Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09673-3

ORIGINAL ARTICLE

Hierarchical fuzzy regression functions for mixed predictors and an application to real estate price prediction

Haydar Demirhan[1] · Furkan Baser[2]

Received: 15 September 2023 / Accepted: 25 March 2024

Abstract

Categorical features appear in datasets from almost every practice area, including real estate datasets. One of the most critical handicaps of machine learning algorithms is that they are not designed to capture the qualitative nature of the categorical features, leading to sub-optimal predictions for the datasets with categorical observations. This study focuses on a new fuzzy regression functions framework, namely hierarchical fuzzy regression functions, that can handle categorical features properly for the regression task. The proposed framework is benchmarked with linear regression, support vector machines, deep neural networks, and adaptive neuro-fuzzy inference systems with real estate data having categorical features from six markets. It is observed that the proposed method produces better prediction performance for real estate price prediction than the benchmark methods in a wide variety of real estate markets. Since we provide all the required software codes to implement the proposed hierarchical fuzzy regression functions framework, our approach offers practitioners a readily applicable, high-performing tool for real estate price prediction and other regression problems involving categorical independent features.

Keywords Categorical features · Deep learning · Market segmentation · Neural networks · Real estate valuation · Support vector machines

1 Introduction

Real estate, one of the indispensable assets of the economy, directly affects many areas, from the financial to the legal systems. For the development of economies and sustainable growth, the most accurate pricing of real estate is crucial. Many economic activities, such as real estate trading, mortgage loans, investment, balance sheet and taxation, depend on real estate prices [1, 2]. Inaccuracy in property price predictions may result in unsuitable investment decisions [3].

The biggest obstacle in accurate real estate valuation is the heterogeneous structure of real estate data [4]. Highly variable characteristics of the real estate market make it challenging to predict real estate values. Real estate datasets include categorical (mostly binominal and nominal) and discrete and continuous (interval scale) features. For example, while the number of rooms is a discrete measurement in the interval scale, the price of a property is a continuous measurement in the interval scale. On the other hand, whether a property is furnished or not is a binary categorical observation and property type is a nominal categorical observation.

As detailed in Sect. 2, there are many different approaches to the property value/price prediction problem in the literature. The hedonic model used for price prediction is based on multiple linear regression [5]. Machine learning (ML) methods such as support vector machines (SVMs) [6] and artificial neural networks (ANNs) [7] are proposed to improve the hedonic model's performance. Fuzzy methods are also introduced for property price

Haydar Demirhan and Furkan Baser have contributed equally to this work.

✉ Haydar Demirhan
haydar.demirhan@rmit.edu.au

Furkan Baser
furkan.baser@ankara.edu.tr

[1] Mathematical Sciences Discipline, School of Science, RMIT University, Melbourne, VIC 3000, Australia

[2] Department of Actuarial Sciences, Faculty of Applied Sciences, Ankara University, 06590 Ankara, Turkey

Published online: 17 April 2024

prediction problems. The most frequently used fuzzy approaches are fuzzy neural networks (FNNs) and the adaptive neuro-fuzzy inference system (ANFIS) [8, 9]. The hybrids of these methods with clustering approaches are employed to segment the real estate market for improved price prediction empirically.

The main shortfall of the methods in the literature is that they are not designed to process the categorical features to preserve their categorical nature [10, 11]. Although there are attempts to improve methods against the existence of categorical features through using different encoding strategies at the data preprocessing step, the gain in prediction performance over the standard implementation is unclear [10]. Most methods process categorical features as discrete measurements in the interval scale. If not paired with suitable approaches, this results in a significant loss of accuracy that translates into economic loss by sub-optimal decisions based on the predictions, disregarding the features' categorical nature. Another issue with the recent literature is the lack of reproducibility due to not providing the software codes for implementing the proposed methods in practice. This limits the applicability of the methods in real estate price prediction and leaves practitioners with sub-optimal methods.

To tackle these problems, first, we aim to develop a method that can handle both interval scale and categorical features and provides us with a lower magnitude of prediction error and lower variability in prediction error than the frequently used methods. We propose a fuzzy regression functions (FRF) approach designed to handle categorical measurements along with interval-scale observations. The proposed approach involves a step that assigns a membership value to each property to represent each property's degree of belonging to each spatial segment based on hierarchical clustering with the generalized Minkowski distance of [12]. This feature of our approach relaxes the requirement of expert knowledge for segmentation and is able to handle categorical information. The generalized Minkowski distance is also employed in other steps of FRFs to handle categorical features accurately. We develop all the required computer codes to implement HFRFs in R software and make them available for immediate implementation through the Internet.

The proposed FRF approach, namely hierarchical fuzzy regression functions (HFRFs), is applied to real estate datasets from six different markets worldwide to predict real estate prices. The performance of the proposed HFRFs is assessed in terms of the magnitude of the absolute prediction error and the amount of variability in the prediction errors. We benchmark the performance of HFRFs with ANFIS, deep neural networks (DNNs), linear regression, and SVMs using the six real estate datasets. Since DNN is a multi-layered version of ANN, we do not consider ANNs separately in this study. HFRFs demonstrate a considerable extent of improvement in real estate prediction performance. The computational cost of the proposed approach is assessed and found suitable for larger-scale valuation systems. The contributions of this study are: i) An FRF approach that handles categorical and interval-scale features accurately is proposed to predict real estate values. HFRFs are not only limited to real estate datasets. It is readily applicable to any dataset with categorical and continuous features for regression problems. ii) We relax the requirement of expert knowledge for segmentation by using hierarchical clustering that captures segmentation information straightforwardly. iii) A comparison of HFRFs, SVMs, ANFIS, linear regression, and DNNs for the real estate values prediction is presented. Thus, information on the performance of the benchmarking methods is also evaluated in this study. iv) Due to the suitable computational cost of HFRFs, we provide practitioners with a method that can be applied to online valuation systems or large-scale prediction problems.

The rest of the article is organized as follows: Sect. 2 is devoted to the literature review. Section 3 presents datasets, descriptive analysis results, the details of FRFs and the proposed HFRF approach, and the goodness-of-fit measures. Section 4 demonstrates the performance and computational cost comparison of the proposed HFRF approach with benchmark methods. Section 5 concludes the article with discussions and recommendations.

2 Literature review

In the real estate market, a common approach to reducing data variability is to cluster the observations into more homogeneous sub-markets [13, 14]. The idea that housing markets should be divided into clusters according to certain characteristics and that these clusters should be included in the price prediction process is an approach advocated in the real estate literature [4, 15, 16]. Performing market segmentation prior to pricing model estimation avoids aggregation bias and ensures accurate parameter estimation and strong model fit [17, 18].

In the real estate market, a sub-market is a cluster in which pricing and related property characteristics differ from another sub-market [15]. It is argued that each sub-market should have its own unique price models [19]. Considering that the coefficient estimates will be biased in a single model representing the whole market, the model estimates for each sub-market are expected to have better pricing performance than an aggregated model [20]. Identifying sub-markets not only improves price prediction accuracy but also helps researchers better model temporal and spatial changes in prices, helps lenders accurately assess credit risk, and reduces home buyers' search costs [4].

In order to identify housing market segmentation, two basic approaches are used: *a priori* information (experience-oriented) classification and data-driven methods. In *a priori* classification, sub-markets are constructed using expert insights into spatial divisions such as administrative areas, socio-economic characteristics, census tract, zip code districts, and physical features [5, 21–23]. Although *a priori* information classification is intuitive and straightforward, the method is relatively inadequate in terms of accuracy, precision, and objectivity [24]. It is also demonstrated by [25] that market segmentation using administrative boundaries might not be effective in mass appraisal. Expert opinions for segmentation in the same area cannot be used as a widely accepted solution, given the fact that real estate agents may differ in their judgment [20]. As [4] emphasize, consumers are guided by the relationship between price and general characteristics as well as the location of the properties when searching for a home. In contrast, data-driven methods are more objective and accurate and reflect the spatial and temporal dynamics of the real estate market [26].

In recent years, studies have focused on the empirical segmentation of the real estate market using various statistical and machine learning methods to overcome the arbitrariness and subjectivity of *a priori* information classification. Principal Component Analysis (PCA), factor analysis, and clustering methods are applied to identify sub-markets as data-driven approaches. Market segmentation is performed based on property type, structural features, neighborhood features, spatial features, or a combination of them. To explain the economic meaning of sub-market segmentation, spatial attributes and economic characteristics of properties are often used for clustering [27, 28]. Watkins [22] use PCA to identify structurally differentiated market segments by obtaining the most common components in the housing stock. Then, the hedonic regression equation is estimated separately for each sub-market. Using PCA to identify sub-markets made significant progress with the use of cluster analysis. Clustering, as an unsupervised learning problem, is used to separate data samples into clusters without needing any prior knowledge of partitioning. Various crisp and fuzzy clustering algorithms also used in the literature to identify housing sub-markets [27, 29–31].

Machine learning methods are used to identify, interpret, and analyze hugely complicated data structures and patterns. When machine learning techniques are utilized appropriately, they provide fast and reliable results and become a tremendous tool that assists decision-making processes [32]. Intelligent automatic valuation systems for real estate have been designed using various methods and models for mass appraisal in a certain area. When expert algorithms are compared to ML methods in predicting property values, ML methods outperform expert algorithms for Poland data [33].

Since SVMs can capture the nonlinear relationship patterns in regression tasks, they are suitable for the price prediction problem. Other optimization algorithms, such as genetic algorithms or particle swarm optimization, are used to find optimal parameter settings for SVM implementation for real estate price prediction [6, 34]. Another mainstream ML algorithm, ANN is also widely employed to predict real estate values. ANNs capture the nonlinear behavior of input variables well. Mach [35] compares the performance of ANNs with multiple regression modeling and observes a similar performance for Poland data. Cetkovic et al. [7] implement ANNs for France, the Czech Republic, and Lithuania using backpropagation with economic and social variables. They observe a satisfactory prediction performance with ANNs. Some of the variables considered in this study, such as gross domestic product (GDP), GDP per capita at market prices, and foreign direct investment, have high variations. ANNs with backpropagation are improved by employing a genetic algorithm for optimization, producing better prediction performance for the Chinese market [36]. In addition to ANNs, other ML methods, such as ElasticNet and XGBoost, are recently employed to predict real estate prices. But ANNs outperform both ElasticNet and XGBoost for Italy datasets [37]. A recent review of ANN and ML methods for real estate price prediction in comparison with the hedonic model is presented by [38]. We consider an improved and more generalized version of ANNs, namely DNNs, in this study.

Lee [11] highlights the importance of considering the nature of the categorical independent features and the weakness of neural networks in processing categorical features and proposes using entity embedding techniques in natural language processing to improve the performance of neural networks. The entity embedding technique leads to market segmentation for prediction using only one model. This is an important consideration since most real estate pricing applications include categorical features.

Overall, ML methods are promising for real estate price prediction and perform better than conventional regression-based models. However, their weaknesses include a lack of an explanation of the mechanism behind the predictions [11] and poor handling of categorical predictors [11]. Different encoding strategies such as one hot, CatBoost, Helmert, target, and ordinal encoding are used for categorical features at the data preprocessing step to improve the performance of ML methods [10]. Although the performance of different methods is compared under these encodings, the performance gain with encoding against the straightforward implementation is unclear.

In the fuzzy domain, FNNs are the most used technique for real estate value prediction dating back to the end of the 1990 s [39, 40]. Artificial neural networks are criticized due to their black-box nature and low generalization capacity due to overfitting [8, 40]. The hedonic price theory is used along with the FNNs. The straightforward implementation of FNNs involves getting crisp inputs, fuzzifying them, calculating fuzzy inference rules by using an ANFIS and defuzzifying the output with centroid defuzzification. Liu et al. [40] get promising results and good generalization capability with this approach. Guan et al. [41] consider an ANFIS approach to property valuation for the US data and compare the accuracy of their ANFIS approach to multiple regression modeling. The ANFIS approach performs slightly worse than a multiple regression approach for the US data. Note that [41] dataset includes categorical variables such as basement type, wall type, and garage type. However, according to descriptive statistics reported in the article, these categorical features are treated as interval-scale measurements after coding. As a neural network-based method, ANFIS does not handle categorical features accurately. This can be the reason for getting better performance from the multiple regression approach. This result highlights the importance of using methods that can handle categorical features in property price prediction. Shi et al. [31] use the Fuzzy k-means (FKM) clustering method in housing market segmentation, and predictions in each cluster are obtained using ANFIS. This study concludes that data-driven market segmentation improves the accuracy of price prediction. Kusan et al. [42] propose a fuzzy logic model for property price prediction in Turkey. Their model is based on a wide range of continuous and categorical variables and shows a satisfactory performance in the presence of categorical variables. Sarip et al. [9] provide a comparison of ANN, ANFIS, and fuzzy least-squares regression (FLSR) methods in the predictions of property prices through simulation-based experiments. In the simulation experiments, some categorical variables are included. As a result, FLSR is identified as a promising method for property valuation. Grid partitioning and subtractive clustering are considered along with ANFIS to create optimum fuzzy rule base sets for forecasting house selling prices [43], and ANFIS with grid partitioning was found promising for forecasting house prices. There are also attempts to build automated valuation systems for real estate using fuzzy rule-based systems [44–46]. However, whether these systems distinguish the categorical nature of the categorical predictors is not clarified.

The previously introduced FRF approaches by [47–49] aim to improve the FRF method against the outliers in interval-scale features. The methods proposed in these studies are not recommended when categorical features are in the dataset; hence, they are unsuitable for real estate price prediction. However, the current study specifically focuses on developing an FRF approach that provides better prediction accuracy for datasets with both interval-scale and categorical (nominal and/or ordinal) features.

Overall, most studies in the literature do not treat categorical variables differently from interval-scale measurements, resulting in the loss of accuracy in the real estate price prediction. This is the main limitation of the literature we address in this study.

3 Datasets and methods

3.1 Data description

Six diverse real estate datasets from Russia, Georgia (the USA), Taiwan, Riga (Latvia), Sao Paulo (Brazil), and Victoria (Australia) are considered in this study. The data sources are available in Table 1.

The datasets contain continuous and categorical predictors with varying sample sizes. Table 2 contains information about each dataset's features and the sample size. Since there are duplications and zero-variance features in the datasets given by the original data sources, we applied a preprocessing step. The resulting datasets are available at https://github.com/haydarde/HFRF. The type of categorical predictors is given in brackets in the first column of Table 2. The names of some predictors are changed to group them into one name across the datasets. We consider the location of each property by latitude and longitude information. The sale price is log-transformed to reduce the large range caused by the local currencies for datasets 1, 2, 4, 5, and 6. Since the Taiwan data have the unit area price, log-transformation is not applied. The most common features across the datasets are the total area and the number of rooms, seen in 4 and 5 out of 6 datasets, respectively. Out of 21 predictors, we have 2 nominal categorical, 5 binary, 7 discrete and 7 continuous predictors across the datasets. Thus, it is crucial to have methods that handle binary and nominal categorical features sufficiently.

Box plots of log-sale price for each dataset are displayed in Fig. 1. The triangle in each box indicates the mean log-sale price, while the horizontal line in each box shows the median log-sale price. All datasets have almost similar log-sale price distributions. The variation of log-sale prices is the least for Victoria, while it is the highest for Russia. The Sao Paulo and Victoria datasets are slightly right-skewed with high sale prices.

The box plots of the log-total area for Georgia, Riga, Russia, and Sao Paulo datasets are given in Fig. 2. While properties in Georgia have the highest variation in total area with a right-skewed distribution, the Russia dataset has the lowest variation in total area. The Sao Paulo dataset

Table 1 Datasets and web links to data sources

Dataset	Source
1. Russia	https://www.kaggle.com/mrdaniilak/russia-real-estate-20182021
2. Georgia (US)	https://www.kaggle.com/yellowj4acket/real-estate-georgia
3. Taiwan	https://www.kaggle.com/quantbruce/real-estate-price-prediction?select=Real+estate.csv
4. Riga (LV)	https://www.kaggle.com/trolukovich/riga-real-estate-dataset
5. Sao Paulo (BR)	https://www.kaggle.com/argonalyst/sao-paulo-real-estate-sale-rent-april-2019
6. Victoria (AU)	https://www.kaggle.com/ruizjme/realestate-vic-sold

Table 2 Features, sample size, number of predictors, and number of categorical predictors in each dataset

Feature	Dataset					
	1	2	3	4	5	6
Latitude	x	x	x	x	x	x
Longitude	x	x	x	x	x	x
Log of price in local currency	x	x	x	x	x	x
Total area	x	x	o	x	x	o
Kitchen area	x	o	o	o	o	o
Number of rooms	x	x	o	x	x	x
Apartment floor	x	o	o	x	o	o
Number of stories	o	o	x	x	o	o
Property type (Nominal categorical)	o	o	o	o	o	x
Building type (Nominal categorical)	x	o	o	x	o	o
Number of bathrooms	o	x	o	o	x	x
Age	o	x	x	o	o	o
In new construction (Binary)	x	o	o	o	o	o
Number of parking spaces	o	x	o	o	x	x
Pool (Binary)	o	x	o	o	x	o
Spa (Binary)	o	x	o	o	o	o
Distance to train station	o	o	x	o	o	o
Condition (Nominal categorical)	o	o	o	x	o	o
EnSuites	o	o	o	o	x	o
Condominium expenses	o	o	o	o	x	o
Elevator (Binary)	o	o	o	o	x	o
Furnished (Binary)	o	o	o	o	x	o
Number of predictors	8	9	5	8	12	6
Number of categorical predictors	2	2	0	2	3	1
Sample size	4990	5678	414	2448	4634	386

In the table, ‘x’ and ‘o’ indicate whether the corresponding feature exists in the dataset or not, respectively. Dataset: 1. Russia, 2. Georgia (US), 3. Taiwan, 4. Riga (LV), 5. Sao Paulo (BR), 6. Victoria (AU)

is notably right-skewed with properties with very large total areas.

The histograms of the number of rooms for Georgia, Riga, Russia, Sao Paulo, and Victoria datasets are given in Fig. 3. All distributions of the number of rooms are right-skewed, except for Victoria, which has a left-skewed distribution.

The bar plots of the categorical predictors in Russia, Georgia, Riga, Sao Paulo, and Victoria datasets are given in Fig. 4. All the categorical variables except the swimming pool and elevator predictors for Sao Paulo are highly imbalanced. In addition to having categorical predictors in the dataset, their imbalancedness adds another level of challenge to accurate modeling. A sufficient model should handle imbalanced categorical variables successfully in this case.

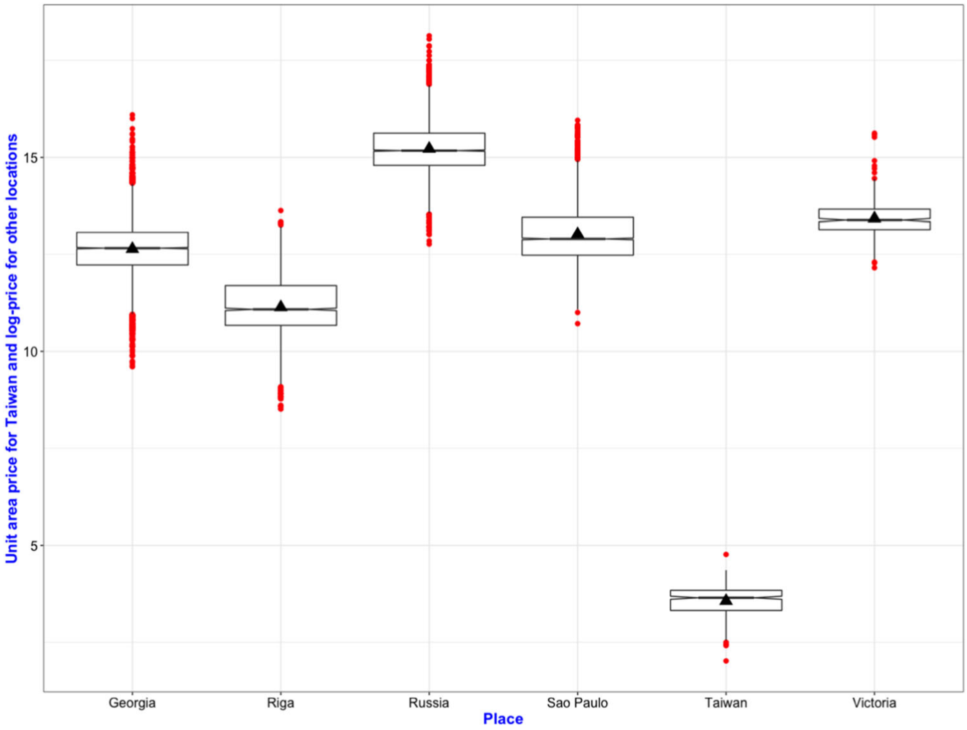

Fig. 1 Box plots of log-sale price for each dataset

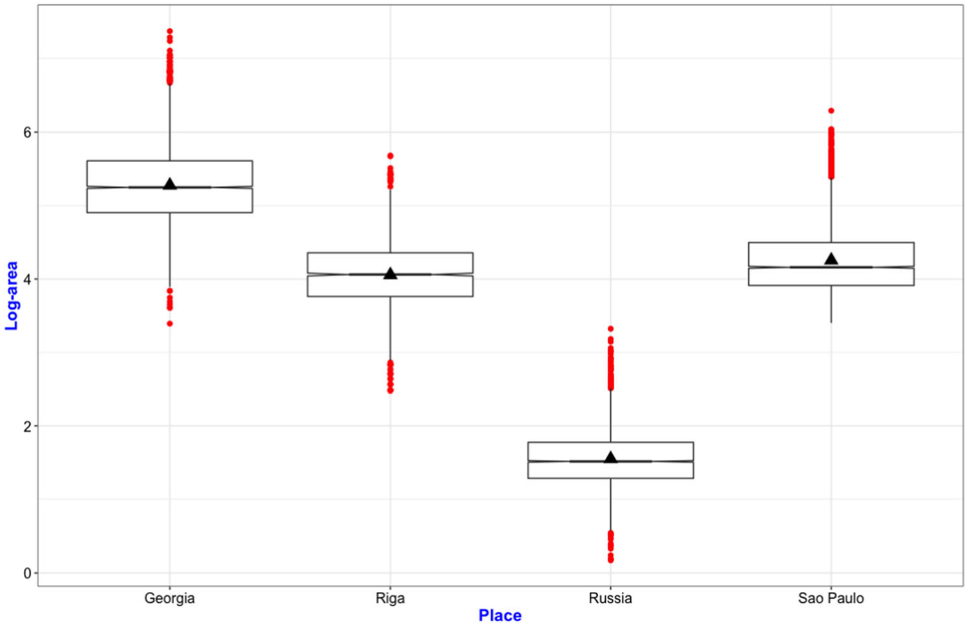

Fig. 2 Box plots of the logarithm of the total area for Georgia, Riga, Russia, and Sao Paulo datasets

3.2 Fuzzy regression functions

The FRF technique is originally developed by [50]. Baser and Demirhan [47] introduced the SVMs into the FRFs to enhance their performance for regression. Then, the noise cluster approach of [51] is used along with SVMs and ANNs within FRF to improve the robustness of FRFs against the outliers for regression [48]. This approach is called FRF with a noise cluster (FRFN). Then, [49] modified FRFNs by implementing robust clustering methods in the FRFN approach, yielding a modified FRFN approach (MFRFN) to further robustify FRFNs against the outliers in the dataset. MFRFN demonstrates a promising estimation accuracy for wind speed estimation [52]. In the time series setting, FRFs provide robust forecasting results against the outliers as well [53, 54]. However, no categorical predictor

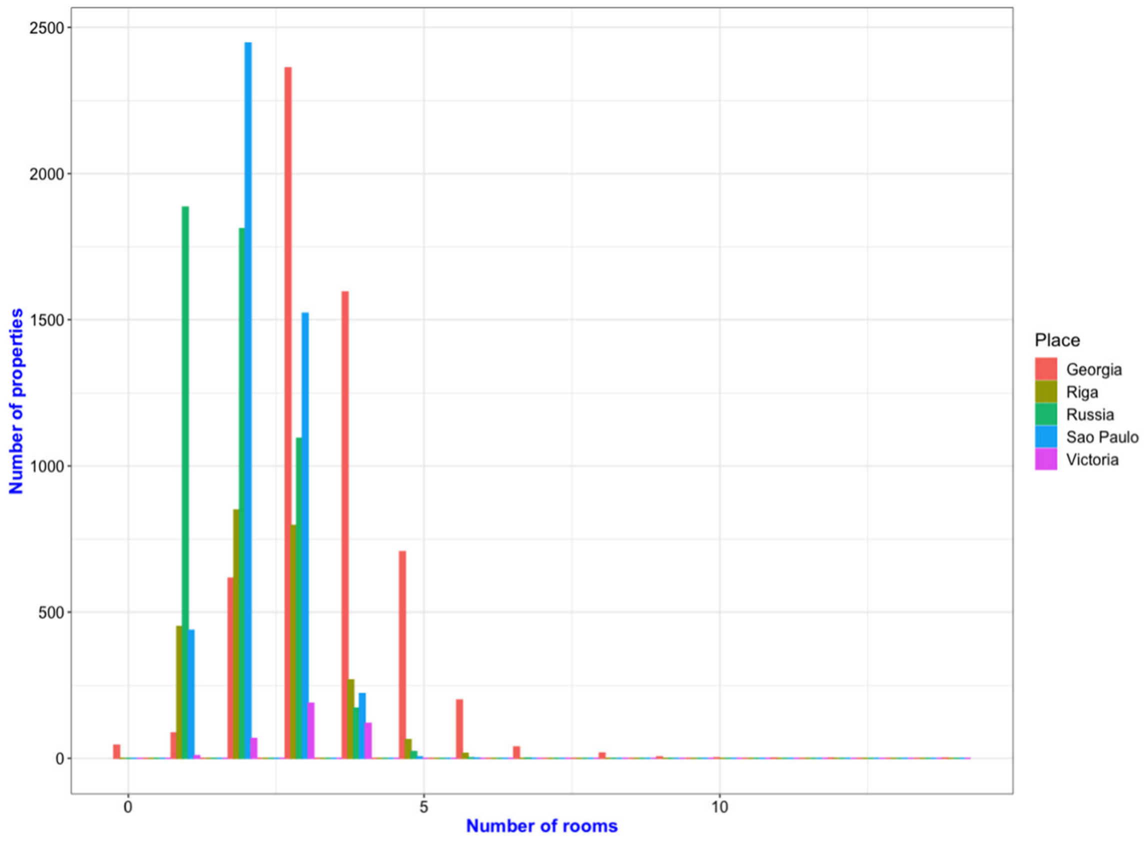

Fig. 3 Histograms of the number of rooms for Georgia, Riga, Russia, Sao Paulo, and Victoria datasets

is considered in any of these works. In this study, we develop an FRF model, namely the hierarchical fuzzy regression functions (HFRF) model, that runs efficiently when binary, multi-class nominal, ordinal and/or interval-scale predictors are in the data.

The objective function to minimize in FRF implementation is

$$f(\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{X}) = \sum_{i=1}^{n} \sum_{k=1}^{c} g_{k,i}(\mu_{k,i}, d_{k,i}), \tag{1}$$

where c is the number of clusters, n is the sample size, $\boldsymbol{X}_{n \times (\ell+1)}$ is the matrix of ℓ independent predictors and the dependent feature, $\mu_{k,i}(\boldsymbol{x}_i, \boldsymbol{v}) = \mu_{k,i} \in [0, 1], \sum_i \mu_{k,i} = 1 \forall k = 1, \ldots, c$ shows membership values, $d_{k,i}(\boldsymbol{x}_i, \boldsymbol{v}_k) = d_{k,i}$ is the Euclidean distance between the cluster centers and observations as the inner product norm:

$$d_{k,i} = ||\boldsymbol{x}_i - \boldsymbol{v}_k||, \tag{2}$$

with the $\ell \times 1$ vector of cluster centers, $\boldsymbol{v}_k(\boldsymbol{x}_i) = \boldsymbol{v}_k$, $k = 1, \ldots, c$. In Eq. (1), the function $g_{k,i}(\cdot)$ indicates the clustering method. For the fuzzy c-means (FCM) clustering, $\mu_{k,i}$ is defined in Eq. (3),

$$\mu_{k,i} = \left\{ \sum_{l=1}^{c} \left(\frac{d_{k,i}}{d_{l,i}} \right)^{2/(m-1)} \right\}^{-1} \tag{3}$$

with the degree of fuzziness $m > 1.1$, and the center of cluster k is defined in Eq. (4):

$$\boldsymbol{v}_k = \frac{\sum_{i=1}^{n} \mu_{k,i}^{m} \boldsymbol{x}_i}{\sum_{i=1}^{n} \mu_{k,i}^{m}}, \tag{4}$$

and $g_{k,i}(\cdot)$ function is given in Eq. (5):

$$g_{k,i}(\mu_{k,i}, d_{k,i}) = \mu_{k,i}^{m} d_{k,i}^{2}, k = 1, \ldots, c. \tag{5}$$

The FRF implementation includes two stages: the fuzzy clustering stage with the training dataset and the fuzzy inference stage with the output of the fuzzy clustering stage on the testing dataset. Learning similarities between observations happens at the clustering stage as it captures the clusters created by the similarities of observations in the data dataset. This corresponds to the segmentation in the real estate valuation domain. Then, the membership values based on the distance between the observations and cluster centers are added to the training dataset as another predictor, and an SVM model is fitted at the inference stage. The fitted SVM model is then used to generate predictions with the test data at the last stage of FRFs [49].

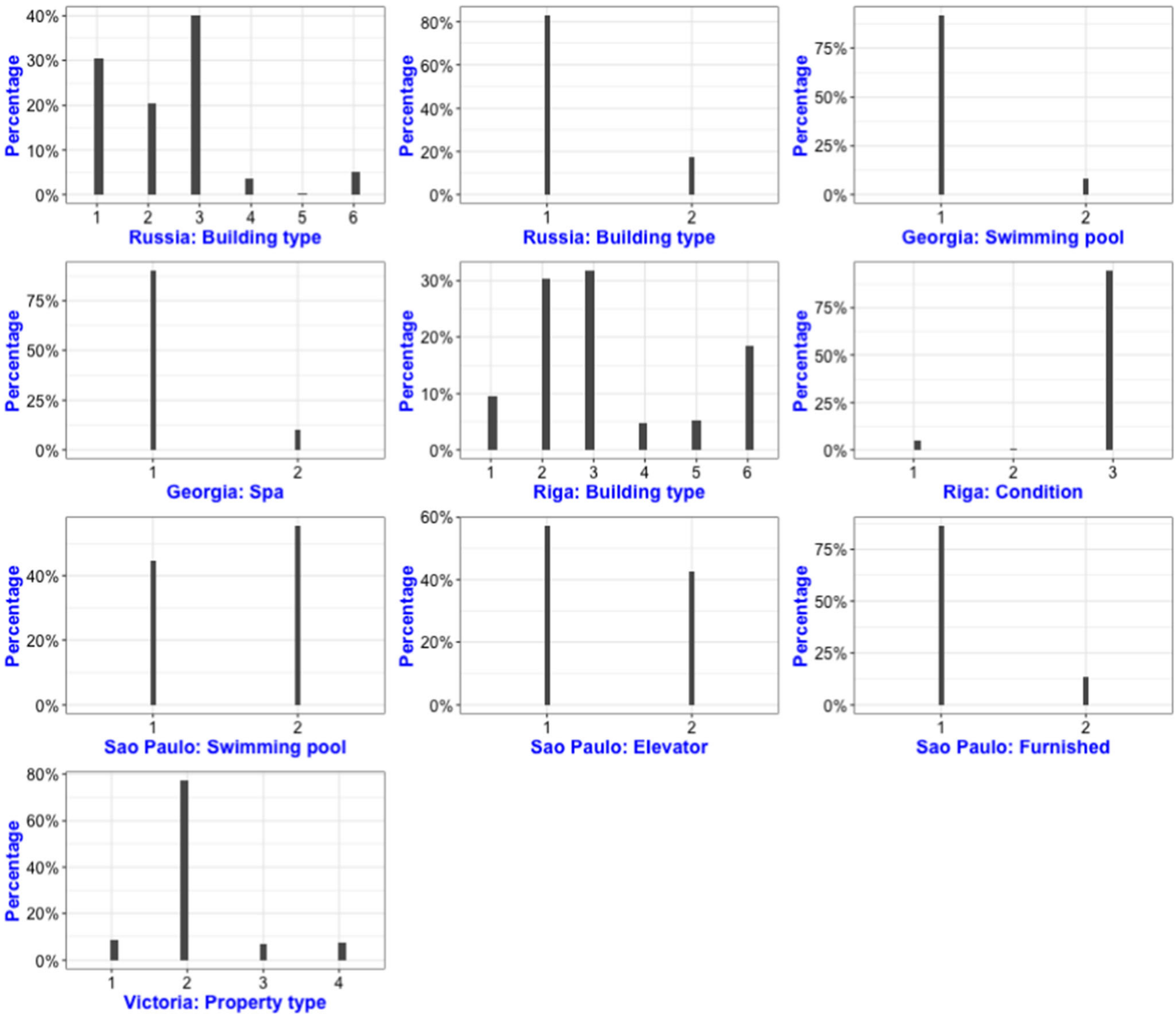

Fig. 4 Bar plots of the categorical variables for Russia, Georgia, Riga, Sao Paulo, and Victoria datasets

However, FRFs rely on the Euclidean distance, which is not an appropriate distance measure for the categorical features.

3.2.1 Hierarchical fuzzy regression functions

It is clear that when some of the predictors are categorical, using the Euclidean distance between the cluster centers and observations from the categorical predictors is unsuitable [12, 55]. The distance is used twice in FRFs. First, it is used at the clustering stage, and then, the distance between cluster centers and observations is fed into the inference stage. Therefore, using a distance metric suitable for mixed data types is crucial. One can either transform all categorical features into binary by dummy coding and use a distance for binary variables or use a Minkowski distance by considering the categorical variables with integer coding [12, 55, 56]. Considering that our FRF method consists of SVM implementation, keeping the variables numerical rather than binary is more beneficial.

The accuracy of the clustering algorithm at the first stage of FRFs significantly impacts the overall performance of the FRF implementation. Therefore, the clustering algorithm of the first stage needs to be suitable for mixed data types and less time-consuming as there are two stages of implementation. Specifically, capturing the similarity among the observations is also important for real estate valuation. Considering these characteristics, a hierarchical clustering algorithm is suitable by employing the generalized Minkowski distance of [12].

To handle nominal, binary, ordinal, and interval-scale features with FRFs, we consider using hierarchical clustering with the generalized Minkowski distance [12] at the first stage of FRFs and calculate the generalized Minkowski distances between observations and centers from the hierarchical clustering to feed additional data to the inference stage in the HFRF framework.

For two vectors $\boldsymbol{z} = (z_i)^T$ and $\boldsymbol{y} = (y_i)^T, i = 1, \ldots, n$ in $\mathbb{R}^n$, the Minkowski distance is defined as in Eq. (6):

$$d^p_{z,y} = \left(\sum_{i=1}^{n} |z_i - y_i|^p \right)^p. \tag{6}$$

The generalized version of the Minkowski distance is defined by [12] for continuous, discrete, quantitative, qualitative, and structural data types. However, since we are working with continuous, discrete, nominal, and ordinal data types, we establish a simpler form of it based on the observed ranges of the $\boldsymbol{x}$ and $\boldsymbol{y}$, which is less time-consuming in implementation.

Let $\boldsymbol{X}_{n \times k} = (x_{ij}), i = 1, \ldots, n; j = 1, \ldots, \ell$, where n is the number of observations, and k is the number of features, be the data matrix and R_j denote

- The absolute range of the column vectors of $\boldsymbol{X}$, $|\max(\boldsymbol{x}_j) - \min(\boldsymbol{x}_j)|$ when $\boldsymbol{x}_j$ is in interval scale (continuous or discrete), and
- The number of levels, $L(\boldsymbol{x}_j)$ when $\boldsymbol{x}_j$ is composed of nominal or ordinal measurements.

Then, the generalized Minkowski distance of order p between two observation vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_r$ is defined in Eq. (7):

$$d^p_{ir}(\boldsymbol{X}) = \left[\sum_{j=1}^{\ell} \left(|x_{ij} - x_{rj}| / R_j \right)^p \right]^p. \tag{7}$$

We run hierarchical clustering at the clustering stage of HFRF with the distance matrix $\boldsymbol{D}^p = (d^p_{ir}), i, r = 1, \ldots, \lfloor \gamma \cdot n \rfloor$, where $\lfloor \cdot \rfloor$ shows the floor function and γ is the proportion of the training sample, and obtain cluster center $\boldsymbol{v}^H_k = (v^H_{kj}), j = 1, \ldots, \ell$ for each cluster $k = 1, \ldots, c$. Then, we compute the membership values of observation vectors to the clusters with Eq. (8):

$$\mu^H_{p,k,i} = \left\{ \sum_{l=1}^{c} \left[\frac{\sum_{j=1}^{\ell} \left(|x_{ij} - v^H_{kj}| / R_j \right)^p}{\sum_{j=1}^{\ell} \left(|x_{ij} - v^H_{lj}| / R_j \right)^p} \right]^{2p/(m-1)} \right\}^{-1}. \tag{8}$$

The membership values from Eq. (8) are merged with the observation matrix of each cluster, $\boldsymbol{\Gamma_k}(\boldsymbol{X}_k \vdots \mu^H_{p,k,i})$ and fuzzy regression functions for each cluster, $f_k(\boldsymbol{\Gamma_k}, \boldsymbol{\beta}_k)$, are fitted using SVMs with radial kernel to get the parameter estimates $\hat{\boldsymbol{\beta}}_k$. The $\hat{\boldsymbol{\beta}}_k$ vectors are used to create fitted values of each observation vector, $\hat{\boldsymbol{y}}_k = (\hat{y}_{k,i}), i = 1, \ldots, \lfloor \gamma \cdot n \rfloor$ in the training sample. Then, the fitted values in the training set are calculated as the weighted averages of the fitted values with weights corresponding to the membership values:

$$\hat{y}_i = \sum_{k=1}^{c} \hat{y}_{k,i} \mu^H_{p,k,i} / \sum_{k=1}^{c} \mu^H_{p,k,i}. \tag{9}$$

For the observations in the test sample, we calculate Eqs. (7) and (8) with observations $\boldsymbol{x}_i, i = \lfloor \gamma \cdot n \rfloor + 1, \ldots, n$ and find the predictions using the fitted SVM model's parameter estimates $\hat{\boldsymbol{\beta}}_k$ for each cluster. Then, the final predictions are calculated using Eq. (9) for the observations $\boldsymbol{x}_i, i = \lfloor \gamma \cdot n \rfloor + 1, \ldots, n$.

The workflow of HFRF is given in Fig. 5. Compared to the flow of FRF implementation, we employ the hierarchical clustering with the generalized Minkowski distance to find the cluster centers and then utilize the generalized Minkowski distance for the membership degrees of the observation to the clusters. This improvement makes FRFs applicable for datasets that include categorical and interval-scale observations. The hierarchical clustering with the generalized Minkowski distance provides us with clusters of market segmentation, and the distance between observation to the clusters relates each real estate to the segments.

3.3 Goodness-of-fit measures

In order to assess the prediction performance of the models, root-mean-squared error (RMSE) and mean absolute error (MAE) are used as defined in Eq. (10):

$$\text{RMSE} = \left(\sum_{i=1}^{N} (\hat{x}_i - x_i)^2 / N \right)^{0.5} \quad \text{and} \quad \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{x}_i - x_i|, \tag{10}$$

where $|\cdot|$ shows absolute value, x_i is the observed value in either training or test sets of size N, $\hat{x}_i$ is the corresponding prediction by a model, and $\bar{x}$ is the mean of either training or test set. The scaled versions of RMSE and MAE, namely rRMSE and rMAE, are obtained by dividing them by $\bar{x}$ in Eq. (11):

$$\text{rRMSE} = \text{RMSE}/|\bar{x}| \quad \text{and} \quad \text{rMAE} = \text{MAE}/|\bar{x}|. \tag{11}$$

rRMSE and rMAE provide an assessment independent of each market's price level.

RMSE and MAE handle prediction errors differently and need to be considered simultaneously for a comprehensive evaluation of the models. MAE measures the average absolute difference between the observed and predicted values to depict the average magnitude of the model's error. On the other hand, RMSE measures the variation in the errors, showing the degree of divergence of the errors. It inflates quicker than MAE for larger errors. The mean magnitude of error needs to be considered

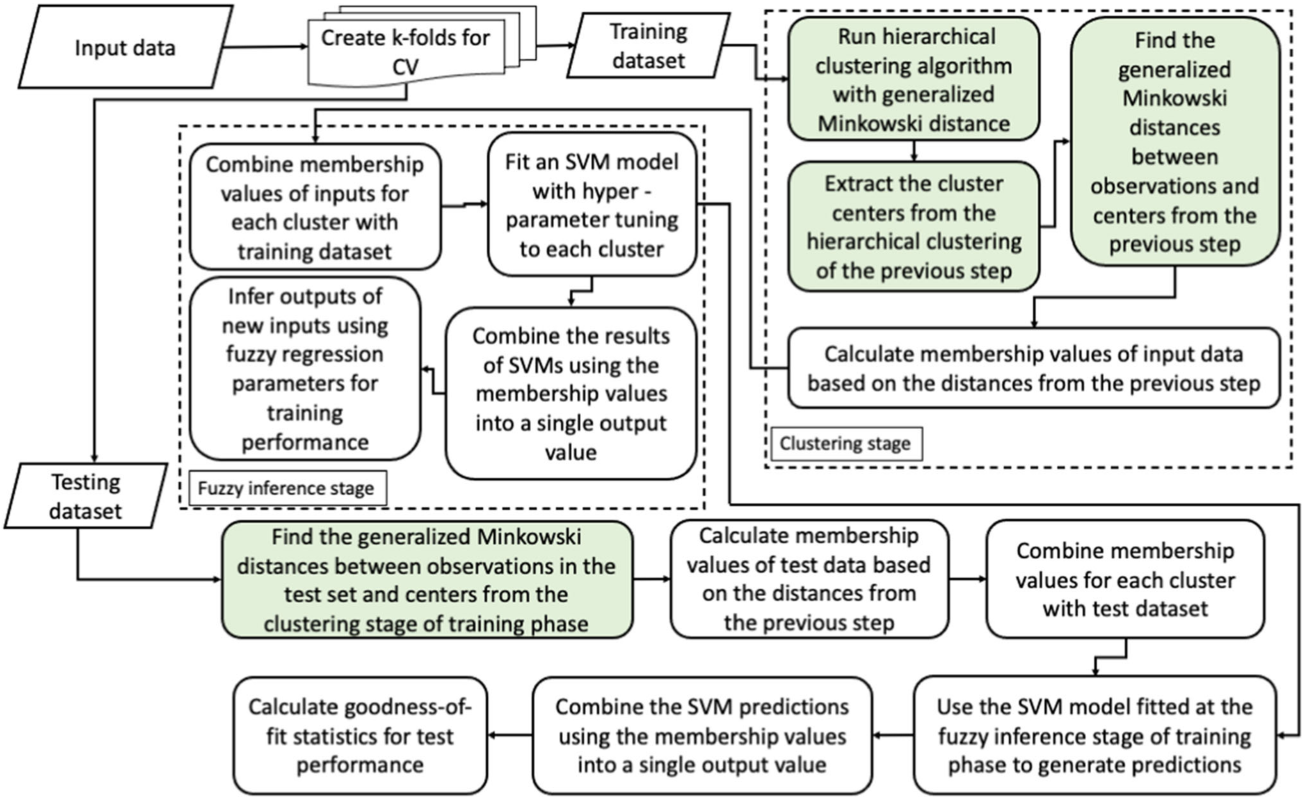
Input data
Create k-folds for CV
Training dataset
Run hierarchical clustering algorithm with generalized Minkowski distance
Extract the cluster centers from the hierarchical clustering of the previous step
Find the generalized Minkowski distances between observations and centers from the previous step
Calculate membership values of input data based on the distances from the previous step
Clustering stage
Combine membership values of inputs for each cluster with training dataset
Fit an SVM model with hyper - parameter tuning to each cluster
Combine the results of SVMs using the membership values into a single output value
Infer outputs of new inputs using fuzzy regression parameters for training performance
Fuzzy inference stage
Testing dataset
Find the generalized Minkowski distances between observations in the test set and centers from the clustering stage of training phase
Calculate membership values of test data based on the distances from the previous step
Combine membership values for each cluster with test dataset
Use the SVM model fitted at the fuzzy inference stage of training phase to generate predictions
Combine the SVM predictions using the membership values into a single output value
Calculate goodness-of-fit statistics for test performance

Fig. 5 Flowchart of the HFRF implementation

simultaneously with the errors' degree of variation to see if the model generates consistently low errors or not. The rescaled versions of MAE and RMSE, rMAE and rRMSE, are also considered since MAE and RMSE are not directly comparable across different markets due to different price levels. However, although the rescaled versions are comparable for multiple markets, they do not reflect the actual magnitude of error for individual markets. Therefore, we investigate both regular and rescaled versions of MAE and RMSE in this study.

4 Results

We implement the proposed HFRF, DNN [57], ANFIS [58], SVMs with the linear and radial kernels (SVM-Lin and SVM-Rad) [59] and linear regression (LinReg) [60] methods with the six real estate datasets presented in Sect. 3.1 to assess the performance of the proposed HFRFs and benchmark it with ANFIS, DNN, SVMs with the linear and radial kernels and linear regression methods, which are frequently applied to tackle the real estate pricing problem in the literature.

All methods are run with hyperparameter tuning via tenfold cross-validation with 80% $(\gamma = 0.8)$ training and 20% test splits. The natural logarithm of real estate prices is used in all methods due to the skewed nature of price data for all datasets except Taiwan data, showing the unit area price differently from others. No transformation is applied to the predictors.

4.1 Hyperparameter tuning

The DNN implementation consists of three dense layers with dropout rates subject to tuning. The first dense layer has an L2 regularizer with a regularization factor of 0.0001. The activation function of the first layer is tuned up considering soft sign and RELU activation functions. The next two layers have RELU activation. The alpha parameter of all RELU activation functions is tuned considering 0.05 and 0.1. The last two layers have He normal initializers. The dropout rate of each layer is tuned by using 0.05, 0.1, and 0.15. The number of units in each layer is tuned by considering 15, 25, and 40. Batch size is tuned with 32, 64, and 128. 200 epochs are run with an early callback monitor based on the mean absolute error. Table 3 shows the final combinations of parameters after the hyperparameter tuning for all datasets.

The ANFIS models are fitted using the Caret and FRBS R packages [61, 62]. In the ANFIS implementation, min and max functions are used as t-norm and s-norm operators, respectively. The Zadeh implication function is used

Table 3 Implementation parameters for DNNs after the hyperparameter tune-up

Dataset	Dropout 1	Dropout 2	Dropout 3	Units	Activation	Alpha	Batch size
1	0.05	0.1	0.05	40	Softsign	0.1	128
2	0.15	0.1	0.05	40	RELU	0.05	128
3	0.05	0.05	0.15	40	Softsign	0.05	32
4	0.05	0.05	0.05	40	Softsign	0.1	32
5	0.1	0.15	0.05	25	RELU	0.1	32
6	0.05	0.15	0.1	25	RELU	0.1	64

[61, 62]. Five layers are considered for forward stage [62]: i) Fuzzification with Gaussian membership function. ii) Inference using the t-norm operator. iii) Calculate the ratio of rules' strength. iv) Estimate parameters. v) Calculate overall output using the sum operator. The least squares method is used for parameter estimation in the backward stage. The step size of the gradient descent is set as 0.01. A hyperparameter tuning effort for ANFIS implementation requires examining 405 combinations of t-norm, s-norm, and implication functions. While the computational cost of tuning is 3.4 days for the smallest dataset, Dataset 3, it inflates to 962.1 days, or 2.6 years, for the largest dataset, Dataset 2. Therefore, the same setting is used across the six regions.

SVMs are also implemented using the Caret package. The linear and radial basis kernels are considered. The regularization parameter C is tuned against RMSE for the linear kernel. For the radial basis kernel, C and σ parameters are tuned against RMSE by random search with 20 replications. The resulting optimal values of C and σ are given in Table 4 for each dataset.

HFRFs are run by the R implementation of Fig. 5 with the codes developed by the authors at https://github.com/haydarde/HFRF. The hierarchical clustering algorithm needs the number of clusters as the essential input. The number of clusters for each dataset is determined based on the silhouette (SIL) index [63]. The degree of fuzziness is tuned considering $m = 1.5, 2, 2.5, 3$ for each dataset. The order of generalized Minkowski distance is tuned through $p = 1.1, 1.2, 1.5, 1.75, 2, 2.25, 2.5$. After the tune-up, the final values of the number of clusters (c), the degree fuzziness (m), and the order of generalized Minkowski distance (p) used for HFRFs are given for each dataset in Table 5. The C and σ parameters of SVM-Rad at the inference stage of HFRFs are tuned against RMSE under each cluster for each dataset by random search with 20 replications. Table 6 shows the optimal values of C and σ for each dataset and the number of clusters shown in Table 5. For each dataset, only c rows of Table 6 are filled with the values of corresponding C and σ hyperparameters.

Table 4 Implementation parameters for SVM-Rad after the hyperparameter tune-up

Dataset	1	2	3	4	5	6
σ	0.086	0.541	0.023	0.027	0.211	0.446
C	5.492	0.827	20.986	50.340	130.426	11.699

Table 5 Implementation parameters for HFRFs for each dataset after the hyperparameter tune-up

	Dataset					
	1	2	3	4	5	6
Number of clusters (c)	2	2	5	3	6	3
Fuzziness degree (m)	3	1.5	1.5	1.5	3	3
Order of generalized Minkowski distance (p)	1.1	1.1	1.1	1.1	1.25	1.1

4.2 Prediction performance

Table 7 presents the RMSE, rRMSE, MAE, and rMAE results of HFRF, DNN, SVM, and ANFIS methods for all datasets. The rescaled error measures rRMSE and rMAE help assess the variation and magnitude of errors in test sets independent of different price ranges across different countries considered in the study. In terms of rMAE, the proposed HFRFs produce the minimum absolute error for real estate price prediction for all datasets. This implies that HFRFs provide us with the lowest magnitude of error in the predicted prices. A useful method is also desired to provide predictions with low variability in addition to the low magnitude of error. Regarding the variation in the errors of price prediction, HFRFs have the lowest variation in price predictions. Only for the Taiwan dataset, SVM-Rad produces a very close rRMSE to HFRF. DNN has the second-best rRMSE for Russia, and SVM-Rad has the second-best rRMSE for other datasets. While ANFIS performs poorly among the considered methods, the hedonic model, LinReg, closely follows SVM-Lin.

Since we apply log-transformation on price data for modeling and neutralize the impact of the mean price in each of the considered real estate markets by reporting the

Table 6 Implementation parameters for SVM-Rad under HFRFs for each cluster associated with each dataset after the hyperparameter tune-up

Cluster	Dataset											
	1		2		3		4		5		6	
	σ	C	σ	C	σ	C	σ	C	σ	C	σ	C
1	0.026	19.125	0.297	3.290	0.188	104.48	0.168	6.194	0.014	736.60	0.022	387.47
2	0.158	20.163	0.043	12.797	0.018	164.35	0.041	20.886	0.013	2.717	0.235	43.109
3					0.122	1.467	0.154	16.738	0.016	382.14	0.758	84.653
4					0.133	3.166			0.058	533.18		
5					0.006	443.73			0.117	8.381		
6									0.142	10.767		

scaled error measures, the real magnitude of the improvement by HFRFs is not quite clear in Table 7. To assess the magnitude of the gain by HFRFs, we find the percent improvement by HFRFs over the second-best model in terms of MAE and report the impact of the percent gain in terms of the average price in each of the considered markets in Table 8.

The improvement in the magnitude of prediction error by the proposed HFRF method ranges between 3.8% in the Russian market and 12.5% in the Sao Paulo market across the compared markets. The proposed HFRF produces price predictions with 6.3% less error than the runner-up method in the Georgia market. This corresponds to a 19,312 USD improvement in the price predictions' margin of error for Georgia, USA, an 11,340 USD better prediction for Sao Paulo, BR, or a 47,418 USD improvement for Victoria, AU. Overall, the proposed HFRF method provides us with significant advancement in the magnitude of real estate price prediction.

Table 9 shows the gain in the variability of the prediction errors by the HFRF method for all datasets based on rRMSE. In terms of rRMSE, the HFRF method is superior to all benchmark methods. The gain in rRMSE is very close to SVM-Rad only for the Taiwan dataset. For Riga, Georgia, Sao Paulo, and Victoria datasets, the gain in the variability of the prediction errors varies between 7% and 14.1%. The lowest gain of 2.8% is recorded for the variability of the prediction errors in the Russian market.

HFRFs are proposed to handle categorical variables better in a dataset with categorical and interval-scale measurements. First, having a 7.7% improvement with the Taiwan dataset that has only interval-scale variables implies that HFRF is a promising method even if there is no categorical variable in the datasets. This is a desired feature of HFRFs. However, we do not observe such a significant improvement in the variability of the prediction errors by the HFRF for the Taiwan dataset. This can also be attributed to not having any categorical variable in this dataset. The least improvement, 3.8%, is seen in the Russia dataset. This dataset has a large sample size and has only 1 categorical variable. In comparison, the Sao Paulo dataset has a similar sample size but 3 categorical variables, and HFRF provides a 12.5% improvement in the error magnitude and a 14.1% improvement in the prediction errors' variability. As we have more categorical variables in a dataset, we can expect better gains in performance with HFRFs. When the number of categorical variables is low with a small-to-moderate sample, such as in the Victoria dataset, the gain with HFRFs is also very promising: 10.9% in error magnitude and 12.1% in error variability.

Figure 6 shows actual and predicted log-sale prices with HFRF for all locations. The 45-degree lines in Fig. 6 display the perfect case of prediction. Generally, HFRF predictions align well with the 45-degree line for all locations. However, we observe outliers for some locations. For Victoria, we have one property located far from the main body of observations in the scatter plot. Since this point is an outlier in the test set and is located in the direction of the trend, we can conclude that HFRF consistently produces a large prediction for this high-priced property. So, HFRF learns from similar observations in the training set to produce a high price prediction for this property. Another notable observation from Fig. 6 is from the Taiwan market. Only one property price is significantly underestimated by HFRF, which is the only significant underestimation by HRFRF in all six markets. This observation is located on top of the main cluster of the points in the scatter plot for Taiwan. Since this property is located in the same region as other top-priced properties in the Taiwan market, we do not anticipate a measurement error for this data point. The reason for HFRF's underestimation is that this property is a single-story dwelling, while other high-priced ones have more than 6 stories.

Table 7 Goodness-of-fit results of HFRF, DNN, SVM, and ANFIS methods for all datasets

Dataset	Method	RMSE	rRMSE	MAE	rMAE
1	HFRF	0.4265	**0.0280**	0.3105	**0.0204**
Russia	LinReg	0.4999	0.0328	0.3776	0.0248
	SVM-Lin	0.5014	0.0329	0.3741	0.0246
	SVM-Rad	0.4416	0.0290	0.3226	*0.0212*
	DNN	0.1925	*0.0288*	0.3244	0.0213
	ANFIS	1.8742	0.1230	1.6501	0.1083
2	**HFRF**	0.3815	**0.0302**	0.2706	**0.0214**
Georgia	LinReg	0.4792	0.0379	0.3463	0.0274
(US)	SVM-Lin	0.4788	0.0379	0.3439	0.0272
	SVM-Rad	0.4195	*0.0332*	0.2887	*0.0228*
	DNN	0.2363	0.0385	0.3486	0.0276
	ANFIS	2.2381	0.1771	2.0092	0.1590
3*	**HFRF**	9.8183	**0.2604**	4.6825	**0.1242**
Taiwan	LinReg	11.0845	0.2939	6.6302	0.1758
	SVM-Lin	11.0197	0.2922	6.3163	0.1675
	SVM-Rad	9.8423	*0.2610*	5.0716	*0.1345*
	DNN	107.1802	0.2745	5.9782	0.1585
	ANFIS	21.4033	0.5676	17.2708	0.4580
4	**HFRF**	0.3246	**0.0292**	0.2379	**0.0214**
Riga	LinReg	0.4115	0.0370	0.3133	0.0282
(LV)	SVM-Lin	0.4157	0.0374	0.3121	0.0281
	SVM-Rad	0.3491	*0.0314*	0.2490	*0.0224*
	DNN	0.1515	0.0350	0.2910	0.0262
	ANFIS	2.5659	0.2307	2.4072	0.2165
5	**HFRF**	0.2425	**0.0186**	0.1779	**0.0137**
Sao Paulo	LinReg	0.3419	0.0262	0.2686	0.0206
(BR)	SVM-Lin	0.3401	0.0261	0.2551	0.0196
	SVM-Rad	0.2907	0.0223	0.2032	*0.0156*
	DNN	0.0796	*0.0217*	0.2147	0.0165
	ANFIS	1.0980	0.0843	0.8318	0.0639
6	**HFRF**	0.2908	**0.0217**	0.2192	**0.0163**
Victoria	LinReg	0.4138	0.0308	0.3314	0.0247
(AU)	SVM-Lin	0.4128	0.0307	0.3079	0.0229
	SVM-Rad	0.3308	*0.0246*	0.2461	*0.0183*
	DNN	0.2113	0.0342	0.3311	0.0247
	ANFIS	1.5203	0.1132	1.4438	0.1075

Bold font indicates the best method, and italic font shows the second-best method in terms of rRMSE and rMAE. *Price per unit area

4.3 Computation time

Figure 7 shows the run times of a single run of the top three methods, HFRF, SVM-Rad, and DNN, with the sample size and number of predictors in each dataset to assess the applicability of the proposed HFRFs in practice. All runs are done with a MacBook Pro computer with an Apple M1 Max chip and 64 GB memory. When there is a hyperparameter tuning effort, the run times in Fig. 7 need to be multiplied by the size of the tuning grid. Among the top three methods, HFRFs are most impacted by increased sample size and the number of predictors. For small-to-moderate samples, such as Taiwan and Victoria datasets, the run time of HFRFs is very close to SVM-Rad and DNNs. However, there is a notable difference between HFRFs and SVM-Rad for large samples such as Georgia or Sao Paulo. On the other hand, for Riga data, which is a moderate sample, HFRF has a better run time. Overall, the reported run times for HFRF have no negative impact on the method's applicability in practice.

5 Conclusion

This study develops a method that performs satisfactorily for the regression task in the presence of categorical and interval-scale measurements in the dataset and particularly focuses on implementing the method for real estate price prediction. Real estate price prediction is one of the important application areas where categorical and interval-scale measurements appear in the dataset. Most machine learning methods, especially distance-based methods, cannot handle the non-numerical information contained in the categorical variables and treat them as numerical, resulting in reduced performance. To propose a solution to this issue, we consider using generalized Minkowski distance along with the hierarchical clustering, as introduced by [12], in the fuzzy regression functions method of [47] with support vector machines (SVMs) to handle the categorical variables better. This approach leads to the proposed hierarchical fuzzy regression functions (HFRF) method. SVMs with radial kernel function are implemented at the inference stage of HFRFs to train the model. SVMs are trained with training data and the membership degree of each observation to the clusters created by the hierarchical clustering with the generalized Minkowski distance. The clustering stage captures the information about similarities between observations. Specifically, it corresponds to market segmentation in the property price prediction problem. Since it is done within HFRFs without requiring user input, it relaxes the requirement for expert knowledge in segmentation. SVMs' parameters are hyper-tuned for each dataset and cluster under the HFRF implementation to achieve the optimum performance for each market segment.

The HFRF method is applied to real estate pricing datasets from six diverse markets from different countries. The error magnitude and variability in prediction with HFRFs are benchmarked against linear regression, SVMs and adaptive neuro-fuzzy inference system (ANFIS)

Table 8 Percent gain in the magnitude of absolute error by HFRFs against the second-best method for each dataset and the impact of the gain on average price

Dataset	1 Russia (RUB)	2 Georgia (USD)	3 Taiwan (TWD)	4 Riga (LVL)	5 Sao Paulo (BRL)	6 Victoria (AUD)
Gain with HFRF	3.8%	6.3%	7.7%	4.4%	12.5%	10.9%
Avr. Price	4,130,406.2	307,735.5	37.710*	67,555.0	453,893.7	676,831.3
Gain in Avr. Price	154,908.8	19,312.4	2.893*	3000.6	56,587.7	74,066.2
Approx. USD gain	1601	19,312	0.091*	4645	11,340	47,418

The currency of each market is noted in brackets. Avr *average*. *Price per unit area

Table 9 Percent gain in the variability of the prediction error by HFRFs against the second-best method for each dataset

Dataset	1 Russia	2 Georgia	3 Taiwan	4 Riga	5 Sao Paulo	6 Victoria
Gain with HFRF	2.8%	9.1%	0.2%	7.0%	14.1%	12.1%

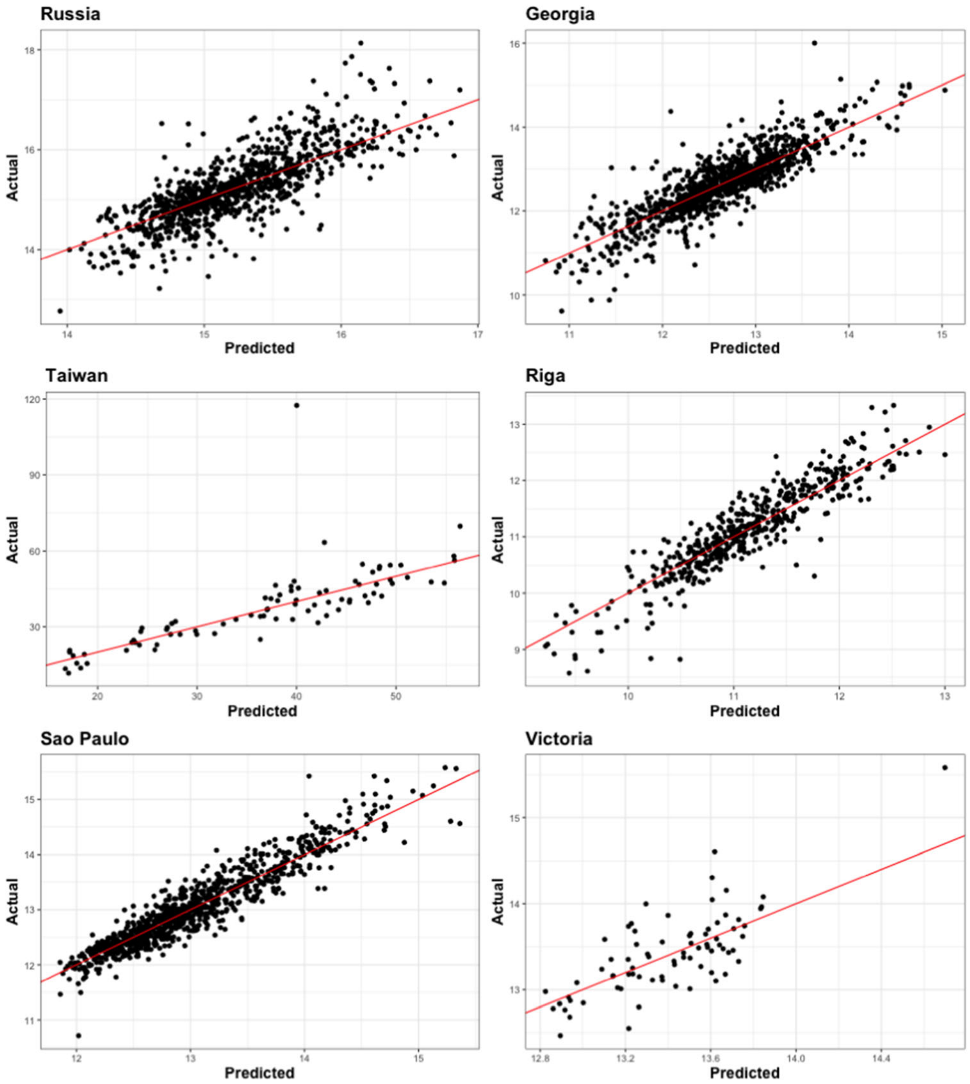

Fig. 6 Scatter plots of actual and predicted log-sale prices with HFRF for all locations. The red 45-degree line indicates a perfect match between predictions and actual observations

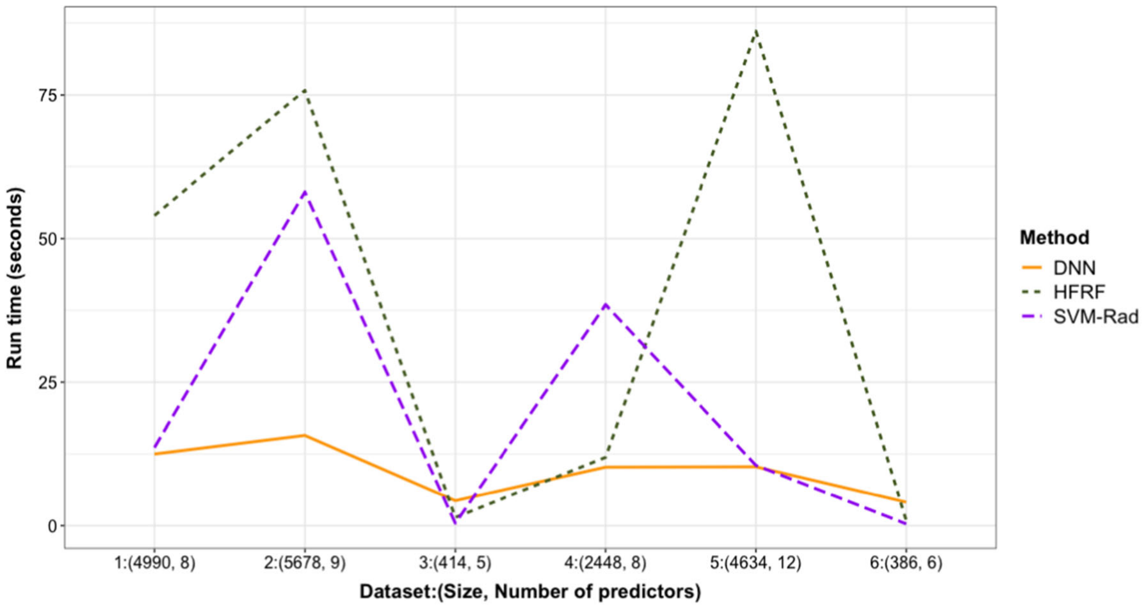

Fig. 7 Run times of the top three best-performing methods, HFRF, SVM-Rad, and DNN, for all datasets

methods that are frequently used for real estate price prediction. In addition to these methods, deep neural networks (DNNs) are also considered in benchmarking as they are more general versions of ANNs with multiple layers. We can summarize the overall conclusions and recommendations of this study as follows:

- The HFRF method produces 3.9% to 12.1% less prediction error than the second-best-performing SVMs with the radial kernel function (SVM-Rad).
- The variation in the prediction errors is improved between 2.8% and 13.3% for the datasets including at least one categorical variable. However, we did not record an improvement in the variability of prediction errors when there is no categorical variable.
- While SVM-Rad is the second-best-performing method, DNNs produce a close performance to them. However, ANFIS and linear regression methods do not deliver satisfactory real estate price prediction performance; hence, these methods are not recommended for use in practice. Consistent with the literature [41], linear regression methods perform better than ANFIS.
- Performance gain with HFRFs depends on the sample size and the number of categorical variables. This translates into the balance between the amount of categorical information and the total information in the sample. As the weight of categorical information increases, HFRFs are expected to provide more gain in performance.
- The computation time of HFRFs is sensitive to the sample size and the number of predictors. However, this does not pose a problem with the applicability of the HFRFs on large samples in a reasonable time frame.
- HFRFs are strongly recommended when any categorical variable is in the dataset for regression tasks. HFRFs are still beneficial in the absence of any categorical variables in data. However, this study does not observe any gain in the variability of prediction errors by HFRF when there is no categorical variable in the dataset.

The study’s main limitation is that the conclusions given here are limited within the scenarios resembled by the considered datasets. However, since the datasets are from six different markets with a sufficient variety of sample sizes, number of predictors, and number of categorical variables, there is no concern about the generalizability of the results.

The real estate price prediction datasets can also include outlier observations. Since we do not focus on the outlier problem in this study, we do not offer any conclusions on the performance of HFRF in the presence of outliers. Considering outliers in the presence of mixed predictor types is a future study.

Author contributions Both authors contributed equally to this work.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The experimental data that support the findings of this study are available in a GitHub repository at https://github.com/haydarde/HFRF.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Pryce G (2013) Housing submarkets and the lattice of substitution. Urban Stud 50(13):2682–2699. https://doi.org/10.1177/0042098013482502
2. Mayer M, Bourassa SC, Hoesli M, Scognamiglio D (2019) Estimation and updating methods for hedonic valuation. J Eur Real Estate Res 12(1):134–150. https://doi.org/10.1108/JERER-08-2018-0035
3. Chen Z, Cho S-H, Poudyal N, Roberts RK (2009) Forecasting housing prices under different market segmentation assumptions. Urban Stud 46(1):167–187. https://doi.org/10.1177/0042098008098641
4. Goodman AC, Thibodeau TG (2007) The spatial proximity of metropolitan area housing submarkets. Real Estate Econ 35(2):209–232. https://doi.org/10.1111/j.1540-6229.2007.00188.x
5. Adair AS, Berry JN, McGreal WS (1996) Hedonic modelling, housing submarkets and residential valuation. J Prop Res 13(1):67–83
6. Wang X, Wen J, Zhang Y, Wang Y (2014) Real estate price forecasting based on SVM optimized by PSO. Optik 125(3):1439–1443
7. Ćetković J, Lakić S, Lazarevska M, Žarković M, Vujošević S, Cvijović J, Gogić M (2018) Assessment of the real estate market value in the European market by artificial neural networks application. Complexity
8. Zhang H, Gao S, Zhang Y, Yang F (2015) Performance evaluation of the listed real estate companies in China based on fuzzy neural networks: the perspective of stakeholders. J Real Estate Pract Educ 18(2):195–215
9. Sarip AG, Hafez MB, Daud MN (2016) Application of fuzzy regression model for real estate price prediction. Malays J Comput Sci 29(1):15–27
10. Gnat S (2021) Impact of categorical variables encoding on property mass valuation. Procedia Comput Sci 192:3542–3550
11. Lee C (2022) Enhancing the performance of a neural network with entity embeddings: an application to real estate valuation. J Hous Built Environ 37(2):1057–1072
12. Ichino M, Yaguchi H (1994) Generalized Minkowski metric for mixed feature-type data analysis. IEEE Trans Syst, Man, Cybern 24(4):698–708
13. Bourassa SC, Hamelink F, Hoesli M, MacGregor BD (1999) Defining housing submarkets. J Hous Econ 8(2):160–183
14. Wilhelmsson M (2004) A method to derive housing sub-markets and reduce spatial dependency. Prop Manag 22(4):276–288
15. Goodman AC, Thibodeau TG (1998) Housing market segmentation. J Hous Econ 7(2):121–143
16. Goodman AC, Thibodeau TG (2003) Housing market segmentation and hedonic prediction accuracy. J Hous Econ 12(3):181–201
17. Manganelli B, Pontrandolfi P, Azzato A, Murgante B (2014) Using geographically weighted regression for housing market segmentation. Int J Bus Intell Data Min 13 9(2):161–177
18. Amédée-Manesme C-O, Baroni M, Barthélémy F, Des Rosiers F (2017) Market heterogeneity and the determinants of Paris apartment prices: a quantile regression approach. Urban Stud 54(14):3260–3280
19. Gabrielli L, Giuffrida S, Trovato MR (2017) Gaps and overlaps of urban housing sub-market: hard clustering and fuzzy clustering approaches. In: Appraisal: from theory to practice, pp 203–219. Springer, Berlin
20. Michaels RG, Smith VK (1990) Market segmentation and valuing amenities with hedonic models: the case of hazardous waste sites. J Urban Econ 28(2):223–242
21. Farber S (1986) Market segmentation and the effects on group homes for the handicapped on residential property values. Urban Stud 23(6):519–525
22. Watkins C (1999) Property valuation and the structure of urban housing markets. J Prop Invest Financ 17(2):157–175
23. Levkovich O, Rouwendal J, Brugman L (2018) Spatial planning and segmentation of the land market: the case of the Netherlands. Land Econ 94(1):137–154
24. Watkins CA (2001) The definition and identification of housing submarkets. Environ Plan A 33(12):2235–2253
25. Zurada J, Levitan A, Guan J (2011) A comparison of regression and artificial intelligence methods in a mass appraisal context. J Real Estate Res 33(3):349–388
26. Wu C, Sharma R (2012) Housing submarket classification: the role of spatial contiguity. Appl Geogr 32(2):746–756
27. Soaita AM, Dewilde C (2019) A critical-realist view of housing quality within the post-communist EU states: progressing towards a middle-range explanation. Hous Theory Soc 36(1):44–75
28. Wu Y, Wei YD, Li H (2020) Analyzing spatial heterogeneity of housing prices using large datasets. Appl Spat Anal Policy 13(1):223–256
29. Guo K, Wang J, Shi G, Cao X (2012) Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. Procedia Comput Sci 9:1299–1305
30. Helbich M, Brunauer W, Hagenauer J, Leitner M (2013) Data-driven regionalization of housing markets. Ann Assoc Am Geogr 103(4):871–889
31. Shi D, Guan J, Zurada J, Levitan AS (2015) An innovative clustering approach to market segmentation for improved price prediction. J Int Technol Inf Manag 24(1):15–32
32. Alkan T, Dokuz Y, Ecemiş A, Bozdağ A, Durduran SS (2023) Using machine learning algorithms for predicting real estate values in tourism centers. Soft Comput 27(5):2601–2613
33. Trawiński B, Telec Z, Krasnoborski J, Piwowarczyk M, Talaga M, Lasota T, Sawiłow E (2017) Comparison of expert algorithms with machine learning models for real estate appraisal. In: 2017 IEEE international conference on innovations in intelligent systems and applications (INISTA), pp 51–54. IEEE

34. Gu J, Zhu M, Jiang L (2011) Housing price forecasting based on genetic algorithm and support vector machine. Expert Syst Appl 38(4):3383–3386
35. Mach Ł (2017) The application of classical and neural regression models for the valuation of residential real estate. Folia Oeconomica Stetinensia 17(1):44–56
36. Sun Y (2019) Real estate evaluation model based on genetic algorithm optimized neural network. Data Sci J 18(36):1–9. https://doi.org/10.5334/dsj-2019-036
37. Rampini L, Cecconi FR (2021) Artificial intelligence algorithms to predict Italian real estate market prices. J Prop Invest Financ 40(6):588–611. https://doi.org/10.1108/JPIF-08-2021-0073
38. Aminuddin AJ, Maimun NHA (2022) A review on the performance of house price index models: Hedonic pricing model vs artificial neural network model. Int J Account 7(39):53–63
39. Bagnoli C, Smith H (1998) The theory of fuzzy logic and its application to real estate valuation. J Real Estate Res 16(2):169–200
40. Liu J-G, Zhang X-L, Wu W-P (2006) Application of fuzzy neural network for real estate prediction. In: International symposium on neural networks, pp 1187–1191. Springer, Berlin
41. Guan J, Zurada J, Levitan A (2008) An adaptive neuro-fuzzy inference system based approach to real estate property assessment. J Real Estate Res 30(4):395–422
42. Kuşan H, Aytekin O, Özdemir I (2010) The use of fuzzy logic in predicting house selling price. Expert Syst Appl 37(3):1808–1813
43. Gerek IH (2014) House selling price assessment using two different adaptive neuro-fuzzy techniques. Autom Constr 41:33–39
44. Del Giudice V, De Paola P, Cantisani GB (2017) Valuation of real estate investments through fuzzy logic. Buildings 7(26):1–22. https://doi.org/10.3390/buildings7010026
45. Yalpir S, Ozkan G (2018) Knowledge-based FIS and ANFIS models development and comparison for residential real estate valuation. Int J Strateg Prop Manag 22(2):110–118
46. Renigier-Biłozor M, Janowski A, d'Amato M (2019) Automated valuation model based on fuzzy and rough set theory for real estate market with insufficient source data. Land Use Policy 87:104021
47. Baser F, Demirhan H (2017) A fuzzy regression with support vector machine approach to the estimation of horizontal global solar radiation. Energy 123:229–240
48. Chakravarty S, Demirhan H, Baser F (2020) Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting. Appl Soft Comput 96:106535
49. Chakravarty S, Demirhan H, Baser F (2022) Modified fuzzy regression functions with a noise cluster against outlier contamination. Expert Syst Appl 205:117717
50. Celikyilmaz A, Turksen IB (2008) Enhanced fuzzy system models with improved fuzzy clustering algorithm. IEEE Trans Fuzzy Syst 16(3):779–794
51. Davé RN, Sen S (2002) Robust fuzzy clustering of relational data. IEEE Trans Fuzzy Syst 10(6):713–727
52. Chakravarty S, Demirhan H, Baser F (2022) Robust wind speed estimation with modified fuzzy regression functions with a noise cluster. Energy Convers Manage 266:115815
53. Bas E, Egrioglu E (2022) A fuzzy regression functions approach based on Gustafson-Kessel clustering algorithm. Inf Sci 592:206–214
54. Bas E (2022) Robust fuzzy regression functions approaches. Inf Sci 613:419–434
55. D'urso P, Massari R (2019) Fuzzy clustering of mixed data. Inf Sci 505:513–534
56. Guha S, Rastogi R, Shim K (2000) Rock: a robust clustering algorithm for categorical attributes. Inf Syst 25(5):345–366
57. Boehmke B, Greenwell BM (2019) Hands-on machine learning with R. CRC Press, New York
58. Faustino CP, Novaes CP, Pinheiro CAM, Carpinteiro OA (2014) Improving the performance of fuzzy rules-based forecasters through application of FCM algorithm. Artif Intell Rev 41:287–300
59. Sammut C, Webb GI (2011) Encyclopedia of machine learning. Springer, New York
60. Montgomery DC, Peck EA, Vining GG (2021) Introduction to linear regression analysis. Wiley, New York
61. Kuhn M (2022) Caret: classification and regression training. R package version 6.0-93. https://CRAN.R-project.org/package=caret
62. Riza LS, Bergmeir C, Herrera F, Benítez JM (2015) frbs: fuzzy rule-based systems for classification and regression in R. J Stat Softw 65(6):1–30
63. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2022) Cluster: cluster analysis basics and extensions. R package version 2.1.4. https://CRAN.R-project.org/package=cluster