The Journal of Real Estate Finance and Economics
https://doi.org/10.1007/s11146-022-09915-y

Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach

Juergen Deppner[1] · **Marcelo Cajias**[1,2]

Accepted: 17 June 2022

Abstract

Data-driven machine learning algorithms have initiated a paradigm shift in hedonic house price and rent modeling through their ability to capture highly complex and non-monotonic relationships. Their superior accuracy compared to parametric model alternatives has been demonstrated repeatedly in the literature. However, the statistical independence of the data implicitly assumed by resampling-based error estimates is unlikely to hold in a real estate context as price-formation processes in property markets are inherently spatial, which leads to spatial dependence structures in the data. When performing conventional cross-validation techniques for model selection and model assessment, spatial dependence between training and test data may lead to undetected overfitting and overoptimistic perception of predictive power. This study sheds light on the bias in cross-validation errors of tree-based algorithms induced by spatial autocorrelation and proposes a bias-reduced spatial cross-validation strategy. The findings confirm that error estimates from non-spatial resampling methods are overly optimistic, whereas spatially conscious techniques are more dependable and can increase generalizability. As accurate and unbiased error estimates are crucial to automated valuation methods, our results prove helpful for applications including, but not limited to, mass appraisal, credit risk management, portfolio allocation and investment decision making.

Keywords Hedonic modeling · Machine learning · Spatial autocorrelation · Spatial cross-validation · Mass appraisal · Automated valuation models

✉ Juergen Deppner
juergen.deppner@irebs.de

Marcelo Cajias
marcelo.cajias@irebs.de

[1] University of Regensburg, IRE|BS International Real Estate Business School, Universitätsstraße 31, 93053 Regensburg, Germany

[2] PATRIZIA AG, Fuggerstraße 26, 86150 Augsburg, Germany

Published online: 13 July 2022

Introduction

Real estate markets feature a spatial dimension that is pivotal to price and rent determination processes. The inherent spatial dependence in the economic value of assets cannot be ignored in hedonic models, as this would lead to spurious and biased results (Anselin, 1988; Can and Megbolugbe, 1997; Basu and Thibodeau, 1998). Guidance on how to account for spatial dependence in linear regression models is vast and remains the subject of many contributions to the hedonic and spatial econometric literature.

Moving from parametric hedonic regression techniques to the universe of non-parametric statistical learning methods, the literature has brought forth a growing body of evidence that machine learning algorithms can provide superior predictive performance for complex spatial regression problems, including various applications to house price estimation (e.g., Kok et al., 2017; Mullainathan and Spiess, 2017; Mayer et al., 2019; Hong et al., 2020; Pace and Hayunga, 2020; Bogin and Shui, 2020). To a great extent, the gains in explanatory power can be attributed to the flexibility of such models. This provides machine learning algorithms with the capability to exploit anisotropic and non-monotonic structures across space, which is of particular benefit when the spatial domain under investigation is a global one, as shown by Pace and Hayunga (2020). While this characteristic is a blessing when reproducing sample data, it can be a curse when predicting out-of-sample data since high flexibility is linked to overfitting, as demonstrated by Mullainathan and Spiess (2017) and Bogin and Shui (2020). Any kind of dependence structures in the data can exacerbate this problem, if not controlled for (Roberts et al., 2017). Thus, all the more surprising, little attention has been paid to the implications of spatial dependence in house prices and rents for the statistical validity of cross-validation (CV) errors, which are widely used to select and assess non-parametric models. For CV errors to be valid estimates of predictive performance, observations must be statistically independent of each other (Bishop, 1995; Brenning, 2005; Varma and Simon, 2006). This assumption is unlikely to hold in a real estate context (Bourassa et al., 2010) because "[…] error variance is not equal to zero but may be a function of spatial proximity among houses", as explained by Can and Megbolugbe (1997).

Two main problems arise when applying random resampling techniques to spatially dependent data. First, spatially structured variation in the residuals may be absorbed by non-causal regressors, consequently leading to the selection of overly complex and overfitted models that do not perform well with unseen data. Second, spatial autocorrelation between training and test observations provides the predictor with information that is assumed to be unavailable during model training, thus inflating estimates of predictive accuracy. In turn, this may hide the first problem as CV errors appear to be legitimate (Brenning, 2012; Roberts et al., 2017). Using such models to predict unseen data can result in substantially lower accuracy than is approximated by CV. When furthermore applied in combination with model-agnostic interpretation techniques to draw inference on the relation between housing value and property features, spurious regression can result in the identification of meaningless relationships.

In response, researchers from geoscientific modeling fields have developed spatially conscious resampling methods to address these problems. However, the adequacy of such techniques for hedonic house price models cannot be blindly assumed since prediction goals may differ. To the best of our knowledge, no research has thus far accounted for spatial dependence in algorithmic hedonic models by applying spatial resampling techniques. We believe that a sound understanding of the implications arising from spatial dependence is of great importance when applying machine learning algorithms to hedonic regression problems. Hence, this study aims to investigate the role of spatial autocorrelation on resampling-based model selection and model assessment of algorithmic hedonic methods, thereby evaluating the efficacy of spatial CV in contrast to non-spatial (i.e., random) CV. By doing so, we demonstrate the pitfalls of resampling-based performance evaluation and intend to raise awareness of the importance of spatially conscious resampling techniques in hedonic house price modeling.

Based on a cross-section of apartment rents in Frankfurt, Germany, we train and evaluate tree-based algorithms using spatial as well as non-spatial CV. We subsequently forecast out-of-sample data to assess the bias in error estimates associated with spatial autocorrelation. The results are put into a broader perspective by benchmarking our machine learning algorithms against a non-spatial ordinary least squares (OLS) and a spatial autoregressive framework, allowing for a relative comparison of bias and predictive performance. Lastly, we analyze the residual spatial autocorrelation to detect signs of overfitting to spatial structures in the data.

To make informed decisions, the precise estimation of house prices and rents is imperative to parties in the real estate industry, such as investors, developers, lenders or regulators. Since CV is commonly used as an “out-of-sample experiment” (Mullainathan and Spiess, 2017) to assess the predictive accuracy of algorithmic hedonic models, a systematic bias in error estimates may have adverse effects on the allocation of both debt and equity (Kok et al., 2017). The results of this study prove helpful in increasing the reliability and generalizability of CV errors, thus containing valuable implications for mass appraisal practices, credit risk management, portfolio allocation as well as investment decision making.

This paper is structured as follows: the section on "Hedonic Modeling of Spatially Structured House Prices and Rents" elaborates the problems of spatially structured data and their implications for hedonic analyses in the most commonly applied parametric as well as non-parametric regression frameworks, thereby providing an overview of the empirical literature on algorithmic hedonic approaches with a focus on applied resampling strategies. In the "Data and Methodology" section, the dataset is presented, followed by a description of the study design and the methodological approach. The empirical results are presented and discussed in the "Results" section and the final "Conclusion" section summarizes the findings of this study.

Hedonic Modeling of Spatially Structured House Prices and Rents

In his 1970 study on urban growth in the Detroit region, W. R. Tobler invoked his well-cited first law of geography, stating that the outcomes of nearby events correlate stronger than those of more distant events. Transferred to a housing context, this implies that the economic value of housing at any given location in geographic space depends, amongst other aspects, on the value of housing in neighboring locations. This deduction is well underpinned by spatial econometric as well as land economic theory for several reasons, such as spatial spillover effects (i.e., adjacency effects) and neighborhood effects (Can, 1992). Moreover, spatial clustering of house prices and rents may originate from a high correlation in the utility of the underlying houses derived from their structural characteristics (Basu and Thibodeau, 1998) and their fixed location in geographic space (Can and Megbolugbe, 1997; Osland, 2010), both of which determine the economic value of housing.

This leads to the conclusion that space is a fundamental factor that drives price formation processes in housing markets, subsequently resulting in two critical characteristics of housing market data: First, spatial autocorrelation, which is spatial dependence in price and rent determination processes; second, spatial heterogeneity, defined as the systematic variation in the behavior of price and rent formation processes across space (Anselin, 1988; Can and Megbolugbe, 1997). As stated by Osland (2010), one can assume that "[…] a mixture of these effects will be present in all housing market cross-section data". This poses important methodological implications on both parametric and non-parametric hedonic regression frameworks, which will be discussed below.

Parametric Hedonic Models

The economic theory of hedonic pricing in a housing context dates to Rosen (1974), who implemented the derivation of implicit prices of hedonic characteristics using a least squares estimator. Due to their efficiency and ease of interpretability, least squares estimators have established themselves as the standard econometric approach to hedonic house price modeling. Likewise, the concept of hedonic price modeling has been successfully transferred and applied to the determination of apartment rents (e.g., Sirmans et al., 1989; Sirmans and Benjamin, 1991; Allen et al., 1995).

Implications of Spatial Dependence

Independent and identically distributed errors with a zero mean and constant variance are crucial Gauss-Markov assumptions to produce consistent and efficient estimates in a least squares context (Wooldridge, 2016). Spatial autocorrelation and spatial heterogeneity in the residuals violate these assumptions, resulting in unreliable confidence intervals and biased t-statistics, which lead to spurious statistical inference (Anselin, 1988; Basu and Thibodeau, 1998). Depending on the underlying

spatial processes causing spatial effects, even point estimates might be biased and lead to erroneous results (Pace and LeSage, 2010).

Moreover, endogeneity is likely to occur due to omitted variable bias, measurement errors in the independent variables or feedback loops induced by adjacency effects. The explanatory power of spatial effects not explicitly reflected in the model specification is picked up by the error term or by covarying explanatory variables instead, leading to biased estimates and non-normality of the errors (LeSage and Pace, 2009). Even if spatial controls are included in the regression equation, the assumption of linearity in the functional form requires their relationship with the dependent variable to be constant across space. However, in the real world, such relationships are seldom linear and monotonic nor isotropic since slopes are likely to vary by distance and direction (Osland, 2010). Non-stationarity across space will persist as spatial heterogeneity in the residuals, violating the crucial OLS assumption of homogeneity in the errors.

It can be concluded that both theory, as well as empirical research, suggests that the Gauss-Markov assumptions underlying traditional OLS estimators cannot be naturally presumed in a real estate context (Can and Megbolugbe, 1997; Bourassa et al., 2010; Cajias and Ertl, 2018), resulting in biased and inconsistent least squares estimates as well as spurious inference.

Accounting for Spatial Dependence

Spatial autoregressive models are the typical statistical instruments to consider spatial effects in parametric frameworks. They control for spatial dependence by explicitly incorporating the underlying correlation structures as spatial lags in their functional form (see Cliff and Ord, 1973; Anselin, 1988; Cressie, 1993; Manski, 1993; Kelejian and Prucha, 1998; LeSage and Pace, 2009). As the necessity to account for spatial effects in linear models is well understood, spatial autoregressive, as well as other spatial modeling alternatives, are widely applied and discussed in a real estate context (e.g., Pace and Gilley, 1997; Case et al., 2004; Militino et al., 2004; Valente et al., 2005; Bourassa et al., 2007, 2010; Osland, 2010; Füss and Koller, 2016; Cajias and Ertl, 2018). Although such methods have been demonstrated to reduce residual spatial autocorrelation if applied carefully, the models continue to be linear, limiting their ability to capture highly complex and multi-dimensional relationships in the formation of house prices and apartment rents.

Non-Parametric Hedonic Models and Cross-Validation

As the real world can be more accurately described by logarithmic, exponential or step functions, the increasing availability of data together with technical progress in computational power has triggered the consideration of more flexible non-parametric machine learning methods for the problem of hedonic house price and rent modeling. In principle, such data-driven approaches do not rely on any a priori assumptions about the distributions of the errors, nor the functional form $f(x)$ that explains i house prices y_i using j regressors x_{ij}, but approximate the shape of $f(x)$ by fitting a

spline to the data (James et al., 2013). However, it is to mention that the lack of a pre-defined additive functional form comes at the cost of inferential insights, as the prediction rules of the algorithms are opaque and cannot be directly interpreted due to their complexity. Moreover, their high flexibility makes them prone to overfitting, which is why modern statistical tools rely on resampling methods for model selection (i.e., selecting an appropriate level of regularization to approximate the shape of $f(x)$ during hyperparameter tuning) and for model assessment (i.e., assessing the test error rate of the selected model $\hat{f}(x)$ to evaluate its performance).

Resampling is typically performed using cross-validation, during which observations are randomly partitioned into mutually exclusive training and test subsets, whereby the predictor is fitted on the training data and evaluated on the respective test data (Stone, 1974; Snee, 1977). This concept can be thought of as creating "[…] an out-of-sample experiment inside the original sample", as described by Mullainathan and Spiess (2017).

In its most simple form, cross-validation randomly divides the data into two subsets, that is a training set and a validation (i.e., holdout) set based on a given percentage split. Subsequently, the model is fitted on the training sample, which is then used to predict the responses from the validation sample. This holdout strategy has been widely applied in the algorithmic hedonic house price literature. Worzala et al. (1995), Din et al. (2001), Peterson and Flanagan (2009) as well as Chiarazzo et al. (2014) use this technique for model assessment of artificial neural networks (ANNs), and Yoo et al. (2012), Kok et al. (2017) as well as Pérez-Rave et al. (2019) to validate different tree-based algorithms such as regression trees (RT), random forest regression (RFR), gradient tree boosting (GTB) and extreme gradient boosting (XGB). Lam et al. (2009), Antipov and Pokryshevskaya (2012), McCluskey et al. (2013) and Bogin and Shui (2020) benchmark different machine learning approaches, including support vector regression (SVR), shrinkage estimators (e.g., LASSO) as well as neural networks and tree-based methods using error estimates from a holdout sample. The applied split ratios vary between 60 to 80% for the training data and 40 to 20% for the test data, respectively. Such holdout strategies are computationally inexpensive and easy to implement. However, the test error rate may be heavily dependent on which observations are held out for validation and used for training, resulting in a potential bias in the error estimates (James et al., 2013).

To address this form of bias, k-fold cross-validation has been introduced to the statistical community (Lachenbruch and Mickey, 1968; Efron, 1983). During k-fold cross-validation, the data is partitioned into k mutually exclusive subsets of equal size. Subsequently, each of the k folds is once used as a test set and the remaining $k - 1$ folds are used to calibrate the model, consequently yielding k estimates of prediction error that are then averaged. This strategy attempts to generate more robust and reliable approximations of out-of-sample predictive performance. In a real estate hedonic context, k-fold cross-validation has gained in popularity during the past decade. Park and Bae (2015), Gu and Xu (2017), Čeh et al. (2018), Chin et al. (2020) as well as Pace and Hayunga (2020) apply k-fold cross-validation to evaluate the performance of tree-based methods. Applications to a broader spectrum of machine learning algorithms, including ANNs, SVR, k-nearest neighbors, shrinkage estimators as well as ensembles of regression trees using boosting and bagging

techniques, can be found in Zurada et al. (2011), Mullainathan and Spiess (2017), Baldominos et al. (2018), Mayer et al. (2019), Hu et al. (2019), Ho et al. (2021), Cajias et al. (2021), as well as Rico-Juan and Taltavull de La Paz (2021). In all those studies, the applied number of folds is either five or ten, except for Mullainathan and Spiess (2017), who set k equal to eight.

Machine learning algorithms excel parametric models in the identification of complex non-linear relationships between the value of real estate and property characteristics, but they are also criticized for their black box character as their inner workings are opaque and comprehensibility as well as direct interpretation of the models are impeded by their complexity. Although recent developments allow insights into these opaque black boxes via model-agnostic interpretation techniques (see Rico-Juan and Taltavull de La Paz, 2021; Lorenz et al., 2022), this constitutes a limitation for the use of machine learning in both, academic research and practice. As stated by Rico-Juan and Taltavull de La Paz (2021), data-driven models might not be consistent with theoretical expectations and may thus have no economic meaning when identified relationships are spurious because "[…] the laws of economics (and the explanatory models that show causality) as well as the limitations econometrics imposes on the models [are ignored], leaving these systems free to make inferences from a combination of data". Spatial autocorrelation is one such econometric constraint that is typically ignored in machine learning applications to house price data.

Implications of Spatial Dependence

Although cross-validation proves to be decisive in reducing bias in error estimates, any kind of resampling technique is subject to one central assumption. By conducting an out-of-sample experiment that draws random observations from the data that are then used to approximate prediction errors, CV attempts to simulate unseen data. For cross-validation to yield unbiased prediction error estimates, statistical independence between training and test observations is required (Bishop, 1995; Brenning, 2005; Varma and Simon, 2006). Consequently, the meaningfulness of the resulting CV errors as a robustness test for out-of-sample predictive performance is highly reduced in spatial modeling fields where the independence assumption is violated (Le Rest et al., 2014). More specifically, spatial dependence structures in the data cause two main problems in the workflow of machine learning algorithms.

First, regressors are often covarying with unexplained spatial dependence structures in the residuals. During hyperparameter tuning (i.e., model selection), the model may overfit these spatial structures to non-causal, but covarying regressors in the attempt to optimize model performance, thereby reducing or completely absorbing unexplained structured covariation from the residuals. This may lead to a selection of overly complex models that can reproduce the training data but may not generalize well to unseen data (Can and Megbolugbe, 1997; Le Rest et al., 2014; Roberts et al., 2017; Meyer et al., 2019). Second, when the sample from the validation fold is drawn from the same dependence structure as the training folds due to spatial proximity, the predictor may obtain information from the spatially autocorrelated test data that is assumed to be unavailable to the model during training. This

unauthorized glimpse on the test data results in approximations of predictive power (i.e., model assessment) that may be overly optimistic and thus not representative for unseen data with different spatial structures (Picard and Cook, 1984; Hastie et al., 2009; Le Rest et al., 2014; Trachsel and Telford, 2016; Roberts et al., 2017; Schratz et al., 2019; Lovelace et al., 2019).

In a real estate context, such cases of poor out-of-sample predictive performance were, for instance, reported by Mayer et al. (2019) for their random forest. Bogin and Shui (2020) report significant overfitting using a random forest with a deviation of 22.1 percentage points in the R^2 compared to in-sample cross-validation errors. Mullainathan and Spiess (2017) demonstrated a similar bias in cross-validation errors for both bagging and boosting with 39.6 percentage points discrepancy in the R^2 of the random forest and 8.7 percentage points in the boosting trees.

This is problematic because, on the one hand, accuracy implied by cross-validation may lead to unjustified confidence in a model's predictive power that cannot be guaranteed when making predictions with unseen data. On the other hand, identified relationships may be spurious and can result in fallacious inferential conclusions when model-agnostic interpretation techniques are applied to interpret the pricing processes of the algorithms.

Accounting for Spatial Dependence

One possible approach to account for spatial dependence structures in the selection and assessment of non-parametric models is by using resampling techniques that split the data strategically by considering spatial proximity among observations rather than randomly. Spatial partitioning can be designed in many ways. However, the general concept is to increase independence between training and test data by clustering or blocking the individual folds across space or by removing training data within a specific distance band of each test point, such that performance is evaluated on more distant events that tend to be less correlated to the training sample (Tobler, 1970; Trachsel and Telford, 2016; Roberts et al., 2017). In a spatial context, such approaches have been introduced to the statistical community under many different terms. These include "spatial cross-validation" (Brenning, 2005, 2012), "spatial leave-one-out cross-validation" (Le Rest et al., 2014), "*h*-block cross-validation" (Trachsel and Telford, 2016), "spatial *k*-fold cross-validation" (Pohjankukka et al., 2017), "spatial buffering", "spatial blocking", "environmental blocking" (Valavi et al., 2018) and "leave-one-cluster-out cross-validation" (Meyer et al., 2019). Following the methodology and terminology of Brenning (2012), we will continue naming this concept spatial cross-validation in the remainder of this study. The conceptual difference between random and spatial partitioning of folds during cross-validation is visualized in Fig. 1 based on an example with three folds.

Since "[…] the adequacy of non-spatial partitioning techniques for spatial datasets can be questioned" as stated by Schratz et al. (2019), spatial cross-validation methods are widely used in scientific fields such as climatology (Trachsel and Telford, 2016), ecology (Bahn and McGill, 2007; Schratz et al., 2019), remote sensing (Brenning, 2012; Meyer et al., 2019) and geosciences (Brenning, 2005). The need for spatial resampling has been stressed repeatedly in those fields, yet, its suitability

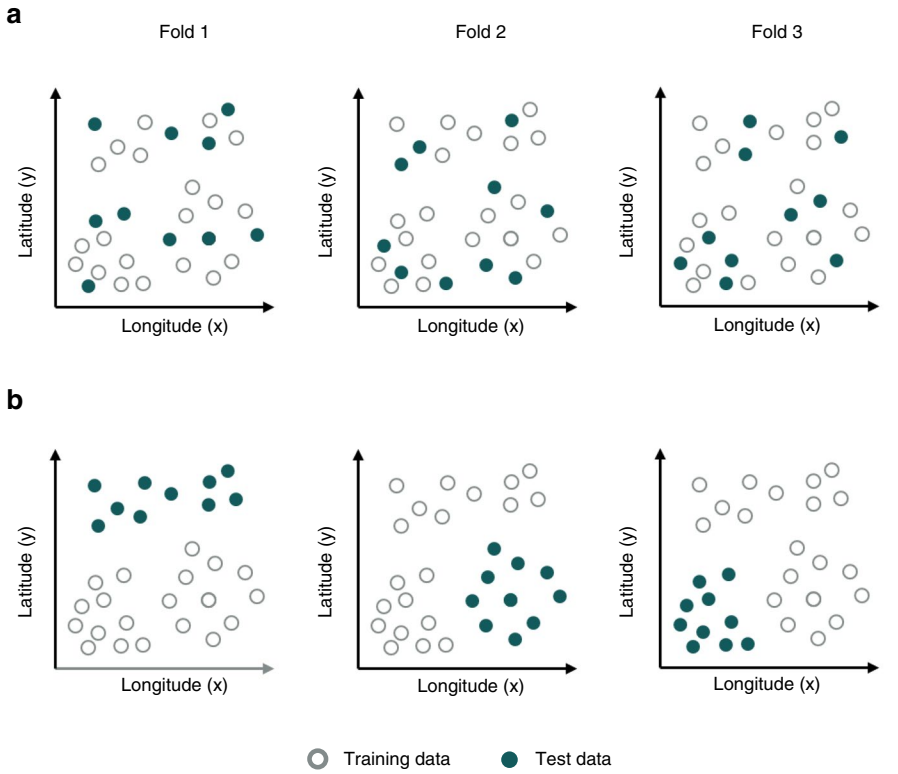

Fig. 1 **a** Random Partitioning of Folds. **b** Spatial Partitioning of Folds. Notes: This figure depicts the conceptual difference between random partitioning and spatial partitioning of folds using *k*-means clustering during cross-validation

and efficacy for real estate data has not been investigated thus far. Findings from other disciplines cannot be easily transferred to a real estate context since the objectives and circumstances under which predictive models are designed may differ. Despite compelling arguments to use spatial cross-validation when modeling data in geographic space, there is good reason to be cautious when applying such methods to a real estate hedonic context.

Spatial partitioning may hide entire ranges or functional relationships of regressors during training, thereby introducing extrapolation to a model that is supposed to interpolate and consequently resulting in overly pessimistic estimates of prediction errors during model assessment (Snee, 1977; Roberts et al. 2017). However, the model can also be underfitted when the selected level of regularization is too high which may result in poor predictions (Kok et al., 2017). This dichotomy is particularly pronounced in real estate related regression tasks where high levels of spatial dependence tend to exist between observations but the prediction goal is usually dominated by the interpolation of existing properties within a delineated market. In cases where both the degree of spatial autocorrelation in the data and the

extrapolation range are low, conventional cross-validation techniques that split the data randomly may be appropriate for performance optimization and evaluation. However, in situations where a model predicts outside the spatial domain of the training data and correlation structures between the residuals and non-causal regressors differ from the structures that were overfitted to non-causal regressors, random partitioning may yield unsatisfactory results (Bahn and McGill, 2007; Roberts et al., 2017).

As shown by Gröbel and Thomschke (2018) and Hong et al. (2020), prediction accuracy also depends on the spatial density of the sample locations. This is in line with Bahn and McGill (2007), who state that "[…] the sparser the existing coverage of sample locations for the dependent variable, the worse the spatial interpolation will perform". In other words, non-spatial CV may perform well in samples with a high spatial density but not so well if the distribution of observations across space is sparse, as this typically increases extrapolation. Furthermore, this implies that bias in prediction accuracy may be a function of distance from the city center since observations usually become sparser farther outside where housing structures are less dense and markets tend to be less active (Gröbel and Thomschke, 2018).

Although many studies on hedonic machine learning approaches exist, spatial cross-validation has so far not been applied to a real estate context, let alone to algorithmic hedonic house price and rent estimation problems. Reported cross-validation errors are almost consistently lower than errors of alternative parametric methods, such as least squares or spatial autoregressive frameworks. In particular tree-based ensemble learners, such as bagging (Breiman, 2001) and boosting (Friedman, 2001), have been shown to be most promising for house price estimation compared to alternative machine learning methods (see Antipov and Pokryshevskaya, 2012; Kok et al., 2017; Mullainathan and Spiess, 2017; Baldominos et al., 2018; Mayer et al., 2019, Hu et al., 2019; Ho et al., 2021). Pace and Hayunga (2020) find evidence that the gains in explanatory power achieved by boosting and bagging algorithms are mainly attributable to the exploitation of spatial structures in the data. Consistent with the previously outlined logic presented by Bahn and McGill (2007) as well as Gröbel and Thomschke (2018), they find the error variance of bagging to increase the farther the model extrapolates to a global domain which could indicate that the model is overfitted to the spatial structures of more frequently observed houses in central districts. This notion is in line with Bogin and Shui (2020) who found a significant degree of overfitting in their random forest measured by a holdout strategy using appraisal records of homes in rural areas.

The extensive scientific debate about spatial dependence in real estate together with the concurrent, steadily growing corpus of literature on machine learning applications for house price and rent predictions motivates us to assess the sign and magnitude of potential bias associated with spatial dependence when using conventional CV methods for model selection and model assessment. Moreover, we investigate whether spatial cross-validation is an appropriate technique to account for spatial autocorrelation in apartment rents when using predictive machine learning algorithms, although the primary intention is to interpolate within a delineated spatial polygon.

Data and Methodology

We first train and cross-validate tree-based algorithms using a cross-section of apartment rents, thereby applying random as well as spatial partitioning during the cross-validation procedure for both, model selection and model assessment. With everything else remaining equal, there should be no substantial difference in the selected hyperparameters nor the cross-validation errors between spatial and non-spatial models if the assumption of spatial randomness was fulfilled. In a second step, we calculate the out-of-sample predictive performance of the models by estimating the data from a holdout sample one quarter ahead. We then analyze the difference between in-sample cross-validation errors and the true out-of-sample prediction errors to assess the bias associated with the respective partitioning techniques. A non-spatial linear model, as well as spatial autoregressive models, are used as points of reference. Third, we evaluate the deviation in bias when excluding spatial control variables from the model specification. Based on the hypotheses elaborated in section two, we would expect the bias to increase in non-spatial modeling frameworks when spatial information is absent due to overfitting unexplained spatial dependence structures in the data to covarying but non-causal regressors. This will be more closely evaluated in a fourth step by analyzing the residual spatial autocorrelation in all model alternatives.

Data Description

Our sample consists of a pooled cross-section of apartment rents from the Frankfurt residential market spanning the period from January 2019 through March 2020. The data were sourced from German multiple listing systems (MLS) and are confined to apartment rentals excluding single, semi-detached and terraced houses, student apartments, senior living accommodations, furnished co-living spaces and short-stay apartments. Data cleaning was performed to account for duplicates, missing values and erroneous data points. The final sample comprises a total of 9256 asking rents observed on a monthly scale, including the properties' most important structural attributes and equipment as well as their coordinates. A typical way to reflect differences in demand for locations in parametric models is to include district fixed effects by means of location dummies such as pre-defined submarkets (e.g., Bourassa et al., 2003, 2007) or attractiveness zones (e.g., Doszyń, 2020) that are specified by real estate experts. In non-parametric machine learning models, the inclusion of spatial coordinates (i.e., latitude and longitude) facilitates the identification of relevant submarkets based on spatial patterns in the data without the need to provide specific location zones. This allows a model to construct more local sub models for the identified areas (see Pace and Hayunga, 2020). The use of continuous coordinates is more efficient because it is computationally less expensive than a matrix of location dummies while at the same time, coordinates have a finer resolution, so the models are not forced into using pre-defined spatial polygons that limit their flexibility. Also, having too many dummy variables or too few observations per location zone may favor overfitting the models. Since our data only contain postcode areas

that have very limited economic meaning, we refrain from the inclusion of location zones and include the observations' coordinates by means of latitude and longitude. Moreover, distances to nearby amenities were added using an Open Street Maps API to control for locational and neighborhood effects.

The building age was calculated relative to the year 2018, and values with a construction date before 1900 were trimmed to avoid disproportionate leverage of those observations. The entry date was transformed into a decimal number in years, and logarithmic transformations were used for the apartment rent and living area. The summary statistics of the features univariate distributions is presented in Table 1. The number of entries in each month is distributed uniformly throughout the sample period without a significant time trend in apartment rents, as shown in Table 2.

The spatial distribution of the data in Fig. 2a does not seem to exhibit any distinct location bias, albeit the spatial density of observations increases toward the city center. As described by Gröbel and Thomschke (2018), this is not surprising as building structures are denser in central areas, which are, moreover, predominantly occupied by younger and more mobile tenants, resulting in higher fluctuation rates and subsequently more frequent rental offers compared to the outskirts. The average distances to the 1, 5, 10, 30, and 100-nearest neighbors amounts to 0.02, 0.04, 0.06, 0.13 and 0.30 km respectively. Aggregated on a ZIP-code level, Fig. 2b indicates that more expensive apartments tend to be clustered in the city center and along the north-south axis. More formally, spatial clustering of apartment rents is confirmed by the semi-variance of the log rent as depicted in Fig. 3a. The empirical Matérn semi-variogram model suggests a spatial autocorrelation range of 0.58 km, which is the distance up to which spatial dependence between observations persists in the data (Cressie, 1993). In other words, an apartment in our sample has on average 168 neighbors that do not satisfy the assumption of independence. This number increases with spatial density and vice versa. The distribution of neighbors within the spatial autocorrelation range is presented in Fig. 3b.

Methodological Approach

Parametric Models

We use an ordinary least squares (OLS) estimator as a non-spatial parametric benchmark model. Written in matrix notation, the multiple linear regression model follows a log-linear functional form of the relationship

$$Y = \alpha + X\beta + \varepsilon \tag{1}$$

with Y being the response vector with n observations of log-transformed apartments rents, α being a fixed intercept, X representing regressor matrix with n rows and p columns, β being the corresponding $n \times 1$ coefficient vector and ε being the random error term vector of length n.

Our modeling approach is based on the principle to avoid overfitting and bias in error estimates to isolate the bias originating from spatial dependence. We thus follow Harrell (2015) and Mayer et al. (2019) and exclude only regressors with almost

Table 1 Summary Statistics

Variable	N	Mean	Median	SD	Min	Max
Continuous						
Rent per month [Euro]	9256	1088.77	940.00	647.97	190.00	10,000.00
Living Area [sqm]	9256	75.20	70.00	35.07	10.00	440.00
Age [years]	9256	44.59	46.00	39.34	−2.00	118.00
Entry date [years]	9256	0.65	0.67	0.35	0.08	1.25
Latitude	9256	50.12	50.12	0.02	50.08	50.21
Longitude	9256	8.66	8.66	0.05	8.49	8.78
Discrete						
Rooms	9256	2.55	2.50	1.00	1.00	8.50
Floor	9256	2.54	2.00	2.82	−0.50	39.00
Dummies [1 = yes, 0 = no]						
Bathtub	9256	0.53	1.00	0.50	0.00	1.00
Refurbished	9256	0.22	0.00	0.41	0.00	1.00
Built-in kitchen	9256	0.71	1.00	0.45	0.00	1.00
Balcony	9256	0.65	1.00	0.48	0.00	1.00
Parking	9256	0.48	0.00	0.50	0.00	1.00
Elevator	9256	0.50	1.00	0.50	0.00	1.00
Terrace	9256	0.13	0.00	0.34	0.00	1.00
Distances						
NUTS centroid [km]	9256	3.65	3.68	1.87	0.01	10.84
Bakery [km]	9256	0.39	0.26	0.41	0.00	1.61
Bar [km]	9256	0.73	0.52	0.64	0.00	2.54
Biergarten [km]	9256	1.16	0.97	0.77	0.02	3.10
Café [km]	9256	0.36	0.25	0.33	0.00	1.31
School [km]	9256	0.31	0.28	0.17	0.02	0.75
Supermarket [km]	9256	0.26	0.22	0.17	0.00	0.75
Bus station [km]	9256	3.13	2.77	1.54	0.09	7.56

This table reports the univariate distributions of 9256 asking rents of residential apartments listed between January 2019 and March 2020 in Frankfurt (Germany), and their observed characteristics after data cleaning. The entry date is represented as a decimal number in years, the building age is calculated relative to the year 2018 and is trimmed for buildings constructed before the year 1900, distances are calculated as the Euclidean distance to the apartment in kilometers, binary variables indicate whether a characteristic is included in the apartment (1) or not (0). N: number of observations, SD: standard deviation, Min: minimum value, Max: maximum value

no predictive power from the hedonic equation but, at the same time, refrain from the inclusion of interaction or quadratic terms to keep the models simple. We test the null hypothesis of spatial randomness in the OLS residuals by calculating the Moran's *I* statistic (Cliff and Ord, 1973) and account for potential spatial dependence using the spatial econometric toolbox.

Opposed to the spatial cross-validation technique (where spatially autocorrelated information is explicitly excluded from the model), the mechanism of spatial econometric models works the exact opposite way by explicitly mapping spatial

Table 2 Listings per Month

Entry Date	N	Mean Rent [Euro]	Mean Rent [Euro/sqm]
Jan-19	632	1129.57	14.57
Feb-19	586	1116.31	14.34
Mar-19	677	1081.28	14.20
Apr-19	576	1118.35	14.39
May-19	685	1135.66	14.54
Jun-19	602	1084.38	14.41
Jul-19	746	1083.23	14.22
Aug-19	755	1014.57	14.20
Sep-19	602	1177.08	14.37
Oct-19	633	1088.54	14.26
Nov-19	613	1005.09	13.85
Dec-19	392	1021.86	14.48
Jan-20	600	1101.10	14.82
Feb-20	611	1096.45	14.75
Mar-20	546	1068.98	14.73

This table reports the occurence of apartment listings throughout the sample period from January 2019 to March 2020 on a monthly scale with the respective mean absolute rent in Euro per month and the mean rent in Euro per square meter per month. N: number of observations

interactions among neighboring observations as spatial lag terms in the functional form of the relationship. The spatial weight matrix W formally defines the spatial relationship between observations. To identify the source of the underlying processes causing spatial effects in the data, a model specification search is conducted following the general to specific approach advocated by LeSage and Pace (2009) and LeSage (2014) starting from the Spatial Durbin Model (SDM)

$$Y = \alpha + \rho WY + X\beta + WX\theta + \varepsilon \tag{2}$$

as well as the Spatial Durbin Error Model (SDEM).

$$Y = \alpha + X\beta + WX\theta + u \tag{3}$$

$$u = \lambda Wu + \varepsilon \tag{4}$$

Subsequently, we perform likelihood-ratio tests to challenge the relevance of the autoregressive coefficients ρ, θ, and λ from the SDM and the SDEM against more specific spatial model alternatives that include only one out of the two interactions respectively (Anselin, 1988; Anselin et al., 1996). As stated by LeSage and Pace (2009), this top-down approach has the advantage that the SDM still produces unbiased coefficient estimates even when the true data-generating process is a more specific model.

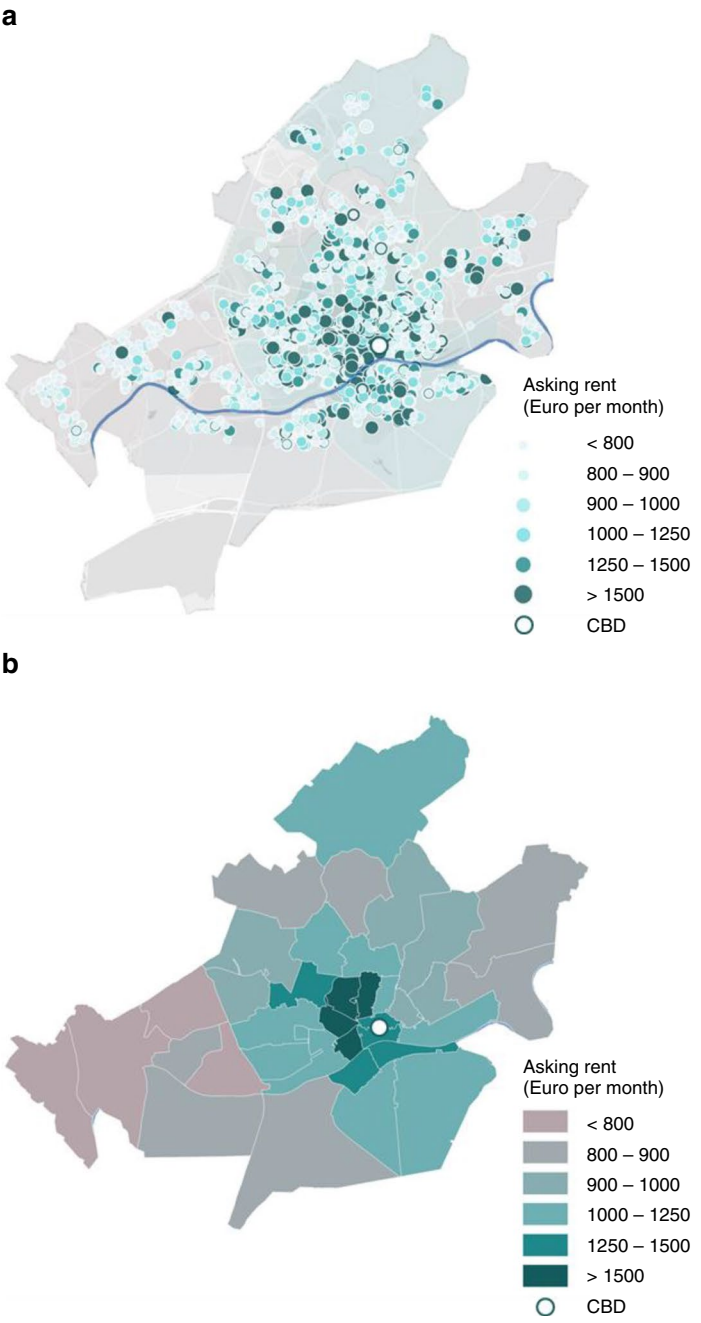

Fig. 2 **a** Spatial Distribution of Apartment Rents. **b** Mean Apartment Rents on ZIP-Code Level. Notes: The upper map depicts the absolute monthly asking rent in Euro per month of each individual observation in our sample of 9256 listings between January 2019 and March 2020 in Frankfurt. The bottom map shows the respective mean asking rents in Euro per month aggregated on a ZIP-code level

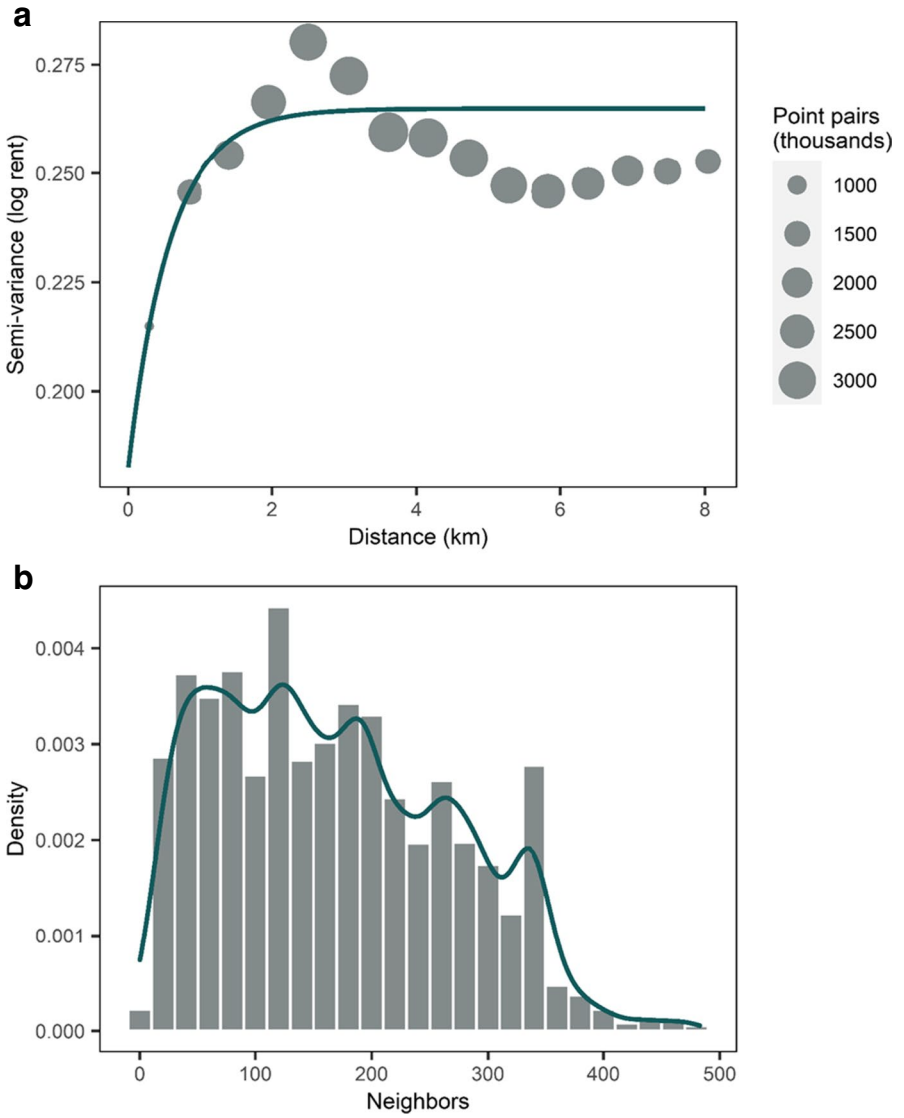

Fig. 3 **a** Semi-variogram of the log rent. **b** Distribution of Neighbors within the Spatial Autocorrelation Range. Notes: The empirical Matérn semi-variogram model suggest a spatial autocorrelation range of 0.58 km, which is the distance up to which spatial autocorrelation persists in the data. The histogram presents the distribution of neighbors within the spatial autocorrelation range

Non-Parametric Models

Among a wide variety of algorithmic hedonic methods evaluated in comparative studies, ensembles of regression trees using bagging and boosting techniques have consistently shown the most promising results concerning predictive power

(see Antipov and Pokryshevskaya, 2012; Kok et al., 2017; Baldominos et al., 2018; Mayer et al., 2019; Hu et al., 2019; Ho et al., 2021; Bogin and Shui, 2020).

The idea behind a regression tree is to stratify the feature space into a set of M disjoint intervals $R_1, R_2, \ldots, R_M$, each of which is assigned a constant c_m as predicted value, being referred to as the leaf or terminal node of the tree (Breiman et al., 1984). Intervals are created by recursive binary partitioning at the nodes t_p, choosing a split-point s of a particular feature x_j in the process of solving a minimization problem that can be expressed as

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}\left(y_i-c_1\right)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}\left(y_i-c_2\right)^2\right] \tag{5}$$

under the conditions that

$$R_1(j,s)=\left\{X|X_j\le s\right\} \tag{6}$$

$$R_2(j,s)=\left\{X|X_j> s\right\} \tag{7}$$

following the notation of Hastie et al. (2009). Each observation is subsequently passed down the tree branches by making binary decisions at each split following the feature values until the data point has reached its final leaf. As single regression trees tend to overfit easily and do not perform well on unseen data, we focus on ensembles of regression trees, more specifically the bagging-based random forest regression introduced by Breiman (2001) and the extreme gradient boosting algorithm developed by Chen and Guestrin (2016) which is an extension of the gradient tree boosting method dating to Friedman (2001).

The bagging algorithm grows a forest of many individual but slightly different trees b using bootstrapped training samples. Instead of pruning the trees, which is typically done to counteract the overfitting of individual regression trees, the trees in a forest are grown deeply, resulting in more terminal nodes with fewer observations being allocated to the same constant c_m as the predicted value. The minimum node size min_{node}, which is the smallest possible number of observations in each leaf, determines the depth of the tree. Deep trees can lead to overfitting, thus reacting very sensitively to changes in the training sample, whereas shallow trees may not pick up information from the data adequately, hence producing models which are underfitted (Kok et al., 2017). Consequently, the relatively deep trees in a forest have a high variance but low bias. The variance is then removed by averaging over the results of the b bootstrap trees to increase robustness (Breiman, 1996; James et al., 2013). Unlike conventional forests, a random forest considers only a randomly selected subset of m predictors from all available predictors p at each split, thereby introducing an additional source of variation into the model to counteract overfitting.

Boosting works somewhat similar, but unlike bagging, where trees are grown simultaneously and independently, boosted trees are grown sequentially using the residuals' information content from preceding trees to continue learning. At each sequence, a new regression tree is fitted to the residuals of the previous tree and is added into the fitted function, thereby iteratively updating the model (James et al.,

2013). Model fit is improved with each iteration until the number of boosting rounds is exhausted. To avoid overfitting boosted trees, a shrinkage parameter η, also referred to as learning rate, is used to slow down the learning process by making the error corrections in each round more conservative. Like the random forest, the extreme gradient boosting algorithm is a more regularized alternative of the gradient boosting technique in the sense that it attempts to decorrelate the individual trees by using only a randomly subsampled portion $\frac{m}{p}$ of features in each round to increase the robustness of the boosting trees.

Generally speaking, model fit tends to increase with higher flexibility, such as a lower min_{node}, or a lower η. However, this is tied to the risk of overfitting the training data, subsequently resulting in poor generalizability of the models. Thus, a level of regularization has to be chosen to limit the flexibility of a model. As manual choice of these hyperparameter combinations is arbitrary and unlikely to yield satisfactory results, model selection is typically conducted using data-driven optimization by iteratively testing different hyperparameter combinations within a pre-defined search space and evaluating their performance based on cross-validation errors in the attempt to minimize a given loss function. Eventually, the set of hyperparameters yielding the lowest cross-validation error rate (that is, the best performance in the out-of-sample experiment) is chosen as the optimal hyperparameter combination (James et al., 2013). This approach is called grid search optimization and is a widely used algorithm for automated hyperparameter tuning. For computational reasons, we restrict our search space to the three main tuning parameters of each model described above. For the random forest, these are the number of bootstrap trees b, the number of available features m at each split and the minimum node size min_{node} in each terminal node. Likewise, the number of boosting rounds n_{rounds}, the column subsample $\frac{m}{p}$ and the shrinkage parameter η are equally important to the boosting algorithm. As optimization criterion, we adopt the minimization of squared residuals from the least squares estimator.

Performance Evaluation

We measure predictive performance (i.e., the true error rate) of our models by predicting out-of-sample data from the first quarter of 2020, which is referred to as the "holdout sample" in the remainder of this study. The remaining "in-sample" data of 2019 is used to calibrate the models and approximate their out-of-sample predictive performance (i.e., the expected error rate). The resampling strategies used for the steps of model selection as well as model assessment are outlined below.

In the parametric world, model selection is performed manually by specifying a functional form of the estimator a priori rather than following a data-driven approach to maximize fit. Moreover, the flexibility of such models is usually restrained by linearity assumptions that make overfitting less of a problem. Thus, prediction errors of linear models are typically estimated by simple re-substitution of the data used for model fitting (Efron, 1983; Simon, 2007). We follow this standard statistical approach to approximate the predictive accuracy of the parametric benchmark

models and calculate the true error rate by regressing the holdout data using the estimated parameters.

As elaborated earlier, data-driven approaches require resampling methods such as cross-validation to calculate fair estimates of predictive performance. To isolate the effects of spatial dependence and allow for a fair comparison between random and spatial partitioning, the resampling strategy is designed with the principle to eliminate any bias resulting from sources other than spatial autocorrelation that could potentially distort our results. Thus, we perform *k*-fold cross-validation for model selection and model assessment to avoid that error estimates are biased by chance due to a specific training or validation set.

It is worth mentioning that the choice of *k* is associated with a bias-variance trade-off, that is, the bias becoming smaller with each additional fold whereas the variance of the error estimates increases at the same time due to a higher correlation of the training sets (Hastie et al., 2009). As suggested by theory and empirical research, a *k* of five or ten proves to be a reasonable compromise in this trade-off, whereby a value of five is only recommended for very large datasets to ensure enough observations for model training (Breiman and Spector, 1992; Kohavi, 1995; Hastie et al., 2009; James et al., 2013). With the primary aim to isolate the effect of spatial autocorrelation, we accept a higher error variance in favor of lowering bias and therefore set *k* to ten, as was also done by Park and Bae (2015), Chin et al. (2020), Hu et al. (2019) as well as Rico-Juan and Taltavull de La Paz (2021).

Further following the logic that information flow between training and test observations leads to biased cross-validation errors, we apply a nested resampling strategy that strictly separates data used for model selection from data used for model assessment. This is important since assessing model performance on the same data used for model selection does not yield an unbiased estimate of prediction error but more of a re-substitution error (Varma and Simon, 2006). Thus, nested resampling consists of two resampling loops, that is the inner resampling loop for hyperparameter tuning (i.e., model selection), which is wrapped within the outer resampling loop for performance evaluation (i.e., model assessment) such that model selection and model assessment is repeatedly performed on mutually exclusive subsamples, thereby simulating independent data throughout the entire workflow of the algorithm (Simon, 2007).

Following this strategy, the resulting cross-validation errors should, at least in theory, provide an unbiased picture of out-of-sample predictive performance to be expected from the models if the assumption of spatial randomness was fulfilled, thus enabling us to disentangle the effects of spatial dependence by using spatial CV.

To implement spatial partitioning in the cross-validation procedure, we apply a *k*-means clustering algorithm as proposed by Brenning (2012). The *k*-means clustering method is a universal and commonly used technique to detect a specified number of *k* clusters among *n* observations based on a given set of features. In a first step, the algorithm randomly chooses *k* centroids in the multi-dimensional feature space. The initial clustering is achieved by allocating each of *n* observation to the "nearest" centroid in the feature space (i.e., by minimizing the Euclidean distance from the feature values to the centroid). The positions of the cluster centroids are then adjusted by taking the mean feature values of each grouping and the clustering

is repeated. The clusters are iteratively adjusted until the allocation doesn't change anymore, so the within-cluster sum of squares is minimized (James et al., 2013).

The goal of spatial cross-validation is to maximize the distance between training and test folds. In this context, a cluster refers to a fold whereby k denotes the number of equally sized folds to be partitioned and the point coordinates (latitude and longitude) represent the features. The feature space is a two-dimensional scatterplot as depicted in Fig. 1. The algorithm arranges the folds in a way that minimizes the average distances within each fold and maximizes the average distance between the folds. This effectively decreases spatial autocorrelation between training and test data.

Using spatial and non-spatial partitioning, our nested resampling strategy provides four alternatives to calculate cross-validation errors which are:

(1) *non-spatial* model selection + *non-spatial* model assessment,
(2) *non-spatial* model selection + *spatial* model assessment,
(3) *spatial* model selection + *spatial* model assessment, and
(4) *spatial* model selection + *non-spatial* model assessment.

The first alternative is the conventional "off-the-shelf" approach typically applied in the hedonic literature, although nesting is not common yet. In contrast, the third option describes a pure spatial approach that should reduce spatial dependence between training and test observations to a minimum but may result in too pessimistic expectations of predictive performance. As the prediction goal in a housing context is not a pure spatial one, the second and fourth alternative could potentially provide a fair compromise in the trade-off between reduction of spatial autocorrelation and the extrapolation range introduced into the model.

To arrive at a final model that can predict the holdout data, all steps of the algorithm need to be executed once again, whereby the cross-validation in the outer loop is replaced by the holdout sample such that the full information from 2019 is used to train a model that predicts apartment rents from the first quarter of 2020 (Varma and Simon, 2006; Simon, 2007). Analogous to the nested resampling for the estimation of prediction error, optimal hyperparameters for the final prediction model are once again derived using spatial and non-spatial grid search CV resulting in two alternatives for the true error rate.

Based on the true error rates, we benchmark predictive performance and determine the bias in error estimates. Model accuracy and precision are assessed using the coefficient of determination (R^2), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean squared error (RMSE). To measure variation in the residuals, we calculate the interquartile range (IQR), the coefficient of dispersion (COD) and an error bucket that includes the proportion of predictions within 10 % of the true value (PE10). The mean percentage error (MPE) is used as a measure of biasedness. Subsequently, the asymptotic properties of all estimators are evaluated by comparing the distributions of error estimates resulting from the respective resampling strategies to the distributions of the true prediction errors.

Results

This section first presents the final model specifications and evaluates differences in hyperparameters selected by the automated grid search CV. Second, the results of the error-based model assessment are reported to determine the bias for the respective resampling strategies, and the asymptotic behavior of the estimators is discussed. Third, we investigate the residual spatial autocorrelation in each of the estimated models to draw conclusions on whether differences in the selected levels of regularization and the related model performance are linked to overfitting spatial structures in the data.

Model Selection

All variables listed in Table 1 were kept in the final model specification of the linear model. Following the principle to avoid overfitting, only the squared term for the building age and no interaction terms were additionally included. The resulting model specification serves as a baseline for all subsequent model alternatives and is hereinafter referred to as model specification "A". We estimate an alternative model specification "B", which does not consider locational and neighborhood characteristics in the regressor matrix, to see how the results change in the absence of spatial controls. All remaining modeling decisions for the linear models outlined below are based on specification A and were adopted for specification B. If not stated otherwise, the presented results refer to specification A. The respective regression outputs of the least squares estimator are shown in Appendix Table 6.

The Moran's I statistic of the OLS residuals rejects the null hypothesis of spatial randomness in price formation processes at a close-to-zero level of significance. The likelihood-ratio (LR) tests of a restriction of the SDM confirms the common factor hypothesis and implies the presence of both endogenous as well as exogenous interaction effects, leading to the acceptance of the SDM. The relevance of the SDEM was further investigated as an alternative spatial model. Again, the LR-tests reject a simplification of the SDEM with p values close to zero. We subsequently consider both the SDM as well as the SDEM as spatially conscious linear model alternatives. As spatial density of observations tapers toward the outskirts, we follow Pace et al. (2000) for the specification of W and choose a κ-nearest neighbors (κ-nn) matrix where each observation has a fixed number of κ neighbors. After evaluating different values for κ between 10 and 100, we eventually set the number of neighbors to 30, as this yields fair error estimates without diluting spatial effects in the lag terms. Overall, results remain robust for different choices of κ as well as for distance-based matrices with different boundaries.

The optimal hyperparameters selected by the grid search CV after 100 evaluations are shown in Table 3. For the random forest, there are no structural differences in the number of trees b nor the number of features m considered at each split, although spatial tuning seems to favor slightly higher values of m. Notable deviations can be observed in the minimum node size that determines the depth and, thus, the complexity of the individual trees in the forest. The non-spatial grid search

Table 3 Optimal Hyperparameters selected by 10-fold Cross-Validation

Fold	b (1)	b (2)	b (3)	b (4)	m (1)	m (2)	m (3)	m (4)	min_{node} (1)	min_{node} (2)	min_{node} (3)	min_{node} (4)
Panel A1: Random Forest including Spatial Controls												
1	650	500	400	400	9	9	12	14	2	2	8	7
2	300	350	550	250	9	9	10	10	1	1	5	8
3	300	500	200	500	9	9	10	10	1	1	6	6
4	300	600	500	450	7	9	9	10	2	1	2	1
5	500	450	300	650	7	7	10	10	1	2	6	5
6	350	500	450	600	9	9	9	12	2	2	7	6
7	600	300	500	500	9	7	12	9	1	1	1	7
8	650	650	650	550	7	9	9	10	1	1	3	7
9	550	650	600	400	9	9	10	9	1	3	4	6
10	550	500	650	500	9	10	9	10	2	1	2	3
Panel B1: Random Forest excluding Spatial Controls												
1	550	650	200	500	5	6	5	6	2	1	9	10
2	250	250	450	550	5	5	5	6	1	1	9	8
3	600	450	500	250	5	5	6	5	2	1	10	8
4	500	550	200	600	5	5	5	6	2	1	8	10
5	600	650	500	500	5	5	5	6	1	2	10	8
6	600	350	450	300	5	5	6	6	2	2	9	9
7	450	650	350	350	5	6	6	6	1	1	9	9
8	400	200	450	600	5	5	6	5	1	1	10	9
9	450	450	600	600	5	5	5	6	1	1	8	10
10	450	550	500	300	6	6	6	6	2	1	9	9

Table 3 (continued)

Fold	n_{rounds} (1)	n_{rounds} (2)	n_{rounds} (3)	n_{rounds} (4)	*m/p* (1)	*m/p* (2)	*m/p* (3)	*m/p* (4)	η (1)	η (2)	η (3)	η (4)
Panel A2: Extreme Gradient Boosting Trees including Spatial Controls												
1	600	550	250	400	72%	58%	65%	45%	0.06	0.07	0.04	0.05
2	650	600	550	250	38%	78%	65%	58%	0.06	0.10	0.02	0.06
3	650	550	200	300	58%	85%	52%	85%	0.10	0.07	0.08	0.05
4	550	600	500	350	78%	78%	58%	85%	0.08	0.08	0.02	0.02
5	600	600	200	550	72%	72%	78%	45%	0.07	0.08	0.04	0.02
6	600	600	450	450	65%	45%	45%	45%	0.07	0.09	0.06	0.03
7	650	550	550	300	72%	78%	45%	65%	0.08	0.09	0.09	0.04
8	500	650	450	350	52%	52%	78%	38%	0.08	0.09	0.02	0.07
9	650	650	500	650	58%	65%	78%	52%	0.05	0.06	0.02	0.02
10	500	550	400	450	78%	45%	72%	65%	0.07	0.08	0.03	0.03
Panel B2: Extreme Gradient Boosting Trees excluding Spatial Controls												
1	500	650	500	600	72%	72%	65%	78%	0.07	0.06	0.01	0.01
2	500	400	300	200	78%	78%	72%	85%	0.05	0.07	0.02	0.03
3	650	650	250	550	58%	65%	72%	85%	0.04	0.04	0.02	0.01
4	650	650	550	250	72%	65%	58%	85%	0.05	0.06	0.01	0.02
5	650	600	550	600	65%	58%	72%	72%	0.04	0.07	0.01	0.01
6	550	600	200	500	65%	65%	85%	65%	0.05	0.05	0.03	0.02
7	400	600	600	550	78%	58%	72%	65%	0.09	0.08	0.01	0.01
8	450	350	500	500	52%	58%	72%	78%	0.07	0.07	0.01	0.01
9	450	650	200	300	78%	65%	72%	85%	0.06	0.04	0.03	0.02
10	450	450	500	600	65%	45%	65%	65%	0.07	0.06	0.01	0.01

This table reports the optimal hyperparameters for each fold of the inner loop in the four alternatives of the nested resampling procedure: *(1) non-spatial tuning + non-spatial validation*, *(2) non-spatial tuning + spatial validation*, *(3) spatial tuning + spatial validation*, *(4) spatial tuning + non-spatial validation* selected by an automated grid search using 10-fold cross-validation. *b*: number of bootstrap trees, *m*: number of features in the column subsample, min_{node}: minimum node size, n_{rounds}: number of boosting rounds, *m/p*: portion of all available features *p* in the column subsample, η: learning rate (eta)

CV consistently prefers a min_{node} between one and two, which is significantly lower compared to the spatial model that has on average a minimum node size of five. A min_{node} of one provides the trees with the flexibility to have virtually infinite vertical growth, allowing them to remove all noise from the data (Kok et al., 2017). Or as expressed by Mullainathan and Spiess (2017), a tree which grows one leaf for each observation in the data "[…] will have perfect fit, but of course this is really perfect overfit", consequently yielding unsatisfactory predictions for unseen data.

A similar pattern can also be observed for the XGB. Again, there are no remarkable differences in the size of the column subsample $\frac{m}{p}$. However, the spatial instantiation of the resample call in the inner loop requires on average only 405 boosting rounds versus 593 boosting rounds for the non-spatial CV. Although the rate η at which the boosting algorithm learns at each round is more conservative in the non-spatial model, a higher number of n_{rounds} indicates excessive error corrections that may result in a model that overfits the residuals. The selection of less complex models compared to the non-spatial tuning persists for model specification B, although the higher complexity of the non-spatial random forest is now even more distinct. These findings corroborate our hypotheses derived from studies in other fields such as Le Rest et al. (2014), Roberts et al. (2017) as well as Meyer et al. (2019), who state that non-spatial partitioning during resampling is associated with the choice of overly complex models if the data exhibits spatial dependence.

Model Assessment

The subsequent section discusses how the differences in model selection affect model accuracy and whether increased complexity is indeed linked to overfitting and vice versa. Therefore, we analyze the bias and the asymptotic properties of our estimated models by comparing the true one quarter ahead prediction error to the expected error rates resulting from the respective resampling strategies outlined in the "*Performance Evaluation*" section. Bias is measured as the difference between the true error rate and the expected error rate. The aggregated performance measures are presented in Table 4.

The re-substitution errors from the linear models are significantly lower than the true error rates on all accounts, which, however, is not surprising (Efron, 1983). Whereas this overoptimism is smallest for the SDM, the error estimates of the SDEM are even more biased than those of the OLS model. This is mainly attributable to the relatively weaker predictive performance of the SDEM, since re-substitution errors of both spatial models are only marginally different from each other. Having said that, it is worth mentioning that spatial autoregressive models are primarily designed for statistical inference rather than out-of-sample predictions.

For the non-parametric models, one can see an improvement in all performance measures compared to the linear models, which is not surprising and in line with the literature (see Antipov and Pokryshevskaya, 2012; Yoo et al., 2012; Gu and Xu, 2017; Kok et al., 2017; Mullainathan and Spiess, 2017; Čeh et al., 2018; Bogin and Shui, 2020; Pace and Hayunga, 2020). With respect to predictive power, the XGB yields the most accurate results, closely followed by the RFR. Interestingly,

Table 4 Error-based Performance Matrix

Method	Resampling Strategy	R^2	MAE	MAPE	MPE	RMSE	PE10	IQR	COD
Panel A: Models including Spatial Controls									
OLS	*holdout*	78.81%	168.24	15.03%	−0.61%	292.14	44.45%	205.39	3.52%
	re-substitution	85.13%	149.79	13.91%	1.64%	251.01	47.21%	199.63	13.73%
SDM	*holdout*	81.62%	153.56	13.77%	−0.91%	272.09	50.54%	182.57	3.25%
	re-substitution	87.50%	134.81	12.64%	1.40%	230.13	51.81%	177.15	12.52%
SDEM	*holdout*	79.08%	166.51	14.70%	−1.57%	290.29	45.65%	198.80	3.45%
	re-substitution	87.46%	134.74	12.63%	1.40%	230.53	51.86%	175.95	12.51%
RFR	*holdout (non-spatial tuning)*	85.13%	143.48	12.79%	−0.33%	244.69	52.31%	169.58	3.14%
	holdout (spatial tuning)	85.14%	143.58	12.81%	−0.30%	244.64	52.48%	171.37	3.15%
	(1) non-spatial/non-spatial	89.42%	116.59	10.93%	1.45%	211.74	59.15%	141.57	11.12%
	(2) non-spatial/spatial	82.62%	157.30	14.35%	0.67%	271.40	45.83%	208.53	14.69%
	(3) spatial/spatial	83.07%	157.41	14.39%	0.58%	267.90	45.27%	208.00	14.69%
	(4) spatial/nonspatial	89.49%	117.93	11.08%	1.42%	211.07	58.65%	144.64	11.26%
XGB	*holdout (non-spatial tuning)*	85.20%	142.45	12.89%	0.71%	244.15	53.04%	165.43	3.15%
	holdout (spatial tuning)	85.21%	142.96	12.83%	0.16%	244.06	52.87%	171.23	3.11%
	(1) non-spatial/non-spatial	90.93%	112.66	10.53%	1.13%	196.08	60.71%	140.46	10.67%
	(2) non-spatial/spatial	83.90%	157.16	14.25%	−0.54%	261.20	44.86%	214.47	14.52%
	(3) spatial/spatial	84.54%	152.67	13.95%	0.00%	255.98	46.21%	202.73	14.19%
	(4) spatial/nonspatial	90.15%	117.11	10.97%	1.12%	204.33	58.27%	149.12	11.14%

Table 4 (continued)

Panel B: Models excluding Spatial Controls									
OLS	*holdout*	74.52%	188.95	16.80%	−0.67%	320.33	38.13%	250.15	3.96%
	re-substitution	80.58%	173.36	15.99%	2.08%	286.89	40.23%	235.13	15.66%
SDM	*holdout*	81.56%	154.06	13.82%	−0.91%	272.49	49.91%	184.02	3.25%
	re-substitution	87.38%	135.79	12.72%	1.41%	231.27	51.77%	175.98	12.59%
SDEM	*holdout*	75.90%	177.71	15.41%	−2.07%	311.53	42.17%	218.24	3.67%
	re-substitution	87.08%	136.36	12.74%	1.42%	233.96	51.74%	175.61	12.60%
RFR	*holdout (non-spatial tuning)*	80.66%	171.26	15.14%	0.17%	279.07	44.05%	219.94	3.76%
	holdout (spatial tuning)	80.59%	172.98	15.30%	0.22%	279.60	43.54%	217.75	3.83%
	(1) non-spatial/non-spatial	85.80%	143.98	13.40%	1.67%	245.33	49.17%	185.16	13.68%
	(2) non-spatial/spatial	80.51%	173.25	15.85%	2.00%	287.38	40.70%	228.44	16.05%
	(3) spatial/spatial	80.29%	172.41	15.73%	1.96%	289.02	40.75%	226.66	15.94%
	(4) spatial/nonspatial	85.60%	146.48	13.62%	1.69%	247.07	48.66%	191.14	13.96%
XGB	*holdout (non-spatial tuning)*	80.33%	176.27	15.85%	1.65%	281.48	42.69%	223.19	3.83%
	holdout (spatial tuning)	79.47%	173.97	15.28%	−1.17%	287.56	43.26%	225.95	3.79%
	(1) non-spatial/non-spatial	85.61%	148.24	13.77%	1.63%	246.96	46.97%	194.32	13.96%
	(2) non-spatial/spatial	81.26%	172.33	15.78%	1.97%	281.83	40.39%	226.85	15.93%
	(3) spatial/spatial	79.75%	173.80	15.31%	−1.18%	292.96	40.78%	229.86	15.77%
	(4) spatial/nonspatial	84.24%	154.54	14.03%	−0.52%	258.41	45.02%	208.12	14.49%

This table reports the performance measures for the *re-substitution* errors in the linear models and the cross-validation errors in the non-parametric models *(1) non-spatial tuning + non-spatial validation*, *(2) non-spatial tuning + spatial validation*, *(3) spatial tuning + spatial validation*, *(4) spatial tuning + non-spatial validation* in comparison to the true predictive performance from the errors of the *holdout* sample. R^2: coefficient of determination, MAE: mean absolute error, MAPE: mean absolute percentage error, MPE: mean percentage error, RMSE: root mean squared error, PE10: error bucket of estimates within 10% of the true value, IQR: interquartile range, COD: coefficient of dispersion. Absolute values are reported in Euro per month

performance measures do not seem noticeably affected by the resampling strategy in the inner loop for hyperparameter tuning despite the higher levels of regularization in the spatial models. Hence, predictive power is almost identical no matter whether spatial dependence has been accounted for during tuning or not.

Distinctive differences between spatial and non-spatial cross-validation can be observed for the outer resampling loop though. Compared to the true predictive performance, non-spatial CV errors are overly optimistic for both the RFR as well as the XGB. In contrast, spatial CV consistently yields overly pessimistic but more reliable approximations of prediction errors compared to non-spatial CV errors. It is, moreover, noteworthy that non-spatial hyperparameter tuning combined with spatial performance evaluation results in the most pessimistic cross-validation errors for all measures except the MAE and the MAPE of the random forest. The understatement of predictive accuracy is not surprising in this case since the model was trained with the objective to interpolate and subsequently validated by extrapolating to a new spatial domain, thus yielding non-optimal results.

The bias in non-spatial cross-validation errors is even more distinctive in panel B of Table 4. In contrast, the spatially conscious cross-validation errors now closely resemble the true prediction errors, thereby reducing bias to a minimum. As already anticipated, the results indicate that over-optimism in non-spatial CV is likely to originate from spatial structures being overfitted to covarying but non-causal regressors during model training. This is particularly noticeable when locational and neighborhood controls are missing such that the spatial information content in the residuals is picked up by other attributes that are structured in space. Accounting for spatial dependence in the inner resampling loop had once again only a minor impact on both model accuracy and bias. Noteworthy, the non-parametric models are now outperformed by the SDM in terms of predictive accuracy, which drops only marginally compared to specification A. This demonstrates the high robustness of the SDM, which is able to capture spatial effects through the spatial lag terms even in a scenario where locational control variables are not available. As stated by Doszyń (2020), machine learning methods require a very good data basis to take advantage of their flexibility, whereas less complex models are more robust in situations where extensive data is not available. The superiority of the SDM in the specification without spatial control variables moreover corroborates the findings of Pace and Hayunga (2020) who demonstrate that most of the improvement in accuracy achieved by machine learning models over parametric spatial models results from exploiting spatial structures in the data by creating spatially disaggregated models.

Figure 4 presents the asymptotic distribution of the absolute percentage error. The density plots reveal a higher variance and a lower kurtosis for prediction errors estimated by spatial CV opposed to the true error distribution. In comparison, non-spatial CV underestimates prediction errors, particularly in the lower tails of the distribution where errors are close to zero. For both the RFR and the XGB, non-spatial error estimates are centered around values that are considerably lower than the true means whereas spatial error estimates are more dispersed. This is affirmed by both the higher deviation of spatial error estimates from the median true error represented by the coefficient of dispersion as well as their larger spread illustrated by the interquartile range. These findings again confirm our expectations derived from

literature in other spatial modeling fields (see Le Rest et al., 2014; Roberts et al., 2017; Schratz et al., 2019).

Residual Spatial Autocorrelation

Finally, we analyze the spatial autocorrelation found in the residuals of the models after calculating the Moran's *I* statistic to investigate whether overfitting is related to the exploitation of spatial dependence structures in the data. Since the relative magnitude of the Moran's *I* is only meaningful for identical spatial weight matrices, we also calculate the Z-scores as a standardized measure, which allows us to compare the spatial autocorrelation of in-sample and out-of-sample residuals (Anselin, 1995). The results are presented in Table 5.

The spatial linear models successfully reduce spatial autocorrelation in the in-sample residuals, although there is still spatial information content left, which is consistent with the findings of Pace and Hayunga (2020). The non-spatial cross-validation errors of the random forest exhibit spatial autocorrelation of roughly the same magnitude as the spatial linear models. Interestingly, spatially cross-validated errors show a significantly higher degree of spatial autocorrelation that even exceeds the Z-score of the simple OLS model. The same applies to the boosted trees, although this method seems to understand spatial structures in the data slightly better. This may seem counterintuitive at first but knowing that non-spatial error estimates are biased downwards, the substantial differences in residual spatial autocorrelation between the spatial and the non-spatial cross-validation errors indicate that spatial partitioning in the outer resampling loop indeed prevents the models from exploiting unexplained spatially autocorrelated information from the test data during training. By and large, the outcomes are consistent for panel B but, unsurprisingly, the magnitude of spatial autocorrelation is in general higher as opposed to specification A, which further substantiates our hypotheses and underlines the importance of spatial cross-validation. Consistent with the results from *section 4.2*, the SDM does have a superior understanding of spatial dependence structures in the data compared to all other models when spatial variables are not considered.

Conclusion

Recent literature has brought forth an increasing body of evidence that demonstrates a superior predictive performance of machine learning algorithms compared to parametric models for complex spatial regression problems involving the estimation of house prices and rents. In non-parametric models, predictive performance is widely measured using resampling techniques such as cross-validation, which can be thought of as an out-of-sample experiment inside the original sample. This requires the statistical independence of the data to yield unbiased and meaningful prediction error estimates that can be used for model selection and model assessment. The inherent spatial dependence in house price and rent formation processes gives reason to question the validity of cross-validation errors in a hedonic context. Hence,

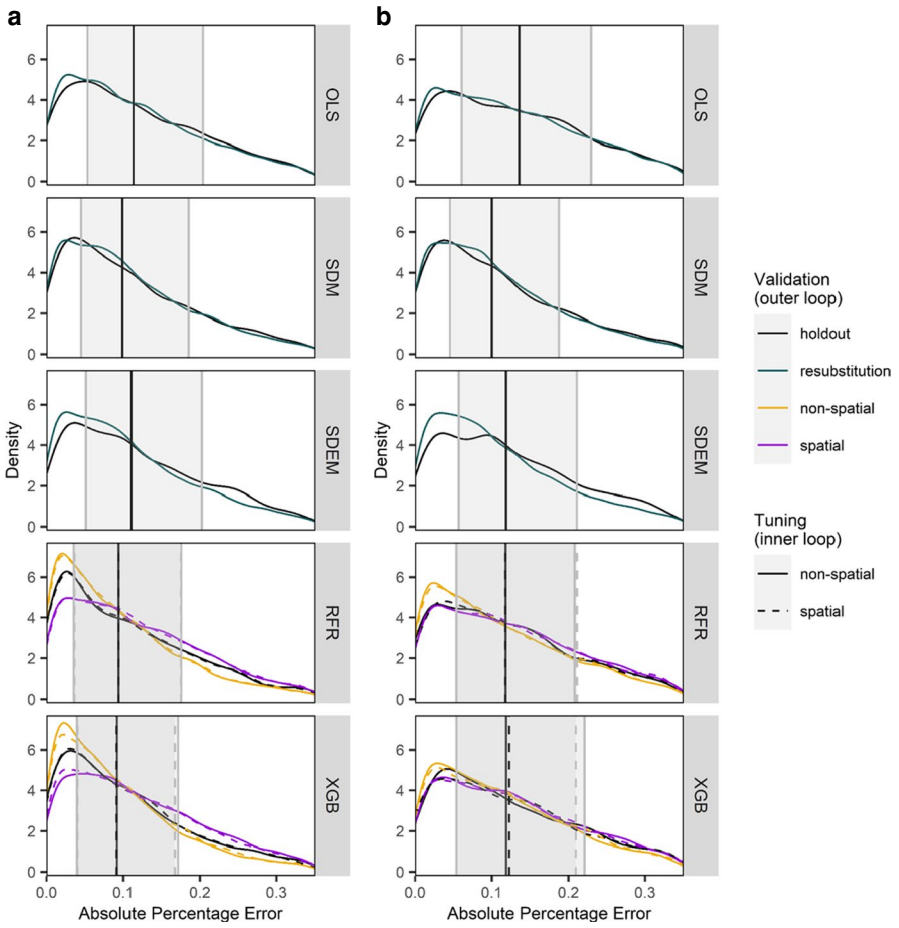

Fig. 4 **a** Distribution of the Absolute Percentage Error for Models including Spatial Controls. **b** Distribution of the Absolute Percentage Error for Models excluding Spatial Controls. Notes: The density plots present the expected distribution of the absolute percentage error resulting from the respective resampling strategies in comparison to the true out-of-sample distribution of the absolute percentage error from the holdout sample. The line type represents the resampling strategy used in the inner loop for model selection (non-parametric models only) and the line color represents the resampling strategy applied in the outer loop for model assessment. The true out-of-sample distribution is represented in black. The shaded areas depict the interquartile range, which is the area between the first quartile and the third quartile of the true absolute percentage error with the middle line representing the median

this study investigates the adequacy of conventional *k*-fold cross-validation for the purpose of model selection and model assessment in an algorithmic hedonic context using tree-based boosting and bagging methods and proposes a spatially conscious alternative that attempts to reduce bias in cross-validation errors by accounting for the spatial proximity of observations.

Despite using a nested resampling strategy and applying column subsampling in our bagging and boosting algorithms to prevent overfitting, our results demonstrate

that failing to account for spatial dependence during the cross-validation procedure still has two undesirable consequences. First, hyperparameter tuning using non-spatial grid search CV favors the selection of overly complex models that overfit spatial dependence structures in the training data, thereby compromising the models' generalizability. Second, performance estimates are artificially inflated through the exploitation of spatial dependence structures during model training, resulting in overly optimistic error estimates when compared to the true prediction errors.

In nested resampling approaches these two problems go hand in hand since the selection of overly complex models in the inner resampling loop is masked by over-optimistic accuracy measures during model assessment in the outer resampling loop. This can lead to spurious confidence in a model that overestimates predictive accuracy as nesting aims to simulate unseen data throughout the entire workflow of an algorithm, therefore suggesting unbiased error estimates (Varma and Simon, 2006). In contrast, spatial grid search CV prefers a higher level of regularization, thereby introducing extrapolation into the models, which results in error estimates that are slightly too pessimistic, yet closer to the true error rates.

An analysis of the residual spatial autocorrelation provides evidence that the spatially conscious cross-validation technique hinders the algorithm from exploiting spatial dependence structures, thereby preventing overfitting. To see how the results vary with the extent to which spatial information is reflected in the feature space, we evaluate a second model alternative that does not consider spatial control variables. In this scenario, over-optimism in predictive accuracy is even more distinctive when spatial autocorrelation is not accounted for, whereas the spatial CV procedure yields almost unbiased estimates of the true prediction error that converge asymptotically closer to the true error distribution.

Despite their flexibility and higher accuracy compared to traditional parametric methods, machine learning techniques are often criticized for their black box character that impedes direct model interpretation as well as for their high computational burden. To empirically illustrate where the costs and benefits of these methods lie, a least squares model as well as a linear spatial autoregressive framework are furthermore used as points of reference to assess predictive accuracy. Whereas the boosting algorithm performs best when spatial controls are reflected in the model, the spatial durbin model outperforms the non-parametric model alternatives in the absence of spatial information in the regressor matrix, which stresses the importance of considering parametric model alternatives besides non-parametric models.

We conclude that in a real estate hedonic context, state-of-the-art cross-validation does not yield unbiased estimates of prediction error even when applying methods that intend to counteract overfitting. Resulting cross-validation errors should rather be interpreted as an estimate of the lower bound of the true error rate. In contrast, spatial cross-validation errors tend to be slightly too pessimistic but more reliable estimates of prediction errors. Likewise, the more conservative spatial cross-validation errors can be regarded as an upper bound of prediction errors.

That being said, in scenarios where the study area is very small and clearly delineated so that spatial dependence structures do not vary significantly (i.e., on the submarket or ZIP-code level), spatial density of observations is high (i.e., CBD or city center), and spatial control variables are numerous, random partitioning of folds

Table 5 Residual Spatial Autocorrelation

		Panel A: Models including Spatial Controls			*Panel B: Models excluding Spatial Controls*		
Method	Resampling Strategy	Morans' *I*	Z-score	p value	Morans' *I*	Z-score	p value
OLS	*holdout*	0.17	55.44	0.00	0.29	95.36	0.00
	re-substitution	0.27	19.30	0.00	0.39	27.48	0.00
SDM	*holdout*	0.03	10.17	0.00	0.03	10.23	0.00
	re-substitution	0.16	11.60	0.00	0.17	11.76	0.00
SDEM	*holdout*	0.03	10.75	0.00	0.04	13.58	0.00
	re-substitution	0.25	17.89	0.00	0.34	23.59	0.00
RFR	*holdout (non-spatial tuning)*	0.17	11.75	0.00	0.33	23.45	0.00
	holdout (spatial tuning)	0.17	11.70	0.00	0.32	22.60	0.00
	(1) non-spatial/non-spatial	0.03	9.79	0.00	0.18	61.22	0.00
	(2) non-spatial/spatial	0.19	64.70	0.00	0.26	87.78	0.00
	(3) spatial/spatial	0.19	62.80	0.00	0.26	87.91	0.00
	(4) spatial/non-spatial	0.03	9.97	0.00	0.19	63.13	0.00
XGB	*holdout (non-spatial tuning)*	0.14	9.52	0.00	0.26	18.53	0.00
	holdout (spatial tuning)	0.14	10.10	0.00	0.33	23.26	0.00
	(1) non-spatial/non-spatial	−0.01	−1.90	0.06	0.15	51.03	0.00
	(2) non-spatial/spatial	0.16	54.67	0.00	0.23	75.07	0.00
	(3) spatial/spatial	0.16	53.59	0.00	0.28	92.21	0.00
	(4) spatial/non-spatial	0.01	2.55	0.01	0.21	71.22	0.00

This table reports the spatial autocorrelation found in the residuals of the models. A positive and significant Morans' *I* signals spatial clustering of similar values whereas a negative and significant Morans' *I* signals alternating values which indicates the presence of spatial outliers and/or spatial heterogeneity. The Z-score serves as a standardized value for comparison of the in-sample and out-of-sample statistics. It is calculated as the difference between the observed value of *I* and the expected value of *I* divided by the standard deviation of *I*, whereby the expected value of *I* is the theoretical mean defined as −1/(N-1), N being the number of observations

may yield fair estimates of predictive performance. However, for typical use cases (i.e., predictions on the city-level or above) where spatial dependence structures and spatial density vary continuously across space, spatial cross-validation should be preferred for model selection and model assessment, since we believe that, in general, the cost of a slightly too pessimistic perception of predictive accuracy is lower than having spurious confidence in a model's capability to predict unseen data. Overstatement of predictive accuracy may withhold appraisers, underwriters, lenders, as well as portfolio and investment managers from appropriately reflecting the uncertainties associated with appraised values in their decision making and risk management, potentially leading to adverse effects in capital allocation.

Future research in this field may apply model-agnostic interpretation techniques and analyze to what extent identified relationships are spurious when spatial dependence is not accounted for to shed light on the role of spatial autocorrelation on the decision-making of the algorithms.

Appendix 1

Table 6 OLS Regression Output

	Panel A: OLS Model including Spatial Controls				
Variable	Estimate	Std. Error	t-value	p value	Significance
(Intercept)	−74.91	5.92	−12.65	0.00	***
Continuous					
Living Area [log]	0.80	0.01	81.98	0.00	***
Age [years]	0.00	0.00	−20.11	0.00	***
Age squared	0.00	0.00	19.59	0.00	***
Entry date [years]	−0.02	0.01	−2.12	0.03	*
Latitude[1]	1.45	0.12	12.17	0.00	***
Longitude[1]	0.67	0.06	11.68	0.00	***
Discrete					
Rooms	0.04	0.00	10.05	0.00	***
Floor	0.00	0.00	5.29	0.00	***
Binary [1 = yes, 0 = no]					
Bathtub	−0.04	0.00	−9.78	0.00	***
Refurbished	0.02	0.01	3.03	0.00	**
Built-in kitchen	0.11	0.01	21.65	0.00	***
Balcony	0.02	0.00	3.84	0.00	***
Parking	0.03	0.01	4.88	0.00	***
Elevator	0.04	0.01	7.64	0.00	***
Terrace	0.04	0.01	5.99	0.00	***
Distances					
NUTS centroid [km]	−0.02	0.00	−13.02	0.00	***
Bakery [km]	−0.01	0.01	−1.22	0.22	
Bar [km]	−0.02	0.00	−3.73	0.00	***
Biergarten [km]	−0.05	0.00	−14.71	0.00	***
Café [km]	−0.02	0.01	−2.11	0.03	*
School [km]	−0.03	0.01	−2.35	0.02	*
Supermarket [km]	0.02	0.01	1.20	0.23	
Bus station [km]	−0.04	0.00	−17.16	0.00	***
Panel B: OLS model excluding Spatial Controls					
(Intercept)	3.12	0.04	83.32	0.00	***
Continuous					
Living Area [log]	0.84	0.01	76.62	0.00	***
Age [years]	0.00	0.00	−18.16	0.00	***
Age squared	0.00	0.00	21.45	0.00	***
Entry date [years]	−0.02	0.01	−2.25	0.02	*

Table 6 (continued)

	Panel A: OLS Model including Spatial Controls				
Variable	Estimate	Std. Error	t-value	p value	Significance
Discrete					
Rooms	0.03	0.00	6.32	0.00	***
Floor	0.01	0.00	6.43	0.00	***
Binary [1 = yes, 0 = no]					
Bathtub	−0.05	0.00	−10.78	0.00	***
Refurbished	0.03	0.01	4.53	0.00	***
Built-in kitchen	0.15	0.01	26.57	0.00	***
Balcony	0.02	0.01	2.94	0.00	**
Parking	0.01	0.01	1.52	0.13	
Elevator	0.10	0.01	16.78	0.00	***
Terrace	0.03	0.01	4.06	0.00	***

This table reports the ordinary least squares (OLS) regression outputs for the model including spatial controls in panel A and the model excluding spatial controls in panel B. The dependent variable is the *log(rent),* independent variables are listed in the left column accordingly. Significance codes: $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', $p < 0.1$ '.', $p > 0.1$ ''. Std. Error: standard error. Panel A: F-statistic 2245.00 (p value: 0.0000), AIC -4450.02, BIC -4276.96. Panel B: F-statistic 2968.00 (p value: 0.0000), AIC -2592.44, BIC -2488.61

[1]The coefficients of the coordinates indicate that the rent in the study area increases on average ceteris paribus by 67% for one degree of longitude to the east and analogous by 145% for on the degree of latitude to the north. We acknowledge that, in a linear context, interpretation of coordinates is very abstract and has limitations such as anisotropy across space. A more practical interpretation is that, in general, rents tend to increase towards the east/north of the city

1. As modeling choices may differ depending on whether analysis or prediction is the main objective of a study, we concentrate primarily on prediction and do not wish to draw any causal inference or conclusions of the market under investigation.
2. Although the use of asking rents can been criticized since they may deviate from actual contract rents, multiple listing systems (MLS) provide a valuable data source for statistical learning applications due to their high frequency of occurrence and timely availability and have been repeatedly used in the algorithmic hedonic literature (e.g., Chiarazzo et al., 2014; Park and Bae, 2015; Baldominos et al., 2018; Gröbel and Thomschke, 2018; Hu et al., 2019; Pérez-Rave et al., 2019; Pace and Hayunga, 2020; Rico-Juan and Taltavull de La Paz, 2021). Considering vacancy rates for dwellings in Frankfurt well below 1 %, we can follow the rationale of Gröbel (2019) and assume that renters are price takers, such that there should be no notable differences between asking and contract rents. Besides that, deviations between asking and contract rents "[…] are not expected to lead to an error bias", especially when hedonic characteristics are controlled for, as stated by Cajias (2018). We hence do not see any reason to question the validity of asking rents, especially in view of the objective of our study.
3. The systematic variation of rent formation processes across space should not introduce bias into cross-validation errors since machine learning algorithms do not assume fixed hedonic pricing coefficients but have the flexibility to differentiate between spatially heterogeneous environments. In this study, spatial heterogeneity is, therefore, put aside and the focus lies on spatial autocorrelation only.
4. Evidently, housing data not only exhibit high levels of spatial but also temporal dependence when pooled across time (Pace et al., 2000). In this study, we leave temporal aspects aside for future research and focus solely on space.

5. One limitation of the selected *k*-means clustering cross-validation strategy (Brenning, 2012) is that full independence between training and test data can only be achieved if the distance between each pair of training and test observation exceeds the spatial autocorrelation range (Brenning, 2005; Le Rest et al., 2014). This is unlikely to be the case for all observations in our resampling instantiation, especially for data points located at the borders of the spatial clusters. Nonetheless, we believe that the number of observations in each test fold is large enough to counteract structural overfitting, such that the impact on aggregated results should be minor. Following the suggestion of Roberts et al. (2017), we refrain from further reducing the number of *k* folds in the cross-validation procedure since this would withhold too much information during training and may introduce unnecessary extrapolation into the models.
6. All analyses and model estimations were executed using the open-source statistical programming language R under the version 4.0.4 (R Core Team, 2021). All machine learning algorithms and resampling techniques were employed using the *mlr3* framework implemented by Lang et al. (2019), which is an ecosystem that facilitates a standardized interface to many existing packages in the R environment. To obtain reproducible results and to ensure that the instantiation of the resampling calls do not vary between the different models, which could distort the results, all outputs were produced with the same random number generator using the *set.seed* function in R.
7. Estimations were executed on a standard 1.80GHz processor with four cores, eight logical processors and eight gigabytes of RAM using a 64-bit Windows operating system. After parallelization, the in-sample estimation of the random forest required between 11 and 20 hours for each of the four estimated resampling alternatives, whereby spatial tuning in the inner resampling loop reduced estimation time by up to 43%. The much more efficient extreme gradient boosting algorithm needed only about 3 to 3.5 hours respectively with the spatial tuner being slightly less time-consuming. During the one quarter ahead prediction, where cross-validation only needs to be performed for hyperparameter tuning in the inner loop, estimation time dropped to 3.5 hours for the bagging algorithm and to approximately 30 minutes for the boosting algorithm. The spatial linear models required less than 30 minutes each and the least squares estimator less than a second. Estimation time was significantly lower for the alternative model specification B for all models.

Acknowledgments The authors especially thank PATRIZIA AG for contributing to this study. All statements of opinions are those of the authors and do not necessarily reflect the opinion of PATRIZIA AG or its associated companies.

Authors' Contributions All authors contributed to the study conception and design, analysis, and material preparation. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data were provided by PATRIZIA AG and are confidential.

Code Availability Custom code using R open-source software and packages (R Core Team, 2021).

Declarations

Conflicts of Interest/Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

Allen, M. T., Springer, T. M., & Waller, N. G. (1995). Implicit pricing across residential rental submarkets. *The Journal of Real Estate Finance and Economics, 11*, 137–151. https://doi.org/10.1007/BF01098658

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-015-7799-1

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis, 27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics, 26*(1), 77–104. https://doi.org/10.1016/0166-0462(95)02111-6

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Bahn, V., & McGill, J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography, 16*(6), 733–742. https://doi.org/10.1111/j.1466-8238.2007.00331.x

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences, 8*(11), 2321. https://doi.org/10.3390/app8112321

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics, 17*(1), 61–85. https://doi.org/10.1023/A:1007703229507

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bogin, A. N., & Shui, J. (2020). Appraisal accuracy and automated valuation models in rural areas. *The Journal of Real Estate Finance and Economics, 60*, 40–52. https://doi.org/10.1007/s11146-019-09712-0

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics, 12*(1), 12–28. https://doi.org/10.1016/S1051-1377(03)00003-2

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics, 35*(2), 143–160. https://doi.org/10.1007/s11146-007-9036-8

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. *The Journal of Real Estate Research, 32*(2), 139–160. https://doi.org/10.1080/10835547.2010.12091276

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review, 60*(3), 291–319. https://doi.org/10.2307/1403680

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees* (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences, 5*, 853–862. https://doi.org/10.5194/nhess-5-853-2005

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *IEEE International Geoscience and Remote Sensing Symposium, 2012*, 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393

Cajias, M. (2018). Is there room for another hedonic model? The advantages of the GAMLSS approach in real estate research. *Journal of European Real Estate Research, 11*(2), 224–245. https://doi.org/10.1108/JERER-07-2017-0025

Cajias, M., & Ertl, S. (2018). Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions? *Journal of Property Investment & Finance, 36*(1), 32–49. https://doi.org/10.1108/JPIF-10-2016-0080

Cajias, M., Willwersch, J., Lorenz, F., & Schaefers, W. (2021). Rental pricing of residential market and portfolio data – A hedonic machine learning approach. *Real Estate Finance, 38*(1), 1–17.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics, 22*(3), 453–474. https://doi.org/10.1016/0166-0462(92)90039-4

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics, 14*, 203–222. https://doi.org/10.1023/A:1007744706720

Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics, 29*(2), 167–191. https://doi.org/10.1023/B:REAL.0000035309.60607.53

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), 168–183. https://doi.org/10.3390/ijgi7050168

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia, 3*, 810–817. https://doi.org/10.1016/j.trpro.2014.10.067

Chin, S., Kahn, M. E., & Moon, H. R. (2020). Estimating the gains from new rail transit investment: A machine learning tree approach. *Real Estate Economics, 48*(3), 886–914. https://doi.org/10.1111/1540-6229.12249

Cliff, A., & Ord, K. (1973). *Spatial autocorrelation*. Pion.

Cressie, N. A. C. (1993). *Statistics for spatial data* (Revised ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119115151

Din, A., Hoesli, M., & Bender, A. (2001). Environmental variables and real estate prices. *Urban Studies, 38*(11), 1989–2000. https://doi.org/10.1080/00420980120080899

Doszyń, M. (2020). Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation. *Journal of European Real Estate Research, 13*(2), 161–179. https://doi.org/10.1108/JERER-11-2019-0040

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78*(382), 316–331. https://doi.org/10.2307/2288636

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Füss, R., & Koller, J. A. (2016). The role of spatial and temporal structure for residential rent predictions. *International Journal of Forecasting, 32*(4), 1352–1368. https://doi.org/10.1016/j.ijforecast.2016.06.001

Gröbel, S. (2019). Analysis of spatial variance clustering in the hedonic modeling of housing prices. *Journal of Property Research, 36*(1), 1–26. https://doi.org/10.1080/09599916.2018.1562490

Gröbel, S., & Thomschke, L. (2018). Hedonic pricing and the spatial structure of housing data – An application to Berlin. *Journal of Property Research, 35*(3), 185–208. https://doi.org/10.1080/09599916.2018.1510428

Gu, G., & Xu, B. (2017). Housing market hedonic price study based on boosting regression tree. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 21*(6), 1040–1047. https://doi.org/10.20965/jaciii.2017.p1040

Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic regression. In *And survival analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hong, J., Choi, H., & Kim, W.-sung. (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. https://doi.org/10.3846/ijspm.2020.11544

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy, 82*, 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. *Springer.* https://doi.org/10.1007/978-1-4614-7138-7

Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with spatial disturbances. *The Journal of Real Estate Finance and Economics, 17*(1), 99–121. https://doi.org/10.1023/A:1007707430416

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International joint Conference on Artificial Intelligence, 2*, 1137–1143.

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management, 43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Lachenbruch, P., & Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*(1), 1–11. https://doi.org/10.2307/1266219

Lam, K. C., Yu, C. Y., & Lam, C. K. (2009). Support vector machine and entropy based decision support system for property valuation. *Journal of Property Research, 26*(3), 213–233. https://doi.org/10.1080/09599911003669674

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software, 4*(44), 1903. https://doi.org/10.21105/joss.01903

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography, 23*, 811–820. https://doi.org/10.1111/geb.12161

LeSage, J. P. (2014). What regional scientists need to know about spatial econometrics. *The Review of Regional Studies, 44*(1), 13–32. https://doi.org/10.52324/001c.8081

LeSage, J. P., & and Pace, R. K. (2009). Introduction to spatial econometrics. CRC Press. https://doi.org/10.1201/9781420064254.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics. Forthcoming*. https://doi.org/10.1111/1540-6229.12397

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. CRC Press. https://doi.org/10.1201/9780203730058

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies, 60*(3), 531–542. https://doi.org/10.2307/2298123

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *The Journal of European Real Estate Research, 12*(1), 134–150. https://doi.org/10.1108/JERER-08-2018-0035

McCluskey, W., McCord, M., Davis, P., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research, 30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Meyer, H., Reudenbach, C., Woellauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling, 411*. https://doi.org/10.1016/j.ecolmodel.2019.108815

Militino, A. F., Ugarte, M. D., & García-Reinaldos, L. (2004). Alternative models for describing spatial dependence among dwelling selling prices. *The Journal of Real Estate Finance and Economics, 29*(2), 193–209. https://doi.org/10.1023/B:REAL.0000035310.20223.e9

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modeling. *The Journal of Real Estate Research, 32*(3), 289–320. https://doi.org/10.1080/10835547.2010.12091282

Pace, R. K., & Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics, 14*(3), 333–340. https://doi.org/10.1023/A:1007762613901

Pace, R. K., & Hayunga, D. (2020). Examining the information content of residuals from hedonic and spatial models using trees and forests. *The Journal of Real Estate Finance and Economics, 60*, 170–180. https://doi.org/10.1007/s11146-019-09724-w

Pace, R. K., & LeSage, J. P. (2010). Omitted variable biases of OLS and spatial lag models. In A. Páez, J. Le Gallo, R. N. Buliung, & S. Dall'erba (Eds.), *Progress in spatial analysis: Methods and applications* ((1st ed., pp. 17–28). Springer. https://doi.org/10.1007/978-3-642-03326-1

Pace, R. K., Barry, R., Gilley, O. W., & Sirmans, C. F. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting, 16*(2), 229–246. https://doi.org/10.1016/S0169-2070(99)00047-3

Park, B., & and 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. https://doi.org/10.1016/j.eswa.2014.11.040.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research, 36*(1), 59–96. https://doi.org/10.1080/09599916.2019.1587489

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *The Journal of Real Estate Research, 31*(2), 147–164. https://doi.org/10.1080/10835547.2009.12091245

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*(387), 575–583. https://doi.org/10.2307/2288403

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science, 31*(10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*. https://doi.org/10.1016/j.eswa.2021.114590

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroeder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55. http://dx.doi.org/https://doi.org/10.1086/260169.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling, 406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Simon, R. (2007). Resampling strategies for model assessment and selection. In W. Dubitzky, M. Granzow, & D. P. Berrar (Eds.), *Fundamentals of data mining in genomics and proteomics, (1st ed., pp. 173–186)*. Springer. https://doi.org/10.1007/978-0-387-47509-7

Sirmans, G. S., & Benjamin, J. D. (1991). Determinants of market rent. *The Journal of Real Estate Research, 6*(3), 357–379. https://doi.org/10.1080/10835547.1991.12090653

Sirmans, G. S., Sirmans, C. F., & Benjamin, J. D. (1989). Determining apartment rent: The value of amenities, services and external factors. *The Journal of Real Estate Research, 4*(2), 33–43. https://doi.org/10.1080/10835547.1989.12090581

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics, 19*(4), 415–428. https://doi.org/10.1080/00401706.1977.10489581

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B: Methodological, 36*(2), 111–147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234–240. https://doi.org/10.2307/143141

Trachsel, M., & Telford, R. J. (2016). Technical note: Estimating unbiased transfer-function performances in spatially structured environments. *Climate of the Past, 12*, 1215–1223. https://doi.org/10.5194/cp-12-1215-2016

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution, 10*(2), 225–232. https://doi.org/10.1111/2041-210X.13107

Valente, J., Wu, S., Gelfand, A., & Sirmans, C. F. (2005). Apartment rent prediction using spatial modeling. *The Journal of Real Estate Research, 27*(1), 105–136. https://doi.org/10.1080/10835547.2005.12091148

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(91). https://doi.org/10.1186/1471-2105-7-91

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *The Journal of Real Estate Research, 10*(2), 185–201. https://doi.org/10.1080/10835547.1995.12090782

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning, 107*(3), 293–306. https://doi.org/10.1016/j.landurbplan.2012.06.009

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *The Journal of Real Estate Research, 33*(3), 349–387. https://doi.org/10.1080/10835547.2011.12091311