Journal of Property Research

VOLUME FORTY NUMBER TWO JUNE 2023

Routledge

Journal of Property Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjpr20

Automated valuation models: improving model performance by choosing the optimal spatial training level

Bastian Krämer, Moritz Stang, Vanja Doskoč, Wolfgang Schäfers & Tobias Friedrich

To cite this article: Bastian Krämer, Moritz Stang, Vanja Doskoč, Wolfgang Schäfers & Tobias Friedrich (2023): Automated valuation models: improving model performance by choosing the optimal spatial training level, Journal of Property Research, DOI: 10.1080/09599916.2023.2206823

To link to this article: https://doi.org/10.1080/09599916.2023.2206823

Published online: 02 May 2023.

JOURNAL OF PROPERTY RESEARCH
https://doi.org/10.1080/09599916.2023.2206823

Automated valuation models: improving model performance by choosing the optimal spatial training level

Bastian Krämer[a], Moritz Stang[a], Vanja Doskoč[b], Wolfgang Schäfers[a] and Tobias Friedrich[b]

[a]International Real Estate Business School, University of Regensburg, Regensburg, Germany; [b]Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

ABSTRACT

The academic community has discussed using Automated Valuation Models (AVMs) in the context of traditional real estate valuations and their performance for several decades. Most studies focus on finding the best method for estimating property values. One aspect that has not yet to be studied scientifically is the appropriate choice of the spatial training level. The published research on AVMs usually deals with a manually defined region and fails to test the methods used on different spatial levels. Our research aims to investigate the impact of training AVM algorithms at different spatial levels regarding valuation accuracy. We use a dataset with 1.2 million residential properties from Germany and test four methods: Ordinary Least Square, Generalised Additive Models, eXtreme Gradient Boosting and Deep Neural Network. Our results show that the right choice of spatial training level can significantly impact the model performance, and that this impact varies across the different methods.

ARTICLE HISTORY
Received 14 October 2022
Accepted 18 April 2023

KEYWORDS
Automated Valuation Models (AVMs); machine learning; spatial training level; model performance; valuation accuracy

Introduction

The academic community has discussed using Automated Valuation Models (AVMs) in the context of traditional real estate valuations and their performance for several decades, and practitioners are also now increasingly scrutinising it. Most studies focus on the comparison of different statistical methods. Accordingly, a large body of literature compares traditional hedonic models with more modern machine learning (ML) approaches or approaches from spatial econometrics (see, e.g. Pace & Hayunga, 2020). These studies aim to identify which method is best suited for estimating real estate values or prices.

Besides the method selection, AVMs can be optimised in many other areas. For example, the selection, cleaning and preparation of data play an important role for the overall performance of the AVM. Another aspect is the choice of spatial level to train the selected methods. This is decisive for determining which data are ultimately included in the estimation of the AVM and, thus, what information is used or ignored. Thanks to georeferencing, models can, in principle, be trained at any level. For example, a model can be trained at the city level, the associated commuter belt, or even nationwide.

CONTACT Bastian Krämer bastian.kraemer@ur.de

However, this aspect has received little to no attention from the academic community until now.

The published research on AVMs usually deals only with a manually defined region and fails to test the methods used on different spatial levels. One reason for this might be that historically, the availability of suitable real estate data[1] for academic purposes has been limited. Therefore, analyses could only be conducted in the limited area where the data was available. However, data availability has improved massively in recent years, so this has become less of a factor (Bankers Association, 2019). In the meantime, there are providers of real estate-related data in almost every country, which centrally force a collection of existing data and make them available for further analysis. Another reason could be the usually assumed heterogeneity of real estate markets. Traditionally, real estate markets are believed to have a certain regionality, meaning that data from other diverging regions would not provide additional explanatory power. However, the fundamental question arises as to whether this heterogeneity is generally present or whether there are not also basic characteristics that apply consistently to all markets. If this is the case, achieving a higher degree of valuation accuracy may be possible by adding further data from different markets.

Therefore, it raises the question of whether considering different spatial levels for training AVMs could be an important and undervalued factor in enhancing their valuation accuracy. Our research aims to answer this question and investigate the influence of training statistical models used for AVMs on different spatial levels.

For this purpose, we compare four different methods trained on four differing spatial levels each and compare the overall performance of the models. Our objective is not primarily a comparison of the methods used, but a specific comparison within the individual methods concerning their performance on different spatial levels. We are interested in whether different methods deliver different results and whether any patterns emerge.

The methods selected represent a collection of regularly used ones in academic studies related to AVMs. In addition to parametric Ordinary Least Square (OLS) regressions, we analyse semi-parametric Generalised Additive Models (GAM), eXtreme Gradient Boosting (XGBoost) algorithms and Deep Neural Networks (DNN) from the field of modern ML. Our analysis is based on a dataset of 1.2 million residential properties across Germany provided by professional real estate appraisers. The four spatial levels are based on the NUTS nomenclature of the European Union. The NUTS (Nomenclature of territorial units for statistics) classification is a hierarchical system for dividing up the economic territory of the EU and the UK. There are four different subdivision levels, called NUTS-0, NUTS-1, NUTS-2 and NUTS-3, which we use to train our models on a country, state, cross-regional, and county level, respectively.[2]

Our research has theoretical and practical implications that collectively help improve AVMs' valuation accuracy. Our findings show that the right choice of spatial training level can significantly influence the model performance of different AVM algorithms, and that this influence varies considerably, depending on the type of method. The results indicate that for parametric and semi-parametric approaches, choosing a relatively small training level is advisable. This shows that the trained OLS and GAM cannot draw additional explanatory power from observations outside a particular region. The results for the two modern ML algorithms are quite different. We observe that they can gain more explanatory power by adding further observations, and that this effect outweighs local heterogeneity. Therefore, we recommend, choosing a generally higher training level for modern ML algorithms.

The contributions of our paper are manifold. First and foremost, our findings provide further evidence that when it comes to applying more traditional versus modern ML methods, fundamental differences in their application should be considered to achieve the best model performance. Our findings indicate that assumptions valid for applying traditional ML methods may not be suitable for modern methods.

This provides real estate researchers and practitioners with new guidelines for using different AVM algorithms, which can help improve the performance of their valuation results. Additionally, our findings also shed light on the question of whether real estate markets are characterised by high local heterogeneity. The results of our OLS and GAM models study suggest significant heterogeneity in local real estate markets. Still, the results of the XGBoost and DNN indicate that there are overall patterns that apply to all real estate markets. In summary, our paper provides a new set of guidelines that can be used to answer various real estate-related questions more accurately. These new guidelines are a starting point for further research into the analysis of real estate markets using modern ML algorithms.

Literature Review

AVMs are computer-based applications that use various statistical and algorithmic approaches to assess the value or price of a property in an automated manner. They can be a cost-effective and rapid alternative to traditional valuation procedures (Schulz et al., 2014).

AVMs emerged mainly from research on hedonic price models (HPM). HPMs were developed to estimate the effects of individual characteristics, so-called marginal prices, of a good on its value or price. By aggregating these marginal prices, the overall value of a good can subsequently be calculated (Chau & Chin, 2002). HPMs were first brought into a real estate context by Lancaster (1966) and Rosen (1974). As Malpezzi (2003) and Sirmans et al. (2005) show, a diverse and dynamic field of research has emerged since then, addressing a wide variety of real-estate-specific issues.

To improve the quality of automated real estate appraisals, the research community's focus in recent years has been almost exclusively on finding the best-fitting method. For this purpose, many approaches were either newly designed, or adapted and applied from other areas. The applied methods cover the full bandwidth of statistical methods and can be classified as parametric, semi-parametric or non-parametric approaches. Regarding parametric approaches, the most common multiple linear regression (MLR) models are applied and tested. Schulz et al. (2014), for example, use a flexible parametric hedonic regression introduced by Bunke et al. (1999) to measure the potential predictive performance of an AVM applied to the housing market of Berlin in Germany. Other examples of parametric approaches can be found at Tse (2002), Pace and LeSage (2004), Páez et al. (2008), Bourassa et al. (2008), Osland (2010) and Zurada et al. (2011).

Semi-parametric approaches can come in a variety of different forms. An often-used semi-parametric approach is the GAM, first introduced by Hastie and Tibshirani (1987). In contrast to traditional MLR models, the GAM can automatically control for nonlinear relations between the dependent and independent variables. An early and prominent application within a real estate context is the study of Pace (1998). The author applies a GAM to a dataset for residential properties in Memphis (Tennessee) and finds that the GAM can outperform parametric and polynomial methods in terms of predictive behaviour.

A more recent example of the GAM can be found in Dąbrowski and Adamczyk (2010). Non-parametric approaches are a category of methods which do not need an a-priori specified functional form regarding the predictor. Instead, the form is learned from the information derived from the data itself. Given this flexibility, non-parametric approaches usually account for non-linearities and interactions within datasets and outperform parametric and semi-parametric approaches (Stang et al., 2022).

Prominent examples of non-parametric approaches include modern machine learning methods like Support Vector Machines, Artificial Neural Networks or Tree-Based Models. A real-estate-specific application of ML methods can be found in Mayer et al. (2019). The authors apply three commonly used basic techniques of modern ML (Random Forrest Regression, Gradient Boosting and Neural Networks) and compare their performance against some more traditional parametric approaches. Their findings show that the non-parametric methods can outperform stricter parametric approaches. Other real-estate-specific applications of non-parametric modelling techniques can be found in Chun Lin and Mohan (2011), Yoo et al. (2012), Antipov and Pokryshevskaya (2012), W. J. McCluskey et al. (2013), Kok et al. (2017), and Yilmazer and Kocaman (2020).

Another aspect with regard to the optimisation of the valuation accuracy of AVMs is, besides the method selection, the choice of spatial level for training the models. The level at which the models are trained implies for which data, and thus ultimately also which information is considered in the context of the valuation and which is not. This could have a strong influence on the results of the models and is therefore a factor that should not be neglected. AVM-related studies currently always focus on a predefined region. The region to which the analyses are limited is in most cases the city level or the immediate surroundings of a city. Yao et al. (2018), for example, focus on the city level of Shenzhen (China), and W. McCluskey et al. (2012) choose the Lisburn District Council area around Belfast (North Ireland) to test their hypotheses. Other authors go a step further and conduct their analysis at the city district level. Baldominos et al. (2018), for example, focused on the Salamanca district of Madrid (Spain), Hong et al. (2020) run their analysis for the Gangnam district of Soul (South Korea), and Yilmazer and Kocaman (2020) run their model at the Mamak district of Ankara (Turkey). However, none of the authors investigates whether the chosen level is also the best one for training the models.

To the best of our knowledge, no study currently that deals with the optimal spatial level for training AVMs. Therefore, we aim to close this gap in the literature and determine the influence of the choice of spatial level on the quality of statistical valuation results. In particular, we are interested in whether this influence is the same for different types of methods (parametric, semi-parametric, non-parametric) or whether there are fundamental differences. In our analysis, we calculate the valuation accuracy of four different statistical methods (OLS, GAM, XGBoost, DNN), each trained at four different spatial levels, and compare their results subsequently.

Data

We base our analysis on a dataset consisting of 1,212,546 residential properties. These observations are distributed across Germany and were collected between 2014 and 2020. The dataset originates from the valuation department of one of Germany's largest mortgage lenders. Table 1 shows the distribution of the data over the observation period.

The data are actual valuation data collected by professional appraisers. We use the assessed market value as our target variable. An overview of the average market values across Germany is provided in Figure 1. The areas with the highest market values are in the Top-7[3] cities and commuter belts. Furthermore, the market values are by far the highest in the south of Germany and tend to be lower in the east.

As hedonic characteristics, we use a set of features describing the properties' structural characteristics, the micro-location and the macro-location. In addition, the year and

Table 1. Observations per year.

	2014	2015	2016	2017	2018	2019	2020
n	196,318	196,403	176,238	163,365	165,106	165,996	149,120
(%)	0.1619	0.1620	0.1453	0.1347	0.1362	0.1369	0.1230

Notes: This table reports the number of observations available for each year. Over the years, the trend is slightly downward. Especially in 2020, the number of observations is lower, due to the COVID restrictions prevailing at that time. Due to the contact restrictions in place, on-site visits by appraisers were limited.

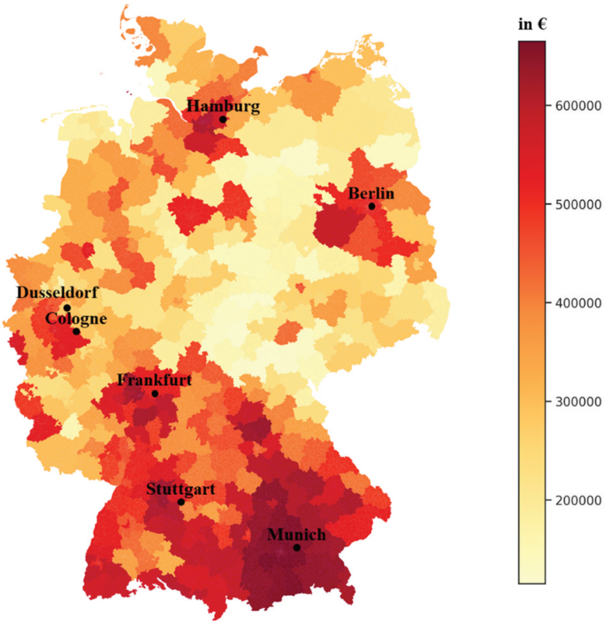

Figure 1. Average market value per district. Notes: This figure shows the average market values per NUTS-3 district. The average was calculated using all available observations within the individual districts. The highest market values are near the major metropolitan regions and in the south of Germany. The substantial discrepancy between the west and east of Germany is striking. The market values observed here are also consistent with other studies (see, e.g., Just & Maennig, 2012), so it can be assumed that the observations used are representative.

quarter of the valuation are used to capture a temporal trend and seasonality. An overview of all the features used and their univariate distribution can be seen in Table 2.[4] Before being used, the dataset was cleaned to account for duplicates, incompleteness, and erroneous data points. There are no correlations of concern within the dataset so that all variables can be integrated accordingly.[5]

Features describing the properties' structural characteristics include the property's subtype, year of construction, modernisation year, living area, lot size, use of the property, quality grade, condition and variable denoting whether the property has

Table 2. Descriptive statistics.

Variable	Unit	Mean	Median	Standard Deviation	Maximum	Minimum
Market value	Integer	228,157.10	200,000.00	141,717.54	3,860,000.00	20100.00
Modernisation year	Integer	1989.10	1988.00	17.19	2020.00	1950.00
Year of construction	Integer	1978.48	1981.00	29.77	2023.00	1900.00
Year of valuation	Integer	2016.82	2017.00	2.00	2020.00	2014.00
Quarter of valuation	Integer	2.45	2.00	1.12	4.00	1.00
Quality grade	Integer	3.12	3.00	0.51	5.00	1.00
Living area	Float	120.31	114.68	51.69	440.00	15.00
Lot size	Float	436.48	323.00	541.66	10,000.00	0.00
Longitude	Float	9.25	8.94	1.90	19.25	5.87
Latitude	Float	50.62	50.74	1.85	55.02	47.40
Micro score	Float	72.73	74.20	14.44	99.85	0.00
Unemployment ratio	Float	4.96	4.17	2.89	26.89	0.04
Time on market	Float	12.27	10.90	4.80	106.00	0.20
Basement condominium	Binary	0.38	0.00	0.48	1.00	0.00
No basement	Binary	0.19	0.00	0.39	1.00	0.00
Basement	Binary	0.44	0.00	0.50	1.00	0.00
Owner-occupied & Non-owner-occupied	Binary	0.09	0.00	0.29	1.00	0.00
Owner-occupied	Binary	0.70	1.00	0.46	1.00	0.00
Non-owner-occupied	Binary	0.21	0.00	0.41	1.00	0.00
Object subtype condominium	Binary	0.38	0.00	0.48	1.00	0.00
Object subtype detached house	Binary	0.42	0.00	0.49	1.00	0.00
Object subtype no detached house	Binary	0.20	0.00	0.40	1.00	0.00
Condition good	Binary	0.38	0.00	0.49	1.00	0.00
Condition disastrous	Binary	0.00	0.00	0.02	1.00	0.00
Condition middle	Binary	0.45	0.00	0.50	1.00	0.00
Condition moderate	Binary	0.02	0.00	0.14	1.00	0.00
Condition bad	Binary	0.00	0.00	0.05	1.00	0.00
Condition very good	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype agglo commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype agglo cbd	Binary	0.13	0.00	0.34	1.00	0.00
Regiotype agglo middle order centre	Binary	0.13	0.00	0.34	1.00	0.00
Regiotype agglo upper order centre	Binary	0.04	0.00	0.19	1.00	0.00
Regiotype rural commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype rural middle order centre	Binary	0.07	0.00	0.26	1.00	0.00
Regiotype rural upper order centre	Binary	0.01	0.00	0.07	1.00	0.00
Regiotype urban commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype urban middle order centre	Binary	0.10	0.00	0.29	1.00	0.00
Regiotype urban upper order centre	Binary	0.07	0.00	0.26	1.00	0.00
NUTS-1	String	-	-	-	-	-
NUTS-2	String	-	-	-	-	-
NUTS-3	String	-	-	-	-	-

Notes: This table reports the descriptive statistics of the dataset. Polytomous variables are one-hot encoded to binary variables to account for the requirements of modern machine learning methods. For the rather traditional methods – OLS and GAM – these polytomous variables are dummy encoded. The numbers were determined on the basis of all available observations. Overall, both structural features and location-describing features were used for model estimation. The selection of the parameters was in accordance with other publications in the AVM literature (see e.g. Metzner & Kindt, 2018). The parameter 'market value' is the dependent variable in the model estimation.

a basement or not. The subtype of a property can be either a 'Condominium', 'Detached house' or 'Not a detached house'. The year of modernisation represents when the last major refurbishment took place. The use of the building describes the possible uses, whereby the characteristics are either 'Owner-occupied & Non-owner-occupied',[6] 'Owner-Occupied' or 'Non-owner-occupied'. The variable describes whether a property can be rented to a third party. The quality of the property is measured via a grade on a scale ranging from 1 (very poor) to 5 (very good). The general condition of the property is represented by a categorical variable with five different categories ranging from disastrous to very good.[7] The variable 'Basement condominium' measures whether an apartment has an extra cellar compartment or not, whereas the 'Basement' and 'No Basement' variables are only valid for detached and non-detached houses.

The features describing the micro-location of the properties are the latitude and longitude, the different regiotypes and the micro score. The regiotype is provided by Acxiom[8] and clusters Germany into ten different area types. In general, Acxiom defines four different spatial types: 'Central-Business-District', 'Agglomeration Area, "Urban Area" and "Rural Area". The last three can be divided further into three sub-categories each ("Upper Centres", "Middle Centres" and "Commuter Belt"). All addresses in Germany can be allocated to one of the ten area types. The individual area types are determined according to the respective settlement structure and population density within the municipality and its surrounding area. The micro score of a location is calculated via a gravity model and reflects the accessibility in the sense of proximity to selected everyday destinations. A more detailed description of the construction of the micro score of a location can be found in Appendix II. In addition, the two socio-economic variables, "unemployment ratio" and "Time-on-Market", are included to represent the properties' macro-location. All are available at the ZIP code level.

The spatial breakdown of our dataset is based on the NUTS nomenclature of the European Union and is done by creating three new features, namely 'NUTS-1', 'NUTS-2' and 'NUTS-3'. The NUTS system was introduced by the European Union and is monitored by Eurostat. The system is used to provide a standardised system of territorial reference for the EU to make it easier to collect, compare and analyse statistics across different regions and countries. The NUTS system is, in general, divided into four levels, each with increasing geographical detail:

- NUTS-0: This level consists of larger regions that are typically based on the country or a group of countries
- NUTS-2: This level consists of basic regions for the application of regional policies within a NUTS-1 region
- NUTS-3: This level consists of small regions for specific diagnoses within a NUTS-2 region

The NUTS nomenclature is used for a wide range of purposes, including monitoring the progress of the EU's cohesion policy, as well as for other economic, social, and environmental statistics.[9]

Germany can generally be divided into a single NUTS-0, 16 NUTS-1, 38 NUTS-2 and 401 NUTS-3 regions. Since only a few observations were available in some NUTS-3 regions, we combined these regions and ended up with 327 NUTS-3 regions for our analysis. Figure 2 provides an overview of the different NUTS regions and the number of observations available for the specific regions. Analysing the NUTS-3 level shows most observations are located around the most significant German metropolitan areas like Berlin, Hamburg and Munich. In addition, the NUTS-2 and NUTS-1 levels indicate that a difference can be observed between west and east Germany, with the east tending to have fewer observations. This is consistent with the widely diverging population figures between these regions. Just and Schäfer (2017) provide a comprehensive introduction to

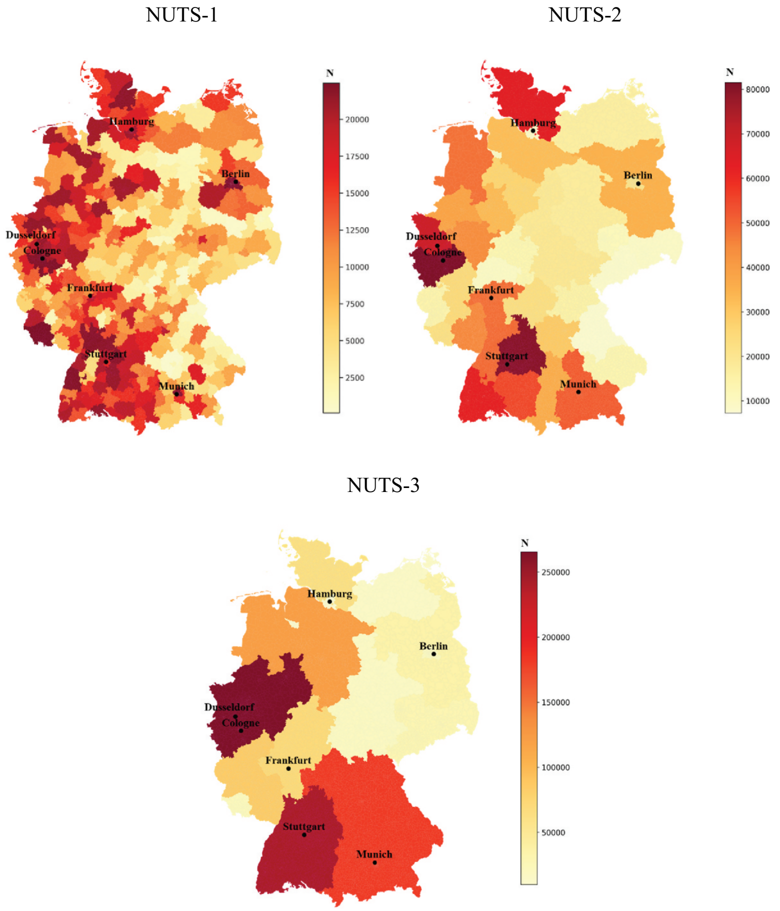

Figure 2. Number of observations per NUTS region. Notes: This figure highlights the observations available for the individual NUTS regions. Fewer observations are available in the eastern part of Germany. This can be explained by the generally lower market activity in these regions. Structurally, these regions are primarily rural and characterised by high out-migration and vacancies. Therefore, the data distribution is not a dataset-specific distortion but a representative reflection of the German residential real estate market.

the structure of the German regions. Just and Maennig (2012) give a more detailed overview of the German real estate markets.

Methodology

Ordinary Least Square Regression – OLS

The first method applied is an Ordinary Least Square Regression (OLS). The main advantage of the OLS is that it is easy to understand and interpret. Therefore, it is the most commonly used machine learning method and is often considered a benchmark. The aim of an OLS is to explain a dependent variable y, with independent variables $x_1, \ldots, x_k$, a-priori unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ and an error term ε:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

for all observations with

$$\mu = E[y] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Thereby, the relationship between the dependent and independent variables is assumed to be linear in parameters, and the error terms ε are considered independent and to have a constant variance. For further information, we recommend Fahrmeir et al. (2013).

Several optimisations were performed to account for locational differences and to achieve the best model performance, including backward stepwise regression, interaction terms and variable transformations.

Generalized Additive Model – GAM

Our second method is a Generalised Additive Model (GAM). It is a further development of the OLS and is based essentially on the concept of the Generalised Linear Model. A monotonic link function $g(.)$ is used to model the relationship between the expected value μ of the dependent variable y and the independent variables $x_1, \ldots, x_k$. The main advantage of the GAM over the OLS is that unspecified, non-parametric smoothing functions $s_j, j \in \{1, \ldots, k\}$, of the covariates can be included in the model:

$$g(\mu) = \beta_o + s_1(x_1) + \ldots + s_k(x_k).$$

For a more extensive description of the GAM, we recommend Wood (2017).

Again, multiple model optimisations were carried out. In addition to the methods mentioned above, different penalised spline types like cubic and thin plane splines were considered. As in the OLS, these optimisations were implemented manually.

Extreme Gradient Boosting – XGBoost

The third method, an eXtreme Gradient Boosting (XGBoost) algorithm, is a tree-based ensemble learning method. Ensemble learning algorithms train many weak learners h_m, in our case, single decision trees, and combine them to form one strong learner h:

$$h(y|x) = \sum_{i=1}^{M} u_m h_m(y|x),$$

with u_m being used to weight the weak learners. M denotes the number of single trees, x is the features space and y the response variable. In boosting, the weak learners h_m are trained sequentially. The algorithm starts with one model and uses the errors made to improve the subsequent trees. In Gradient boosting, the gradient descent algorithm is used to add new trees to minimise the loss of the model. The eXtreme Gradient Boosting is a computationally effective and highly efficient version of Gradient Boosting. The advantage of XGBoost is that it can recognise very complex patterns within a large amount of data. However, it is unclear from the model structure why a certain result occurs. The eXtreme Gradient Boosting is a computationally effective and highly efficient version of Gradient Boosting. The XGBoost can automatically detect complex non-linearities or higher-order interactions within a large dataset, with fewer manual optimisations than the OLS and GAM. Hastie et al. (2001) provide a detailed description of tree-based methods, ensemble learning and gradient boosting.

Deep Neural Network – DNN

Lastly, we consider deep neural networks (DNN), a popular and performant machine learning technique. DNNs are designed from biological neural networks (Pham, 1970), like the human brain, and consist of multiple layers, which are typically densely connected. Each layer consists of numerous neurons, each processing the weighted output of all (hence the term dense) neurons of the previous layer, combined with a bias value, and applies a so-called activation function onto this linear combination. To capture this formally, consider a neuron in a given layer. Let n be the number of neurons in the previous layer. For $i \in \{1, \ldots, n\}$, let z_i be the output of the i-th neuron in the previous layer and let w_i be the according weight. Furthermore, let f be the activation function of the current neuron and b the bias term. Then, the output of the neuron is

$$f\left(b + \sum\nolimits_{i=1}^{n} z_i w_i\right).$$

A DNN then consists of multiple such neurons and layers.

To train a DNN for a specific task and data, the weights and biases are adapted. The data is passed through the DNN in batches in a forward-propagation step. A prediction is calculated, for each datum in a batch, and the predictions are evaluated regarding loss function. The weights and biases are then adjusted using gradient descent to minimise the loss function. After all the data is passed through the DNN once, we say one epoch has passed. After many epochs, the DNN is trained, and predictions for a new object can be obtained by passing the object through the DNN again.

Finding the right architecture of a DNN for the task at hand is an essential yet tedious task. We use the hyperparameter optimisation framework Optuna (Akiba et al., 2019) to find suitable architectures for each region. In particular, we allow Optuna to choose the number of layers, the number of neurons per layer and the activation function per layer. Furthermore, we allow Optuna to choose the dropout rate per layer, which controls how many neurons per layer are activated.

The advantages of deep neural networks are that they are very flexible and adapt automatically to all data. Therefore, they can capture complex non-linearities and higher-order interactions by themselves. Besides that, compared to other modern machine learning approaches, deep neural networks require less computation power to produce reliable results. For more information about DNNs, see Goodfellow et al. (2016).

Testing concept

An extending window approach is implemented according to Mayer et al. (2019) to evaluate the predictive performance of the models. Figure 3 illustrates the testing concept.

The first iteration divides the dataset into a training set with observations from Q1/2014 to Q4/2019 and a test set from Q1/2020. In the following steps, the data of the tested quarter is added to the training set, and the models are retrained and tested on data from the next quarter. The advantages of this approach are that all algorithms are tested on unseen data and thus produce unbiased, robust results. Furthermore, the testing approach provides a realistic testing scenario. Table 3 presents the number of training and test observations for each iteration.[10]

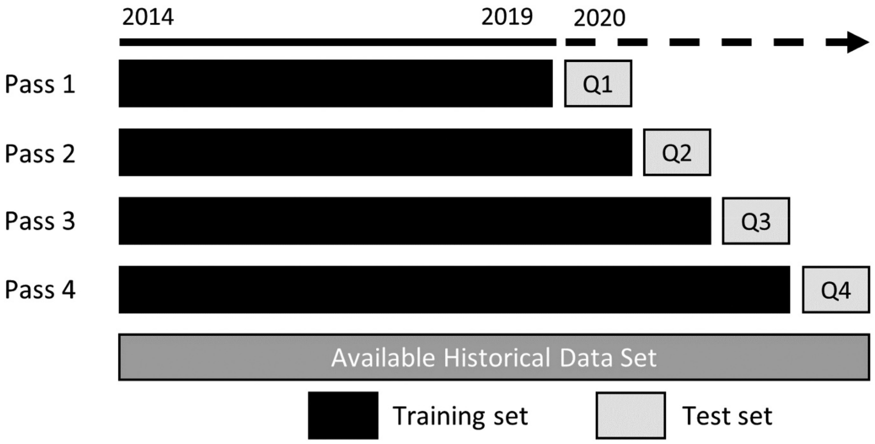

Figure 3. Extending window approach. Notes: This figure visualises the applied extending window approach-testing strategy. The strategy is the right choice for this study, as it best reflects the test procedure of conventional AVM providers. AVM providers usually update their models quarterly basis.

Table 3. Training and test observations.

Data split	Q1	Q2	Q3	Q4
Training	1,063,426	1,106,866	1,141,612	1,180,741
Test	43,440	34,746	39,129	31,805

Notes: This table shows the number of training and test observations over the four quarters of 2020. The number of training data increases over the quarters by the number of test data from the previous quarter. With regard to the test data, it can be seen in particular that fewer observations are available in Q2 and Q4. This can be attributed to COVID restrictions which made it difficult to conduct assessment visits, especially shortly after the pandemic outbreak (Q2) and during the winter (Q4).

Evaluation metrics

We compute the Mean Absolute Percentage Error (MAPE) and the Median Absolute Percentage Error (MdAPE) as accuracy measures for each model. Unlike Mayer et al. (2019), we use the relative rather than the absolute error measures to better compare the different spatial levels. To obtain an overall picture of the strength and weaknesses of the algorithms, we additionally provide the proportion of predictions within 10 and 20 per cent (PE(x)) following Cajias et al. (2019) and Stang et al. (2022). A detailed description of all metrics can be found in Table 4.

Table 4. Evaluation metrics.

Error	Formula	Description
Mean Absolute Percentage Error (MAPE)	$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$	Mean of all absolute percentage errors. A lower MAPE signals higher overall prediction accuracy in percent.
Median Absolute Percentage Error (MdAPE)	$MdAPE(y,\hat{y}) = median\left(\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert\right)$	Median of all absolute percentage errors. A lower MdAPE denotes a higher precision in percent without being sensitive to outliers.
Error buckets (PE(x))	$PE(x) = 100\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert < x$	Percentage of predictions where the relative deviation is less than $x\%$, with x being 10 and 20. A larger PE(x) signals a lower variation in the predictions.

Notes: This table reports the evaluation metrics used to determine the valuation accuracy of the different algorithms. All four metrics are regularly used to assess the quality of AVMS. The choice of several metrics in total allows a more differentiated statement to be made than would be the case with just one metric.

Results

Our study aims to find out whether the choice of spatial level for training statistical models has an influence on their performance, whether this influence is the same for all methods, or whether there are differences between more traditional and modern ML methods. In contrast to other publications, the main focus is not on which method performs best overall but on an intra-method comparison to determine which spatial level seems best suited for which method. This enables finding out whether the assumed local heterogeneity of real estate markets is also reflected in the results of the valuation methods or whether greater valuation accuracy can be achieved by adding further observations from other submarkets. For this purpose, two traditional approaches (OLS & GAM) as well as two modern ML approaches (XGBoost & DNN) are each trained for different spatial levels (NUTS-0, NUTS-1, NUTS-2, NUTS-3).

Below, we show the results for all four methods. To achieve comparability and to be able to make a valid statement, we evaluate the results on an aggregated level. For this purpose, we first provide a table for each method that shows the individual evaluation metrics for the four spatial levels of all test observations. For the metrics in the 'NUTS-3' row, for example, all test data is predicted with the different models calculated at the NUTS-3 level. Finally, the metrics are calculated for the nationwide aggregated residuals. For the other three levels, the procedure is then the same.

Furthermore, four maps are shown for each method. The maps are a cartographic representation of the results of the MAPE on a NUTS-3 level from the tables presented earlier. The representation allows for more detailed interpretations concerning regional performance. For example, it allows us to determine whether the results differ across different regions and whether general data availability plays a role.

Results of the ordinary least squares regression

The OLS results presented in Table 5 yield a clear pattern: The smaller the spatial level, the better the performance. Regarding the MAPE, the NUTS-3 models, which divide Germany into 327 submarkets, are more than three percentage points better than the NUTS-0 model, which considers Germany one overall market. In relative terms, this represents a performance increase of 18.0%. The PE-ratio also shows that the NUTS-3 models are far superior to the NUTS-0 model.

he cartographic representation in Figure 4 illustrates the results from Table 5. It can be seen that the lower the spatial training level, the better the MAPE for each region. The maps further show that the increased performance at the aggregate level can be attributed to improved performance in the eastern parts of Germany. In addition, the German North Sea Island group around Sylt stands out on the top left of the maps. Here, it can be seen that the performance in the NUTS-3 models is much better than in the NUTS-0 model. Very distinct peculiarities characterise the real estate market on Sylt and the surrounding islands. Residential properties are traded there only at top prices, and there is a strong dependency between the property's specific location and its value.

In summary, the OLS can only capture local effects of the German residential real estate market when trained on a small spatial level. Therefore, it is advisable to use the smallest possible spatial level, in our case NUTS-3, for training the OLS. These results also make sense in theory since the OLS is generalising in its structure and, therefore, can hardly (or not at all) take into account the local characteristics of individual regions if training is done on a global level. For the NUTS-0 model, the coefficients of the OLS are smoothed by too many individual and inconsistent regional effects, leading to a significant deterioration in performance. In the case of an OLS, it should always be ensured that only regional data are used to determine the coefficients and, ideally, that different submarkets are delimited from one another in advance.

Table 5. OLS – model prediction errors of the year 2020 throughout Germany.

Models	MAPE	MdAPE	PE(10)	PE(20)
$\text{OLS}_{\text{NUTS-0}}$	0.2023	0.1521	0.3423	0.6236
$\text{OLS}_{\text{NUTS-1}}$	0.1914	0.1454	0.3577	0.6473
$\text{OLS}_{\text{NUTS-2}}$	0.1852	0.1407	0.3688	0.6612
$\text{OLS}_{\text{NUTS-3}}$	0.1714	0.1294	0.3985	0.7004

Notes: This table reports the model prediction errors for the OLS. The results are evident across all metrics and show that model performance improves with a decreasing spatial training level. This result confirms the correctness of the proceeding that in parametric approaches, a data selection that is as granular as possible must be conducted in each case.

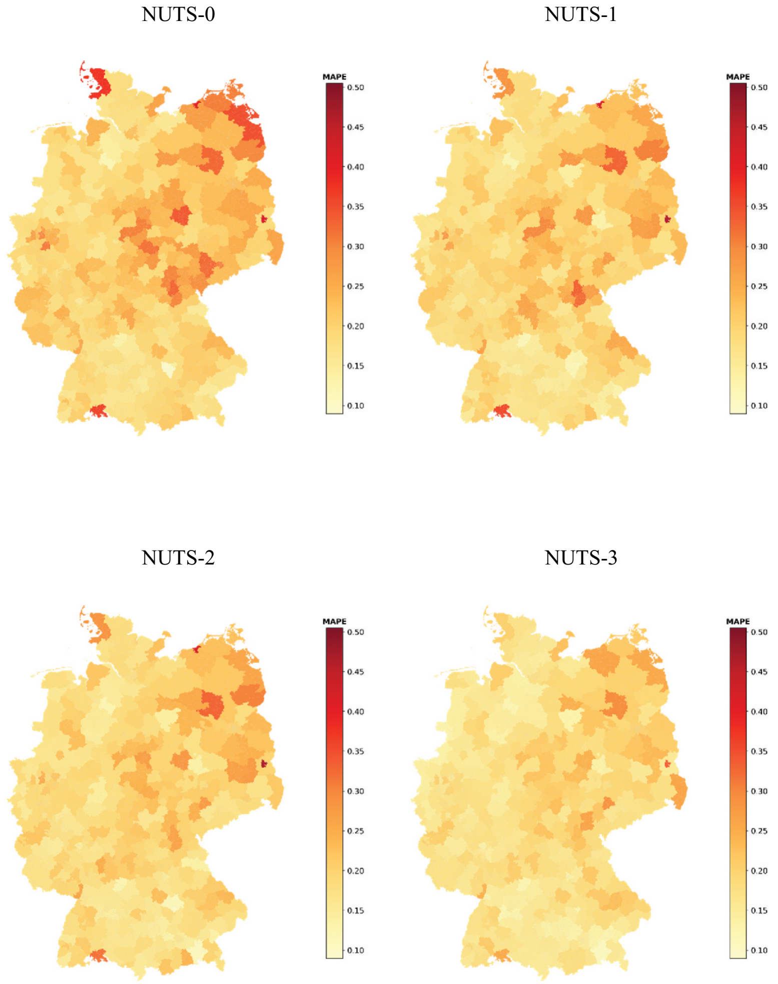

Figure 4. MAPE of the different OLS models. Notes: This figure visualises the MAPE of the four different OLS models. The maps show the average absolute percentage error obtained when applying the individual models within a given region. For a granular representation, the 327 NUTS-3 regions were selected as the corresponding levels of representation. The representation of the scale is chosen so that the minimum and maximum are the largest and smallest errors, respectively, of all four methods.

Results of the generalised additive model

The results for the GAM, shown in Table 6, are also clear and similar to those for the OLS. The more granular the spatial level for training, the better the estimation accuracy. This is true for all four evaluation metrics used. This time, the MAPE at the NUTS-3 level is 23.8% better than the NUTS-0 model. If we look at the general performance of the GAM and compare it with the results of the OLS, we see that the GAM is generally able to correctly estimate the market values of the properties better. The use of non-linear functions, which characterises the GAM, results in a performance boost. However, it is interesting to note that this effect only comes into play at a granular level. While the relative difference between the MAPEs of the

Table 6. GAM – model prediction errors of the year 2020 throughout Germany.

Models	MAPE	MdAPE	PE(10)	PE(20)
$\mathrm{GAM}_{\text{NUTS-0}}$	0.1971	0.1423	0.3641	0.6504
$\mathrm{GAM}_{\text{NUTS-1}}$	0.1832	0.1339	0.3852	0.6800
$\mathrm{GAM}_{\text{NUTS-2}}$	0.1734	0.1273	0.4044	0.7028
$\mathrm{GAM}_{\text{NUTS-3}}$	0.1592	0.1160	0.4398	0.7426

Notes: This table reports the model prediction errors for the GAM. The results are also clear across all metrics and similar to the results of the OLS. They show that model performance improves with a decreasing spatial training level. Again, the implication is that the smallest spatial level should be chosen to achieve the best model performance.

NUTS-0 models of the OLS and the GAM is only 2.6%, it increases continuously and amounts to 7.7% at the level of the NUTS-3 models.

The cartographic representation in Figure 5 shows the same picture as the OLS. Once again, it is noticeable that the estimation accuracy in the eastern part of Germany can be

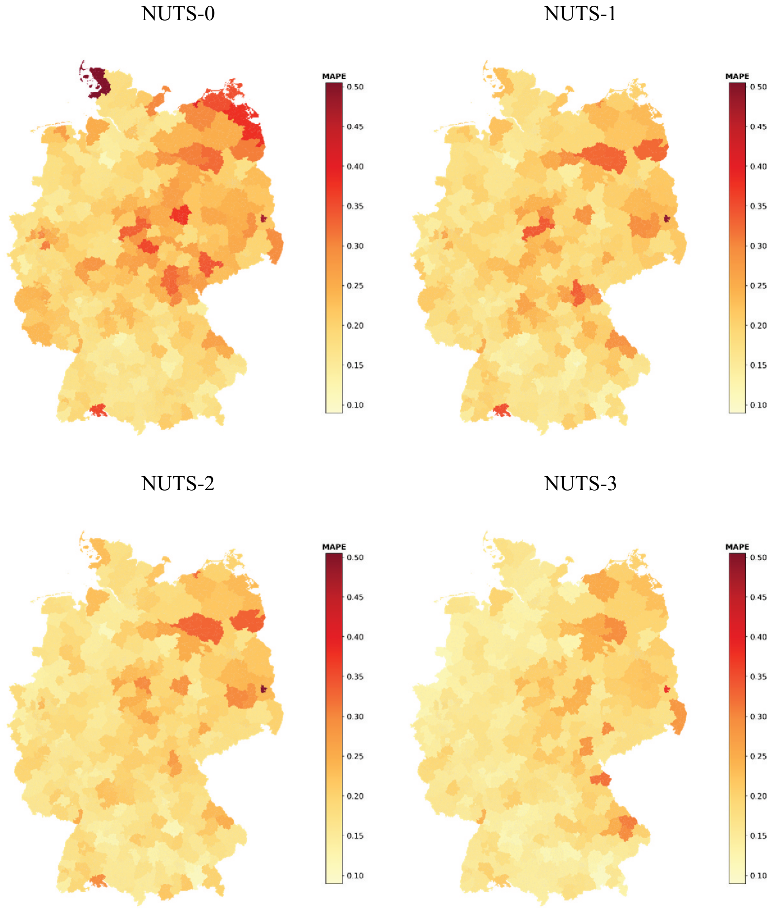

Figure 5. MAPE of the different GAM models. Notes: This figure depicts the MAPE of the four different GAM models. The maps show the average absolute percentage error obtained when applying the individual models within a given region. For a granular representation, the 327 NUTS-3 regions were selected as the corresponding levels of representation. The representation of the scale is chosen so that the minimum and maximum are the largest and smallest errors, respectively, of all four methods.

improved by implementing the method on a granular level. Furthermore, the group of islands around Sylt stands out again. It implies that the smaller the spatial level for training the model, the better the performance.

In summary, the feedback for the GAM is the same as that for the OLS. On a higher spatial level, the GAM does not capture the complexity and heterogeneity of the individual residential real estate markets in a single model as accurately as on a granular level. Therefore, when using a GAM for estimating residential property values, the smallest possible level should be used for training.

Results of eXtreme gradient boosting

Compared to the first two methods, the results of the XGBoost yield a different picture. The evaluation metrics from Table 7 show that the performance is similar on all four NUTS levels, and the greatest accuracy is achieved this time on the NUTS-1 level and not, as with the OLS and the GAM, on the NUTS-3 level. This is interesting because, as shown in the literature review, most academic studies on ML algorithms focus on the NUTS-3 level. This spatial level yields the worst performance in our case. Relative to the NUTS-1 level, the NUTS-3 level based on MAPE is 2.9% worse regarding valuation accuracy. Although the differences between the individual metrics are only minor in absolute values, if these are considered in relative terms, then a small performance boost is shown by the correct choice of the spatial level.

The analysis of the maps from Figure 6 shows, in particular, that in the parts of Germany where few observations are available (see Figure 2), the choice of a higher spatial level for training the models leads to a performance improvement. It is an important implication that in regions where little data are available, it can be useful in the case of the XGBoost to include data from other surrounding districts. This represents an essential difference between the results of the GAM and the OLS. For them, especially in parts of Germany with low data availability, the results deteriorate with a higher spatial level for training the models.

In summary, the heterogeneity of local real estate markets can still be detected by the XGBoost when trained at a higher spatial level, as the XGBoost can combine several local

Table 7. Xgboost – model prediction errors of the year 2020 throughout Germany.

Models	MAPE	MdAPE	PE(10)	PE(20)
$XGB_{NUTS\text{-}0}$	0.1426	0.1077	0.4693	0.7780
$XGB_{NUTS\text{-}1}$	0.1402	0.1064	0.4739	0.7869
$XGB_{NUTS\text{-}2}$	0.1407	0.1071	0.4719	0.7850
$XGB_{NUTS\text{-}3}$	0.1442	0.1107	0.4578	0.7733

Notes: This table reports the model prediction errors for the XGBoost. Here, too, the results are the same across all evaluation metrics. Unlike the first two methods, however, the model performance of the XGBoost does not improve with a decreasing spatial training level but is relatively constant across all levels. The best performance is achieved at the NUTS-1 level, indicating that the XGBoost can gain more explanatory power by adding more data.

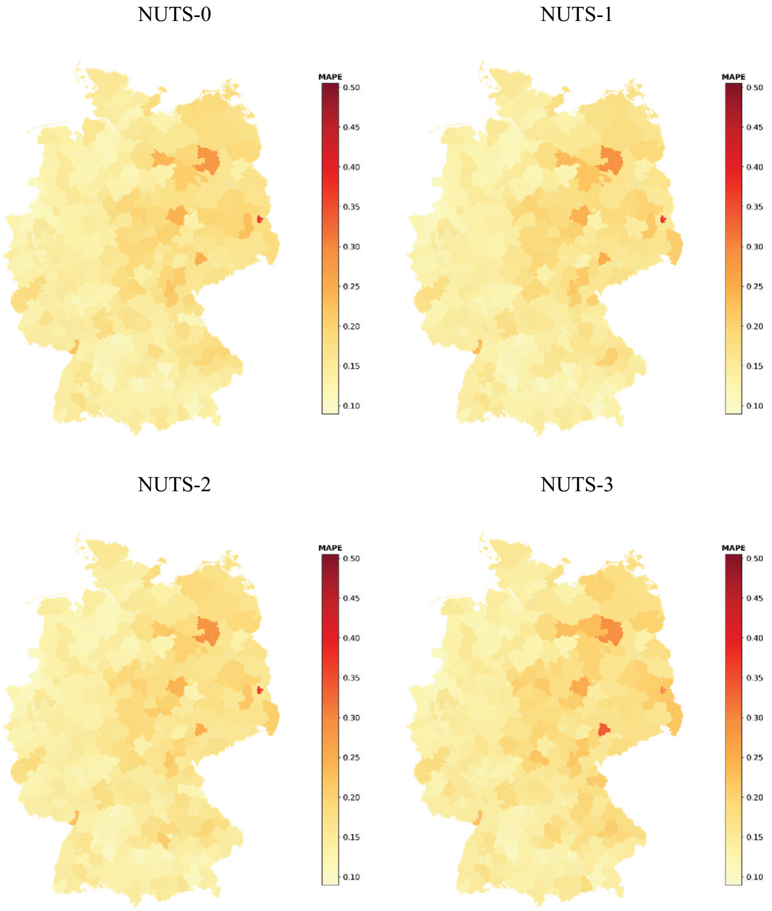

Figure 6. MAPE of the different XGB models. Notes: This figure visualises the MAPE of the four different XGBoost models. The maps show the average absolute percentage error obtained when applying the individual models within a given region. For a granular representation, the 327 NUTS-3 regions were selected as the corresponding levels of representation. The representation of the scale is chosen so that the minimum and maximum are the largest and smallest errors, respectively, of all four methods.

models into one large global model. In some cases, the use of additional data even leads to a further improvement of the estimation accuracy as more nuanced relationships can be learned, and the risk of overfitting decreases. Therefore, unlike for the OLS and the GAM, the NUTS-3 level is not the optimal spatial level for training the XGBoost, but the NUTS-1 level. However, the results of Table 7 also show that there seems to be a limit regarding the optimal size of the spatial level. The results at NUTS-0 level are still better than those at NUTS-3 level but not as good as on NUTS-1 and NUTS-2 level.

Results of the neural network

Finally, in analysing the results of the DNN, we again see a different picture. The evaluation metrics presented in Table 8 show that the DNN can improve its

Table 8. DNN – model prediction errors of the year 2020 throughout Germany.

Models	MAPE	MdAPE	PE(10)	PE(20)
$DNN_{NUTS\text{-}0}$	0.1551	0.1080	0.4700	0.7620
$DNN_{NUTS\text{-}1}$	0.1542	0.1090	0.4648	0.7595
$DNN_{NUTS\text{-}2}$	0.1595	0.1142	0.4471	0.7448
$DNN_{NUTS\text{-}3}$	0.1656	0.1176	0.4356	0.7281

Notes: This table reports the model prediction errors for the DNN. The results show that model performance can be increased by choosing the highest possible spatial training level. Unlike the XGBoost results, the increase in performance is much more significant. The results indicate that the DNN can only show its strength with a certain amount of data.

valuation accuracy as the spatial training level increases. This is the exact opposite of the OLS and GAM results and a different result than the XGBoost. Although the results of the MAPE indicate that the NUTS-1 level performs best here as well, the three other metrics yield a slightly different picture for this specific algorithm. They evaluate the NUTS-0 level as the best suited. In principle, therefore, the situation between the NUTS-0 and NUTS-1 levels is quite similar, influenced only by marginal changes. Compared to the other modern ML algorithm, the XGBoost, the number of observations used to optimise the algorithm seems more important. This is also logical from the point of view of the complexity of the method. The DNN can only show its strength in recognising non-linear relationships and multi-layer interactions if a sufficiently large number of observations is available. This finding is also in line with those of Nghiep and Al (2001), which show, based on a dataset for Rutherford County, Tennessee, that neural networks perform better than multiple regression analysis only with increasing dataset size.[11]

The visual representation of the results in Figure 7 yields a similar picture to the XGBoost results. By choosing a higher training level for the DNN, the valuation performance can be increased, especially in areas with few observations. Again, the same implication emerges as with the XGBoost: in regions where little data is available, including data from other surrounding districts can be helpful. Concerning the four algorithms used for the analysis, this can only be empirically proven for the modern ML algorithms, which represents a significant contribution to the literature of this study.

In summary, the DNN can only estimate property values as accurately as possible once a certain number of observations has been used. Adding more observations, therefore, outweighs the effect of local heterogeneity. Thus, the DNN can independently generate additional explanatory power for a specific real estate market, even from data outside the specific market. Regional effects can therefore be more effectively detected, extracted and extrapolated by modern ML algorithms. The reason could be that neural networks are designed to handle large data sets and can benefit from more data by learning more nuanced relationships. Besides that, larger data sets help to reduce the variance of the neural networks' predictions. This variance reduction helps prevent overfitting, making the model more robust and accurate. Concerning the DNN, it is advisable to choose as high as possible a training level or to maximise the available observations for training the algorithm.

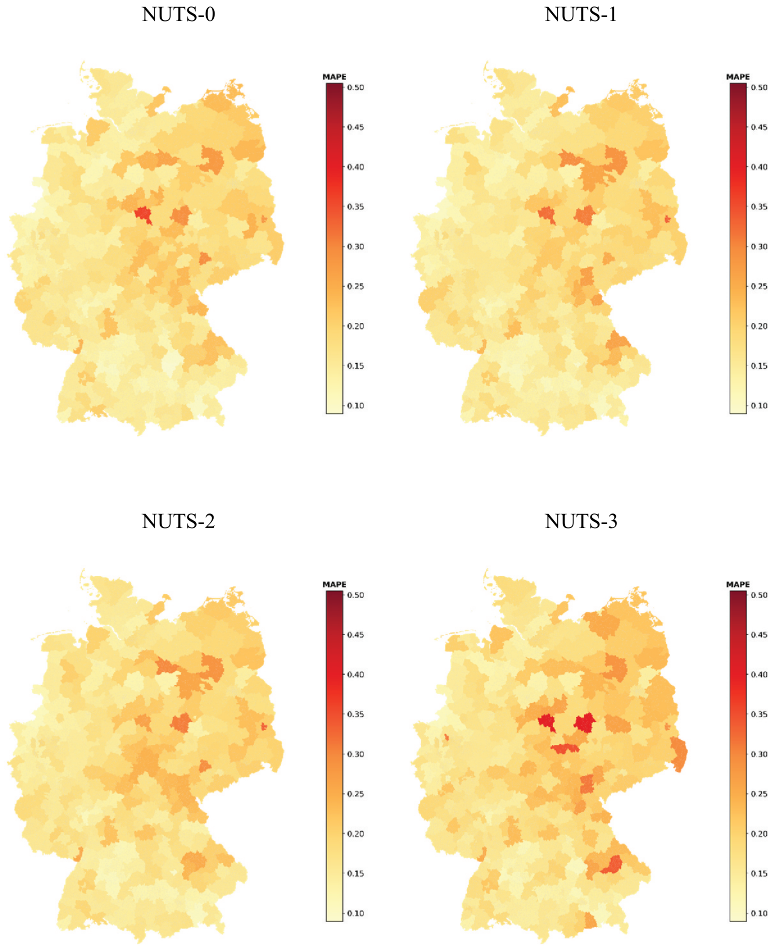

Figure 7. MAPE of the different DNN models. Notes: This figure visualises the MAPE of the four different DNN models. The maps show the average absolute percentage error obtained when applying the individual models within a given region. For a granular representation, the 327 NUTS-3 regions were selected as the corresponding levels of representation. The representation of the scale is chosen so that the minimum and maximum are the largest and smallest errors, respectively, of all four methods.

Conclusion

This study is intended to answer whether the right choice of the appropriate spatial level for training AVM algorithms also plays an important and underestimated role in improving the valuation accuracy of AVMs. We use a dataset of 1.2 million residential properties across Germany to test our hypotheses for four different typical AVM algorithms (OLS, GAM, XGBoost, DNN). All four are each trained on four different spatial levels, after which the results are evaluated. The four spatial levels are based on the NUTS nomenclature of the European Union. We use the NUTS-0, NUTS-1, NUTS-2 and NUTS-3 levels to train our models on a country, state, cross-regional, and county level, respectively.

Our results indicate that the correct choice of spatial training level can significantly influence the model performance, and that this can vary considerably, depending on the type of method. Concerning the OLS results, selecting a training level that is as granular as possible is the only way to ensure that the most accurate valuations are attained. There are regional differences and, thus, certain heterogeneities, which the OLS can only recognise as accurately as possible if they are locally limited.

The results for the GAM yield a similar picture to the OLS. The model performance correlates positively with a smaller spatial training level. Accordingly, the same findings can be generated for the parametric and the semi-parametric approaches. These confirm the correctness of the trend in academic publications and in practice of choosing the most granular analysis level possible for traditional econometric methods. These two methods cannot draw additional explanatory power from observations that lie outside a region. On the contrary, they even suffer from it.

The results of the two applied modern ML algorithms are quite different. Concerning the XGBoost, the evaluation metrics show that the choice of the most suitable spatial level can be made with relative indifference. Although there are marginal differences concerning the evaluation accuracy, these are only minor compared to OLS and GAM. In contrast to the parametric and semi-parametric approaches, the non-parametric XGBoost shows that the performance increases slightly with increasing spatial training levels. The NUTS-1 level seems the most appropriate level. This trend can be observed even more clearly in the results of the DNN. Here, it can be seen that the performance does not decrease with an increasing training level, as is the case with the OLS and the GAM, but it improves.

Concerning the two modern ML algorithms, they can gain a higher degree of explanatory power by adding further observations, and this effect outweighs that of local heterogeneity. In particular, their ability to recognise and map non-linear relationships and multi-layered interactions allows them to exploit overlapping effects of different regions to achieve more accurate real estate valuations. This is particularly evident in regions where there are few observations. In these cases, training a modern ML algorithm with additional regions is advisable to benefit from their basic commonalities.

In summary, the right training level should always depend on the method. For parametric and semi-parametric methods we recommend using a spatial level that is as granular as possible for training the models, since these can only separate local heterogeneities from each other to a limited extent. For non-parametric modern ML methods, however, we generally recommend a higher training level. These complex methods can detect regional differences independently and separate them. Furthermore, they benefit from the fact that there are basic commonalities in the functioning of local real estate markets, which can be used to increase their explanatory power. Concerning the practical application and implementation of AVM algorithms, this offers the additional advantage that the higher training level means fewer models must be trained and calibrated overall. For example, less effort is required for data preparation and processing. Thus, efficiencies can be increased for AVM providers operating nationwide, and significant economic advantages can be achieved.

Our findings empower real estate researchers to make more informed decisions about the appropriate spatial level when using and analysing different machine learning algorithms. As such, the main contribution of this paper is to update the standard guidelines for applying both traditional econometric and modern ML algorithms and setting new guidelines. On top of that, the contributions of our paper are not limited to scientific purposes but also provide practitioners in the field of AVM application with a new set of guidelines that can help them to improve the accuracy of their AVMs and reduce their implementation efforts at the same time.

Notes

1. To avoid a structural break within the dataset, the data should ideally come from one source or have been collected according to the same criteria.
2. The models could not be analysed at an even smaller spatial level because of data availability.
3. The Top-7 are the most important cities in Germany, namely Berlin, Munich, Hamburg, Frankfurt, Cologne, Dusseldorf and Stuttgart. Their importance is based on their market size and market activity. They can be seen as the most liquid and dynamic real estate markets in Germany.
4. Table A1 in Appendix I explaines the individual variables.
5. The correlation matrix is available on request.
6. Applies if the property is both partly owner-occupied and partly non-owner-occupied (e.g. single-family home with an attached rental unit).
7. The assessment of the two variables, 'condition' and 'quality grade', was performed by professional appraisers during the property inspection process.
8. Acxiom is an American provider of international macroeconomic and microeconomic data. Further information can be found at: https://www.acxiom.com/.
9. Further information about the NUTS nomenclature can be found at https://ec.europa.eu/eurostat/web/nuts/background.
10. The train/test split was selected to include as much data as possible in the training set, as some NUTS-3 regions have limited data. If an alternative split were employed, it would result in inconsistent results for these regions. Our study focuses on a nationwide comparison, rather than individual metropolitan regions, which typically have a large and dense amount of data. This approach is consistent with other studies that compare algorithms on a national scale, such as Stang et al. (2022).
11. However, unlike in our study, these authors only work at a county level and only vary the data available within the county. In our case, the amount of data is varied by adding observations from other spatial levels.
12. https://www.openstreetmap.org/.
13. See https://wiki.openstreetmap.org/wiki/Map_features.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Bastian Krämer is doctoral student and project team member in the Real Estate Management group at the International Real Estate Business School (IRE|BS) of the University of Regensburg, Germany. Bastian holds a master's degree in mathematics from the Technical University of Munich, Germany. His research focuses on applications of machine learning in the field of real estate.

Moritz Stang is doctroal student and project team member in the Real Estate Management group at the International Real Estate Business School (IRE|BS) of the University of Regensburg,

Germany. Moritz holds a master's degree in Real Estate Management from the Real Estate Business School (IRE|BS) of the University of Regensburg. His research objective lays on Automated Valuation Models (AVMs) and the use of Big Data solutions in the real estate industry.

Vanja Doskoč is a doctoral student and project team member in the Algorithm Engineering group of the Hasso Plattner Institute (HPI). Vanja obtained a master's degree in technical mathematics from the TU Wien (Vienna University of Technology), Austria. In his research he focuses on understanding behaviourally correct language learning under various restrictions as well as applications of deep learning and evolutionary algorithms on different real world problems.

Wolfgang Schäfers is Professor and Chair of the Department of Real Estate Management at the International Real Estate Business School (IRE|BS) of the University of Regensburg. His research focuses on real estate valuation, applications of machine learning and sentiment analysis.

Tobias Friedrich is a professor at the University of Potsdam and the head of the Algorithm Engineering group of the Hasso Hasso Plattner Institute (HPI). His research interest lies in algorithm engineering, probabilistic methods, artificial intelligence, data science, network science, distributed algorithms and graph algorithms.

ORCID

Moritz Stang http://orcid.org/0000-0003-4452-3148

Data availability statement

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631).

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

Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences*, *8*(11), 2321. https://doi.org/10.3390/app8112321

Bankers Association, M. (2019). *The State of Automated Valuation Models in the Age of Big Data*. Washington DC.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2008). Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions. *Social Science Research Network Electronic Journal*. https://doi.org/10.2139/ssrn.1090147

Bunke, O., Droge, B., & Polzehl, J. (1999). Model Selection, Transformations and Variance Estimation in Nonlinear Regression. *Statistics*, *33*(3), 197–240. https://doi.org/10.1080/02331889908802692

Cajias, M., Willwersch, J., & Lorenz, F. (2019). *I know where you will invest in the next year – Forecasting real estate investments with machine learning methods*. European Real Estate Society (ERES). ERES. https://ideas.repec.org/p/arz/wpaper/eres2019_171.html

Chau, K. W., & Chin, T. L. (2002). A Critical Reveiw of the Literature on the Hedonic Pricing Model and Its aplication to the Housing Market in Penang. In *Proceedings of the The Seventh Asian Real Estate Society Conference*.

Chun Lin, C., & Mohan, S. B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis*, *4*(3), 224–243. https://doi.org/10.1108/17538271111153013

Dąbrowski, J., & Adamczyk, T. (2010). Application of GAM additive non-linear models to estimate real estate market value. *Geomatics and Environmental Engineering*, *4*(2), 55–62.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer-Verlag.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Handy, S. L., & Clifton, K. J. (2001). Evaluating Neighborhood Accessibility: Possibilities and Practicalities. *Journal of Transportation and Statistics*, *4*(2), 67–78.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Hong, J., Choi, H., & Kim, W. (2020). A House Price Valuation based on Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544

Huang, Y., & Dall'erba, S. (2021). Does Proximity to School Still Matter Once Access to Your Preferred School Zone Has Already Been Secured? *The Journal of Real Estate Finance and Economics*, *62*(4), 548–577. https://doi.org/10.1007/s11146-020-09761-w

Just, T., & Maennig, W. (Eds.). (2012). *Understanding German Real Estate Markets*. Springer. https://doi.org/10.1007/978-3-642-23611-2

Just, T., & Schaefer, P. (2017). Germany's Regional Structure. *Understanding German Real Estate Markets*, 41–57. https://doi.org/10.1007/978-3-642-23611-2

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*, *43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *The Journal of Political Economy*, *74*(2), 132–157. https://doi.org/10.1086/259131

Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. *Housing Economics and Public Policy*, 67–89. Original work published 2003. https://doi.org/10.1002/9780470690680.ch5

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

McCluskey, W., Davis, P., Haran, M. [., McCord, M. [., & McIlhatton, D. [. (2012). The potential of artificial neural networks in mass appraisal: The case revisited. *Journal of Financial Management of Property and Construction*, *17*(3), 274–292. https://doi.org/10.1108/13664381211274371

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*, *30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Metzner, S., & Kindt, A. (2018). Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties. *International Journal of Housing Markets and Analysis*, *11*(1), 73–100. https://doi.org/10.1108/IJHMA-02-2017-0018

Nghiep, N., & Al, C. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research*, *22*(3), 313–336. https://doi.org/10.1080/10835547.2001.12091068

Nobis, C., & Kuhnimhof, T. (2018). *Mobilität in Deutschland – MiD: Ergebnisbericht, Bonn*.

Osland, L. (2010). An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*, *32*(3), 289–320. https://doi.org/10.1080/10835547.2010.12091282

Pace, R. K. (1998). Appraisal Using Generalized Additive Models. *Journal of Real Estate Research*, *15*(1), 77–99. https://doi.org/10.1080/10835547.1998.12090916

Pace, R. K., & Hayunga, D. (2020). Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests. *The Journal of Real Estate Finance and Economics*, *60*(1–2), 170–180. https://doi.org/10.1007/s11146-019-09724-w

Pace, R. K., & LeSage, J. (2004). Spatial Statistics and Real Estate. *The Journal of Real Estate Finance and Economics*, *29*(2), 147–148. https://doi.org/10.1023/b:real.0000035307.99686.fb

Páez, A., Long, F., & Farber, S. (2008). Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques. *Urban Studies*, *45*(8), 1565–1581. https://doi.org/10.1177/0042098008091491

Pham, D. T. (1970). Neural Networks in Engineering. *WIT Transactions on Information and Communication Technologies*, *6*, 3–36. https://doi.org/10.2495/AI940011

Powe, N. A., Garrod, G. D., & Willis, K. G. (1995). Valuation of urban amenities using an hedonic price model. *Journal of Property Research*, *12*(2), 137–147. https://doi.org/10.1080/09599919508724137

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *The Journal of Political Economy*, *82*(1), 34–55. https://doi.org/10.1086/260169

Schulz, R., Wersing, M., & Werwatz, A. (2014). Automated valuation modelling: A specification exercise. *Journal of Property Research*, *31*(2), 131–153. https://doi.org/10.1080/09599916.2013.846930

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), 1–44. https://doi.org/10.1080/10835547.2005.12090154

Stang, M., Krämer, B., Nagl, C., & Schäfers, W. (2022). From human business to machine learning —Methods for automating real estate appraisals and their practical implications. *Zeitschrift Für Immobilienökonomie*, 1–28. https://doi.org/10.1365/s41056-022-00063-1

Tse, R. Y. C. (2002). Estimating Neighbourhood Effects in House Prices: Towards a New Hedonic Model Approach. *Urban Studies*, *39*(7), 1165–1180. https://doi.org/10.1080/00420980220135545

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

Yang, J., Bao, Y., Zhang, Y., Li, X., & Ge, Q. (2018). Impact of Accessibility on Housing Prices in Dalian City of China Based on a Geographically Weighted Regression Model. *Chinese Geographical Science*, *28*(3), 505–515. https://doi.org/10.1007/s11769-018-0954-6

Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, *22*(2), 561–581. https://doi.org/10.1111/tgis.12330

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, *99*, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

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

Zurada, J., Levitan, A., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, *33*(3), 349–388. https://doi.org/10.1080/10835547.2011.12091311

Appendix

Appendix I - Table of explanatory variables

Table A1. Feature description.

Variable	Description
Market value	Market value of the property determined by appraiser.
Modernisation year	Year of the last major refurbishment.
Year of construction	Year in which the property was built.
Year of valuation	Year in which the property was assessed.
Quarter of valuation	Quarter in which the property was assessed.
Quality grade	Grade concerning the quality of the property ranging from 1 (very poor) to 5 (very good).
Living area	Size of the property in square meters.
Lot size	Size of the property plot in square meters.
Longitude	Longitude of the property.
Latitude	Latitude of the property.
Micro score	Rates the quality of the micro location.
Unemployment ratio	Variable that describes the unemployment ratio on a zip code level.
Time on market	Measurement of the length of real estate listings in weeks at the zip code level.
Basement	Variable that describes whether the property has a basement or not.
Owner-occupied	Variable that describes whether the property is rented or owner-occupied.
Object subtype	Variable that describes whether the property is a condominium, a detached single-family house, or a townhouse.
Condition	Variable that describes the general condition of the property.
Regiotype	Variable that describes the type of area in which the property is located.
NUTS	Variable that specifies the NUTS 1/2/3 region associated with the property.

Note: The table provides an overview of the variables used.

Appendix II – Micro Score

The micro score of a location is calculated via a gravity model and reflects the accessibility in the sense of proximity to selected everyday destinations. A gravity model is a standard method for approximating the accessibility of a location and is based on the assumption that nearby destinations play a more significant role in everyday life than more distant destinations (Handy and Clifton (2001). The relevant points-of-interest (POIs) are selected from the findings of Powe et al. (1995), Metzner and Kindt (2018), Yang et al. (2018), Nobis and Kuhnimhof (2018) and Huang and Dall'erba (2021) and are provided in Table A2.

Our gravity model can be described using an activity function $f(A_p)$ and a distance function $f(D_{i,p})$:

$$A_{i,p} = \sum f(A_p) f(D_{i,p}).$$

Here, $A_{i,p} \in [0, 100]$ denotes the accessibility of point i for the POI p, whereby the activity function $f(A_p)$specifies the relative importance of POI p, with $f(A_p) \in [0, 1]$. The function $f(D_{i,p})$ measures the travel time from point i to the POI p by using a non-symmetric sigmoidal distance function. The travel time was obtained for the selected POIs via Open Street Map[12] and normalised using the following function:

$$L(x) = \frac{K}{(1 + Qe^{0.5x})^{\frac{1}{v}}},$$

Table A2. Features of the micro score of a location.

Points-of-Interests	Category	Description
University	Education & Work	University campus: institute of higher education
School	Education & Work	Facility for education
Kindergarten	Education & Work	Facility for early childhood care
CBD	Education & Work	Centre of the next city
Supermarket	Local Supply	Supermarket – a large shop with groceries
Marketplace	Local Supply	A marketplace where goods are traded daily or weekly
Chemist	Local Supply	Shop focused on selling articles for personal hygiene, cosmetics, and household cleaning products
Bakery	Local Supply	Place for fresh bakery items
ATM	Local Supply	ATM or cash point
Hospital	Local Supply	Facility providing in-patient medical treatment
Doctors	Local Supply	Doctor's practice/surgery
Pharmacy	Local Supply	Shop where a pharmacist sells medications
Restaurant	Leisure & Food	Facility to go out to eat
Café	Leisure & Food	Place that offers casual meals and beverages
Park	Leisure & Food	A park, usually urban (municipal)
Fitness Centre	Leisure & Food	Fitness Centre, health club or gym
Movie Theatre	Leisure & Food	Place where films are shown
Theatre	Leisure & Food	Theatre where live performances take place
Shopping Mall	Leisure & Food	Shopping Centre – multiple shops under one roof
Department Store	Leisure & Food	Single large shop selling a large variety of goods
Subway Station	Transportation	City passenger rail service
Tram Station	Transportation	City passenger rail service
Railway Station	Transportation	Railway passenger only station
Bus Stop	Transportation	Bus stops of local bus lines
E-Charging Station	Transportation	Charging facility for electric vehicles

Note: The descriptions of the selected Points-of-Interest is based on the explanations of Open Street Map.[13].

where $K, Q \in \mathbb{R}$ and $v \in \mathbb{R}^+$ are defined for all possible distances $x \in \mathbb{R}$. Furthermore, we have:

$$K = (1 + Q)^{1+v},$$

$$Q = v \cdot \exp(B \cdot x^*),$$

$$v = \frac{\exp(B \cdot x^*) - 1}{\ln(y_i) - 1},$$

where x^* denotes a feature specific point of inflection and y^* is 0.5.