BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 04/25 pp603–620
DOI: 10.26599/BDMA.2024.9020019
Volume 7, Number 3, September 2024

House Price Prediction: A Multi-Source Data Fusion Perspective

Yaping Zhao, Jichang Zhao, and Edmund Y. Lam*

Abstract: House price prediction is of utmost importance in forecasting residential property prices, particularly as the demand for high-quality housing continues to rise. Accurate predictions have implications for real estate investors, financial institutions, urban planners, and policymakers. However, accurately predicting house prices is challenging due to the complex interplay of various influencing factors. Previous studies have primarily focused on basic property information, leaving room for further exploration of more intricate features, such as amenities, traffic, and social sentiments in the surrounding environment. In this paper, we propose a novel approach to house price prediction from a multi-source data fusion perspective. Our methodology involves analyzing house characteristics and incorporating factors from diverse aspects, including amenities, traffic, and emotions. We validate our approach using a dataset of 28 550 real-world transactions in Beijing, China, providing a comprehensive analysis of the drivers influencing house prices. By adopting a multi-source data fusion perspective and considering a wide range of influential factors, our approach offers valuable insights into house price prediction. The findings from this study possess the capability to improve the accuracy and effectiveness of house price prediction models, benefiting stakeholders in the real estate market.

Key words: price prediction; real estate; data mining; data fusion; machine learning

1 Introduction

House price prediction plays a crucial role in the research area focused on forecasting residential property prices. As economic development progresses, the demand for higher quality housing has increased, underscoring the growing importance of accurate house price prediction. The implications of accurately predicting house prices extend to various stakeholders, including real estate investors, financial institutions, urban planners, and policymakers. Accurate predictions not only contribute to market surveillance, but also empower sellers to determine optimal pricing strategies and assist potential buyers in making well-informed decisions.

However, accurately predicting house prices poses a considerable challenge due to the complex interplay of factors that influence them. The multifaceted nature of these factors makes it difficult to comprehensively measure and precisely predict house prices. Consequently, achieving comprehensive measurements and accurate predictions in house price remains a formidable task.

Previous studies have predominantly focused on basic property information, such as the quantity of rooms and the total floor area, to predict house prices[1]. However, more intricate factors, including the surrounding amenities, traffic conditions, and social sentiments, have received limited attention, leaving room for further exploration and research.

In this paper, we present house price prediction from

• Yaping Zhao and Edmund Y. Lam are with Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong 999077, China. E-mail: zhaoyp@eee.hku.hk; elam@eee.hku.hk.

• Jichang Zhao is with School of Economics and Management, Beihang University, Beijing 100191, China. E-mail: jichang@buaa.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 2023-11-24; revised: 2024-03-12; accepted: 2024-03-22

a multi-source data fusion perspective. Our methodology involves studying the characteristics of the house, as well as considering factors such as amenities, traffic, and emotions in the surrounding environment, as Fig. 1 shows. To validate our approach, we conduct an analysis of 28 550 real-world transactions in Beijing, China, providing a comprehensive analysis of the drivers influencing house prices. While Zhao et al.[2] laid the foundation for multi-source data fusion in house price prediction, we delve deeper into studying house features and explores additional predictive methods, resulting in a more comprehensive solution for accurate house price prediction.

The primary contributions of our research are outlined below:

- The examination of the correlation between various house features and their respective influence on house prices. We rank these features based on their importance, providing valuable insights into the factors that significantly affect house prices.
- The precision evaluations of different machine learning models in the task of multi-source data fusion for house price prediction. We compare and analyze the results obtained from support vector machines, linear regression, XGBoost regression, and random forest regression. Additionally, we make an attempt to explore different variants of multi-layer perceptron and investigate their performance in this particular task.
- The ablation study to highlight the unexpected economic influence of different aspects, including amenities, traffic, and emotions, on house prices.

By adopting a multi-source data fusion perspective and considering a wide range of influential factors, our approach provides valuable insights into house price prediction. The findings from this study have the potential to enhance the accuracy and effectiveness of house price prediction models, ultimately benefiting various stakeholders in the real estate market.

The subsequent sections of this paper will delve into the related work (Section 2), methodology (Section 3), experiments (Section 4), and conclusions (Section 5), further elucidating our approach and presenting the implications of our findings.

2 Related Work

The prediction of house prices is frequently tackled as the assessment of a diverse commodity, distinguished by a blend of utility-bearing attributes[3, 4]. Consequently, the monetary value assigned to a house can be perceived as a numerical depiction of a collection of these attributes. Extensive research endeavors have been undertaken in the past few decades to examine the interconnection between house prices and their corresponding attributes. For instance, Król[5] explored the relationship between apartment prices and noteworthy attributes through the application of hedonic analysis in Poland. In Türkiye,

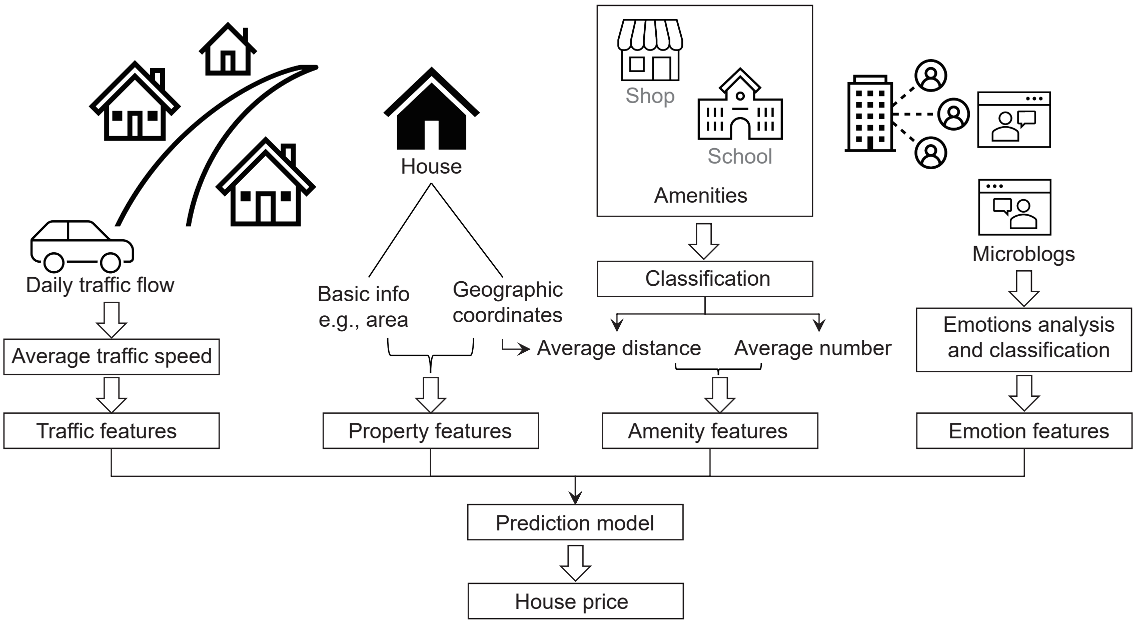


Fig. 1 Framework of the multi-source data fusion for house price prediction.

Refs. [6, 7] examined the positive and negative effects of different house features on house values. Kryvobokov and Wilhelmsson[8] ascertained the relative significance weights of location attributes influencing apartment market values in Donetsk, Ukraine. Ottensmann et al.[9] contrasted location metrics, including distances and travel time to the Central Business District (CBD) and multiple employment centers, in order to comprehend the influence of residential location on house prices in Indianapolis, Indiana, USA. Ozalp and Akinci[10] identified housing and environmental attributes that impact residential real estate sale prices in Artvin, Türkiye. These investigations, among numerous others, have delved into the connection between house prices and diverse attributes, culminating in the advancement of house price prediction methodologies that estimate prices based on inputted attributes.

Existing house price prediction methods, based on their underlying methodologies, estimate house prices by considering a variety of constituent attributes and are typically applied directly to the entire dataset. Several studies have followed this approach. For instance, Gu et al.[11] harnessed Support Vector Machines (SVM)[12] to predict house prices, exhibiting promising outcomes using cases from China. Wang et al.[13] introduced a novel model based on SVM to predict average house prices across different years, showcasing the effective utilization of Particle Swarm Optimization (PSO) to determine SVM parameters. Park and Bae[14] devised a general prediction model utilizing machine learning techniques, such as RIPPER, Naive Bayesian, and AdaBoost, comparing their classification accuracy performance.

Nevertheless, these models frequently neglect the impact of house location and its surroundings on prices, resulting in suboptimal prediction performance as the dataset size expands. Recent studies have shifted towards local perspectives in house price prediction, serving as viable alternatives and extensions to traditional modeling approaches. Among these studies, Bourassa et al.[15] compared various methods to incorporate spatial dependence into house price prediction. Case et al.[16] emphasized the significance of incorporating transactions from nearest neighbors for accurate predictions. Gerek[17] devised two adaptive approaches, considering grid partitioning and sub-clustering. Montero et al.[18] explored model variations to capture spatial effects in house prices, proposing a mixed model that accounted for spatial autocorrelation, spatial heterogeneity, and nonlinearities. The findings highlighted the effectiveness of nonlinear models in house price prediction.

While some recent studies incorporate factors, such as infrastructure[19] and neighborhoods[20–25], they rely on limited factors and overlook the intricate complexities associated with a multi-source data model.

Despite the extensive research on the house price prediction problem, our work diverges from most existing studies in several aspects. Firstly, our house dataset[2], as illustrated in Table 1, encompasses a more comprehensive range of transaction records and house attributes compared to the datasets utilized in previous studies. This enables us to conduct a more thorough exploration of the impact of various attributes on house prices and enhances our understanding of the prediction problem. Secondly, we approach house price prediction from a unique perspective by incorporating multi-source data fusion techniques. Through this approach, we have made a discovery, revealing that different features from diverse perspectives, such as amenities, traffic, and emotions[2], unexpectedly exert a socioeconomic influence on house prices. This novel finding adds a new dimension to the understanding of the factors driving house price dynamics.

3 Methodology

In accordance with the findings of PATE[2], our research aims to provide an in-depth analysis of multi-source data and its application in accurately predicting house prices per square meter. To achieve this, we commence with a comprehensive data collection and pre-processing phase, as outlined in Section 3.1. Subsequently, we extract 27 distinct features from the raw dataset, followed by the computation of feature correlations, which are detailed in Section 3.3. Furthermore, we conduct a detailed examination of the significance of each feature in relation to house prices, as documented in Section 3.4. Finally, we employ a range of diverse methods for house price prediction, including SVM[12], linear regression[30], XGBoost[31], random forest[32], and Multi-Layer Perceptron (MLP)[33], as elucidated in Section 3.5.

It is important to emphasize that our primary objective is to explore the potential of data fusion in

Table 1 Comparisons with prior studies on house price prediction, our research stands out in terms of the comprehensiveness of our multi-source data and features.

Reference	Number of data (⩾ 10 000)	Property		Amenity					Traffic	Emotion
		Basic Info	Geo-Info	Transport	Education	Hospital	Shop	Tourism		
[3]	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗
[5]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
[6]	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗
[7]	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
[8]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗
[9]	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗
[10]	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗
[14]	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
[15]	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
[16]	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
[18]	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
[26]	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗
[27]	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗
[28]	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
[29]	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

predicting house prices by integrating various factors such as property characteristics, amenities, traffic, and social emotions. Rather than introducing novel prediction techniques, our focus lies in assessing the efficacy and advantages of data fusion. As exemplified in previous studies that have employed multi-source data for house price prediction[2, 25], we adopt well-established prediction models. Through the experimental results presented in Section 4, we demonstrate that several of these methods have already achieved satisfactory performance and provided a comprehensive evaluation of the impact of different factors derived from multi-source data on house prices.

3.1 Data collection and pre-processing

To facilitate our house price prediction task from a multi-source data fusion perspective, we undertake comprehensive data collection and pre-processing procedures. This subsection presents the various steps involved in acquiring and preparing the data for analysis.

3.1.1 Property data collection

We initiate the data collection process by sourcing house transaction data from online platforms using Python web scraping techniques. We utilize the Requests library in Python to implement HTTP requests and access real estate websites to retrieve the HTML content of the property details pages. By analyzing the HTML structure of these pages, we determine the specific locations from which to extract the relevant information. For each house transaction, we extract pertinent features that reflect the fundamental characteristics of the properties, such as the availability of elevators and the number of bathrooms. Furthermore, we employ advanced geocoding techniques from Baidu Maps to accurately derive and store the geographic coordinates for each property based on the provided address. Overall, by collecting data from diverse online sources, we amass a dataset comprising 28 550 house transactions[34], each accompanied by its corresponding basic property features.

3.1.2 Amenities extraction

Considering the large number of properties in our dataset, storing detailed information such as names and addresses of nearby facilities after web scraping requires significant storage space. Therefore, we retain only the counts and mean distances of amenities for statistical analysis. Our web scraping strategy involves iterating over the existing real estate data in the database and using the latitude and longitude values as parameters to request web services. We retrieve information on surrounding facilities within a one-kilometer radius of each property, leveraging the geographical coordinates obtained earlier and the capabilities of Baidu Maps. These amenities are classified into five distinct categories: Transportation, tourist sites, educational establishments, healthcare centers, and dining establishments. As part of the preprocessing stage, we compute two key features for each property: The overall count and the mean distance to each amenity type. These features provide valuable information regarding the availability and proximity of amenities in the vicinity of the properties.

3.1.3 Traffic data acquisition

To incorporate the impact of transportation efficiency on house prices, we leverage Baidu Maps to capture detailed traffic flow data around each property. We collect granular traffic speed data at five-minute intervals throughout the day, from 6 a.m. to midnight. Processing this data involves calculating the average traffic speed metric for each property, which encapsulates the transportation dynamics in the vicinity. This feature provides a quantitative indicator of the accessibility and convenience associated with the location[35, 36].

3.1.4 Emotional sentiment analysis

In order to incorporate the emotional aspect into our predictive model, we acquire microblog posts. Following the social sentiment analysis method proposed by Fan et al.[37], we analyze the emotional content of each post and categorize them into five distinct emotional states: anger, dislike, happiness, sadness, and fear. For each property, we compute the distribution of these emotional sentiments, thereby deriving a set of features that reflect the emotional ambiance associated with it. This innovative approach enables us to integrate a layer of emotional intelligence into our dataset, offering a more comprehensive perspective on the factors influencing house prices.

3.2 Feature selection

In our comprehensive research endeavor, we diligently collect and preprocess a vast amount of data from multiple sources, resulting in the extraction of a total of 27 features. These features encompass various crucial aspects that play a significant role in determining house prices. We include detailed information, notations, and descriptions in Table 2 to provide a clear understanding of these features.

Our feature selection process meticulously integrates a spectrum of factors pivotal to determining house prices. It spans essential property attributes, such as bedroom count, bathroom count, and elevator presence, alongside proximity to key amenities like transportation, educational institutions, healthcare centers, dining options, and tourist attractions, acknowledging their critical role in enhancing a property's desirability and convenience. Additionally, leveraging Baidu Maps, we assess traffic dynamics by monitoring average traffic speeds, acknowledging that superior transportation links can significantly elevate property values. Furthermore, to enrich our model with a nuanced perspective, we analyze emotional sentiments from microblog posts, categorizing them into emotions, such as anger, dislike, happiness, sadness, and fear, thereby incorporating the emotional ambiance into our assessment of factors influencing house prices. This holistic approach ensures a comprehensive analysis of both tangible and intangible elements that affect property valuations.

As our primary objective is to accurately predict house prices, we designate the feature "`Price`" (as depicted in Table 2) as the dependent variable, denoted by y. The remaining features, represented as x_i, where $i = 0, 1, \ldots, 25$, are treated as independent variables. Each of these independent variables encompasses valuable information that contributes to the predictive power of our model. By incorporating this rich set of features, we aim to develop a robust and accurate predictive model for house prices.

3.3 Feature correlation

Understanding the relationships between different features is crucial in our analysis. We employ the Pearson's correlation coefficient, denoted as r_{pq}, to quantify the correlation between two features p and q[38]. The coefficient is calculated as follows:

$$r_{pq} = \frac{\sum\limits_{j=1}^{n}(p_j-\overline{p})(q_j-\overline{q})}{\sqrt{\sum\limits_{j=1}^{n}(p_j-\overline{p})^2}\sqrt{\sum\limits_{j=1}^{n}(q_j-\overline{q})^2}} \tag{1}$$

where n represents the sample size. The individual sample points of features p and q are denoted as p_j and q_j, respectively. The sample mean $\overline{p}$ is calculated as $\overline{p} = \frac{1}{n}\sum_{j=1}^{n} p_j$, and a similar calculation is for $\overline{q}$.

The Pearson's correlation coefficient ranges between -1 and 1. A value in proximity to 1 signifies the strong positive correlation between the features, while a value near -1 suggests the strong negative correlation, implying that a feature is more "opposite" in nature. A

Table 2 Features collected and extracted from multiple sources for house price prediction, along with descriptions[2].

Category	Feature	Description
Property	Year	Construction year of the building
	Elvt	Presence of an elevator in the building
	RmNum	Number of bedrooms in the house
	HllNum	Number of living and dining rooms in the house
	KchNum	Number of kitchens in the house
	BthNum	Number of bathrooms in the house
	Lat	Latitude coordinate of the house
	Lng	Longitude coordinate of the house
Amenity	TspNum	Number of surrounding transportation infrastructures
	TspDst	Average distance to surrounding transportation infrastructure
	AtrNum	Number of surrounding tourist attractions
	AtrDst	Average distance to surrounding tourist attractions
	EdcNum	Number of surrounding education and training institutions
	EdcDst	Average distance to education and training institutions
	HthNum	Number of surrounding healthcare infrastructures
	HthDst	Average distance to surrounding healthcare infrastructure
	RstNum	Number of surrounding restaurants
	RstDst	Average distance to surrounding restaurants
	RtlNum	Number of surrounding retail goods and services
	RtlDst	Average distance to surrounding retail goods and services
Traffic	TrfV	Average value of daily traffic speeds
Emotion	AgrPct	Percentage of anger in all emotions
	DstPct	Percentage of detestation in all emotions
	HppPct	Percentage of happiness in all emotions
	SadPct	Percentage of sadness in all emotions
	FeaPct	Percentage of fear in all emotions
Price	Price	Price per square meter of the house in Renminbi (RMB)

value close to 0 denotes the weak correlation between the features.

3.4 Feature importance

The concept of feature importance is crucial in evaluating the usefulness and value of every attribute in the decision tree construction. The frequency with which an attribute is employed in making pivotal decisions within decision trees directly influences its relative significance. By explicitly estimating the importance of different features, we can rank and compare them.

To estimate the importance of features in the context of predicting house prices, we utilize the random forest library in Python. This approach allows us to evaluate the comparative significance of every feature and gain insights into their contributions to the predictive model.

3.5 Prediction model

In our study, our choice of machine learning models is carefully curated to encompass a diverse array of algorithms, each renowned for its strengths in predictive analytics within the domain of house price forecasting. Specifically, we employ SVM (we employ four distinct machine learning methods: SVM)[12], linear regression[30], XGBoost[31], and random forest[32]. This selection is substantiated by a thorough review of existing literature[2, 25, 39], where these models have been identified for their high precision and reliability in the prediction of housing prices. Each model’s unique capabilities—ranging from SVM’s proficiency in modeling complex, non-linear relationships, linear regression’s transparency and ease of interpretation, XGBoost’s exceptional handling of varied data types and structures, to random forest's robustness against overfitting through its ensemble approach—collectively contribute to a comprehensive analytical framework.

In addition to these traditional machine learning models, we delve into the realm of deep learning by experimenting with MLP of varying architectures. This exploration is motivated by the deep learning methodology's potential for capturing intricate patterns in large-scale data. For the MLP configuration, given the dataset includes 27 features, the initial layer is designed with 32 neurons. This design strategy aims to expand the dimensionality from the original feature space to a more complex representation, potentially uncovering deeper relationships within the data. Subsequent layers are configured to either expand or contract in size, adhering to a strategy of doubling or halving the neuron count, with the centerpiece layer being the most expansive. This architectural design facilitates a deep and nuanced processing of features through the network, allowing for a sophisticated synthesis of information.

As a practical illustration, our three-layer MLP model adopts a configuration of "27-32-16-8-1", signifying the progression from the input layer, through intermediate layers, and culminating in the output layer for price prediction. In contrast, a more complex, five-layer model is structured as "27-32-64-32-16-8-1", showcasing an initial expansion followed by a gradual contraction, mirroring the network's attempt to distill the most salient features for accurate price estimation.

These configurations reflect a deliberate effort to harness the multifaceted nature of our dataset, derived from multiple sources, for the prediction of house prices. By integrating traditional machine learning models with deep learning architectures, we aim to leverage the unique advantages of each approach, thereby establishing a robust and versatile predictive model. This meticulous approach to model selection and configuration underlines our commitment to advancing the predictive accuracy and interpretability of house price forecasting models, ensuring they are well-equipped to navigate the complexities inherent in real estate data.

3.6 Evaluation

For the prediction accuracy measurement, five evaluation metrics are employed: R-squared (R^2), adjusted R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Those metrics offer comprehensive insights into the effectiveness of the models.

The formulas for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} \tag{2}$$

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{n - k - 1}\right] \tag{3}$$

$$\text{MAE} = \frac{\sum_{j=1}^{n} |y_j - \hat{y}_j|}{n} \tag{4}$$

$$\text{MSE} = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n} \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n}} \tag{6}$$

The actual values of the dependent variable and their corresponding predicted values are denoted by y_j and $\hat{y}_j$, respectively. The mean value of all the observed dependent variable values is represented by $\bar{y}$. The term k denotes the count of independent variables, except for the constant term. These evaluation metrics enable us to assess the accuracy and performance of our prediction models.

4 Experiment

4.1 Feature distribution

In Fig. 2, the multi-source features are represented using boxplots, providing a visual depiction of their distributions. To further examine the distributions of the multi-source features, Fig. 3 presents histograms. The histograms reveal certain characteristics of the features. Notably, features, such as `Price` and `Year`, exhibit highly skewed distributions. On the other hand, features like `EdcNum` and `EdcDst` appear to follow a normal distribution, while other features show either a normal or bimodal distribution of data, except for `Elvt`, `RmNum`, `KchNum`, and `BthNum`, which are discrete variables.

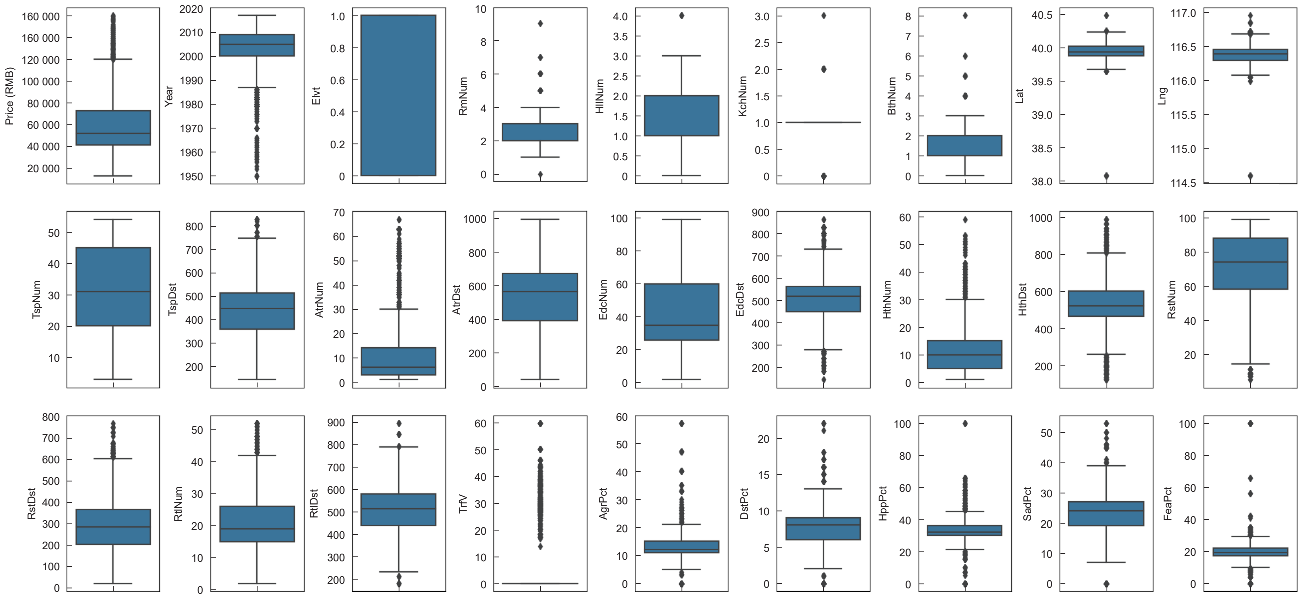


Fig. 2 Boxplots of the multi-source features with interesting trends or statistics.

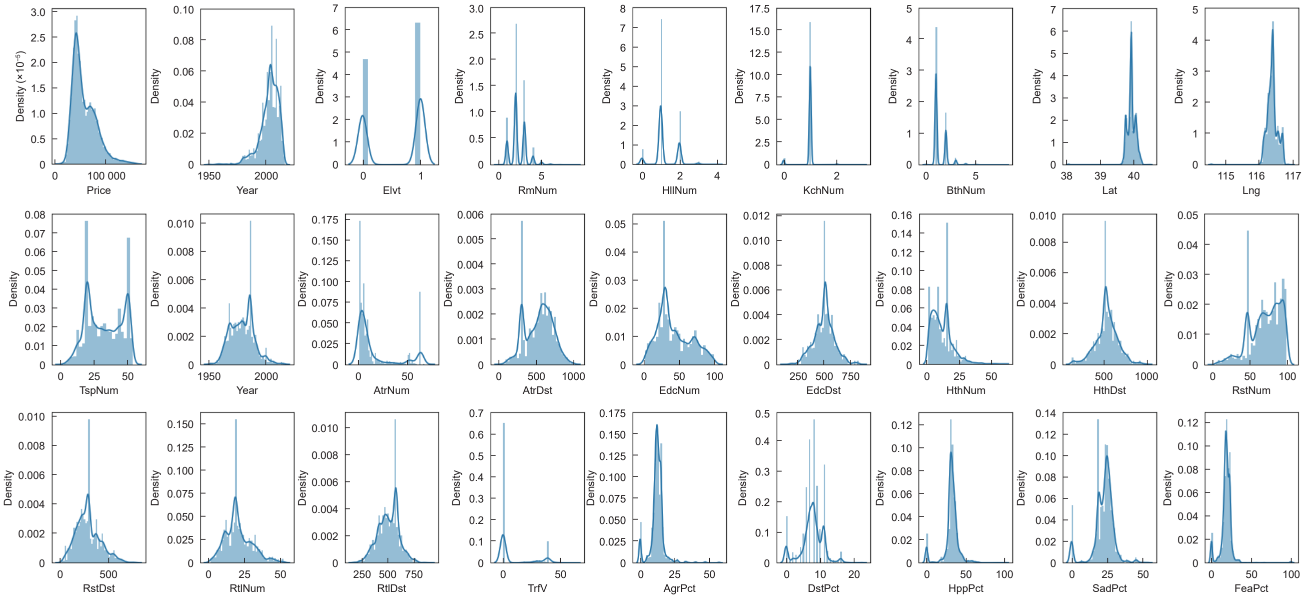


Fig. 3 Plots illustrate the distributions of the multi-source features. The histograms also show that features, like `Price` and `Year` have highly skewed distributions. Also `EdcNum` and `EdcDst` look to have a `normal` distribution and other features seem to have norma or bimodel ditribution of data except `Elvt`, `RmNum`, `KchNum`, and `BthNum` (which are discrete variables).

4.2 Feature correlation

According to Fig. 4, the absolute value of the correlation coefficient ($|r|$) between `Price` and other features lies within the interval of [0, 0.5], indicating a weak linear correlation. This implies that relying solely on one or a few features is unlikely to accurately predict house prices. Instead, a collective gathering of multiple features from various sources is necessary and effective for accurate prediction.

Furthermore, in terms of the correlation with house prices, we also observe correlations between features within the same category. Specifically, the following interesting patterns are observed:

(1) Property features, such as `RmNum` (number of bedrooms) and `BthNum` (number of bathrooms), exhibit a correlation coefficient of 0.62, indicating substantial association between the quantity of bedrooms and bathrooms within a residence.

(2) Amenity features exhibit diverse relationships.

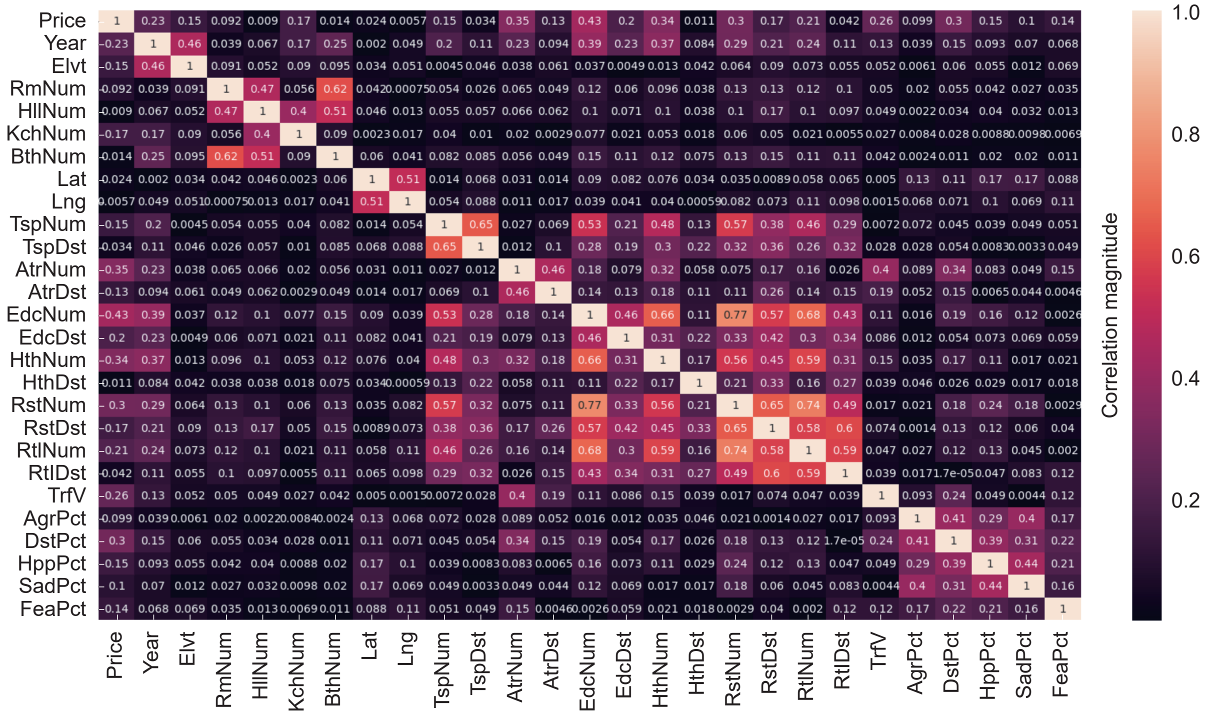


Fig. 4 Association between multi-source features, quantified by the absolute magnitude of the value of *r* from Eq. (1), commonly known as Pearson's correlation coefficient[38]. From this correlation matrix, we can see that if we only consider individual features, there is no significant correlation between any feature and `Price`. However, we can also observe that there may be strong correlations between some features (excluding `Price`). For instance, we see RmNum and BthNum, EdcNum and RstNum are highly correlated features.

For instance, the correlation coefficient between `EdcNum` (number of educational institutions) and `RstNum` (number of restaurants) is 0.77, between `RstNum` and `RtlNum` (number of retailing facilities) it is 0.74, and between `EdcNum` and `RtlNum` it is 0.68. These suggest that educational institutions usually coexist with restaurants and retail establishments in close proximity. Furthermore, restaurants and retailing establishments are frequently clustered within the same vicinity.

(3) The traffic value `TrfV` exhibits a relatively strong relationship with `AtrNum` (number of tourist attractions), indicating that regions with a high concentration of tourist attractions tend to have high vehicular traffic.

Notably, the features `AtrNum`, `EdcNum`, `HthNum`, `RstNum`, `TrfV`, and `DstPct` exhibit a higher correlation score with Price. To delve deeper into the relationship between each individual feature and the price, we present Fig. 5. As depicted in Fig. 5, while there is a faint hint of a linear fit in the overall trend of the data, it is not evident and can be considered negligible.

4.3 Feature importance

To assess the significance of various attributes in the random forest regression model, Fig. 6 provides a ranking. The following observations can be made:

(1) The geographical coordinates of the house, denoted by the latitude (`Lat`) and longitude (`Lng`), emerges as the most influential factors affecting the house price.

(2) Considering all the features related to amenities, the average count of nearby tourist attractions (`AtrNum`) and education institutions (`EdcNum`) prove to be the most significant. This implies that the availability of sightseeing opportunities and educational proximity are crucial considerations for potential house buyers.

(3) The year of construction (`Year`), which denotes the vintage of the property, holds the fifth position in terms of its significance in impacting the housing price.

(4) The mean proximity to adjacent transportation facilities (`TspDst`) emerges as the sixth most notable factor influencing the price.

(5) Interestingly, the average magnitude of daily traffic velocities (`TrfV`) attains the 24th position

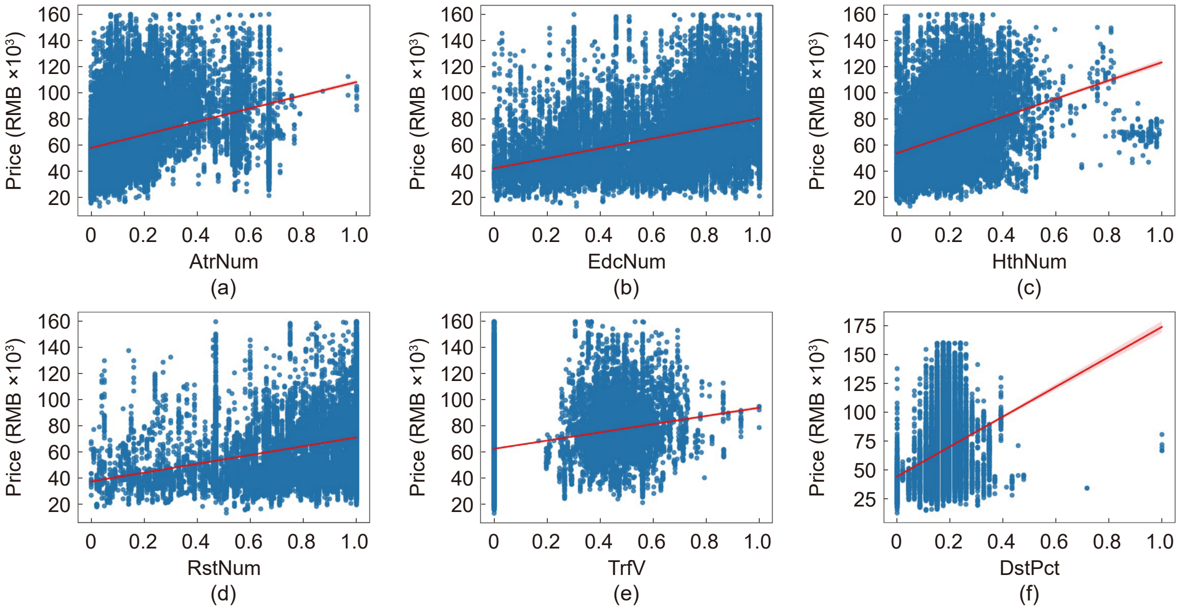


Fig. 5 Plots of the selected features against `Price`, where the features are with a higher correlation score with `Price` according to the correlation matrix in Fig. 4.

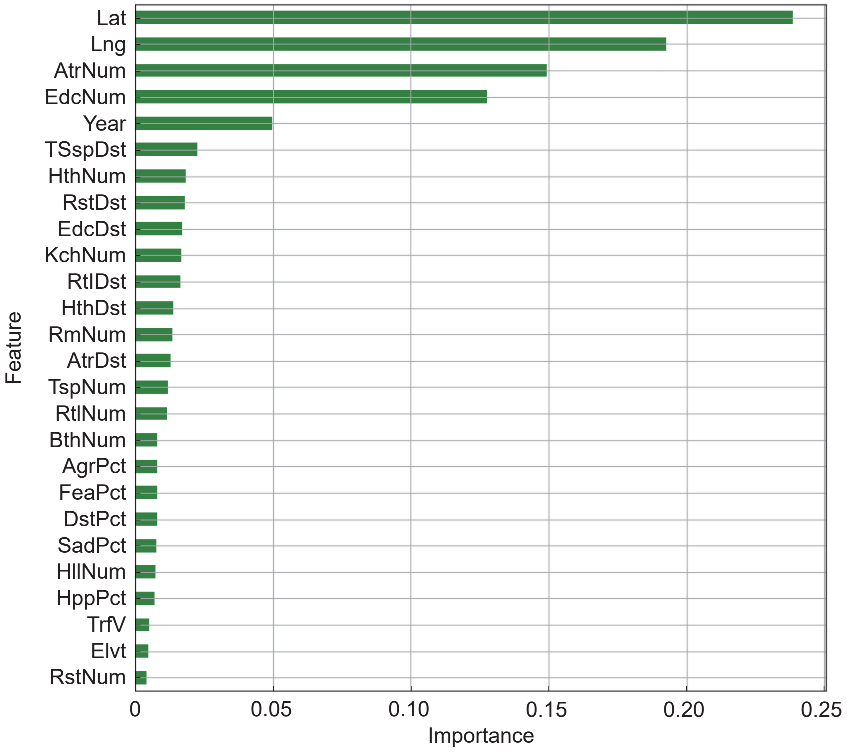


Fig. 6 Ranking of features in the random forest regression model, assessing their importance in predicting house prices. Several key observations can be made from the ranking: (1) The dominant factor exerting influence on house prices is the geographical location, characterized by latitude (`Lat`) and longitude (`Lng`). (2) Pertaining to amenities, the average count of nearby tourist attractions (`AtrNum`) and educational institutions (`EdcNum`) emerges as highly influential, underscoring the significance of sightseeing opportunities and educational accessibility for potential purchasers. (3) The year of construction (`Year`) assumes the fifth position in terms of importance, signifying its impact on pricing. (4) The average distance to transportation infrastructure (`TspDst`) ranks as the sixth most noteworthy factor. (5) Interestingly, the average daily traffic speeds (`TrfV`) demonstrate relatively lower significance, suggesting that proximity to public transportation outweighs concerns regarding traffic congestion.

among the 26 examined attributes, suggesting that while proximity to public transportation is valued by house consumers, the level of traffic congestion in the vicinity holds relatively less importance.

These findings shed light on the relative importance of various features in predicting house prices, providing valuable insights for real estate market analysis and decision-making.

4.4 Prediction model

Before applying the prediction model, we conduct experiments to detect and mitigate the effects of multicollinearity using Variance Inflation Factor (VIF) and Principal Component Analysis (PCA). The VIF value indicates the severity of multicollinearity, with values closer to 1 suggesting a lighter degree, while higher values indicate a stronger presence. Typically, a VIF equal to or greater than 10 is considered too large. In our study, as presented in Table 3, all variables exhibit VIF values below 5, indicating the absence of severe multicollinearity.

Additionally, we perform PCA to explore the results further. The PCA results, depicted in Fig. 7, indicate that only Component 1 contains a relatively substantial amount of information, while the remaining components experience a rapid decline in variance. Even Component 1, which captures more information than other components, only explains approximately 25% of the variance, suggesting suboptimal performance. Moreover, the other components contain even less information. Therefore, our data can be directly proceeded with the linear regression.

The dataset consists of 28 550 data points, comprising a dependent variable, denoted as y, and 26 independent features denoted as x_i (where i ranges from 0 to 25). To train and evaluate our prediction models, we randomly partition the dataset. This involves the allocation of 70% of the dataset as the training set, while the remaining 30% served as an independent testing set.

We employ a diverse set of four distinct machine learning techniques to predict house prices and conduct a comprehensive analysis of their outcomes. Following the training phase, we obtain models based on SVM, linear regression, XGBoost, and random forest algorithms. To facilitate a comprehensive comparison among the linear, XGBoost, and random forest regression models, we present their outcomes on the testing set visually in Fig. 8. Some observations can be made:

Table 3 Results of VIF analysis for the variables included in our study. The variables are listed in descending order based on their VIF values, which indicate the extent of multicollinearity among the variables.

Variable	VIF
RstNum	4.160 353
EdcNum	3.863 456
RtlNum	3.081 778
RstDst	2.597 066
TspNum	2.573 293
HthNum	2.210 560
RtlDst	2.096 533
BthNum	2.004 899
TspDst	1.992 980
Year	1.898 299
RmNum	1.825 553
AtrNum	1.802 491
HllNum	1.764 224
Lat	1.608 337
DstPct	1.596 789
Lng	1.547 029
SadPct	1.503 720
HppPct	1.473 639
Elvt	1.464 674
EdcDst	1.433 411
AgrPct	1.392 132
AtrDst	1.362 955
KchNum	1.307 281
TrfV	1.235 336
HthDst	1.219 479
FeaPct	1.170 280

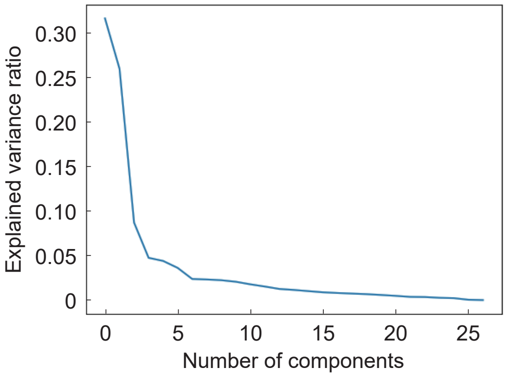


Fig. 7 Results of PCA for the dataset used in our study.

(1) Figures 8a, 8b, and 8c illustrate the disparities between the ground-truth prices and the predicted values from the linear regression, XGBoost, and random forest regression models, respectively. In

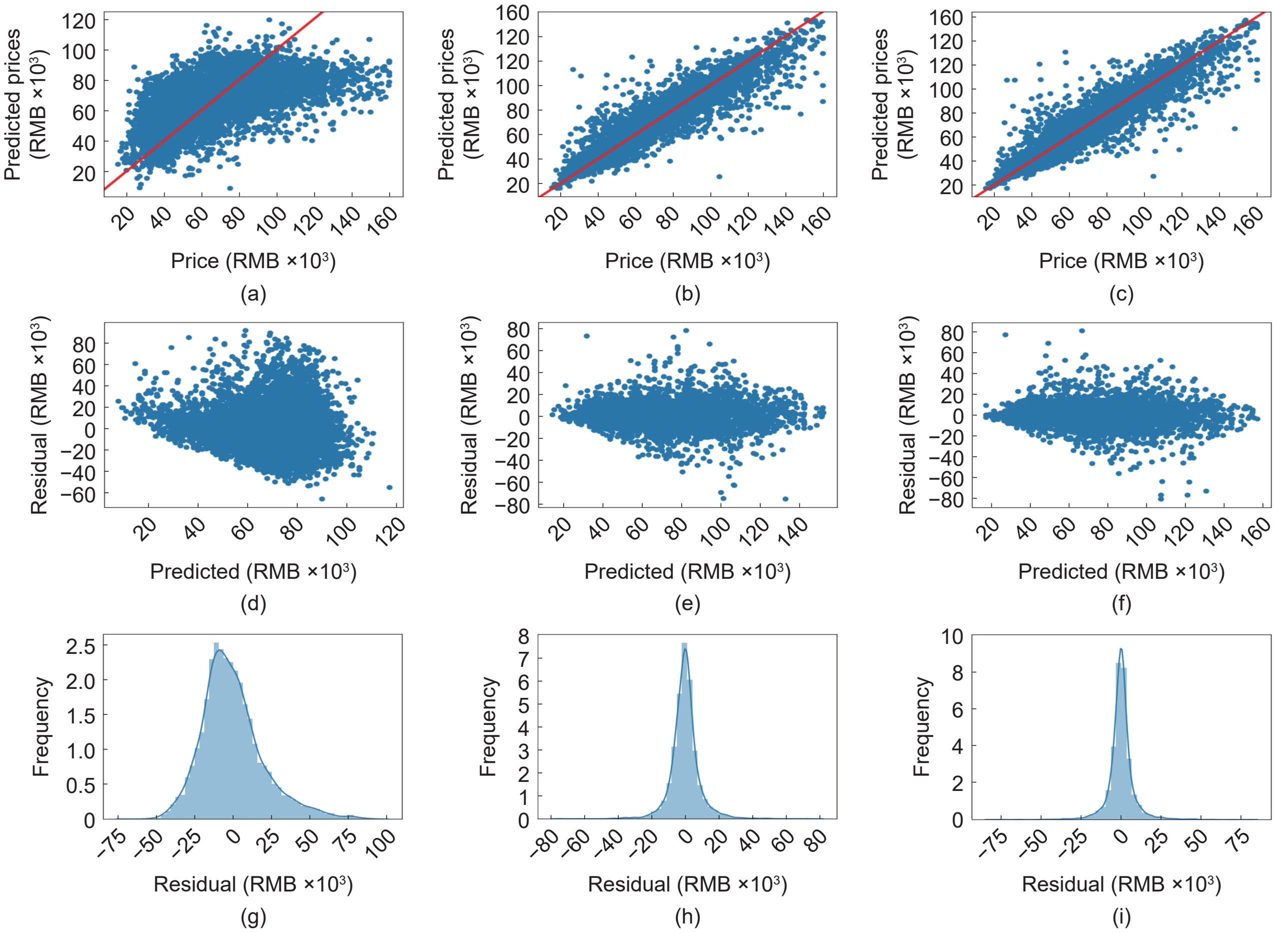


Fig. 8 Comparative analysis of house price prediction models: linear regression, XGBoost, and random forest. The top row presents the disparities between actual prices and predicted values from (a) linear regression, (b) XGBoost regression, and (c) random forest models. The red line represents the *x*=*y* reference. The middle row illustrates the residuals, showcasing deviations between actual prices and predicted values for (d) linear regression, (e) XGBoost regression, and (f) random forest models. The bottom row displays the error distribution histograms for (g) linear regression, (h) XGBoost regression, and (i) random forest regression models.

contrast to the linear regression model, which exhibits a tendency to overestimate actual prices, the XGBoost and random forest models demonstrate a balanced dispersion of data points around the unity line ($x = y$). This indicates that the predictions of the XGBoost and random forest models are more accurate.

(2) Figures 8d, 8e, and 8f display the residuals, representing the disparities between the ground-truth prices and the predicted values from the linear regression, XGBoost, and random forest models. The residuals of the XGBoost and random forest models are evenly distributed around zero point, indicating a favorable performance. Conversely, the linear regression model exhibits numerous noticeable outliers, particularly around the predicted value of 80 000.

(3) Figures 8g, 8h, and 8i depict the histograms of the errors for the linear regression, XGBoost, and random forest regression models, respectively. The errors in all models exhibit a normal distribution. Nevertheless, the error variance in the linear regression model is greater than those of the XGBoost and random forest models.

These comparisons offer valuable insights into the efficacy of different prediction models for house price estimation. Table 4 provides the evaluation and comparison of all the models. From Table 4, we can observe that the performance of MLP surpasses that of SVM and linear regression, but it still has limitations. Even the 15-layer MLP performs inferior to XGBoost and random forest. Furthermore, as the depth of the MLP increases, its performance tends to improve.

Table 4 Evaluation and comparision of all the models.

Model	R^2	MAE	RMSE
SVM	−0.5579	19 833	25 243
Linear regression	0.3651	15 235	20 057
3-layer MLP	0.3756	14 999	19 889
5-layer MLP	0.4091	14 426	19 189
7-layer MLP	0.4510	14 005	18 734
9-layer MLP	0.4615	13 707	18 926
11-layer MLP	0.4903	13 884	18 186
13-layer MLP	0.5414	12 502	17 064
15-layer MLP	0.5525	12 302	16 318
XGBoost	0.8770	5721	8829
Random forest	0.8934	4932	8219

Specifically, as we increase the number of layers from 3 to 15, the depth of the MLP expands fivefold and leads to an approximate 20% reduction in error.

4.5 Ablation study

To assess the influence of various features derived from a fusion of multi-source data for house price prediction, we examine the performance of the SVM, linear regression, XGBoost, and random forest models on both the training and testing datasets.

As presented in Table 5, in contrast to the scenario of utilizing solely property features (with only P), the addition of any supplementary feature (amenity, traffic, or emotions) demonstrates performance improvement. Furthermore, the following observations are noted: (1) The utilization of comprehensive attributes encompassing diverse dimensions such as property characteristics, amenities, transportation, and emotional factors leads to the attainment of optimal performance levels. (2) Notably, the omission of amenity-related features induces the most pronounced decrease in performance, surpassing the impact observed from excluding traffic or emotional features. This observation accentuates the influential role of amenity features compared to traffic and emotions. (3) The exclusion of traffic-related attributes exhibits a comparatively moderate effect on performance. This can be attributed to the unidimensional nature of traffic features, which inherently contributes to their relatively lower influence on the overall predictive capability. Nevertheless, even the inclusion of this one-dimensional traffic feature contributes to performance enhancement.

These findings highlight the importance of incorporating various types of data, particularly amenity features, in accurately predicting house prices. The ablation study provides valuable insights into the relative contributions of different feature categories, facilitating a deeper understanding of the underlying factors influencing house prices.

4.6 Sensitivity analysis

To provide further insights into the contribution of each data source to the prediction accuracy, we conduct a sensitivity analysis. The analysis aims to assess the impact of variations in the input variables on the target variable, using the adopted random forest prediction model, which demonstrates the best performance according to Table 4.

In this analysis, we focus on 26 input variables and predict the target variable, denoted as y, using the random forest model. We address two key questions during the sensitivity analysis:

- Impact of a 5% independent increase in input variables: We examine the influence of a 5% increase in each input variable on the target variable. By systematically varying the input variables while keeping other factors constant, we determine the resulting changes in the predicted price.
- Impact of a 5% independent decrease in input variables: Similarly, we investigate the impact of a 5% decrease in each input variable on the target variable. By reducing the input variables' values while holding other factors constant, we measure the corresponding changes in the predicted price.

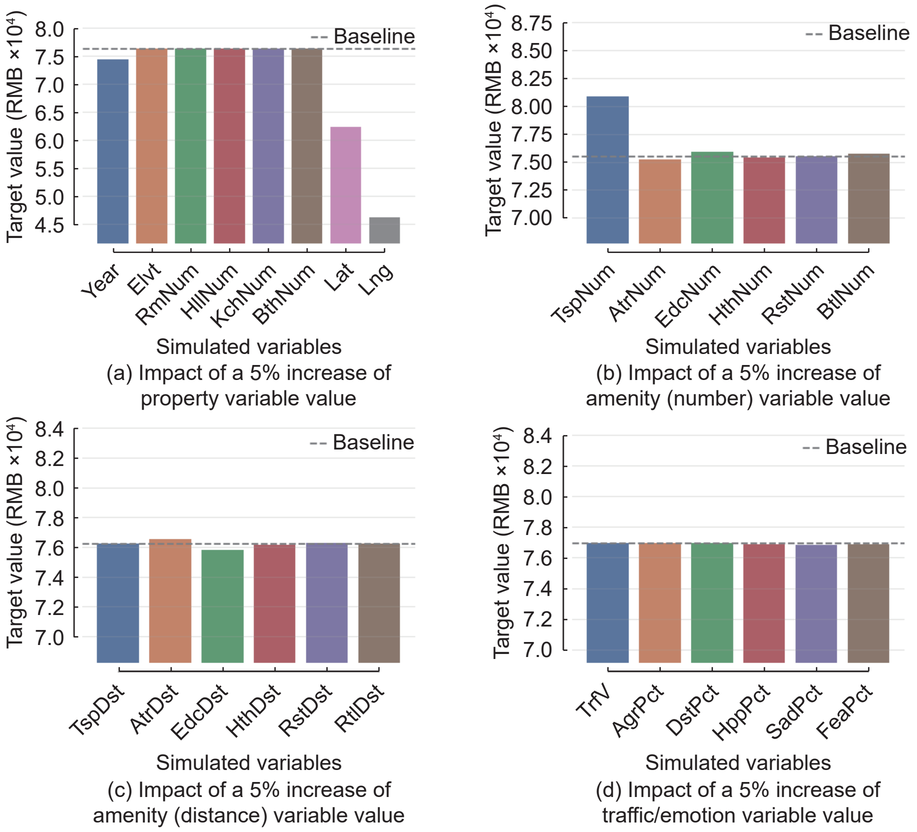
The outcomes of these sensitivity analyses are presented in Figs. 9 and 10, respectively. Figure 9 illustrates the results pertaining to the impact of a 5% independent increase in the input variables on the target variable. Conversely, Fig. 10 showcases the effects of a 5% independent decrease in the input variables on the target variable.

By examining these figures, we made several key discoveries that shed light on the influence of different variables on house prices. These findings contribute to a deeper understanding of the factors affecting the prediction accuracy. The following important observations are made:

(1) The changes in the Latitude (`Lat`) and Longitude (`Lng`) variables have a significant impact on house prices. This finding is reasonable since variations in latitude and longitude can result in houses being located further away from the city center. In real-life scenarios, house prices are generally highly correlated

Table 5 Experiments with with diverse setups: (1) Utilizing solely property features, referred to as with only P; (2) incorporating property, traffic, and emotional features, excluding amenity features, referred to as without A; (3) incorporating property, amenity, and emotional features, excluding traffic features, referred to as without T; (4) incorporating property, amenity, and traffic features, excluding emotional features, referred to as without E; and (5) employing all features encompassing property, amenity, traffic, and emotions, referred to as with PATE. The directions of the arrows indicate whether higher or lower values are better, and bold text highlights these better values.

Data	Method	R^2 ↑	Adjusted R^2 ↑	MAE ↓	MSE ↓	RMSE ↓
Training set	SVM with only P	−0.0148	−0.0152	20 251	672 425 564	25 931
	SVM without A	−0.0122	−0.0129	20 219	670 742 089	25 898
	SVM without T	−0.0075	−0.0088	20 177	667 666 607	25 839
	SVM without E	−0.0085	−0.0095	20 189	668 287 594	25 851
	SVM with PATE	**−0.0082**	**−0.0095**	**20 182**	**668 044 046**	**25 847**
	Linear regression with only P	0.1674	0.1671	18 284	551 713 399	23 489
	Linear regression without A	0.2520	0.2515	16 947	495 668 829	22 264
	Linear regression without T	0.3730	0.3722	15 499	415 469 247	20 383
	Linear regression without E	0.3636	0.3629	15 582	421 702 012	20 535
	Linear regression with PATE	**0.3797**	**0.3789**	**15 391**	**411 032 381**	**20 274**
	XGBoost regression with only P	0.9095	0.9095	5206	59 965 069	7744
	XGBoost regression without A	0.9184	0.9183	4941	54 069 367	7353
	XGBoost regression without T	0.9331	0.9330	4416	44 350 499	6660
	XGBoost regression without E	0.9319	0.9318	4477	45 145 802	6719
	XGBoost regression with PATE	**0.9343**	**0.9342**	**4387**	**43 549 356**	**6599**
	Random forest with only P	0.9721	0.9721	2480	18 466 444	4297
	Random forest without A	0.9722	0.9722	2461	18 372 367	4286
	Random forest without T	0.9726	0.9725	2448	18 143 797	4259
	Random forest without E	0.9726	0.9726	2456	18 135 870	4265
	Random forest with PATE	**0.9726**	**0.9726**	**2448**	**18 133 433**	**4258**
Testing set	SVM with only P	−0.0124	−0.0134	19 899	641 522 761	25 328
	SVM without A	−0.0099	−0.0116	19 871	639 963 349	25 297
	SVM without T	−0.0049	−0.0078	19 827	636 760 159	25 234
	SVM without E	−0.0059	−0.0084	19 840	637 453 688	25 247
	SVM with PATE	**−0.0056**	**−0.0086**	**19 833**	**637 188 037**	**25 243**
	Linear regression with only P	0.1626	0.1618	17 905	530 648 157	23 036
	Linear regression without A	0.2437	0.2424	16 696	479 250 332	21 892
	Linear regression without T	0.3591	0.3572	15 324	406 118 449	20 152
	Linear regression without E	0.3510	0.3494	15 358	411 213 070	20 278
	Linear regression with PATE	**0.3651**	**0.3632**	**15 235**	**402 302 484**	**20 057**
	XGBoost regression with only P	0.8560	0.8558	6244	91 267 181	9553
	XGBoost regression without A	0.8646	0.8644	6080	85 802 314	9263
	XGBoost regression without T	0.8751	0.8747	5773	79 153 827	8897
	XGBoost regression without E	0.8740	0.8737	5814	79 830 419	8935
	XGBoost regression with PATE	**0.8770**	**0.8766**	**5721**	**77 956 264**	**8829**
	Random forest with only P	0.8867	0.8866	5037	71 763 746	8471
	Random forest without A	0.8893	0.8891	4967	70 132 674	8374
	Random forest without T	0.8920	0.8917	4941	68 382 700	8269
	Random forest without E	0.8929	0.8926	4941	67 855 085	8237
	Random forest with PATE	**0.8934**	**0.8931**	**4932**	**67 547 377**	**8219**



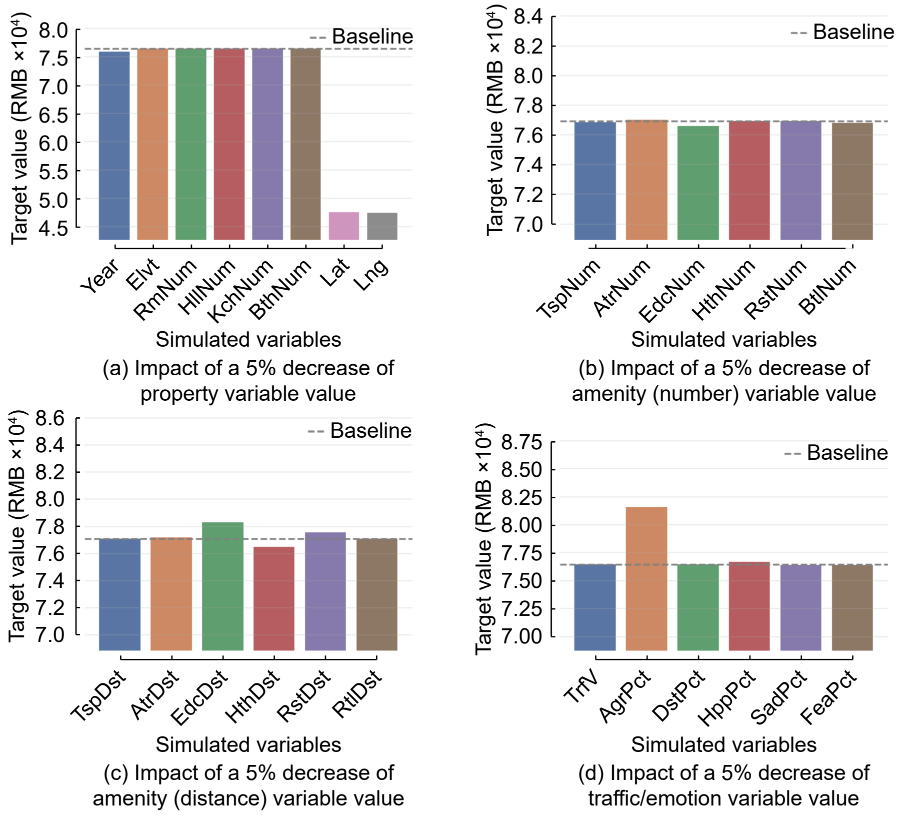
(a) Impact of a 5% increase of property variable value

(b) Impact of a 5% increase of amenity (number) variable value

(c) Impact of a 5% increase of amenity (distance) variable value

(d) Impact of a 5% increase of traffic/emotion variable value

Fig. 9 Impact of a 5% independent increase in input variables on the target variable.



(a) Impact of a 5% decrease of property variable value

(b) Impact of a 5% decrease of amenity (number) variable value

(c) Impact of a 5% decrease of amenity (distance) variable value

(d) Impact of a 5% decrease of traffic/emotion variable value

Fig. 10 Impact of a 5% independent decrease in input variables on the target variable.

with location.

(2) An increase in the number of nearby transportation facilities, represented by the variable `TspNum`, leads to a noticeable increase in house prices. This observation is also logical, as areas with better transportation accessibility typically command higher property values.

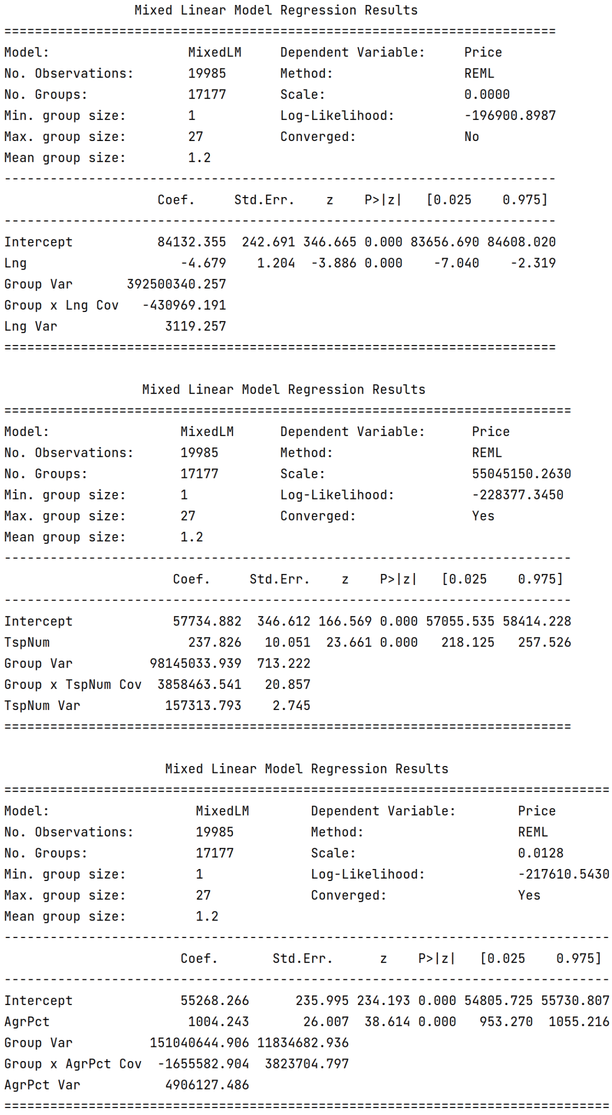
(3) Surprisingly, a decrease in the `ArgPct` variable, which represents the percentage of anger among all emotions, is associated with an increase in house prices. This finding is intriguing and unexpected, suggesting a potentially interesting relationship between emotional expressions and property values.

These discoveries highlight the significance of specific variables in influencing house prices. The sensitivity analysis provides valuable insights into the relative importance and unexpected associations between the variables and the target variable, enhancing our understanding of the prediction model's accuracy and the underlying dynamics of the housing market.

4.7 Mixed-effects model

Figure 4 implies the potential presence of batch effects and other stochastic influences, thereby indicating that utilizing a mixed-effects model could be a more suitable alternative. Furthermore, the Sensitivity Analysis conducted in Section 4.6 reveals that the variables `Lat`, `Lng`, `TspNum`, and `AgrPct` exhibit considerable impacts on the target variable. Given these findings, we have chosen to utilize Linear Mixed Effects (LME)[40] models to analyze the data, as presented in Fig. 11.

Although we have obtained initial experimental outcomes, it is important to acknowledge that the model's performance is significantly inferior to that of the XGBoost and random forest methods. These more advanced algorithms have demonstrated superior predictive capabilities throughout our study.



```
                 Mixed Linear Model Regression Results
========================================================================
Model:                  MixedLM     Dependent Variable:     Price
No. Observations:       19985       Method:                 REML
No. Groups:             17177       Scale:                  0.0000
Min. group size:        1           Log-Likelihood:         -196900.8987
Max. group size:        27          Converged:              No
Mean group size:        1.2
------------------------------------------------------------------------
                    Coef.     Std.Err.    z    P>|z|   [0.025    0.975]
------------------------------------------------------------------------
Intercept           84132.355  242.691 346.665 0.000 83656.690 84608.020
Lng                    -4.679    1.204  -3.886 0.000    -7.040    -2.319
Group Var       392500340.257
Group x Lng Cov   -430969.191
Lng Var              3119.257
========================================================================

                  Mixed Linear Model Regression Results
==========================================================================
Model:                 MixedLM      Dependent Variable:      Price
No. Observations:      19985        Method:                  REML
No. Groups:            17177        Scale:                   55045150.2630
Min. group size:       1            Log-Likelihood:          -228377.3450
Max. group size:       27           Converged:               Yes
Mean group size:       1.2
--------------------------------------------------------------------------
                      Coef.     Std.Err.    z    P>|z|   [0.025    0.975]
--------------------------------------------------------------------------
Intercept             57734.882  346.612 166.569 0.000 57055.535 58414.228
TspNum                  237.826   10.051  23.661 0.000   218.125   257.526
Group Var          98145033.939  713.222
Group x TspNum Cov  3858463.541   20.857
TspNum Var           157313.793    2.745
==========================================================================

                     Mixed Linear Model Regression Results
===============================================================================
Model:                   MixedLM        Dependent Variable:        Price
No. Observations:        19985          Method:                    REML
No. Groups:              17177          Scale:                     0.0128
Min. group size:         1              Log-Likelihood:            -217610.5430
Max. group size:         27             Converged:                 Yes
Mean group size:         1.2
-------------------------------------------------------------------------------
                       Coef.       Std.Err.      z    P>|z|   [0.025    0.975]
-------------------------------------------------------------------------------
Intercept              55268.266      235.995 234.193 0.000 54805.725 55730.807
AgrPct                  1004.243       26.007  38.614 0.000   953.270  1055.216
Group Var          151040644.906 11834682.936
Group x AgrPct Cov  -1655582.904  3823704.797
AgrPct Var           4906127.486
===============================================================================
```

Fig. 11 Mixed linear model regression results.

5 Conclusion

In this paper, we present a comprehensive analysis of house price prediction from a multi-source data fusion perspective. By incorporating property features, amenity data, traffic information, and social emotions, we aim to uncover the underlying factors influencing house prices. Through extensive experiments, we demonstrate that the integration of various types of data improves the predictive precision of the models. The comprehensive consideration of property features, amenity data, traffic information, and social emotions yields the highest predictive accuracy.

These findings can guide real estate buyers, sellers, and policymakers in making informed decisions. Moreover, it stirs and benefits social and economic analysis[41]. Future research can explore additional data sources[42] and advanced modeling techniques[43–45] to further enhance the accuracy of house price prediction models.

References

[1] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French, Real estate appraisal: A review of valuation methods, *J. Prop. Investment Finance*, vol. 21, no. 4, pp. 383–401, 2003.

[2] Y. Zhao, R. Ravi, S. Shi, Z. Wang, E. Y. Lam, and J. Zhao, PATE: Property, amenities, traffic and emotions coming together for real estate price prediction, in *Proc. IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–10.

[3] 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.

[4] H. Kuşan, O. Aytekin, and İ. Özdemir, The use of fuzzy logic in predicting house selling price, *Expert Syst. Appl.*, vol. 37, no. 3, pp. 1808–1813, 2010.

[5] A. Król, Application of hedonic methods in modelling real estate prices in Poland, in *Data Science, Learning by Latent Structures, and Knowledge Discovery*, B. Lausen, S. Krolak-Schwerdt, and M. Böhmer, eds. Berlin Germany: Springer, 2015, pp. 501–511.

[6] R. Yayar and D. Demir, Hedonic estimation of housing market prices in Turkey, *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, vol. 43, pp. 67–82, 2014.

[7] H. Selim, Determinants of house prices in Turkey: Hedonic regression versus artificial neural network, *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2843–2852, 2009.

[8] M. Kryvobokov and M. Wilhelmsson, Analysing location attributes with a hedonic model for apartment prices in Donetsk, Ukraine, *Int. J. Strategic Prop. Manage.*, vol. 11, no. 3, pp. 157–178, 2007.

[9] J. R. Ottensmann, S. Payton, and J. Man, Urban location and housing prices within a hedonic model, *J. Reg. Anal. Policy*, vol. 38, no. 1, pp. 19–35, 2008.

[10] A. Y. Ozalp and H. Akinci, Correction to: The use of hedonic pricing method to determine the parameters affecting residential real estate prices, *Arab. J. Geosci.*, vol. 11, no. 1, p. 2, 2018.

[11] J. Gu, M. Zhu, and L. Jiang, Housing price forecasting based on genetic algorithm and support vector machine, *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3383–3386, 2011.

[12] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, Support vector machines, *IEEE Intell. Syst. Their Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[13] X. Wang, J. Wen, Y. Zhang, and Y. Wang, Real estate price forecasting based on SVM optimized by PSO, *Optik*, vol. 125, no. 3, pp. 1439–1443, 2014.

[14] B. Park and J. K. Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2928–2934, 2015.

[15] S. Bourassa, E. Cantoni, and M. Hoesli, Predicting house prices with spatial dependence: A comparison of alternative methods, *J. Real Estate Res.*, vol. 32, no. 2, pp. 139–160, 2010.

[16] B. Case, J. Clapp, R. Dubin, and M. Rodriguez, Modeling spatial and temporal house price patterns: A comparison of four models, *J. Real Estate Finance Econ.*, vol. 29, no. 2, pp. 167–191, 2004.

[17] I. H. Gerek, House selling price assessment using two different adaptive neuro-fuzzy techniques, *Autom. Constr.*, vol. 41, pp. 33–39, 2014.

[18] J. M. Montero, R. Mínguez, and G. Fernández-Avilés, Housing price prediction: Parametric versus semi-parametric spatial hedonic models, *J. Geogr. Syst.*, vol. 20, no. 1, pp. 27–55, 2018.

[19] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z. H. Zhou, Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering, in *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 1047–1056.

[20] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, and J. Yuan, Sparse real estate ranking with online user reviews and offline moving behaviors, in *Proc. 2014 IEEE Int. Conf. Data Mining*, Shenzhen, China, 2014, pp. 120–129.

[21] Walking the walk: How walkability raises home values in U.S. cities, Report, National Association of City Transportation Officials, USA, https://policycommons.net/artifacts/4498224/walking-the-walk/5300886/, 2023.

[22] E. Washington and E. Dourado, The premium for walkable development under land use regulations, *SSRN Electron. J.*, doi: 10.2139/ssrn.3169535.

[23] W. Wang, S. Yang, Z. He, M. Wang, J. Zhang, and W. Zhang, Urban perception of commercial activeness from satellite images and streetscapes, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 647–654.

[24] D. Hristova, L. M. Aiello, and D. Quercia, The new urban success: How culture pays, *Front. Phys.*, vol. 6, p. 27, 2018.

[25] M. de Nadai and B. Lepri, The economic value of neighborhoods: Predicting real estate prices from the urban environment, in *Proc. IEEE 5th Int. Conf. Data Science and Advanced Analytics*, Turin, Italy, 2018, pp. 323–330.

[26] M. Kuntz and M. Helbich, Geostatistical mapping of real estate prices: An empirical comparison of kriging and cokriging, *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 9, pp. 1904–1921, 2014.

[27] A. S. Adair, J. N. Berry, and W. S. McGreal, Hedonic modelling, housing submarkets and residential valuation, *J. Prop. Res.*, vol. 13, no. 1, pp. 67–83, 1996.

[28] G. Gao, Z. Bao, J. Cao, A. K. Qin, and T. Sellis, Location-centered house price prediction: A multi-task learning approach, *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, p. 32, 2022.

[29] A. Can, Specification and estimation of hedonic housing price models, *Reg. Sci. Urban Econ.*, vol. 22, no. 3, pp. 453–474, 1992.

[30] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. 6th ed. Hoboken, NJ, USA: John Wiley & Sons, 2021.

[31] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[32] A. Liaw and M. Wiener, Classification and regression by randomForest, *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[33] M. Riedmiller, Advanced supervised learning in multi-layer perceptrons—From backpropagation to adaptive learning algorithms, *Comput. Stand. Interfaces*, vol. 16, no. 3, pp. 265–278, 1994.

[34] Y. Zhao, S. Shi, R. Ravi, Z. Wang, E. Y. Lam, and J. Zhao, H4M: Heterogeneous, multi-source, multi-modal, multi-view and multi-distributional dataset for socioeconomic analytics in the case of Beijing, in *Proc. IEEE 9th Int. Conf. Data Science and Advanced Analytics*,

Shenzhen, China, 2022, pp. 1–10.

[35] B. Zhou, J. Liu, S. Cui, and Y. Zhao, Large-scale traffic congestion prediction based on multimodal fusion and representation mapping, in *Proc. IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–9.

[36] B. Zhou, J. Liu, S. Cui, and Y. Zhao, A large-scale spatio-temporal multimodal fusion framework for traffic prediction, *Big Data Mining and Analytics*, doi: 10.26599/BDMA.2024.9020020.

[37] R. Fan, J. Zhao, Y. Chen, and K. Xu, Anger is more influential than joy: Sentiment correlation in weibo, *PLoS One*, vol. 9, no. 10, p. e110184, 2014.

[38] I. Cohen, Y. Huang, J. Chen, and J. Benesty, Pearson correlation coefficient, in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.

[39] C. H. R. Madhuri, G. Anuradha, and M. V. Pujitha, House price prediction using regression techniques: A comparative study, in *Proc. Int. Conf. Smart Structures and Systems*, Chennai, India, 2019, pp. 1–5.

[40] J. C. Pinheiro and D. M. Bates, Linear mixed-effects models: Basic concepts and examples, in *Mixed-Effects Models in Sand S-Plus*, J. C. Pinheiro and D. M. Bates, eds. New York, NY, USA: Springer, 2000, pp. 3–56.

[41] Y. Zhao, Z. Wang, and E. Y. Lam, Improving source localization by perturbing graph diffusion, in *Proc. IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–9.

[42] Y. Zhao and E. Y. Lam, SASA: Saliency-aware self-adaptive snapshot compressive imaging, In *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Seoul, Korea, Republic of, 2024, pp. 2370–2374.

[43] Y. Zhao, S. Zheng, and X. Yuan, Deep equilibrium models for snapshot compressive imaging, in *Proc. 37th AAAI Conf. Artificial Intelligence*, Washington, DC, USA, 2023, pp. 406.

[44] Y. Zhao, G. Li, and E. Y. Lam, Cross-camera human motion transfer by time series analysis, in *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Seoul, Korea, Republic of, 2024, pp. 4985–4989.

[45] Y. Zhao, H. Zheng, Z. Wang, J. Luo, and E. Y. Lam, Manet: Improving video denoising with a multi-alignment network, in *Proc. 2022 IEEE Int. Conf. Image Processing*, Bordeaux, France, 2022, pp. 2036–2040.

Yaping Zhao is currently a PhD candidate at Department of Electrical and Electronic Engineering, the University of Hong Kong (HKU), China, supervised by Prof. Edmund Y. Lam. Prior to joining HKU, she received the BS degree from Beihang University under the supervision of Prof. Jichang Zhao, China in 2018, and the MS degree from Tsinghua University under the supervision of Prof. Lu Fang, China in 2021. Her research interests encompass computational imaging, computer vision, machine learning, and artificial intelligence.

Jichang Zhao received the BEng and PhD degrees from Beihang University, China in 2008 and 2014, respectively. He is currently a full professor at School of Economics and Management, Beihang University, China. His research interests include computational social science and complex systems.

Edmund Y. Lam received the BS (with distinction), MS, and PhD degrees in electrical engineering from Stanford University, USA. He is currently a professor at Department of Electrical and Electronic Engineering, the University of Hong Kong, China. Concurrently, he is also the director of the Computer Engineering Program and the founding director of the Imaging Systems Laboratory at The University of Hong Kong, China. He has also been a visiting associate professor at Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA. His research focuses on computational optics and imaging, which span from the design of algorithms and systems to applications, especially in semiconductor manufacturing and biomedicine. In addition to advancing inverse imaging techniques for image reconstruction and quality enhancement, he has also been a pioneer in applying artificial intelligence to computational imaging, developing deep learning algorithms that significantly improve image resolution and noise suppression in holographic microscopes.