Computational Economics (2023) 61:317–340
https://doi.org/10.1007/s10614-021-10209-3

Unfolding Beijing in a Hedonic Way

Wei Lin[1] · Zhentao Shi[2,3] · Yishu Wang[3] · Ting Hin Yan[3]

Accepted: 27 September 2021 / Published online: 28 October 2021

Abstract

The housing market is of tremendous importance to the Chinese economy. Housing prices in a metropolitan like Beijing are determined not only by the structural attributes of housing units, but also by the externality stemming from local amenities and their investment potential. Meanwhile, the traditional hedonic pricing model fails to capture the latter two aspects due to its inability to capture the spatial and temporal dimensions of the housing market. In this paper, we augment the traditional model by introducing machine learning algorithms that can handle the additional complexity arising from time and space. Using a transaction-level dataset of housing prices in Beijing, we compare the performance of random forest and Gradient Boosting Machine (GBM) with Ordinary Least Squares (OLS), k-Nearest Neighbor (KNN) and local polynomial. We find that GBM significantly outperforms the other methods in spatial prediction and sequential forecast. GBM's superior predictive capacity indicates the potential of machine learning techniques in reducing the costs of real estate appraisal, alleviating information asymmetry in the housing market, and improving people's welfare given the implications of home ownership on economic inequality and social mobility.

Keywords Boosting · Housing price · Nonlinear estimation · Machine learning · Prediction · Regression trees

Shi acknowledges the financial support from the Hong Kong Research Grants Council No.14500118.

✉ Zhentao Shi
zhentao.shi@gatech.edu

[1] Institute for Economic and Social Research, Jinan University, Guangzhou 510632, People's Republic of China

[2] School of Economics, Georgia Institute of Technology, 205 Old C.E. Building, 221 Bobby Dodd Way, Atlanta, GA 30332, USA

[3] Department of Economics, The Chinese University of Hong Kong, 928 Esther Lee Building, Sha Tin, New Territories, Hong Kong SAR, People's Republic of China

And then there were shots of the city as a whole, shots that included both faces of the city during the Change: earth flipping, revealing the other side studded with skyscrapers with sharp, straight contours; men and women energetically rushing to work; neon signs lighting up the night; towering apartment buildings, cinemas, nightclubs full of beautiful people. But there were no shots of where Lao Dao worked.
—*Folding Beijing* by Jingfang Hao; translated by Ken Liu

1 Introduction

Jingfang Hao, Ph.D. in economics and the author of *Folding Beijing*, gives a compelling account of the social inequality in Beijing in her award-winning ultra-unreal fiction which we quote above. In the real world, however, it is home ownership that divides Beijing residents into folds. China terminated its welfare-based housing allocation system and switched to a market-based one in 1998. Since then, housing has grown into the largest asset class for urban residents in China. An apartment has become not only a dwelling unit, but also a means of storing wealth. This is especially true for those residing in China's first-tier cities such as Beijing, where strong speculative demand has sent house prices into an ascending spiral in the aftermath of the 2008 financial crisis (Guo and Huang 2010).

Besides investment motives, housing affects China's economic inequality through its impact on people's access to public goods. In China, the compulsory education system assigns students to schools based on their residential districts. As a result, the housing prices in districts where schools are of high quality have been driven up (Li and Fu 2010; Feng and Lu 2013; Wen et al. 2017). This in effect penalizes students from less privileged economic backgrounds and is detrimental to social mobility. Furthermore, as China's health insurance is based on *hukou*,[1] the level of public health services available also depends on where one resides. Moreover, the housing market is essential to local governments' public finance since the revenue from the sale of land lease rights is the main source of income to support local governments' debt servicing payment (Ding 2003; Hsing 2010; Wu et al. 2012; Lu and Sun 2013), hinting the critical role that housing prices play in macroeconomic stability.

Many empirical studies have focused on the housing prices in Beijing. Examples include the impact of new land and residential property policies on housing prices (Zheng and Kahn 2008), the implications of an emerging market for "green" real estate on housing prices (Zheng et al. 2012b), and the influence of increased government investment in local public goods on the housing market (Zheng and Kahn 2013), to name just a few. Nonetheless, while these papers look at causal effects of policy changes on Beijing's housing prices, none touches upon the important topic of devising accurate housing price predictions, which is the focus of this paper.

[1] *Hukou* is China's household registration system, which requires people to live and work where they have official permission to do so (Liu 2005).

As prices of individual properties vary drastically on account of some unique characteristics they possess, being able to accurately predict housing prices is central to improving economic stability and to making better informed decisions in Beijing's housing market. To this end, this paper considers the hedonic pricing model put forward by Rosen (1974), where the author defines hedonic prices as "the implicit values of attributes that are revealed to economic agents from observed prices of differential products and the specific amounts of characteristics associated with them." This definition suggests that real estate properties can be viewed as heterogeneous goods, each featuring a distinctive package of characteristics pertaining to location, housing structure and amenities found in the neighborhood. The hedonic pricing model has been adopted widely in housing research, see, *inter alia*, Can (1992), Berry et al. (2003), Ong et al. (2003).

To predict housing prices, we start with using Ordinary Least Squares (OLS) to estimate a linear hedonic model. In the literature, OLS has been widely employed in constructing housing price indices and analyzing the determinants of housing prices (Sirmans et al. 2005). Although OLS is computationally straightforward and easy to understand, it suffers from inflexibility as it imposes linearity restrictions on the model.

Trained as economists, we are aware of the drawbacks of OLS for this prediction task. Besides the attributes of an apartment, which naturally compose the first set of variables of choice in a hedonic model, the variety of amenities of a metropolitan like Beijing must also be priced in when transactions are conducted. These amenities include, but are not limited to, convenience of transportation, availability of good schools and quality healthcare services, and prospect of future development in the surrounding districts. Furthermore, transaction prices are also influenced by macroeconomic factors such as the return to real estate investment relative to other asset classes, which is especially relevant given the low interest rate and quantitative easing over the past decade. It is in capturing these nuanced factors where a linear model falls short. One can enumerate many more factors without exhausting the full list, let alone the infeasibility to collect such a large amount of data. Fortunately, these factors can be encapsulated into two additional dimensions to augment the standard hedonic model, namely a spatial dimension and a temporal dimension. The spatial dimension is designed to capture the externality from neighborhood amenities (e.g. schools, hospitals, and the public transportation networks, etc.), whereas the temporal dimension seeks to summarize the financial calculations an economic agent makes when a transaction is conducted.

As a first attempt to incorporate spatial and temporal factors, we proceed with using *k*-Nearest Neighbor (KNN) and local polynomial methods on a partial linear model, which sits in the middle ground between a linear method and a machine learning algorithm. The partial linear model (Robinson 1988) is one of the simplest semiparametric methods in practice, for it mitigates the curse of dimensionality when there are a multitude of factors influencing the dependent variable. Anglin and Gencay (1996) apply the partial linear model to estimate a hedonic pricing model using data from the UK. On the other hand, KNN (Fix and Hodges 1951) has been extensively applied in nonparametric density and regression estimation, and has set foot on hedonic pricing models in housing research (Pace 1993). Several studies

have been conducted to analyze the performance of KNN using different distance measures. The results, however, are at best mixed. Therefore, in this paper, we present the KNN estimation models and results using various popular distance metrics such Euclidean distance and Manhattan distance. The earliest work on local polynomial regression dates back to that of Nadaraya (1964) and Watson (1964). The local polynomial regression is a technique that fits a polynomial on a "local" subset of the data. Local polynomial can improve the function estimation in regions with sparse observation (Härdle et al. 2004), a problem commonly encountered with the KNN estimation. In terms of housing research, Meese and Wallace (1991) are among the pioneers who adopt local polynomial techniques in estimating hedonic pricing equations. Using KNN or local polynomial to estimate the nonparametric component of the partial linear model offers extra flexibility in functional forms relative to OLS.

In terms of predictive accuracy, modern machine learning algorithms often possess an edge over conventional parametric and nonparametric methods. In this paper, we utilize off-the-shelf machine learning methods based on regression trees. Regression trees, also known as decision trees, are popularized by Breiman et al. (1984) and have witnessed rapid growth in the last decade. Regression trees fit almost every kind of traditional statistical models. Interpretability of the tree structure is a strong reason for their popularity among practitioners, but so are good prediction accuracy, fast computation speed, and wide availability of software. In spite of these advantages, regression trees have several limitations. They are prone to overfitting and the splitting rules are strongly dependent on the training data, which means a small change in the training data may lead to very different trees. Furthermore, they are not good at approximating smooth functions such as a straight line.

To bypass the limitations of standard regression trees and enhance the predictive performance, we experiment with Random Forest (RF) and Gradient Boosting Machine (GBM). Proposed by Breiman (2001), RF combines several randomized decision trees and aggregates their predictions by averaging. It is based on the bagging algorithm and uses ensemble learning techniques. GBM (Friedman 2001) adds at each step a new decision tree that best reduces the loss function. It is capable of regularizing the variance by combining many decision trees. Although RF improves on single decision trees, its prediction accuracy of RF on complex problems is usually inferior to gradient-boosted trees. Many empirical examples have shown GBM beats other machine learning methods and semiparametric and parametric methods (Cheng et al. 2021; Golden et al. 2019; Ogutu et al. 2011). Our empirical results also confirm such relative performance between GBM and RF.

The data used in this paper is extracted from *Lianjia*, a Chinese real estate brokerage company founded in 2001. This paper uses out-of-sample R-squared (OOS-R^2) and Root Mean Prediction Squared Error (RMPSE) to measure the predictive power of different estimation methods. Furthermore, RMPSE shares the same monetary unit of the dependent variable and thus is easy to interpret.

With the advent of big data and access to advanced computing methods, we are not the first team who invoke machine learning methods in predicting housing

prices. Early work includes Park and Bae (2015), who use classical machine learning algorithms for classification on housing in Fairfax County, Virginia, United States, and Lim et al. (2016), who use neural networks to predict condominium prices in Singapore. More recently, a few studies emerge to compare the performance of RF approach with conventional OLS. For instance, Hong et al. (2020) conduct a comparative study, using RF to appraise residential property values in the district of Gangnam, South Korea. Čeh et al. (2018) compare the predictive power of RF with OLS in predicting apartment prices in Ljubljana, the capital of Slovenia. However, these papers treat the prediction machinery as a set of mechanical routines but do not attempt to take economics into their considerations. As a consequence, the economic implications of their findings remain unclear. In addition, although Park and Bae (2015), Lim et al. (2016), Hong et al. (2020) and Čeh et al. (2018) employ some housing attributes in their prediction models, they do not consider flexible functions that allow for time and space, which turn out to be critical factors in our analysis.

With a hedonic pricing model backed by economic principles, we find that GBM reaps substantial benefits in reducing RMPSE. In particular, relative to the OLS benchmark, a one-standard-deviation improvement in spatial prediction amounts to 3.9 years of income for an average Beijing resident, while the corresponding figure for sequential forecast is 2.3 years of income. These suggest that accurate housing price predictions would help *unfold* Beijing in light of the profound impact of home ownership on people's welfare.

The rest of the paper is organized as follows. Section 2 describes the data. Section 3 elaborates on the different prediction models and estimation algorithms used in spatial prediction, followed by the empirical results. Section 4 explores sequential forecast and how it can facilitate transaction decisions. Section 5 concludes. The data cleaning process and a discussion of the spatial autoregressive models (SAR) are relegated to the Appendix. Furthermore, our data and code are hosted on `Github`[2] to enable replications of our work.

2 Data

Our raw data from *Lianjia* is publicly available.[3] *Lianjia* is one of the biggest brokers in China's housing market. It operates an online real estate service platform, through which listing information is circulated to match buyers and sellers, in lieu of a Multiple Listing Service system that is more commonly found in North America and other parts of the world. On the other hand, obtaining accurate housing price data in China is fraught with difficulties as many buyers and sellers report fake transaction prices to local housing authorities (Wu et al. 2014) and real estate contracts that contain actual transaction details are confidential. Therefore, *Lianjia*'s online listing information provides a better trade-off between the accuracy,

[2] https://github.com/ishwang1/Beijing-Housing-prediction.

[3] https://www.kaggle.com/ruiqurm/lianjia.

timeliness and accessibility of housing price information among different data sources (Wang et al. 2018).

Each row of our dataset contains an apartment's transaction price along with a list of its attributes.[4] Most transactions were made during 2011–2017, with a small fraction of them dated either January 2018 or as early as 2009–2010. Table 2 in the Appendix enumerates the variables in the original dataset. After removing observations that contain missing values and outliers, which is common in electronically recorded data, we maintain an effective sample size of 120,403. This is still considerably larger than most sample sizes in conventional econometric studies. The loss of observations is mainly caused by missing DOM (number of active days on the market), in particular among those traded before 2014.

The variables are summarized in Table 3. We choose price, whose unit is Chinese Yuan per square meter (CNY/m^2), in its level form as the dependent variable for direct economic interpretability. The set of explanatory variables x_i is a long vector of 39 variables, including 11 continuous or integer variables (See Table 3(a)) and 28 dummy variables (See Table 3(b) and (c)). As will be discussed in Sect. 3, a few variables are excluded from x_i, namely, the straight line distance to the city center (Tiananmen Square) $\texttt{distc}_\texttt{i}$, daily (integer) time variable $\texttt{t}_\texttt{i}$, year-quarter dummies $\{\texttt{YQ}_\texttt{i}^\texttt{l}\}_{\texttt{l}=1}^{\texttt{q}}$, and geographic coordinates longitude $\texttt{Lng}_\texttt{i}$ and latitude $\texttt{Lat}_\texttt{i}$.[5]

Time and space are the key variables of interest in our study and their effects on house prices are visualized in Figs. 1 and 2, respectively. Figure 1 plots the average housing price against time, with the X-axis being the time of transaction and the Y-axis being either the monthly or quarterly version of average $\texttt{price}_i$. The figure shows that the two average prices exhibit a nonlinear trend over time. Despite a general upward trend until the beginning of 2017, a simplistic linear trend model would miss the nuances of short-term price fluctuations. Figure 2 plots the effect of location on housing prices over time, with the X-axis being the longitude and the Y-axis being the latitude in each subgraph. Apartments of the highest prices are concentrated in the city center, followed by those in the northwest of the city thanks to the clustering of top Chinese universities like Peking University and Tsinghua University, along with the technology hub Zhongguancun in this region. Intuitively, location, represented by longitude and latitude, does not affect housing prices in a linear fashion.

3 Spatial Prediction

In this section, we compare the performance of different models and estimation methods on spatial prediction. To do so, we first split the whole sample randomly into a training sample that accounts for 75% of all observations, denoted by $\mathcal{T}$, and a

[4] An example is available at https://bj.lianjia.com/chengjiao/101084782030.html.

[5] The logarithmic form, $\log(\texttt{price}_i)$, is also commonly used in hedonic housing price models (Owusu-Ansah 2011). When we use $\log(\texttt{price}_i)$ as the dependent variable as robustness checks, the results from all estimation methods (available upon request) are similar to those using $\texttt{price}_i$, and therefore we omit them here.

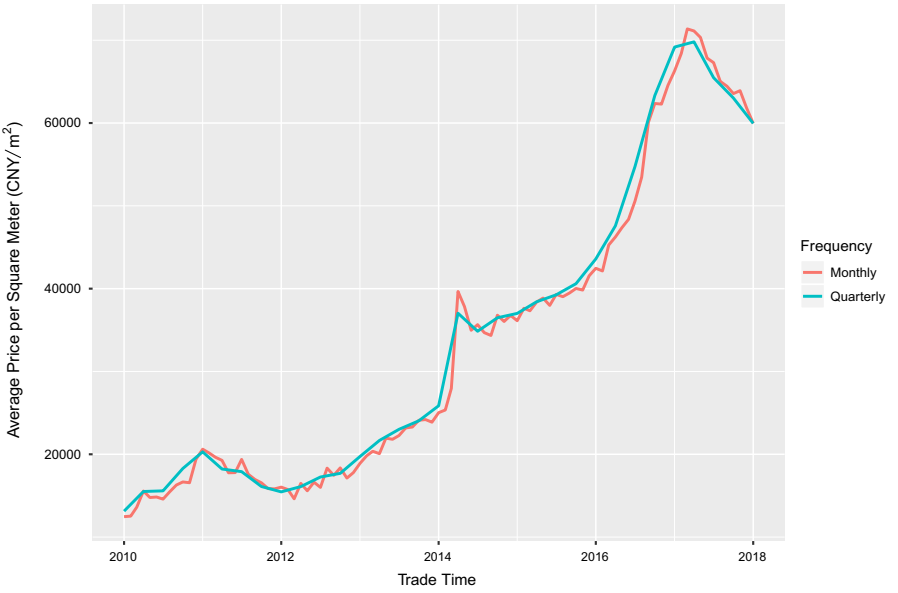

Fig. 1 Average housing prices over time

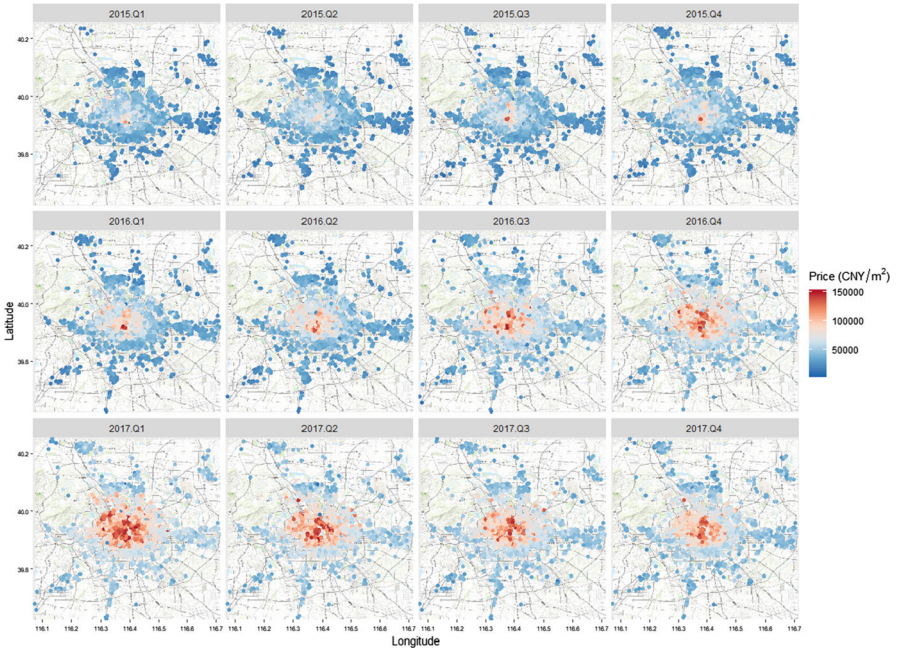

Fig. 2 Housing prices across time and space

testing sample that accounts for the remaining 25% of all observations, denoted by $\mathcal{P}$. Then we estimate various models using the training sample and make predictions in the testing sample. Finally, we evaluate the out-of-sample performance of these models by comparing the predicted and realized housing prices in the testing sample. We use OOS-R^2 and the RMPSE as our evaluation criteria.

We call this exercise *spatial prediction* because it can be viewed as a thought experiment where the apartment in question has been transacted and the researcher is tasked to predict the transaction price. It is in contrast to *sequential forecast* in Sect. 4, where the researcher predicts what the transaction price will be when the apartment is to be sold in the near future.

3.1 Ordinary Least Squares

To begin with, we consider the hedonic regression model

$$y_i = x_i'\beta + \texttt{distc}_\texttt{i}\gamma + \sum_{l=1}^{q} \alpha_l \texttt{YQ}_\texttt{i}^\texttt{l} + \epsilon_i, \tag{1}$$

where the straight line distance to the city center $\texttt{distc}_\texttt{i}$ and the year-quarter dummies $\{\texttt{YQ}_\texttt{i}^\texttt{l}\}_{\texttt{l}=1}^{\texttt{q}}$ are included in addition to the characteristics x_i. The nonlinear relationship between the average housing prices and the time trend shown in Fig. 1 motivates the use of year-quarter fixed effects in place of the continuous time variable $\texttt{t}_\texttt{i}$. Furthermore, as demonstrated in Fig. 2, the effects of location cannot be simply modeled using longitude and latitude as two linear terms, and we therefore exclude geographic coordinates in the hedonic OLS regression. Instead, the effect of location is absorbed, to some extent, by the dummy variables of the 12 Beijing districts in x_i. Besides the district dummies, $\texttt{distc}_\texttt{i}$ also bears geographic information and is included to capture the fact that houses in the city center are more expensive than those in suburban areas. We estimate $\hat{\beta}$, $\hat{\gamma}$ and $\{\hat{\alpha}_l\}_{l=1}^{q}$ from the training sample $\mathcal{T}$, and then plug them into the testing sample $\mathcal{P}$ to obtain out-of-sample predictions. The OLS estimation of the linear regression model serves as the benchmark for comparison.

3.2 K-Nearest Neighborhood

3.2.1 Spatial K-Nearest Neighborhood

A spatial hedonic pricing partial linear model can be specified as

$$y_i = x_i'\beta + \sum_{l=1}^{q} \alpha_l \texttt{YQ}_\texttt{i}^\texttt{l} + \gamma(\texttt{Lng}_\texttt{i}, \texttt{Lat}_\texttt{i}) + \epsilon_i, \tag{2}$$

where $\gamma(\cdot)$ is an unknown function of geographic coordinates. The rest of the model remains linear.[6] Taking conditional expectation given $(\texttt{Lng}_\texttt{i}, \texttt{Lat}_\texttt{i})$, we have

[6] In theory, it is desirable to consider a fully nonparametric specification $y_i = f(x_i, \{\texttt{YQ}_\texttt{i}^\texttt{l}\}_{l=1}^{q}, \texttt{Lng}_\texttt{i}, \texttt{Lat}_\texttt{i}) + \epsilon_i$ for an unknown function $f(\cdot)$. In practice, however, such generality

$$E[y_i|\mathtt{Lng_i},\mathtt{Lat_i}] = E[x_i|\mathtt{Lng_i},\mathtt{Lat_i}]'\beta + \sum_{l=1}^{q}\alpha_l E\left[\mathtt{YQ}_\mathtt{i}^\mathtt{l}|\mathtt{Lng_i},\mathtt{Lat_i}\right] + \gamma(\mathtt{Lng_i},\mathtt{Lat_i}). \tag{3}$$

Subtract (3) from (2), we have

$$y_i - E[y_i|\mathtt{Lng_i},\mathtt{Lat_i}] = (x_i - E[x_i|\mathtt{Lng_i},\mathtt{Lat_i}])'\beta + \sum_{l=1}^{q}\alpha_l\left(\mathtt{YQ}_\mathtt{i}^\mathtt{l} - E\left[\mathtt{YQ}_\mathtt{i}^\mathtt{l}|\mathtt{Lng_i},\mathtt{Lat_i}\right]\right) + \epsilon_i. \tag{4}$$

KNN is a popular nonparametric method used to locally estimate the conditional mean, which is the average of its k-nearest observations. To find out these nearest neighbors, the most straightforward distance metric is the Euclidean distance, also known as the L_2 norm or Ruler distance, by which the distance from observation j to i is defined as

$$d_i(j) = \sqrt{(\mathtt{Lng_j} - \mathtt{Lng_i})^2 + (\mathtt{Lat_j} - \mathtt{Lat_i})^2}. \tag{5}$$

An alternative measure is the Manhattan distance, also known as L_1 norm, Taxicab norm, Rectilinear distance, or City block distance, which was conceived by Hermann Minkowski in 19th-century Germany. This distance represents the sum of the absolute differences of the geographic coordinates:

$$d_i(j) = |\mathtt{Lng_j} - \mathtt{Lng_i}| + |\mathtt{Lat_j} - \mathtt{Lat_i}|. \tag{6}$$

Hence i's k-nearest neighborhood is the set $J_i = \{j \in \mathcal{T} : j \neq i, d_i(j) \leq d_i^{(k)}\}$, where $d_i^{(k)}$ is the k-th smallest value in $\{d_i(j) : j \in \mathcal{T}, j \neq i\}$.

We call this method *spatial* KNN, for the nonparametric function $\gamma(\cdot)$ contains only spatial information. Let $\overline{y}_{\mathrm{S},i} = k^{-1}\sum_{j\in J_i} y_j$ be the spatial average of i's k-nearest neighborhood, and $\overline{x}_{\mathrm{S},i}$ and $\overline{\mathtt{YQ}}_{\mathtt{S,i}}^{\mathtt{l}}$ are defined similarly. Plugging these estimates of conditional expectations into (4), we run OLS using the regression equation

$$y_i - \overline{y}_{\mathrm{S},i} = (x_i - \overline{x}_{\mathrm{S},i})'\beta + \sum_{l=1}^{q}\alpha_l(\mathtt{YQ}_\mathtt{i}^\mathtt{l} - \overline{\mathtt{YQ}}_{\mathtt{S,i}}^{\mathtt{l}}) + \epsilon_i \tag{7}$$

in the training sample $\mathcal{T}$ to obtain the estimates of $\hat{\beta}$ and $\{\hat{\alpha}_l\}_{l=1}^q$. Then for each i in the testing sample $\mathcal{P}$, we predict the housing price as

Footnote 6 continued
would suffer the curse of dimensionality in finite sample estimation as x_i involves many control variables along with two more continuous regressors $\mathtt{Lng_i}$ and $\mathtt{Lat_i}$. This partial linear specification here is a widely-used approach to balance the generality and dimensionality (Robinson 1988).

$$\widehat{y}_{\mathrm{S},i} = (x_i - \overline{x}_{\mathrm{S},i})'\hat{\beta} + \sum_{l=1}^{q} \hat{\alpha}_l(\mathtt{YQ}_{\mathtt{i}}^{\mathtt{l}} - \overline{\mathtt{YQ}}_{\mathrm{S},\mathtt{i}}^{\mathtt{l}}) + \overline{y}_{\mathrm{S},i}. \tag{8}$$

When choosing the tuning parameter k for spatial KNN, we have to wrestle with the problem of singularity stemming from the lack of variation in the dummy variables in a small neighborhood. For example, when k is small, all nearby homes are located in the same district or near the same subway station, so a column of zeros will be generated in (7) after local demeaning. In this dataset, we find that $k = 200$ avoids the singularity issue on randomly split training sample for Euclidean distance, while $k = 400$ rules out singularity for Manhattan distance. When k is larger, out-of-sample prediction gets worse. On the contrary, when k gets smaller, singularity is likely to emerge.

3.2.2 Spatial-temporal K-Nearest Neighborhood

Rather than choosing a relatively large k, a solution to this singularity problem is to incorporate the temporal dimension in the model. In particular, we can remove the year-quarter dummies and add the continuous time variable $\mathtt{t_i}$ into the unknown function $\tilde{\gamma}(\cdot)$ as

$$y_i = x_i'\beta + \tilde{\gamma}(\mathtt{Lng_i}, \mathtt{Lat_i}, \mathtt{t_i}) + \epsilon_i. \tag{9}$$

The idea behind this solution is that now the neighbors, re-defined by both the spatial and temporal dimensions, are cut into much finer grids. We call this model (9) spatial-temporal partial linear model, due to the presence of $\mathtt{t_i}$ in the unknown nonparametric function. This model can also be estimated by KNN. The only difference from the spatial KNN is the definition of the distance, which now becomes

$$\tilde{d}_i^{\lambda}(j) = \sqrt{(\mathtt{Lng_j} - \mathtt{Lng_i})^2 + (\mathtt{Lat_j} - \mathtt{Lat_i})^2 + [\lambda(\mathtt{t_j} - \mathtt{t_i})]^2} \tag{10}$$

for the Euclidean distance, and

$$\tilde{d}_i^{\lambda}(j) = |\mathtt{Lng_j} - \mathtt{Lng_i}| + |\mathtt{Lat_j} - \mathtt{Lat_i}| + \lambda|\mathtt{t_j} - \mathtt{t_i}| \tag{11}$$

for the Manhattan distance. Here λ is the new tuning parameter prescribing the weight of the temporal dimension relative to the spatial dimension. Similarly, we call this method spatial-temporal KNN and estimate it in the training sample by OLS with the KNN locally demeaned variables, and hence predictions can be made in the testing sample as

$$\widehat{y}_{\mathrm{ST},i} = (x_i - \overline{x}_{\mathrm{ST},i})'\hat{\beta} + \overline{y}_{\mathrm{ST},i}, \tag{12}$$

where $\overline{y}_{\mathrm{ST},i}$ is the spatial-temporal k-nearest neighborhood average. $\overline{x}_{\mathrm{ST},i}$ is defined similarly.

In order to ensure that the training sample makes up 75% of the whole sample, we apply a 4-fold cross validation (CV) on the training sample to determine the optimal tuning parameters k and λ. The optimal set of tuning parameters is the pair

that minimizes the OOS-R^2, which turns out to be $k = 15$ and $\lambda = 0.0005$ for both Euclidean distance and Manhattan distance. The chosen k is substantially smaller than that for the spatial KNN, suggesting enhanced local predictive performance by spatial-temporal KNN. Singularity will be encountered for $\lambda < 0.0005$, because a small λ causes the distance measures specified in (10) and (11) to collapse into that in (5) and (6). In other words, a small λ induces spatial-temporal KNN to degenerate into spatial KNN.

3.3 Local Mean and Polynomial

3.3.1 Spatial Local Mean and Polynomial

Besides the KNN approach, an alternative nonparametric method to estimate the partial linear model (2) is the Nadaraya-Watson (NW) kernel regression, also known as the spatial local mean method. Unlike the KNN approach, where we exploit the information of the k nearest neighbors using equal weighting, the NW kernel regression is essentially a weighted average using kernel functions. Formally, let

$$\overline{y}_{\mathrm{S},i} = \sum_{j \in \mathcal{T}} w_i(j) y_j / \Big(\sum_{j \in \mathcal{T}} w_i(j)\Big),$$

with the weight $w_i(j) = K\Big((\mathtt{Lng_j} - \mathtt{Lng_i})/h\Big) K\big((\mathtt{Lat_j} - \mathtt{Lat_i})/h\big)$, where $K(\cdot)$ is some kernel function for some given bandwidth h which serves as the tuning parameter. For computational simplicity, we choose the Epanechnikov kernel as $K(\cdot)$. Moreover, $\overline{x}_{\mathrm{S},i}$ and $\overline{\mathtt{YQ}}^{\mathtt{l}}_{\mathrm{S,i}}$ are defined similarly. The OLS estimates $\hat{\beta}$ and $\{\hat{\alpha}_l\}_{l=1}^{q}$ of (7) in the training sample $\mathcal{T}$ are consequently the partial linear spatial local mean estimates. Accordingly, for each i in the testing sample $\mathcal{P}$, we can still predict the housing price by using equation (8).

Local polynomial estimation is akin to the kernel method, with the difference being that the former uses polynomial to further capture the local effect. Now let $\gamma(\mathtt{Lng_i}, \mathtt{Lat_i})$ in a partial linear model (2) take the following form with the order of polynomial equaling to one ($p = 1$),

$$\begin{aligned}\gamma(\mathtt{Lng_i}, \mathtt{Lat_i}) = \gamma_0(\mathtt{Lng_i}, \mathtt{Lat_i}) + \gamma_1^{\mathtt{Lng}}(\mathtt{Lng_i}, \mathtt{Lat_i}) \cdot (\mathtt{Lng_j} - \mathtt{Lng_i}) \\ + \gamma_1^{\mathtt{Lat}}(\mathtt{Lng_i}, \mathtt{Lat_i}) \cdot (\mathtt{Lat_j} - \mathtt{Lat_i})\end{aligned} \tag{13}$$

for all $i, j \in \mathcal{T}$. When $j = i$, the above equation degenerates into $\gamma(\mathtt{Lng_i}, \mathtt{Lat_i}) = \gamma_0(\mathtt{Lng_i}, \mathtt{Lat_i})$. In this case, at location i, by Weighted Least Squares (WLS) we have

$$\hat{\gamma}_0(\mathtt{Lng_i}, \mathtt{Lat_i}) = R(Z_i' \Omega_i Z_i)^{-1} Z_i' \Omega_i (y_{\mathcal{T}} - X_{\mathcal{T}} \beta - \mathtt{YQ}_{\mathcal{T}} \alpha),$$

where Z_i is a matrix with all polynomial terms as columns (including the first column of ones for the intercept) in (13) and all $j \in \mathcal{T}$ as rows, $\Omega_i = \mathrm{diag}(w_i(j))$ for all $j \in \mathcal{T}$, $R = (1, 0, 0)$, and $(y_{\mathcal{T}}, X_{\mathcal{T}}, \mathtt{YQ}_{\mathcal{T}})$ is the training sample in matrix form. To estimate β, we simply need to define $\overline{y}_{\mathrm{S},i} = R(Z_i' \Omega_i Z_i)^{-1} Z_i' \Omega_i y_{\mathcal{T}}$, and define $\overline{x}_{\mathrm{S},i}$ and

$\overline{\mathtt{YQ}}^{1}_{\mathtt{S},\mathtt{i}}$ similarly. Thus, we use the training sample $\mathcal{T}$ to run the regression specified in (7). For each i in $\mathcal{P}$, the out-of-sample prediction is achieved by (8). Here local mean estimation can be regarded as a special case of local polynomial method where the order of polynomial equals zero ($p = 0$).

It is well-known that while increasing the order of local polynomial method has the advantage of reducing the bias of local effects, the variance, and thus the total prediction error may increase. Furthermore, as with KNN, while out-of-sample performance declines when h increases, h cannot be too small due to the singularity issue. Hence, when a higher order of p introduces more terms, we must invoke a larger bandwidth h to avoid local singularity, which is another factor contributing to the deterioration of predicting performance. In the reported results, we restrict the order of polynomial to be $p = 0$ and $p = 1$. The bandwidth h is chosen as $h = 0.04$ and $h = 0.07$, respectively, to ensure the best out-of-sample performance and to circumvent the singularity issue.[7]

3.3.2 Spatial-temporal Local Mean and Polynomial

Similarly, by adding a temporal dimension to the local effect as shown in model (9), we derive the spatial-temporal local mean and polynomial methods. Let $\overline{y}_{\mathrm{ST},i}$ and $\overline{x}_{\mathrm{ST},i}$ be similarly defined as their counterparts in spatial local mean and polynomial, with the only difference being the construction of the weights $w_i(j) = K\Big((\mathtt{Lng}_\mathtt{j} - \mathtt{Lng}_\mathtt{i})/h_s\Big)K\big((\mathtt{Lat}_\mathtt{j} - \mathtt{Lat}_\mathtt{i})/h_s\big)K\big((\mathtt{t}_j - \mathtt{t}_\mathtt{i})/h_t\big)$, where h_s is the spatial bandwidth and h_t is the temporal bandwidth.

Furthermore, let $\tilde{\gamma}(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_\mathtt{i})$ take the following form

$$\begin{aligned}\tilde{\gamma}(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_i) = \tilde{\gamma}_0(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_i) &+ \tilde{\gamma}_1^{\mathtt{Lng}}(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_i) \cdot (\mathtt{Lng}_\mathtt{j} - \mathtt{Lng}_\mathtt{i}) \\ &+ \tilde{\gamma}_1^{\mathtt{Lat}}(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_i) \cdot (\mathtt{Lat}_\mathtt{j} - \mathtt{Lat}_\mathtt{i}) \\ &+ \tilde{\gamma}_1^{\mathtt{t}}(\mathtt{Lng}_\mathtt{i}, \mathtt{Lat}_\mathtt{i}, \mathtt{t}_i) \cdot (\mathtt{t}_j - \mathtt{t}_\mathtt{i})\end{aligned} \tag{14}$$

for all $i, j \in \mathcal{T}$. Following the notation of the spatial local polynomial method, let $\tilde{Z}_i$ denote the matrix form of all terms in (14). Hence, at location i, by WLS we have

$$\overline{y}_{\mathrm{ST},i} = R(\tilde{Z}_i' \Omega_i \tilde{Z}_i)^{-1} \tilde{Z}_i' \Omega_i y_{\mathcal{T}},$$

where $\overline{x}_{\mathrm{ST},i}$ is defined similarly. Therefore, for all $i \in \mathcal{P}$, the spatial-temporal local polynomial predictor is exactly the same as (12), where $\hat{\beta}$ is estimated from the training sample $\mathcal{T}$.

For the same reason stated for spatial-temporal KNN, we continue to apply a 4-fold CV on the training sample to determine the optimal bandwidths h_s and h_t for $p = 0$ and $p = 1$. Our results show that for $p = 0$, the best choices of bandwidth are

[7] We also experimented estimation under a higher order of p. For $p = 2$ and $p = 3$, the results are very similar to those with $p = 1$, and when $p > 3$, the performance deteriorates. Those results are available upon request.

$h_s = 0.05$ and $h_t = 30$, and for $p = 1$, the best choices are $h_s = 0.3$ and $h_t = 50$. Cases where $p \geq 2$ are infeasible in our dataset, because the local singularity issue cannot be avoided even for very large bandwidths. The reason lies in that the number of polynomial terms increases too fast when there are three dimensions combined in $\tilde{\gamma}(\cdot)$.

3.4 Random Forest

To evaluate the predictive power of more sophisticated machine learning methods over the linear model and the partial linear model, we first turn to regression-tree based RF. A regression tree algorithm recursively partitions the space of regressors as follows:

1. For each regressor, find a cut point that splits the sample into two partitions that minimize the Sum of Squared Residuals (SSR) computed from fitting the mean value of housing price in each partition.
2. Go through all regressors to find the one that minimizes the SSR, and then split this variable into two dummy variables.
3. Repeat Step 1-2 until it reaches the pre-specified *tree depth*, a tuning parameter regularizing the number of repeated sequential splitting.

In contrast to OLS which only accommodates linear terms, regression trees can split the data directly based on the time variable and geographic coordinates. As a result, we can include x_i, $\mathtt{t_i}$, $\mathtt{Lng_i}$ and $\mathtt{Lat_i}$ to estimate a regression tree. The prediction for a given observation is simply the sample average of the housing prices in the partition where the observation lies in. To highlight the importance of incorporating location information in predicting housing prices, we consider two models, one with longitude and latitude and the other without.

In spite of the flexibility, the regression tree is known to produce unstable paths of selected variables and partition thresholds. In addition, its prediction is sensitive to a small perturbation to the training data. To bypass these limitations, several methods are proposed, for example, Bootstrap Averaging (Bagging), Boosting, and RF. We implement RF as a first attempt. The algorithm of RF goes as follows:

1. Generate a bootstrapped sample by random selection from the training data set with replacement.
2. Randomly select m regressors out of all the regressors.
3. Fit a regression tree with these m explanatory variables and some *tree depth* (number of terminal nodes) on the bootstrapped training sample, and then save the prediction for the target point.
4. Repeat Step 1-3 until it reaches the pre-specified *number of trees*, and lastly obtain the average of predictions from all these trees is the final prediction.

Following the convention in the literature, *the number of trees* is set to be 500. Thus, the remaining tuning parameters for RF are the *number of selected variables m* and *tree depth*. Here we specify the grid for m as $\{10, 11, \cdots, 50\}$ and for the *tree depth* as $\{8000, 8100, \cdots, 12000\}$. To find the optimal set of tuning parameters, we use a

10% subsample and conduct a 5-fold CV on it. Using the 10% subsample, instead of the entire training sample, accelerates the search for the optimal set of tuning parameters, as RF is not sensitive to the sample size. As it turns out, RF without coordinates chooses (29, 10500), while RF with coordinates chooses (30, 9800).

3.5 Gradient Boosting Machine

As mentioned above, another machine learning method to enhance regression tree is boosting. We adopt GBM from the family of boosting methods. A variety of boosting algorithms have emerged in the econometrics literature in recent years (Bai and Ng 2009; Shi 2016; Phillips and Shi 2021). GBM is an off-the-shelf machine learning algorithm which can be implemented as follows:

1. To predict y_i, we use the explanatory variables in the training sample to grow a tree and save the fitted values of each individual i in the training sample, denoted as $\hat{r}_i^{(0)}$. The prediction is $f_i^{(0)} = \alpha \cdot \hat{r}_i^{(0)}$, where $\alpha \in [0, 1]$ is the *shrinkage level*, the second tuning parameter for this method. The iteration index is initiated as $l = 1$.
2. In the l-th iteration, taking $e_i^{(l-1)} = y_i - f_i^{(l-1)}$ as the new dependent variable, we use the same explanatory variables in the training sample to grow a new tree $\hat{r}_i^{(l)}$. The prediction is updated as $f_i^{(l)} \leftarrow f_i^{(l-1)} + \alpha \cdot \hat{r}_i^{(l)}$ and the iteration index is updated as $l \leftarrow l + 1$.
3. Repeat Step 2 until $l > M$, where M is the number of iterations, the third GBM tuning parameter.

Similarly, we apply GBM both with and without coordinates. Likewise, to find the optimal tuning parameters with a lower time cost, we conduct a 5-fold CV on a 10% random subsample, as GBM is not sensitive to the sample size either. Three tuning parameters are involved in GBM, namely, (i) the *tree depth* (number of splits), (ii) the *shrinkage level*, and (iii) the *number of boosting iterations*. We specify the grid for tuning parameters as (i) $\{10000, 10100, \cdots, 20000\}$, (ii) $\{0.001, 0.005, 0.01, 0.05, 0.1\}$, and (iii) $\{10, 15, \cdots, 50\}$, respectively. It turns out that GBM without coordinates chooses $(12000, 0.05, 25)$, while GBM with coordinates shares the same set of optimal tuning parameters except the *tree depth* being 19,400.

3.6 Empirical Results

We estimate in the training sample by OLS, KNN, local mean and polynomial, RF and GBM. OOS-R^2 and RMPSE in the testing sample are reported in Panel A of Table 1. Among these estimation methods, GBM with coordinates produces the highest OOS-R^2 and the lowest RMPSE, while RF with coordinates ranks the second best, which illustrates the significant advantage of utilizing machine learning methods in predicting housing prices. For spatial semiparametric methods, the performances are similar, with spatial-temporal KNN with Manhattan distance ranking the best among these methods. Overall, spatial local polynomials and

Table 1 Prediction performance in testing sample

Method	Panel A: Spatial Pred.		Panel B: Seq. Forecast	
	OOS-R^2	RMPSE	OOS-R^2	RMPSE
OLS	0.898	7546	0.915	6636
Spatial KNN (Euc. Dist.)	0.905	7269	0.916	6587
Spatial KNN (Man. Dist.)	0.903	7338	0.916	6601
Spatial-temporal KNN (Euc. Dist.)	0.935	6018	NA	NA
Spatial-temporal KNN (Man. Dist.)	0.937	5908	NA	NA
Spatial Local Polynomial ($p = 0$)	0.901	7435	0.914	6678
Spatial Local Polynomial ($p = 1$)	0.900	7467	0.914	6669
Spatial-temporal Local Poly. ($p = 0$)	0.927	6381	NA	NA
Spatial-temporal Local Poly. ($p = 1$)	0.912	7021	NA	NA
RF (without Coordinates)	0.948	5387	NA	NA
RF (with Coordinates)	0.951	5224	0.949	5134
GBM (without Coordinates)	0.959	4797	NA	NA
GBM (with Coordinates)	**0.966**	**4335**	**0.957**	**4734**

A boldface number highlights the best performance in each column

The unit of RMPSE is CNY$/m^2$. For Panel A, the entire sample is randomly split into a 75% training sample and a 25% testing sample. For Panel B, observations before November 2017 are used as the training sample, and those in December 2017 are used as the testing sample

spatial-temporal local polynomials perform worse than their counterpart KNN estimates.

As RMPSE is denominated in CNY$/m^2$, the same unit as housing prices, the results suggest that one standard error in GBM prediction (without coordinates) is 2,749 CNY$/m^2$ lower than that of OLS, and 1,111 CNY$/m^2$ lower than that of spatial-temporal KNN with Manhattan distance, respectively. The inclusion of coordinate information further reduces GBM's RMPSE by 462 CNY$/m^2$, indicating that location is a critical factor in spatial prediction. Unsurprisingly, OLS performs the worst, as the linear terms cannot capture the sophistication and subtlety of hedonic price determination. While semiparametric methods are better at addressing the spatial and temporal complexity than OLS, they are not nearly as good as the machine learning methods.

To provide an economic interpretation of the reduced RMPSE, for an average-sized apartment of 82.5m^2, the 3211 CNY$/m^2$ improvement in prediction associated with GBM (with coordinates) over OLS amounts to a whopping CNY264,908. With an average annual disposable income of CNY67,392 in Beijing, reported by China's National Statistics Bureau in 2019, the gain from GBM is equivalent to 3.9 years of disposable income for an average resident of Beijing. Similarly, GBM's improve-

ment relative to spatial-temporal KNN (Manhattan distance) is equivalent to 2 years of average annual income. The increase in predictive accuracy achieved by the machine learning method significantly reduces the information asymmetry in the housing market and makes property transactions more transparent for both buyers and sellers.

GBM with coordinates is based on the regression trees. These trees partition the coordinates and generate a map of geographic location premium in Beijing. The heat map presented in Fig. 3 is constructed with continuous variables in x_i fixed at their means and the categorical or dummy variables in x_i set to their modes. The GBM model predicts on a grid raster spanned by longitude and latitude with an interval of 0.01 degree. This map represents the value-added by the location, net of other attributes of the apartments. The highest premium emerges around the university and technology clusters in Haidian District, which is even darker in color than the city center. This is in contrast to Fig. 2, where housing in the city center is more expensive than that in Haidian District. Notice that Fig. 3 is conditional on features of a representative apartment, whereas Fig. 2 shows unconditional housing prices.

The difference between the two figures reflects the paradox of *school-zone apartments*. Although the amenities of apartments near good school districts are

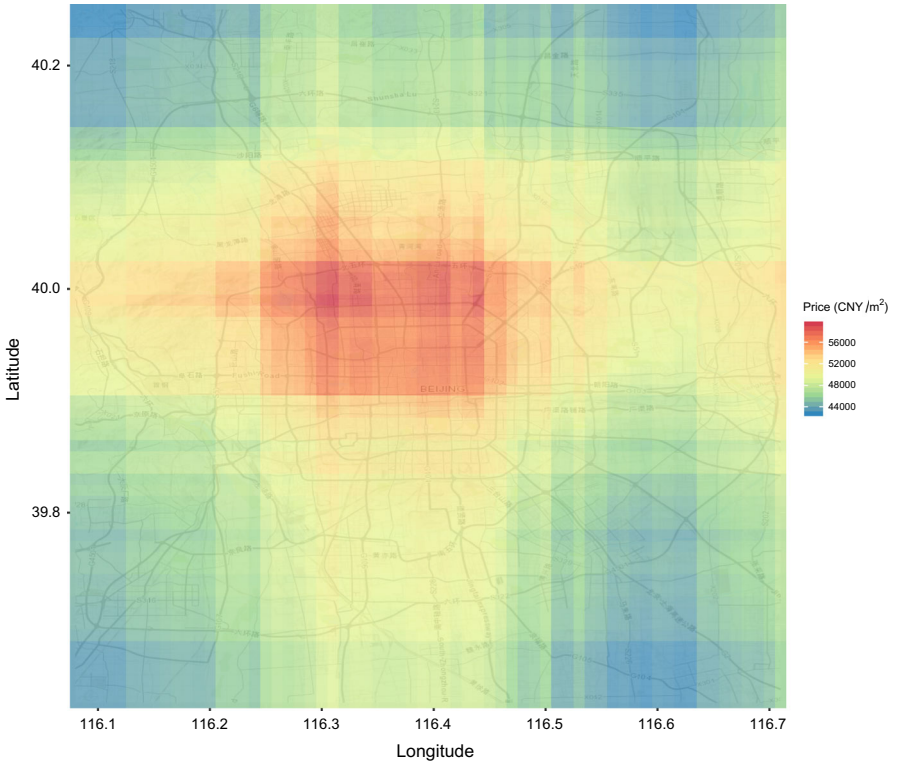

Fig. 3 GBM prediction on coordinate raster

poorer, they still enjoy a premium due to the high demand for schools nearby. Chinese parents' zeal for children's schooling is a well-known national phenomenon. Despite the relative affluent educational resources in Beijing, parents still rush to snap old and shabby apartments in order to secure a position for their children in one of the top schools.[8] While our current approach summarizes all location-specific amenities, which is more general and comprehensive than singling out schooling premium, we may consider explicitly modeling school information in future studies.

4 Sequential Forecast

In this section, we compare the sequential forecast performance of the five methods described earlier. Being able to forecast future prices is directly relevant to the transaction decisions. To implement the sequential forecast procedure, we set all the data before November 2017 as the training sample, and those in December 2017 as the testing sample. In sharp contrast to spatial prediction, the training sample and the testing sample here are partitioned over time. Consisting of 2,972 observations, the testing sample is relatively small. To capture temporal heterogeneity, we assume that for every three consecutive months (that is, a quarter) the time fixed effect remains the same. This implies that we can estimate the time fixed effect of the testing sample, using the observations in the fourth quarter of 2017 (that is, October and November 2017) in the training sample.

OLS follows the same procedure as before, with the only difference being the way the sample is partitioned. Spatial KNN also follows the same procedure with $k = 200$ for Euclidean distance and $k = 500$ for Manhattan distance to circumvent singularity. For local polynomial, the same bandwidths as in spatial prediction are chosen. Spatial-temporal KNN and Spatial-temporal local polynomials are unavailable because we do not have a spatial-temporal neighborhood for observations in the testing sample. Regarding RF and GBM, we only consider the model with geographic coordinates. We drop the time variable $\mathtt{t}_\mathtt{i}$, as the testing sample $\mathtt{t}_\mathtt{i}$ is out of the range in the training sample. Rather, we include the year-quarter dummy variables $\{\mathtt{YQ}_\mathtt{i}^\mathtt{l}\}_{\mathtt{l}=1}^{q}$ in the model. To attain the data-driven tuning parameters, we further carve out the observations in September 2017 as validation data and use the sample before that cut-off as the training data. Based on the validation data, RF chooses $m = 38$ and *tree depth*=10,100 from the same grid as in the spatial prediction. For GBM, 16,900 *tree splits*, a *shrinkage level* of 0.05, and 15 *boosting iterations* are determined to be the optimal tuning parameters in the grid system.

Panel B in Table 1 reports the sequential forecast performance. Notice that for OLS, semiparametric methods, and RF, their RMPSEs are smaller than those in the spatial prediction. This is caused by the smaller heterogeneity that can be found in the testing sample when data are concentrated within a single month. Compared

[8] By employing a conventional DID technique along with a boundary discontinuity analysis, Zheng et al. (2012a) finds that Beijing parents' additional willingness to pay for a flat in the top 5% of school districts range between $24,452 and $54,186.

with OLS, spatial KNN with Euclidean distance only poses a small gain of 49 CNY$/m^2$ in sequential forecast, whereas other semiparametric methods even show worse predictive accuracy. In contrast, GBM again performs the best, slashing the RMPSE by 1902 CNY$/m^2$ relative to OLS. For an average-sized apartment of $82.5m^2$, a reduction of one standard deviation amounts to 2.3 years of income, given Beijing's average annual disposable income of CNY67,392. Further, GBM outperforms RF, which is the second best predictor, by 400 CNY$/m^2$, corresponding to 0.5 years of income.

In both spatial prediction and sequential forecast, we observe sizable reduction in uncertainty by machine learning techniques, especially GBM, relative to linear regressions or partial linear semiparametric methods. Since both RF and GBM are based on regression trees, they inherit the advantage of data-driven control variable interactions, which generate much flexible nonlinear patterns. Moreover, the empirical observation that GBM outperforms other methods lies in that GBM essentially adds predictors to the ensemble and follows the sequence in correcting preceding predictors to arrive at an accurate predictor at the end of the procedure. It aims at fitting a new predictor in the residual errors committed by the preceding predictor. Utilizing the gradient descent to pinpoint the challenges in the learners' predictions used previously, the previous error is highlighted.

Unlike RF which involves random resampling of the data, GBM is a deterministic algorithm. Theoretical comparison between the two methods is out of the scope of this paper though, the econometric literature has witnessed theoretical investigations (Phillips and Shi 2021; Shi and Huang 2021) for the boosting-type procedures in asymptotic frameworks. Accumulating empirical evidence, such as Cheng et al. (2021), also demonstrate GBM's strong performance relative to RF in low-frequency macroeconomic indicator forecasts.

By combining one weak learner with the next learner, the error is reduced significantly over time. The strong predictive performance of machine learning methods supports the prospect that they may help economic agents to make better-informed financial decisions. Given the importance of housing to people's welfare in Beijing, this new set of statistical techniques, fused with sound economic modeling, may bring profound social benefits.

5 Conclusion

Based on a hedonic pricing model augmented by spatial and temporal factors, five estimation methods are compared for their ability to predict Beijing housing prices. In both spatial prediction and sequential forecast, we find machine learning methods such as RF and GBM to be superior to OLS and partial linear semiparametric methods like KNN and local polynomial. Our study shows that an off-the-shelf machine learning algorithm, especially GBM, can enhance predictive accuracy significantly. Therefore, machine learning techniques hold the potential to reduce the costs of real estate appraisal, ameliorate information asymmetry in the housing market, and accelerate mortgage loan decisions. As a possible industrial application, this research paves the way for an online platform that provides housing price

predictions and forecasts. While we focus on a specific city Beijing, with available data the methods are readily applicable to other cities in the world.

Appendix

See Tables 2, 3.

Table 2 Variables in the dataset

Variable name	Description	Example
Lng	Longitude using the BD09 protocol	116.475489
Lat	Latitude using the BD09 protocol	40.01952
Cid	Community ID	1111027376244
tradeTime	Date of transaction	2016-08-09
DOM	Number of active days on the market	1464
followers	Number of people follow the transaction	106
totalPrice	Total price	415.0 (10k CNY)
price	Average price by square meter	31680 (CNY/m^2)
square	Area of house in square meter	131.0 (m^2)
livingRoom	Number of living room	2
drawingRoom	Number of drawing room	1
kitchen	Number of kitchen	1
bathroom	Number of bathroom	1
floorType	Unknown, ground, low, middle, high, roof	high
floorTotal	Total floor number of the building	26
buildingType	Tower, bungalow, combination of plate and tower, plate	tower
constructionTime	Year of construction	2005
renovationCondition	Other, rough, simplicity, hardcover	simplicity
buildingStructure	Unknown, mixed, brick and wood, brick and concrete, steel, steel-concrete composite	steel-concrete composite
ladderRatio	How many ladders a resident have on average	0.217
elevator	Have or not have elevator	1
fiveYearsProperty	If the owner owns the property for less than 5 years	0
subway	If subway nearby	1
district	District ID	7
communityAverage	Average price of community	56021 (CNY/m^2)

Table 3 Summary statistics

Variable	Mean	S.D.	Min	1st Q.	Median	3rd Q.	Max
(a) Continuous variables and integer variables							
price	54314.39	23716.75	6272	36912	49866	67397	150000
square	82.46	36.15	10.70	57.87	73.57	97.28	586.00
livingRoom	2.02	0.77	0	1	2	2	7
drawingRoom	1.14	0.50	0	1	1	1	5
kitchen	0.99	0.12	0	1	1	1	3
bathroom	1.18	0.42	0	1	1	1	6
floorTotal	13.26	7.79	1	6	11	19	63
ladderRatio	0.38	0.18	0.01	0.25	0.38	0.50	3.33
age	16.70	9.02	−2	10	15	23	67
communityAverage	63472.46	22077.32	14773	46204	58976	75789	183109
DOM	33.98	52.64	1	1	13	46	1464
followers	31.30	47.16	0	5	16	39	1143

(b) Dummy variables

Variable	Mean	Variable	Mean
fType.ground	0.08	bType.bungalow	4.98×10^{-5}
fType.low	0.20	bType.plate	5.55×10^{-1}
fType.middle	0.38	bType.tower	2.59×10^{-1}
fType.high	0.22	bType.plate&tower	1.86×10^{-1}
fType.roof	0.12	bStructure.brick&wood	1.66×10^{-4}
		bStructure.brick&concrete	4.20×10^{-2}
rCond.rough	0.03	bStructure.steel&concrete	5.88×10^{-1}
rCond.simplicity	0.40	bStructure.steel	6.56×10^{-4}
rCond.hardcover	0.57	bStructure.mixed	3.69×10^{-1}
elevator	0.58	5YrsProperty	0.64
subway	0.60		

(c) District summary

ID	Name	Obs.	Avg. Price	ID	Name	Obs.	Avg. Price
1	Dongcheng	6556	76667	8	Haidian	13955	68708
2	Fengtai	11310	46819	9	Shijingshan	3955	42986
3	Yizhuang	1084	36806	10	Xicheng	11580	85988
4	Daxing	5677	37840	11	Tongzhou	4780	37746
5	Fangshan	1796	30760	12	Mentougou	1277	25878
6	Changping	14456	36889	13	Shunyi	3057	32048

Table 3 continued

(c) `District` summary

ID	Name	Obs.	Avg. Price	ID	Name	Obs.	Avg. Price
7	`Chaoyang`	40920	54452				

In addition to these continuous variables, `Lng` is longitude, `Lat` is latitude, and `t` is time (daily)

`floorType` is abbreviated as `fType`, `renovationCondition` as `rCond`, `buildingType` as `bType`, `buildingStructure` as `bStructure` and `fiveYearsProperty` as `5YrsProperty`

Data Cleaning

Data removal To remove missing variables and outliers from the data extracted from *Lianjia*, we take the following two steps:

1. First, we remove all observations with "NA" or with any other forms of missing coding. For instance, we remove observations with "unknown" in `floorType` and `buildingStructure`; we also remove observations with "other" in `renovationCondition`.
2. Second, we remove blatant manual input errors. For example, we remove observations where `bathroom` has a value of "1999-2009", which obviously refers to a time period. We also remove observations with `ladderRatio` greater than 10^7.

Data construction and transformation While most variables in this data set can be used directly in our estimation models, some variables need to be constructed or transformed.

1. We transform the categorical variables into a series of dummy variables. That is, if an observation falls into one category, we set the corresponding dummy variable to one. The following variables are transformed in this fashion: (1) `floorType`, (2) `renovationCondition`, (3) `buildingType`, (4) `buildingStructure`, and (5) `district`.
2. We construct the variable `age` by taking the difference in years between `constructionTime` and `tradeTime`.
3. We transform `tradeTime` (originally formatted as date) into a sequence of integers, denoted by $\mathtt{t} = 1, 2, \cdots, \mathtt{T}$, where T represents the last date in the sample.
4. We construct a series of year-quarter dummy variables from `tradeTime`, where each represents a year-quarter combination of the date of transaction. The year-quarter combination is denoted by $\mathtt{YQ}^{\mathtt{l}}$, where `l` stands for `l`-th year-quarter combination in the data.

Spatial Autoregressive Models

When we consider spatial dataset as ours, it is tempting to apply spatial econometric methods. Spatial Autoregressive (SAR) model is the most commonly used specification in spatial econometrics. However, with big data, SAR loses its appeal in modeling and prediction. Consider the SAR model in matrix form:

$$y = \rho Wy + X\beta + \mathtt{YQ}\alpha + \epsilon.$$

where W is the adjacent matrix and ρ is the associated scalar coefficient. Give the large training sample, it is computationally cumbersome to estimate this SAR model in the training sample, where $W_{\mathcal{T}}$ is an $n_{\mathcal{T}} \times n_{\mathcal{T}}$ matrix and $n_{\mathcal{T}}$ is the sample size of $\mathcal{T}$. Compared with maximum likelihood estimation, the instrumental variable approach is more computationally friendly due to its linear nature. Although imposing a sparse structure onto $W_{\mathcal{T}}$ will help alleviate computational burden, when dealing with big data SAR is infeasible for the out-of-sample (OOS) prediction. To see this, the OOS prediction $\hat{y}_{\mathcal{P}}$ conceptually goes as

$$\hat{y} = \begin{pmatrix} \hat{y}_{\mathcal{T}} \\ \hat{y}_{\mathcal{P}} \end{pmatrix} = (I_n - \hat{\rho}W)^{-1}(X\hat{\beta} + \mathtt{YQ}\hat{\alpha}),$$

where $W = \begin{pmatrix} W_{\mathcal{T}} & W_{\mathcal{TP}} \\ W_{\mathcal{PT}} & W_{\mathcal{P}} \end{pmatrix}$ which is an $n \times n$ matrix, and n is the total number of observations of both the training sample and the testing sample. It demands 116GB (= 64 bit floating-point $\times 120403^2$) to store such a large square matrix W, where 120, 403 is our sample size. The need of memory is beyond a standard single computer or a small computing node, not to mention the multiple copies generated in the intermediate steps of linear and nonlinear operations such as matrix multiplication and inversion. Sparsity of W itself will not help either, because even though $I_n - \hat{\rho}W$ can be a sparse matrix, its inverse remains dense in general. We are unable to apply SAR to this big dataset for working with large dense matrices at such a magnitude is infeasible given our computational resources.

Funding Shi acknowledges the financial support from the Hong Kong Research Grants Council No.14500118.

Data availability Raw dataset is publicly available at https://www.kaggle.com/ruiqurm/lianjia.

Code availability Github repository for replication at https://github.com/ishwang1/Beijing-Housing-prediction.

Declarations

Conflicts of interest None of the coauthors have conflict of interest or completing interests.

References

Anglin, P. M., & Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics, 11*(6), 633–648.

Bai, J., & Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics, 24*(4), 607–629.

Berry, J., McGreal, S., Stevenson, S., Young, J., & Webb, J. (2003). Estimation of apartment submarkets in dublin, ireland. *Journal of Real Estate Research, 25*(2), 159–170.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Statistics/Probability Series. Belmont, California, U.S.A.: Wadsworth Publishing Company.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics, 22*(3), 453–474.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-information, 7*(5), 168.

Cheng, K., Huang, N., & Shi, Z. (2021). Survey-based forecasting: To average or not to average. In *Behavioral Predictive Modeling in Economics*, pp. 87–104. Springer.

Ding, C. (2003). Land policy reform in china: Assessment and prospects. *Land Use Policy, 20*(2), 109–120.

Feng, H., & Lu, M. (2013). School quality and housing prices: Empirical evidence from a natural experiment in shanghai, china. *Journal of Housing Economics, 22*(4), 291–307.

Fix, E., & Hodges, L. (1951). Discriminatory analysis: nonparametric discrimination: consistency properties. *Report No. 4, USAF School of Aviation Medicine, Randolph Field, Texas, Feb*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics,* 1189–1232.

Golden, C. E., Rothrock, M. J., Jr., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting listeria spp. prevalence in the environment of pastured poultry farms. *Food Research International, 122,* 47–55.

Guo, F., & Huang, Y. S. (2010). Does "hot money" drive china's real estate and stock markets? *International Review of Economics & Finance, 19*(3), 452–466.

Härdle, W. K., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Newyork: Springer.

Hong, J., Choi, H., & Kim, W.-S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in south korea. *International Journal of Strategic Property Management, 24*(3), 140–152.

Hsing, Y.-T. (2010). *The Great Urban Transformation: Politics of Land and Property in China*. New York: Oxford University Press.

Li, X., & Fu, W.-Y. (2010). Investigation of the capitalization of municipal government infrastructure investment on housing market: Hedonic model based on guangzhou housing price data. *Geographic Research, 29*(7), 1269–1280.

Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 518–522. IEEE.

Liu, Z. (2005). Institution and inequality: the hukou system in china. *Journal of Comparative Economics, 33*(1), 133–157.

Lu, Y., & Sun, T. (2013). *Local government financing platforms in China: A fortune or misfortune?* International Monetary Fund.

Meese, R., & Wallace, N. (1991). Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. *Real Estate Economics, 19*(3), 308–332.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & its Applications, 9*(1), 141–142.

Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings*, Volume 5, pp. 1–5. BioMed Central.

Ong, S. E., Ho, K. H. D., & Lim, C. H. (2003). A constant-quality price index for resale public housing flats in Singapore. *Urban Studies, 40*(13), 2705–2729.

Owusu-Ansah, A. (2011). A review of hedonic pricing models in housing research. *Journal of International Real Estate and Construction Studies, 1*(1), 19.
Pace, R. K. (1993). Nonparametric methods with applications to hedonic models. *The Journal of Real Estate Finance and Economics, 7*(3), 185–204.
Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications, 42*(6), 2928–2934.
Phillips, P.C.B., & Shi, Z. (2021). Boosting: Why you can use the HP filter. *International Economic Review, 62*(2), 521–570.
Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal Econometric Society*, 931–954.
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55.
Shi, Z., & Huang, J. (2021). Forward-selected panel data approach for program evaluation. *Journal of Econometrics*. forthcoming.
Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics, 195*(1), 104–119.
Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature, 13*(1), 1–44.
Wang, X., Li, K., & Wu, J. (2018). House price index based on online listing information: The case of china. *Available at SSRN 3223256*.
Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.
Wen, H., Xiao, Y., & Zhang, L. (2017). School district, education quality, and housing price: Evidence from a natural experiment in hangzhou, china. *Cities, 66,* 72–80.
Wu, J., Deng, Y., & Liu, H. (2014). House price index construction in the nascent housing market: the case of china. *The Journal of Real Estate Finance and Economics, 48*(3), 522–545.
Wu, J., Gyourko, J., & Deng, Y. (2012). Evaluating conditions in major chinese housing markets. *Regional Science and Urban Economics, 42*(3), 531–543.
Zheng, S., Wu, J., Kahn, M. E. & Deng, Y. (2012a). Estimating the value of educational quality in china using beijing school district assignment policies.
Zheng, S., Wu, J., Kahn, M. E., & Deng, Y. (2012b). The nascent market for "green" real estate in beijing. *European Economic Review,56*(5), 974–984.
Zheng, S., & Kahn, M. E. (2008). Land and residential property markets in a booming economy: New evidence from beijing. *Journal of Urban Economics, 63*(2), 743–757.
Zheng, S., & Kahn, M. E. (2013). Does government investment in local public goods spur gentrification? evidence from beijing. *Real Estate Economics, 41*(1), 1–28.