s:fi
RESEARCH

Swiss Finance Institute
Research Paper Series
N°18-76

Estimation and Updating Methods for Hedonic Valuation

Michael Mayer
Consult AG Bern

Steven C. Bourassa
Florida Atlantic University

Martin Hoesli
University of Geneva, University of Aberdeen, Kedge Business School, and Swiss Finance Institute

Donato Flavio Scognamiglio
University of Berne

Estimation and Updating Methods for Hedonic Valuation

Michael Mayer

Consult AG Bern, Zurich, Switzerland

Steven C. Bourassa

School of Urban and Regional Planning and School of Public Administration, Florida Atlantic University, Boca Raton, FL, USA

Martin Hoesli

Geneva Finance Research Institute and Swiss Finance Institute, University of Geneva, Geneva, Switzerland; Business School, University of Aberdeen, Aberdeen, UK; Kedge Business School, Bordeaux, France

Donato Scognamiglio

IAZI AG, Zurich, Switzerland; Faculty of Business, Economics and Social Sciences, University of Bern, Bern, Switzerland

Abstract

Purpose – We use a large and rich data set consisting of over 123,000 single-family houses sold in Switzerland between 2005 and 2017 to investigate the accuracy and volatility of different methods for estimating and updating hedonic valuation models.

Design/methodology/approach – We apply six estimation methods (linear least squares, robust regression, mixed effects regression, random forests, gradient boosting, and neural networks) and two updating methods (moving and extending windows).

Findings – The gradient boosting method yields the greatest accuracy while the robust method provides the least volatile predictions. There is a clear trade-off across methods depending on whether the goal is to improve accuracy or avoid volatility. The choice between moving and extending windows has only a modest effect on the results.

Originality/value – This paper compares a range of linear and machine learning techniques in the context of moving or extending window scenarios that are used in practice but which have not been considered in prior research. The techniques include robust regression, which has not previously been used in this context. The data updating allows for analysis of the volatility in addition to the accuracy of predictions. The results should prove useful in improving hedonic models used by property tax assessors, mortgage underwriters, valuation firms, and regulatory authorities.

Keywords Hedonic models, Appraisal accuracy, Appraisal volatility, Machine learning, Robust regression, Mixed effects models, Random forests, Gradient boosting, Neural networks

Paper type Research paper

JEL codes R31, C45, C53

1 Introduction

Hedonic models are widely used for property valuation purposes. They use information about a sample of properties that transacted to estimate models that are then used to predict the values of out-of-sample properties that did not transact. They are a valuable tool for property tax appraisers, mortgage underwriters, valuation firms, and regulatory authorities. Popular online resources, such as Zillow.com in the United States, rely on hedonic models to provide regularly updated estimates of property values that are accessible to the public. Here we explore two types of questions regarding the methods used to estimate the models used for prediction purposes.

The first question has to do with the method used to estimate the model. The standard approach is to estimate a linear model with ordinary least squares (OLS) regression. However, a variety of other techniques have been developed that offer some potential advantages over the standard approach. These include robust and mixed effects regression and various machine learning techniques, such as artificial neural networks, gradient boosting, and random forests. The second question has to do with the data used for estimation purposes. The typical approach in the house price prediction literature is to use one sample of data without taking into account the practical issue of updating over time. A more realistic approach would consider multiple samples that change as data are added for subsequent time periods. In this context, one strategy is to add new data as they become available while retaining all historical data; this is referred to as the *extending window* approach. The second strategy is to delete the oldest data when new data are added; this is the *moving window* approach. We compare the above-mentioned methods for estimating models using both extending and moving windows.

The most simple and common way to statistically model house prices is based on OLS regression of the (log) price on property characteristics and environmental variables assessing the quality of the property's location. Such hedonic models are described, for example, in Bourassa *et al*. (2003), Sirmans *et al*. (2005), Malpezzi (2008), and Schulz *et al*. (2014).

In order to deal with outliers, non-normality, and heteroscedasticity frequently seen in the data used to fit such models, different types of robust regressions have been found to

be useful in the context of hedonic modelling. We focus here on methods designed to address outliers and related data problems, such as in Peña and Ruiz-Castillo (1984) or Bourassa *et al*. (2016). To our knowledge, no previous research has applied robust techniques to the problem of out-of-sample house price prediction.

One important issue when modelling house prices is that of accurately measuring a property's location. In our case, location variables are measured at a relatively high level of aggregation, i.e., at the level of the municipality. In order to better account for spatial information in the data, the classical linear model can be extended to a hierarchical or multilevel (mixed effects) model by adding the municipality and possibly other higher-level administrative units as random intercepts in the model equation. Such models are, for example, applied by Brown and Uyar (2004), Ciuna *et al*. (2017), and Keskin *et al*. (2017) in the framework of hedonic price modelling. Numerous publications, such as Orford (2002), Goodman and Thibodeau (2003), Bourassa *et al*. (2003), Case *et al.* (2004), and Bourassa *et al*. (2007, 2010), use related approaches in the context of market segmentation.

Over the past several decades, with the advent of machine learning, modern regression techniques like artificial neural networks (Rumelhart *et al.*, 1986), random forests (Breiman, 2001) and gradient boosting (Friedman, 2000) have been introduced to the statistical community (Hastie *et al*., 2001; James *et al*., 2014; Efron and Hastie, 2016). If carefully applied, these modelling techniques can be more accurate than the standard approach because they automatically learn relevant transformations, nonlinearities, and high-order interactions among the predictor variables, although at the price of reduced interpretability. General applications of modern machine learning in econometrics are described in Varian (2014) and Mullainathan and Spiess (2017).

These modelling techniques are becoming more and more popular, including for house price modelling. Applications in this field include: Worzala *et al*. (1995), Din *et al*. (2001), Peterson and Flanagan (2009), Zurada *et al*. (2011), McCluskey *et al*. (2013), and Chiarazzo *et al*. (2014) for neural networks; Yoo *et al*. (2012) and Antipov and Pokryshevskaya (2012) for random forests; Kagie and Van Wezel (2007), Lu *et al*. (2017), Gu and Xu (2017), and Sangani *et al*. (2017) for boosting; and a vast selection of blog posts and contributions on the machine-learning competition platform *kaggle.com*. Most of the published research on

machine learning applications to house price prediction focuses on comparing one method, such as artificial neural networks, with the traditional OLS estimation. In a small number of cases, researchers have compared multiple machine learning techniques (Zurada *et al*., 2011; Antipov and Pokryshevskaya, 2012). In most but not all cases, researchers have concluded that machine learning techniques yield more accurate predictions than standard linear models. However, these methods have been criticized for their complexity and lack of transparency (see, e.g., Din *et al*., 2001; McCluskey *et al*., 2013).

The aim of this paper is to compare the precision of six methods (traditional linear regression, robust regression, mixed effects regression, gradient boosting, random forests, and neural networks) applied to both moving and extending window models using a large and rich data set covering over 123,000 houses sold between 2005 and 2017 in Switzerland. Instead of working with a single static data set, our models are repeatedly updated quarter by quarter by either a moving window or extending window strategy and evaluated on the following quarter to ensure a fair comparison and to resemble real life applications as closely as possible. This allows us to investigate volatility as well as accuracy of appraisals over time, an aspect that is typically ignored both in the literature as well as in Kaggle competitions, but highly relevant in practice.

Hence, the main contributions of this paper are: (1) to compare multiple important estimation methods; (2) to consider robust regression techniques that have not previously been applied in this context; (3) to repeatedly update our data and re-estimate the models in a manner that replicates real-life applications; (4) to consider the volatility of predictions over time, in addition to accuracy; and (5) to compare two data updating methods. Our analysis shows that there is a trade-off between accuracy and stability of price predictions. Based on most criteria, such as the percentages of predictions within 10 or 20 per cent of the sale price, gradient boosting is most accurate, followed by the mixed-effects model. The robust linear regression method yields the least volatile predictions, closely followed by the standard model and then the mixed-effects model. The choice of extending versus moving windows to update the model data has only a modest impact on the results.

The remainder of the paper is organized as follows. Section 2 describes the data and presents our estimation methods. The results are discussed in section 3. A final section provides some concluding remarks.

2 Data and methods

2.1 Data

We focus on a sample of 123,090 transactions of single-family houses sold at arm's length in Switzerland between 2005 and the second quarter of 2017 (except for the volatility analysis, for which we added data from the third quarter of 2017). The data were provided by the Informations- und Ausbildungszentrum für Immobilien AG (IAZI), a property valuation firm located in Zurich. Among other things, IAZI produces hedonic house price indexes and appraisals based on a majority of property transactions in Switzerland (Bourassa *et al.*, 2008; Bourassa *et al.*, 2010). Table I summarizes the univariate distributions of the raw characteristics and how they were represented in the models (typically by a log transformation). The median transaction is for a home built in 1980 with 5.5 rooms, 151 m^2 of living area and a 564 m^2 lot, which sold for CHF 780,000. From the two-room "rustic" in Ticino to the 15-room luxury villa on the shores of Lake Geneva, the data set covers a very wide range of properties and provides a representative sample of the Swiss housing market.

Table I. Descriptive statistics for sale price and property characteristics (*n* = 123,090)

Variable	*Mean*	*Standard deviation*	*Median*	*Minimum*	*Maximum*	*Transformation*
Sale price (CHF millions)	0.95	0.71	0.78	0.1	16	log
Living area (m^2)	163	60	151	30	1,180	log
Volume (m^3)	917	375	845	110	7,506	log (volume/ living area)
Lot size (m^2)	705	907	564	50	85,727	log

Number of rooms (excluding kitchen and bathrooms)	5.78	1.40	5.5	2.0	15.0	log
Number of bathrooms	2.07	0.75	2	1	7	log
Number of garages	1.0	0.9	1	0	7	root
Building age (years)	40.4	44.7	31.0	1.0	815	log
Condition of building (1=best to 4=worst)	2.1	0.7	2	1	4	none
Quality of building (1=best to 4=worst)	1.99	0.7	2	1	4	none
Quality of micro location (2=best to 4=worst)	3.02	0.6	3	2	4	none
Luxurious house (0=no, 1=yes)	0.03	–	0	0	1	none
Second home (0=no, 1=yes)	0.05	–	0	0	1	none
Single-family home (0=no, 1=yes)	0.66	–	1	0	1	none

Transactions occurred quite regularly over our full sample period (Table II provides the transaction counts and percentages for each year). In our models, we represented the transaction quarter either by dummy variables or, for the tree-based models, as a decimal number (in years).

Table II. Transactions per year, 2005 to mid-2017 (1,000s)

	2005	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*
n	10.0	11.0	11.6	10.2	11.1	11.1	11.2	10.0	7.9	7.6	8.2	8.6	4.6
%	8.2	9.0	9.4	8.3	9.0	9.0	9.1	8.1	6.4	6.2	6.7	7.0	3.7

The data set is enriched by environmental variables available at the municipal level in order to model the effect of location (Table III). Confidentiality restrictions imposed by the data provider mean that no finer level of geo-referencing (e.g., using postcodes or spatial coordinates) is available.

Table III. Descriptive statistics of municipality characteristics (*n* = 123,090)

Variable	*Mean*	*Standard deviation*	*Median*	*Minimum*	*Maximum*	*Transformation*
Travel time to large city (minutes)	42.07	29.64	36.00	0.00	286.00	log
Travel time to medium city (minutes)	16.96	14.14	14.00	0.00	159.00	log
Primary sector employment (proportion)	0.07	0.10	0.03	0.00	0.91	log
Secondary sector employment (proportion)	0.25	0.14	0.23	0.00	0.95	log
Forest area (proportion)	0.24	0.14	0.23	0.00	0.93	root
Industry area (proportion)	0.06	0.06	0.04	0.00	0.45	log
Tourist destination (0=no, 1=yes)	0.04	–	0.00	0.00	1.00	none
Number of doctors per 1,000 people	3.31	4.43	1.96	0.12	327.87	log
Number of food stores per 1,000 people	1.00	0.75	0.85	0.10	13.19	log
Proportion with university degree	0.15	0.09	0.12	0.00	0.44	log
Number of criminal offences per 1,000 people	49.67	32.75	43.15	0.00	545.07	log
Unemployment rate	0.01	0.01	0.01	0.00	0.05	log
Number of welfare recipients per 100 people	2.69	1.92	2.23	0.17	11.62	log
Foreigner proportion	0.22	0.10	0.21	0.00	0.60	log
Average federal tax load per capita (CHF)	1260	2713	796	77	87846	log
Average taxable income (CHF 1,000s)	34.33	24.32	30.54	12.49	785.87	log
Population (1,000s)	17.15	48.47	5.13	0.03	402.76	log

Vacancy rate	0.02	0.02	0.01	0.00	0.14	log
Number of houses per capita	0.10	0.07	0.09	0.00	3.71	log
Rental price level in 1990 (CHF)	816	187	785	203	1755	log
Beside lake (0=no, 1=yes)	0.16	–	0.00	0.00	1.00	none

Note: Some variables were shifted by a small positive amount before taking natural logarithms to increase distributional symmetry and to avoid exact zeros.

2.2 Data updating strategies

In order to provide up-to-date appraisals, hedonic models are periodically updated with new transactions. There are two types of updating strategies depending on whether old transactions are removed from the data: the moving window strategy based on a time window containing sufficient transactions for estimation purposes or the extending window strategy (see Figure 1 for a schematic overview). With respect to accuracy, our a priori sense is that less flexible techniques like the standard linear regression model would benefit from shorter windows (as some price effects will change over time), while flexible techniques like tree-based models would benefit from longer windows.

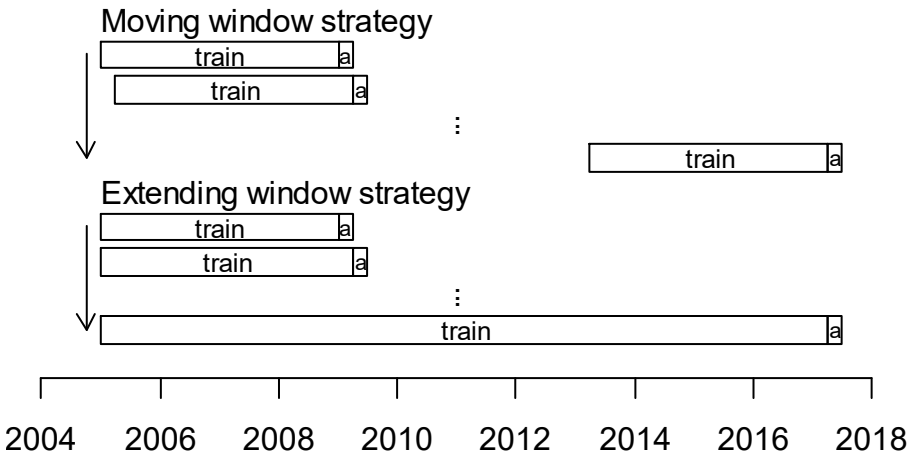

Figure 1. Schematic illustration of the two data selection strategies

Note: "a" refers to the quarter used to evaluate the appraisals.

In our case, both strategies begin with a four-year training data set with transactions from the first quarter of 2005 to the last quarter of 2008 to fit the models. For the moving window strategy, we repeatedly shift this window by one quarter and refit the models. For the extending window strategy, instead of shifting by one quarter, we keep adding the new quarterly data to the training data. This is repeated until the training data set ends at the first quarter of 2017. In this way, 34 different (yet overlapping) training sets are available for each of the two strategies. The model performance is evaluated always on the quarter following the training period. Every quarter from the beginning of 2009 until mid-2017 is used once for evaluation purposes.

2.3 Modelling techniques

All models and analyses were calculated with the statistical software R, version 3.4.3 (R Core Team, 2017). Data related decisions like the selection and transformation of independent variables or the choice of relevant tuning parameters were based on one single four-year data window selected from the middle of the full time range and kept fixed for all other model calculations.

The reference model is a linear regression model, $\log P_i = \beta x_i + \varepsilon_i$, fitted with base R function `lm` (OLS), where P_i is the transaction price for property i, x_i is the regressor vector derived from the property characteristics, transaction quarters (dummy coded), and a set of environmental variables describing the municipality to which property i belongs (see Tables I and III for details about specific transformations applied). Variable selection was done manually by removing only variables with virtually no predictive power (based on *t*-values very close to zero), following the suggestions in Harrell (2001). Quadratic terms were added very cautiously with the aim of keeping the model relatively simple. For the same reason, no interactions were added to the model.

The selected model specification was then used to fit a robust linear regression with the aim to better deal with outliers in the independent or dependent variables. We used the `lmrob` function in the R package `robustbase` (Maechler *et al*., 2017), version 0.92-8, that implements an MM-type robust regression suggested in Yohai (1987) and Koller and Stahel (2011). In contrast to OLS, which minimizes the sum of the squared errors, the MM

estimator down weights outliers in an iterative manner. This means that outliers have less influence over the estimation. This method offers excellent robustness properties while being almost as efficient as OLS under normal errors.

There are more than 2,000 municipalities in Switzerland. As described in Table III, our data set includes 21 variables measuring characteristics of municipalities; even with these variables, some residual bias will be left at the municipality level. One way to at least partly remove this bias without introducing considerable overfit is to extend the classical linear model by adding random effects at one or more spatial levels. In our case, we used nested spatial random intercepts at cantonal (state or provincial), regional (smaller than a canton, but larger than a commune), and communal (municipal) levels. These mixed effects models were fitted by the function `lmer` in the R package `lme4`, version 1.1.17 (Bates *et al*., 2015). The model formula was selected starting with the final specification for the standard linear model and then iteratively removing fixed municipality characteristics with *t*-values close to zero.

Besides these three linear models, we considered some of the most frequently used basic techniques of modern machine learning in the context of regression: random forests, gradient boosting, and neural networks. The first two of these methods are ensembles of decision trees. A decision tree is a collection of binary questions about the covariables (e.g., is the living area smaller than 220 square meters?) and predictions are found by the average response of all observations sharing the same answers to these binary questions (see Hastie, 2001, for more information). Figure 2 illustrates a simple decision tree with house price in CHF millions as response and (untransformed) model variables as covariables. While simple to interpret, single decision trees typically do not provide very accurate results and very small changes to the input can lead to big jumps in the predictions.

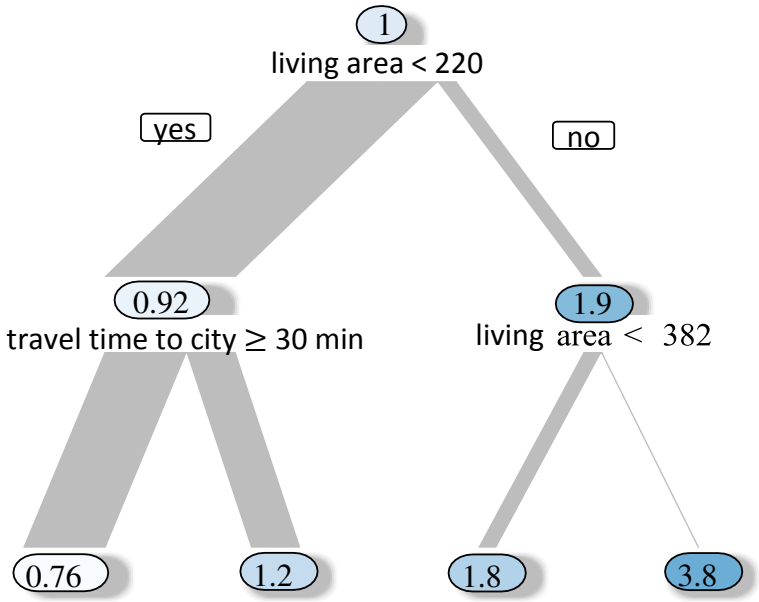

Figure 2. A simple decision tree of depth two

Notes: The ovals contain the average response in CHF millions of all observations following the same path. A house with 150 square meters of living area and 20 minutes travel time to the nearest large city costs CHF 1,200,000 on average.

Better results are usually obtained by random forests, which – in the context of regression – are averages of many slightly different, very deep decision trees. The trees differ for two reasons. First, each decision split of each tree is found by considering only a random subset of m covariables. Second, each tree is calculated on a bootstrap sample from the model data, introducing an additional source of variability. One advantage of random forests is that they perform well even when all parameters are set to typical default values. Another advantage is that fair prediction accuracies can be approximated without the need for cross-validation from rows not selected by the bootstrap. We used these "off-the-shelf" accuracies to select the main tuning parameter m. The number of trees was set to a time saving 500. In R, different random forest implementations are available. The results shown were found by the R package `ranger`, version 0.9.0 (Wright and Ziegler, 2017).

Another way to combine multiple decision trees is gradient (tree) boosting. A shallow decision tree is first fitted to the model data. Then, the residuals are fitted by a new decision tree to correct the mistakes made by the initial tree. This is repeated many times

until cross-validation performance stops improving. The final predictions are made by taking an average of all predictions from all trees. Gradient boosting typically outperforms random forests if its many tuning parameters are carefully selected. We did this by iteratively going through different choices of the main tuning parameters and selecting the best combination by five-fold cross-validation, a strategy that is called "GridSearchCV" (see Raschka and Mirjalili, 2017). To calculate boosted trees, we used the `lightgbm` package (Ke, 2018), version 2.1.0, a highly efficient alternative to the popular XGBoost (Chen and Guestrin, 2016) algorithm. We tuned the learning rate, the number of boosting rounds, different aspects determining the tree size, and the proportion of rows and covariables selected in the calculation of each tree (row and column subsampling).

Finally, an artificial neural network extends the classical linear regression by adding additional structure to "learn" the optimal representation of covariables (non-linearities, interactions, transformations) autonomously from the data. This is done by adding intermediate layers of derived variables (called "hidden nodes") whose values are non-linearly transformed weighted sums of all variables on the previous layer. As for gradient boosting, the selection of tuning parameters such as the learning rate (determining how aggressively the model parameters are adjusted by adding new data rows), the architecture (how many hidden layers with how many hidden nodes each), regularization measures (dropout, L2 penalization), and the number of epochs (how many times each data row is presented to the algorithm) was done by GridSearchCV. As optimizer we used mini-batch stochastic gradient descent without momentum. The neural networks were calculated by the function `h2o.deeplearning` in the R package h2o, version 3.16.0.2 (The H20.ai team, 2017).

All models use the natural logarithm of the transaction price as dependent variable (see, e.g., Yacim and Boshoff, 2018, with respect to specification of artificial neural network models). Results are reported on this scale if not otherwise mentioned. Further note that all covariables were prepared in order to be suitable for the linear models (e.g., using log transformations and decorrelating volume and living area by taking their ratio).

3 Results

In this section, we first describe the data-driven model decisions. Then, we discuss the performance (accuracy) of the methods for both the moving and extending window strategies. Finally, we study the volatility of appraisals over time for each method and strategy.

3.1 Models

Following the model selection strategies outlined above, the models include most of the transformed variables listed in Tables I and III, with the following differences across model techniques. As noted above, the tree-based models use the transaction quarter as a numerical variable rather than as a dummy-coded factor. The linear models are enriched by adding squared terms for the age of the property and the building condition, as well as for the proportion of the population with a university degree. The tree-based models did not require the dummy for a municipality being a tourist destination, while the linear models did not benefit from inclusion of the unemployment rate, the number of food stores per 1,000 people, the number of houses per capita, or the percentage of industrial area, so those variables were not included in the relevant models. In addition, the mixed effects model did not require inclusion of the number of doctors per 1,000 people, the travel time to a medium-sized city, or the number of welfare recipients per 100 people, thanks to the random locality effects.

The main tuning parameter m (number of randomly picked variables to determine the best split at each split) of the random forest was set to 13. No other decisions were made for the random forest.

The boosted trees worked best with 1,400 boosting rounds at a learning rate of 0.02. The maximal tree size was set to 127 leaves. No row subsampling was applied and each tree was calculated by using a random subset of 40 per cent of all covariables.

The neural networks were trained for 30 epochs at a learning rate of 0.005. No regularization was necessary (no dropout, no L2 penalties). The optimal architecture found

by GridSearchCV consisted of two hidden layers (the first with 30 hidden nodes, the second with five).

3.2 Accuracy

For each model, we calculate accuracy measures on the logarithmic one quarter ahead prediction errors, namely the absolute value of the mean of *e* ("absMean", which is a measure of bias), the root mean square error ("rmse"), the mean absolute error ("mae"), the median absolute error ("medae"), and finally the proportion of predictions within 10 and 20 per cent, respectively, of the actual transaction price ("within10%", "within20%"). We did not focus on a single accuracy measure (e.g., rmse) since not all models optimize the same objective function and thus focusing on a single measure would be unfair for some methods. In order to see if descriptive differences for each accuracy measure and both data selection strategies could be explained by pure luck, we compared the results between methods in a pairwise manner using two-sided, exact paired permutation t-tests at the 5 per cent level of significance.

Figure 3 depicts the results over time for the moving window strategy according to the first four criteria mentioned above. The gradient boosting machine approach followed by the linear mixed effects model outperform the other techniques, while the random forest method as well as OLS and robust linear regressions do worst. Table IV (moving window strategy) and Table V (extending window strategy) show averages over time for all accuracy measures. Overall, the choice of the data selection strategy had only a minor impact on accuracy with a small advantage for the moving window strategy, except for the gradient boosting machine method which seems to benefit slightly from the expanding size of the window (see Figure 4).

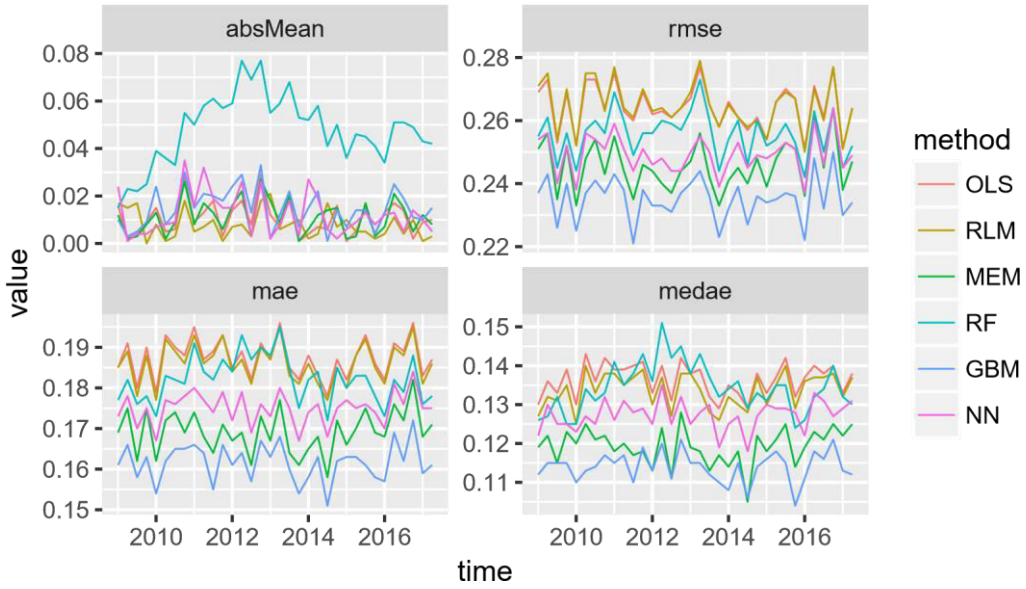

Figure 3. Comparison of the accuracy of methods (moving window strategy)

Notes: The methods are: OLS: ordinary least squares estimation of standard linear model; RLM: robust linear model; MEM: mixed effects model; RF: random forest; GBM: gradient boosting machine; and NN: neural network. The four accuracy criteria are: the absolute mean of the error, the root mean square error, the mean absolute error, and the median absolute error, respectively.

Since we evaluate model performance on the quarter following the training data periods, a bias of the same magnitude as the most recent market movement is expected for all models. But how can the much larger bias of the random forest (and thus also its unexpectedly bad performance) be explained?

Table IV. Average accuracy for the moving window strategy across all evaluation quarters

	absMean	*rmse*	*mae*	*medae*	*within10%*	*within20%*
OLS	0.011	0.264	0.187	0.136	0.383	0.660
RLM	**0.008***	0.265	0.186	0.134	0.391	0.666
MEM	0.011	0.246	0.169	0.119	0.432	0.713
RF	0.048	0.256	0.182	0.134	0.391	0.670
GBM	0.015	**0.235***	**0.162***	**0.114***	**0.451***	**0.729***
NN	0.012	0.249	0.175	0.127	0.410	0.690

Notes: The best method for each accuracy measure is indicated in bold and * means it was significantly better than all other methods. Since the response is logarithmic price, the values of the first four accuracy measures can (approximately) be read as percentage errors.

Table V. Average accuracy for the extending window strategy across all evaluation quarters

	absMean	*rmse*	*mae*	*medae*	*within10%*	*within20%*
OLS	0.011	0.266	0.189	0.137	0.383	0.657
RLM	**0.008***	0.267	0.188	0.135	0.387	0.661
MEM	0.011	0.248	0.171	0.121	0.430	0.708
RF	0.063	0.258	0.186	0.139	0.378	0.656
GBM	0.015	**0.234***	**0.160***	**0.113***	**0.455***	**0.732***
NN	0.015	0.250	0.175	0.126	0.410	0.692

Notes: The best method for each accuracy measure is indicated in bold and * means it was significantly better than all other methods. Since the response is logarithmic price, the values of the first four accuracy measures can (approximately) be read as percentage errors.

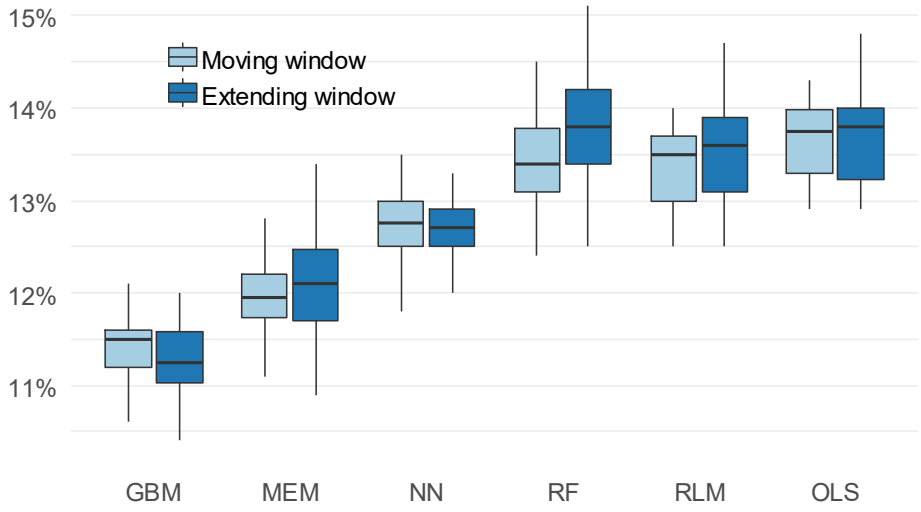

Figure 4. Boxplots of median absolute errors across all 34 evaluation quarters

Note: Outliers not shown to increase readability.

The reason is that, in our data setting, the random forests seem to be unable to pick up the usually weak effects of the transaction quarter, no matter which random forest implementation we used and how we represent the transaction quarter (numerically or with dummy variables). Thus, for the random forest, the typical bias on the evaluation quarter does not only represent the market movement from the *end* of the model period to the evaluation quarter, but rather from the *middle* of the model period. Partial dependence plots may help to identify the issue (see Figure 5). Such plots depict the marginal effect of a variable on the response and their use is suggested in Friedman (2000) to shed some light on black box models like gradient boosting machines or artificial neural networks. While in

our case the marginal effect of a strong predictor like living area is quite similar across all modelling techniques, the transaction quarter is almost flat for the random forest, thus revealing that technique's problematic property.

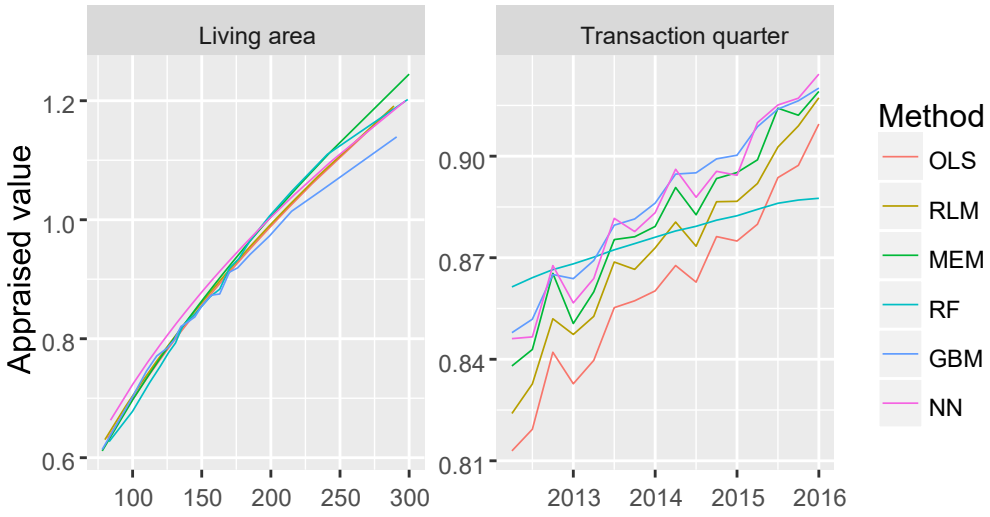

Figure 5. Partial dependence plots for living area and transaction quarter

Notes: Both plots are calculated on the four-year data window ending in Q1 2016 using the moving window strategy (back transformed to the original raw data scale; appraisals in CHF millions).

3.3 Volatility of individual appraisals

In the banking world, in order to assess the risk associated with loans, the value of a house might be reappraised on a regular basis, each year or quarter, by the most current version of the bank's automated valuation software. Ideally, changes in the appraised value of a given property would mainly follow market trends and not exhibit large jumps that are due to changes over time in the data structure used to calculate the models. Thus, besides accuracy, an important feature of a statistical model in the area of automated valuation is the volatility of individual appraisals over time. We investigate this aspect by estimating the value of all properties sold after the last training data window using all models, time periods, and data selection strategies.

To gain a visual impression of volatility in the appraised values, Figure 6 shows estimated values over time for four randomly selected properties and the moving window strategy. Clearly, the jumps over time are considerably smaller for the OLS and robust linear

regressions. The values generated from the neural network model exhibit erratic behavior – despite the model's acceptable performance with respect to accuracy.

Figure 6. Appraisals over time for four properties (moving window strategy)

To quantify the volatility of such curves, we calculated absolute differences of (log) appraised values from one quarter to the next for each method, both data selection strategies, and all 2,773 transactions of 2017q2 and 2017q3 which are outside any model calculation window. Table VI shows summary statistics for the moving window strategy. The OLS and robust linear regressions do almost equally well with a slight advantage for the robust regression, closely followed by the mixed effects model. The jumps for the tree-based methods are on average about twice as large as for the linear methods. The neural network method clearly yields the worst results. The ranking of the methods is similar for the extending window strategy (see Table VII). Except for the neural network method, the extending window strategy tends to yield slightly less volatile results, especially for the gradient boosting approach, which seems to benefit from the richer data (see Figure 7). In order to supplement the descriptive comparison, we compared average jump heights per

property across methods by means of two-sided approximate paired permutation t-tests at the 5 per cent level of significance. The jumps related to the robust regression were significantly smaller than those for all other methods for both strategies.

Table VI. Summary statistics of (absolute) jumps per method for the moving window strategy

	Mean	*Standard deviation*	*Median*	*Maximum*
OLS	0.012	0.008	0.010	0.059
RLM	**0.011***	0.007	0.009	0.045
MEM	0.015	0.013	0.013	0.258
RF	0.023	0.021	0.018	0.405
GBM	0.033	0.028	0.026	0.369
NN	0.059	0.047	0.049	0.457

Notes: Evaluated on 2,773 observations times 33 quarterly differences. The best method is indicated in bold and * indicates that the mean was significantly lower than those of all other methods. Since jumps are calculated in logarithmic differences, these values can be interpreted as approximate percentages.

Table VII. Summary statistics of (absolute) jump heights per method for the extending window strategy

	Mean	*Standard deviation*	*Median*	*Maximum*
OLS	0.011	0.008	0.009	0.045
RLM	**0.010***	0.007	0.009	0.038
MEM	0.012	0.009	0.011	0.160
RF	0.021	0.019	0.017	0.359
GBM	0.027	0.024	0.021	0.318
NN	0.064	0.052	0.052	0.727

Note: Evaluated on 2,773 observations times 33 quarterly differences. The best method is indicated in bold and * indicates that the mean was significantly lower than those of all other methods.

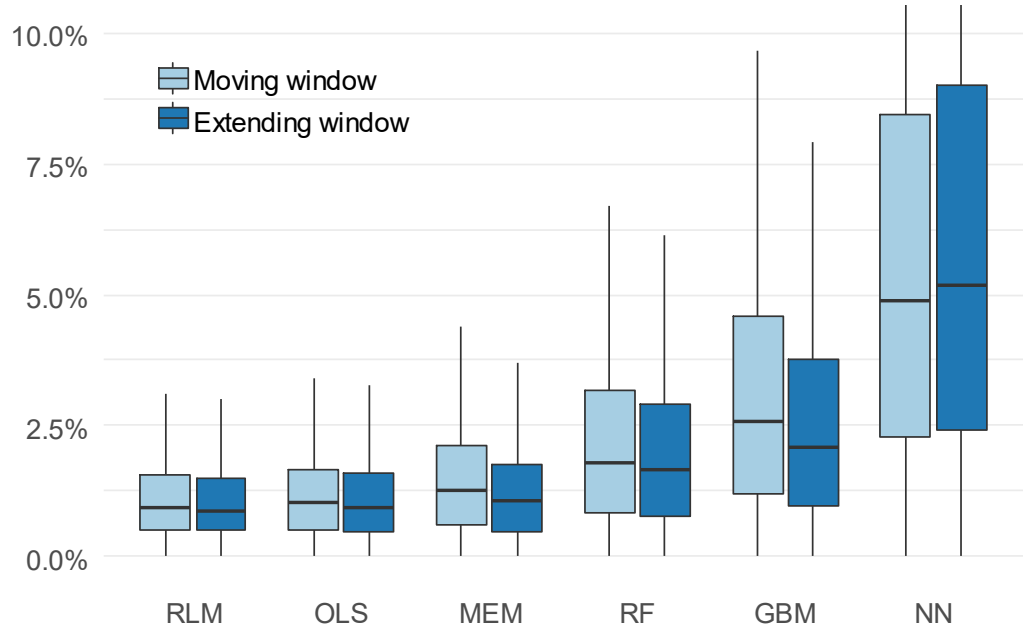

Figure 7. Boxplots of (absolute) jump heights of 2,773 observations evaluated over 34 quarters each

Note: Outliers not shown to increase readability.

4 Conclusion

With respect to accuracy, the gradient boosting approach outperforms the other estimators, followed by the mixed effects regression, the neural network method, and the random forest approach. The robust and OLS regression methods perform the worst. Random forest models suffer large biases because they have trouble capturing the market trend, a severe problem in real world applications where a model is fitted strictly on historical data and then applied to the current market.

Thanks to their simplicity, the three linear models clearly provide less volatile appraisals over time than the three "black box" models. The robust regression method performs best with respect to volatility. Thus, in settings where properties are periodically re-appraised (for instance for refinancing purposes or for risk assessment) with regularly updated

models, linear models offer considerable advantages compared to the tree-based methods and especially to the very erratic results generated by neural network models.

Consequently, if the sole aim is high precision, then gradient boosting decision trees seem to be the appropriate choice. When volatility in repeated appraisals is important, too, the mixed effects model provides a good compromise. The mixed effects model also avoids the complexity and lack of transparency of the machine learning methods. If reducing volatility is of key importance, robust regression models should be selected.

The choice between a moving or extending window approach has only a modest impact on the results. The moving window approach seems attractive for the standard linear and robust regression methods. Such a strategy permits a small increase in accuracy while volatility is not affected. The gradient boosting and neural network approaches' accuracy and volatility tend to be better with the extending window strategy. Mixed effects models perform slightly worse but are less volatile under the extending window approach.

References

Antipov, E. and Pokryshevskaya, E. (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.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015), "Fitting linear mixed-effects models using lme4", *Journal of Statistical Software*, Vol. 67 No. 1, pp. 1-48.

Bourassa, S., Cantoni, E. and Hoesli, M. (2007), "Spatial dependence, housing submarkets, and house prices", *Journal of Real Estate Finance and Economics*, Vol. 35 No. 2, pp. 143-160.

Bourassa, S., Cantoni, E. and Hoesli, M. (2010), "Predicting house prices with spatial dependence: a comparison of alternative methods", *Journal of Real Estate Research*, Vol. 32 No. 2, pp. 139-159.

Bourassa, S., Cantoni, E. and Hoesli, M. (2016), "Robust hedonic price indexes", *International Journal of Housing Markets and Analysis*, Vol. 9 No. 1, pp. 47-65.

Bourassa, S., Hoesli, M. and Peng, V. (2003), "Do housing submarkets really matter?" *Journal of Housing Economics*, Vol. 12 No. 1, pp. 12-18.

Bourassa, S., Hoesli, M. and Scognamiglio, D. (2010), "Housing finance, prices, and tenure in Switzerland", *Journal of Real Estate Literature*, Vol. 18 No. 2, pp. 263-282.

Bourassa, S., Hoesli, M., Scognamiglio, D. and Sormani, P. (2008), "Constant-quality house price indexes for Switzerland", *Swiss Journal of Economics and Statistics*, Vol. 144 No. 4, pp. 561-575.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32.

Brown, K. and Uyar, B. (2004), "A hierarchical linear model approach for assessing the effects of house and neighborhood characteristics on housing prices", *Journal of Real Estate Practice and Education*, Vol. 7 No. 1, pp. 15-24.

Case, B., Clapp, J., Dubin, R. and Rodriguez, M. (2004), "Modeling spatial and temporal house price patterns: a comparison of four models", *Journal of Real Estate Finance and Economics* Vol. 29 No. 2, pp. 167-191.

Chen, T. and Guestrin, C. (2016), "Xgboost: a scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco*, pp. 785-794.

Chiarazzo, V., Caggiani, L., Marinelli, M. and Ottomanelli, M. (2014), "A neural network based model for real estate price estimation considering environmental quality of property location", *Transportation Research Procedia*, Vol. 3, pp. 810-817.

Ciuna, M., Salvo, F. and Simonotti, M. (2017), "The multilevel model in the computer-generated appraisal: a case in Palermo", in d'Amato, M. and Kauko, T., *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis*, Springer International Publishing, New York, pp. 225-261.

Din, A., Hoesli, M. and Bender, A. (2001), "Environmental variables and real estate prices", *Urban Studies*, Vol. 38 No. 11, pp. 1989-2000.

Efron, B. and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press, New York.

Friedman, J. (2000), "Greedy function approximation: a gradient boosting machine", *Annals of Statistics*, Vol. 29 No. 5, pp. 1189-1232.

Goodman, A. and Thibodeau, T. (2003), "Housing market segmentation and hedonic prediction accuracy", *Journal of Housing Economics*, Vol. 12 No. 3, pp. 181-201.

Gu, G. and Xu, B. (2017), "Housing market hedonic price study based on boosting regression tree", *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 21 No. 6, pp. 1040-1047.

Harrell, F. (2001), *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, New York.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014), *An Introduction to Statistical Learning: With Applications in R*, Springer, New York.

Kagie, M. and Van Wezel, M. (2007), "Hedonic price models and indices based on boosting applied to the Dutch housing market", *Intelligent Systems in Accounting, Finance and Management*, Vol. 15 No. 3-4, pp. 85-106.

Ke, G. (2018), *lightgbm: Light Gradient Boosting Machine*, R package version 2.1.0.

Keskin, B., Dunning, R. and Watkins, C. (2017), "Modelling the impact of earthquake activity on real estate values: a multi-level approach", *Journal of European Real Estate Research*, Vol. 10 No. 1, pp. 73-90.

Koller, M. and Stahel, W. (2011), "Sharpening wald-type inference in robust regression for small samples", *Computational Statistics & Data Analysis*, Vol. 55 No. 8, pp. 2504-2515.

Lu, S., Li, Z., Qin, Z., Yang, X. and Goh, R., (2017), "A hybrid regression technique for house prices prediction", in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 319-323.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. and di Palma, M. (2017), *robustbase: Basic Robust Statistics*, R package version 0.92-8.

Malpezzi. S. (2008), "Hedonic pricing models: a selective and applied review", in O'Sullivan, T. and Gibb, K. (Eds), *Housing Economics and Public Policy*, Blackwell, Oxford, UK, pp. 67-89.

McCluskey, W., McCord, M., Davis, P., Haran, M. and McIlhatton, D. (2013), "Prediction accuracy in mass appraisal: a comparison of modern approaches", *Journal of Property Research*, Vol. 30 No. 4, pp. 239-265.

Mullainathan, S. and Spiess, J. (2017), "Machine learning: an applied econometric approach", *Journal of Economic Perspectives*, Vol. 31 No. 2, pp. 87-106.

Orford, S. (2002), "Valuing locational externalities: a GIS and multilevel modelling approach", *Environment and Planning B*, Vol. 29 No. 1, pp. 105-127.

Peña, D. and Ruiz-Castillo, J. (1984), "Robust methods of building regression models: an application to the housing sector", *Journal of Business & Economic Statistics*, Vol. 2 No. 1, pp. 10-20.

Peterson, S. and 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.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Raschka, S. and Mirjalili V. (2017), *Python Machine Learning*, Packt Publishing, Birmingham, UK.

Rumelhart, D., Hinton, G. and Williams, R. (1986), "Learning internal representations by error propagation", in Rumelhart, D., McClelland, J. and the PDP Research Group (Eds), *Parallel Distributed Processing: Computational Models of Cognition and Perception*, Volume 1: Foundations, MIT Press, Cambridge, MA and London, UK, pp. 318-362.

Sangani, D., Erickson, K. and Hasan, M. (2017), "Predicting Zillow estimation error using linear regression and gradient boosting", in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 530-534.

Schulz, R., Wersing, M. and Werwatz, A. (2014), "Automated valuation modelling: a specification exercise", *Journal of Property Research*, Vol. 31 No. 2, pp. 131-153.

Sirmans, G., Macpherson, D. and Zietz, E. (2005), "The composition of hedonic pricing models", *Journal of Real Estate Literature*, Vol. 13 No. 1, pp. 3-43.

The H2O.ai team (2017), *h2o: R Interface for H2O*, R package version 3.16.0.2.

Varian, H. (2014), "Big data: new tricks for econometrics", *Journal of Economic Perspectives*, Vol. 28 No. 2, pp. 3-28.

Worzala, E., Lenk, M. and Silva, A. (1995), "An exploration of neural networks and its application to real estate valuation", *Journal of Real Estate Research*, Vol. 10 No. 2, pp. 185-201.

Wright, M. and Ziegler, A. (2017), "ranger: a fast implementation of random forests for high dimensional data in C++ and R", *Journal of Statistical Software*, Vol. 77 No. 1, pp. 1-17.

Yacim, J. and Boshoff, D. (2018), "Impact of artificial neural networks training algorithms on accurate prediction of property values", *Journal of Real Estate Research*, Vol. 40 No. 3, pp. 375-418.

Yohai, V. (1987), "High breakdown-point and high efficiency robust estimates for regression", *The Annals of Statistics*, Vol. 15 No. 2, pp. 642-656.

Yoo, S., Im, J. and Wagner, J. (2012), "Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County, NY", *Landscape and Urban Planning*, Vol. 107 No. 3, pp. 293-306.

Zurada, J., Levitan, A. and Guan, J. (2011), “A comparison of regression and artificial intelligence methods in a mass appraisal context”, *Journal of Real Estate Research*, Vol. 33 No. 3, pp. 349-387.

Swiss Finance Institute

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public–private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.

swiss:finance:institute

c/o University of Geneva, Bd. Du Pont d'Arve 42, CH-1211 Geneva 4
T +41 22 379 84 71, rps@sfi.ch, www.sfi.ch