university of groningen

Appraiser-based automated valuation: A case study of valuing buy-to-let properties in the Netherlands

Master Thesis, MSc Real Estate Studies

Faculty of Spatial Sciences, University of Groningen

Date: 28-02-2022

Author

D.P.W. (Daan) van der Hoeven

S3419630

d.p.w.van.der.hoeven@student.rug.nl

Supervisor

M. (Mark) van Duijn

mark.van.duijn@rug.nl

Assessor

A. J. (Arno) van der Vlist

a.j.van.der.vlist@rug.nl

Supervisor Envalue

W. (Wessel) van Loon

wessel.vanloon@envalue.com

Abstract

Automated valuation methodologies are valuable tools in complementing traditional valuation methods because of their potential in terms of accuracy, efficiency, and reliability. Concurrently, however, these methodologies lack sufficient explainability to be applied in valuation practice. This paper, therefore, proposes a valuation methodology incorporating appraiser expertise to investigate the potential in terms of accuracy, explainability, and reliability. We investigate the implementation of such a methodology in automating buy-to-let property valuations in the Netherlands. To do so, we compare the proposed appraiser-based methodology to other methodologies described in the forecasting literature. We compare these methodologies by forecasting the constituents of market value in the single period capitalization method: vacant possession value, market rent, and gross income multiplier using a unique dataset comprising buy-to-let property transactions in the Netherlands. Using transaction data on all three market value constituents, we find the proposed appraiser-based methodology to have slightly lower accuracy while having similar reliability and possessing more explainability compared to existing methodologies.

Keywords:
Automated valuation
Machine learning
Monotonic constraints

Table of contents

Glossary

Abbreviation	Full form
AI	Artificial intelligence
ANN	Artificial neural network
API	Application programming interface
AVM	Automated valuation model
AVS	Automated valuation services
BAG	Basisregistratie adressen en gebouwen
CART	Classification and regression tree
CV	Cross-validation
DLM	Dynamic linear model
DT	Decision tree
GB	Gradient boosting
GIM	Gross income multiplier
HPM	Hedonic pricing model
HTM	Hierarchical trend model
IQRat	Inter-quartile range in ratios
LMDPE	Log median prediction error
LRMSE	Log root mean squared error
MAE	Mean absolute error
ML	Machine learning
mmMAPE	Max-min mean absolute prediction error
mmPER	Max-min percentage error range
MR	Market rent
MV	Market value
NRVT	Register of Real Estate Appraisers Netherlands
RF	Random forest
RMSE	Root mean squared error
RICS	Royal Institution of Chartered Surveyors
SMBO	Sequential model-based global optimization
SPCM	Single period capitalization method
TPE	Tree Parzen Estimator
VPV	Vacant possession value
XGB	Extreme gradient boosting
XGBMC	Extreme gradient boosting with monotonic constraints

1. Introduction

Property valuations are valuable for various purposes in the real estate industry. Government agencies, for instance, get periodic property valuations for tax purposes, while financial institutions require property valuations for estimates of collateral value (IAAO, 2018). Traditionally, professional appraisers conduct these valuations manually. Using either the *asset-based approach, cost approach, income approach,* or *market approach,* professional appraisers estimate *market value* (MV), defined as the price resulting from an arms-length transaction between an informed and willing buyer and seller (IVSC, 2019: 18; TEGOVA, 2020: 27). Manual valuations, however, are time-consuming and expensive (Schulz et al., 2014). Because of recent developments in property data and computing power, property valuations are therefore shifting toward more automated methods (Jordan & Mitchell, 2015). Described as automated valuation models (AVMs), these methodologies provide an estimate of value using large amounts of property data (Schulz et al., 2014). AVMs can outperform manual valuations in terms of accuracy (Horváth et al., 2016), while also being more efficient and reliable (Kok et al., 2017).

Both *econometric* and *machine learning* (ML) methodologies are capable of the automated valuation of real estate (Kroon & Francke, 2021). Traditionally, econometric models are adopted for this purpose, such as the hedonic pricing model (HPM; see Malpezzi (2003), for an extensive overview). Although explainable and reliable, econometric models require a priori model specifications before parameter estimation, limiting their flexibility as an AVM (Robinson & Sanderford, 2017). In the presence of forecasting complexity, ML methodologies have more effective ways of modeling relationships (Varian, 2014). Specifically, ML methodologies do not require a priori model specification, as these methodologies search for a functional form and parameter values simultaneously (Mullainathan & Spiess, 2017). In contrast to parameter estimation, ML methodologies revolve around finding an optimal prediction, making them more accurate in forecasting purposes compared to econometric models (Breiman, 2001b; Ceyhan, 2017). Nguyen & Cripps (2001), and Peterson & Flanagan (2009), for example, compare HPM with artificial neural networks (ANNs; a ML model) and conclude ANNs to be more accurate. Other authors (Antipov & Pokryshevskaya, 2012; Zurada et al., 2011) compare a wider range of econometric and ML models and argue ML models perform better when dealing with heterogeneous data. While these authors argue for the higher accuracy of ML models, most of these (non-parametric) methodologies lack sufficient explainability for valuation practice (McCluskey et al., 2013).

More recently, several authors investigate combining econometric and ML methodologies in a *hybrid methodology* (see, e.g., Ceyhan (2017), and Kroon & Francke (2017)). By replacing the rigid a priori specified function of an econometric model with ML methodologies, they argue that prediction accuracy can increase through higher flexibility while maintaining the structure of an econometric model. In line with these authors, Schimert & Wineland (2010) argue that determining the influence of input features on the predicted output through ML estimation while maintaining the econometric

structure able to capture temporal effects performs best in terms of accuracy and reliability compared to stand-alone econometric and ML models. Although these studies show hybrid methodologies outperform stand-alone models in terms of prediction accuracy, however, Ceyhan (2017) does argue that explainability declines compared to stand-alone econometric methodologies.

Explainability is a contemporary topic in the practical applicability of AVMs. Schulz & Wersing (2021), in their discussion on the applicability of AVMs as automated valuation services (AVS), argue that users of these methodologies require an understanding of the results. The uncertainty in reported information based on AVMs, they argue, should be reported in a manner that's comprehensible to its users. Similarly, in the context of valuation practice, AVMs require capabilities to communicate the origination of calculated MV constituents. As dictated in current Dutch regulatory practice, transparency of valuation methodologies to clients is essential (Fakton, 2021). To comply with current valuation standards (IVSC, 2019), therefore, AVMS require explainability (RICS, 2021). Complementing this literature, we will investigate the overall *accuracy*, *explainability*, and *reliability* of a methodology that permits the influence of appraiser expertise and can therefore be applicable to valuation practice. Specifically, we will explore the *extreme gradient boosting with monotonic constraints* (XGBMC) methodology by predicting the MV constituents of buy-to-let properties in the Netherlands. For real estate valuation, predicting the MV constituents entails the automation of the *single period capitalization method* (SPCM; RICS, 2020a), the most common method in valuing buy-to-let properties (Fakton, 2021)[1].

Extreme gradient boosting (XGB) is a scalable tree-boosting system that controls for overfitting, thus providing better out-of-sample performance compared to traditional tree-based methodologies (Chen & Guestrin, 2016). Monotonicity, in addition, refers to shape constraints that limit the influence of input features on the predicted output, therefore acting as regularization (You et al., 2017). Stated differently, a positive monotonic relationship between X and Y implies that for an increase in X, the predicted variable Y cannot decrease (Bartley et al., 2019). Developing monotonic constraints (MCs) through appraiser expertise improves the explainability of an AVM while maintaining flexible ML properties that make these methodologies attractive (Gupta et al., 2018). Additionally, MCs are argued to improve the generalization of an AVM to unseen valuation 'cases', which is more pronounced when dealing with thin market data (You et al., 2017).

The current study develops an appraiser-based XGB methodology, incorporating appraiser expertise to develop MCs and comparing it to several econometric and ML methodologies. We compare these methodologies on their accuracy, explainability, and reliability by forecasting the three constituents of MV: *vacant possession value* (VPV), *market rent* (MR), and *gross income multiplier*

[1] The single period capitalization method is part of the income approach to valuation. Appraisers calculate market value using the single period capitalization method by capitalizing market rent for a representative single period. Other approaches to market value estimation are divided into the four categories asset-based approach, cost approach, income approach, and market approach (RICS, 2020a).

(GIM). Specifically, based on buy-to-let transaction data from the Netherlands, we compare XGBMC to a parametric HPM, non-parametric random forest (RF), gradient boosting (GB), and XGB, and a hybrid methodology comprising HPM with an XGB estimation procedure. The primary research question of this paper is as follows.

> *How do appraiser-based automated valuation methodologies compare to traditional econometric and machine learning methodologies?*

We answer this research question through several steps. First, comparing AVMs for automation of the SPCM requires a definition of current valuation practice in the Netherlands and how AVMs fit into that context. We therefore adopt the structure of Hilgers et al. (2021) with insights from Dutch (NRVT, 2021) and international (RICS, 2020a) regulatory practice for this purpose, providing an overview of current valuation practice and how AVMs can complement real estate appraisers. Then, we will proceed by providing an elaboration on the derived dependent variable to be estimated by AVMs, MV, and a review of its constituents. Second, a chapter will be devoted to explain the functioning of the proposed AVM methodologies. Specifically, we follow the combined structure of Hinrichs et al. (2021) and Schulz & Wersing (2021), with insights from other authors (e.g., Beimer & Francke (2019); Kroon & Francke (2021); Mullainaithan & Spiess, (2017)). Finally, to provide a valid comparison of the proposed methodologies, we measure accuracy, explainability, and reliability. We follow the structure of Steurer et al. (2021) to empirically measure accuracy and reliability, providing a comprehensive overview of model performance. To measure explainability we adopt conceptual arguments from an appraiser's perspective using insights from Arrieta et al. (2020).

The paper is further structured as follows. Section 2 discusses current valuation practice, how AVMs fit into that context, and defines the three constituents of MV as outlined above. Section 3 contains a theoretical background to each proposed methodology. Section 4 describes the Dutch housing market, data sources, and the steps taken to process the transaction data that are used in this study. Section 5 provides the modeling strategy, an overview of methodological designs, and performance measures. Finally, Section 6 describes the results, after which conclusions and a discussion are provided in Section 7.

2. Valuation framework

Adopting AVMs for real estate valuation can aid professional appraisers in their daily conduct. For an AVM to be worth considering, however, it must fit into current valuation practice. This section will, therefore, discuss valuation practice and how AVMs can fit into that context. First, we will discuss current international standards dictating requirements on professional valuations. Based on these requirements, we will then discuss how AVMs can complement professional appraisers. Finally, definitions will be given to the constituents of MV to be estimated by the AVMs. The topics discussed

in this section serve as a relevant framework for the methodological comparison in the following sections.

2.1 Valuation practice

Professional appraisals are important to various interrelated processes in the property market, including performance measurement, acquisition, and disposal, by acting as a surrogate for transaction prices (McAllister et al., 2003). In the Netherlands, professional valuations are bound to regulations in line with the European (EVS; TEGOVA, 2020) and International Valuation Standards (IVS; IVSC, 2019)[2]. Central to these standards are the risk and liability associated with valuations, and the responsibility appraisers have in conducting these valuations (RICS, 2018). Valuations are to be conducted, for instance, by qualified professionals who must be able to show skill, knowledge, diligence, and ethical behavior. Appraisers have a responsibility to conduct thorough research on the valued property, including internal (e.g., site visitation, structural integrity review, etc.) and external (e.g., land use review, ownership rights, etc.) audits[3]. In doing so, appraisers have a responsibility to communicate all findings of their audits, as these serve as the basis for the final value estimate. The Royal Institution of Chartered Surveyors (RICS; 2020a: 106) describe the importance of due diligence in conducting valuations as the following: "it is the responsibility of the valuer to ensure that they have undertaken an appropriate level of due diligence… as this may have a very significant impact on value." Valuations are inherently uncertain (French & Gabrielli, 2004), and, in addition, clients incur financial risk based on the value estimates from these professionals (Krause et al., 2020). In case of negligence, the liability of negative effects based on the estimates of value resides with these professionals (RICS, 2018; TEGOVA, 2020).

Given the scrutiny in conducting professional appraisals, it is surprising to see systematic differences between appraised and transacted values. MSCI (Walvekar & Kakka, 2019), for instance, report that appraisal and transaction values in the Netherlands have a weighted average absolute difference of 10% in 2019. Historically, differences between these values arise from various sources. McAllister et al. (2003) discuss that appraiser behavior strongly influences appraisal value, which can cause differences with transaction values. For example, historic appraisals influence current appraisals through an 'anchoring bias'. Specifically, appraisal smoothing occurs as appraisers anchor onto the previously appraised value, even though previous appraisal values does not necessarily reflect current value (Clayton et al., 2001). Additionally, professional appraisers employ methodologies requiring historic transaction data and are hence lagging. Appraisers are therefore slow in adopting non-transaction-based information (e.g., market changes) into appraisals (McAllister, 2003). Apart from appraiser behavior, client influence on independent appraisals is another potential issue associated with

[2] The main regulator of the European and International Valuation Standards in the Netherlands is the Register of Real Estate Appraisers Netherlands (NRVT, 2021).
[3] The Royal Institution of Chartered Surveyors (RICS, 2019) describes the due diligence process for the UK and the Register of Real Estate Appraisers Netherlands (NRVT, 2017) for the Netherlands.

manual valuations. Crosby et al. (2018) compare differing client needs and conclude on the existence of significant differences in appraisal outcomes. Although this result is contingent on circumstances (e.g., appraisal salience, market context), it raises a potential issue with manual valuations.

In this context, AVMs are mentioned as valuable tools in conducting valuations. Mooya (2011) argues for instance that the ability of AVMs to provide standardized valuations cause them to be more sophisticated and accurate than manual valuations. The author further states there to be no reason for AVMs not to completely replace appraisers. More recently, a report by the RICS (Scheurwater, 2017) describes two sides to the debate on the role of AVMs in valuation practice. On the one hand, and in line with Mooya (2011), reasons are mentioned like those described above to argue AVMs to be feasible in replacing manual valuations. On the other hand, valuations are argued to be "part art, part science" (Scheurwater, 2017: 25). Although AVMs can replace the "science" part of valuation, streamlining the routine aspects of the profession, AVMs will not be able to replace the "art" of valuing, others argue.

AVMs affect the valuation industry. There is still, however, an ongoing debate on the liability of valuations. It is unclear, for example, who assures the value estimates by AVMs and, if AVMs cause significant loss to clients acting on AVM estimates, who is held accountable (RICS, 2021). For now, at least, Scheurwater (2018) argues AVMs to be tools for most valuations, complementing the appraiser. Improving transparency and setting standards for assessing the quality of AVMs, such as those proposed by the IAAO (2018), can expand the usefulness of AVMs in the valuation process.

2.2 Market value

The estimation of MV is the primary objective of commercial valuations. MV is defined as (IVSC, 2019: 18; TEGOVA, 2020: 27):

"The estimated amount for which an asset or liability should exchange on the valuation date between a willing buyer and a willing seller in an arm's length transaction, after proper marketing and where the parties had each acted knowledgeably, prudently, and without compulsion."

The "estimated amount" excludes any special terms, considerations, and concessions and "an arm's length transaction" presumes the transaction to be between unrelated parties, each acting independently (IVSC, 2019: 19). Aligning with current valuation practice and associated regulations as outlined above requires an AVM that automates a methodology currently applied by professional appraisers. To estimate MV, appraisers use three methods: the market approach, income approach, and cost approach. Although all methods are feasible in calculating MV, we will investigate the automation of the income approach, more specifically the SPCM (RICS, 2020a), as it is most commonly applied in valuing buy-to-let properties (Fakton, 2021). Applying an AVM to the constituents of the SPCM to MV estimation has the additional benefit of allowing appraiser adjustments when model results become unreliable (e.g.,

in case of nonstandard properties; Hilgers et al., 2021). The parameters used in the SPCM are threefold. First, VPV is defined as (Fakton, 2021: 36):

"The estimated selling price based on buyer's costs, free of rent and other expenditures… is synonymous with the term 'market value freehold with vacant possession'"[4]

Meaning VPV is similar in concept to MV, but is calculated using the special assumption that the property is vacant. Special assumptions are assumptions that either (1) assume facts that differ from the actual facts existing at the date of valuation, or (2) that would not be made by a typical market participant in a transaction on the date of valuation (RICS, 2020a: 36). The special assumption of vacant possession is used in the valuation of buy-to-let properties as a benchmark measure and is typically higher than MV. Second, MR is defined as (IVSC, 2019: 21):

"The estimated amount for which an interest in real property should be leased on the valuation date between a willing lessor and a willing lessee on appropriate lease terms in an arm's length transaction, after proper marketing and where the parties had each acted knowledgeably, prudently and without compulsion."

Or in other words, MR is the value expected to be paid as rent for a property (TEGOVA, 2020: 30). The conceptual framework in support of the definition of MR is similar to that of MV described above. Specifically, in the definition, "estimated amount" excludes any special terms, considerations, and concessions, while market participants typically agree on "the appropriate lease terms" for the type of property on the date of valuation (IVSC, 2019: 21). Finally, the GIM is defined as (IVSC, 2003: 406):

"The ratio between the sale price or value of a property and the average [gross] annual income or income expectancy. It is applied to income to arrive at capital [market] value."

Also described as *years' purchase* in the Commonwealth, appraisers determine appropriate levels of GIM through comparable transactions using factors reflecting risk levels (RICS. 2020b).

To accurately estimate these parameters, appraisers require three types of attributes associated with buy-to-let properties. First, market attributes are all conditions associated with the market during which the property is valued. Second, spatial attributes are the locational characteristics. Finally, structural attributes entail all physical characteristics of the property. Based on relationships between the property's attributes, input variables, and MV, we construct Figure 1 (Fakton, 2021). Specifically, in valuing buy-to-let properties, the input parameters VPV, MR, and GIM are estimated based on the

[4] Translated from the Dutch definition: "De geschatte verkoopprijs op basis van de kosten koper, vrij van huur en overige lasten… is synoniem voor de term 'marktwaarde vrij van huur en gebruik'" (Fakton, 2021: 36).

market, spatial, and structural attributes of the property. By selecting reference properties possessing similar attributes, appraisers compare and adjust these input parameters. Determining MV using the input parameters, in turn, requires consideration of various ratios, potentially further adjusting the input parameters (Fakton, 2021). VPV, for instance, influences MV in two ways (for background on how VPV and MV differ, see Conijn & Schilder (2011)). First, appraisers compare the MR-to-VPV ratio to the GIM. Specifically, appraisers use the MR-to-VPV ratio to determine an appropriate level of the GIM based on comparable properties. Due to higher probabilities of vacancies, properties yielding higher MR-to-VPV ratios carry higher risk. As demonstrated by the risk-return tradeoff (Markowitz & Blay, 2014), higher risks must be compensated with a higher GIM. The MV-VPV ratio, finally, serves as a reference in determining accurate levels of MV[5]. Based on these ratios, appraisers calculate the MV of buy-to-let properties using MR-to-GIM in the SPCM.

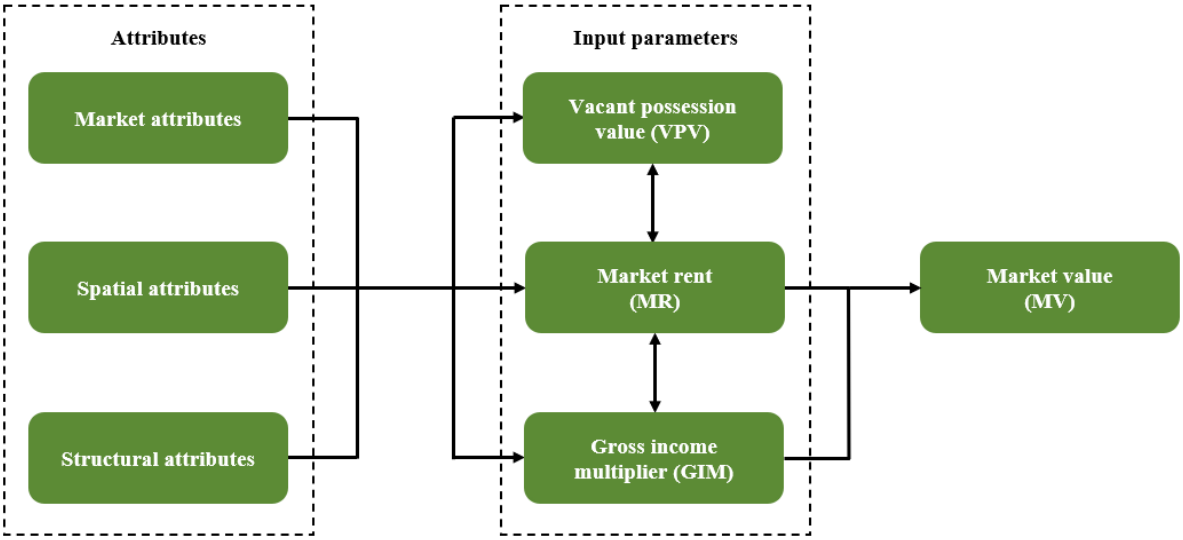

Figure 1. Relationships between transaction attributes (market, spatial, and structural), input parameters (vacant possession value, market rent, and gross initial yield) and market value.

3. Automated valuation methodologies

This section reviews the proposed AVMs we use to predict the MV constituents as outlined above[6]. Specifically, we discuss the more traditional HPM and newer ML methodologies in this section. Within ML methodologies, we discuss the tree-based methodologies RF, GB, and XGB. Additionally, we provide background to the hybrid methodology comprising HPM with an XGB estimation procedure

[5] Based on a property's attributes, appraisers determine appropriate market value-vacant possession value ratios. The Register of Real Estate Appraisers Netherlands (NRVT, 2021) provides extensive discussion on the vacant possession parameter in valuing Dutch buy-to-let properties.
[6] Various methodologies suitable as an automated valuation methodology exist. For the purposes of comparing our proposed methodology, however, these are out of the scope of this paper. Interested readers are referred to other authors, including Malpezzi (2002; hedonic pricing models), and Shalev-Shwartz & Ben-David (2014; machine learning methodologies).

and the newly proposed development of an XGB methodology with MCs (XGBMC). The proposed methodologies are shown in Figure 2. First, however, we discuss the method used in comparing the proposed methodologies: cross-validation (CV).

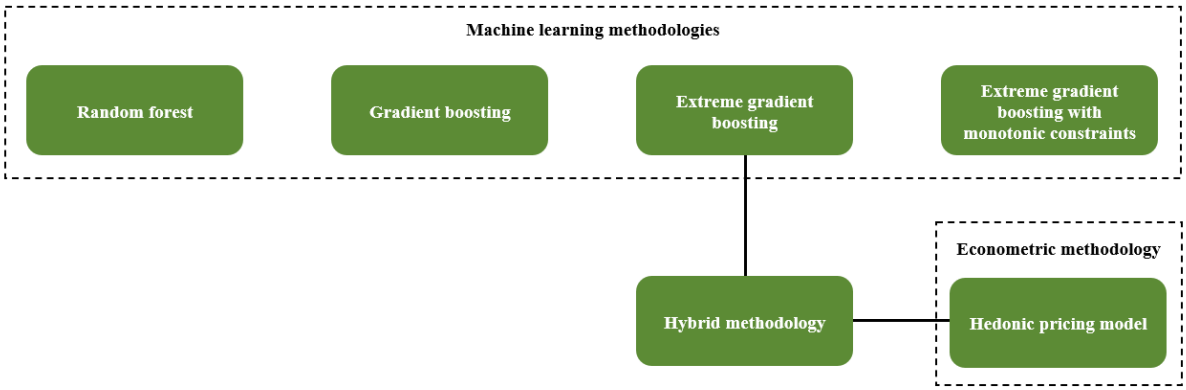

Figure 2. Automated valuation methodologies, divided into machine learning methodologies (random forest, gradient boosting, extreme gradient boosting, and extreme gradient boosting with monotonic constraints), an econometric methodology (hedonic pricing model), and a hybrid methodology comprising the hedonic pricing model with an extreme gradient boosting estimation procedure.

3.1 Comparing methodologies

Predictions made by AVMs contain inherent uncertainty; reporting a single prediction value does not provide sufficient information regarding an AVMs performance. Linear-based methodologies, like the HPM, can calculate standard error outputs due to parametric model assumptions. Confidence intervals can therefore be calculated within these methodologies, enabling an interpretation of model accuracy and reliability (Krause et al., 2020). Non-linear methodologies, like the other methodologies proposed in this paper, however, do not share the same properties as methodologies from the former type and are hence unable to calculate uncertainty measures directly. To compare both types of methodologies, therefore, we adopt a resampling procedure. In contrast to other, error-based procedures, resampling procedures do not require assumptions on symmetry and normality of the error distribution (Krause et al., 2020). One such procedure is bootstrap resampling. Bootstrap resampling refers to the statistical technique of drawing sample observations from the original data, returning drawn samples to the original data (i.e. replacement), until reaching the sample size (Davison & Hinkley, 1997). In the context of AVMs, methodologies are fit on this bootstrap sample and tested on the remaining data (i.e. out-of-bag sample). This procedure is repeated various times, allowing the computation of accuracy and reliability metrics.

Although resulting in good estimates of model accuracy and reliability, bootstrap resampling is computationally expensive. We therefore choose to adopt another resampling procedure: CV, an umbrella term for various out-of-sample testing techniques (Steurer et al., 2021). CV techniques prevent overfitting by splitting the data into training and testing parts, using the training data to fit a particular model, while the model accuracy and reliability are obtained from the testing, or hold-out data

(Goodfellow et al., 2015). The overall performance in k-fold CV (a CV technique) using k data splits, or folds, is calculated by taking the average performance across k trials (Steurer et al., 2021). CV techniques are generally more efficient than other resampling techniques (see, e.g., Nakatsu, 2021) [7]. Additionally, Yang (2007) argues that CV techniques allow comparison of methodologies that differ in their specification procedures, such as parametric and non-parametric methods. Specifically, Yang (2007) argues that using CV, under several conditions (e.g., appropriate splitting ratio), selects the best model with probability converging to 1.

Within this paper, we apply a nested repeated k-fold CV strategy to measure the accuracy and reliability of the proposed methodologies. Repeating CV ensures that the results obtained are not arbitrarily based on the specific splits. We further choose a nested CV strategy as most ML methodologies require tuning of hyperparameters. Hyperparameters, in contrast to parameters, are manual settings of methodologies commonly used in ML. Because of the manual nature of hyperparameters, we therefore require a similar CV technique in the training data to optimize the hyperparameters of each ML methodology[8]. Figure 3 displays our repeated k-fold CV strategy for the performance measurement in terms of accuracy and reliability. Specifically, we choose to repeat ten-fold CV three times. We optimize hyperparameters by repeating five-fold CV three times during each aforementioned fold and repeat in the training data (hence, nested CV). In line with Steurer et al. (2021), we obtain in-sample hyperparameter set performance using the root mean squared error (RMSE). Using the metrics described in Section 5.3, we calculate model accuracy and reliability on the hold-out sample for each fold and repeat. We determine aggregate accuracy performance for each methodology and performance measurement based on the average across folds and repeats. By calculating error statistics across folds and repeats, we determine model reliability (further described in Section 5.3).

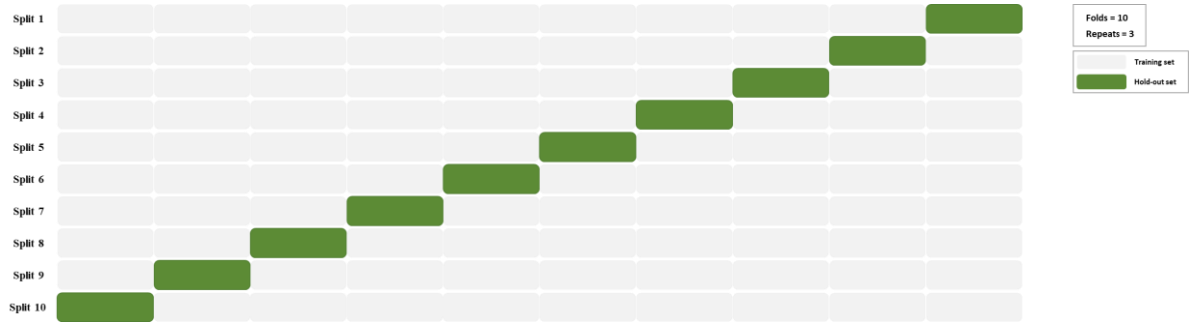

Figure 3. Repeated k-fold cross-validation strategy for the performance measurement in terms of accuracy and reliability using ten folds and three repeats.

[7] Although cross-validation is more computationally efficient than bootstrap resampling, we found an average throughput time of model training and evaluation of three days per methodology and market value constituent.
[8] Setting hyperparameters is standard in machine learning methodologies as these determine the accuracy and reliability of outcomes. See Probst et al. (2019) for an overview of hyperparameter settings in machine learning methodologies.

3.2 Hedonic pricing model

Formalized by Rosen (1974), HPMs relate observed outcomes to implicit or hedonic prices. Sheppard (1999) discusses that the value of heterogeneous products, such as properties, can be disentangled into an underlying vector of attributes. MV constituents, in our example, can therefore be regressed on property attributes (Malpezzi, 2002). The implicit pricing of these attributes is what constitutes an HPM. Apart from standard econometric issues associated with the HPM (see, e.g., Sheppard, 1999), however, constrains exist when predicting MV constituents using HPMs due to several issues. First, the HPM requires ex-ante decisions on relevant variables and functional form, making it less flexible compared to other methodologies (Robinson & Sanderford, 2017)[9]. Additionally, specification errors are probable as it is inconceivable to construct an exhaustive HPM including all relevant relationships between variables (Peterson & Flanagan, 2009). In this study, we denote the prediction of the constituents of MV using the HPM methodology as:

$$MV^{VPV,MR,GIM} = X\beta + \delta \quad (\mathbf{1})$$

In Equation (1), $MV^{VPV,MR,GIM}$ represent MV constituents, X a vector of market, spatial, and structural attributes, β a vector of associated coefficients to be estimated, and δ time-fixed effects. Although standard for the HPM, the error-term, or *noise*, is not displayed as the error-term is estimated using the procedure described in Section 3.1 and further elaborated in Section 5.3.

3.3 Random forest

RF is a tree-based methodology introduced by Breiman (2001a). In the context of house price prediction, RFs are made up of many decision trees (DT), also called classification and regression trees (CART). A simple decision tree is shown in Figure 4. The figure shows a regression tree consisting of both *nodes* and *branches*. Nodes make up the feature selection mechanisms that best split the data into non-overlapping regions (Schulz & Wersing, 2021). Branches represent outcomes, connecting nodes based on this feature selection. Starting at the top, the *root node* is the first node of the tree and involves the initial feature selection criterium. These splits continue in other nodes, called *child nodes*, until a terminal node containing the value estimate, called a *leaf*, is reached (Mullainathan & Spiess, 2017). The value estimate is the mean value of all training data instances that meet the categorical requirements of feature selections in the respective leaves. We can describe this process, similar to the HPM, as products of dummy variables (e.g., $X_1 = 1_{\text{Bedrooms}>2} * 1_{\text{property size}>100}$), with corresponding coefficients (e.g., € 300.000). Recursively evaluating features in nodes to best split the data is called the training process

[9] To compare our appraiser-based methodology, we adopt a standard linear hedonic pricing model. More flexible econometric specifications are, however, capable of modeling complex non-linear relationships. For derivations of these econometric methodologies see, for example, Gao & Li (2013), Malikov & Sun (2017), and Debarsy & Lesage (2021).

and is conducted by measuring performance in terms of *information gain* (e.g., mean squared error; Breiman, 2001a). Narrowing down feature selection to be more specific, thereby increasing *tree depth*, will cause training samples to be predicted more accurately. Although more tree depth will lead to accurate results *in-sample*, it comes at the cost of lower *out-of-sample* prediction accuracy; a phenomenon known as *overfitting* (Mullainathan & Spiess, 2017). Optimizing out-of-sample prediction accuracy involves limiting the size of the tree through *regularization* (see, e.g., Varian (2014)), and by aggregating trees.

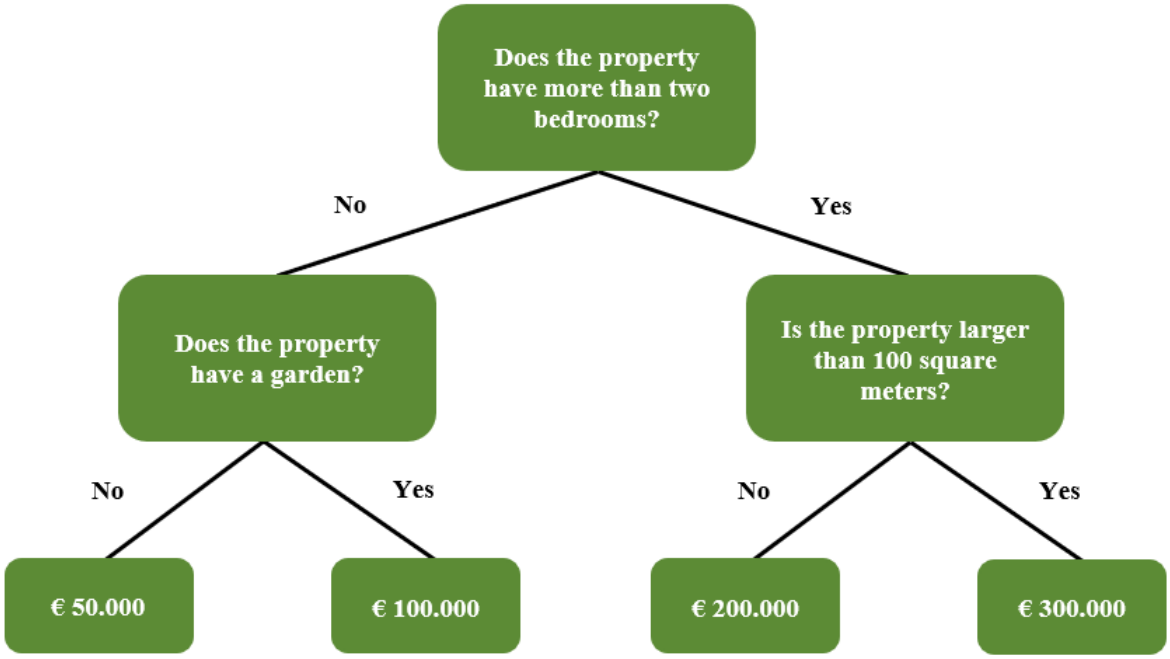

Figure 4. Simple regression tree consisting of a *root node*, two *child nodes*, and four terminal nodes, or *leaves*.

RF improves individual DT accuracy by growing an ensemble of regression trees in a *forest* (Breiman, 2001a). If individual trees are uncorrelated, RFs can cancel out the influence of noisy data during training through a bagging approach. In the context of RF development, bagging, or *bootstrap aggregating*, refers to the random selection of training data for each tree with replacement and combining individual results through aggregation (Breiman, 1996). This process makes RFs achieve higher accuracy than any individual tree. In this study, we denote the prediction of the three constituents of MV using the RF methodology as (Schulz & Wersing, 2021):

$$MV^{VPV,MR,GIM} = \frac{1}{T}\sum_{s=1}^{S}\sum_{t=1}^{T}\mathbb{1}(x \in S_{s,t})\,\theta_{s,t}^{VPV,MR,GIM} \quad (\mathbf{2})$$

In Equation (2), $MV^{VPV,MR,GIM}$ represent MV constituents and $\mathbb{1}(x \in S_{s,t})$ dictates an indicator function that equals one if the property's characteristics, denoted as x, fall into the set S_s of tree t. The mean values of the MV constituents associated with the set S_s of tree t are then denoted as $\theta_{s,t}^{VPV,MR,GIM}$. Aggregating the estimates of trees is performed by taking the average values generated by all trees, denoted as T (Breiman, 1996).

3.4 Gradient boosting

GB functions similar to RF, relying on the premise of aggregating individual predictors (Friedman, 2001). In contrast to RF, however, GB differs on several elements that make it more accurate in regression-type problems, such as valuing real estate (Graczyk et al., 2010). First, contrary to a bagging approach in RF, GB adopts a *boosting* approach involving the aggregation of *weak learners* (Hastie et al., 2017: 337–384). These weak learners involve shallow trees, called *stumps*, where the first tree is fitted as standard. All additional trees are then fitted to the residuals of predecessor trees in a sequential learning process, minimizing the loss function using a *gradient descent algorithm*. Graphically, this sequential process can be represented using Figure 5.

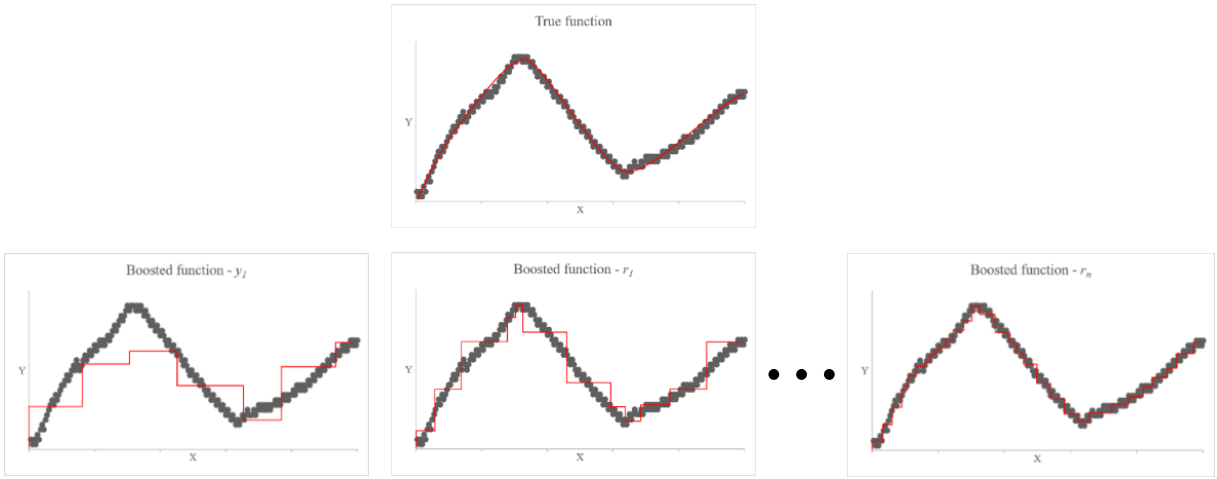

Figure 5. Gradient boosting learning process. The upper figure shows the true function the gradient boosting learning process tries to approximate. The lower figures display the boosted function after each iteration. Specifically, the first lower figure shows the boosted function after the approximation of y_i and the additional lower figures display the boosted functions after fitting the residuals of predecessor trees in a sequential learning process.

As shown in the figure, the gradient boosted function approximates the true function through each sequential iteration. To prevent overfitting, individual trees are only partially considered in the aggregated model through a *learning rate* (Schulz & Wersing, 2021), leaving space for additional trees to improve the aggregate model (Chen & Guestrin, 2016). In this study, we denote the prediction of the three constituents of MV using the GB methodology as:

$$MV^{VPV,MR,GIM} = y_1^{VPV,MR,GIM} + \eta \sum_{j=1}^{n} r_j \quad (\mathbf{3})$$

In Equation (3), $MV^{VPV,MR,GIM}$ represent MV constituents. $y_1^{VPV,MR,GIM}$ denotes MV constituent predictions made by the first regressor, r the residuals of regressor j, and η the learning rate[10].

3.5 Extreme gradient boosting

Introduced by Chen & Guestrin (2016), XGB is an extension of the GB methodology. XGB, however, adopts several contrasting properties. First, XGB does not require the sequential model fitting of individual trees. Instead, it adopts a parallel algorithm for optimal split finding. In contrast to RF and GB, XGB adds a regularization term to the standard loss function used for split optimization in decision-tree-based methodologies (Chen & Guestrin, 2016: 2):

$$\Omega = \gamma L + \frac{1}{2}\lambda \|w\|^2 \quad (\mathbf{4})$$

In Equation (4), γ and λ are user-specified regularization terms[11], L denotes the number of leaves (i.e. terminal nodes), and $\|w\|$ represents the vector norm of leaf weights, or predicted values. γ is meant to encourage tree *pruning* (i.e. decrease tree complexity), while larger values of λ are meant to decrease the sensitivity to individual weight observations w. Higher values of Ω, therefore, decrease the likelihood a tree specification ends up in the aggregated model. When the regularization parameter Ω is set to 0, predicting the MV constituents $MV^{VPV,MR,GIM}$ using the XGB methodology falls back to the GB methodology. In addition to this regularization term, XGB incorporates a sparsity-aware algorithmic measure to deal with sparse (e.g., missing) input data. Specifically, XGB uses non-missing values to learn the best direction to handle sparse values. These unique features of XGB improve out-of-sample performance compared to GB, while also enhancing training speed (Chen & Guestrin, 2016)[12].

3.6 Hybrid methodology

To combine the benefits of stand-alone econometric and ML methodologies, Ceyhan (2017) developed two hybrid methodologies consisting of a hierarchical trend model (HTM) with an iterative estimation procedure using ANNs and RFs. As described by Kroon & Francke (2021), adopting the methodology in a later study, the hybrid model combines the *structure* of econometric methodologies with the

[10] For a more detailed derivation of the gradient boosting methodology, see Friedman (2001).
[11] α is a median-based third regularization parameter and functions similar to the mean-based λ as both decrease sensitivity to individual weight observations. Although not explicitly described in Chen & Guestrin (2016), implementations of the extreme gradient boosting methodology include all three regularization parameters.
[12] In this study, we focus on the most important differences between the extreme gradient boosting methodology and other methodologies. The exact derivation of extreme gradient boosting and its differences with GB are outside the scope of this paper. Interested readers are referred to Chen & Guestrin (2016).

flexibility of ML methodologies. The hybrid methodologies proposed by these authors are based on a methodology adopted in Schimert & Wineland (2010) to predict engine wear. Specifically, the authors estimate the residuals of a RF methodology with a dynamic linear model (DLM) using a Kalman filter. The combined methodology outperformed the stand-alone models in terms of prediction accuracy.

Similarly, other authors combine the linear capabilities of econometric methodologies with nonlinear capabilities of ML. Zhang (2003), for example, combines an autoregressive integrated moving average (ARIMA) time series model with ANNs and argues combining both yielding better results. In a later study, Smyl (2020) argues econometric methodologies to be able to measure the main level and seasonality, while ML methodologies are capable of capturing non-linear trends. In line with these authors, we will investigate the merit of combining both types of methodologies in a hybrid methodology. Specifically, we develop a methodology combining the structure of an HPM with the flexibility of an XGB estimation procedure. We estimate the constituents of MV using the hybrid methodology as follows:

$$MV^{VPV,MR,GIM} = X\beta + \delta + \varepsilon \quad (\mathbf{5})$$

$$where: \varepsilon = f(X)$$

In Equation (5), $MV^{VPV,MR,GIM}$ represent MV constituents, X a vector of market, spatial, and structural attributes, β a vector of associated coefficients to be estimated, δ time-fixed effects, and ε the residuals, or *noise*. The residuals are then estimated by the XGB methodology using the function $f(X)$. Specifically, we first train an HPM without modification to capture the linear component. The resulting nonlinear residuals (ε) between the predicted MV constituents ($\widehat{MV}^{VPV,MR,GIM}$) and the actual MV constituents ($MV^{VPV,MR,GIM}$) in the training data are then estimated using the XGB methodology.

3.7 Monotonic constraints

MCs provide guarantees on how the output should depend on input (Canini et al., 2016). For example, when appraising a property, all else being equal, appraisers may want to constrain the MV of a property to be monotonically increasing or decreasing with certain attributes. MCs can be formally denoted as the following:

$$\textbf{Positive monotonicity} - \text{ For all } x_2 \geq x_1, MV(x_2) \geq MV(x_1) \quad (\mathbf{6})$$

$$\textbf{Negative monotonicity} - \text{ For all } x_2 \geq x_1, MV(x_2) \leq MV(x_1) \quad (\mathbf{7})$$

Meaning MV constituents are non-decreasing (non-increasing) with increasing values of monotonically constraint input attributes x. Graphically, monotonicity can be represented using Figure 6. As shown in

the figure, in the example of positive monotonicity, output values do not have to exclusively increase with higher values of monotonically constraint input attributes, but cannot decrease.

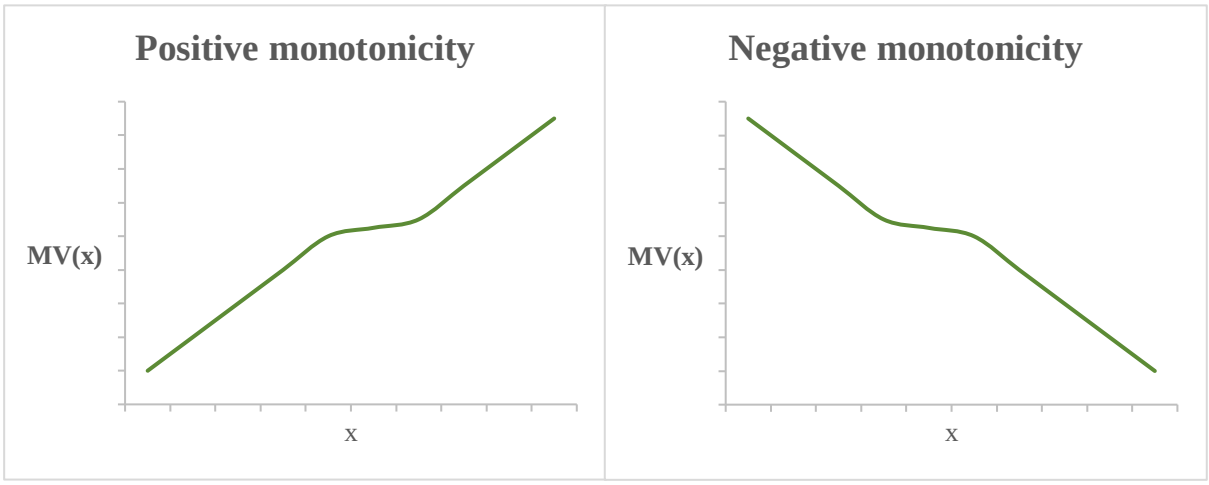

Figure 6. Monotonicity. Imposing positive (negative) monotonicity implies that for an increase in a positively (negatively) monotonically constraint attribute x, market value constituents MV(x) cannot decrease (increase).

Imposing monotonicity on input acts as regularization, producing more stable model results, improving generalizability (You et al., 2017), and increasing explainability (Gupta et al., 2018). Unlike other regularization measures, MCs do not require tuning other than the binary decision of whether and in what direction to apply shape constraints. Additionally, constraining the model specification by using MCs provides a clear semantic meaning to users (e.g., appraisers) and clients on how each shape-constrained attribute affects MV constituents. Restricting models a priori by imposing monotonicity, however, raises a potential issue. Specifically, the selection of features to monotonically constrain requires strong prior belief and expertise of the regression problem. Enforcing monotonicity on features that are not purely monotonically related to output variables will unnecessarily hamper flexibility compared to similar, non-monotonically constraint methodologies.

In deciding what features to monotonically constrain in the XGB methodology, we use the expertise of appraisers at a valuation firm (see Appendix 1: Envalue valuation firm), for which arguments are provided in Section 5.2. When carefully developed, the proposed XGBMC methodology can improve generalizability while also increasing explainability compared to existing methodologies.

4. Data

4.1 Market and data sources

Our case study is situated in the Netherlands, comprising buy-to-let property transactions. The housing market of the Netherlands has experienced large price developments over the last few years. For instance, transaction prices, in general, have risen by more than 5% per annum since 2016 (CBS, 2021a). Most recently, price developments were more pronounced with an average increase in transaction prices of 18% in the third quarter of 2021 compared to the same quarter in 2020 (NVM, 2021). In total, an

average of approximately 227,000 residential property transactions occurred in the Netherlands per year in the period 2019–2020 (CBS, 2021b). For our case study, we obtain a sample of 31,888 vacant transactions and 7,595 rental transactions in 2021. Additionally, we obtain a sample of 1,275 investment transactions from the years 2019, 2020, and 2021. These transactions are retrieved from a valuation firm (see Appendix 1: Envalue valuation firm) and processed into separate datasets. The raw data received contains information on the VPV (in case of vacant transactions), MR (in case of rental transactions), and GIM (in case of investment transactions). We measure VPV and MR per square meter of usable floor area to investigate the effect of constraining certain attributes through MCs (elaborated further in Section 5.2).

We also obtain property characteristics. In the case of investment transactions, these include construction year, the Leefbaarometer[13], theoretical rental income, whether the property is single- or multifamily, transaction date, number of residential units, and usable floor area. In the case of rental transactions, these include construction year, days on the market, the Leefbaarometer, the number of rooms, object volume, whether the property is single- or multifamily, transaction date, type- and subtype of object, and usable floor area. We receive vacant transaction data with similar characteristics as the rental transaction data with the additional feature of whether a property was transacted with purchaser's costs. Raw transaction data for each type are then enriched, cleaned, and transformed, elaborated below.

4.2 Data enrichment

The received data has 7 relevant features for the investment dataset, 10 relevant features for the rental dataset, and 11 relevant features for the vacant transaction dataset. Prediction accuracy, however, improves with additional input features (Steurer et al., 2021), such as locational attributes and environmental characteristics. To increase the number of features in the raw transaction data, we establish application programming interface (API) connections with real estate-related sources in the Netherlands[14]. All of the established connections and retrieved features are displayed in Table 1. Specifically, we expand our datasets by extracting the residence ID, address coordinates (EPSG:4289 format), and usable floor area from Basisregistratie Adressen en Gebouwen (BAG; Kadaster, 2021). We adopt usable floor area estimates from BAG to enrich missing usable floor area information received in the original datasets. Based on the retrieved residence ID, we obtain the energy label from the EP-online database, maintained by the Netherlands Enterprise Agency (NEA, 2021). The coordinates of BAG are translated into geodetic latitude and longitude and used as input for the Walk Score API. Walk Score (2021) provides a walkability score from an address (in terms of coordinates) to nearby amenities by evaluating walking routes, which we adopt in each of our considered datasets.

[13] The Leefbaarometer measures the livability of a district based on five indicators: amenities, inhabitants, physical environment, properties, and safety. For background on the Leefbaarometer and its specific measurement, see Leidelmeijer et al. (2019).

[14] Python code used in developing application programming interface (API) connections are obtainable from the author.

Based on address information, we obtain the names and codes of the neighborhoods, districts, municipalities, and provinces from Public Services on the Map (PDOK, 2021). Using this information, we measure locational attributes through data available at Statistics Netherlands. Specifically, we obtain the average distance in a neighborhood to eleven amenities and populations of districts, municipalities, and neighborhoods based on district, municipality, and neighborhood codes from Statistics Netherlands (2021c). Additionally, we obtain average household size, housing stock, average property value as set by municipalities, percentage of single- and multi-family homes, and the percentage of owner-occupied and private/socially rented properties on the neighborhood level.

Table 1. Features retrieved from public real estate sources using application programming interface (API) connections.

Source	Source description	Extracted features
Basisregistratie Adressen en Gebouwen	Basisregistratie Adressen en Gebouwen is a Dutch registry of all addresses and buildings in the Netherlands, maintained by the Netherlands' Cadastre, Land Registry, and Mapping Agency (Kadaster, 2021).	Address coordinates – EPSG:4289 format Residence ID Usable floor area
EP-online	EP-online is the official Dutch database where energy advisors register energy labels and indicators related to energy performance. The Netherlands Enterprise Agency maintains the EP-online database.	Energy label
Public Services on the Map	Public Services on the Map is a platform that makes geographically related data from Dutch governmental bodies publicly available (PDOK, 2021).	District code District name Municipality code Municipality name Neighborhood code Neighborhood name Province code Province name
Statistics Netherlands	Statistics Netherlands is a national Dutch statistical office maintaining data on various topics, including real estate-related information (CBS, 2021c).	Average household size Average property value Distance to eleven amenities[15] District population Housing stock Municipality population Neighborhood population Percentage owner-occupied and private/socially rented Percentage single-/multifamily homes
Walk Score	The Walk Score is a number measuring the walkability of an address to nearby amenities by evaluating walking routes. Scores are provided in a range from 0–100, with higher scores indicating better walkability (Walk Score, 2021).	Walk Score

Note: Table 1 displays features extracted from application programming interface (API) sources to enrich raw transaction data.

[15] The specific amenities are not displayed for concision, but include daycare, fire brigade, general practitioner, highway, hospital, movie theater, out-of-school care, pharmacy, primary school, supermarket, and train station.

4.3 Data cleaning

Data on all three types of transactions were received in a raw format. To investigate the additional explainability of adding MCs in the XGB methodology and how the accuracy and reliability of this proposed methodology compares to existing methodologies, we choose to balance the transaction data through several steps (see Gelman & Hill (2007: 199–231) for elaboration on data balance). Rental and vacant transaction data were received in a similar format, therefore being processed simultaneously. Investment transaction data were received with differing features; we, therefore, describe the cleaning procedure of investment transactions separately.

In the case of investment transactions, we take the following steps. First, we restrict investment cases based on usable floor area, residential units, and net purchase price. Specifically, we restrict investment transactions between 50 and 15,000 square meters of usable floor area, removing 49 investment transactions. Additionally, we remove 100 investment transactions that consist of more than 500 residential units. Third, we remove 239 transactions that have a net purchase price of lower than € 80,000 or larger than € 5,000,000. Finally, we restrict transactions with a GIM between 10 and 100, removing one transaction. Predicting the GIM of these properties is inherently uncertain due to the small number of market participants and therefore transaction data. Finally, we remove 402 investment transactions with missing values on certain features.

In case of rental and vacant transactions, we take the following steps. First, we remove uncommon properties in terms of usable floor area. Specifically, we remove all properties smaller than 20 and larger than 250 square meters of usable floor area. Based on this requirement, we remove 70 rental transactions and 787 vacant possession transactions. We further restrict rental prices between € 8 and € 40 per square meter of usable floor area, per month, and transaction prices between € 1,400 and € 10,000 per square meter of usable floor area. This restriction causes us to remove 233 rental transactions and 3,723 vacant transactions. Predicting the MR and VPV of these properties is inherently uncertain due to the small number of market participants and therefore transaction data. Finally, we remove transactions with missing values on certain features, resulting in the removal of 3,779 rental transactions and 9,344 vacant transactions.

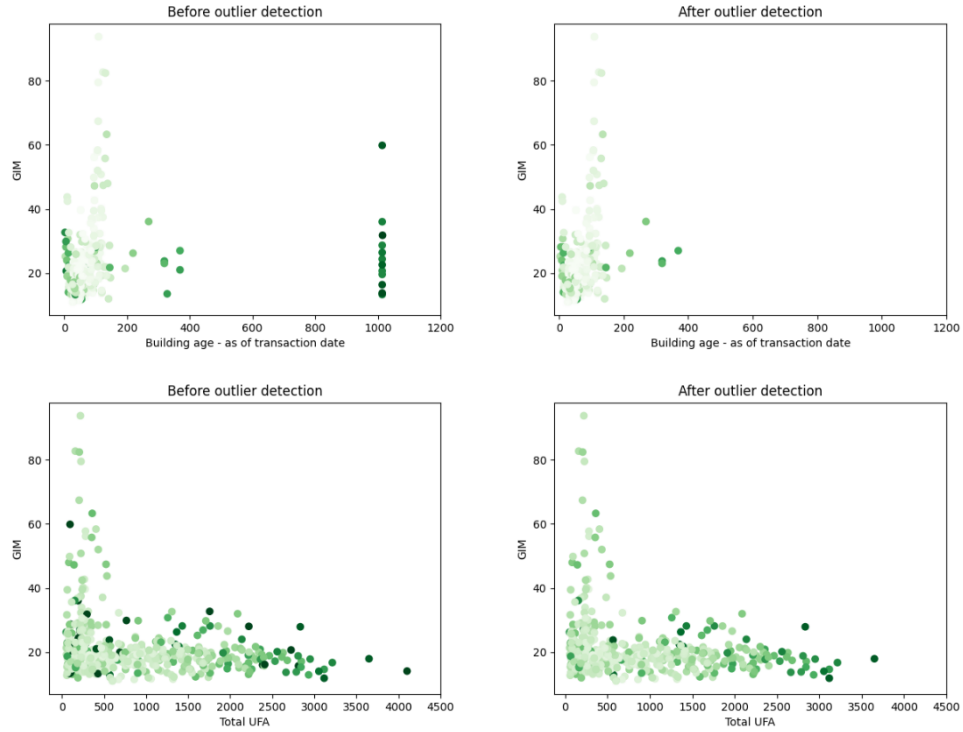

Figure 7. Building age, total usable floor area, and gross income multiplier distributions in the investment transaction dataset before (left) and after (right) outlier detection using the isolation forest algorithm. Darker dots indicate higher anomaly scores.

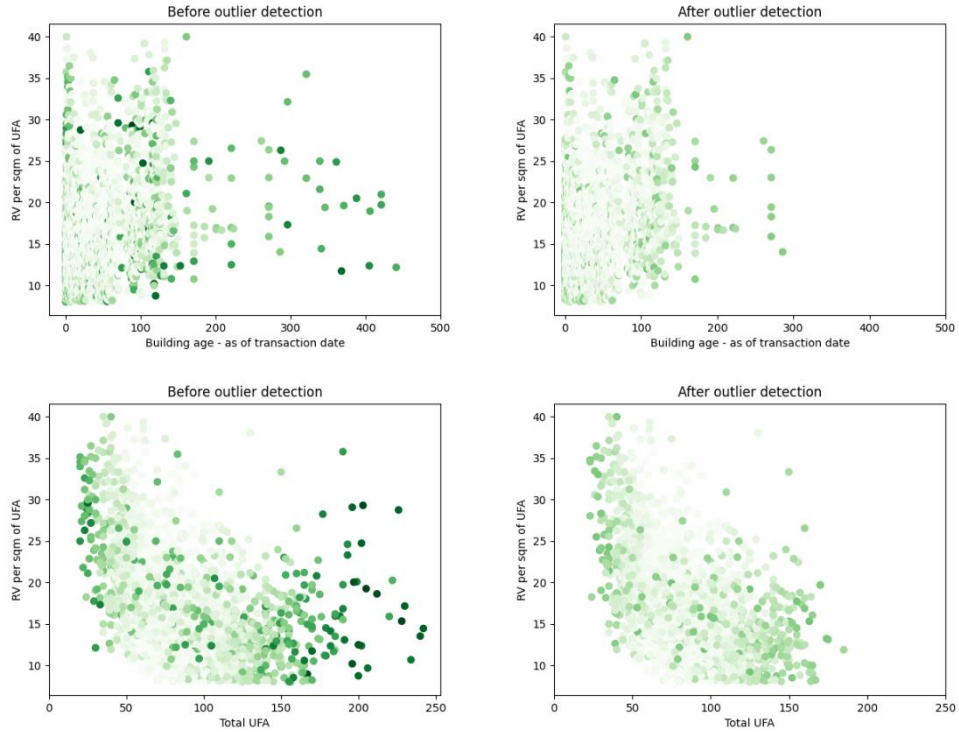

Figure 8. Building age, total usable floor area, and market rent distributions in the rental transaction dataset before (left) and after (right) outlier detection using the isolation forest algorithm. Darker dots indicate higher anomaly scores.

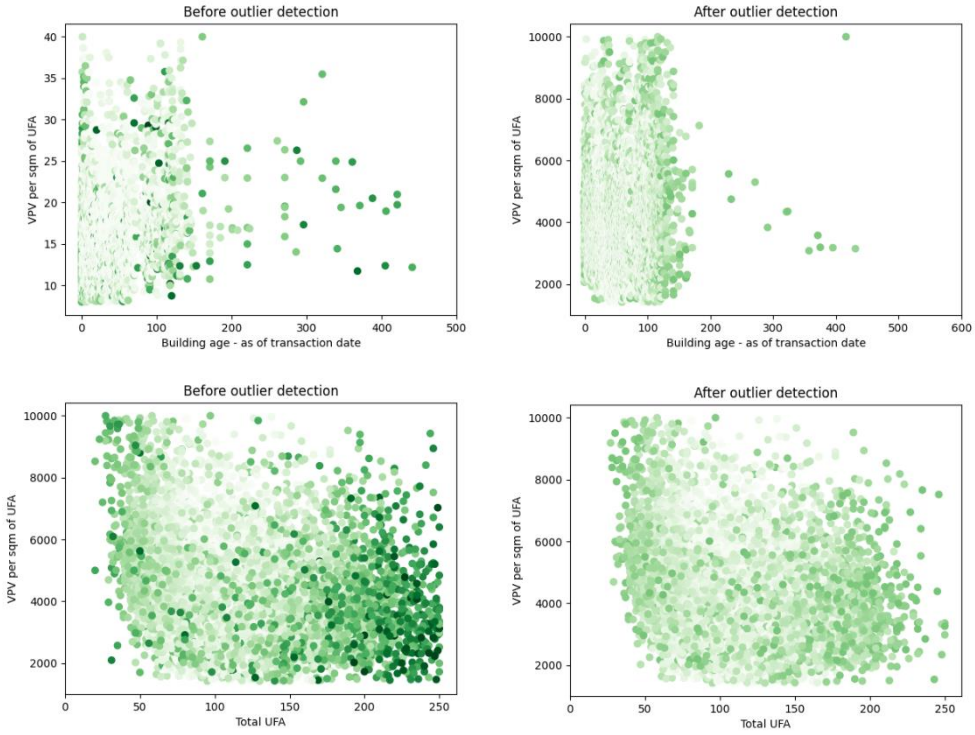

Figure 9. Building age, total usable floor area, and vacant possession value distributions in the vacant transaction dataset before (left) and after (right) outlier detection using the isolation forest algorithm. Darker dots indicate higher anomaly scores.

Based on Liu et al. (2008) and in line with Steurer et al. (2021), we use an isolation forest (IF) to detect and remove outliers. An IF is an unsupervised anomaly detection algorithm that isolates cases that have different characteristics from normal instances (Liu et al., 2008). Using the IF algorithm, we calculate anomaly scores based on the average number of splits needed to isolate each observation, deleting the worst performers. We train the IF algorithm on each dataset and remove the approximately 5% worst performers of remaining transactions in each dataset, resulting in the removal of 25 investment transactions, 176 rental transactions, and 902 vacant transactions. Figures 7, 8, and 9 display usable floor area, building age, and MV constituents before and after outlier removal for each transaction type, where darker dots indicate higher anomaly scores.

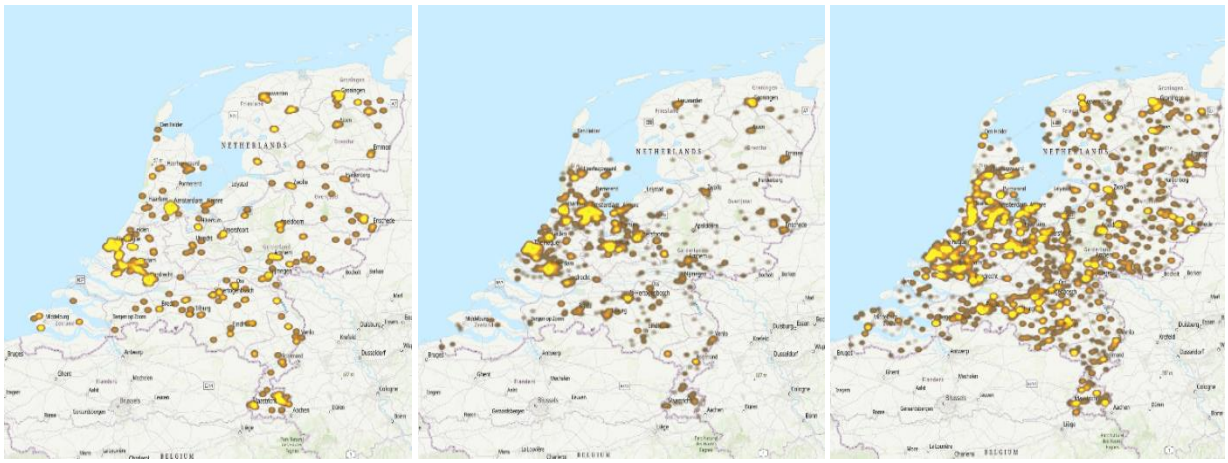

Figure 10. The locational distribution of investment transactions (left), rental transactions (middle) and vacant transactions (right).

After the cleaning procedure, 459 transactions remain in the investment transaction dataset, 3,337 transactions in the rental transaction dataset, and 17,132 transactions in the vacant transaction dataset. The locational distribution of the transactions after the cleaning procedure is displayed in Figure 10. All features used in the prediction of each MV constituent are displayed in the tables below. Specifically, Table 2 presents the summary statistics for investment transactions, Table 3 for rental transactions, and Table 4 for vacant transactions. For the investment transactions, we have aggregated provinces into four regions (north, south, east, and west) due to limited data availability. For the rental transactions, the MR is measured per square meter of usable floor area, per month, and for the vacant transactions, the VPV is measured per square meter of usable floor area. In line with Schulz & Wersing (2021), rooms are measured as the sum of living and bedrooms, as the classification of rooms is arbitrary. Building age is the difference between the transaction date and construction year, measured in years. The negative minimum building age value in the vacant transaction summary statistics is therefore due to the property being transacted before delivery in the future. Energy labels and Leefbaarometer values are categorically encoded, where higher values indicate better energy performance and Leefbaarometer values, respectively.

To capture locational attributes, we have included distance to eleven amenities and nine characteristics of the neighborhoods of transacted properties. Additionally, the population of the district, neighborhood, and municipality is included in the investment transaction dataset, and the population of the neighborhood and municipality is included in the rental and vacant transaction dataset. For investment transaction data, distance to amenities, transaction months, and regions are not reported for concision. Similarly, for rental and vacant transaction data, distance to amenities, the provincial code, type, sub type, and transaction months are not displayed for concision.

Table 2. Investment transactions descriptive statistics.

Feature	Mean	Median	Std. Dev.	Min	Max
Gross income multiplier	21.04	18.58	9.83	10.93	93.79
Average usable floor area per unit	82.85	79.60	24.67	19.75	189.00
Building age	56.29	46.00	39.10	3.00	369.00
Total residential units	10.49	7.00	9.24	1.00	45.00
Total theoretical rental income	85,187.07	62568.00	71,596.96	3,813.48	336,212.90
Total usable floor area	855.94	560.00	761.74	50	3650
Walk Score	77.36	79.00	16.30	16.00	99.00
Population					
District	15,797	11,485	15,225	465	109,805
Municipality	234,508	103,581	278,370	10,477	873,338
Neighborhood	3,989	2,845	3,973	175	28,870
Neighborhood level					
Average household size	1.95	1.90	0.35	1.30	3.30
Average property value	247,205	210,000	133,165	82,000	1,262,000
Housing stock	1,982	1,484	1,905	69	14,222
Percentage multifamily homes	0.55	0.59	0.32	0.00	1.00
Percentage owner occupied	0.50	0.51	0.20	0.05	0.98
Percentage ownership unknown	0.00	0.00	0.01	0.00	0.15
Percentage private rent	0.20	0.17	0.14	0.00	0.62
Percentage single family homes	0.45	0.41	0.32	0.00	1.00
Percentage social rent	0.30	0.28	0.18	0.00	0.92
Energy label		3.00		0.00	6.00
Leefbaarometer		5.00		1.00	8.00
Type single-family	0.27				
Type multifamily	0.73				
Transaction year 2019	0.87				
Transaction year 2020	0.11				
Transaction year 2021	0.02				

Note: Table 2 displays the descriptive statistics of the investment transaction data after cleaning and enriching, described above. Usable floor area is measured in square meters. Building age is measured as of the transaction date. Energy label and Leefbaarometer are ordinally encoded. Theoretical rental income and average property value are measured in euros. An object is either single-family, multifamily, or single/multifamily. Distance to amenities, transaction months, and regions are not reported for concision. The number of transactions is 459.

Table 3. Rental transactions descriptive statistics.

Feature	Mean	Median	Std. Dev.	Min	Max
Market rent	16.96	16.25	5.47	8.00	40.00
Building age	40.92	28.00	41.05	0.00	286.00
Days on the market	38.40	24.00	45.33	1.00	533.00
Number of rooms	3.15	3.00	1.04	1.00	10.00
Object volume	243.98	230.00	86.20	60.00	723.00
Total usable floor area	82.88	80.00	25.62	23.00	185.00
Walk Score	80.06	84.00	16.86	0.00	100.00
Population					
Neighborhood	3,977	3,055	3,390	50	28,870
Municipality	325,175	162,543	297,427	9,362	873,338
Neighborhood level					
Average household size	1.92	1.90	0.37	1.10	3.40
Average property value	309,691	283,300	137,479	97,000	2,065,000
Housing stock	2006	1541	1716	22	14,222
Percentage multifamily homes	0.64	0.72	0.30	0.00	1.00
Percentage owner occupied	0.46	0.45	0.20	0.00	0.97
Percentage ownership unknown	0.01	0.00	0.02	0.00	0.36
Percentage private rent	0.26	0.23	0.18	0.00	1.00
Percentage single family homes	0.36	0.28	0.30	0.00	1.00
Percentage social rent	0.27	0.24	0.19	0.00	1.00
Energy label		5.00		0.00	10.00
Leefbaarometer		5.00		1.00	8.00
Type single family	0.15				

Note: Table 3 displays the descriptive statistics of the rental transaction data after cleaning and enriching, described above. Market rent is measured per square meter of usable floor area, per month in euros. Usable floor area is measured in square meters. Building age is measured as of the transaction date. In line with Schulz & Wersing (2021), the number of rooms is measured as the sum of living and bedrooms. Energy label and Leefbaarometer are ordinally encoded. Object volume is measured in cubic meters. An object is either single-family (=1) or multifamily (=0). Distance to amenities, provincial code, type, sub type, and transaction month are not displayed for concision. The number of transactions is 3,337.

Table 4. Vacant transactions descriptive statistics.

Feature	Mean	Median	Std. Dev.	Min	Max
Vacant possession value	3,831.01	3611.11	1,278.05	1,421.05	10,000.00
Building age	47.59	44.00	30.86	-2.00	431.00
Days on the market	28.05	22.00	23.37	1.00	352.00
Number of rooms	4.50	5.00	1.33	1.00	13.00
Object volume	401.21	387.00	142.82	74.00	1613.00
Total usable floor area	111.51	110.00	33.95	27.00	250.00
Walk Score	67.79	71.00	19.75	0.00	100.00
Population					
Neighborhood	3,594	2,720	3,076	35	28,870
Municipality	164,225	89,999	204,533	931	873,338
Neighborhood level					
Average household size	2.17	2.20	0.36	1.10	3.70
Average property value	273,462	254,000	104,599	77,000	1,624,000
Housing stock	1,656	1,241	1,467	20	14,222
Percentage multifamily homes	0.36	0.27	0.29	0.00	1.00
Percentage owner occupied	0.60	0.62	0.20	0.00	1.00
Percentage ownership unknown	0.00	0.00	0.01	0.00	0.36
Percentage private rent	0.13	0.09	0.12	0.00	1.00
Percentage single family homes	0.64	0.73	0.29	0.00	1.00
Percentage social rent	0.27	0.24	0.18	0.00	0.97
Energy label		4.00		0.00	11.00
Leefbaarometer		6.00		1.00	8.00
Purchaser's costs	0.99				
Type single family	0.70				

Note: Table 4 displays the descriptive statistics of the vacant transaction data after cleaning and enriching, described above. Vacant possession value is measured per square meter of usable floor area, in euros. Usable floor area is measured in square meters. Building age is measured as of the transaction date. In line with Schulz & Wersing (2021), the number of rooms is measured as the sum of living and bedrooms. Energy label and Leefbaarometer are ordinally encoded. Object volume is measured in cubic meters. Average property values are also measured in euros. An object is either single-family (=1) or multifamily (=0). Distance to amenities, provincial code, type, sub type, and transaction month are not displayed for concision. The number of transactions is 17,132.

4.4 Variable transformation

In line with common practice in the AVM literature, we log-transform the predicted variables VPV, MR, and GIM to improve the overall performance of the proposed methodologies. After prediction, we

retransform the predicted variables using the smearing adjustment factor introduced by Duan (1983), in line with Schulz & Wersing (2021) and Steurer et al. (2021), to deal with retransformation bias:

$$E\left(\widehat{VPV}, \widehat{MR}, \widehat{GIM}\right) = \hat{\varphi} \exp\left[ln\left(\widehat{VPV}, \widehat{MR}, \widehat{GIM}\right)\right] \quad (\mathbf{8})$$

$$where\ \hat{\varphi} = \frac{1}{n}\sum_{i=1}^{n} exp\ (\varepsilon_i^{VPV,MR,GIM})$$

where the expected values of VPV, MR, and GIM are calculated by taking the exponent of the log-transformed values and adjusted using an adjustment factor $\hat{\varphi}$. $\hat{\varphi}$ is calculated by taking the average of the error term exponent values for VPV, MR, and GIM estimates, denoted as ε.

Finally, in line with Steurer et al. (2021), all input features have been scaled and centered to lie between 0 and 1 during the training and testing of the methodologies. Feature scaling is common in ML applications, as to not distinguish between features of lower magnitude (e.g., number of bedrooms) and features of higher magnitude (e.g., building age). Although tree-based methodologies are insensitive to feature scaling, applying it is essential to accurately compare all proposed methodologies. Without scaling these features, the latter type of features weighs more decisively in determining VPV, MR, and GIM than the former in non-tree-based methodologies. Additionally, the training time of the proposed methodologies decreases by using scaling.

5. Design

As described earlier, we will train and compare the following methodologies: HPM, RF, GB, XGB, a hybrid methodology comprising HPM with an XGB estimation procedure, and XGBMC. These methodologies will be optimized to predict the MV constituents of the SPCM to MV estimation: GIM, MR, and VPV[16]. First, we will discuss the hyperparameter optimization procedure, discussing what methods are adopted in tuning the methodologies. In doing so, we will also provide an overview of the most important tunable hyperparameters in each methodology. Finally, we elaborate on the MCs in the XGBMC methodology.

5.1 Hyperparameter optimization

Each considered methodology requires optimization of hyperparameters. Optimizing these hyperparameters requires a comparison of various model specifications, as the most optimal combination is highly dependent on the obtained data. Four approaches are feasible for optimizing hyperparameters, (1) manual, (2) grid search, (3) randomized search, and (4) Bayesian optimization. Manual optimization involves manually adjusting hyperparameters after each training and performance

[16] Python code used in developing each methodology are obtainable from the author.

evaluation iteration. This method allows the exploration of specific hyperparameter regions that seem promising for the problem at hand. Another commonly adopted hyperparameter optimization method is grid search. Grid search involves defining a set of values for each hyperparameter requiring tuning. The set of training trials is then determined based on assembling every possible combination of hyperparameter values. In comparison with manual search, this approach does not require manual adjustment of the hyperparameters after each iteration. Grid search does, however, suffer from the curse of dimensionality, as the number of combinations and thus training iterations increases exponentially with the number of hyperparameter values (Bergstra & Bengio, 2012). Comparable to grid search, randomized search, the third method, draws hyperparameter combinations from a search space. In contrast to grid search, however, randomized search draws independent, random draws, thus not suffering from the dimensionality issue described above. This contrasting feature allows randomized search to be more efficient in high-dimensional search spaces, compared to grid search (Bergstra & Bengio, 2012). Both these methods work in optimizing hyperparameters but are relatively inefficient as new hyperparameters are not chosen on the basis of the results of previous hyperparameter settings. In other words, grid and random search take independent draws from the search space and therefore do not learn from past iterations.

Shahriari et al. (2015) describe the importance of methodologies learning from past iterations when facing a greater degree of optimization complexity. Learning from past iterations is inherent to Bayesian optimization, the fourth method. Bayesian optimization involves adopting a surrogate probability model, next to the objective function. Bayesian optimization centers on this probability model $p(score|configuration)$ that is updated through historical $(score|configuration)$ pairs (Bergstra et al., 2013). The most common surrogate probability model used in Bayesian optimization is the Sequential Model-Based Global Optimization (SMBO) algorithm with the Tree Parzen Estimator (TPE) modeling strategy. Based on Bayes' theorem, SMBO with the TPE modeling strategy can be defined as (Bergstra, 2011: 4):

$$p(VPV, MR, GIM|x) = \frac{p(x|VPV, MR, GIM) * p(VPV, MR, GIM)}{p(x)} \quad (\mathbf{9})$$

$$where\ p(x|VPV, MR, GIM) = \begin{cases} l(x) & if\ VPV, MR, GIM < (VPV, MR, GIM)^* \\ g(x) & if\ VPV, MR, GIM \geq (VPV, MR, GIM)^* \end{cases}$$

Where $p(VPV, MR, GIM|x)$ denotes the probability of estimate VPV, MR, or GIM, given a candidate of hyperparameters *x*. In contrast to Gaussian Processes (GP, another modeling strategy; see Bergstra et al., 2011), TPE involves not estimating $p(VPV, MR, GIM|x)$ directly, but through $p(x|VPV, MR, GIM)$ and $p(VPV, MR, GIM)$. $p(x|VPV, MR, GIM)$, in turn, is defined by the densities $l(x)$ and $g(x)$. Specifically, $l(x)$ is the density formed using the hyperparameter observations x where the

corresponding loss was less than $(VPV, MR, GIM)^*$, the threshold value of the objective function. $(VPV, MR, GIM)^*$ is determined in the TPE algorithm as a quantile γ of the observed values VPV, MR, and GIM, such that $p(VPV, MR, GIM < (VPV, MR, GIM)^*) = \gamma$. All hyperparameter observations x where the value of the objective function is greater than the threshold are contained in the density $g(x)$. Optimization using the TPE is done through a selection function or criterium. In the case of TPE, the most common selection criterion is Expected Improvement (EI), defined as (Bergstra, 2011: 4):

$$EI_{(VPV,MR,GIM)^*}(x) = \frac{\gamma(VPV, MR, GIM)^* l(x) - l(x) \int_{-\infty}^{(VPV,MR,GIM)^*} p(VPV, MR, GIM)\partial(VPV, MR, GIM)}{\gamma l(x) + (1-\gamma) g(x)}$$

$$\propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1} \quad (\mathbf{10})$$

Where the last expression denotes that hyperparameters sets *x* with high probability under *l(x)* and low probability under *g(x)* are preferred. As described in Bergstra et al. (2011), due to the tree-structured form of l and g shown in Equation (10) above, the TPE algorithm can draw many hyperparameter candidates *x* based on *l* and evaluate them according to *g(x)*/*l(x)*. Each iteration results in the candidate of hyperparameters x^* with the greatest EI. Bergstra et al. (2013) argue that adopting the surrogate probability model yields significant gains in overall optimization efficiency, as only the candidate of hyperparameters yielding the highest EI is then evaluated in the original objective function.

We apply the SMBO algorithm with the TPE modeling strategy to tune the hyperparameters of each methodology. Specifically, we obtain the aforementioned algorithm and modeling strategy using the *Hyperopt* library[17]. The Hyperopt library allows CV techniques, all ML methodologies described in this paper, and the TPE modeling strategy. These aspects of the Hyperopt library allow a relatively straightforward implementation of Bayesian optimization. The library does, however, require the specification of the distribution function for each hyperparameter. Example distribution functions include uniform and normal distributions. Generally, it is equitable to assume that for numerically valued hyperparameters, each value has equal probability to be most optimal for regression problems. This means that for numerically valued hyperparameters, the uniform probability distribution, or variations thereof, is most applicable.

We list the most important hyperparameters for each ML methodology tunable using Bayesian optimization in Table 4. As mentioned earlier, optimizing hyperparameters using Bayesian optimization requires setting probability distributions for each hyperparameter. These probability distributions aid the TPE modeling strategy in identifying promising hyperparameter values by using EI, elaborated above (Bergstra et al., 2011). For continuous hyperparameters, uniform distributions are most commonly adopted in the TPE, which we also adopt here. For categorical hyperparameters, we adopt

[17] Documentation on the Hyperopt library can be found here: https://github.com/hyperopt/hyperopt.

the choice distribution, where the algorithm decides on the hyperparameter value for each iteration. Finally, when hyperparameter values involve various data types (e.g., float and integer), we nest uniform distributions in a choice distribution.

Table 4. Hyperparameters of machine learning methodologies adopted in this study.

Methodology	Hyperparameter	Definition
Random forest	Bootstrap	Whether or not to apply bootstrap aggregation (true/false). Setting this hyperparameter to false means models train without replacing the sample through each iteration.
	Maximum number of features	The number of features to consider when splitting individual decision trees at each node.
	Maximum tree depth	The maximum depth of individual decision trees; a higher maximum depth is equivalent to the expansion of more nodes.
	Minimum samples leaf	The minimum number of training cases (i.e., transactions) required to be at a leaf (i.e., terminal) node.
	Minimum samples split	The minimum number of training cases (i.e., transactions) required to split an internal node.
	Number of estimators	The number of decision trees in the aggregate forest.
Gradient boosting	Learning rate	The learning rate determines the contribution of each tree in the aggregate model. The number of estimators and the learning rate faces a tradeoff; a higher learning rate requires a lower number of estimators
	Maximum number of features	The number of features to consider when splitting individual tree boosters at each node.
	Maximum tree depth	The maximum depth of individual tree boosters; a higher maximum depth is equivalent to the expansion of more nodes.
	Minimum samples leaf	The minimum number of training cases (i.e., transactions) required to end up in a leaf (i.e., terminal) node.
	Minimum samples split	The minimum number of training cases (i.e., transactions) required to split an internal node.
	Number of estimators	The number of boosting stages.
	Subsample	The fraction of the total training data used for training individual tree boosters.
Extreme gradient boosting	Alpha (α)	Median-based regularization term. Higher values of alpha decrease sensitivity to individual weight observations.
	Column sample by tree	Similar to the maximum number of features in the RF and GB methodologies; the number of features to consider when splitting individual tree boosters at each node.

Gamma (γ)	Regularization term. The minimum loss reduction required to further split a leaf node in an individual booster tree. Higher gamma values decrease the chance of overfitting the training data.
Lambda (λ)	Mean-based regularization term. Higher values of lambda decrease sensitivity to individual weight observations.
Learning rate	The learning rate determines the contribution of each tree in the aggregate model. The number of estimators and the learning rate faces a tradeoff; a higher learning rate requires a lower number of estimators.
Maximum tree depth	The maximum depth of individual tree boosters; a higher maximum depth is equivalent to the expansion of more nodes.
Minimum child weight	Similar to minimum samples split in the RF and GB methodologies; the minimum number of training cases (i.e., transactions) needed to split an internal node.
Number of estimators	The number of boosting stages.
Subsample	The fraction of the total training data used for training individual tree boosters.

Note: Table 4 displays the hyperparameters of the machine learning methodologies optimized using Bayesian optimization, with associated definitions. We use the Hyperopt, Scikit-Learn, and XGBoost Python libraries to implement the machine learning methodologies with Bayesian optimization.

5.2 Monotonic constraints

We develop MCs for the XGBMC methodology using the expertise of appraisers. As data on the three types of transactions differ, monotonically constraint features can differ per MV constituent of the SPCM. Additionally, MR and VPV are measured per square meter of usable floor area as described in Section 4.1. We will analyze the addition of MCs in the XGB methodology by comparing the proposed methodology to a similar, non-monotonically constraint XGB methodology. We develop MCs for ten features based on arguments provided by a valuation firm in the Netherlands (see Appendix 1: Envalue valuation firm).

First, for MR and VPV per square meter of usable floor area, object volume and the number of rooms are positively constraint as these increase the attractiveness of each square meter of usable floor area. Specifically, properties being equal in terms of usable floor area and other aspects, but differing sizes of object volume and number of rooms will yield different MR and VPV per square meter of usable floor area. Higher object volume and more rooms, in this example, will lead to higher MR and VPV per square meter of usable floor area. Second, as MR and VPV are measured per square meter of usable floor area, total usable floor area is negatively constraint. Properties equal in most aspects, but differing sizes of total usable floor area, will generally yield different MR and VPV per square meter of usable floor area. Higher total usable floor areas result, all else being equal, in lower expected values of MR and VPV per square meter of usable floor area due to the lower influence of land value on total property value. Third, distance to amenities are negatively constraint for all three MV constituents. Specifically,

a larger distance to the eleven amenities is equivalent to a less attractive location, resulting in lower expected values of GIM, MR, and VPV. In contrast, the Walk Score is positively constraint for all three MV constituents. The Walk Score measures the walkability of properties to nearby amenities. As higher Walk Scores indicate a more attractive location, we expect, all else being equal, higher values of GIM, MR, and VPV. The energy label is positively constraint for all MV constituents. Higher energy labels indicate better energy performance, increasing the attractiveness of a property and hence, all else being equal, resulting in higher values of GIM, MR, and VPV.

On the regional level, the Leefbaarometer measures the attractiveness of a district. More attractive districts yield higher values of GIM, MR, and VPV and are therefore positively constraint for all three MV constituents. Similarly, the population of a municipality is an indicator of the size of a municipality. Larger municipalities are more attractive and are therefore expected to yield, all else being equal, higher values of GIM, MR, and VPV and are therefore positively constraint for all three MV constituents. For MR and VPV, specifically, days on the market are negatively constraint. Days on the market measures the attractiveness of a property to the market. Generally, more days on the market indicate lower attractiveness of the property and hence, all else being equal, a lower expected MR and VPV per square meter of usable floor area. All the MCs per feature and MV constituent of the SPCM are displayed in Table 5.

Table 5. Monotonic constraints in the extreme gradient boosting with monotonic constraints methodology for the gross income multiplier, and market rent and vacant possession value measured per square meter of usable floor area.

Feature	**Monotonic constraint (GIM)**	**Monotonic constraint (MR)**	**Monotonic constraint (VPV)**
Days on the market	-	Negative	Negative
Distance to amenities	Negative	Negative	Negative
Energy label	Positive	Positive	Positive
Leefbaarometer	Positive	Positive	Positive
Municipality population	Positive	Positive	Positive
Number of rooms	-	Positive	Positive
Object volume	-	Positive	Positive
Total residential units	Positive	-	-
Total usable floor area	-	Negative	Negative
Walk Score	Positive	Positive	Positive

Note: Table 5 displays monotonic constraints in the extreme gradient boosting with monotonic constraints methodology. Monotonic constraints are developed using appraiser expertise at a valuation firm (see Appendix 1: Envalue valuation firm) and provided for each market value constituent: gross income multiplier, market rent, and vacant possession value. Distance to amenities entails distance to all eleven amenities considered in this paper. We use the XGBoost Python library to implement monotonic constraints in the extreme gradient boosting methodology. We analyze the addition of monotonic constraints in the extreme gradient boosting methodology by

comparing the proposed methodology to a similar, non-monotonically constraint extreme gradient boosting methodology.

5.3 Performance measurement

To accurately compare methodologies in our case study, we adopt various performance metrics. In line with Steurer et al. (2021), we use seven metrics to measure accuracy. Each of these metrics satisfies the symmetry condition, meaning interchanging the actual and predicted values do not change the absolute values of each metric, nor its interpretation. The metrics used are of the seven classes average bias, absolute difference, absolute ratio, squared difference, squared ratio, percentage ratio, and quantile ratios (Steurer et al., 2021). Additionally, we describe how we measure the reliability and explainability of each methodology.

Average bias. In contrast to other metrics of accuracy and reliability, average bias metrics can be both negative and positive. To measure average bias, we adopt the log median prediction error (LMDPE). The closer this metric is to zero, the better the performance of a methodology:

$$LMDPE = med\left[\ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)\right] \qquad \forall \, \mathrm{i} = 1 \ldots \mathrm{n} \quad (\mathbf{11})$$

Where n represents the total number of observations, $MV^{VPV,MR,GIM}$ the actual and $\widehat{MV}^{VPV,MR,GIM}$ the predicted values of the MV constituents in the SPCM of observation i. LMDPE is median-based, meaning the measure is more robust to large prediction errors, favoring methodologies that have low average bias. Symmetry is imposed in the LMDPE metric through $ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)$.

Absolute difference. Absolute difference ratios measure the average absolute error of the predictions by a methodology. In comparison to squared difference metrics (discussed below), absolute difference metrics limit the influence of large prediction errors. To measure the absolute difference, we adopt the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|MV_i^{VPV,MR,GIM} - \widehat{MV}_i^{VPV,MR,GIM}\right| \quad (\mathbf{12})$$

Where n represents the total number of observations, $|MV^{VPV,MR,GIM}|$ the absolute actual and $\left|\widehat{MV}^{VPV,MR,GIM}\right|$ the absolute predicted values of the MV constituents in the SPCM of observation i. MAE is a common measure adopted in the forecasting literature (see, e.g., McCluskey et al., 2013; Zurada et al., 2011), satisfying the symmetry condition through $\left|MV_i^{VPV,MR,GIM} - \widehat{MV}_i^{VPV,MR,GIM}\right|$.

Absolute ratio. In comparison to absolute difference metrics, absolute ratio metrics measure the prediction error as ratios, providing for a more relative performance metric. As with the absolute

difference metric, absolute ratio metrics are less sensitive to outliers. We measure the absolute ratio using a symmetric adaptation to a widely adopted metric in the literature, the mean absolute prediction error (MAPE). Specifically, we adopt the max-min mean absolute prediction error (mmMAPE):

$$mmMAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\max\left(MV_i^{VPV,MR,GIM}, \widehat{MV}_i^{VPV,MR,GIM}\right)}{\min\left(MV_i^{VPV,MR,GIM}, \widehat{MV}_i^{VPV,MR,GIM}\right)} - 1\right) \quad (\mathbf{13})$$

Where n represents the total number of observations, $MV^{VPV,MR,GIM}$ the actual and $\widehat{MV}^{VPV,MR,GIM}$ the predicted values of the MV constituents in the SPCM of observation i. mmMAPE, compared to MAPE, imposes symmetry by replacing $\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}} - 1$ with the max-min operation shown in Equation 13.

Squared difference. Squared difference ratios heavily penalize large prediction errors. This metric is especially important in analyzing the attractiveness of the proposed methodologies to valuation practice, as (individual) large prediction errors are unfavorable. To measure the squared difference, we adopt the RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(MV_i^{VPV,MR,GIM} - \widehat{MV}_i^{VPV,MR,GIM}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{e}_i^2\right)^2} \quad (\mathbf{14})$$

Where n represents the total number of observations, $MV^{VPV,MR,GIM}$ the actual and $\widehat{MV}^{VPV,MR,GIM}$ the predicted values of the MV constituents in the SPCM and ê the error term of observation i.

Squared ratio. Squared ratios, in line with squared difference metrics, heavily penalize large prediction errors. Steurer et al. (2021) argue that the symmetry condition is even more important for squared ratio metrics compared to absolute ratio metrics. Asymmetric absolute ratio metrics (e.g., MAPE), weigh prediction errors in the left and right tail of the error distribution asymmetrically. Squaring these errors, they argue, amplifies this asymmetry error. Within this case study, we adopt the log root mean squared error (LRMSE):

$$LRSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)\right]^2} \quad (\mathbf{15})$$

Where n represents the total number of observations, $MV^{VPV,MR,GIM}$ the actual and $\widehat{MV}^{VPV,MR,GIM}$ the predicted values of the MV constituents in the SPCM of observation i. Symmetry is imposed in the LRSME metric, similar to the LMDPE metric, through $ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)$.

Percentage ratio. Percentage ratios count the percentage of prediction error ratios that are outside a pre-specified limit. Practitioners frequently use these ratios when assessing valuation accuracy. For instance, if 20 is the result of a percentage ratio with a pre-specified limit of 10 percent, that means the error rate is above 10 percent in 20 percent of the cases. Percentage ratios, therefore, incorporate the uncertainty associated with value estimates described in Section 2.1. To measure the percentage ratio, we adopt the max-min percentage error range (mmPER):

$$mmPER = 100 \left| \frac{\max\left(MV_i^{VPV,MR,GIM}, \widehat{MV}_i^{VPV,MR,GIM}\right)}{\min\left(MV_i^{VPV,MR,GIM}, \widehat{MV}_i^{VPV,MR,GIM}\right)} - 1 \right| > x \qquad \forall \, \mathrm{i} = 1 \ldots \mathrm{n} \quad (\mathbf{16})$$

Where n represents the total number of observations, $MV^{VPV,MR,GIM}$ the actual and $\widehat{MV}^{VPV,MR,GIM}$ the predicted values of the MV constituents in the SPCM of observation i, and x determines the cut-off percentage error rate. Manual valuations are expected to not deviate more than 10% from the actual MV (Crosby, 2000), meaning we will use that as the cut-off percentage error rate for the mmPER ratio.

Quantile ratio. AVMs for real estate valuation require insight into the dispersion of the prediction error distribution, which quantile ratios measure. Quantile ratios are more robust metrics of dispersion in contrast to variance-based metrics such as squared difference ratios, providing a valuable additional performance metric (Steurer et al., 2021). To measure the quantile ratio, we adopt the inter-quartile range in ratios (IQRat):

$$IQRat = ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)_{75} - ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)_{25} \qquad \forall \, \mathrm{i} = 1 \ldots \mathrm{n} \quad (\mathbf{17})$$

Where n represents the total number of observations, $ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)_{75}$ the 75$^{\text{th}}$ percentile and $ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)_{25}$ the 25$^{\text{th}}$ percentile of the natural logarithm of the actual values $MV^{VPV,MR,GIM}$ and predicted values $\widehat{MV}^{VPV,MR,GIM}$ of the MV constituents in the SPCM of observation i. Symmetry is imposed in the IQRat metric, similar to LMDPE, through $ln\left(\frac{MV_i^{VPV,MR,GIM}}{\widehat{MV}_i^{VPV,MR,GIM}}\right)$.

Reliability. We measure reliability using a novel method for estimating the error distribution across folds and repeats. Specifically, we calculate the standard error of each accuracy performance metric described above across all 30 CV iterations (elaborated in Section 3.1). Based on the standard errors, we can construct error distributions for each accuracy performance metric. In doing so, we align with comparable efforts to measure reliability besides standard prediction intervals, such as conformal prediction intervals in Schulz & Wersing (2021), the Diebold-Mariano test statistic in Hinrichs et al. (2021), and bootstrap resampling in Krause et al. (2020).

Explainability. We measure explainability using insights from interpretable ML, where explainability is understood as "given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand" (Arrieta et al., 2020: 85). In our case study, the audience consists of both appraisers and their respective clients. We, therefore, measure explainability of the proposed methodologies by considering how well appraisers can understand and explain these methodologies to clients. Based on an overview by Arrieta et al. (2020), we focus on the explainability of each methodology in terms of *algorithmic transparency*, *decomposability*, and *simulatability*. Algorithmic transparency refers to the ability of model users (e.g., appraisers) to understand the process of producing any given output from input data. Decomposability is more stringent in that methodologies should be understandable without requiring additional tools. Finally, simulatability refers to the readability of a model, hence being a measure of complexity. Each of these three classes contains its predecessors; a simulatable methodology is therefore also decomposable and algorithmically transparent. We rank each methodology in terms of these measures from an appraiser's perspective.

6. Results

6.1 Gross income multiplier

Table 8 provides the accuracy and reliability per methodology and performance metric for the investment transaction data, where the most accurate models according to each metric are highlighted in bold. The table shows a general increase in performance across all metrics for more flexible methodologies. Using our novel method for measuring reliability, we find all methodologies to have high reliability, with low standard errors for each performance metric[18]. In terms of accuracy, we observe the largest relative differences in the central tendency (average bias) metric (LMDPE). Specifically, central tendency is lowest for the GB methodology, with the XGB methodology performing 33.8% worse and the XGBMC methodology performing 58.8% worse than the GB methodology. GB is therefore most symmetric in bias and robust to outliers according to the LMDPE metric. Differences between the GB, XGB, and XGBMC methodologies are relatively small for the remaining metrics. Accuracy in terms of dispersion (MAE) is the highest for the XGB methodology, with both the GB and XGBMC methodologies performing 0.7% worse than the XGB methodology. For the absolute ratio (mmMAPE), the XGB methodology also performs better in terms of both accuracy and reliability compared to other methodologies. Specifically, the GB methodology performs 0.6% worse and the XGBMC methodology performs 1.2% worse than the XGB methodology on this metric. Differences

[18] Low standard errors in cross-validation indicate minor differences between hold-out sample performance across folds. Minor differences in hold-out sample performance can be due to various reasons; one of these reasons is the use of balanced datasets through our data cleaning procedures. For an elaboration on data balance and its implications, see Gelman & Hill (2007: 199–231).

between these three methodologies are small compared to the hybrid methodology that performs 62.4% worse than the XGB methodology on this metric.

The more established squared difference metric (RMSE), penalizing large errors more heavily, favors the XGB methodology in terms of accuracy, but differences are relatively small to the XGBMC methodology that performs 1.2% worse on this metric. Similar patterns hold for the squared ratio metric (LRMSE), where XGB is deemed more accurate with small differences compared to the XGBMC methodology. In terms of the percentage ratio (mmPER), 53.7% of predictions are outside ± 10% of the GIM for the XGB methodology, while XGBMC performs slightly worse with 54.1% of predictions falling outside the pre-specified deviation limit. Finally, in contrast to other metrics, the dispersion of the prediction error (IQRat) is the lowest for the XGBMC methodology, with a relatively small 0.5% difference from the XGB methodology on this metric.

Table 8. Gross income multiplier - Accuracy and reliability per methodology.

Methodologies	**LMDPE**	**MAE**	**mmMAPE**	**LRMSE**	**RMSE**	**mmPER**	**IQRat**
HPM	-0.0327 (0.0062)	4.364 (0.173)	0.216 (0.009)	0.240 (0.008)	7.435 (0.008)	0.628 (0.014)	0.264 (0.014)
RF	-0.0504 (0.0069)	3.935 (0.218)	0.195 (0.009)	0.224 (0.009)	7.068 (0.009)	0.571 (0.022)	0.207 (0.022)
GB	**-0.0260** **(0.0052)**	3.477 (0.133)	0.166 (0.005)	0.196 (0.006)	6.367 (0.006)	0.553 (0.013)	0.210 (0.013)
XGB	-0.0348 (0.0052)	**3.454** **(0.180)**	**0.165** **(0.007)**	**0.193** **(0.007)**	**6.132** **(0.007)**	**0.537** **(0.019)**	0.197 (0.019)
XGBMC	- 0.0413 (0.0053)	3.477 (0.179)	0.167 (0.007)	0.195 (0.007)	6.203 (0.007)	0.541 (0.017)	**0.196** **(0.017)**
Hybrid methodology	-0.0444 (0.0082)	5.262 (0.202)	0.268 (0.010)	0.287 (0.010)	9.119 (0.009)	0.678 (0.012)	0.310 (0.012)

Note: The dependent variable is the log-transformed gross income multiplier, adjusted using Duan's (1983) adjustment factor before calculating performance metrics. Standard errors are calculated using permutation tests and are displayed in parentheses. The most accurate methodologies according to each metric are highlighted in bold.

Explainability differs for each methodology. Empirical methodologies, like the baseline HPM proposed in this paper, use statistical inference to determine how the output is related to input. Additionally, the uncertainty of the estimates can be determined using probability theory. Predictions made by ML methodologies, on the other hand, are not statistically interpretable due to the many parameters that relate the output to input. To approximate the statistical inference techniques used in econometric methodologies, we determine feature importance using permutation tests, a feature importance

technique[19]. Specifically, for the hold-out sample in the first iteration of the CV procedure described in Section 3.1, we calculate the relative importance of each feature in three consecutive steps. First, we randomly shuffle data points in one input feature while keeping the other input features constant. Second, we calculate new predictions of the MV constituents with the shuffled data points for one of the input features. Third, we calculate feature importance by computing the change in the forecasting performance compared to the unshuffled data predictions.

The ten most contributing features in the case of the GIM are displayed in Figures 11 and 12 below[20]. As can be seen in the figures, total theoretical rental income, municipality population, average property value on the neighborhood level, and whether a house is owner-occupied rank first in one or two of the methodologies. Specifically, total theoretical rental income ranks highest for the HPM and GB methodologies, the municipality population for the RF methodology, average property value on the neighborhood level for the XGB and XGBMC methodologies, and owner-occupied housing status for the hybrid methodology[21]. Other high-ranking features are housing stock on the neighborhood level, whether a house is socially rented, total usable floor area, the transaction month, distance to the nearest highway, and the Leefbaarometer.

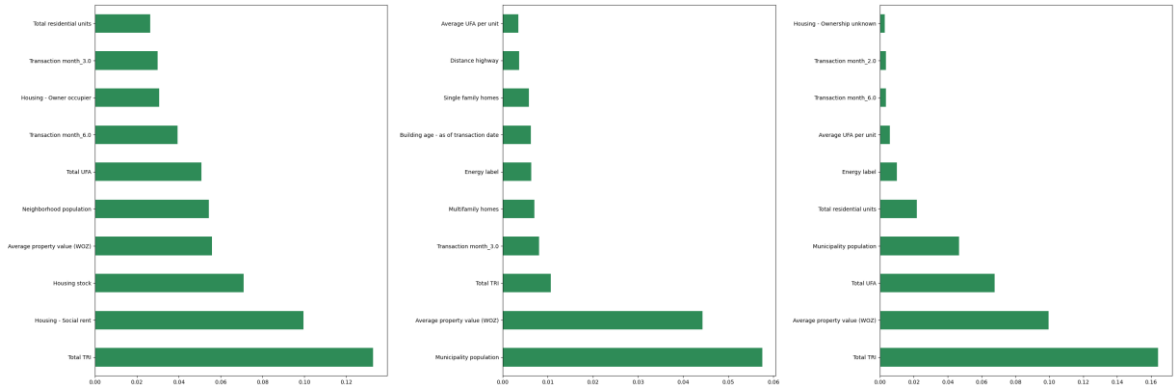

Figure 11. Hedonic pricing model (left), random forest (middle), and gradient boosting (right) feature importance in predicting the gross income multiplier. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

[19] Various feature importance techniques exist that seek to determine the relative importance of features in determining the predictions made by machine learning methodologies. See Saarela & Jauhiainen (2021) for derivations and comparison of these techniques.
[20] For the gross income multiplier and the other market value constituents, scales of the figures differ; feature importance can therefore not be compared between methodologies.
[21] We calculate feature importance for the hybrid methodology during the extreme gradient boosting estimation procedure.

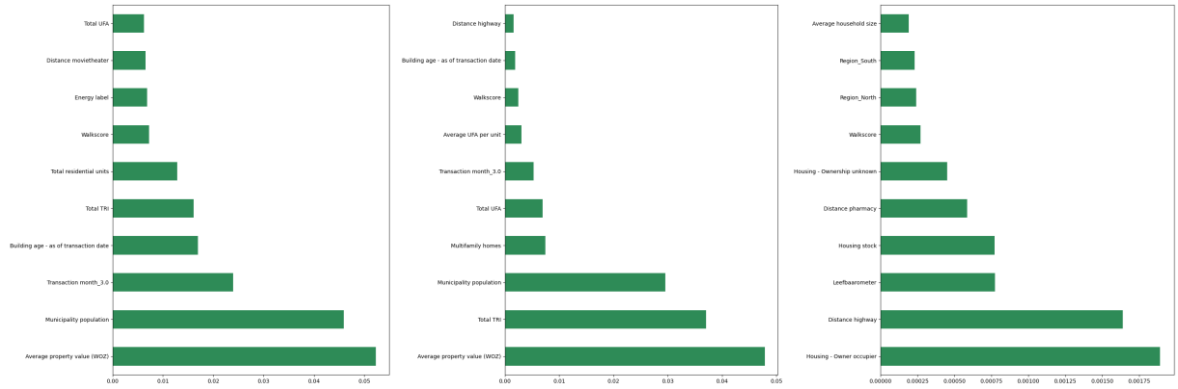

Figure 12. Extreme gradient boosting (left), extreme gradient boosting with monotonic constraints (middle), and hybrid methodology (right) feature importance in predicting the gross income multiplier. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

6.2 Market rent

Table 9 provides the accuracy and reliability per methodology and performance metric for the rental transaction data, where the most accurate models according to each metric are highlighted in bold. Similar to the investment transaction data, we observe a general increase in performance across all metrics for more flexible methodologies. All methodologies show high reliability, with low standard errors for each performance metric. In terms of specific methodologies and in contrast to the investment transaction data, the GB methodology performs best across all metrics. In line with the investment transaction data, we observe the largest relative differences in the central tendency (average bias) metric. Specifically, central tendency is lowest for the GB methodology, with the XGB methodology performing 4.3% worse and the XGBMC methodology performing 25.6% worse than the GB methodology. GB is therefore most symmetric in bias and robust to outliers according to the LMDPE metric. In contrast to the investment transaction data, however, other metrics also favor the GB methodology with small differences with the XGB and XGBMC methodologies. Specifically, accuracy in terms of dispersion (MAE) is the highest for the GB methodology, with the XGB methodology performing 1.4% worse and the XGBMC methodology performing 2.5% worse than the GB methodology. In terms of the absolute ratio (mmMAPE), the GB methodology also performs better in terms of accuracy compared to other methodologies. Both the XGB and XGBMC methodologies perform 0.6% worse than the GB methodology on this metric. In line with the investment transaction data, differences between these three methodologies are small compared to the hybrid methodology that performs 61.2% worse than the GB methodology on this metric.

In terms of the squared ratio metric (LRMSE) that penalizes large errors more heavily, the GB methodology is again favored over other methodologies. Differences are relatively small, however, to the XGB methodology that performs 2.0% worse and the XGBMC methodology that performs 3.3% worse on this metric. Similar patterns hold for the squared difference metric (RMSE), where GB is deemed more accurate with small differences compared to the XGB and XGBMC methodologies. In

terms of the percentage ratio (mmPER), 48.1% of predictions are outside ± 10% of the MR for the GB methodology, 48.8% for the XGB methodology, and 49.1% of predictions falling outside the pre-specified deviation limit for the XGBMC methodology. Finally, the dispersion of the prediction error (IQRat) is lowest for the GB methodology, with a relative small difference of 0.5% from the XGB and XGBMC methodologies on this metric.

Table 9. Market rent - Accuracy and reliability per methodology.

Methodologies	**LMDPE**	**MAE**	**mmMAPE**	**LRMSE**	**RMSE**	**mmPER**	**IQRat**
HPM	-0.0151	2.346	0.158	0.181	3.190	0.558	0.219
	(0.0014)	(0.023)	(0.002)	(0.002)	(0.002)	(0.006)	(0.006)
RF	-0.0126	2.134	0.142	0.165	2.889	0.532	0.206
	(0.0010)	(0.033)	(0.002)	(0.002)	(0.002)	(0.007)	(0.007)
GB	**-0.0117**	**1.961**	**0.129**	**0.153**	**2.691**	**0.481**	**0.183**
	(0.0011)	**(0.015)**	**(0.001)**	**(0.001)**	**(0.001)**	**(0.005)**	**(0.005)**
XGB	-0.0122	1.988	0.132	0.156	2.730	0.488	0.184
	(0.0009)	(0.015)	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)
XGBMC	-0.0147	2.010	0.133	0.158	2.759	0.491	0.184
	(0.001)	(0.016)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
Hybrid methodology	-0.0339	3.193	0.222	0.242	4.306	0.665	0.295
	(0.002)	(0.039)	(0.003)	(0.002)	(0.002)	(0.006)	(0.006)

Note: The dependent variable is the log-transformed market rent, adjusted using Duan's (1983) adjustment factor before calculating performance metrics. Standard errors are calculated using permutation tests and are displayed in parentheses. The most accurate methodologies according to each metric are highlighted in bold.

The relative feature importance for the ten most contributing features in determining MR is calculated using permutation tests as described in Section 6.1 and displayed in Figures 13 and 14 below. As can be seen in the figures, whether a house is of type single family, total usable floor area, and days on the market rank highest in one or several methodologies. Specifically, whether a house is of type single family ranks highest for the HPM methodology, total usable floor area ranks highest for the RF, GB, XGB, and XGBMC methodologies, and days on the market ranks highest for the hybrid methodology. Other high-ranking features include whether a house is of subtype corner house and terraced house, municipality population, average property value on the neighborhood level, building age, and the energy label.

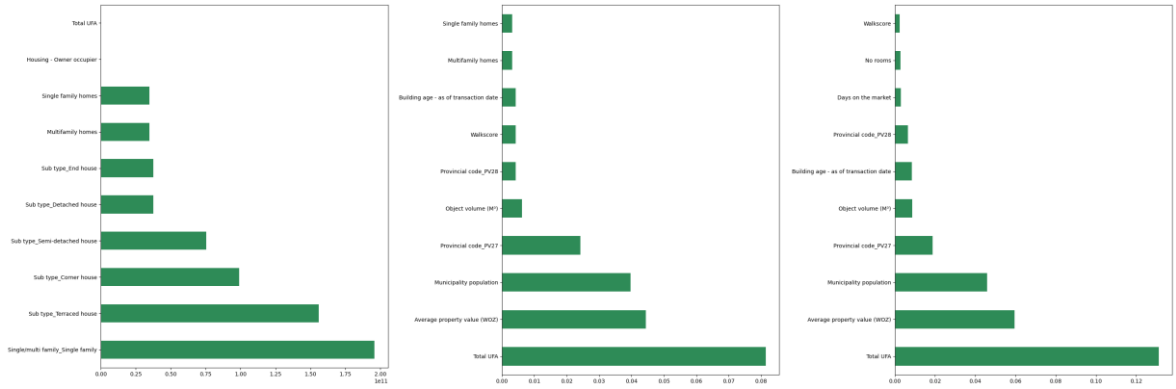

Figure 13. Hedonic pricing model (left), random forest (middle), and gradient boosting (right) feature importance in predicting the market rent. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

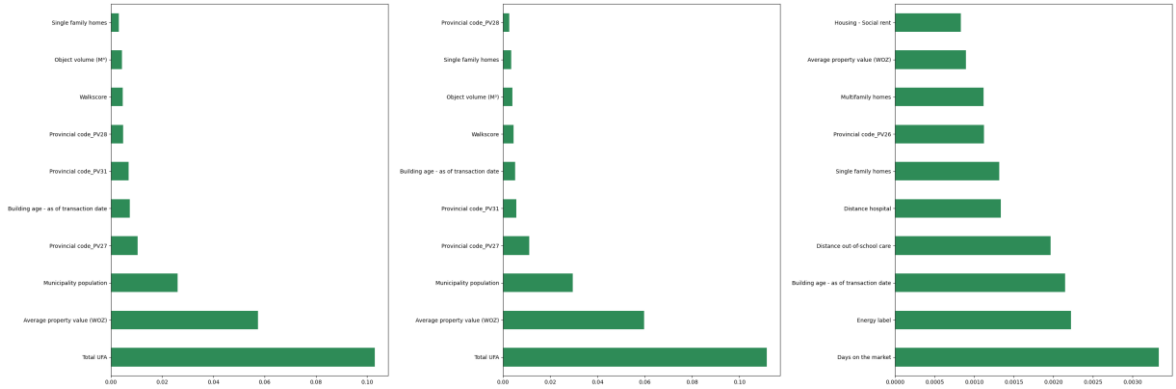

Figure 14. Extreme gradient boosting (left), extreme gradient boosting with monotonic constraints (middle), and hybrid methodology (right) feature importance in predicting the market rent. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

6.3 Vacant possession value

Table 10 provides the accuracy and reliability per methodology and performance metric for the vacant transaction data, where the most accurate models according to each metric are highlighted in bold. Similar to the other MV constituents, we observe a general increase in performance across all metrics for more flexible methodologies. All methodologies show high reliability, with low standard errors for each performance metric. In terms of specific methodologies and in line with the rental transaction data, the GB methodology performs best across most metrics. In contrast to the rental transaction data, however, central tendency is lowest for the XGB methodology according to the LMDPE metric, meaning XGB is most symmetric in bias and robust to outliers. In line with the other MV constituents, we observe the largest differences in this metric, where the GB and XGBMC methodologies perform 8.5% and 7.3% worse than the XGB methodology in terms of central tendency, respectively. Other metrics favor the GB methodology, with small differences with the XGB and XGBMC methodologies. Specifically, accuracy in terms of dispersion (MAE) is the highest for the GB methodology, although differences are small with the XGB methodology performing 1.1% worse and the XGBMC

methodology performing 1.8% worse than the GB methodology on this metric. In terms of the absolute ratio (mmMAPE), the GB methodology also performs better in terms of accuracy compared to other methodologies. Differences are relatively small, however, with the XGB methodology performing 1.2% worse and XGBMC methodology performing 1.8% worse than the GB methodology on this metric. Differences between these three methodologies are small compared to the hybrid methodology that performs 89% worse than the GB methodology on this metric.

In terms of both the squared ratio metric (LRMSE) and squared difference metric (RMSE) that penalize large errors more heavily, the GB methodology is again favored over other methodologies. For both metrics, however, differences are small compared to the XGB and XGBMC methodologies. Specifically, the XGB methodology performs 0.7% worse on the LRMSE metric and 0.9% worse on the RMSE metric compared to the GB methodology. Additionally, the XGBMC methodology performs 1.4% worse on the LRMSE metric and 1.6% worse on the RMSE metric compared to the GB methodology. In terms of the percentage ratio (mmPER), 44.2% of predictions are outside ± 10% of the VPV for the GB methodology, 44.6% for the XGB methodology, and 45.0% of predictions falling outside the pre-specified deviation limit for the XGBMC methodology. Finally, the dispersion of the prediction error (IQRat) is the lowest for the GB methodology, with a relative small difference of 1.2% from the XGB methodology and 1.8% from the XGBMC methodology.

Table 10. Vacant possession value - Accuracy and reliability per methodology.

Methodologies	**LMDPE**	**MAE**	**mmMAPE**	**LRMSE**	**RMSE**	**mmPER**	**IQRat**
HPM	-0.0145	530.669	0.154	0.178	774.693	0.554	0.216
	(0.0006)	(2.430)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
RF	-0.0113	469.019	0.139	0.162	642.345	0.514	0.198
	(0.0006)	(4.172)	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)
GB	-0.0089	**402.438**	**0.117**	**0.141**	**558.984**	**0.442**	**0.166**
	(0.0004)	**(1.678)**	**(0.001)**	**(0.001)**	**(0.001)**	**(0.002)**	**(0.002)**
XGB	**-0.0082**	407.027	0.119	0.142	563.746	0.446	0.168
	(0.0004)	(2.037)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
XGBMC	-0.0088	409.632	0.119	0.143	567.947	0.450	0.169
	(0.0004)	(1.804)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Hybrid methodology	-0.0306	795.639	0.237	0.257	1243.809	0.684	0.313
	(0.0009)	(6.332)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)

Note: The dependent variable is the log-transformed vacant possession value, adjusted using Duan's (1983) adjustment factor before calculating performance metrics. Standard errors are calculated using permutation tests and are displayed in parentheses. The most accurate methodologies according to each metric are highlighted in bold.

The relative feature importance for the ten most contributing features in determining VPV is calculated using permutation tests as described in Section 6.1 and displayed in Figures 15 and 16 below. As can be seen in the figures, whether a house is of type single family, average property value, and total usable floor area rank highest in one or several methodologies. Specifically, whether a house is of type single family ranks highest for the HPM methodology, average property value ranks highest for the RF, GB, XGB, and XGBMC methodologies, and total usable floor area ranks highest for the hybrid methodology. Other high-ranking features include building age, energy label, municipality population, object volume, and whether a house is of sub type terraced house and semi-detached house.

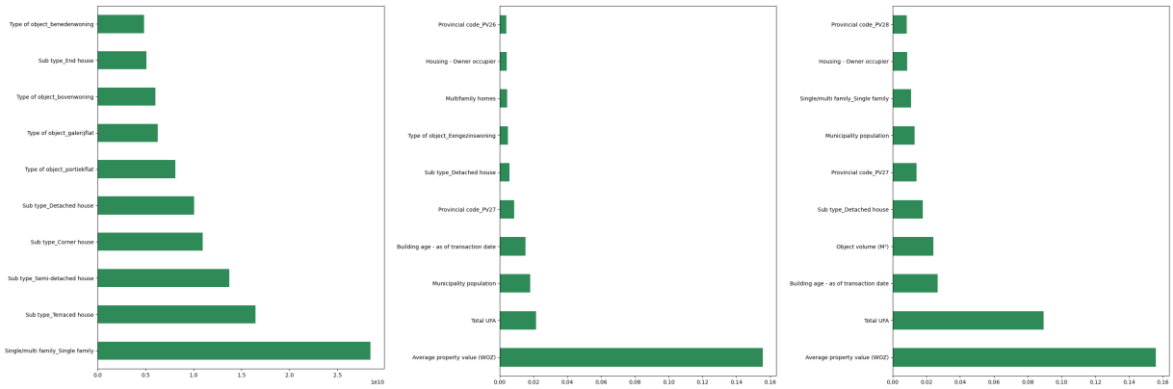

Figure 15. Hedonic pricing model (left), random forest (middle), and gradient boosting (right) feature importance in predicting the vacant possession value. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

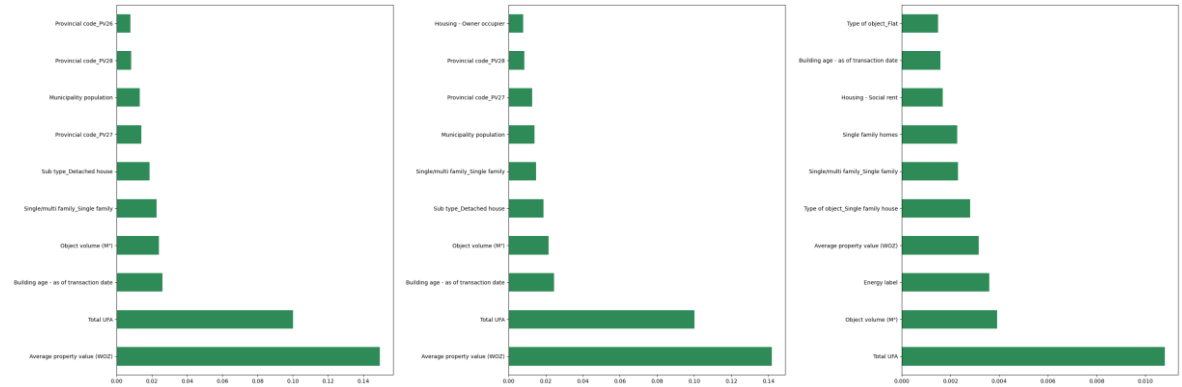

Figure 16. Extreme gradient boosting (left), extreme gradient boosting with monotonic constraints (middle), and hybrid methodology (right) feature importance in predicting the vacant possession value. Feature importance is determined using permutation tests (see Saarela & Jauhiainen (2021) for derivations).

6.4 Overall comparison

Based on the accuracy and reliability results in the previous sections, we summarize the performance of each methodology in Table 11. Additionally, explainability is compared on the basis of the algorithmic transparency, decomposability, and simulatability described in Arrieta et al. (2020) from an appraiser's perspective. In terms of accuracy and reliability, we conclude GB to perform best in the case of MR and

VPV. The XGB and XGBMC methodologies, however, perform very similarly for both MV constituents with small relative differences in most metrics. In the case of the GIM, the XGB methodology performs best across most metrics. For this MV constituent too, however, we observe small relative differences on most metrics with the GB and XGBMC methodologies. Comparing XGBMC to XGB, specifically, we find the XGBMC methodology to perform slightly worse with small relative differences in most accuracy metrics and MV constituents and reliability being generally very similar. The RF methodology ranks fourth in terms of both accuracy and reliability across most metrics and MV constituents, followed by the HPM methodology. The hybrid methodology structurally ranks last in terms of both accuracy and reliability with large relative differences in most metrics and MV constituents compared to best-performing methodologies.

In terms of explainability, we argue for the following differences between methodologies. First, due to its interpretable and uncomplex specification, we argue the HPM to be both algorithmically transparent, decomposable, and simulatable. Appraisers can understand predictions due to the reporting of regressor values (algorithmically transparent), the readability (decomposability), and direct understanding of the prediction process (simulatability) due to the uncomplex nature of the HPM. Second, the decision-tree-based methodologies potentially meet similar requirements but lack all three explainability dimensions due to their complexity. Specifically, due to the ensemble of decision trees in the RF, GB, XGB, and hybrid methodologies, appraisers are no longer able to wholly understand predictions made by these methodologies without requiring mathematical background (simulatability), readability is lost due to model complexity (decomposability), and appraisers are no longer able to obtain a direct understanding of the prediction process in each methodology (algorithmic transparency). The XGBMC methodology, however, is argued to meet some requirements for algorithmic transparency. Specifically, MCs provide semantic meaning to appraisers on how shape-constrained attributes contribute to predictions. Appraisers can therefore understand the prediction process beyond that offered in similar, non-constrained methodologies. Table 11 summarizes the accuracy, reliability, and explainability per methodology.

Table 11. Overall comparison of methodologies according to their accuracy, reliability, and explainability (measured through algorithmic transparency, decomposability, and simulatability).

Methodologies	Accuracy	Reliability	Algorithmic transparency	Decomposability	Simulatability
HPM	-	-	+	+	+
RF	+/-	+/-	-	-	-
GB	+	+	-	-	-
XGB	+	+	-	-	-
XGBMC	+	+	+/-	-	-
Hybrid methodology	-	-	-	-	-

Note: Table 11 presents the overall comparison of the methodologies per market value constituent in terms of accuracy, reliability, and explainability. Accuracy and reliability are determined based on the results shown in the previous sections and summarized as best (+), semi-best (+/-), and worst-performing (-). Explainability is compared on algorithmic transparency, decomposability, and simulatability and based on conceptual arguments from an appraiser's perspective. The methodologies are summarized as satisfying (+), semi-satisfying (+/-), or not satisfying (-) the explainability dimension.

7. Conclusions and discussion

In this paper, we conducted a case study of valuing buy-to-let properties in the Netherlands by forecasting the three constituents of MV in the SPCM: GIM, MR, and VPV. We focused specifically on developing a valuation methodology that incorporates appraiser expertise through MCs, the XGBMC methodology, to investigate its potential in terms of accuracy, explainability, and reliability. The proposed appraiser-based methodology was compared on these three aspects to a similar, non-monotonically constraint XGB methodology, the econometric HPM, RF, GB, and a hybrid methodology comprising HPM with an XGB estimation procedure. Several important insights arise from our case study.

First, we compared the accuracy and reliability of the considered methodologies using seven symmetric metrics proposed by Steurer et al. (2021). Depending on the metric and MV constituent, best-performing methodologies differ, but we identify several general trends. Specifically and as argued in Steurer et al. (2021), accuracy and reliability increase when more transaction data are available. Results were therefore generally better in the case of VPV compared to MR and GIM. In terms of specific methodologies, the most accurate and reliable methodology for MR and VPV was the GB methodology, although differences were marginal with the XGB and XGBMC methodologies for most metrics. This finding can be attributed to the use of relatively balanced transaction data for all three MV constituents within our case study, which also resulted in the overall high reliability across metrics and MV constituents. In future research, it can be interesting to investigate how the XGBMC compares to existing methodologies using imbalanced transaction data. Additionally, as the GB methodology was

generally found most accurate in this study, the incorporation of MCs or similar regularization efforts in the GB methodology is an interesting topic for future research.

In the case of GIM, the XGB methodology was found to be most accurate and reliable for most metrics. The appraiser-based XGBMC is similar to the XGB methodology in terms of reliability, but slightly worse than the XGB methodology in terms of accuracy for most metrics and MV constituents, thus indicating the cost of imposing constraints on the model specification. The hybrid methodology performed worse than other considered methodologies in terms of both accuracy and reliability in the case of all three MV constituents. The considered hybrid specification shows signs of overfitting training data, therefore performing worse on unseen transactions. More complex model specifications and other estimation procedures can help prevent this issue. This is an area of future research.

Second, explainability is difficult to objectively measure and compare between methodologies. Within this paper, we took an appraiser's perspective in conceptually determining the level of explainability of each methodology. Based on three categories of explainability proposed in Arrieta et al. (2020), we conclude the XGBMC methodology, apart from the econometric HPM, to offer the most explainability from an appraiser's perspective. Imposing MCs can therefore be a valuable tool in making MV predictions more transparent to both appraisers and clients. Explainability was measured, however, using conceptual arguments as empirical measures are lacking in the literature. We leave the exploration of empirical explainability to future research.

Third, although the XGBMC methodology was shown to be slightly less accurate than the unconstrained XGB methodology, the XGBMC methodology is equally reliable and more explainable according to the explainability dimensions adopted in this study. Appraisers, therefore, can benefit from a slightly less accurate but equally reliable and more explainable model specification, compared to existing methodologies. Increasing the explainability of methodologies without incurring the loss of accuracy is a topic for future research.

Finally, according to Scheurwater (2017), it is unlikely that AVMs will entirely substitute manual valuations in the near future. Specifically, current legal frameworks dictate that appraisers should be able to understand and explain valuation methodologies. As described in the AVM roadmap outlined by the RICS (2021), it is currently unclear when AVM output is within the scope of international appraisal guidelines (i.e. International Valuation Standards; IVSC, 2019) due to explainability issues. Because of insufficient explainability, appraisers are unable to understand the valuation process through AVMs, rendering them inappropriate for professional use. Nonetheless, the proposed methodology in this study improves on the explainability of existing methodologies, bringing it closer to professional applicability. Although AVMs are unlikely to substitute manual valuations in the near future, they can complement professional appraisers in creating a more efficient valuation process.

Ethical considerations

The research described in this paper is conducted by considering several ethical principles. Based on the principles put forth by the VSNU (2018), *transparency* was guaranteed by reporting on all aspects of the research process. This includes, first, a thorough, separate chapter on the data preparation phase, including sources and decisions made in terms of data management. Additionally, detailed descriptions of methodologies used in this paper are contained in the appendices. *Independence* and *responsibility* are ensured by taking all steps of the research independently. Supervision was provided by both the supervisor at the University of Groningen and Envalue (see Appendix 1: Envalue valuation firm). Both supervisors, however, were not involved in conducting the research itself, and responsibility for its outcomes remains by the author. *Responsibility* was further taken for the obtained data by treating it confidentially, including the decision to not publicly share the data obtained. Finally, all general ethical guidelines and standards as put forth by VSNU (2018) were adhered to during the research process.

Literature

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*, vol. 39, no. 2, pp. 1772–1778.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. 2020. "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, vol. 58, pp. 82–115.

Bartley, C., Liu, W., & Reynolds, M. 2019. "Enhanced random forest algorithms for partially monotone ordinal classification", *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, HI, USA.

Beimer, J., & Francke, M. 2019. "Out-of-sample house price prediction by hedonic price models and machine learning algorithms", *Real Estate Research Quarterly*, vol. 18, no. 2, pp. 13–20.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. 2011. "Algorithms for hyper-parameter optimization", *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, Granada, Spain.

Bergstra, J., & Bengio, Y. 2012. "Random search for hyper-parameter optimization", *Journal of Machine Learning Research*, vol. 13, pp. 281–305.

Bergstra, J., Yamins, D., & Cox, D. D. 2013. "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures", *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, USA.

Breiman, L. 1996. "Bagging predictors", *Machine Learning*, vol. 24, no. 2, pp. 123–140.

Breiman, L. 2001a. "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32.

Breiman, L. 2001b. "Statistical modeling: The two cultures", *Statistical Science*, vol. 16, no. 3,

pp. 199–215.

Canini, K., Cotter, A., Gupta, M. R., Fard, M. M., & Pfeifer, J. 2016. "Fast and flexible monotonic functions with ensemble of lattices", *30th Conference on Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

CBS. 2021a. *Housing market visualizations*, retrieved from Statistics Netherlands: https://www.cbs.nl/nl-nl/visualisaties/huizenmarkt-in-beeld

CBS. 2021b. *Number of owner-occupied property transactions increases by more than 29% in first quarter*, retrieved from Statistics Netherlands, https://www.cbs.nl/nl-nl/nieuws/2021/16/aantal-transacties-koopwoningen-ruim-29-procent-hoger-in-eerste-kwartaal

CBS. 2021c. *Open data*, retrieved from Statistics Netherlands, https://www.cbs.nl/en-gb/onze-diensten/open-data

Ceyhan, C. 2017. "Evaluation of machine learning techniques for house valuations", MSc thesis, Erasmus University, Rotterdam.

Chen, T., & Guestrin, C. 2016. "XGBoost: A scalable tree boosting system", *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA.

Clayton, J., Geltner, D., & Hamilton, S. W. 2001. "Smoothing in commercial property valuations: Evidence from individual appraisals", *Real Estate Economics*, vol. 29, no. 3, pp. 337–360.

Conijn, J., & Schilder, F. 2011. "How housing associations lose their value: The value gap in the Netherlands", *Journal of Property Management*, vol. 29, no. 1, pp. 103–119.

Crosby, N. 2000. "Valuation accuracy, variation and bias in the context of standards and expectations", *Journal of Property Investment & Finance*, vol. 18, no. 2, pp. 130–161.

Crosby, N., Devaney, S., Lizieri, C., & McAllister, P. 2018. "Can institutional investors bias real estate portfolio appraisals? Evidence from the market downturn", *Journal of Business Ethics*, vol. 147, no. 3, pp. 651–667.

Davison, A. C., & Hinkley, D. V. 1997. *Bootstrap methods and their application*, Cambridge University Press, Cambridge, UK.

Debarsy, N., & Lesage, J. 2021. "Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices", *Journal of Business and Economic Statistics*, pp. 1–35.

Duan, N. 1983. "Smearing estimate: A nonparametric retransformation method", *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 605–610.

Fakton. 2021. *Handboek modelmatig waarderen marktwaarde* [Handbook for the model-based appraisal of market value], retrieved from the Ministry of the Interior and Kingdom Relations, https://www.woningmarktbeleid.nl/documenten/publicaties/2020/10/30/handboek-marktwaardering-2020

Francke, M., & Kroon, D. 2021. *Machine learning in real estate valuation*, Ortec

Finance technical report, Rotterdam.

French, N., & Gabrielli, L. 2004. "The uncertainty of valuation", *Journal of Property Investment & Finance*, vol. 22, no. 6, pp. 484–500.

Friedman, J. H. 2001. "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232.

Gao, Y., & Li, K. 2013. "Nonparametric estimation of fixed effects panel data models", *Journal of Nonparametric Statistics*, vol. 25, no. 3, pp. 679–693.

Gelman, A., & Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, Cambridge, UK.

Goodfellow, I., Bengio, Y., & Courville, A. 2015. *Deep Learning*, MIT Press, MA, USA.

Graczyk, M., Lasota, T., Trawiński, B., & Trawiński, K. 2011. "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal", *Proceedings of the second international conference on intelligent information and database systems: Part II"*, Hue, Vietnam.

Gupta, M. R., Bahri, D., Cotter, A., & Canini, K. 2018. "Diminishing returns shape constraints for interpretability and regularization", *32nd Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

Hastie, T., Tibshirani, R., & Friedman, J. 2017. *The elements of statistical learning: Data mining, inference, and prediction (second edition)*, Springer, NY, USA.

Hilgers, B., Brantsma, J., & Boeve, J. 2021. "Towards automating commercial real estate valuations? An experiment within valuation practice", *Real Estate Research Quarterly: Taxatiespecial*, vol. 20, no. 2, pp. 8–16.

Hinrichs, N., Kolbe, J., & Werwatz, A. 2021. "Using shrinkage for data-driven automated valuation model specification - a case study from Berlin", *Journal of Property Research*, vol. 38, no. 2, pp. 130–153.

Horváth, Á., Imre, B., & Sápi, Z. 2016. "The international practice of statistical property valuation methods and the possibilities of introducing automated valuation models in Hungary", *Financial and Economic Review*, vol. 15, no. 4, pp. 45–64.

IAAO. 2018. *Standard on automated valuation models (AVMs)*, retrieved from the International Association of Assessing Officers, https://www.iaao.org/media/standards/AVM_STANDARD_2018.pdf.

IVSC. 2003. *Glossary of terms for international valuation standards*, retrieved from the International Valuation Standards Council: https://www.icjce.es/images/pdfs/TECNICA/C02%20-%20IASB/C210%20-%20IVSC%20-%20Normas/27-glossary.pdf

IVSC. 2019. *International valuation standards*, retrieved from the International Valuation Standards Council, https://www.rics.org/globalassets/rics-website/media/upholding-professional-standards/sector-standards/valuation/international-valuation-standards-rics2.pdf

Jordan, M. I., & Mitchell, T. M. 2015. "Machine learning: Trends, perspectives, and prospects",

Science, vol. 349, no. 6245, pp. 255–260.

Kadaster, 2021. *BAG API individuele bevragingen*, retrieved from the Netherlands' Cadastre, Land Registry and Mapping Agency: https://www.kadaster.nl/zakelijk/producten/adressen-en-gebouwen/bag-api-individuele-bevragingen

Kok, N., Koponen, E., & Martínez-Barbosa, C. A. 2017. "Big data in real estate? From manual appraisal to automated valuation", *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211.

Krause, A., Martin, A., & Fix, M. 2020. "Uncertainty in automated valuation models: Error-based versus model-based approaches", *Journal of Property Research*, vol. 37, no. 4, pp. 308–339.

Leidelmeijer, K., Middeldorp, M., & Marlet, G. 2019. *Leefbaarheid in Nederland 2018*, retrieved from the Ministry of the Interior and Kingdom Relations, https://www.leefbaarometer.nl/home.php

Li, S., & Chen, C. 2015. "A regularized monotonic fuzzy support vector machine model for data mining with prior knowledge", *IEEE Transactions on Fuzzy Systems*, vol. 23 no. 5, pp. 1713–1727.

Liu, F. T., Ting, K. M., & Zhou, Z. 2008. "Isolation forest", *Eighth IEEE International Conference on Data Mining*, Pisa, Tuscany, Italy.

Malikov, E., & Sun, Y. 2017. "Semiparametric estimation and testing of smooth coefficient spatial autoregressive models", *Journal of Econometrics*, vol. 199, no. 1, pp. 12–34.

Malpezzi, S. 2003. "Hedonic pricing models: A selective and applied review", in O'Sullivan, T. & Gibb, K. (Eds.), *Housing economics and public policy: Essays in honour of Duncan Maclennan*, Hoboken, NJ: Blackwell

Markowitz, H. M. & Blay, K. A. 2014. *Risk-return analysis: The theory and practice of rational investing*, vol. 1, New York, NY: McGraw-Hill.

McAllister, P., Baum, A., Crosby, N., Gallimore, P., & Gray, A. 2003. "Appraiser behaviour and appraisal smoothing: Some qualitative and quantitative evidence", *Journal of Property Research*, vol. 20, no. 3., pp. 261–280.

McCluskey, W. J., McCord, M., Davis, P.T., Haran, M., & McIlhatton, D. 2013. "Prediction accuracy in mass appraisal: A comparison of modern approaches", *Journal of Property Research*, vol. 30, no. 4, pp. 239–265.

Mooya, M. 2011. "Of mice and men: Automated valuation models and the valuation profession", *Urban Studies*, vol. 48, no. 11, pp. 2265–2281.

Mullainathan, S., & Spiess, J. 2017. "Machine learning: An applied econometric approach", *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106.

Nakatsu, R. T. 2021. "An evaluation of four resampling methods using in machine learning classification", *IEEE Intelligent Systems*, vol. 36, no. 3, pp. 51–57.

NEA. 2021. *EP-online*, retrieved from the Netherlands Enterprise Agency: https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/hulpmiddelen-tools-en-

inspiratie-gebouwen/ep-online

Nguyen, N., & Cripps, A. 2001. "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks", *Journal of Real Estate Research*, vol. 22, no. 3, pp. 313–336.

NRVT. 2017. *Reglement bedrijfsmatig vastgoed NRVT* [Regulations commercial real estate NRVT], retrieved from the Register of Real Estate Appraisers Netherlands, https://www.nrvt.nl/wp-content/uploads/2015/07/Reglement-Bedrijfsmatig-Vastgoed-15-6-2017-1.pdf

NRVT. 2021. *Regelgeving* [Regulations], retrieved from the Register of Real Estate Appraisers Netherlands, https://www.nrvt.nl/regelgeving/kamer-en-reglementen/

NVM. 2021. *Marktinformatie koopwoningen* [Market information owner-occupied housing], retrieved from the Dutch Cooperative Association of Realtors and Valuers in Realty: https://www.nvm.nl/wonen/marktinformatie/

PDOK. 2021. *PDOK location server*, retrieved from Public Services on the Map: https://www.pdok.nl/restful-api/-/article/pdok-locatieserver-1

Peterson, S., & Flanagan, A. 2009. "Neural network hedonic pricing models in mass real estate appraisal", *Journal of Real Estate Research*, vol. 31, no. 2, pp. 147–164.

Probst, P., Boulesteix, A., & Bischl, B. 2019. "Tunability: Importance of hyperparameters of machine learning algorithms", *Journal of Machine Learning Research*, vol. 20, pp. 1–32.

RICS. 2018. *Risk, liability and insurance in valuation work*, retrieved from the Royal Institution of Chartered Surveyors: https://www.rics.org/globalassets/rics-website/media/upholding professional-standards/sector-standards/valuation/risk-liability-and-insurance-in-valuation work-2nd-edition-rics.pdf

RICS. 2019. *Home survey standard*, retrieved from the Royal Institution of Chartered Surveyors, https://www.rics.org/globalassets/rics-website/media/qualify/home-survey-standard-nov-2020.pdf

RICS. 2020a. *RICS valuation - global standards*, retrieved from the Royal Institution of Chartered Surveyors, https://www.rics.org/globalassets/rics-website/media/upholding-professional-standards/sector-standards/valuation/rics-valuation--global-standards.pdf

RICS. 2020b. *Valuing residential property purpose built for renting*, retrieved from the Royal Institution of Chartered Surveyors, https://www.rics.org/globalassets/rics-website/media/upholding-professional-standards/sector-standards/valuation/valuing-residential-property-purpose-built for-renting-1st-edition-rics2.pdf

RICS. 2021. *Automated valuation models roadmap for RICS members and stakeholders*, retrieved from the Royal Institution of Chartered Surveyors, https://www.rics.org/nl/upholding-professional-standards/sector-standards/valuation/automated-valuation-models/

Robinson, S. J., & Sanderford, A. R. 2017. "Hedonic models and the inclusion of conditions of sale in commercial real estate transactions: A review of the literature", *Journal of Real Estate*

Literature, vol. 25, no. 2, pp. 311–326.

Rosen, S. 1974. "Hedonic prices and implicit markets: Product differentiation in pure competition", *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55.

Saarela, M., & Jauchiainen, S. 2021. "Comparison of feature importance measures as explanations for classification models", *Springer Nature Applied Sciences*, vol. 3, no. 272.

Scheurwater, S. 2017. *The future of valuations*, retrieved from the Royal Institution of Chartered Surveyors: https://www.rics.org/globalassets/rics-website/media/knowledge/research/insights/future-of-valuations-insights-paper-rics.pdf

Scheurwater, S. 2018. *Automated valuation models - Which future?*, retrieved from the Royal Institution of Chartered Surveyors: https://www.rics.org/nl/news-insight/latest-news/news-opinion/automated-valuation-models--which-future

Schimert, J., & Wineland, A. 2010. "Coupling a dynamic linear model with random forest regression to estimate engine wear", *Annual Conference of the Prognostics and Health Management Society*, Portland, OR, USA.

Schulz, R., Wersing, M., & Werwatz, A. 2014. "Automated valuation modelling: A specification Exercise", *Journal of Property Research*, vol. 31, no. 2, pp. 131–153.

Schulz, R., & Wiersing, M. 2021. "Automated valuation services: A case study for Aberdeen in Scotland", *Journal of Property Research*, vol. 38, no. 2, pp. 154–172.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. 2015. "Taking the human out of the loop: A review of Bayesian optimization", *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175.

Shalev-Shwartz, S., & Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*, Cambridge University Press, Cambridge, UK.

Sheppard, S. 1999. "Hedonic analysis of housing markets", In Cheshire, P. & Mills, E. S., (Eds.) *Handbook of Regional and Urban Economics* (Vol. 3), North-Holland, Amsterdam, The Netherlands.

Smyl, S. 2020. "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting", *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85.

Steurer, M., Hill, R. J., & Pfeifer, N. 2021. "Metrics for evaluating the performance of machine learning based automated valuation models", *Journal of Property Research*, vol. 38, no. 2, pp. 99–129.

TEGOVA. 2020. *European valuation standards*, retrieved from The European Group of Valuers' Associations: https://tegova.org/european-valuation-standards-evs

Varian, H. R. 2014. "Big data: New tricks for econometrics", *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28.

VSNU. 2018. *Netherlands code of conduct for research integrity*, retrieved from the Association of Universities in The Netherlands, https://www.vsnu.nl/en_GB/research-integrity

Walk Score. 2021. *Walk Score – About us*, retrieved from Walk Score:

https://www.walkscore.com/about.shtml

Walvekar, G., & Kakka, V. 2020. *Private real estate: Valuation and sale price comparison 2019,* retrieved from Morgan Stanley Capital International: https://www.msci.com/www/research-paper/private-real-estate-valuations/01032853837

Yang, Y. 2007. "Consistency of cross validation for comparing regression procedures", *The Annals of Statistics*, vol. 35, no. 6, pp. 2450–2473.

You, S., Ding, D., Canini, K., Pfeifer, J., & Gupta, M. R. 2017. "Deep lattice networks and partial monotonic functions", *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

Zhang, G. P. 2003. "Time series forecasting using a hybrid ARIMA and neural network model", *Neurocomputing*, vol. 50, pp. 159–175.

Zurada, J., Levitan, A. S., & Guan, J. 2011. "A comparison of regression and artificial intelligence methods in a mass appraisal context", *The Journal of Real Estate Research*, vol. 33, no. 3, pp. 349–388.

Appendix 1: Envalue valuation firm

This thesis is written in collaboration with Envalue, a Dutch valuation firm focusing on commercial real estate valuations. Envalue is a recently established valuation firm (Est. 2020) with a strong emphasis on data-driven valuations. In contrast to the traditional valuation process, Envalue strives to conduct valuations within a short timeframe by focusing on various data-driven solutions. First, Envalue builds and expands databases for each commercial valuation segment (e.g., buy-to-let, office). These efforts entail a focus on both data quantity (i.e., number of transactions) and quality (i.e., extensive, and valid information on each transaction). Second, Envalue strives for data-driven valuation solutions to complement the valuer. These data-driven valuation solutions are reflected in the development of model-based valuations using the expertise of the valuer, or the so-called *hybrid approach* to valuations. Finally, streamlining valuations by automating all routine processes. For example, appraisers receive comparable references through algorithmic calculations and aspects of valuation reports are automated using data-driven connectivity (e.g., through application programming interfaces (APIs)), enabling the valuer to focus on the non-routine aspects of commercial real estate valuations.

Envalue is part of the BORON Valuation Group, a collective of six data-driven real estate firms. One closely related firm within the BORON Valuation Group is Property Pass. Originally developed by Deloitte (then called AXIOM), Property Pass is a digital data-sharing platform providing a central database of verified real estate data. Specifically, Property Pass enables sharing of structured rental data with banks in compliance with the European Central Bank's current guidelines. Property Pass is currently in a pilot phase with three major Dutch banks (i.e., ABN AMRO, ING Group, Rabobank) using SBR Nexus[22]. The central database resulting from these efforts is made available to other firms in the BORON Valuation Group, including Envalue. In line with the developments at Envalue, this database ensures readily available, extensive, and verified data to streamline the valuation process.

[22] SBR Nexus develops and provides market standards for sharing professional data. For an overview of their conduct, see https://www.sbrnexus.nl/over-ons.