huizhou, Xiangyang and so on, cannot realize significance at the 10% confidence level. It is clear from Figure 4 that not all housing price indexes can be fitted well by per capita income of residents. As for most cities, the explanatory power is poor. There are 17 cities whose goodness of fit is even lower than 10%, and the number of cities whose goodness of fit is higher than 50% is only 16. It implies that PPI cannot account for most of the variations in HPI, which is not consistent with the feature of a rational real estate market. However, this irrational phenomenon is in line with Chinese current real estate development status. In other words, there exist house price bubbles. Only under the condition of deep marketization, the housing price is obviously proportional to the per capita income. But the government's intervention policies play an important role in leading to the fluctuations of housing prices nowadays, besides the market's spontaneous adjustment.

**4.3. Efficiency Test.** Based on the complexity-entropy binary causal plane method, the residual in model (16) is used as the bridge to test the efficiency of Chinese housing price. In the practical application, we make the delay time  $\tau = 1$  as usual. And it is suggested that the embedding dimension  $D$  should satisfy  $M \gg D!$ , where  $M$  is the number of samples. In our study,  $M = 150$ , so the embedding dimension can take 2, 3,

4, or 5. By comparison, we determine the optimal embedding dimension  $D = 5$ . This process can be illustrated by Figure 5. There are different outcomes of 8 embedding dimensions as seen in Figure 5. When the dimension is gradually increased from 2 to 5, the correlation between permutation entropy and complexity is still the same, that is, a negative correlation. However, when the embedding dimension  $D$  is larger than 5, the relationship between them is suddenly reversed, that is, a positive correlation. Therefore, it is valid to set  $D = 5$ . In addition, as the embedding dimension changes, the degree of divergence for different points also changes. We can find that it can distinguish each state to the greatest extent at  $D = 5$ , and the effect of remaining dimensions gradually declines, such as  $D = 2$  and  $D = 8$ . It is shown on the graph that all points are concentrated in a narrow area. It is easy to explain this phenomenon. When the value of embedding dimension is small, there are few states. For example, there are only two states when  $D = 2$ , which is naturally difficult to distinguish. In an extreme case, there is only one state when  $D = 1$ , and the results of all states must be the same. It does not make any difference. On the other hand, when the value of embedding dimension is large, we can get a large number of states. For example, there is 40320 states when  $D = 8$ , but we only have 150 samples. These samples can only touch a limited number of states. Other states are not accessible at all, which results in the poor performance of divergence. Also, this explains why  $M \gg D!$  is usually required.

After determining the optimal embedding dimension, we begin to calculate the complexity and efficiency of real estate markets. Table 2 shows the descriptive statistics of them.

For intuitive expression, we plot the results into a complexity-entropy binary causal plane. Figure 6 shows the complexity-efficiency binary map of 67 cities' real estate

FIGURE 3: Regression coefficients  $\beta$ .FIGURE 4: Goodness of fit  $R^2$ .

market. It can be found that there is higher efficiency and lower complexity at the lower right corner, in which the real estate market is more efficient and simpler. Otherwise, the market efficiency is low and the market structure is complex.

According to the empirical results, China does not have completely efficient real estate market, that is, no real estate market is completely random. The city with the highest efficiency in Chinese real estate market is Tangshan (No. 48), whose efficiency has been over 0.91. Also, this city has the lowest complexity, which is only 0.103. On the contrary, the real estate market of Shenzhen city (No. 44) is the least

efficient and most complex. To further study the relationship between the efficiency and complexity of the real estate market, Table 3 lists the top 10 cities in four extreme cases. They are the most efficient, the least complex, the least efficient, and the most complex, respectively. It is easy to see from the cities listed in Table 3 that there is a significant relationship between the efficiency of real estate markets and the complexity of them.

Market efficiency indicator  $H_S$  reflects whether the real estate market price can fully respond to the market information. It has been proved by our empirical test that the real

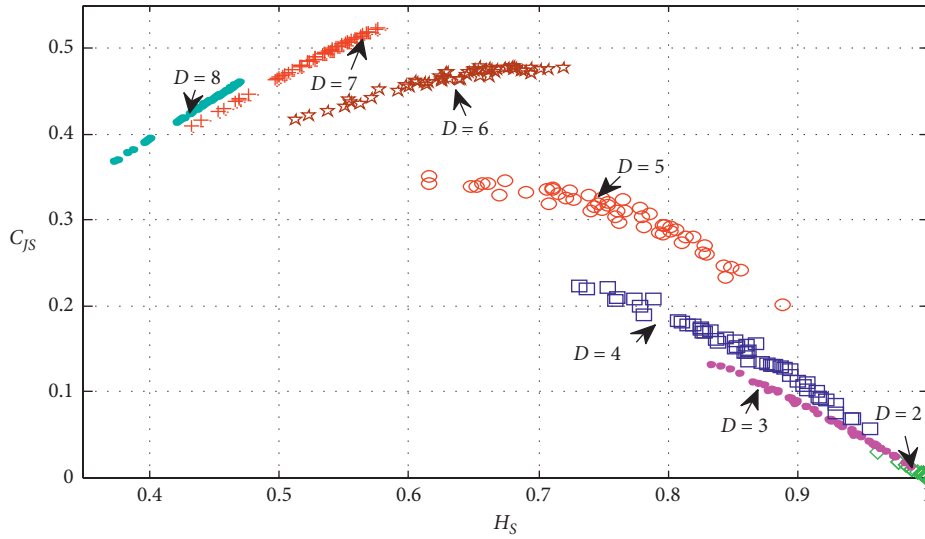


FIGURE 5: The selection of embedding dimension.

TABLE 2: The simple descriptive statistics of efficiency and complexity.

	Max	Min	Mean	Median	Std
Efficiency $H_S$	0.9104	0.5906	0.7862	0.7924	0.0660
Complexity $C_{JS}$	0.2753	0.1030	0.1994	0.2021	0.1030

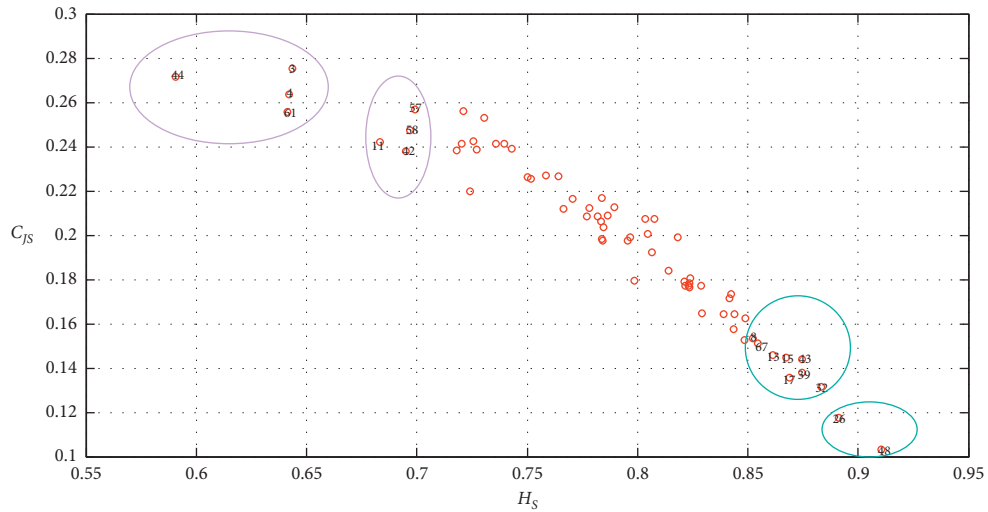


FIGURE 6: Complexity-entropy binary causal plane of Chinese real estate market.

estate markets of 67 main cities do not reach weak form efficiency, which implies that housing prices have a poor performance in responding to the market information and the liquidity of information is not enough. This is related to the nature of the real estate market. Compared with other capital markets, the real estate market has many unique characteristics, such as industrial barriers to entry, regional differences, vulnerable to policy, and housing heterogeneity. Also, there is a monopoly phenomenon in the land development rights. Under the function together with the aforementioned factors, there is less probability to form the

housing price just by the complete market competition. The efficiency of market information transmission becomes low, so it cannot meet the requirements of a weak form effective market. Furthermore, it is much difficult to quickly make the market weakly effective and meet the basic housing needs of residents only by relying on the spontaneous regulation of market. Therefore, considering the importance of real estate market to the national economy and people's livelihood, it is indispensable for the government to implement some reasonable and effective regulation measures and policy guidance.

TABLE 3: Top 10 cities in four extreme cases in Chinese market.

Features	Cities
Top 10 cities with the most efficient real estate markets	Tangshan (0.9104), Kunming (0.8913), Nanchong (0.8833), Shao guan (0.8747), Quanzhou (0.8746), Hefei (0.8691), Haikou (0.8675), Guilin (0.8614), Zunyi (0.8545), and Dandong (0.8522)
Top 10 cities with the least complex real estate markets	Tangshan (0.1030), Kunming (0.1174), Nanchong (0.1316), Hefei (0.1357), Quanzhou (0.1379), Shao guan (0.1439), Haikou (0.1447), Guilin (0.1460), Zunyi (0.1511), and Dandong (0.1526)
Top 10 cities with the least efficient real estate markets	Shenzhen (0.5906), Yueyang (0.6415), Beijing (0.6420), Beihai (0.6437), Guangzhou (0.6831), Shanghai (0.6951), Yantai (0.6968), Xuzhou (0.6992), Jining (0.7180), and Xiamen (0.7203)
Top 10 cities with the most complex real estate markets	Beihai (0.2735), Shenzhen (0.2715), Beijing (0.2635), Xuzhou (0.2567), Yueyang (0.2557), Huizhou (0.2532), Yantai (0.2473), Wuhan (0.2424), Guangzhou (0.2421), and Jinan (0.2413)

The market complexity indicator  $C_{JS}$  represents the complexity of the market structure itself. There are always multilevel markets or various factors affecting the market price, especially nonmarket factors derived from government intervention. It has been illustrated by the complexity of real estate markets in main Chinese cities. Cities with a relatively complex real estate market structure in China are mainly concentrated in first-tier cities, represented by Beijing, Shanghai, Guangzhou, and Shenzhen. These cities have more advantageous resources, thus leading to a massive influx of population and then forming a sharp expansion of housing demand and pushing housing prices further. At this time, the spontaneous regulation of the market could not meet the needs of residents. In order to coordinate the unbalanced allocation of resources and stabilize social development, the government has to take a series of regulation measures, which increased the complexity of the market structure inevitably. As for those less developed cities, the complexity of their real estate market structures is relatively low because of stable supply and demand relationship, better spontaneous regulation ability, and less government intervention. The lower  $C_{JS}$  is, the simpler the market structure is and the easier it is to be improved. Conversely, the more complex the market structure is, the more difficult it is to be governed. For the market with complex structure, it is necessary to analyze different factors more comprehensively and take targeted combination measures. Generally, market efficiency is highly correlated with complexity, but not always. The previous empirical results have revealed this point. Due to the disadvantages brought by complexity market structure, it is more difficult to deal with the higher complexity market when the degree of efficiency is at the same level.

*4.4. Efficiency Comparison of Chinese and American Real Estate Markets.* The development of American real estate market has gone through many years, most of which was relatively stable and made great contributions to the healthy and rapid development of American economy. Moreover,

American market has experienced a huge crisis, especially the global financial crisis caused by the burst of the real estate bubble in 2007. After the crisis, the real estate was still an important engine of economic growth in America. With the recovery of American economy, the real estate industry also took a turn for the better. From the perspective of economic cycle, the development of the real estate market in the United States has gone through a complete economic cycle. Moreover, the housing system of the United States was recognized as a relatively successful system by housing experts from all over the world at that time [31]. The United States adopted government intervention, but did not “take all risks.” It not only had appropriate social security of the government but also made the residents do their best to afford housing consumption.

Although the national conditions and economic development stages are obviously different, China and United States are both big countries, and there is heterogeneity of development between different cities or regions. After the reform and opening up in 1978, China took the road of unbalanced development and adopted the development strategy of “give priority to supporting the development of the eastern coastal areas.” After years of development and accumulation, this strategy directly led to the regional imbalance of economy among the central, western, and eastern regions. Even in the same region, regional central cities and other noncentral cities have differences. Compared with the noncentral city, the central city gathers more resources and capital, promotes the development of local economy, attracts more people to gather in the central city, and improves the local house price. This is the same in China and the United States. Song and Gao [32] pointed out that due to the influence of geographical factors and economic development level, the American housing price growth of nine regions divided by the federal real estate industry monitoring office had huge difference, even among different states in the same region. However, even in this unbalanced development situation, the real estate industry in the United States has become the driving force of economic growth, while providing sufficient social security. Therefore, it has important

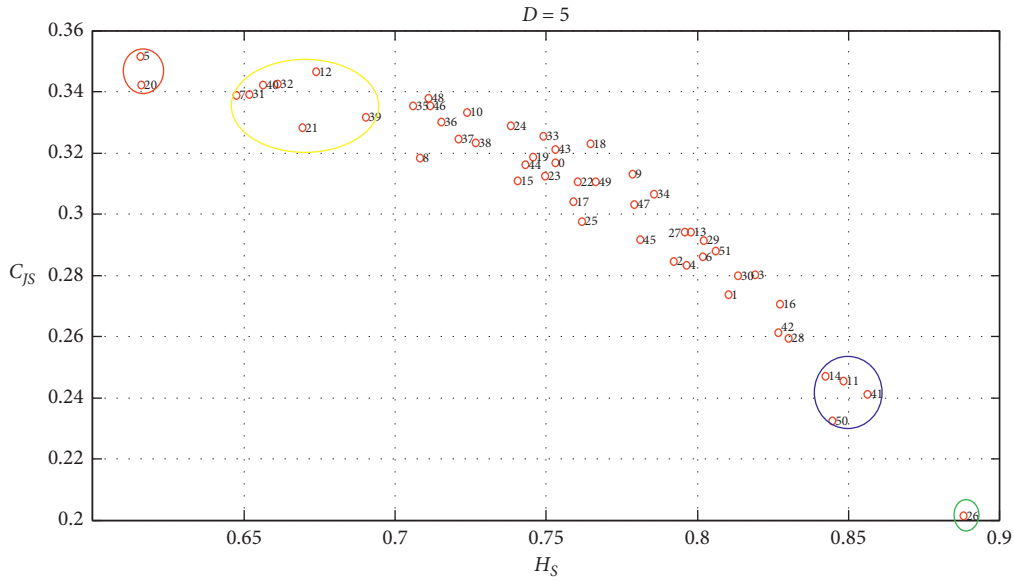


FIGURE 7: Complexity-entropy binary causal plane of American real estate market.

TABLE 4: Top 10 states in four extreme cases in American market.

Features	States
Top 10 states with the most efficient real estate markets	Mississippi, South Carolina, Georgia, West Virginia, Idaho, North Carolina, Indiana, South Dakota, Arkansas, and Nebraska
Top 10 states with the least complex real estate markets	Mississippi, West Virginia, South Carolina, Georgia, Idaho, North Carolina, South Dakota, Alaska, and Nebraska
Top 10 states with the least efficient real estate markets	California, Massachusetts, Connecticut, New Hampshire, Rhode Island, New Jersey, Maryland, Hawaii, Pennsylvania, and New York
Top 10 states with the most complex real estate markets	California, Hawaii, New Jersey, Massachusetts, Rhode Island, New Hampshire, Connecticut, Washington, Virginia, and New York

reference and warning value for Chinese real estate industry if we take it as the object of comparative study.

In order to study the efficiency of the American real estate market, we select 50 states and the District of Columbia as our research objects. Similar to China, we also studied the efficiency of American overall real estate market. We collected relevant data from the first quarter of 1975 to the first quarter of 2014, summing up to 157 quarters. Individual missing data were supplemented by linear interpolation, and we made seasonal adjustments to eliminate the influence of seasonal factors. Using the same econometric model and test method, we obtained the efficiency and complexity of American real estate market. As can be seen from Figure 7, the real estate market of each state and the overall market in America have not reached a weak efficient state, which is the same as China.

The state with the highest degree of efficiency in American real estate market is Mississippi, whose efficiency reaches 0.886, while its complexity is only 0.201 at the lowest complexity level in America. On the other hand, the state with the least efficient and most complex real estate market is

California. Its efficiency and complexity are 0.616 and 0.352, respectively. The significant negative correlation between the efficiency and the complexity in American market is consistent with that in Chinese market. To further illustrate the relationship between efficiency and complexity, Table 4 shows the top 10 states in four extreme cases in American real estate market.

Combined with Tables 3 and 4, it can be seen that the relationship between the efficiency of the real estate market and the level of regional economic development in China is contradictory to that in America. In China, those cities with high efficiency and low efficiency are most of the third-tier or below cities, such as Tangshan (No. 48), Shao guan (No. 43), Guilin (No. 13), Zunyi (No. 67), and so on. The first-tier or second-tier cities with developed economy and high comprehensive level are always less efficient and more complex in real estate market, such as Shenzhen (No. 44), Beijing (No. 4), Guangzhou (No. 11), Xuzhou (No. 57), and so on. As for the United States, it is just the opposite. Those states with higher efficiency and lower complexity are mostly located in the east and west coasts, such as Mississippi (No. 26), South



Carolina (No. 41), Georgia (No. 11), West Virginia (No. 50), and northern Idaho (No. 14). These states all belong to the area whose economy is relatively developed and population is dense. The less efficient areas are located in the middle of the United States and with poor economy. So how to ensure the efficiency and orderliness of the real estate market while developing the economy is what we can learn from the United States.

From the perspective of the relationship between the efficiency and complexity of the real estate market, the performance of Chinese market is consistent with that of American market. It can be seen from Tables 3 and 4 that cities with higher efficiency tend to have lower complexity, and the ranking order is almost the same. In order to clarify the generality of this correlation, we rank the efficiency and complexity and calculate their Spearman order statistical correlation coefficient. Then, we obtain the rank correlation coefficient of efficiency and complexity of Chinese real estate market and American real estate market. The correlation coefficients are  $-0.975$  and  $-0.966$ , respectively, which explains the obvious negative correlation between the efficiency and complexity of the real estate market well.

## 5. Conclusions and Suggestions

By the complexity-entropy binary causal plane method, we measure the efficiency and complexity of Chinese and American real estate markets. Our conclusions can be summarized as follows:

- (1) Neither 67 main cities' real estate markets in China nor the overall real estate market has reached weak form efficiency. Both the efficiency and complexity vary with each subject. There is significantly negative correlation relationship between the degree of efficiency and complexity because their Spearman rank correlation coefficient has reached  $-0.975$ . In Chinese cities, those with a high efficiency degree of real estate market are generally second-tier, third-tier, and lower cities, while those with a relatively developed economy such as Beijing, Shanghai, Shenzhen, and Guangzhou show a lower efficiency degree of real estate market and a higher complexity degree.
- (2) The market efficiency and complexity of the United States are similar to that of China. Specifically, the real estate market in all states and the overall American market have not reached weak form efficiency. The Spearman rank correlation coefficient between the efficiency and complexity is  $-0.966$ , showing the significant negative correlation. Different from China, the areas with high efficiency degree of American real estate market are basically located in the east and west coasts which are relatively developed and densely populated. On the contrary, the areas with low efficiency are located in the middle with relatively low level of economic

development. Moreover, American real estate market is about as efficient as Chinese, but significantly more complex.

Combining the development history of American and Chinese housing finance, it can be found that housing price fluctuation of both countries has close relationship with the enforcement of some housing finance policies. For example, the irrational surges of housing prices over a period are mostly due to the government's efforts to stimulate the economy. Those interventions from government can lead to an excessive prosperity of the real estate market. From the perspective of policy connotation, the more efficient the market is, the less the government's intervention is. The lower degree of efficiency implies the less freedom of the market. In detail, some systems or rules set by the government affect the market's independent regulation. Therefore, on the one hand, it is necessary to formulate and implement some innovative regulation policies to guide the development of real estate market; on the other hand, it is required to find out the deficiencies of the current relevant policies and systems and then improve them. According to the conclusions of our study and the achievements of the housing price control policies issued by Chinese and American governments over the years, we put forward following suggestions to improve the efficiency of the real estate market in China:

- (1) The government should improve mechanisms to enhance information transparency of the real estate market. If the real estate market is efficient, prices can respond adequately and quickly to other market information. At present, all main cities and the overall real estate market in China have not reached weak form efficiency. This phenomenon indicates that there is information asymmetry in the market and the information transmission mechanism among investors, sellers, and buyers is not perfect. The government can make more detailed rules to help information disclosure in the process of real estate transactions, so as to guarantee the liquidity, accuracy, and timeliness of information. Furthermore, it can reduce the information search cost, guide rational investment, and reduce the possibility of releasing false information by developers and speculators. Only in this way, we can ensure fair real estate market transaction and establish an information disclosure mechanism with openness, fairness, and impartiality to promote the healthy development of housing industry.
- (2) The government should advocate both renting a house and buying a house to live in so as to promote the establishment of multilevel housing system. In some developed countries, such as Germany, the rental rate is as high as 58%, but most residents still can live and work happily. The supply of a large number of rental housing has not only increased the well-being of residents but also played an important role in maintaining social stability. This harmonious

phenomenon is inseparable from the perfect rental management system. As for China, there is large population, which results in strong demand for houses. However, because of the high housing price, many people, especially those young people freshly entering the workplace, cannot afford to buy own houses and have no choice but to rent a house. In order to meet the housing needs of low-income groups and maintain the healthy development of housing economy, policy guarantee should be accompanied by legislative regulation. Drawing on the classification supply system of the United States, the government should move faster to put in place a housing system that ensures supply through multiple sources, provides housing support through multiple channels, and encourages both housing purchase and renting. This will make us better placed to meet the housing needs of all of our people.

- (3) The government should take targeted and differentiated measures. In the light of the empirical validation results in 67 cities, the efficiency degree and complexity degree of the real estate market vary from cities. Therefore, the regulation of real estate market in each city cannot be treated as the same; the government should take these factors into consideration, such as economic level, institutional environment, resource allocation, and the supply and demand of urban residents. Targeted and differentiated measures can play the role of regulating housing prices and stabilizing the market more effectively. For instance, there is a big difference in the housing inventory between the first- or second-tier cities and the third- or fourth-tier cities in China. For the third- and fourth-tier cities, destocking should be emphasized, while the first- and second-tier cities should actively prevent the risk of real estate bubble.
- (4) The government should strengthen early monitoring and warning and then prevent international financial risks. In the context of economic globalization, changes in the international situation or the economic situation of a certain country often affect all over the world. Chinese trade with other countries has become increasingly close, and it is more impossible to stay independent. It is memorable that the trade war between Japan and the United States made a rapid rise in the yen over 2 decades ago. To tackle this crisis, the Japanese government decided to implement easy monetary policy and encourage local residents to invest real estate industry, making Japanese economy grow with bubbles. Finally, the burst of the bubble economy not only led to the Japanese house prices to decline dramatically but also caused the Japanese economy to collapse. We can learn from this trade war that we should pay close attention to foreign exchange risks, actively guard against the exchange rate risks caused by foreign capital inflow and domestic capital outflow,

and strengthen risk early warning so as to prevent and tackle risks timely.

## Appendix

The software applied to our work is MATLAB.

The syntax about how to calculate the permutation entropy and statistical complexity measure is presented as follows:

```
function permutation()
clear;
clc;
data=xlsread('ppires'); %input data
[m,n]=size(data); for j=1:n
z=j %select a series
rs0=data(:,z); %rs0=data(:,1)+data(:,2)
D=5 % embedding dimension
tao=1 % time delay
P=perms([1:1:D]); % all Permutations of D
np=length(P(:,1)); % the number of permutations D!
rnp=zeros(np,1); % the number of patterns, initial value=0
for i=1:m-(D-1)*tao
rs=rs0(i:tao:i+(D-1)*tao); %find the initial vector used for rank
[B,IX]=sort(rs); % rank by ascending order and give the original serial number, or called mode
[b,IX1]=sort(P * IX); % Judge mode number according to product maximum
kp=IX1(end); %give the order under this ordinal pattern
rnp(kp)=rnp(kp)+1; end
ppi=rnp/(m-(D-1)*tao) %calculate the probability of each pattern
sum(ppi);
ppi1=ppi+(ppi==0);
PE=-sum(ppi.*log(ppi1)) %calculate permutation entropy
HS=PE/log(np) % maximum-entropy formalism
ppi0=[1;zeros(np-1,1)];
JP=-sum((ppi+1/np)/2.*log((ppi+1/np)/2))-(PE+log(np))/2;
JPO=-sum((ppi0+1/np)/2.*log((ppi0+1/np)/2))-(0+log(np))/2;
CJS=JP/JPO*HS
std(rs0)
hs(j)=HS; cjs(j)=CJS; end
hold on
```

```

box on
grid on
scatter(hs,cjs,'r')
xlabel('H_S')
ylabel('C_{JS}')
hs
cjs
hold on
%plot(hs./cjs)
plot(hs)
xlswrite('csi',[hs;cjs]')
%xlswrite('acjsl',[hs;cjs]')
xlabel('\it H_S')
ylabel('\it C_{JS}')

```

## Data Availability

The data used to support the findings of this study are available from the Wind database, Chinese economic and social big data research platform, Federal Reserve System(H), All-America Economic Survey, and the Bureau of Economic Analysis.

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the Humanities and Social Science Planning Fund from Ministry of Education (16YJAZH078), the Fundamental Research Funds for the Central Universities of China (grant nos. CCNU19TS062, CCNU19A06043, and CCNU19TD006), and the raising initial capital for High-Level Talents of Central China Normal University (30101190001).

## References

- [1] R. Meese and N. Wallace, "Testing the present value relation for housing prices: should I leave my house in san Francisco?," *Journal of Urban Economics*, vol. 35, no. 3, pp. 245–266, 1994.
- [2] J. Clayton, "Are housing price cycles driven by irrational expectations," *The Journal of Real Estate Finance and Economics*, vol. 14, no. 3, pp. 341–363, 1997.
- [3] K. E. Case and R. J. Shiller, "Forecasting prices and excess returns in the housing market," *Real Estate Economics*, vol. 18, no. 3, pp. 253–273, 1990.
- [4] J. M. Abraham and P. H. Hendershott, "Bubbles in metropolitan real estate markets," *NBRE Working Papers*, vol. 7, no. 35, pp. 171–192, 1994.
- [5] S. R. Grenadier, "The persistence of real estate cycles," *The Journal of Real Estate Finance and Economics*, vol. 10, no. 2, pp. 95–119, 1995.
- [6] P. Kunzel, *Inefficiencies in the Real Estate: Implications for the Price Dynamics*, Ph.D. dissertation, The George Washington University, Washington, DC, USA, 2004.
- [7] K. Wang, Y. Zheng, and H. Liu, "Review of the efficiency of real estate market in China," *China Land Science*, vol. 20, no. 5, pp. 54–59, 2006.
- [8] L. Xie, *Empirical Research on the Efficiency of Shanghai Real Estate Market*, East China Normal University, Shanghai, China, 2014.
- [9] A. Pan and H. Wang, "Inefficiencies of the real estate market in China: theory and empirical analysis," *Finance and Economics*, vol. 7, pp. 55–63, 2008.
- [10] Q. Meng and R. Chen, "On real estate price bubbles of China: an empirical study based on Markov switching model," *Financial Research*, vol. 2, pp. 105–120, 2017.
- [11] K. Guo and Y. Huang, "Problems and solutions of China's real estate market development on international comparisons," *Finance & Trade Economics*, vol. 1, pp. 5–22, 2018.
- [12] C. Chen, "Theoretical and practical analysis of the ratio of house price to income," *Times Finance*, vol. 5, pp. 18–20, 2014.
- [13] K. E. Case and R. J. Shiller, "Is there a bubble in the housing market?," *Brookings Papers on Economic Activity*, vol. 2003, no. 2, pp. 299–362, 2003.
- [14] L. Zunino, M. Zanin, O. A. Pérez, D. G. Pérez, and O. A. Rosso, "Forbidden patterns, permutation entropy and stock market inefficiency," *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 14, pp. 2854–2864, 2009.
- [15] L. Zunino, M. Zanin, B. M. Tabak, D. G. Pérez, and O. A. Rosso, "Complexity-entropy causality plane: a useful approach to quantify the stock market inefficiency," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 9, pp. 1891–1901, 2010.
- [16] O. A. Rosso, H. A. Larrondo, M. T. Martin et al., "Distinguishing noise from chaos," *Physical Review Letters*, vol. 99, no. 15, Article ID 154102, 2007.
- [17] L. Gulko, "The entropic market hypothesis," *International Journal of Theoretical and Applied Finance*, vol. 2, no. 3, pp. 293–329, 1999.
- [18] D. Matesanz and G. J. Ortega, "A (econophysics) note on volatility in exchange rate time series," *International Journal of Modern Physics C*, vol. 19, no. 7, pp. 1095–1103, 2008.
- [19] W. A. Rizzo, "The role of the informational efficiency in the dotcom bubble," *Social Science Electronic Publishing*, vol. 17, no. 5, pp. 373–380, 2008.
- [20] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, Article ID 174102, 2002.
- [21] M. Riedl, N. A. Müller, and N. Wessel, "Practical considerations of permutation entropy," *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 249–262, 2013.
- [22] A. Ortiz-Cruz, E. Rodriguez, and C. J. Alvarez-Ramirez, "Efficiency of crude oil markets: evidences from informational entropy analysis," *Energy Policy*, vol. 41, pp. 365–373, 2012.
- [23] G. Oh, S. Kim, and C. Eom, "Market efficiency in foreign exchange markets," *Physica A: Statistical Mechanics and its Applications*, vol. 382, no. 1, pp. 209–212, 2007.
- [24] M. Staniek and K. Lehnertz, "Parameter selection for permutation entropy measurements," *International Journal of Bifurcation and Chaos*, vol. 17, no. 10, pp. 3729–3733, 2007.
- [25] M. Matilla-García and M. Ruiz Marín, "Detection of non-linear structure in time series," *Economics Letters*, vol. 105, no. 1, pp. 1–6, 2009.
- [26] L. Faes, A. Porta, M. Javorka et al., "Efficient computation of multiscale entropy over short biomedical time series based on

- linear state-space models,” *Complexity*, vol. 2017, Article ID 1768264, 13 pages, 2017.
- [27] Z. Chen, Y. Li, H. Liang et al., “Improved permutation entropy for measuring complexity of time series under noisy condition,” *Complexity*, vol. 2019, Article ID 1403829, 12 pages, 2019.
- [28] P. W. Lamberti, M. T. Martin, A. Plastino, and O. A., “Intensive entropic non-triviality measure,” *Physica A: Statistical Mechanics and Its Applications*, vol. 334, no. 1-2, pp. 119–131, 2004.
- [29] D. Rosso, M. Varela-Entrecanales, A. Molina-Pico et al., “Patterns with equal values in permutation entropy: do they really matter for biosignal classification?,” *Complexity*, vol. 2018, Article ID 1324696, 15 pages, 2018.
- [30] H. Shaobo, S. Kehui, and W. Huihai, “Modified multiscale permutation entropy algorithm and its application for multiscroll chaotic systems,” *Complexity*, vol. 21, no. 5, pp. 52–58, 2016.
- [31] Q. Wu, “Referring international experience to rationalize China’s housing development ideas,” *Economic Issues*, vol. 28, no. 2, pp. 13–15, 2006.
- [32] Y. Song and L. Gao, “The prosperity and risk of American real estate industry and its impact on American economy,” *American Studies*, vol. 20, no. 3, pp. 65–76, 2006.

## Research Article

# Grey Spectrum Analysis of Air Quality Index and Housing Price in Handan

**Kai Zhang, Yan Chen , and Lifeng Wu **

*College of Management Engineering and Business, Hebei University of Engineering, Handan 056038, China*

Correspondence should be addressed to Yan Chen; 964129856@qq.com

Received 7 August 2019; Accepted 18 September 2019; Published 25 November 2019

Guest Editor: Francesco Tajani

Copyright © 2019 Kai Zhang 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.

To analyze the relationship between air quality index (AQI) and housing price, six relationship indexes between air quality index and housing price were calculated using grey spectrum theory, specifically grey association spectrum, grey cospectrum, grey amplitude spectrum, grey phase spectrum, grey lag time length, and grey condense spectrum. Three main change periods were extracted. There was a negative correction between the air quality and the housing price in Handan. The results provide a basis for the government's measures to prevent haze.

## 1. Introduction

The real estate industry drives the economy. The added value of real estate and related industries accounts for 15.8% of China's gross domestic product (GDP), the share of real estate in GDP is 3% in America, and the data for Japan are 11%. Beijing–Tianjin–Hebei, the Yangtze River Delta, the Pearl River Delta, Chengdu–Chongqing, and the middle reaches of the Yangtze River are China's five largest city clusters. Their 11% land area contributes 55% to the economy and 47% to commercial housing sales (Boao 21<sup>st</sup> Century Real Estate Forum 19<sup>th</sup> Annual Conference). Research on housing price has been performed. Housing prices have been driven by land price and have had a strong effect on land price [1]. The increase in population is expected to reduce the land area per capita, resulting in the rise of housing price [2]. There are commercial housing and social housing in China, and commercial housing has been the focus of most research works [3]. High housing price will cause social conflicts, and the Chinese government has launched a series of measures to limit housing price [4]. For the government to formulate more accurate measures of housing price, a long short-term memory approach is proposed to predict the housing price [5].

In addition to creating a lot of jobs, economic development has also created a lot of problems, such as air pollution.

Handan, a city in southern Hebei Province, is shown in Figure 1. Gross domestic product and total construction output showed the same trend from 2015 to 2018 (Figure 2). Economic growth slowed in 2017 as the country adjusted its industrial structure. The development of the construction industry has raised the housing price and brought about the problem of environmental pollution. AQI in Handan from 2014 to 2018 is shown in Figure 3. As shown, polluted months were more than that of good months. Air pollution reduces the quality of human environment.

Air pollution is still a major problem faced by the world today. Air quality is affected by a variety of factors. Many scholars have performed research on air quality, including studies on the influence of tree species on air quality improvement [6]. Despite this, it is generally agreed that human socioeconomic activities have the greatest impact on the environment [7, 8]. Air quality problems are common in China, and haze is not evenly distributed: the developed eastern regions face more serious air quality problems than western regions [9]. Under the Chinese government's strict emission measures, air quality has improved greatly [10]. Analysis of factors related to air pollution is the key to haze control. The relationship between outdoor air pollution and sick building syndrome symptoms is researched [11]. The problem of air pollution has been taken seriously in recent years in China, and equipment to measure air quality have

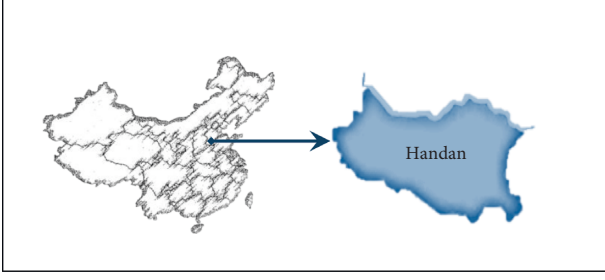


FIGURE 1: Location of Handan in Hebei Province.

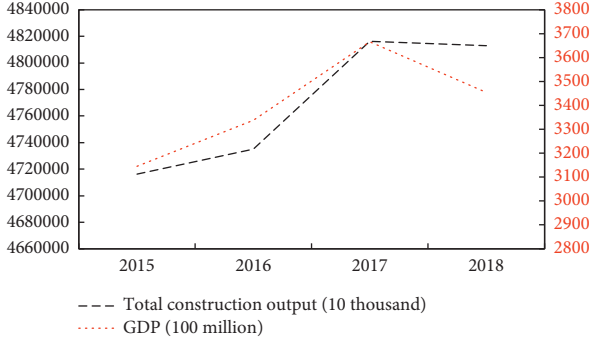


FIGURE 2: Total construction output and gross domestic product (GDP) from 2015 to 2018 in Handan.

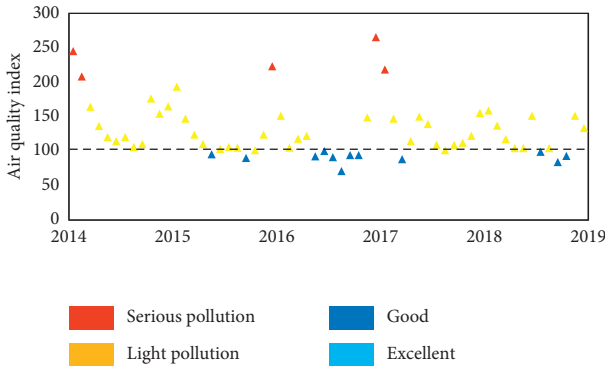


FIGURE 3: Air quality index (AQI) in Handan from 2014 to 2018.

been installed in many cities. Air quality data are incomplete. Grey system theory can address the small sample sizes and poor quality of data [12]. Scholars have performed further research projects. The traditional grey model has been used to provide particulate matter information for the roadside inhabitants [13]. The grey model with fractional order accumulation has been used to predict air quality [14], and grey relational analysis has been used to determine whether carbon price has multiple timescales [15].

The purpose of this study is to assess the relationship between AQI and housing price. The mechanism by which air quality affects housing price has been studied, and suggestions for government programs are provided here.

The paper has the following structure. The calculation process of grey spectrum theory is presented in Section 2. The grey spectrum theory used in real-world cases is presented in Section 3. The paper is summarized in Section 4.

## 2. Method

The degree of correlation in grey relational analysis was here substituted for the correlation number in spectrum analysis, and then grey association spectrum, grey cospectrum, grey condense spectrum, grey phase spectrum, grey lag time length, and other indicators were found [16, 17]. Results are shown in Figure 4. The grey spectrum algorithm is presented below.

*Step 1.* Standardization:

$X_i = \{X_i(t), t = 1, 2, \dots, m; i = 1, 2\}$  is the discrete time series ( $m$  represents the sample size and  $i$  is the sequence name). To make each sequence comparable,  $X_i$  is standardized. The results are as follows:

$$d_i = \frac{1}{m-1} \sum_t^{m-1} |X_i(t+1) - X_i(t)|, \quad t = 1, 2, \dots, m-1, \quad (1)$$

$$y_i = \left\{ \frac{X_i(t)}{d_i}, \quad t = 1, 2, \dots, m \right\}.$$

*Step 2.* The increment sequence is defined as follows:

$$\Delta y_i = \{\Delta y_i(t) = y_i(t+1) - y_i(t), \quad t = 1, 2, \dots, m-1\}. \quad (2)$$

*Step 3.* The correlation coefficient of each time period:

$$\rho_{12} = \begin{cases} \text{sgn}(\Delta y_1(t)\Delta y_2(t+\tau)), & \frac{\min(|\Delta y_1(t)|, |\Delta y_2(t+\tau)|)}{\max(|\Delta y_1(t)|, |\Delta y_2(t+\tau)|)}, \\ 0, & (\Delta y_1(t)\Delta y_2(t+\tau) = 0), \end{cases} \quad (3)$$

where  $\text{sgn}(\Delta y_1\Delta y_2(t+\tau)) = \begin{cases} 1 & (\Delta y_1\Delta y_2(t+\tau) > 0) \\ 0 & (\Delta y_1\Delta y_2(t+\tau) = 0) \\ -1 & (\Delta y_1\Delta y_2(t+\tau) < 0) \end{cases}$ ;  $\Delta y_1$  is the controlled sequence of  $\Delta y_2$ ,  $\Delta y_2$  is calculated in the same way as  $\Delta y_1$ ; and  $\tau$  is the time difference factor.

*Step 4.* The correlation coefficient is defined as follows:

$$r_{12}(\tau) = \frac{1}{m-\tau} \sum_{t=1}^{m-\tau-1} \rho_{12}(t), \quad \tau = 0, 1, 2, \dots, \Delta. \quad (4)$$

The order of the model is represented by  $\Delta$ . On the basis of experience,  $\Delta$  is taken as  $1/10 \sim 1/3$  of the sequence length [18].

*Step 5.* Grey association spectrum and grey cospectrum are calculated. According to spectral theory and equation (4), the across spectrum is decomposed into real and imaginary parts. Real parts are represented by  $P_{12}(k)$ , and the latter is represented by  $Q_{12}(k)$ . The sequence changes are shown as waves, and the wave number of the sequence is represented by  $k$ . The formulas are as follows:

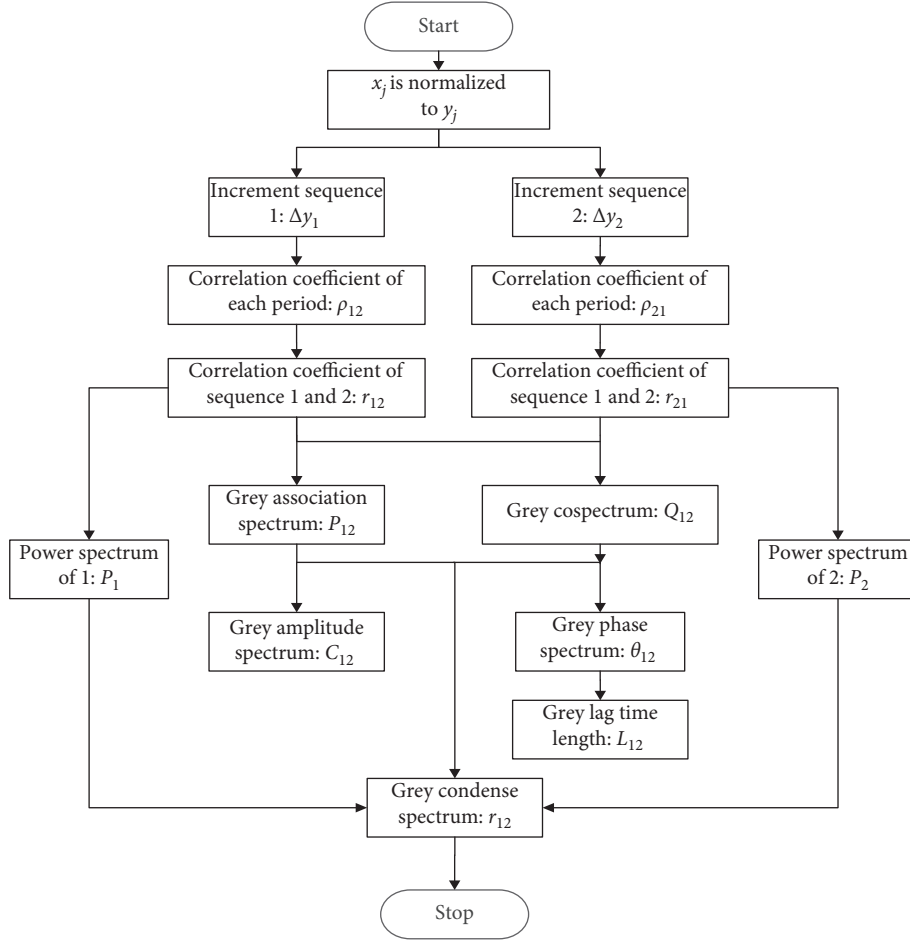


FIGURE 4: Flowchart of grey spectrum analysis.

$$P_{12}(k) = \frac{1}{\Delta} \left\{ r_{12}(0) + \sum_{\tau=1}^{\Delta-1} [r_{12}(\tau) + r_{21}(\tau)] \cos \frac{k\pi}{\Delta} \tau + r_{12}(\Delta) \cos(k\pi) \right\},$$

$$Q_{12}(k) = \frac{1}{\Delta} \sum_{\tau=1}^{\Delta-1} [r_{12}(\tau) - r_{21}(\tau)] \sin \frac{k\pi}{\Delta} \tau, \quad k = 0, 1, 2, \dots, \Delta.$$
(5)

Grey amplitude spectrum  $C_{12}(k)$ , grey phase spectrum  $\theta_{12}(k)$ , grey lag time length  $L_{12}(k)$ , and grey condense spectrum  $r_{12}(k)$  are calculated using the grey association spectrum and grey cospectrum:

$$C_{12}(k) = \sqrt{P_{12}^2(k) + Q_{12}^2(k)},$$

$$\theta_{12}(k) = \arctan \frac{Q_{12}(k)}{P_{12}(k)},$$

$$L_{12}(k) = \frac{\Delta \theta_{12}(k)}{\pi k},$$

$$r_{12}(k) = \text{sgn}\{P_1(k)P_2(k)\} \frac{\min\{P_{12}^2(k) + Q_{12}^2(k), |P_1(k)P_2(k)|\}}{\max\{P_{12}^2(k) + Q_{12}^2(k), |P_1(k)P_2(k)|\}},$$
(6)

where  $-1 \leq r_{12}(k) \leq 1$ ,  $P_1(k)$ , and  $P_2(k)$  are the grey association spectrum of two sequences of the  $k$ -th wave respectively, i.e., power spectrum.

Grey condense spectral  $r_{12}(k)$  is different from the definition of condense spectral values in spectrum analysis. The paper expands the range of grey condense spectral value, a positive value here indicates a positive correlation, and a negative value does the opposite.  $r_{12}(k)$  gets closer to  $-1$ , and the negative correlation between  $X_1$  and  $X_2$  is stronger. Similarly, as  $r_{12}(k)$  gets closer to  $1$ , the positive correlation between  $X_1$  and  $X_2$  is stronger.  $r_{12}(k) = 0$  expresses that  $X_1$  and  $X_2$  are not related.

As in spectral analysis theory, grey association spectrum reflects the same direction change components of the two sequences spectrum. Grey cospectrum represents reverse change component. Grey amplitude spectrum is composed of two components. The grey phase spectrum reflects phase difference of spectrum component of two sequences. The values range from  $-\pi$  to  $+\pi$ , and  $\pm\pi$  indicating the two grey spectrum components have exact opposite directions. In this, the grey association spectrum is zero, and the grey cospectrum is at its maximum. But if the grey phase spectrum is zero, the two grey spectrum components have

TABLE 1: Grey association and cospectrum of AQI and housing price in Handan.

Wave number	$P_1$	$P_2$	$P_{12}$	$Q_{12}$	$R_{12}$	$H_{12}$	$L_{12}$	Period
1	0.0525	0.1734	-0.0199	0.0104	0.0555	-0.4821	-1.5347	20.00
2	0.0937	0.1105	-0.0201	-0.0333	0.1463	1.0274	1.6351	10.00
3	0.1520	0.0506	-0.0199	0.0216	0.1118	-0.8258	-0.8761	6.67
4	0.1008	0.0848	0.0077	0.0023	0.0076	0.2901	0.2309	5.00
5	0.1112	0.0847	0.0191	-0.0154	0.0640	-0.6802	-0.4330	4.00
6	0.0784	0.0625	0.0182	-0.0190	0.1411	-0.8088	-0.4291	3.33
7	0.0891	0.0873	0.0403	-0.0123	0.2282	-0.2969	-0.1350	2.86
8	0.1120	0.0471	0.0087	-0.0173	0.0713	-1.1054	-0.4398	2.50
9	0.0744	0.0546	-0.0050	-0.0018	0.0068	0.3426	0.1212	2.22
10	0.0872	0.0571	0.0033	0.0000	0.0022	0.0000	0.0000	2.00

$P_1$  is the AQI power spectrum,  $P_2$  is the housing price power spectrum,  $P_{12}$  is the grey association between AQI and housing price,  $Q_{12}$  is the grey cospectrum,  $R_{12}$  is the grey condense spectrum,  $H_{12}$  is the grey phase spectrum, and  $L_{12}$  is the grey lag time length.

exactly the same direction. The grey cospectrum is zero, and the grey association spectrum is at its maximum. There is a close relationship between the six kinds of grey spectrum.

### 3. Case and Analysis

We selected a city in southern Hebei Province, Handan. AQI and housing price were here analyzed from April 2015 to December 2018. The AQI data were collected from China's air quality online monitoring platform [19]. Housing price data is from reference [20]. AQI is the main sequence, and housing price is the secondary sequence. Grey spectrum theory is used to calculate six grey spectrum indicators in Handan. The results of the analysis of the relationship between AQI and housing price are shown in Table 1.

Extracting the main period of AQI change, in this paper,  $\Delta = 10$ . AQI power spectrum of Handan is shown in Figure 5.

As shown in Table 1 and Figure 5, the three main periods in Handan AQI were extracted. The first period lasted 6.67 months; the second 2.5 months, and the third 4 months. The amplitudes of 1st to 3rd periods were 0.1520, 0.1120, and 0.1112, respectively and tended to decline. The first period of Handan housing price lasted 20 months; the second 10 months; and the third 2.86 months. The amplitudes of 1st to 3rd periods were 0.1734, 0.1105, and 0.0873, respectively. The results are shown in Table 2. We can see that the first period has the largest amplitude and the third period has the smallest, indicating that housing price contributed the most to the change in AQI in the first period.

Grey association spectrum analysis was performed in Table 1 and Figure 6. AQI and housing price have different correlations in different frequency segments. Codirectional changes were observed lasting for 2.86 months, 4 months, and 3.33 months. As shown in Figure 7, grey cospectrum was used to reflect the reversed changes component lasting for 6.67 months, 20 months, and 5 months. As shown, the negative correlation period between AQI and housing price was longer than the positive period.

The phase angle of AQI before or after housing price was analyzed using the grey phase spectrum. As shown in

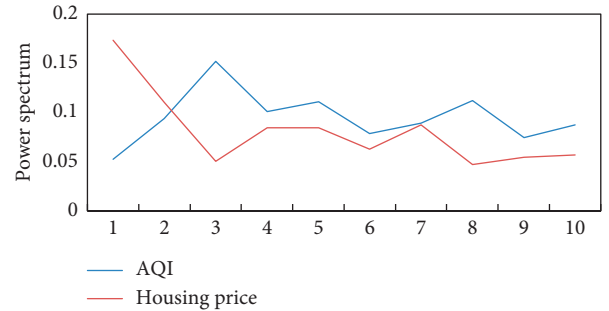


FIGURE 5: Spectrum of AQI and housing price power in Handan.

TABLE 2: Main period and amplitude of AQI and housing price in Handan.

Period	AQI		Housing price	
	Length	Amplitude	Length	Amplitude
1	6.67	0.1520	20	0.1734
2	2.5	0.1120	10	0.1105
3	4	0.1112	2.86	0.0873

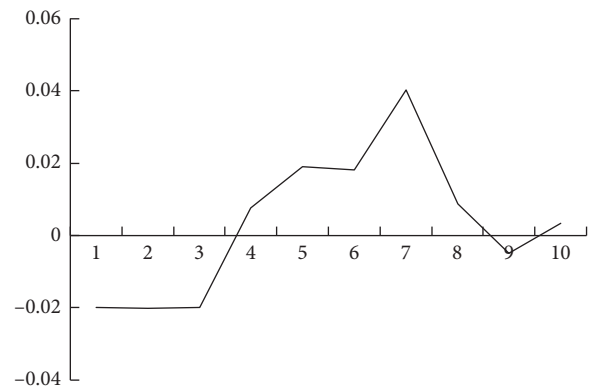


FIGURE 6: Grey association spectrum analysis of AQI and housing price.

Table 1 and Figure 8, the impact of housing price on AQI is both ahead and behind. The maximum values were 1.0274 rad and 1.1054 rad. Next, how long each value



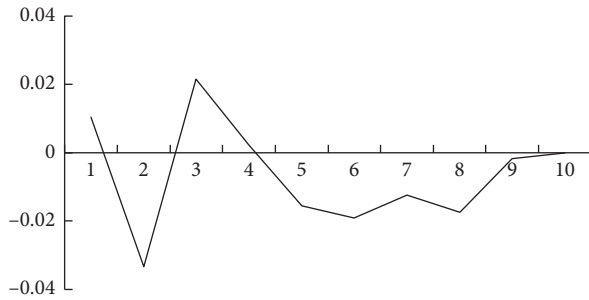


FIGURE 7: Grey cospectrum of AQI and housing price.

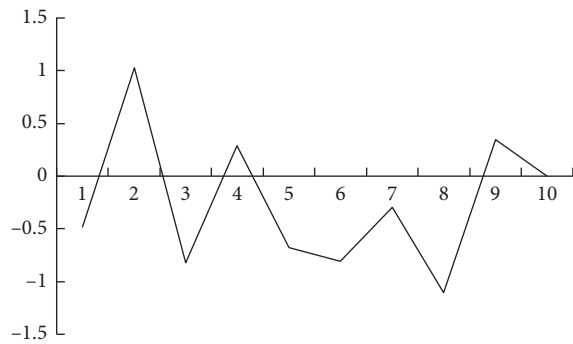


FIGURE 8: Grey phase spectrum of AQI and housing price.

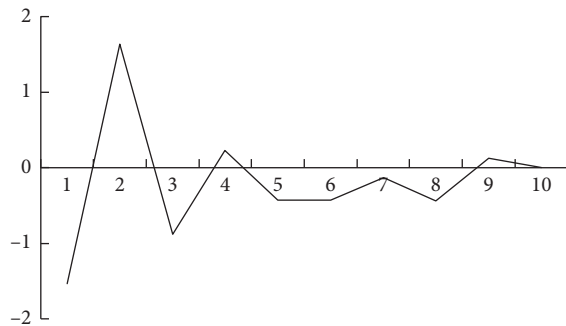


FIGURE 9: Grey lag time length of AQI and housing price.

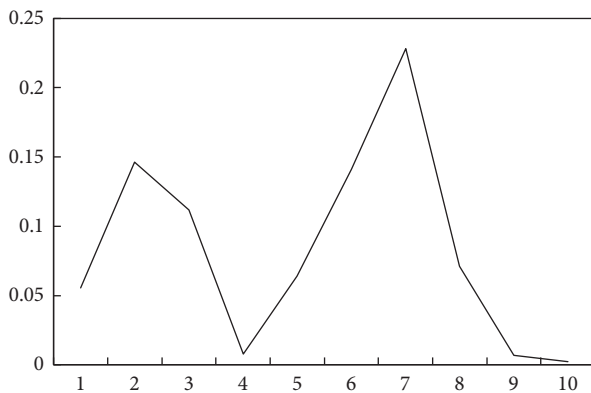


FIGURE 10: Grey condense spectrum of AQI and housing price.

remained behind or ahead was found using grey lag time length. As shown in Table 1 and Figure 9, the lag time between AQI and housing price was  $-1.5347$  to  $1.6351$

months, the maximum time housing price remained ahead of AQI was  $1.6351$  months, and the maximum time it lagged behind was  $1.5347$  months.

The calculated results of grey condense spectrum are shown in Figure 10 and Table 1. The highest grey condense spectrum values were  $2.86$  months,  $10$  months, and  $3.33$  months and the amplitude of these were  $0.0403$ ,  $-0.0201$ , and  $0.0182$ . These results show that AQI and housing price fluctuated closely across a 2- to 10-month period.

#### 4. Conclusions and Future Work

The main periods of AQI and housing price are extracted, and they range from 2 to 20 months. The amplitudes of AQI and housing price are both measured for the first period. This shows that AQI and housing price changed the most during the first period. Under the control of government policies, housing prices are gradually stabilizing and the air quality tends to improve. The negative correlation between AQI and housing price is obvious. Due to the strict implementation of environmental protection policy, housing construction cycles have lengthened. Because of environmental problems, work stoppages are common. In addition to the abovementioned reasons, there are further reasons for air pollution. First, geographic location determines the degree of air circulation. The smog in cities on the plains does not last long. Cities like Handan are located in a mountainous area where the haze does not easily diffuse. Second, building density affects air circulation. The area of urbanization continues to expand, and more and more people have come to live in downtown areas, which further worsen air pollution. Third, a large population and limited land incentivizes the construction of taller buildings, which also affects the timely circulation of air.

The empirical results given here provide a reference that the government may use to prevent haze. The relationship between AQI and housing price in Handan is only analyzed using grey spectrum theory. Whether there is a consistent connection in other areas has not been verified and needs further research.

#### Data Availability

All data used to support the findings of this study are included within the article.

#### Conflicts of Interest

All authors declare that they have no conflicts of interest.

#### Acknowledgments

The relevant research studies in this paper are supported by the National Natural Science Foundation of China (nos. 71871084 and 71401051), the Excellent Young Scientist Foundation of Hebei Education Department (no. SLRC2019001), and the project of high-level talent in Hebei Province.

## References

- [1] H. M. Long and W. Guo, "Analysis on the relationship between housing price and land price in China by VAR model," *Mathematics in Economics*, vol. 26, no. 2, pp. 52–58, 2009.
- [2] C. Choi, H. Jung, and L. Su, "Population structure and housing prices: evidence from Chinese provincial panel data," *Emerging Markets Finance and Trade*, vol. 55, no. 1, pp. 29–38, 2019.
- [3] C. Jin and M. J. Choi, "The causal structure of land finance, commercial housing, and social housing in China," *International Journal of Urban Sciences*, vol. 23, no. 6, pp. 286–299, 2019.
- [4] Z. Du and L. Zhang, "Home-purchase restriction, property tax and housing price in China: a counterfactual analysis," *Journal of Econometrics*, vol. 188, no. 2, pp. 558–568, 2015.
- [5] R. Liu and L. Liu, "Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm," *Soft Computing*, 2019.
- [6] R. Samson, R. D. Grote, C. Calfapietra et al., *Urban Trees and Their Relation to Air Pollution*, Springer International Publish AG, Basel, Switzerland, 2017.
- [7] M. Mao, H. Liu, and X. U. Honghui, "The key factor research of haze with the combined application of the multi element data," *Acta Scientiae Circumstantiae*, vol. 33, no. 3, pp. 806–813, 2013.
- [8] X. Li, W. Zheng, L. Yin, Z. Yin, L. Song, and X. Tian, "Influence of social-economic activities on air pollutants in Beijing, China," *Open Geosciences*, vol. 9, no. 1, pp. 314–321, 2017.
- [9] M. Azimi, F. Feng, and C. Zhou, "Air pollution inequality and health inequality in China: an empirical study," *Environmental Science and Pollution Research*, vol. 26, no. 12, pp. 11962–11974, 2019.
- [10] J. He, S. Gong, Y. Yu et al., "Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities," *Environmental Pollution*, vol. 223, pp. 484–496, 2017.
- [11] C. Sun, J. Zhang, Y. Guo et al., "Outdoor air pollution in relation to sick building syndrome (SBS) symptoms among residents in Shanghai, China," *Energy and Buildings*, vol. 174, pp. 68–76, 2018.
- [12] L. Wu, S. Liu, L. Yao, and S. Yan, "The effect of sample size on the grey system model," *Applied Mathematical Modelling*, vol. 37, no. 9, pp. 6577–6583, 2013.
- [13] T.-Y. Pai, K. Hanaki, and R.-J. Chiou, "Forecasting hourly roadside particulate matter in Taipei county of Taiwan based on first-order and one-variable grey model," *CLEAN—Soil, Air, Water*, vol. 41, no. 8, pp. 737–742, 2013.
- [14] L. Wu, N. Li, and Y. Yang, "Prediction of air quality indicators for the Beijing-Tianjin-Hebei region," *Journal of Cleaner Production*, vol. 196, pp. 682–687, 2018.
- [15] B. Zhu, L. Yuan, and S. Ye, "Examining the multi-timescales of European carbon market with grey relational analysis and empirical mode decomposition," *Physica A: Statistical Mechanics and Its Applications*, vol. 517, pp. 392–399, 2019.
- [16] S. Liu, Y. Yang, and L. Wu, *Grey System Theory and Its Application*, Science Press, Beijing, China, 2014.
- [17] Q. Huang and X. Zhao, *Theory and Method of River Runoff Time Series Analysis and Prediction*, Yellow River Water Conservancy Press, Zhengzhou, China, 2008.
- [18] J. Wang, "Time series periodic analysis based on energy association degree," *System Engineering Theory Practice*, vol. 9, pp. 83–85, 1998.
- [19] China Air Quality Online Monitoring Platform, <https://www.aqistudy.cn/historydata/daydata.php?city=%E9%82%AF%E9%83%B8&month=201>.
- [20] Anjuke, <https://www.anjuke.com/fangjia/>.

## Research Article

# Developing Statistical Optimization Models for Urban Competitiveness Index: Under the Boundaries of Econophysics Approach

Cem Çağrı Dönmez<sup>1</sup> and Abdulkadir Atalan<sup>1,2</sup>

<sup>1</sup>Department of Industrial Engineering, Marmara University, Istanbul, Turkey

<sup>2</sup>Department of Mechanical Engineering, Bayburt University, Bayburt, Turkey

Correspondence should be addressed to Cem Çağrı Dönmez; [cem.donmez@marmara.edu.tr](mailto:cem.donmez@marmara.edu.tr)

Received 30 August 2019; Accepted 8 October 2019; Published 20 November 2019

Guest Editor: Marco Locurcio

Copyright © 2019 Cem Çağrı Dönmez and Abdulkadir Atalan. 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.

The purpose of this research was to establish the urban competitiveness index (UCI) by using the statistical optimization method for the econophysics approach. With this technique, economic data regarding urban areas and the factors affecting UCI have been determined. The research covers 30 urban centres located in 15 countries worldwide. Urban centres with the gross domestic product per capita of \$10,000 or more were taken into consideration. The significant levels of the factors were determined with the statistical optimization method, and optimum values were calculated with the developed optimization models. Re-index values were calculated and compared with the results of PricewaterhouseCoopers and World Economic Forum. According to the results, the high UCI value of these locations depends not only on economic data but also on high values of social factors. Thus, those locations are becoming the centre of attraction for investments and capital with increasing competitiveness.

## 1. Introduction

During the last forty years, new developments in the dynamics of the urban system in physics, the main incentive has come from mathematicians and physicists mainly adaptive cognitive systems in the economical perspective who started applying their models of emerging properties to engineering then to social sciences. which would give a global framework for explaining systems dynamics in a wide range of fields of knowledge [1]. Economic development research has been moving to data-driven approaches within the methodology of natural science, statistical physics, and complexity sciences [2–4], which makes it possible to introduce new metrics that surpass the traditional economic measures in revealing current economic status and predicting future economic growth, with applications to economic development [5, 6], trading behaviour [7], poverty [8, 9], inequality [10, 11], unemployment [12, 13], and industrial structure [5, 14]. Economists and physicists have also introduced a variety of nonmonetary metrics to

quantitatively assess the country's economic diversity and competitiveness by measuring intangible assets of the economic system [15, 16], allowing for quantifying the economies' hidden potential for future development [17, 18] in near real time and at low cost [19].

Many economic methods have been used for the analysis of urban economy problems. One of these methods is the econophysics approach. Econophysics was first introduced in India as a word, "Economics" and "Physics," a conference on statistical physics in 1995 [20, 21]. "Econophysics" is a new discipline insufficiently rooted in economic theory and empirical observation. Another peculiarity of new complex system theory is to focus on the emergence of properties at a macrolevel resulting from the interactions between individual behaviour at a microlevel [1]. Although this term deals with the relations between economics, physics, mathematics, and finance, it basically takes into account the interaction of theories of physics with the economy [20]. This method, which is generally recommended for macroeconomics, is intended to be used for microstructures of

financial markets. Moreover, it is seen that the econophysics method is widely used by the statistical method as a result of research studies because in the characteristic structures of economics data, financial time series leads to statistical features and empirical studies [22, 23]. In this research, the econophysics method was used for statistical physics to address urban competitiveness. As a matter of fact, the data are measured locally rather than globally. Additionally, economists and physicists have applied network and statistical methods to reshape the understanding of international trade that the knowledge about exporting to a destination diffuses among related products and geographic neighbours [24]. More recent works on econophysics and complexity are summarized by review papers [20, 25, 26] and books [20, 27].

The processes of economic, social, and political change constantly affect the economic balance in the urban centres and reveal the necessity of renewing and planning them in terms of competitiveness [28]. Not only national changes, but the fact that they have centres that can govern themselves, accelerate and shape this need. The European Union (EU) is at the forefront, especially for the growing cities of developing countries to be at a competitive level. According to general belief, urban areas with less uncertainty and imbalance in the economy are more competitive than other cities [29–31]. This situation contributes to the rapid change and development of the urban. However, it is seen that the rapid changes of economic structures lead to uncertainty and imbalance in the economies of urban areas, and uncertainty and imbalance dominate in real economies [32]. Within this information, it is seen that the economic theories used are inadequate to account for sudden changes in the economy (fluctuations, sudden movements in the parts and mobility in the capital markets, etc.) [33]. The uncertainty and imbalance of the fluctuations in the economy have led to the emergence of new concepts and approaches. But, there is no specific model to reduce uncertainty and imbalance in the economy, and there are only temporarily developed models. The main reason for this is that the economic structures are complex and the change in the economy is fast. In such cases, the structures that form in the economies are transforming into stochastic models on a statistical basis [21]. In this work, we aimed to solve the problems of uncertainty and imbalance in the economy of urban areas by choosing the contemporary method of the optimization statistical physics techniques to reach the concept of econophysics.

Towards quantifying the complexity of a country's economy, the pioneering attempt was made by Hidalgo and Hausmann [16] who modelled the international trade flows as "Country-Product" networks and derived the Economic Complexity Index (ECI) by characterizing the network structure through a set of linear iterative equations, coupling the diversity of a country (the number of products exported by that country) and the ubiquity of a product (the number of countries exporting that product). The intuition behind this new branch of studies is that the cross-country income differences can be explained by differences in economic complexity, which is measured by the diversity of a country's "capabilities" [15, 16, 19]; the statistical properties of

financial markets have begun to apply physics concepts economically by attracting physicists [34]. Econophysicists have tried to use the theoretical approach of physical statistics to understand empirical findings [35]. Generally, many methods known in statistical physics have been applied to characterize time evolution in stock prices and exchange rates [34]. In the first examination, econophysics can be seen as wide physics models in the economy, that is, more quantitative approach, statistical models, concepts, and calculation methods. The macroscopic properties of intrinsic microscopic interactions cause a complex system in the economy. At this point, econophysics is associated with physical complexity (complex structures), reducing uncertainty and imbalance in the economy [36]. Immediately after those studies described above, Tacchella and research colleagues [37] developed a new statistical approach which defines a country's fitness and a product's complexity by the fixed points of a set of nonlinear iterative equations [38], where the complexity of products is bounded by the fitness of the less competitive countries exporting them. Further, Cristelli and research colleagues [18] studied the heterogeneous dynamics of economic complexity and found, in the fitness-income plane, strong explanatory power of economic development in the laminar regime and weak explanatory power in the chaotic regime. Based on this observation, they argued that regressions are inappropriate in dealing with this heterogeneous scenario of economic development and further proposed a selective predictability scheme to predict the evolution of countries. Nevertheless, these economic complexity indicators are not perfect, for example, ECI suffers from criticisms on its self-consistency, fitness depends on the dimension of the phase space of the heterogeneous dynamics of economic complexity [17, 18], and a new variant of the fitness method, called minimal extremal metric, can perform even better for a noise-free dataset [39]. Recently, Mariani and research colleagues [40] quantitatively compared the ability of ECI and fitness in ranking countries and products and further investigated a generalization of the "Fitness-Complexity" metric.

In general, the uncertainty and imbalance in the economy have been considered to be reduced by the Gaussian law [41]. This approach is not enough to reduce uncertainty and imbalance in the economy. This law is suitable for the use in economically viable microscopic models, such as the principle of randomly proportional effect. However, with the application of microdimension models for the complex structure of economies, healthy results cannot be obtained. In another study, the econometric approach was taken as quantum statistics. In this work, entropy, temperature, free energy, and the Hamiltonian concepts in probability theory are integrated into the quantum statistical method for mathematical financing [42]. As a result, the uncertainty of the phase transitions in mathematical finance has been found to depend on the temperature of the distribution, although the precision of the mathematical financial data is not known due to the uncertainty of the phase change. In order to acquire a more tangible result in this statistical analysis, optimization models with statistical physics have been used to obtain information about the competitiveness

of urban areas and their future competitiveness. Thus, it is inevitable to take the econophysics approach with physical statistics. The recent econophysical approach is that economic phenomena are related to statistical physics models [43].

Urban competitiveness structures are very difficult to delimitate. The trading networks are evolving through time. Furthermore, urban structures are open (the exchanges with their environment enable, for instance, the introduction of technical or social innovations) and are overlapping, as an articulation of interlocked networks [1]. Understanding how economies develop to prosperity and figuring out the best indicators that reveal the status of economic development are long-standing challenges in economics [13, 44], which have far-reaching implications to practical applications. Traditional macroeconomic indicators, like gross domestic product (GDP), are widely applied to reveal the status of economic development; however, calculating these economic census-based indicators is usually costly, resource consuming, and thus a long-time delay [45].

Other techniques have been used to create urban competitiveness index (UCI) besides the statistical physics method. The competitiveness index of four UK cities, Birmingham, Liverpool, Glasgow, and Belfast, was established using the Delphi method and analytic hierarchy process with the multicriteria analysis method. In order to assess the urban competitiveness with the method used, city competitiveness was recommended to be strengthened and renovated and synergies of business strategies were to be established [46]. In support of this work, the benefits of urban renewal and change have been discussed through the creation of new creative industries, corporations, and institutions in the cities [47]. Another study concerning mid-sized French cities is aimed at increasing the competitiveness of the cities with both the efficient use of metropolitan assets and the resources of the land of the cities [48]. One of the different approaches of the study of urban competitiveness has measured two factors for examining the urban competitiveness power of Malaysian cities which are the diminished cultural life aspects of the cities and the marginalized flights to the economic suburbs [49].

On the basis of the studies made for urban competitiveness, the most important factors affecting the urban economy are considered by the national economy. The reason behind this is that the current state of the global economy and capital constraints are increasing the competition between them by putting pressure on the cities of the world [50]. However, when urban competitiveness ties to the factors of the country's economy, it creates a gap in its competitiveness [51]. This is because the presence of new competitors and the variability of targets for cities are linked not only to the country's factors but also to the internal and external factors. For this reason, cities' competitiveness needs to be understood in terms of the strengths and weaknesses of the factors that affect their competitors. In order to establish the UCI, the econophysics approach has been used, considering the internal and external factors of the cities. The key contribution of this research is to compare the urban competitiveness of cities with how they are

positively affecting the social lives of the people as well as how competitive the cities are in attracting investment and capital to them.

Urban competitiveness is structurally complex and multidimensional, and it differs depending on the variety of factors used in the developed methods [52]. Generally, the factors derived from Eurostat [53] are taken into consideration in such studies. However, in the work done, some restrictions were applied to the selected cities. For example, Sáez and Perriñez [50] studied that the population of selected cities was more than 100,000. Thus, differences in the number and type of factors used in studies can correlate with the methods used. The number of factors and cities used in the study varies from 1 to 199 [54–56]. The cities were studied with their per capita income which are thought to be influenced by urban competitiveness are 10,000 dollars and more are taken into consideration. The generated UCI is compared with the cities based on an artificial index.

The main focus of this research is a comparison of how competitive big cities are in terms of locating and attracting investment. Since urban competitiveness is a complex, multidimensional issue, it is aimed to develop the UCI by developing optimization statistical method for the econophysics approach. This work contributes to the literature in the field of countries growing complexity, and the UCI was first explored by the econophysics approach in this research.

This work consists of four main parts. In Section 1, a literature study was carried out by examining the econophysics theory and UCI. Section 2 underlines the importance of the factors and responses that constitute the UCI. Development of optimization models for UCI is discussed in this section. The results of optimization models by statistical analysis for UCI and interpretations of this study are handled in Section 3. Finally, Section 4 includes conclusions and future research.

## 2. Methodology

*2.1. Urban Competitiveness Factors.* Econophysics in urban economy has been defined as a new discipline by using statistical physics methods on urban competitiveness problems. In this case, urban economy with a complex structure can be understood with a new approach, the econophysics. The data obtained in the social, economic, and cultural fields constituting the urban economy have been interpreted by the econophysics approach. The results obtained by the econophysics methodology are used to evaluate the complexity of irregular forms, and nowadays, economists are able to apply them in the field of UC up to urban development models.

Urban economy is the biggest element that constitutes the macro economy of a country. In the structure of the region where the urban economy is located, work force, productivity, education etc., are directly connected to the factors. The stability of the UCI depends not only on these factors but also on the macroeconomic structure of the country where the economy is located. Cities, which are metropolises, also bring urban competitiveness as they have large economies. The strength of the economy of

metropolitan cities depends on many elements. In particular, urban areas' commercial performance and productivity play an important role in urban competitiveness because of its contribution to the urban economy.

A number of methods have been developed for the measurement of UCI, and these methods have been used as an indicator system and weight unit and UCI in general. In the 2011-2012 Global Urban Competitiveness Report, a model of the UC output was created, focusing only on the economic parameters of cities [57]. The model is formulated accordingly, and the result is achieved. The economic size, efficiency, grade, quality, density, and effect of cities are taken into account in the form [58]. For the World Bank, there are four key factors in the evaluation report on the UC which were highlighted: economic structure, human resources, regional wealth, and factors of institutional structure. UCI was created by giving some values to these factors [59]. In order to calculate the urban competitiveness indices of the four cities of United Kingdom, investment, finance, social capital, improvement, and use/occupant's possible and physical environment factors were determined. These factors are regarded as operational components in that study. Calculated factor scores did not capture any city superiority, but some cities had the highest scores for certain factors [46].

In another research, nine different factors were considered fewer than three main headings to form the UCI. These factors are determined as transportation, health, basic education, economy activities, labor force, higher education, business sophistication, knowledge society, and information society of the cities [50]. As a result of the statistical analysis, it was understood that the outcome in the infrastructure, not in the singularities, belong to these factors. However, the inadequacies of this study are that the factors should be studied, not as a group but as an interaction and singularity. It can be concluded the factors must be studied individually and interactively before the group is formed. For example, the factor "a" that affects the competence of urban competitiveness and the factor "b" with little or no effect should be considered as singular, and they should be calculated as significant or insignificant as a result of interactions. Moreover, ignoring the elements of the economy from the factors that influence a city's competitive power triggers the inadequate results to be achieved. It is also said that a city's competitive power is inversely proportional to the economic development. However, this proposal does not find a scientific infrastructure. Basically, the fact that the economic activity factor is statistically significant among the factors cogitated means that the relationship between urban competitiveness and the urban economy is strong.

In this work, 38 indicators were mentioned under six main factors (see Table 1). The data employed in this research cover the years of 2016 and 2017 [53, 55, 60, 61]. It was designed to establish UCI by considering the factors that are expressed statistically significant in previous studies. The fact that there are different factors in one study suggests the existence a complex structure for UCI and not a linear one. Having said that it is possible to increase the numbers of these factors and indicators; the mathematical optimization

TABLE 1: Indicators of the affecting factors of UCI.

Factors	Indicators	Notations
Education, health, and training	Education system	ET
	Higher education	
	Primary and secondary level	
	Information and knowledge society	
Labor and transport	Trained persons rate	LT
	Healthcare infrastructure	
	Health policies	
	Old person employment rate	
	Young person employment rate	
Technology and industry	Female employment rate	TI
	Unemployment rate	
	Transport infrastructure	
	Shipping transport variety	
Market size	Industrial structures and standard	MS
	Technology infrastructure	
	Technological development	
	Information technology	
	Production technology	
Product efficiency	Company variety	PE
	Product variety	
	Market economy	
	Market growth rate	
	Sales cycles	
	Company earnings	
Financial service	Product variety	FS
	Product quality	
	Product supply-demand level	
	Production speed	
	Cost of products	
Financial service	Product consumption time	FS
	Easy-to-carry	
	Product type (portability)	
	Stability of financial system	
Financial service	Financial development	FS
	Bank index	
	Currency and credit	
	Investment convenience	
Financial service	Capital rate	FS

model can be obtained and applied to become more complex and stochastic. The analysis of this model will be very difficult and time consuming, resulting in having distance from optimum values. Statistical analysis is to be maintained by determining factors that are important to reduce this complexity and constitute many subunits as the target factor. Thus, in order to establish the UCI, the econophysics approach has been used by considering the internal and external factors of the cities. In this model, population, energy, and environmental factors were ignored. The study covers 30 cities located in 15 countries worldwide. In statistical analysis to be performed, for each factor, the gross domestic product per capita was used for 30 different cities over \$10,000 [60].

*2.2. Urban Competitiveness Response Variables.* Variations in the economic, environmental, and social structures are evident in determining urban competitiveness factors. In this context, the urban competitiveness levels of the analyses in the scientific literature are still being modelled. From a researcher's point of view, a city can be artificially measured according to its competitive power using different methods and theoretical models. Advantages and disadvantages of each method for urban competitiveness are stated, and they consequently contribute to finding the most reliable system.

In this probe, a number of ways have been identified in determining the factors affecting urban competitiveness and the affected responses. Researchers working on urban competitiveness assume that factors are more directly influenced by the UCI value. When computing the UCI value, they referred to the values attributed to the factors and the formation of an artificial index. The indexes formed in some research are calculated according to the Nash theory and Freudenberg method. In such academic work, the relative weights of the indicators resulting from the weighting of the factors have been calculated. In other words, weighting calculations are made according to the weight of each indicator to obtain relative weights.

In another case study, UCI values were calculated by using a closed formulation method by giving a certain range of values to the factors and comparing the competitiveness of the cities. For example, in the global competitiveness report prepared by Ni in 2014, factors were evaluated between 1 and 7. Subsequently, the urban competitiveness latent formula consisting of six different factors was obtained [57]. In the competitiveness index developed for the 24 cities of Lithuania, the steps and basic characteristics of the composite index were used to measure urban competitiveness [62]. Similarly, the urban competitiveness forces of 23 US cities were measured by statistical analysis taking into account only 3 different factors [63].

When we measure the urban competitiveness in our perspective, it is thought that the factors indirectly affect the UCI value, not directly. The reason for this is that there are some responses that these factors influence. We think that UCI values will be formed in the direction of the results obtained from these responses. There are two main responses to this study. These are the gross domestic product (UGDP) and the gross domestic product per capita (UGDP-PC) of the urban areas considered for the UCI. The reason for taking these responses into consideration is the competitiveness of the cities and the measurement of the cities by their economic growth. Some researchers have considered these responses as factors for the strength of urban competitiveness. They have even argued that the economic growth of cities and the strength of urban competitiveness are directly proportional [64, 65].

*2.3. Development of Optimization Models for UCI.* The areas of use of mathematical optimization models vary. In particular, these models are used to solve the problems that cost, and benefit dilemmas are taken into consideration.

Mathematical modelling is carried out with the aim of reducing the cost and maximizing the benefits. In this study, statistical analysis of the mathematical model was performed. There are three basic things that make up mathematical models. In a mathematical model, the objective function, constraints, and decision-variable signs are considered. There are three stages in the mathematical model created in this study. In the first step, the equations obtained as a result of statistical analysis are considered as objective functions. The next step is to obtain the data from the lower and upper limits that constitute the constraints. The fact that the index values used for the factors are greater than zero causes the decision variables to be greater than zero. In the optimization model developed for this study, the values of the decision variables will be greater than zero.

As a final step, the optimal values required for a city to have competitiveness have been calculated. A desirability function was developed to compute UCI scores of cities for the comparative outcome of the previous urban competitiveness index. Through these functions, comparisons of competitiveness of cities were made.

Attention should be paid to the desirability value when achieving optimum results. Before constructing the optimization models, it is necessary to consider the function of desirability according to the results to be obtained as a result of statistical analysis. The factors affecting the response function directly affect the desirability function [66]. In short, it is desirable that the factors affecting the main response values are at the desired values. This is measured by the value of desirability. The best result is obtained as this value goes from zero to one when calculating desirability.

To find  $n$  factors' values, the function of each response value is expressed as

$$y_i = f_i(x_1, x_2, \dots, x_k), \quad i = 1, 2, \dots, n. \quad (1)$$

The desirability function is specified as

$$d_i = d_i(y_i) = d_i(y_i(x)), \quad (2)$$

where  $d(y)$  takes a value between 0 and 1. If  $d_i(y_i) = 1$ , this is the desired best value [67, 68], but if  $d_i(y_i) = 0$ , it is the worst and undesirable value.

General desirability is expressed as follows:

$$d_i(y_i) = \left( \frac{y_i - l_i}{u_i - l_i} \right)^{w_i}, \quad l_i \leq y_i \leq u_i, \quad (3)$$

where  $l_i$  and  $u_i$  are the lower and upper specification limit of the responses, the power  $w_i$  corresponds to the weighted factor, and  $w_i$  is the parameter that determines the shape of  $d_i(y_i)$ . There are three purposes for the value of desirability. These are maximum, minimum, and target values in the objective function. In this study, formula (4) was considered in order to achieve maximum values of the objective functions. There are three different conditions for each of these three situations.

For maximization problems, the desirability functions are

$$d^{\max} = \begin{cases} 0, & \text{if } y_i \leq l, \\ \left( \frac{y_i - l}{u - l} \right)^{w_i}, & \text{if } y_i \leq u, y_i \geq l, \\ 1, & \text{if } y_i \geq u. \end{cases} \quad (4)$$

For minimization problems, the desirability function is

$$d^{\min} = \begin{cases} 0, & \text{if } y_i \geq u, \\ \left( \frac{u - y_i}{u - l} \right)^{w_i}, & \text{if } y_i \leq u, y_i \geq l, \\ 1, & \text{if } y_i \leq l. \end{cases} \quad (5)$$

For target, the desirability function is

$$d^{\text{target}} = \begin{cases} \left( \frac{y_i - l}{u - l} \right)^{w_i}, & \text{if } y_i \leq u, y_i \geq l, \\ \left( \frac{u - y_i}{u - l} \right)^{w_i}, & \text{if } y_i \leq u, y_i \geq l, \\ 0, & \text{if } y_i \leq l \text{ and } y_i \geq u. \end{cases} \quad (6)$$

At the same time, for these three cases, nominal the best (NTB) for the target, smaller the better (STB) for the minimum, and larger the better (LTB) for the maximum conditions are defined in previous research studies [69]. The overall desirability function equation to be obtained by considering these three conditions is calculated by the geometric mean (see equation (7)). The use of the geometric mean is due to the fact that more than one dimensionless individual desirability scales arise. The individual desirability scales are merely a whole using a geometric mean and merges them into one desirability.

$$d = (d_1 * d_2 * d_3 * \dots * d_n)^{1/n}. \quad (7)$$

The overall desirability function includes the upper and lower bound values of the factors that have an effect on the response. UCI was created separately for each factor. UCI formula was obtained by geometric mean of these factors.

The parameters analyzed as a result of statistical analysis were global-based so as to construct the UCI:

$$\text{UCI} = \frac{[(c_1 + \sum_{i=1}^n p_{i,k} r_k) \text{UGDP}_i]}{c_2}, \quad (8)$$

and

$$\text{UCI} = \frac{[(c_1 + \sum_{i=1}^n p_{i,k} r_k) (\text{GDP PC})_i]}{c_2}, \quad (9)$$

where UCI is the urban competitiveness index,  $c_1$  is the regression constant,  $c_2$  is the normalization constant,  $p_i$  and  $k$  are competitiveness parameters,  $r_k$  is the regression multiplier,  $\text{UGDP}_i$  is the gross domestic product of urban, and  $(\text{UDGP PC})_i$  is the gross domestic product per capita of urban.

In this investigation, optimal UCI for each urban area needs to be established in order for cities to have competitive power. Contemplating two objective functions, it was aimed to maximize the urban competitiveness of cities. Both optimization models contain the same constraints. For the calculation of GDP belonging to the urban area (closed formula),

$$\begin{aligned} & \text{maximize} \quad \text{UGDP}_i \\ & \text{subject to} \\ & l \leq x_j, \\ & x_j \leq u, \\ & 0 \leq x_j. \end{aligned} \quad (10)$$

Optimization equation of the UGDP-PC (closed formula):

$$\begin{aligned} & \text{maximize} \quad \text{UGDP PC}_i \\ & \text{subject to} \\ & l \leq x_j, \\ & x_j \leq u, \\ & 0 \leq x_j. \end{aligned} \quad (11)$$

With this objective function, the two response variables are reduced to a single form. Thus, whichever response is maximum is the optimum value for UCI. The optimization model that maximizes the urban competitiveness index is formulated as follows:

$$\begin{aligned} & \text{maximize}_{\text{UCI}} \left\{ \sum_{y_i}^{y_f} \left[ \frac{[(c_1 + \sum_{i=1}^n p_{i,k} r_k) (\text{UDGP}_i)] / (c_2)}{y_f - y_i} \right], \sum_{y_i}^{y_f} \left[ \frac{[(c_1 + \sum_{i=1}^n p_{i,k} r_k) (\text{UDGP PC}_i)] / (c_2)}{y_f - y_i} \right] \right\}, \\ & \text{subject to} \left[ \frac{(\sum_{y_i}^{y_f} \min\{l_{ij}^y, \dots, l_{ij}^{y_f}\})}{(y_f - y_i)} \right] \leq x_{ij} \left[ \frac{(\sum_{y_i}^{y_f} \max\{u_{ij}^y, \dots, u_{ij}^{y_f}\})}{(y_f - y_i)} \right] \geq x_{ij}, \quad 0 \leq x_{ij}, \end{aligned} \quad (12)$$



where the lower limit score “ $l_{ij}$ ” and the upper limit score “ $u_{ij}$ ” notations are used.  $x_{ij}$  is the competitiveness parameter type ( $j$ : *ET, L, TMS, G, F*),  $y_i$  (the first year of data use) and  $y_f$  (the last year of data use) are symbolizations that indicate which data match the function.

### 3. Results of Optimization Models by Statistical Analysis for UCI

The calculation of the UCI is based on the data of the GDP and GDPPC of the countries they are affiliated with. This is because taking the factors that affect the power of UCI on a country basis has a more accurate result. The results of statistical analysis for UCI indicate that the determination of factors and responses is consistent in the general sense. Some changes and arrangements have been made on the writing of the article. The Minitab statistical analysis program was used for analysis of variance test of data in this research.

According to Table 2, it is understood that all the factors are effective on the UGDP and these factors are meaningfully significant. In addition, the significance values of the factors are considered to be effective on UGDP-PC. From these factors, only the effect of the ET and FS is less pronounced.

Firstly, Figure 1 shows how the UGDP-PC is affected by the factors. It is stated that the education factor is a constant effect and that the increase of product productivity especially affects the negative direction. The line consisting of grey points shows the average values of UGDP-PC and UGDP in Figures 1 and 2.

Figure 2 shows how the UGDP is affected by the parameters. As the scores of education, employment, and financial factors increase after a certain point, the UGDP tends to be negative. It is observed that the level of economy, the size of the market, and the productivity of the product increased the level of economy of the urban area.

In order to have the competitiveness power of the cities, the factors must have optimum values as in Figure 3. In this figure, which gives the optimum values, it is assumed that the parabolic impression of the curves belonging to the factors has more than one objective function. In such cases, improved optimization models are also defined as multi-objective optimization models. The reason for the parabolic view is that if one factor takes the maximum value for one objective function, the other has the minimum value for the objective function. In this case, it is difficult for the desirability value to be achieved at the optimum level. In addition, the optimization models established in this study show that the curves are stochastic as being parabolic. The reason for being stochastic is that the next step is unknown (steady state). A stochastic model was observed on the graph, and the parabola was observed.

According to the optimization models, the best results for the UGDP and UGDP-PC are given in Table 3. The grey area shown in Figure 3 varies with the values taken by the factors. It shows the values that can be taken in the feasible set or area of objective functions of decreasing or increasing this field. The grey regions represent the settings in which the corresponding response variable has a low value or even zero desirability. More feasible space (white region) was formed for the UGDP value, while the grey area in GDPPC was higher. The upper

TABLE 2: The analysis of variance test for UGDP and UGDP-PC.

Factors	Sample size	UGDP		UGDP-PC	
		Prob.	Status	Prob.	Status
ET	30.0	0.011*	Sig. ( $p < 0.05$ )	0.995	Not sig.
L	30.0	0.003**	Sig. ( $p < 0.01$ )	0.007**	Sig. ( $p < 0.01$ )
T	29.0	0.001**	Sig. ( $p < 0.01$ )	0.000**	Sig. ( $p < 0.01$ )
MS	30.0	0.049*	Sig. ( $p < 0.05$ )	0.015*	Sig. ( $p < 0.05$ )
G	30.0	0.010*	Sig. ( $p < 0.05$ )	0.002**	Sig. ( $p < 0.01$ )
F	28.0	0.000**	Sig. ( $p < 0.01$ )	0.397	Not sig.

Note. \*\* $p < 0.01$  and \* $p$  value  $< 0.05$  are significant factors, respectively; otherwise insignificant.

and lower limits of the grey areas depend on the lower and upper limits of the constraints. In the direction of optimum results, the UCI power will increase only if the urban average UGDP-PC is less than \$50,523.50 and the size of the UGDP is over \$419 million. These results reveal an urban area needs to approach these values to compete with other urban areas in economic and social areas.

Considering the optimum values of the data, new competitiveness indexes of the cities are established as shown in Table 4. The components of urban structure with the econophysics approach index (EI) are clearly observed, applied in the emerging UCI research which affects the UCI, especially the human factor. The indexes calculated by statistical physics approach of econophysics for each urban were compared with 3 different sources. PwC indexes considered the economic structures of cities, while UN-Habitat data have considered both the economic structure and sustainable development of cities. In the study conducted for UN-Habitat, indexes have been formed in two different areas considering the economic dimensions and sustainable developments of cities. In addition to the economic dimensions of cities, biodiversity, urban mobility, technological structures, and urban planning methods which constitute the substructure of cities are considered to create indexes for cities. In our study, an index for each city was created by combining both fields in one area.

According to the comparison made with UCI, it has been determined that the method we have developed has a limited variation in some cities. In cities where changes are excessive, some factors are due to low index values belonging to developing countries. In general, the best results within the scope of UCI are spotted in cities, for example, in Europe, Far East Countries, and the United States, in terms of attractiveness for investment. As expected, top seats are occupied by “global cities,” Tokyo, Madrid, Paris, Osaka, London, Seoul, and New York. These cities are major economic centres and also have significant economic leadership positions around the world. Tokyo, New York, Los Angeles, London, Osaka, Paris, Washington DC, Seoul, Madrid, and Philadelphia are among the top 10 cities. This

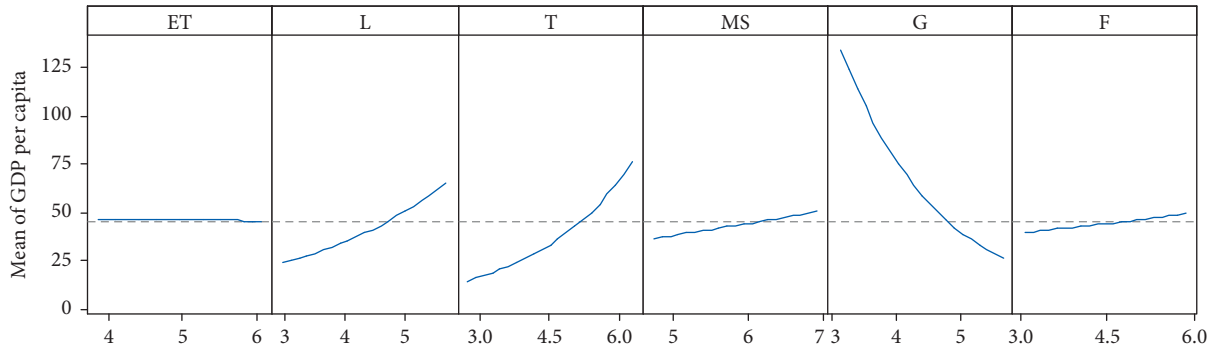


FIGURE 1: The main effects of factors for UGDP-PC.

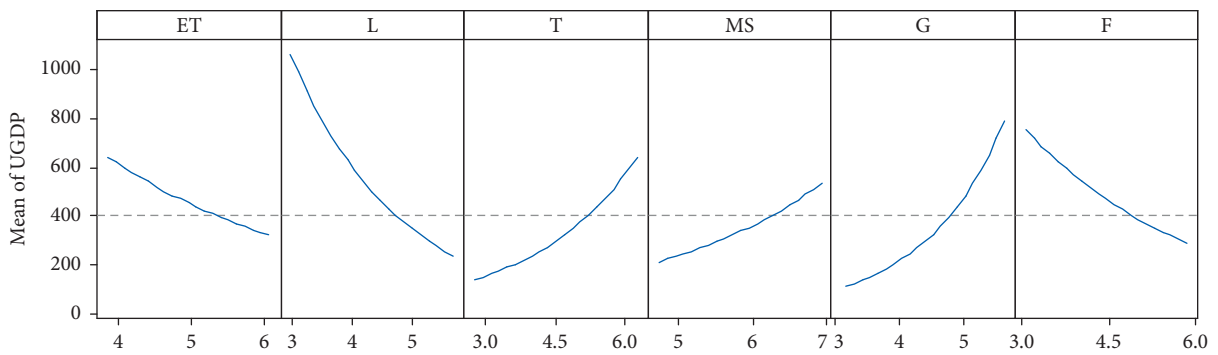


FIGURE 2: The main effects of factors for UGDP.

		ET	L	T	MS	G	F
Optimal	High	6.090	5.690	6.280	6.940	5.640	5.840
D: 0,9994	Cur	[4.3814]	[5.5912]	[6.280]	[4.710]	[5.640]	[5.840]
Predict	Low	3.860	2.970	2.750	4.710	3.140	3.040

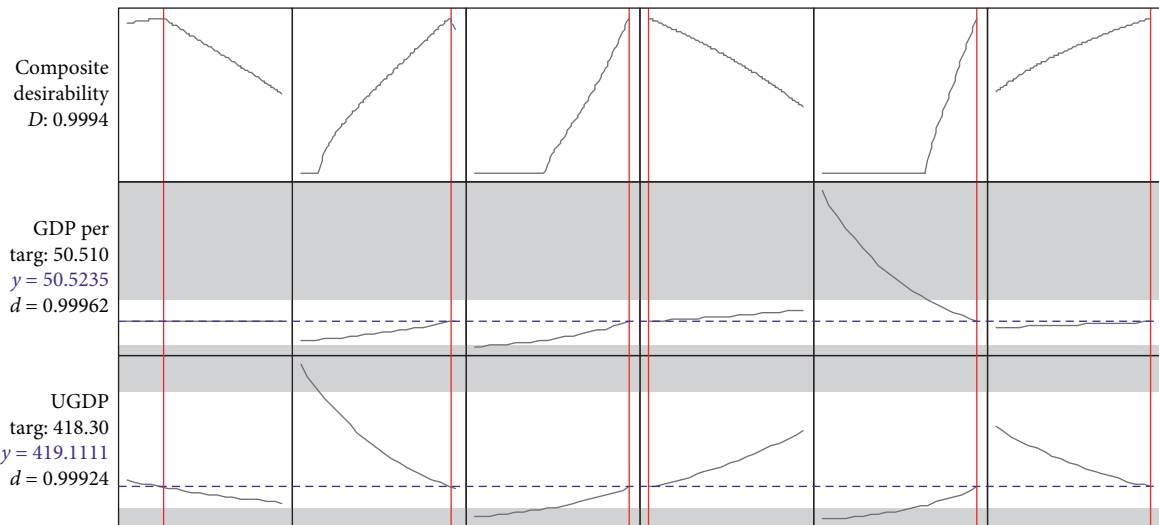


FIGURE 3: The optimum values of both response variables.

indicates the potential for being economically strong, attractive, and competitive cities for international investment. In terms of having the index of competitiveness power, the

city of Buenos Aires has the lowest competitiveness, while the city with the highest competitiveness power was defined as Paris.

TABLE 3: Optimum values of factors and response variables.

ET	L	T	MS	G	F	UGDP-PC	UGDP
4.381	5.591	6.28	4.71	5.64	5.84	50.235	419.111

TABLE 4: Comparisons of ranking of UCI.

Urban	UCI	PwC	*UN-Habitat	**UN-Habitat	EI
Tokyo	7.8353	1	7	3	3
New York	6.5580	2	1	1	7
Los Angeles	6.5569	3	2	14	9
Chicago	6.5565	4	15	9	14
London	7.0359	5	4	2	5
Paris	9.3289	6	18	8	1
Osaka	7.8352	7	9	—	4
Mexico City	3.4330	8	29	26	27
Philadelphia	6.5568	9	19	17	10
São Paulo	3.4424	10	6	—	26
Washington DC	6.5570	11	27	10	8
Boston	6.5564	12	16	4	16
Buenos Aires	0.9988	13	30	27	30
Dallas	6.5567	14	8	21	12
Moscow	4.6313	15	25	24	24
Hong Kong	5.0950	16	11	11	22
Atlanta	6.5566	17	23	16	13
San Francisco	6.5565	18	5	12	14
Houston	6.5560	19	10	7	19
Miami	6.5562	20	14	23	18
Seoul	6.8346	21	12	6	6
Toronto	6.5271	22	22	15	20
Detroit	6.5568	23	24	25	10
Seattle	6.5563	24	20	13	17
Shanghai	5.0944	25	13	18	23
Madrid	8.3354	26	28	22	2
Singapore	3.5065	27	3	5	25
Sydney	6.2848	28	26	19	21
Mumbai	1.9699	29	17	—	29
Istanbul	3.0736	30	21	20	28

Note. The cities considered for UN-Habitat were selected among the top 200 cities. The ranking indexes of the cities that are not on the list are written to the next city in the list of Table 4. PwC: PricewaterhouseCoopers [70]. \*UN-Habitat: the United Nations Human Settlements Programme, Global Urban Economic Competitiveness [71]. \*\*UN-Habitat: the United Nations Human Settlements Programme, Global Urban Sustainable Competitiveness [71]. EI: econophysics index.

#### 4. Conclusions and Future Research

The statistical physics of the econophysics method was shown as an important approach for UCI in this study. This is the first time that the power of UCI has been linked to the social and economic factors with the application of statistical physics of the econophysics method. In order to establish the UCI, human, industry, and technology factors have to be taken into consideration in the economic parameters of the urban area. In this research, the level of education of people, urban technology, size of urban market, and urban economic development have been statistically evaluated. The reason for dealing with factors other than economic factors

is that they are uniquely positioned to provide dynamic environments in which the cities operate. In addition, the vast urban areas of the world are sources of basic knowledge and innovation. This makes them central to the globalizing world economy. However, this globalizing context means increasing competition between cities in order to secure the limited resources available.

The fact that the factors that are considered to influence UCI are in large numbers triggers the artificial level of the methods to be developed. The main reason for the display of the city's competitive talents is that it makes it attractive for investment and capital. Most of the work done for urban regeneration is at the regional level, resulting in the weakening of the competitiveness of those cities as a result of global changes. In this work, competitive indexes of big cities in the world were created. The limitation of the factors taken into consideration is based on the previous studies. However, among these factors, the UCI was formed by ignoring the political structures of the countries, and therefore regulations and laws. But, it should be noted that the competitiveness of the countries is the effect of social and political structures.

There are differences in the results when we take into consideration the data of the UCI obtained by the new method (econophysical approach method) that we have chosen in this study according to the report of the World Economic Forum in 2014 and PricewaterhouseCoopers (PwC) under the UC report. It is evident that the uncertainty and imbalance in the economies are less in the cities where UCI is high. At the same time, it is foreseen that with this approach, the high index of UC will be strong in future economies. At the same time, with the use of this approach, it is predicted that the future economies of the cities will be strong by having a high UC index.

Nevertheless, the fact that the UCI is low can be said to be a negative one in terms of the economy of the urban areas. As a result, social and economic factors are linked to UCI; tangible results have been obtained with the numerical data and the methods used for this research for the first time.

In the later stages of this research, an interactive statistical analysis of parametric and nonparametric factors will be performed by using the Box–Behnken technique, which is among the design of experiment methods by determining the different levels of the factors involved in this work. The reason for using this method is to show how the optimization models to be achieved are not linear and that the stochasticity and the factors influence the responses in the superficial graphs. In addition, it is aimed to enrich the optimum results obtained with the optimization mathematical models to be created with the scenarios.

#### Data Availability

The numerical data used to support the findings of this study have been deposited in the Global Urban Competitiveness Report 2011-2012. Also, we used World Bank urban development data sources. The numerical data used to support the findings of this study were supplied by Marmara University Scientific Research Institute under project no: FEN-

C-DRP110618-0344 and so cannot be made freely available. Requests for access to these data should be made to Mr. Abdulkadir Atalan, Bayburt University, Engineering Faculty, Department of Mechanical Engineering, Turkey, aatalan@bayburt.edu.tr.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] A. Bretagnolle, E. Daudé, and D. Pumain, "From theory to modelling: urban systems as complex systems," *Cybergeo: Revue Européenne de Géographie/European Journal of Geography*, 2006.
- [2] L. Einav and J. Levin, "Economics in the age of big data," *Science*, vol. 346, no. 6210, Article ID 1243089, 2014.
- [3] D. S. Hamermesh, "Six decades of top economics publishing: who and how?," *Journal of Economic Literature*, vol. 51, no. 1, pp. 162–172, 2013.
- [4] C. A. Hidalgo, "Disconnected, fragmented, or united? a transdisciplinary review of network science," *Applied Network Science*, vol. 1, no. 1, p. 6, 2016.
- [5] J. Gao, B. Jun, A. Pentland, T. Zhou, and C. A. Hidalgo, "Collective learning in China's regional economic development," 2017, <http://arxiv.org/abs/1709.05392>.
- [6] C. A. Hidalgo and R. Hausmann, "A network view of economic development," *Developing Alternatives*, vol. 12, no. 1, pp. 5–10, 2008.
- [7] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using Google Trends," *Scientific Reports*, vol. 3, no. 1, p. 1684, 2013.
- [8] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.
- [9] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.
- [10] D. Hartmann, M. R. Guevara, C. Jara-Figueroa, M. Arístarán, and C. A. Hidalgo, "Linking economic complexity, institutions, and income inequality," *World Development*, vol. 93, pp. 75–93, 2017.
- [11] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The collaborative image of the city: mapping the inequality of urban perception," *PLoS One*, vol. 8, no. 7, Article ID e68400, 2013.
- [12] A. Llorente, M. Garcia-Herranz, M. Cebrian, and E. Moro, "Social media fingerprints of unemployment," *PLoS One*, vol. 10, no. 5, Article ID e0128692, 2015.
- [13] J. Yuan, Q.-M. Zhang, J. Gao et al., "Promotion and resignation in employee networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 444, pp. 442–447, 2016.
- [14] C. A. Hidalgo, B. Klinger, A.-L. Barabasi, and R. Hausmann, "The product space conditions the development of nations," *Science*, vol. 317, no. 5837, pp. 482–487, 2007.
- [15] R. Hausmann, C. A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, and M. A. Yildirim, *The Atlas of Economic Complexity: Mapping Paths to Prosperity*, MIT Press, Cambridge, MA, USA, 2014.
- [16] C. A. Hidalgo and R. Hausmann, "The building blocks of economic complexity," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10570–10575, 2009.
- [17] M. Cristelli, A. Gabrielli, A. Tacchella, G. Caldarelli, and L. Pietronero, "Measuring the intangibles: a metrics for the economic complexity of countries and products," *PLoS One*, vol. 8, no. 8, Article ID e70726, 2013.
- [18] M. Cristelli, A. Tacchella, and L. Pietronero, "The heterogeneous dynamics of economic complexity," *PLoS One*, vol. 10, no. 2, Article ID e0117174, 2015.
- [19] J. Gao and T. Zhou, "Quantifying China's regional economic complexity," *Physica A: Statistical Mechanics and Its Applications*, vol. 492, pp. 1591–1603, 2018.
- [20] A. Chakraborti, I. M. Toke, M. Patriarca, and F. Abergel, "Econophysics review: I. Empirical facts," *Quantitative Finance*, vol. 11, no. 7, pp. 991–1012, 2011.
- [21] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge, UK, 2016.
- [22] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, John Wiley & Sons, Hoboken, NJ, USA, 2015.
- [23] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Routledge, Abingdon, UK, 2018.
- [24] B. Jun, A. Alshamsi, J. Gao, and C. A. Hidalgo, "Relatedness, knowledge diffusion, and the evolution of bilateral trade," 2017, <http://arxiv.org/abs/1709.05392>.
- [25] J.-P. Huang, "Experimental econophysics: complexity, self-organization, and emergent properties," *Physics Reports*, vol. 564, pp. 1–7, 2015.
- [26] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, "Econophysics: financial time series from a statistical physics point of view," *Physica A: Statistical Mechanics and Its Applications*, vol. 279, no. 1–4, pp. 443–456, 2000.
- [27] S. Sinha, A. Chatterjee, A. Chakraborti, and B. K. Chakrabarti, *Econophysics: An Introduction*, John Wiley & Sons, Hoboken, NJ, USA, 2010.
- [28] N. Smith, "New globalism, new urbanism: gentrification as global urban strategy," *Antipode*, vol. 34, no. 3, pp. 427–450, 2002.
- [29] M. Cimoli and J. Katz, "Structural reforms, technological gaps and economic development: a Latin American perspective," *Industrial and Corporate Change*, vol. 12, no. 2, pp. 387–411, 2003.
- [30] P. R. Dickson, "Toward a general theory of competitive rationality," *Journal of Marketing*, vol. 56, no. 1, pp. 69–83, 1992.
- [31] S. Szymanski, "The economic design of sporting contests," *Journal of Economic Literature*, vol. 41, no. 4, pp. 1137–1187, 2003.
- [32] L. Nowzohour and L. Stracca, *More Than a Feeling: Confidence, Uncertainty and Macroeconomic Fluctuations*, European Central Bank, Frankfurt, Germany, 2017.
- [33] S. Adzic and O. Sedlak, "Economic modelling and theory of fuzzy sets application in macroeconomic planning within the process of transition," *Yugoslav Journal of Operations Research*, vol. 8, no. 2, pp. 331–342, 1998.
- [34] H. E. Stanley, L. A. N. Amaral, D. Canning, P. Gopikrishnan, Y. Lee, and Y. Liu, "Econophysics: can physicists contribute to the science of economics?," *Physica A: Statistical Mechanics and Its Applications*, vol. 269, no. 1, pp. 156–169, 1999.
- [35] M. Gallegati, S. Keen, T. Lux, and P. Ormerod, "Worrying trends in econophysics," *Physica A: Statistical Mechanics and Its Applications*, vol. 370, no. 1, pp. 1–6, 2006.
- [36] M. Kravchenko, "Structural balance as a basis of the economic sustainability of an enterprise," *World Scientific News*, vol. 57, no. 6, pp. 300–308, 2016.

- [37] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero, "A new metrics for countries' fitness and products' complexity," *Scientific Reports*, vol. 2, p. 723, 2012.
- [38] G. Caldarelli, M. Cristelli, A. Gabrielli, L. Pietronero, A. Scala, and A. Tacchella, "A network analysis of countries' export flows: firm grounds for the building blocks of the economy," *PLoS One*, vol. 7, no. 10, Article ID e47278, 2012.
- [39] Q. Mao, K. Zhang, W. Yan, and C. Cheng, "Forecasting the incidence of tuberculosis in China using the seasonal autoregressive integrated moving average (SARIMA) model," *Journal of Infection and Public Health*, vol. 11, no. 5, pp. 707–712, 2018.
- [40] M. S. Mariani, A. Vidmer, M. Medo, and Y.-C. Zhang, "Measuring economic complexity of countries and products: which metric to use?," *The European Physical Journal B*, vol. 88, no. 11, p. 293, 2015.
- [41] B. Mandelbrot, "The variation of some other speculative prices," *The Journal of Business*, vol. 40, no. 4, pp. 393–413, 1967.
- [42] V. P. Maslov, "Econophysics and quantum statistics," *Mathematical Notes*, vol. 72, no. 5-6, pp. 811–818, 2002.
- [43] C. Schinckus, "Economic uncertainty and econophysics," *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 20, pp. 4415–4423, 2009.
- [44] N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," *Science*, vol. 328, no. 5981, pp. 1029–1031, 2010.
- [45] J.-H. Liu, J. Wang, J. Shao, and T. Zhou, "Online social activity reflects economic status," *Physica A: Statistical Mechanics and Its Applications*, vol. 457, pp. 581–589, 2016.
- [46] S. Singhal, S. McGreal, and J. Berry, "An evaluative model for city competitiveness: application to UK cities," *Land Use Policy*, vol. 30, no. 1, pp. 214–222, 2013.
- [47] T. A. Hutton, "Trajectories of the new economy: regeneration and dislocation in the inner city," *Urban Studies*, vol. 46, no. 5-6, pp. 987–1001, 2009.
- [48] S. Puissant and C. Lacour, "Mid-sized French cities and their niche competitiveness," *Cities*, vol. 28, no. 5, pp. 433–443, 2011.
- [49] J. Abdullah, "City competitiveness and urban sprawl: their implications to socio-economic and cultural life in Malaysian cities," *Procedia-Social and Behavioral Sciences*, vol. 50, pp. 20–29, 2012.
- [50] L. Sáez and I. Periañez, "Benchmarking urban competitiveness in Europe to attract investment," *Cities*, vol. 48, pp. 76–85, 2015.
- [51] M. Lu and Z. Chen, "Urbanization, urban-biased policies, and urban-rural inequality in China, 1987–2001," *The Chinese Economy*, vol. 39, no. 3, pp. 42–63, 2006.
- [52] L. Salvati and P. Serra, "Estimating rapidity of change in complex urban systems: a multidimensional, local-scale Approach," *Geographical Analysis*, vol. 48, no. 2, pp. 132–156, 2016.
- [53] P. Annoni, L. Dijkstra, and N. Gargano, *EU Regional Competitiveness Index 2016*, Italy, 2017.
- [54] R. Huggins and N. Clifton, "Competitiveness, creativity, and place-based development," *Environment and Planning A: Economy and Space*, vol. 43, no. 6, pp. 1341–1362, 2011.
- [55] N. Pengfei and H. Qinghu, "Comparative research on the urban competitiveness," *Chinese Academy for Social Science*, 2017.
- [56] J. Shen, "Cross-border urban governance in Hong Kong: the role of state in a globalizing city-region," *Professional Geographer*, vol. 56, no. 4, pp. 530–543, 2004.
- [57] N. Pengfei and P. K. Kresl, *Global Urban Competitiveness Report (2011-2012)*, Edward Elgar Publishing, Cheltenham, UK, 2014.
- [58] N. Pengfei and H. Qinghu, *Comparative Research on the Global Urban Competitiveness*, 2018.
- [59] E. D. Viorica, *Urban Competitiveness Assessment in Developing Country Urban Regions: The Road Forward*, The World Bank, Washington, DC, USA, 2000.
- [60] Klaus Schwab, "The global competitiveness report 2018," World Economic Forum Reports 2018, <http://www3.weforum.org/docs/GCR2018/05FullReport/TheGlobalCompetitivenessReport2018.pdf>.
- [61] N. Pengfei, M. Kamiya, R. Ding, M. Kamiya, and R. Ding, "Global urban competitiveness: comparative analysis from different perspectives," in *Cities Network along the Silk Road*, pp. 51–64, Springer, Singapore, 2017a.
- [62] J. Bruneckiene, A. Guzavicius, and R. Cincikaite, "Measurement of urban competitiveness in Lithuania," *Engineering Economics*, vol. 21, no. 5, pp. 493–508, 2010.
- [63] P. Kresl and B. Singh, "Urban competitiveness and US metropolitan centres," *Urban Studies*, vol. 49, no. 2, pp. 239–254, 2012.
- [64] S. Iyer, M. Kitson, and B. Toh, "Social capital, economic growth and regional development," *Regional Studies*, vol. 39, no. 8, pp. 1015–1040, 2005.
- [65] Z. Yuan, X. Zheng, L. Zhang, and G. Zhao, "Urban competitiveness measurement of Chinese cities based on a structural equation model," *Sustainability*, vol. 9, no. 4, p. 666, 2017.
- [66] F.-C. Wu, "Optimization of correlated multiple quality characteristics using desirability function," *Quality Engineering*, vol. 17, no. 1, pp. 119–126, 2004.
- [67] N. R. Costa, J. Lourenço, and Z. L. Pereira, "Desirability function approach: a review and performance evaluation in adverse conditions," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 2, pp. 234–244, 2011.
- [68] Z. Hu, M. Cai, and H.-H. Liang, "Desirability function approach for the optimization of microwave-assisted extraction of saikosaponins from *Radix Bupleuri*," *Separation and Purification Technology*, vol. 61, no. 3, pp. 266–275, 2008.
- [69] B. John, "Application of desirability function for optimizing the performance characteristics of carbonitrided bushes," *International Journal of Industrial Engineering Computations*, vol. 4, no. 3, pp. 305–314, 2013.
- [70] PricewaterhouseCoopers, "Which are the largest city economies in the world and how might this change by 2025?," 2009.
- [71] N. Pengfei, M. Kamiya, and W. Haibo, "The global urban competitiveness report 2017-2018—housing prices: changing world cities," 2017, <https://unhabitat.org/global-urban-competitiveness-report-2017-2018-launched/>.