

Hinrichs, Nils; Kolbe, Jens; Werwatz, Axel

Working Paper

AVM and high dimensional data: Do ridge, the lasso or the elastic net provide an "automated" solution?

FORLand-Working Paper, No. 22 (2020)

Provided in Cooperation with:

DFG Research Unit 2569 FORLand "Agricultural Land Markets – Efficiency and Regulation",
Humboldt-Universität Berlin

Suggested Citation: Hinrichs, Nils; Kolbe, Jens; Werwatz, Axel (2020) : AVM and high dimensional data: Do ridge, the lasso or the elastic net provide an "automated" solution?, FORLand-Working Paper, No. 22 (2020), Humboldt-Universität zu Berlin, DFG Research Unit 2569 FORLand "Agricultural Land Markets - Efficiency and Regulation", Berlin, <https://doi.org/10.18452/21263>

This Version is available at:

<https://hdl.handle.net/10419/227605>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by-nc-nd/4.0/>



AVM and High Dimensional Data - Do Ridge, the Lasso or the Elastic Net provide an 'automated' solution?

Nils Hinrichs, Jens Kolbe, Axel Werwatz

FORLand-Working Paper 22 (2020)

Published by

DFG Research Unit 2569 FORLand, Humboldt-Universität zu Berlin
Unter den Linden 6, D-10099 Berlin
<https://www.forland.hu-berlin.de>



Tel +49 (30) 2093 46845, Email gabriele.wuerth@agrار.hu-berlin.de

Agricultural Land Markets – Efficiency and Regulation

AVM and High Dimensional Data – Do Ridge, the Lasso or the Elastic Net provide an ‘automated’ solution

Nils Hinrichs, Jens Kolbe and Axel Werwatz*

*Institut für Volkswirtschaftslehre und Wirtschaftsrecht, Technische Universität Berlin,
Straße des 17. Juni 135, 10623 Berlin, Germany, jkolbe@tu-berlin.de and axel.werwatz@tu-berlin.de.

Abstract

In this paper, we apply Ridge Regression, the Lasso and the Elastic Net to a rich and reliable data set of condominiums sold in Berlin, Germany, between 1996 and 2013. We their predictive performance in a rolling window design to a simple linear OLS procedure. Our results suggest that Ridge Regression, the Lasso and the Elastic Net show potential as AVM procedures but need to be handled with care because of their uneven prediction performance. At least in our application, these procedures are not the ‘automated’ solution to Automated Valuation Modeling that they may seem to be.

Keywords: Automated valuation, Machine learning, Elastic Net, Forecast performance

JEL Classification: R31, C14

1 Introduction

Automated Valuation Models (AVMs) use recorded transaction or listing information to fit a statistical model of the relationship between prices and characteristics of properties. Once the model is fitted, the market value of any property with given characteristics can easily be predicted. The challenge lies in the modeling stage. It is exacerbated by the ‘nature’ of properties and property data. Properties are very heterogeneous and their value depends on a multitude of factors. Thus, to be reasonably accurate, AVMs need to be based on data with detailed information on property characteristics. At the same time, turnover in property markets is typically rather low. Statistically speaking, this translates into a regression problem with relatively few observations but a relatively large set of predictors. Very flexible and very data hungry models such as nonparametric regression or neural networks thus quickly run into the ‘curse of dimensionality’ in this situation: their estimates of the price–characteristics relationship become very imprecise and their predictions therefore can become quite inaccurate. Thus, simply ‘letting the data speak’ is not a real option in AVM regression modeling.

In this paper, we thus follow a different modeling approach. We focus on flexible parametric regression models and estimate them by constrained least squares from a rich and reliable data set of condominium transactions. By working with a basis expansion of the set of original property characteristics, these models have the ability to adapt quite flexibly to the relationship of these characteristics and the transaction price contained in the data. The two constrained least squares estimators, –ridge regression and the LASSO–, make sure that this flexibility does not get out of control. They both ‘regularize’ the problem by constraining the total size of the regression coefficients. By shrinking some regression coefficients (ridge regression) or setting some to zero (LASSO) the procedures make sure that the coefficients of the important relationships can be estimated precisely and that precious degrees of freedom are not wasted on coefficients of unimportant predictors. Which coefficients are constrained, and thus which predictors are deemed unimportant,

is determined in a data-driven, automated way – thus entirely in the spirit of automated valuation.

We apply ridge regression and the LASSO to a data set containing detailed information on all condominiums sold between 1996 and 2013 in Germany’s capitol Berlin. All information is retrieved from the actual sales contracts and can be regarded as reliable. We thus imagine the situation of an AVM service that delivers property valuations not to the general public but to professional customers such as mortgage lenders who demand and reward accuracy. Such an AVM service needs rich data and may even merge information on a given property from various databases – for instance, by enhancing records on a specific property with indicators on upkeep derived from images of the property or indicators on neighborhood amenities from a web mapping service. In short, while our data is quite detailed, the challenge that is at the heart of our approach (the number of observed properties being small relative to the dimension of the set of observed determinants) is likely to be even more intense in other current and future AVM applications.

The remainder of this paper is organized as follows. Section 2 discusses our methodology for developing our AVM regressions and how we will be assessing and comparing their performance. Section 3 presents the data. Section 4 explains the model specification results and the performance assessment obtained from the validation step. Section 5 concludes. Details of the analysis are relegated to the Appendix.

2 Methodology

2.1 Theoretical Framework

The aim of AVM is providing model-based predictions of the market value of a residential property. The market value is the price one should expect in an arm’s-length transaction between informed and willing buyers and sellers. We denote this price as P_{jt} for property j at time t . This value depends on

the property’s structural and location characteristics. Those that are observed in the AVM’s data base are summarized in the vector \mathbf{x}_{jt} . In our data set, \mathbf{x}_{jt} includes –among other things– a condominium’s living space, number of bedrooms or location within the building.

A reasonable criterion for assessing predictive quality is the mean squared prediction error. It is well-known that in this case the optimal predictor of P_{jt} , given \mathbf{x}_{jt} , is the conditional expectation $E[Y_{jt}|\mathbf{x}_{jt}]$. The goal of an AVM is thus to provide a good estimate of this conditional expectation. Denote such an estimate of a candidate model k as $\hat{E}[Y_{jt}|\mathbf{x}_{jt}] = \hat{m}_k(\mathbf{x}_{jt})$. For any estimated candidate model, the mean squared prediction error for a property with characteristics \mathbf{x}_{jt} can be decomposed as (Hastie et al. (2009), p. 223)

$$E[(P_{jt} - \hat{m}_k(\mathbf{x}_{jt}))^2|\mathbf{x}_{jt}] = \text{Var}(P_{jt}|\mathbf{x}_{jt}) + E\{\hat{m}_k(\mathbf{x}_{jt}) - E[P_{jt}|\mathbf{x}_{jt}]\}^2 + \text{Var}[\hat{m}_k(\mathbf{x}_{jt})] \quad (1)$$

The first part on the right hand side of equation (1) is the variance of prices of properties with the given the bundle of characteristics \mathbf{x}_{jt} . It does not depend on $\hat{m}_k(\mathbf{x}_{jt})$ and is thus not affected by the quality of any candidate AVM. It is therefore often referred to as the ‘irreducible error’. However, enlarging the set of features included in \mathbf{x}_{jt} *will* in general decrease this conditional variance and the mean squared prediction error. It explains why there is a strong incentive for AVM providers to collect detailed information on properties. It can be expected that in the future the dimension of the feature vector \mathbf{x}_{jt} will continue to grow. The second part of the decomposition is the squared bias of the estimated candidate model $\hat{m}_k(\mathbf{x}_{jt})$. This term can itself be decomposed into two parts, the estimation bias and the model bias, highlighting the fact that the quality of any AVM model has two key aspects: the quality of the functional form assumptions of the model and the quality of the estimation procedure used to fit the model. The latter, the estimation quality, takes center stage in the third and final component of (1), the variance of the candidate model estimator.

In the remainder of the paper, we assume that the set of available covariates \mathbf{x}_{jt} may be large but that it is fixed at the moment. Hence, the first term of the decomposition (1) then is indeed ‘irreducible’ and AVM prediction

quality entirely hinges on keeping the second and third terms as low as possible. Therein, however, lies a well-known trade-off: candidate models with flexible functional forms keep the model bias low but often suffer from large estimation error because of their complexity. The relatively few parameters of parsimonious candidate models, on the other hand, can typically be estimated precisely but their restrictive functional forms may result in a large model bias. The key to AVM is thus to find a model that delivers the best compromise between bias and variance.

We illustrate this trade-off and our candidate solutions in a highly simplified situation. Suppose that living space were the only continuous characteristics, that there were only two locations A and B and only two time periods, 1 and 2. Then an extremely simple candidate model is the following linear, additively separable model (LASM)

$$m_{LASM}(\mathbf{x}_{jt}) = \beta_0 + \beta_1 \text{space}_{jt} + \beta_2 \text{loc.B}_{jt} + \beta_3 \text{period2}_{jt} \quad (2)$$

where `loc.B` and `period2` are accordingly defined dummy variables. The model presupposes, for instance, that the effect of space is the same at both locations and in both time periods and is linear (i.e., independent of its own level). This functional form strongly restricts the ability of the model to adapt to the data. It is thus likely to suffer from a substantial model bias. The four parameters of the model, however, are easily estimated by least squares. Indeed, least squares will provide unbiased estimates of the model parameters in (2).¹ Estimation bias will thus be zero in this case and the squared bias term in (1) is entirely due to the model bias for this candidate model. Because of its simplicity and parsimony, estimation variance, the third term in (1), will be very small. A particularly clear expression of this can be obtained if the variance of $\hat{m}_{LASM}(\mathbf{x}) = \mathbf{x}^T \hat{\beta}^{LS}$ is not considered at a particular property with characteristics \mathbf{x}_{jt} (as in (1)) but instead averaged over the properties in the

¹More precisely, least squares will provide unbiased estimates of the parameters of the Best Linear Predictor, as which we implicitly define 2. See Hastie et al. (2009), p. 19.

training data:

$$\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \text{Var}(\hat{m}_{LASM}(\mathbf{x}_{it})) = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \text{Var}(\mathbf{x}_{it}^T \hat{\beta}^{LS}) = \frac{\sigma^2}{nT} p \quad (3)$$

This expression, which holds under the assumption of a constant conditional variance $\text{Var}(P_{jt}|\mathbf{x}_{jt}) = \sigma^2$, shows that the estimation variance is linearly increasing in the number of regression parameters $p + 1$. The latter is also the degrees of freedom of this procedure.

At the other end of the model complexity spectrum is the nonparametric regression model. In our simplified illustration, it is given as

$$m_{NPM}(\mathbf{x}_{jt}) = m(\text{space}_{jt}, \text{loc.B}_{jt}, \text{period2}_{jt}) \quad (4)$$

This model imposes no parametric functional form on the data and allows for arbitrary nonlinear effects of continuous regressors like `space` and for arbitrary interactions among the effects of all explanatory variables. This highly flexible model thus has very little if any model bias. It can be estimated by local smoothing procedures like kernel regression, k -nearest neighbor regression or smoothing splines. However, regardless of which estimator is applied, its variance increases dramatically with p for a given sample size, the well-known ‘curse of dimensionality’.

An alternative model, that allows for a nonlinear effect of floor space (via a cubic polynomial) and an interaction between floor space and location in a parametric way is given by:

$$\begin{aligned} \hat{m}_{FLM}(\mathbf{x}_{jt}) = & \beta_0 + \beta_1 \text{space}_{jt} + \beta_2 \text{space}_{jt}^2 + \beta_3 \text{space}_{jt}^3 \\ & + \beta_4 \text{loc.B}_{jt} + \beta_5 \text{period2}_{jt} \\ & + \beta_6 \text{loc.B}_{jt} \times \text{space}_{jt} \\ & + \beta_7 \text{loc.B}_{jt} \times \text{space}_{jt}^2 + \beta_8 \text{loc.B}_{jt} \times \text{space}_{jt}^3 \end{aligned}$$

This ‘compromise’ model is similar in spirit to the very flexible parametric model we actually apply below and which we describe in the following sub-

section. It is substantially more flexible than the simple additively separable linear model, reducing the risk of a large model bias. On the other hand, it is linear in the parameters (thus the subscript *FLM* for flexible linear model) and is thus easily estimated by least squares. Because it has eight parameters (rather than just four), its variance will be slightly larger than that of $\hat{m}_{LASM}(\mathbf{x}_{jt})$ and may thus offer a better bias-variance balance. In our empirical work, we take model (2.1) to its limits with regard to flexibility by considering nonlinear basis functions of the original continuous regressors and by including almost all possible interactions. We then use three constrained least squares procedures, ridge regression, the Lasso and the Elastic Net, to automatically shrink estimated parameters in order to keep the variance in check. The main advantage of this approach is that it offers an automatic, data-driven and computationally tractable solution to the key question in AVM specification search: arriving at an estimated model that presents a good compromise between bias and variance and is thus a sound basis for fast and accurate valuations.

2.2 A very flexible parametric model

We will now describe the details of our approach. Suppose first that there were only continuous property features, such as floor space, and that for each transacted property we would observe p such characteristics: X_1, X_2, \dots, X_p . In order to allow for enough flexibility of the functional forms of the relationships between price and these continuous characteristics, we employ a basis expansion in these variables. The model for the conditional mean of prices becomes:

$$m_{FLP}(\mathbf{x}_{jt}) = \beta_0 + \sum_{m=1}^M \beta_m h_m(X_{1t}, \dots, X_{pt}) \quad (5)$$

This model is linear in the parameters β_1, \dots, β_M but via the basis functions $h_m(X_1, \dots, X_p)$ we allow for non-linearity. We use the following basis functions:

$$h(X_{1t}, \dots, X_{pt}) = \begin{cases} t_\theta(X_j) & \text{for each } j \text{ and each } \theta \\ t_{\theta_1}(X_j) \cdot t_{\theta_2}(X_k) & \text{for each } j \text{ and } k \text{ if } \theta_1 \neq \theta_2 \end{cases} \quad (6)$$

where $t_\theta(X)$ denotes a Box-Cox-transformation of the following form²

$$t_j^{(\theta)} = \begin{cases} x_j^\theta & \text{for } \theta \in \{\pm 2, \pm 1, \pm 0.5\} \\ \log(x_j) & \text{for } \theta = 0 \end{cases}. \quad (7)$$

That is, the first type of basis function considers one explanatory variable at a time, transformed in one of the seven possibilities offered by the class of Box-Cox-transformations. Since we have $p = 18$ continuous characteristics³, this results in $18 \times 7 = 126$ basis functions of the form $h(X_1, \dots, X_p) = t_\theta(X_j)$. It is easiest to think about them as a set of transformed versions of the original explanatory variables. Using `space` to illustrate, the model will include `space`⁻², `space`⁻¹, `space`^{-0.5}, `log(space)`, `space`^{0.5}, `space` and `space`², each with its own β coefficient.

The second type of basis functions $h(X_1, \dots, X_p) = t_{\theta_1}(X_j) \cdot t_{\theta_2}(X_k)$ considers interactions between any of the 126 Box-Cox transformed original explanatory variables. As an example, $t_{\theta_1=0.5}(\text{space}) \cdot t_{\theta_2=0}(\text{age}) = \text{space}^{0.5} \times \log(\text{age})$ is a standard interaction terms of two transformed versions of floor space and age of the property. Products of transformed versions of the same explanatory variable are also considered, provided the transformation parameters differ between the two factors. Using $X_j = \text{space}$ to illustrate, $t_{\theta_1=0.5}(X_j) \cdot t_{\theta_2=-2}(X_j) = \text{space}^{0.5} \times \text{space}^{-2} = \text{space}^{-1.5}$ is such an example.⁴ Because we start with 126 transformed variables, this gives us a total of $\binom{126}{2} = 7875$ such product variables.⁵ 54 of these are constant (for example the interaction $X_j^2 \cdot X_j^{-2} = 1$) and 72 duplicate an already existing transformed variable (e.g. $X_j^1 \cdot X_j^{-0.5} = X_j^{0.5}$). These need to be removed, leaving us with 7,749 additional variables in the model from the interaction terms. Hence, the basis expansion from the continuous regressors gives a total of $126 + 7749 = 7875$ terms in (5).

²Box and Cox, 1964

³A detailed description of the continuous and discrete characteristics in our data is given in section 3. We refer to them synonymously as ‘characteristics’, ‘covariates’, ‘features’, ‘explanatory variables’ and ‘predictors’.

⁴To the contrary, $t_{\theta_1=0.5}(X_j) \cdot t_{\theta_2=0.5}(X_j) = \text{space}^{0.5} \times \text{space}^{0.5} = \text{space}^1$ is *not* a proper product as both factors use the same transformation parameter, i.e. $\theta_1 = \theta_2 = 0.5$.

⁵The interactions also widen the spectrum of polynomials included for each continuous covariate, which now also include x^θ for $\theta \in \{\pm 3, \pm 2.5, \pm 1.5\}$.

Next, we consider the 81 discrete, binary features in our data which we generically denote as D_j . They include dummy variables for time periods (years) and regions (districts of Berlin). They are all entered with their own coefficient. Moreover, all pairwise interactions are also created for them, leading to $\binom{81}{2} = 3,240$ additional binary variables of the form $D_{jk} = D_j \cdot D_k$ for all $j \neq k$.⁶ While these interactions also serve the purpose of enhanced model flexibility, they are easily interpretable and allow for different valuations regarding certain characteristics across time, space and in combination with other characteristics.⁷ Our flexible parametric model now reads:

$$\begin{aligned}
 m_{FLP}(\mathbf{x}_{jt}) &= \beta_0 + \sum_{m=1}^{7875} \beta_m h_m(X_{1t}, \dots, X_{pt}) \\
 &\quad + \sum_{j=1}^{18} \beta_j^d D_{jt} + \sum_{\text{all } j, k: j \neq k} \beta_{j,k}^d D_{jt} \cdot D_{kt}
 \end{aligned}$$

Moreover, we include pairwise interactions between the 81 uninteracted binary variables and the 18 untransformed continuous variables, creating 1,458 additional covariates. The interactions, again, further enhance model flexibility by allowing for different effects of continuous regressors for instance across time and space. Our count of predictors and parameters on the right hand

⁶Because certain characteristics never occur simultaneously in the data, some of these binary interactions are zero for all observations and consequently dropped in the estimation stage.

⁷For example, the value of basement storage space may be different across Berlin's districts or may have changed over time. Also, certain characteristics may lose or gain value in combination with others. Paying a premium for basement storage space may be less attractive if the apartment itself contains a storage rooms.

side of our model now is $7875 + 81 + 1458 = 9414$ and the model becomes:

$$\begin{aligned}
m_{FLP}(\mathbf{x}_{jt}) &= \beta_0 + \sum_{m=1}^{7875} \beta_m h_m(X_{1t}, \dots, X_{pt}) \\
&+ \sum_{j=1}^{81} \beta_j^d D_j + \sum_{\text{all } j, k: j \neq k} \beta_{j,k}^d D_{jt} \cdot D_{kt} \\
&+ \sum_{j=1}^{81} \sum_{k=1}^{18} \beta_j^{dx} D_{jt} \cdot X_{kt}
\end{aligned} \tag{8}$$

At last, we include a set of detailed binary indicators for a property's location and time of sale that we do not interact with other discrete or continuous regressors. Specifically, these are dummy variables indicating the borough (German: *Ortsteil*) of a property and its quarter of sale.⁸ We do this because location is typically seen as one of the most important determinants of a property's value and because we use data from several years of a dynamic property market. On the other hand, it is neither plausible nor feasible to let the effect of, say, floor size on property prices change every quarter or every borough. This is why we include the more crudely coded year dummies and district dummies as part of the fully interacted binary indicators described in the previous paragraph and the more finely graded borough and quarter dummies as a separate group of un-interacted indicators that close out the specification. Thus, the final model is

⁸There are 96 boroughs in Berlin and 40 quarters in each estimation sample.

$$\begin{aligned}
m_{FLP}(\mathbf{x}_{jt}) &= \beta_0 + \sum_{m=1}^{7875} \beta_m h_m(X_{1t}, \dots, X_{pt}) \\
&+ \sum_{j=1}^{81} \beta_j^d D_j + \sum_{\text{all } j, k: j \neq k} \beta_{j,k}^d D_{jt} \cdot D_{kt} \\
&+ \sum_{j=1}^{81} \sum_{k=1}^{18} \beta_j^{dx} D_{jt} \cdot X_{kt} \\
&+ \sum_{j=1}^{95} \beta_j^b B_{jt} + \sum_{j=1}^{39} \beta_j^q Q_{jt}
\end{aligned} \tag{9}$$

At this point, 9,549 covariates are available for predicting the price. Clearly, this is a very flexible model but, in terms of (3), a specification that is likely to suffer from very imprecise estimates, if it can be fit at all. A may thus be gained in estimation precision and therefor prediction accuracy if the model complexity can be reduced in the right way. This is what Ridge regression and the Lasso try to achieve. To simplify the notation for our discussion of these two procedures, we rewrite the model simply as

$$m_{FLP}(\mathbf{x}_{jt}) = \sum_{m=1}^M \beta_m z_{m,jt} \tag{10}$$

where all predictors on the right hand side of (9) have been consecutively numbered and are denoted as z_m with the corresponding consecutively numbered coefficients simply denoted as $\beta_j, j = 1, \dots, M = 9549$

2.3 Regularization

The model described in the previous section can, in principle, adapt very flexibly to the patterns in the data. This flexibility comes at the price of a huge number of parameters. Fitting this linear model via Least Squares results in an estimation bias of zero. However, this may neither be computationally

possible not statistically desirable.⁹ Indeed, there may be a sizeable gains in terms of variance reduction by using a biased estimator that ‘regularizes’ the complexity of the model by minimizing the sum of squared residuals subject to a constraint on flexibility. This is achieved by shrinking the coefficients of ‘unimportant’ regressors. From (1) above it is clear that such a biased but less variable estimator may result in an overall reduction of the expected squared prediction error. Indeed, the question of how many and which regression coefficients should be shrunk can be seen as synonymous with variable or model selection. Ridge regression, the Lasso and the Elastic Net offer an objective answer to this question, basically by shrinking coefficients of regressors that provide little information either because they hardly vary across properties or are highly correlated with other regressors and thus provide no additional information. It is reasonable to assume that both phenomena occur in the problem at hand. The correlation between several characteristics in the original dataset is considerable, for example between the *size of all residential and shared units* and the *size of all residential units*. The transformations and interactions described above only aggravate the problem. If, for instance, a continuous variable is multiplied by a dummy variable that predominantly contains the value 1, this new variable is a nearly perfect copy of the original continuous feature. Although only nonlinear transformations are used, correlations between covariates and their transformations can still be high. Since the only variables excluded from the analysis are those unusable due to missing values or purely administrative, redundant information is potentially contained in the data which motivates the need for regularization.

2.4 Ridge regression

Ridge regression, introduced by Hoerl and Kennard (1970), regularizes the problem by imposing penalties on the size of the estimated parameters, resulting in a unique solution even if the regressor matrix is not of full rank. The ridge estimate for the parameter vector β is found by minimizing the residual

⁹Computationally, if the number of parameters exceeds the sample size, there is no unique least squares solution. See Fahrmeir et al. (2009), p. 61.

sum of squares, subject to the sum of the squared parameters not exceeding a specified threshold t . To make the parameters independent of scale, all covariates are standardized.

$$\widehat{\beta}^R = \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \beta_0 - \sum_{m=1}^M \beta_m z_{m,it})^2 \right\} \text{ s.t. } \sum_{m=1}^M \beta_m^2 \leq t \quad (11)$$

Equivalently, this may be written as (citehastie2009, p. 63),

$$\widehat{\beta}^R = \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \beta_0 - \sum_{m=1}^M \beta_m z_{m,it})^2 + \lambda \sum_{m=1}^M \beta_m^2 \right\} \quad , \quad (12)$$

since for any threshold $t \geq 0$, a tuning parameter $\lambda \geq 0$ exists, which results in exactly the same estimator. The choice of the tuning parameter controls the amount of parameter shrinkage. Choosing a very large threshold t results in equivalence between Ridge regression and OLS. The smaller t (or, equivalently, the larger λ), the more shrinkage is performed.¹⁰

In matrix notation, the Ridge solution is given by

$$\widehat{\beta}^R = (\mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}^T \mathbf{p} \quad . \quad (13)$$

where \mathbf{p} is the (centered) vector of prices and the regressor matrix \mathbf{Z} does not include a column of ones for the intercept and standardized versions of the $z_m, m = 1, \dots, M$. Ridge estimation is biased towards zero. The size of the bias depends on the choice of the tuning parameter λ . We choose λ by cross validation. The negative effect of introducing a bias may be outweighed by the fact that Ridge regression, in general, has a lower estimation variance than unconstrained least squares, see Zou and Hastie (2005).

¹⁰No restriction is imposed on the parameter β_0 to make the procedure independent of the particular level of the average price, see Hastie et al. (2009), p. 64. In fact, centering the vector of prices allows estimation of the model without an intercept and recovering it afterwards.

Graphically, the ridge constraint forms a circle in a two-parameter problem (see the red circle in Figure 1). The Ridge solution is the smallest value of the residual sum of squares function within that circle. The picture illustrates why the Ridge estimator has, in general, a smaller variance than unconstrained least squares (OLS). The OLS solution can, depending on the data, lie anywhere in the plane spanned by β_1 and β_2 , while the Ridge solution is guaranteed to lie within the depicted circle.¹¹ While Ridge regression may thereby achieve a better bias-variance compromise and thus a superior predictive performance, it does not perform variable selection as it generally does not shrink coefficients all the way to zero. This is done by the Least Absolute Shrinkage and Selection Operator (Lasso).

2.5 Lasso regression

Lasso employs a slightly altered constraint which limits the sum of the *absolute values* of the coefficients. This ensures that not only are coefficients shrunk but some are set to exactly zero, turning the procedure into a ‘Selection Operator’. The Lasso was first introduced in Tibshirani (1996) to combine the favorable properties of Ridge regression with the desire to find an interpretable model. The Lasso estimator is defined by

$$\begin{aligned} \widehat{\beta}^L &= \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \sum_{m=1}^M \beta_m z_{m,it})^2 \right\} \text{ s.t. } \sum_{m=1}^M |\beta_m| \leq t \\ &= \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \sum_{m=1}^M \beta_m z_{m,it})^2 + \lambda \sum_{m=1}^M |\beta_m| \right\} . \end{aligned} \quad (14)$$

No closed form solution exists for the estimator in case of more than one predictor, but the problem is convex and a solution can be found iteratively.

Figure 1 shows that the constraint of the Lasso has the form of a diamond with edges even for a two parameter problem. If the solution occurs at such an edge, a parameter is set to zero, resulting in variable selection. The number of

¹¹Ridge regression is able to find a solution for $M > N$ problems and handles highly correlated variables more effectively than unconstrained least squares.

edges rises exponentially. For two parameters, the depicted diamond has four edges. For three parameters, the constraint region is a cube with eight edges. Each additional parameter doubles the number of edges, creating numerous opportunities for parameters to be set to zero¹². The amount of parameter shrinkage and variable selection depends on the choice of the tuning parameter. If a large enough threshold t (or, equivalently, a small enough tuning parameter λ) is chosen, the OLS solution falls within the depicted diamond and neither shrinkage nor selection are performed. A large value of λ or a small threshold t lead to substantial shrinkage with many parameters being set to zero. We select λ by cross validation.

The Lasso exhibits several shortcomings, though. Its nonzero parameter estimates tend to be biased towards zero. Also, when dealing with highly correlated variables the Lasso performs poorly. Slight changes of the tuning parameter λ can have wild effects on estimated parameters, and from a group of highly correlated variables, the Lasso tends to select some and discard others in a rather random fashion. The Elastic Net, a Generalization of the Lasso tries to overcome these shortcomings at the cost of an additional tuning parameter.

2.6 Elastic Net regression

The Elastic Net (introduced by Zou and Hastie (2005)) is a compromise between Ridge and Lasso. Its constraint on the parameters is a mixture between the L_1 penalty of Lasso and the L_2 penalty of Ridge regression. For any

¹²See Hastie et al. (2015), p. 12.

$\alpha \in [0, 1]$ the Elastic Net estimator is defined as

$$\widehat{\beta}^E = \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \sum_{m=1}^M \beta_m z_{m,it})^2 \right\} \quad (15)$$

$$s.t. \sum_{m=1}^M (\alpha |\beta_m| + (1 - \alpha) \beta_m^2) \leq t$$

$$= \arg \min_{\beta} \left\{ \sum_{i=1}^n \sum_{t=1}^T (P_{it} - \sum_{m=1}^M \beta_m z_{m,it})^2 + \lambda \sum_{m=1}^M (\alpha |\beta_m| + (1 - \alpha) \beta_m^2) \right\} . \quad (16)$$

Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$) are special cases of the Elastic Net. We choose its two tuning parameters α and λ by cross validation.

The contours of the Elastic Net constraint are depicted and compared to those of Ridge and Lasso in Figure 1. With the Lasso's constraint the Elastic

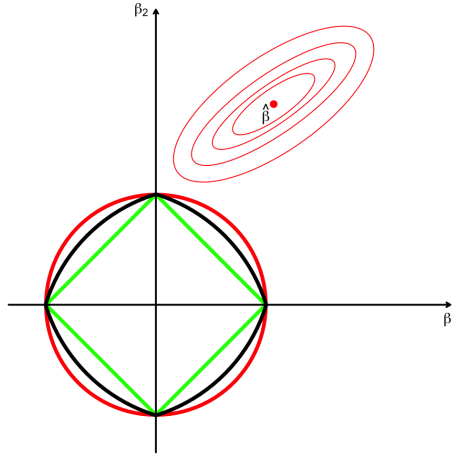


Figure 1: Comparison between the constraint regions of the Lasso (green), Ridge (red) and Elastic Net (black). The ellipses are the contours of the RSS-function. $\widehat{\beta}$ is the OLS solution. (Own representation based on Zou and Hastie, 2005)

Net constraint shares the edges, which are indispensable for variable selection. Otherwise, the constraint region has a rounded form like the Ridge constraint.

The Elastic Net is thus able to perform variable selection and parameter shrinkage simultaneously, like the Lasso, and is an improvement in two key aspects: the number of variables selected by the Elastic Net is not limited by the sample size, and, when dealing with groups of highly correlated covariates, the Elastic Net tends to include or discard the entire group together. Therefore, the Elastic Net has become especially popular for problems with very wide data ($M \gg N$) where the true model is sparse, like in genomics¹³ However, the Elastic Net does not solve the issue that nonzero coefficients are biased towards zero.

2.7 Prediction Performance

We apply Ridge Regression, the Lasso and the Elastic Net to estimate model (10) and compare the predictive performance of the estimated models with a rolling window design. For each window, we split the data into an estimation and validation sample. Each estimation sample contains 10 years (or 40 quarters) of transactions data. The first estimation window contains all condominiums sold between the 1st quarter of 1996 (1996Q1) and the last quarter of 2005 (2005Q4). The model fitted to this data by either Ridge Regression, the Lasso or the Elastic Net is then used to predict the prices of properties sold in the following quarter, 2006Q1, our first validation sample.¹⁴ We then shift the time window by one quarter to the right, resulting in an estimation window from 1996Q2 until 2006Q1 and a validation or test sample of 2006Q2. We proceed this way time until we reach our 32nd and final estimation window. It extends from 2003Q3 to 2013Q3 and is used to fit models that predict the properties in our last validation sample, condominiums sold in the fourth quarter of 2013. This way, we obtain 32 validation samples that can be used to assess and compare the predictive performance of the three procedures, as well

¹³See Hastie et al. (2009), p. 662.

¹⁴We thereby assume that typically no data is available on transactions in the current quarter in which valuations are formed because of the rather protracted process of finalizing a property transaction. The assumed one quarter delay in data processing is too optimistic for the administrative data collection procedure behind our data where contracts are entered into the transaction data base with an average delay of two quarters.

as a very simple linear model as in equation (2). This model, which we simply refer to as the OLS procedure, uses floor space and its square, the number of rooms and district and year dummies as regressors.

The starting point of our assesment of predictive performance of our candidate procedures is the individual prediction error that can be computed for any observation in a validation sample for each of our four candidate procedures:

$$\hat{e}_{jt}^k = P_{jt} - \hat{P}^k(\mathbf{x}_{jt}) \quad \text{for } k \in \{\text{Ridge,Lasso,Elastic Net, OLS}\} \quad (17)$$

where P_{jt} is the actual transaction price of property j and \mathbf{x}_{jt} is its vector of (untransformed) characteristics. Negative errors imply that the AVM prediction is larger than the actual transaction price, while positive errors imply the AVM prediction is smaller than the actual price.

From these prediction errors, we compute several summary measures of performance. The root mean squared prediction error (RMSPE)

$$\text{RMSPE}_k = \frac{1}{N} \sum_{j=1}^N (\hat{e}_{jt}^k)^2$$

is the empirical analogue of the expected squared prediction error in equation (1) above, apart from using the square root to transform the criterion to conventional EUROS. Here and in all other performance measures, the sum is understood to run over all properties in all validation samples even though we computed the criteria also for individual validation quarters. We furthermore complement this L_2 measure with the following L_1 measures that are built on absolute rather than squared errors: the mean absolute prediction error (MAPE) and the median absolute prediction error (MDAPE)

$$\text{MAPE}_r = \frac{1}{N} \sum_{i=1}^N |\hat{e}_{jt}^k| \quad \text{MDAPE}_r = \text{Med}|\hat{e}_{jt}^k|_{i=1}^N.$$

We also consider the relative errors

$$\hat{e}_{jt}^{k*} = \frac{P_{jt} - \hat{P}^k(\mathbf{x}_{jt})}{P_{jt}}$$

and compute their average, the mean absolute relative prediction error (MARPE), and their median, the median absolute relative prediction error (MDARPE). Because all these four mean and median performance measures are based on absolute prediction errors, they can be used to check for systematic over- or underprediction. A different perspective on prediction performance is offered by the fraction of predictions which don't over- or underestimate the actual sales price by more than 10 or 20 percent, respectively,

$$Q_{10} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(|\hat{e}_{jt}^k| \leq 0.1) \quad , \quad (18)$$

$$Q_{20} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(|\hat{e}_{jt}^k| \leq 0.2) \quad . \quad (19)$$

which we also report below. Depending on the AVM application, some of these criteria will have more relevance than others when discriminating between different candidate models. They are all neutral in the sense that over- and underprediction is treated symmetrically. In some AVM applications, other criteria have to be considered. If the AVM is used for instance for risk management purposes, such as the evaluation of loss severity for a portfolio of mortgages, banks are likely to be concerned about large overvaluations and may even favour an AVM method that has the tendency to err on the cautious side.

Finally, in all prediction performance comparisons, the issue arises whether the dominant performance of a method only holds in the particular test- or validation sample or whether the conclusion can be extended to the underlying population from which more observations could be obtained in the future. In order to extend the scope of our comparison beyond the properties in the validation samples, we use the test proposed by Diebold and Mariano (1995) and improved by Harvey et al. (1997) to examine the signif-

ificance of differences in predictive accuracy. This test considers the sample mean loss differential which is computed in the following way. First, we take the individual prediction errors \hat{e}_{jt}^k as described in equation (16), for models $k \in \{Ridge, Lasso, ElasticNet, OLS\}$. In order to examine the significance of predictive accuracy, the test is based on functions $d(\hat{e}_{jt}^{k_1}, \hat{e}_{jt}^{k_2})$ of matched prediction errors. In our application, we use squared pwer loss function. If methods 1 and 2 were of equal predictive quality (the null hypothesis), the sample mean loss differential

$$\bar{d} = \frac{1}{T} \sum_{t=1}^T d(\hat{e}_{jt}^{k_1}, \hat{e}_{jt}^{k_2}) \quad (20)$$

should be approximately normally distributed with mean μ and $2\pi f_d(0)/T$ as variance where $f_d(0)$ is the spectral density of the loss differential at frequency 0. A one sided test can be used to test whether model (1) performs better than model (2) and vice versa. The Diebold-Mariano test statistic

$$S_1 = \frac{\bar{d}}{\sigma_d} \quad (21)$$

under the null hypothesis is standard normally distributed. The modification by Harvey et al. (1997) uses an unbiased estimator for the variance of \bar{d} and takes the critical values from Student's t-distribution with $(n - 1)$ degrees of freedom. For details see Diebold and Mariano (1995) and Harvey et al. (1997).

3 Data

3.1 The original dataset

The data used for this analysis is provided by Berlin’s Committee of Valuation Experts (Gutachterausschuss für Grundstückswerte, GAA). The GAA registers each transaction of real estate by reporting all major details of the sales contracts into a transaction database (Automatisierte Kaufpreissammlung, AKS). Here, only an excerpt of the entire database is used, limiting the dataset to roughly 190,000 condominiums sold between 1996 and 2013. Each transaction is described by its price, time, location and detailed information on the characteristics of the condominium and the building, as well as the contracting parties.

Tables 3 through 7 show the summary statistics for all relevant variables. On average, a condominium was worth 131,729.80 € or 1,658.48 € per m^2 taking all transactions in the sample into account. Not surprisingly, most apartments had some cooking spaces, a toilet and a heating system. About half of all transactions have a balcony and a storage room. About 72% have access to a basement (either as owner or separate use privilege). On the contrary, garages and parking lots are very seldom part of a transaction (0.03% and 0.06% respectively).

Table 5 reports summary statistics for characteristics of the building complex the condominium is part of. We have information on the number of commercial units within the building and can distinguish between high rise apartment complexes and rather small units. On average, buildings (and hence the condominiums) are over 74 years old when they get sold. This corresponds to the construction phase after World War II. The number of sales appears to be relative constant over the years ranging from y .

In total, 106 variables characterize a transaction in the original dataset. Of which, 19 variables are continuous variables. The remaining ones are discrete variables treated as common factor variables in the model. As carried out above in section 2, there are two possible ways of how factor variables can become part

of a model. On the one hand factor variables can enter the regression equation only once (represented by their corresponding dummy variables) or on the other hand, several times as part of interactions with for instance continuous variables like the floor size. In most cases, we generated a lot of new interacted variables for each factor. But for some variables, namely subdistricts and quarter dummies, this would have lead a computationally unfeasible size of the design matrix. r

The location is determined by the condominium's address and a set of variables indicating to which block, subdistrict and district it belongs. It is also reported whether the condominium lies in the former West Berlin or East Berlin. The address of each observation is used to obtain lateral and longitudinal coordinates in the Soldner system¹⁵. The GAA regularly evaluates the quality of each residential location with regard to the attractiveness for owners and residents (stadträumliche Wohnlage), ranging from simple to excellent. This information is also provided for each observation. We included all locational information in the final regression models¹⁶. The only exception, as stated above, are the subdistrict dummies which enter the regression equation only once.

The condominiums themselves are described by their living area, number of bedrooms, location within the building and equipment. Also, the data contains information on additional space belonging to the condominium, like a balcony, hobby room, car park or storage space. It is also documented if the condominium is in poor state of repair.

The buildings containing the condominiums are characterized by the location within their blocks, their year of construction, number of storeys and the number and total space of all residential and nonresidential units. As with the condominiums themselves, poor state of repair is also documented.

Further, the data contains information on the legal status of the contracting parties, the type of contract, subsidies and the occupational status of the condominium. Some additional variables of purely administrative purpose are also

¹⁵See also für Stadtentwicklung (2015).

¹⁶This holds, of course, only for the more elaborate regression model (Ridge, Lasso and Elastic Net). We specified a very parsimonious baseline OLS model.

contained in the original dataset, but are of no relevance to the transaction and are therefore ignored.

The date of transaction was used to generate year and quarter dummies yielding a total of 90 factors (18 years and 72 quarters). For computational reasons, quarter dummies enter the regression only once (not interacted with other variables).

3.2 Data cleansing

The goal in preparing the dataset for the forthcoming analysis is to lose as little information as possible, while at the same time retaining a computationally feasible level of complexity.

Unlike in other works on similar datasets, no adjustment is made for inflation or changes in the real estate market¹⁷. Such effects are handled by including detailed information on the time of transaction in the model.

Variables are deleted if they are either of purely administrative nature, if their values were not continuously reported throughout the observed time period, or if they contain excessive amounts of missing values for other unknown reasons. For computational reasons, dummy variables indicating to which block and subdistrict an observation belongs are discarded, leaving the lateral and longitudinal coordinates and dummy variables indicating the district and affiliation to East or West Berlin as variables describing an observation's location. In addition, the distance to the city center is calculated, with the Pariser Platz serving as Berlin's center.

To allow log-transformation, several continuous variables need to be slightly adapted: the minimal value for the condominium's storey, distance to the city center, the bought share of the complex, the number of commercial units within the complex and their total size, and the number of independent auxiliary units are all set from zero to 0.1.

Regarding spaces detached from a condominium, German law (specifically the *Wohnungseigentumsgesetz*) differentiates between ownership (*Sondereigen-*

¹⁷See Kolbe et al. (2012), for example, who use house price indexes to convert prices.

tum) and the right of separate use (Sondernutzungsrecht). The original dataset distinguishes between these two legal terms for basements, attics, garages, parking spots and hobby rooms. For this analysis, this distinction is not adopted, but rather it is only stated whether the occupant has exclusive access to such a detached space, regardless of ownership.

Observations are only deleted from the dataset if they are atypical with regard to the contract, clearly misspecified or lack critical information. Thus, forced auctions and transactions where a condominium is only partially sold are removed. So, too, are condominiums where the reported average size of bedrooms is less than 8 m^2 , and two cases where the size is unknown, but the reported number of bedrooms are 43 and 67. Transactions for less than 20 € per m^2 are also deleted, since such a low price can only result from misspecification or substantial irregularities not displayed by the available information. If neither the size, nor the number of bedrooms is known, the observation is deleted. The same goes for observations where the apartment complex's area, number of storeys and total living space of the complex are jointly unknown. Most likely due to spelling errors or changes in the street name since the time of transaction, 19 observations are deleted because they cannot be located. Thus, 1672 observations are deleted, resulting in a loss of less than one percent and leaving 186,923 observations.

3.3 Imputation of missing values

Eight variables contain a manageable amount of missing values or values clearly indicating that the information is unknown or misspecified. Deleting all observations with any missing values would contradict the goal of retaining as much information as possible. The easiest approach would consist of filling in the median value of the non-missing values, but in the case of variables describing real estate, it is reasonable to assume that one can do better by exploiting dependencies between the characteristics for filling in the missing values¹⁸. In order to recover the missing information, linear regression imputation is used,

¹⁸Hastie et al. (2009) p. 332

with the exception of the building's age, where missing values are recovered by a sort of nearest neighbour algorithm. The danger with regression imputation lies in artificially driving up the correlation between covariates and underestimating the variability of the variable containing missing values¹⁹. However, the correlation between the characteristics in question is already very high, and no variable shows more than six percent missing values. Thus, the danger of distortion is acceptable for such a procedure. More elaborate approaches as stochastic regression imputation or multiple imputation²⁰ would have increased the already high computational burden of the given research question. Imputation is rather seen as a necessity, having to be dealt with in a quick and efficient fashion in order to secure the goal of minimal information loss. Thus, the used regression models are chosen by picking the variables deemed most likely to have high correlation with the dependent variable in question, and no transformations or interactions are included.

On 10,728 occasions the size of all residential units of the building in question is either missing or reported as smaller than the sold condominium. The area of the apartment complex, the number of residential units and the number of independent auxiliary units serve as predictors to recover the missing values. Where the area is also unknown, a prediction is made by using only the remaining two covariates.

623 occasions of an unknown area of the apartment complex are imputed via the building's number of storeys, the number of residential units and of independent auxiliary units.

The number of the building's storeys is unknown 942 times and is predicted via the area of the complex, the number of residential units and whether or not the building is equipped with an elevator.

The condominium's living space (250 missing values) and the number of bedrooms (2,582 missing values) serve as predictors for each other (observations where this information is jointly missing have already been excluded).

The total space of all residential and shared units is reported as smaller than the sold condominium on 10,669 occasions. These missing values are predicted

¹⁹van Buuren (2012), p. 12

²⁰See van Buuren (2012), chapters 1.3.5 and 1.4

via the total space of all residential units and the number of residential and shared units.

50 occasions of the number of residential units being reported as zero are imputed using the total size of all residential units.

The building's age is unknown for 365 observations. For this particular variable, regression imputation is unsuitable, and a neighborhood search is used to impute the missing values. Starting with a radius of 25 meters, all observations within a circle around the observations of interest are identified. The most frequent value of the surrounding buildings' age is taken as prediction for the missing value. If no observations are within the circle, the radius is widened by an additional 25 meters, until all missing values have been filled in. Although this method can (and on multiple occasions surely does) lead to poor predictions due to the heterogeneity of the Berlin map with regard to the era in which buildings were erect, it is considered more valuable to keep the observation and risk a misspecified building's age, rather than lose all the observation's information on the numerous other characteristics.

Different approaches exist to check the validity of the imputation results. All methods have in common that they compare whether the distribution of the variables before and after recovering the missing values remain comparable. Popular methods are graphical checks, numerical summaries or statistical tests²¹. Again, pointing to the fact that this is by no means the focus of this work, the easiest of these methods is used here and numerical summaries are employed for checking the imputation results. As reported in table 8, the range, quartiles and means are hardly affected for all eight variables subject to imputation, suggesting a successful imputation process.

A complete overview of all variables used for further analysis is given in tables 3 through 7. After the mentioned additions and transferring categorical variables into dummies, each transaction is now described by 19 metric variables (including the price) and 81 dummy variables (not including reference categories).

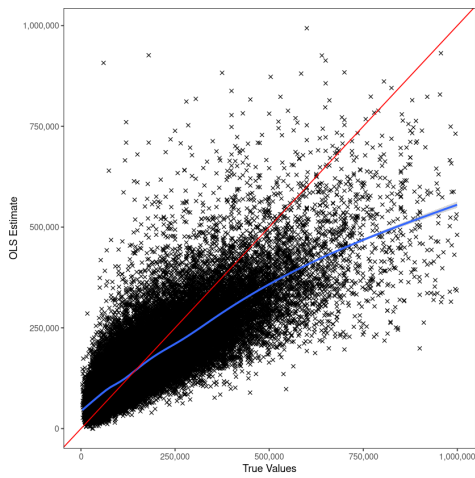
²¹See Nguyen et al. (2017)

4 Results

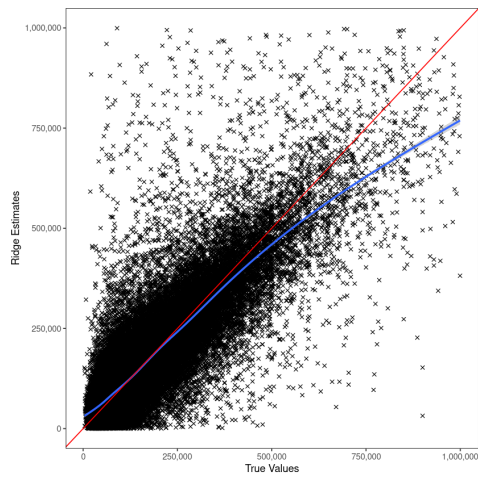
As described in subsection (2.7), we performed our prediction comparison in a rolling window design. Table 9 in the Appendix gives details about the size of the estimation and validation samples in each of our 32 prediction windows. On average, there are about 104,000 observations in an estimation sample and about 2,900 observations in a validation sample. In total, over all windows and validation sample, we have obtained 93,436 prediction errors for each of our four candidate procedures. We first present results from analyzing all these prediction errors together. Subsequently, we also look at results broken down by prediction windows.

We begin with a graphical analysis of the predictions underlying the prediction errors. The scatterplots in Figure 2 show how closely the predicted values of each procedure mimic their targets, the actual sales prices of the properties that were predicted by the AVM methods. Observations where either the actual price or the prediction exceeded 1,000,000€ are excluded from the plot to increase readability in the range of prices that include the bulk of the data. In each scatterplot, we have included a 45° line to help in identifying properties that are under- or overvalued by a candidate AVM. We have also included a univariate nonparametric estimate of the relationship between transaction prices and AVM predictions (blue line). Indeed, the comparison of the 45° line and the nonparametric smooth reveals the main finding of this part of the analysis. All models tend to overestimate low priced condominiums while we see a strong tendency to underestimate observations from the right tail of the price distribution. OLS appears to have the strongest tendency underestimate the values of more expensive properties. This is to be expected as it is the least flexible procedure with the greatest risk of a model bias. Ridge regression shows a particularly close agreement (i.e., small bias) between the nonparametric smooth and the 45° line for properties up to 500,000 €, whereas the Elastic Net seems to have the overall tightest agreement between nonparametric smooth and 45° line. On the other hand, Ridge regression displays a strong variation about the 45° line, especially for lower priced condominiums.

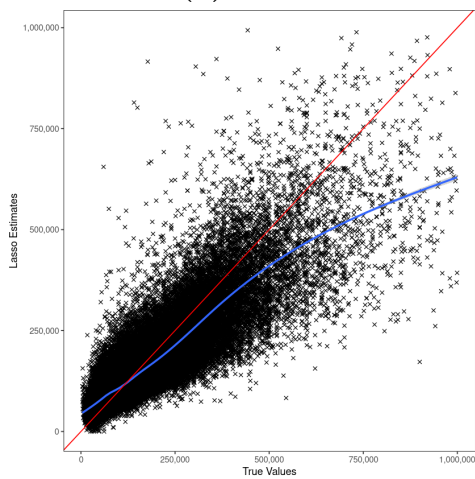
Figure 2: Scatterplots of true transaction prices vs. model estimates. Note: The blue line is a nonparametric fit, the red line is the 45° line .



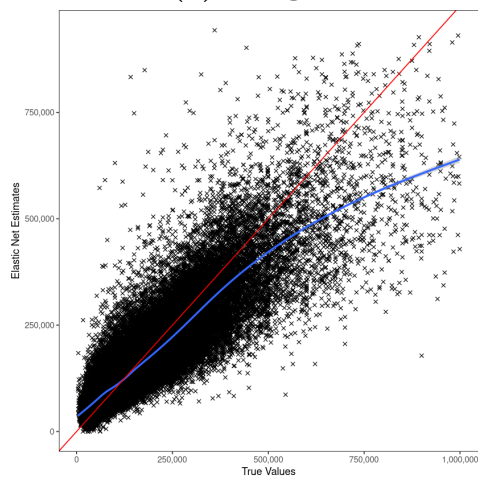
(1) OLS



(2) Ridge



(3) LASSO



(4) Elastic Net

The findings from the inspection of the scatterplots in Figure 2 are largely underscored by the summary measures of the prediction errors reported in Table 1. We have highlighted the result of the best performing procedure in

Table 1: Error metrics for all models aggregated over all forecast quarters.

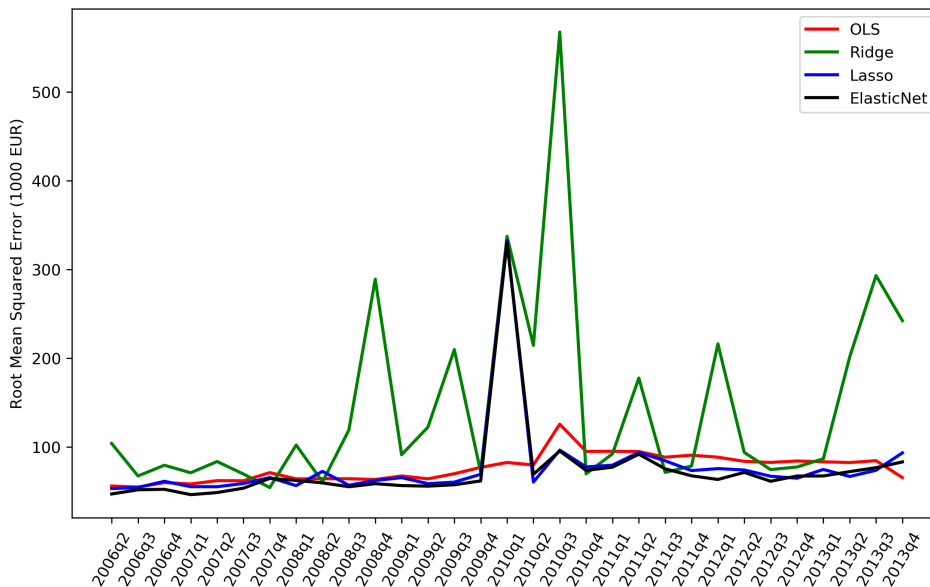
	RMSPE	MAPE	MDAPE	MARPE	MDARPE	Q10	Q20
OLS	80169	47085.1	31996.4	0.4544	0.3029	0.1689	0.3378
Ridge	181934	46816.3	25947.7	0.4444	0.2463	0.2330	0.4265
Lasso	88755	41841.9	27943.0	0.3997	0.2646	0.1959	0.3859
ElasticNet	85746	39176.4	25667.6	0.3619	0.2474	0.2085	0.4121

bold face for each criterion. Strikingly, OLS, the simplest AVM procedure has the lowest root mean squared prediction error. Apparently, it can compensate for its visible bias in Figure 2 by its very small estimation variance indicated by formula (3) above. Its model bias also shows up in all the mean and median absolute prediction error criteria where it always performs worst. Also as expected from Figure 2, the Elastic Net performs best with regard to all bias criteria and outperforms Lasso, the procedure it generalizes, with regard to all criteria. A particularly uneven performance is offered by Ridge regression. It displays the best performance with regard to the Q_{10} and Q_{20} criteria. That is, Ridge regression achieves the largest fraction of predictions within 10 or 20 percent of the actual sales price. However, it is by far the worst procedure in terms of the root mean squared prediction error. Apparently, it performs really well for ‘standard’ properties in the center of the distribution but extremely poorly for some properties at the fringe.

Indeed, this becomes even more evident in figure 3 which plots the RMSPE for each procedure over the course of the rolling window design. The RMSPE values for Ridge regression (connected by a line to enhance readability) show very large bumps for some evaluation periods²². In these periods, there were some properties for which Ridge regression delivered extremely false predictions. These errors are highlighted by RMSPE because of the built-in squaring

²²We see a similar bump in one quarter also for the Elastic Net.

Figure 3: Root mean squared error for every validation quarter of the rolling window design

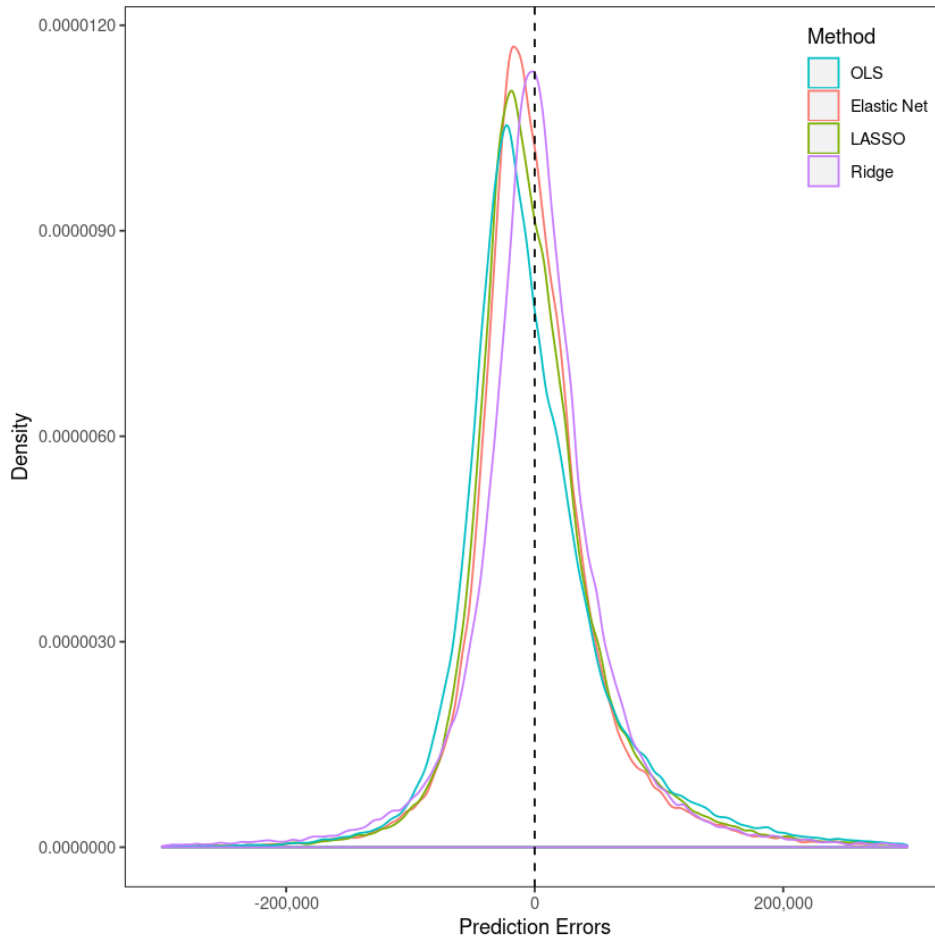


of prediction errors. Ridge regression shrinks coefficients but does not set them to zero. Its prediction formula thus carries a large number of nonzero parameters, thereby increasing the risk for such a bad outcome if properties display extreme values in the transformed feature space.

The uneven prediction performance of the Ridge regression AVM is also visible in Figure 4. It shows kernel densities of the prediction errors of all four models.²³ We have restricted the range of the prediction errors to the interval $-300,000$ to $300,000$. The extreme errors of Ridge regression are deliberately not shown as they would dwarf any differences that may exist within the bulk of the errors. The low median absolute errors of Ridge are reflected by its density being almost centered at zero (stemming from its high flexibility). Its large fractions of Q_{10} or Q_{20} are visible in its density being relatively tightly and symmetrically centered around the median. Indeed, from the perspective of Figure 4, which omits its extremely poor performance at the fringes, Ridge regression looks like a winner. All procedures, except for OLS, start with a

²³We used the Epanechnikov kernel and calculated the bandwidth according to the proposal by Venables and Ripley (2002).

Figure 4: Density plots of prediction errors of all models across all prediction quarters.



large number of parameters and then either shrink some of them or even set some of them to zero. Indeed, with their ability to do the latter, the Lasso and the Elastic Net offer the promise of ‘automatic’ variable selection. We explore the patterns of shrinkage and variable exclusion achieved by these procedures in Figures 5, 6 and 7. These plots are based on summing the estimated coefficients of each transformed feature z_m for the respective model across all prediction periods. In the plots, we ordered them by size. Hence, features are visible that often (i.e. in several prediction windows) ‘survive’ the shrinkage or selection algorithms of the procedures. We normalized the vertical axes of the figures by subtracting the minimum and dividing by the difference between maximum

and minimum over all coefficient sums. Our feature importance measure is thus exactly one at the maximum. As all variables were normalized prior to estimation, the size of their coefficients are comparable. The figures show that across all procedures, the squared floor size is the strongest predictor in this sense. Indeed, floor size appears ‘successful’ in various (interacted) forms and emerges as the most robust determinant in these AVMs.

Lastly, we performed an adjusted Diebold-Mariano-Test as described at the end of subsection 2.7. In a first run, we tested whether there is a significant difference between the prediction performance of any two models. In a second step, we performed a one-sided test which evaluated if one model has a significantly *better* performance than the other. After comparing each model with every other model, it turns out that except for Ridge every other model has no significant better forecast performance than OLS. Elastic Net performs better than the Lasso estimator. Table 2 reports the test results a “+” indicates that the null hypothesis can not be rejected and “-” if the hypothesis was rejected. If the two sided test found no significant difference between two models (“+”), the second test is discarded. These findings are in line with the error metrics. We use a squared loss function for testing the hypotheses and as OLS produces the smallest squared prediction errors the results are not surprising.

Table 2: Results of the Diebold-Mariano Tests. A ‘-’ indicates that the Null hypotheses was rejected on a 5% significance level.

	OLS		Lasso	
	indifferent	better than	indifferent	better than
OLS			+	
Lasso	+			
Ridge	-	+	-	+
Elnet	+		-	-

	Ridge		Elastic Net	
	indifferent	better than	indifferent	better than
OLS	-	-	+	
Lasso	-	-	-	+
Ridge			-	+
Elnet	-	-		

5 Conclusion

To be successful, AVMs need to combine rich data on past transactions with statistical procedures that can exploit the information in this data about the relationship of property prices and its determinants. The number of possible determinants can already be large at present and is likely to considerably grow in the future in the wake of digitization. At the same, the nature of property markets will keep the number of transactions per quarter at a relatively low level. This poses an interesting challenge for AVM modeling as statistical candidate procedures need to be able to cope with this situation of high-dimensional X and relatively small n . At the heart of this challenge is the well known trade off between bias and variance of the procedures. In this paper, we considered three procedures that are designed to ‘automatically’ solve this challenge and arrive at an estimated AVM that is a good compromise between bias and variance: Ridge Regression, the Lasso and the Elastic Net. All three start from a very flexible parametric model with a large number of transformed and interacted features. They then ‘tame’ this model by regularization, i.e. by shrinking regression coefficients or even setting some to zero. They achieve this via constrained least squares estimation of the coefficients, where the severity of the constraint on model flexibility can be chosen in a automatic, data-driven way via cross-validation.

We applied these procedures to a rich and reliable data set of condominium sales in Berlin, Germany, between 1996 and 2013 and compared their predictive performance in a rolling window design with each other and a simple linear OLS procedure. Stunningly, the very parsimonious OLS procedure make up for its apparent model bias with its very low estimation variance, resulting in the lowest overall root mean squared error. The more flexible procedures outperform OLS in terms of their smaller bias and do well in the center of the data but may do very poorly at the fringes. This result is showcased most vividly by the performance of Ridge regression which is a ‘winner’ in the center but delivers some extremely false predictions for other properties which ruins its performance in the mean squared prediction error sense. Hence, our results

suggest that Ridge Regression, the Lasso and the Elastic Net show potential as AVM procedures but need to be handled with care and are not the 'automated' solution to Automated Valuation Modeling that they may seem to be.

Acknowledgements

We are grateful to the German Science Foundation (DFG) for financial support via DFG Research Unit 2569: Agricultural Land Markets - Efficiency and Regulation.

Appendix

Table 3: Summary statistics for transacted condominiums. Condominiums's prices, space and equipment. Prices are in €, living spaces in m^2 . If only the mean is reported, the corresponding variable is a dummy, indicating whether the stated characteristic is available.

	Mean	Std. Dev.	Min.	Median	Max.
Price	131,729.80	120,912.50	2,000	98,134	7,000,000
Living space	73.93	32.75	12.00	66.44	564.27
Bedrooms	2.55	1.04	1	2	16
Poor state of repair	0.01				
<i>Equipment</i>					
Collective heating (<i>Reference</i>)	0.99				
Furnace	0.01				
No heating	0.0002				
Kitchen	0.87				
Cooking cell	0.10				
Pantry kitchen	0.01				
Bathroom	0.92				
Shower	0.08				
Separate toilet	0.14				
Hall	0.16				
Corridor	0.86				
Storage room	0.55				
Balcony	0.47				
Loggia	0.20				
Studio	0.001				
Basement	0.72				
Attic storage room	0.01				
Hobby room	0.01				
Garden	0.03				
Garage	0.03				
Parking spot	0.06				

Table 4: Summary statistics for transacted condominiums. Type of condominium and location within the building. If only the mean is reported, the corresponding variable is a dummy, indicating whether the stated characteristic is available.

	Mean	Std. Dev.	Min.	Median	Max.
<i>Location within the building</i>					
Storey	2.24	2.07	0	2	25
Upper floor (<i>Ref.</i>)	0.82				
Ground floor	0.17				
Elevated ground floor	0.008				
Lowered ground floor	0.002				
Basement floor	0.001				
Souterrain	0.0002				
<i>Type of condominium</i>					
Regular apartment (<i>Ref.</i>)	0.90				
Penthouse	0.001				
Duplex apartment	0.03				
Attic apartment	0.07				
Terrace apartment	0.002				
Shop apartment	0.001				

Table 5: Summary statistics for transacted condominiums. Characteristics of the building. Areas and sizes are in m^2 . If only the mean is reported, the corresponding variable is a dummy, indicating whether the stated characteristic is available.

	Mean	Std. Dev.	Min.	Median	Max.
Area of the complex	4,628.12	7,483.26	91.00	1,433.00	78,170.00
Bought share of the complex	0.04	0.05	0.00	0.02	0.94
Number of storeys	5.42	2.66	1	5	55
Number of allunits	77.93	98.75	1	38	786
Number of residential units	70.87	91.98	1	34	721
Number of commercial units	1.27	2.55	0	0	53
Size of all units	4,778.61	5,813.78	22	2,530	48,266
Size of all residential units	4,611.92	5,643.04	18	2,386	47,224
Size of all commercial units	167.70	634.19	0	0	18,251
Number of independent units	5.79	19.67	0	0	221
Building's age	74.76	37.30	1	80	240
Poor state of repair	0.003				
Leasehold	0.01				
Elevator	0.33				
<i>Type of building</i>					
Detached house (<i>Ref.</i>)	0.16				
House at block's edge	0.57				
Row house	0.12				
Semi detached house	0.001				
Townhouse	0.002				
Courtyard building	0.14				

Table 6: Summary statistics for transacted condominiums. Characteristics of the contract. The transaction day is measured in days before the last transaction, plus one. If only the mean is reported, the corresponding variable is a dummy, indicating whether the stated characteristic is available.

	Mean	Std. Dev.	Min.	Median	Max.
Private buyer	0.95				
Bought by legal entity (<i>Ref.</i>)	0.05				
Private seller	0.39				
Sold by legal entity (<i>Ref.</i>)	0.61				
Social housing	0.16				
Tax benefits	0.02				
Guaranteed rent	0.04				
Conversion to condominium (<i>Ref.</i>)	0.62				
No conversion	0.17				
Partial conversion	0.21				
Bought by tenant	0.05				
Rented out	0.29				
Unoccupied (<i>Ref.</i>)	0.66				
<i>Time of transaction</i>					
Transaction day	2,941.1	1,894.27	1	2,866.5	6,475
1996 (<i>Ref.</i>)	0.03				
1997	0.06				
1998	0.07				
1999	0.05				
2000	0.04				
2001	0.05				
2002	0.04				
2003	0.06				
2004	0.04				
2005	0.07				
2006	0.06				
2007	0.04				
2008	0.05				
2009	0.05				
2010	0.08				
2011	0.07				
2012	0.08				
2013	0.06				

Table 7: Summary statistics for transacted condominiums. Condominia’s location. If only the mean is reported, the corresponding variable is a dummy, indicating whether the stated characteristic is available.

	Mean	Std. Dev.	Min.	Median	Max.
Longitude	23,046.23	5,843.43	5,173.70	23,019.00	47,324.40
Latitude	19,755.76	5,011.72	1,856.40	19,958.40	36,076.30
Distance to city center	6,831.52	3,736.20	0.00	5,853.98	25,975.53
East Berlin	0.35				
West Berlin (<i>Ref.</i>)	0.65				
<i>District</i>					
Mitte (<i>Ref.</i>)	0.13				
Friedrichshain-Kreuzberg	0.10				
Pankow	0.13				
Charlottenburg-Wilmersdorf	0.18				
Spandau	0.05				
Steglitz-Zehlendorf	0.08				
Tempelhof-Schöneberg	0.13				
Neuköln	0.06				
Treptow-Köpenick	0.05				
Marzahn-Hellersdorf	0.02				
Lichtenberg-Hohenschönhausen	0.03				
Reinickendorf	0.05				
<i>Experts’ location rating</i>					
Simple	0.45				
Average	0.29				
Good	0.23				
Excellent (<i>Ref.</i>)	0.03				

Table 8: Descriptive statistics of variables containing missing data, before and after imputation.

	Min.	1st Quart.	Median	Mean	3rd Quart.	Max.
Space of all residential units						
Before imputation	18	1484	2433	4723	5743	47224
After imputation	18	1467	2386	4611	5474	47224
Space of all units						
Before imputation	22	1564	2578	4900	5937	48266
After imputation	22	1536	2530	4777	5708	48266
Number of residential units						
Before imputation	1	20	34	70.89	83	721
After imputation	1	20	34	70.87	83	721
Area of the apartment complex						
Before imputation	91	848	1433	4633	5288	78170
After imputation	91	848	1433	4628	5282	78170
Number of storeys						
Before imputation	1	4	5	5.425	6	55
After imputation	1	4	5	5.421	6	55
Condominium's living space						
Before imputation	15.40	53.10	66.45	73.93	87.10	564.27
After imputation	15.40	53.10	66.44	73.93	87.10	564.27
Number of bedrooms						
Before imputation	1	2	2	2.548	3	16
After imputation	1	2	2	2.551	3	16
Year of construction						
Before imputation	1775	1905	1935	1940	1968	2014
After imputation	1775	1905	1935	1940	1968	2014

Table 9: Summary statistics for forecast windows.

Year	Quarter	Observations		Hyperparameter λ		
		Training	Validation	Ridge	Lasso	Elnet
2006	2	93,880	2,416	7.47	0.75	1.49
	3	95,870	2,421	7.22	0.72	1.44
	4	97,337	3,561	7.04	0.70	1.41
2007	1	99,834	2,085	6.89	0.69	1.38
	2	98,979	2,751	7.36	0.74	1.47
	3	100,658	2,804	7.07	0.71	1.41
2008	4	101,710	3,105	7.12	0.71	1.42
	1	102,599	2,296	6.82	0.68	1.36
	2	100,497	2,682	7.82	0.73	1.46
2009	3	101,192	2,481	6.89	0.70	1.39
	4	101,299	2,714	7.21	0.72	1.44
	1	101,320	2,047	6.85	0.69	1.37
2010	2	96,567	2,775	7.28	0.73	1.46
	3	97,709	2,540	7.06	0.71	1.41
	4	98,112	2,952	7.26	0.73	1.45
2011	1	98,817	2,570	6.93	0.69	1.39
	2	98,235	3,315	7.28	0.73	1.46
	3	99,769	3,448	7.11	0.71	1.42
2012	4	101,342	3,801	7.43	0.74	1.49
	1	103,226	3,184	6.95	0.70	1.39
	2	103,660	3,723	7.27	0.73	1.45
2013	3	105,518	3,892	7.13	0.71	1.43
	4	107,371	4,560	7.31	0.73	1.46
	1	109,927	4,470	6.97	0.70	1.39
2013	2	111,320	3,188	7.24	0.72	1.45
	3	112,899	3,628	7.12	0.71	1.42
	4	114,657	3,943	7.24	0.72	1.45
2013	1	116,906	3,138	6.96	0.70	1.39
	2	116,970	3,244	7.19	0.72	1.44
	3	118,758	2,820	7.18	0.72	1.44
	4	119,843	882	7.21	0.72	1.44
	Mean	104,030.1	2,919.9			
	Sum		93,436			

Figure 5: Scaled feature importance of LASSO across all forecast periods.

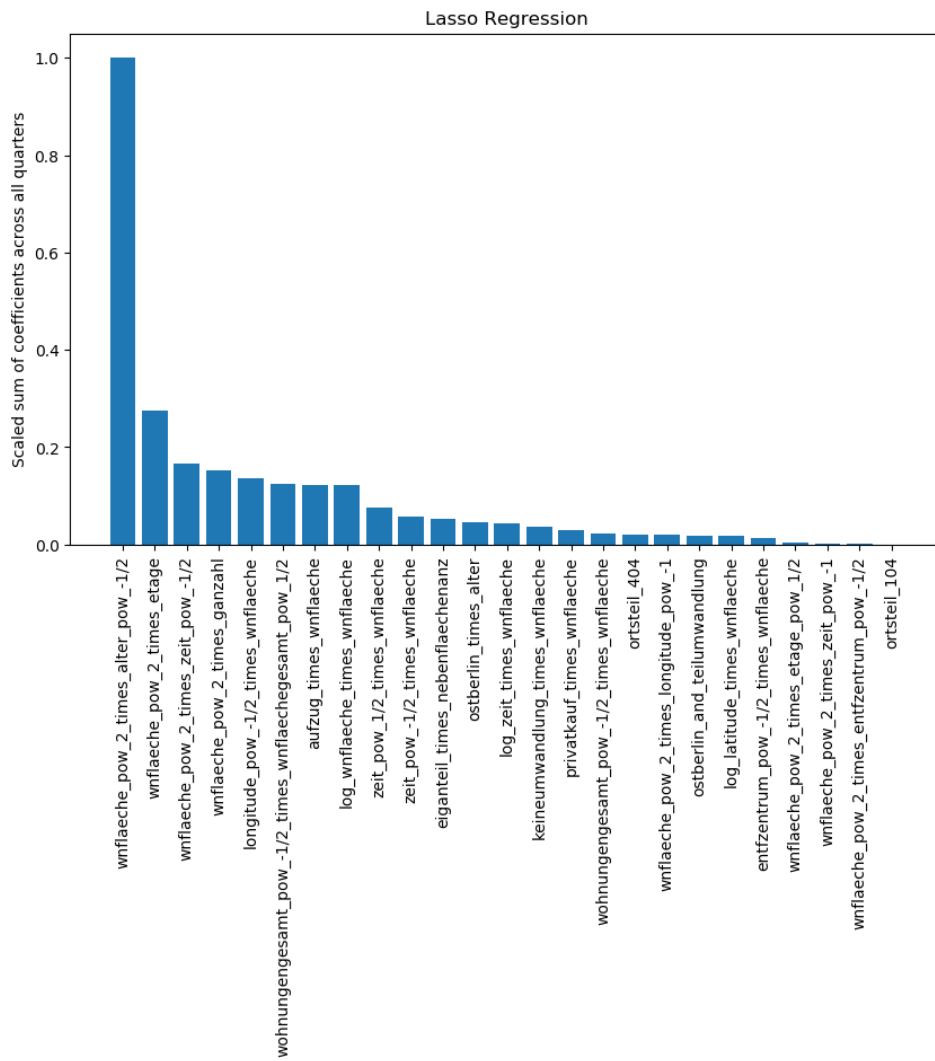


Figure 6: Scaled feature importance of Ridge across all forecast periods.

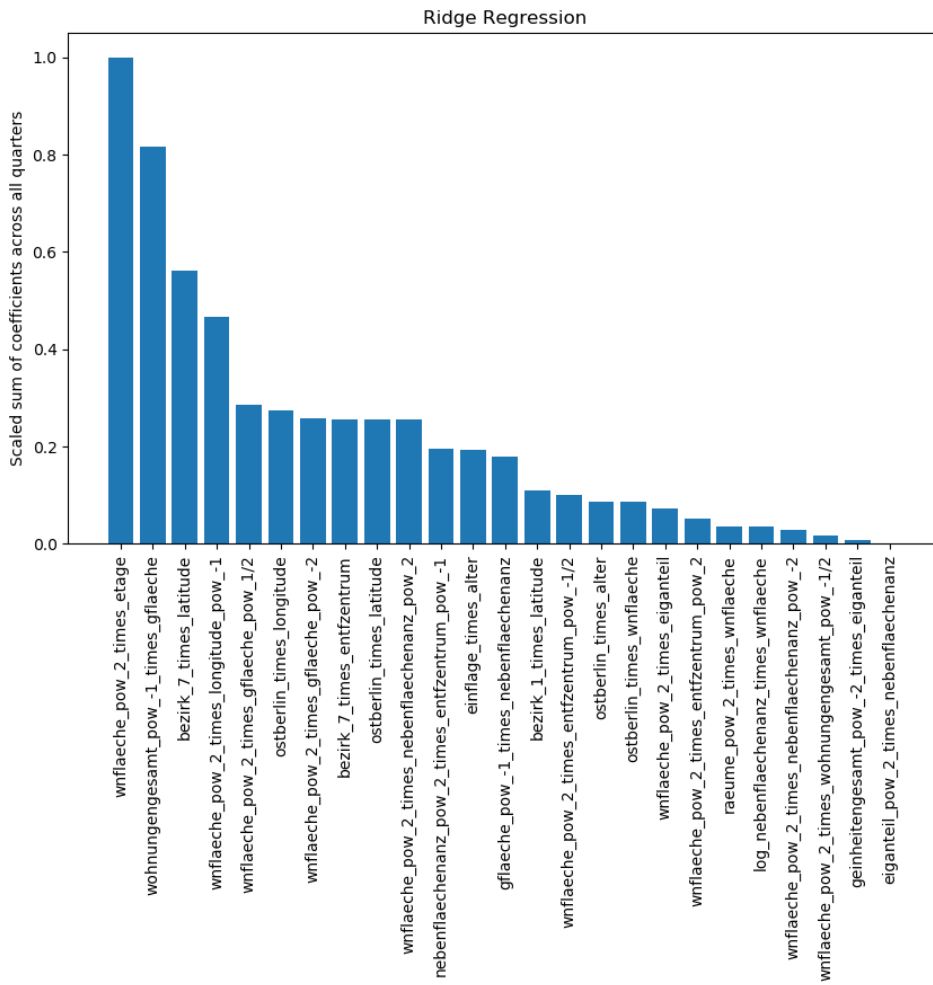
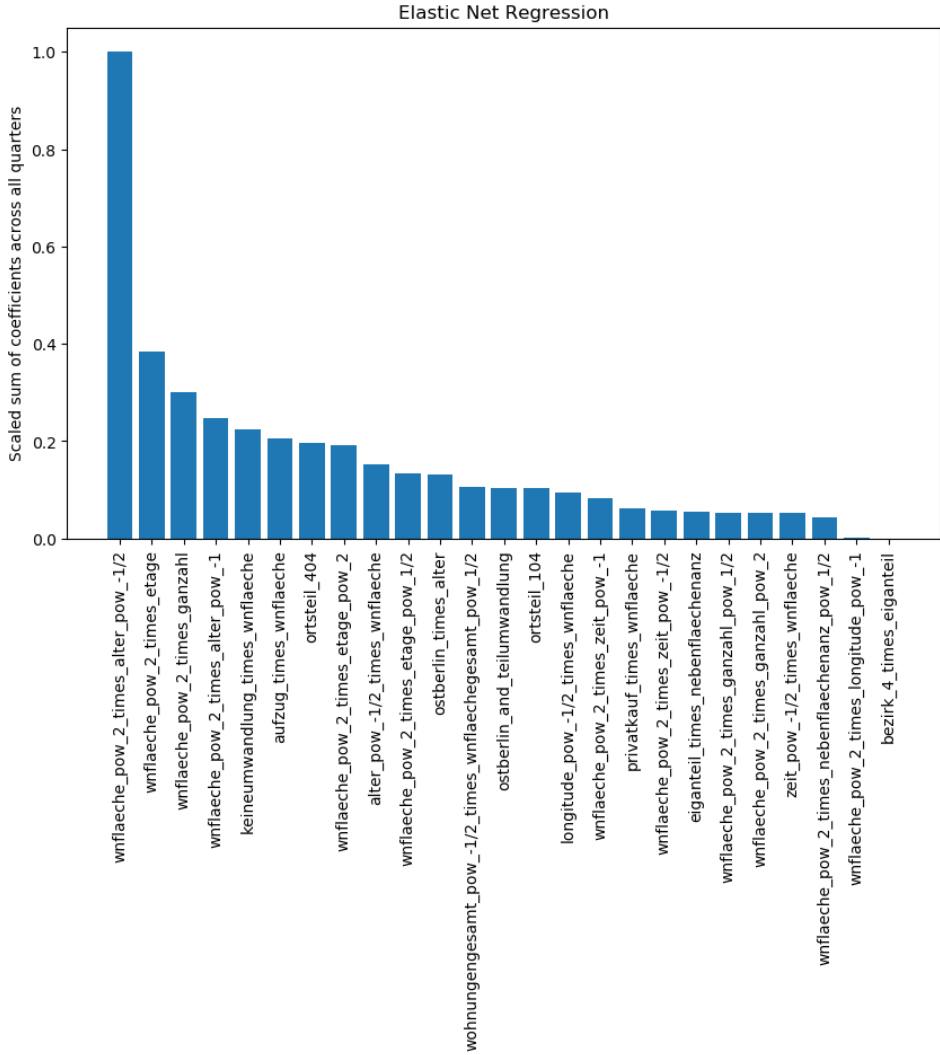


Figure 7: Scaled feature importance of Elastic Net across all forecast periods.



References

- Diebold, F. X. and Mariano, R.: 1995, Comparing forecast accuracy, *Journal of Business and Economics*.
- Fahrmeir, L., Kneib, T. and Lang, S.: 2009, *Regression — Modelle, Methoden und Anwendungen. (Zweite Auflage)*, Springer.
- für Stadtentwicklung, S.: 2015, Ein neues Lagebezugssystem für Berlin — Fragen und Antworten zu Modernisierungen im Vermessungswesen Berlin, http://www.stadtentwicklung.berlin.de/geoinformation/landesvermessung/etrs89/download/6_FAQ.pdf. [accessed: 02.03.2018].
- Harvey, D., Leybourne, S. and Newbold, P.: 1997, Testing the equality of prediction mean squared errors, *International Journal of forecasting* **13**(2), 281–291.
- Hastie, T., Tibshirani, R. and Friedman, J.: 2009, *The Elements of Statistical Learning — Data Mining, Inference and Prediction. (Second Edition)*, Springer.
- Hastie, T., Tibshirani, R. and Wainwright, M.: 2015, *Statistical Learning with Sparsity — The Lasso and Generalizations*, Springer.
- Hoerl, A. E. and Kennard, R. W.: 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1), 55–67.
- Kolbe, J., Schulz, R., Wersing, M. and Werwatz, A.: 2012, Location, location, location: Extracting location value from house prices, *DIW Berlin Discussion Papers* **1216**.
- Nguyen, C. D., Carlin, J. B. and Lee, K. J.: 2017, Model checking in multiple imputation: an overview and case study, *Emerging Themes in Epidemiology* **14**(8).
- Tibshirani, R.: 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

- van Buuren, S.: 2012, *Flexible Imputation of Missing Data*, Chapman&Hall/CRC.
- Venables, W. N. and Ripley, B. D.: 2002, *Modern Applied Statistics with S*, Springer.
- Zou, H. and Hastie, T.: 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Methodological)* **67**(2), 301–320.