Journal of Property Research

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

Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data

Prodosh E. Simlai

To cite this article: Prodosh E. Simlai (2021) Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data, Journal of Property Research, 38:4, 305-336, DOI: 10.1080/09599916.2021.1890187

To link to this article: https://doi.org/10.1080/09599916.2021.1890187

JOURNAL OF PROPERTY RESEARCH
2021, VOL. 38, NO. 4, 305–336
https://doi.org/10.1080/09599916.2021.1890187

Routledge
Taylor & Francis Group

Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data

Prodosh E. Simlai

Department of Economics and Finance, Nistler College of Business and Public Administration, University of North Dakota, USA

ABSTRACT

In this paper, we introduce machine-learning (ML) methods to evaluate one of the key concepts of real estate analysis – the prediction of housing prices in the presence of a large number of covariates. We use several supervised ML tools that are based on regularisation methods – notably Ridge, LASSO, and Elastic Net regressions – and discuss their relative performance in comparison to conventional OLS-based methods. Our empirical results show that the supervised ML methods provide a comprehensive description of the determinants of owner-occupied housing values in the census tracts of California. We find that, compared to the familiar worlds of OLS and WLS, the Ridge, LASSO, and Elastic Net regressions provide relatively better out-of-sample predictions. Among the benefits of shrinkage-based ML methods are their ability to resolve such issues as variable selection and overfitting.

ARTICLE HISTORY
Received 5 May 2020
Accepted 10 February 2021

KEYWORDS
Housing prices; regularisation variable; selection overfitting; predictive accuracy

1. Introduction

The use of effective valuation models and the accurate prediction of real estate prices are fundamental issues in computer-assisted mass appraisal (CAMA).[1] In the literature, researchers have developed alternative automated valuation models that can be used for the mass appraisal of various types of properties.[2] For example, previous studies have used multiple regression analysis (MRA) where the property prices and hedonic attributes of the properties are jointly included in a parametric framework (Kauko et al., 2008). Researchers have also suggested the use of non-parametric methods, along with locally weighted regressions (Bitter et al., 2007) and artificial neural networks (McCluskey et al., 2013; Worzala et al., 1995), amongst other approaches. The use of information content from a large number of structural, locational, neighbourhood, and environmental characteristics has become a standard practice for accurately predicting property values in MRA (Sirmans et al., 2005). Following the seminal contribution of Rosen (1974), existing works in hedonic housing price evaluation typically utilise a number of attributes and features,[3] and the results are fraught with various econometric issues including specification error (Peterson & Flanagan, 2009), overfitting

CONTACT Prodosh E. Simlai prodosh.simlai@und.edu

(Chernozhukov et al., 2017), and appropriate model selections (Bork & Møller, 2018). In this paper, we examine the accuracy of shrinkage-based regression methods, which are part of broader machine-learning (ML) algorithms, for the prediction of median housing prices at the aggregated level using a sample of 7904 census tracts in California. We find that ML tools are a reliable means for assessing the mass appraisal valuation accuracy of median housing prices having a large number of potential determinants, and that the results obtained with such methods are less prone to overfitting and thus provide better out-of-sample performance.

The convention followed by existing influential empirical analyses of hedonic housing price functions is to provide alternative econometric models and a formal procedure for evaluating them, and ultimately to recommend a particular model (see, e.g., reviews by Vanderford et al., 2005; Xiao, 2017). In the spirit of Friedman (1953, p. 4), one can argue that in the absence of a consensus, it is not clear whether the suggested final model represents 'what is' in the data and not 'what ought to be.' Even in the presence of rigorous evaluation procedures, the suggested particular model or specific method tends to be narrow and can go untested. A public policymaker may want to know whether property values would change if an environmental policy is altered and by how much. A potential real estate investor who is interested in effectively managing her mortgage portfolio, or a family in the process of deciding where to live, may want to identify census tracts where housing units are undervalued. In answering these questions, it is imperative that we accurately predict median housing values in the presence of a set of relevant explanatory variables. The added precision resulting from the combination of hedonic regression and shrinkage-based ML tools can be useful for addressing the two aforementioned questions regarding property valuation and mass appraisal. Our approach thus provides an additional opportunity to evaluate automated valuation models used in CAMA and accurately predict real estate prices from an industry perspective that addresses the needs of a diverse set of actors such as buyers, sellers, banks, real estate professionals, mortgage portfolio managers, policymakers, etc. Single-property appraisals and partial-interest appraisals are outside the scope of this paper.

Our main goal is to investigate the determinants of owner-occupied housing values in census tracts of California by using hedonic regression-based parametric learning methods. Apart from this methodological innovation, we aim to improve the predictive accuracy of hedonic housing price functions that include several socio-economic and locational attributes in conjunction with key pollution exposure indicators. In recent years, researchers have begun to recognise the potential of ML techniques, which are quickly emerging as an important addition to the econometric toolbox (see Athey, 2018; Varian, 2014 for a review). The attractiveness of ML methods arises from their unique role in uncovering new patterns in data and their ability to improve out-of-sample prediction (Mullainathan & Spiess, 2017; Stock & Watson, 2019; Valier, 2020). As computing power has become cheap, the ML tools have become increasingly useful in a wide range of applications (Bajari et al., 2015) including real estate analysis (Borde et al., 2017; Perez-Rave et al., 2019).[4]

Unlike traditional econometric modelling, which specialises in processing data, ML methods focus on developing tools or algorithms that can actually learn from the data. These tools for learning from data can be classified as supervised, unsupervised and semi-supervised (see James et al., 2017 for a detailed discussion). Under supervised ML,

statistical models are used for estimating or predicting an output variable (also known as a response or dependent variable) based on one or more input variables (also known as predictor or independent variables). Familiar statistical techniques such as regression and classification are examples of supervised ML tools. In contrast, in unsupervised ML, the relationships are learned from data consisting of input variables without any supervising output variables. Examples of unsupervised learning include clustering, mixture models, and graphical models, etc. Finally, under semi-supervised learning, the distribution of input variables is learned using a large set of unlabelled data and the resulting information is used to analyse a small set of labelled data. In this paper, we focus on supervised ML and utilise model-based parametric learning methods that are most natural for real estate applications as they allow for flexible relationships and complex functional forms.

We use three supervised ML tools and discuss their relative performance. These tools, which are based on penalised regressions and examples of the shrinkage method, include Ridge regression (Hoerl & Kennard, 1970), LASSO (Least absolute shrinkage and selection operator) regression (Tibshirani, 1996), and Elastic Net regression (Zou & Hastie, 2005). Our empirical results, using the census tracts of California, show that the difference is tangible. These shrinkage-based ML methods implement a comprehensive description of owner-occupied housing prices in the presence of a large number of covariates. Compared to conventional econometric approaches such as OLS (ordinary least squares) and WLS (weighted least squares), the regularisation-based ML approaches provide relatively better out-of-sample predictions and prevent the regression model from overfitting. The caveat is that regularisation method-based estimators, such as the Ridge regression, have smaller variances, but unlike their unbiased OLS or WLS counterparts, they have some small biases. This bias–variance trade-off is instrumental in the better predictive performance of the underlying model specifications (for a textbook-level reference see Stock & Watson, 2019).

Following James et al. (2017), we utilise the convention of a standard ML exercise and split the available data into a training set (or estimation sample) and a testing set (or prediction sample). Consequently, in our supervised learning methods, in order to learn from the data, we start with a set of observations to train (or teach) the ML method with the goal of accurately predicting the output for new data that is yet to be seen. Once we learn from the training set and the model is trained, we evaluate how well the method works for the new data or testing set. The success of ML tools depends on the ability of the model to make accurate predictions based on the testing set. Our framework is flexible enough to accommodate new patterns in variables not previously utilised in the real estate literature. Among the benefits of shrinkage-based ML methods are their ability to resolve such common issues as appropriate variable selection and overfitting in a data-rich environment.

The rest of the paper is organised as follows: section two reviews the prior literature and explains what motivates our idea; section three briefly describes the data and methodology used throughout the paper; section four presents our main empirical analysis; and, finally, section five concludes the paper.

2. Motivation and prior literature

There is a strong theoretical foundation in the treatment of housing as a differentiated commodity and the evaluation of housing prices using relevant attributes and characteristics that are heterogeneous. In one of the earliest theoretical works on consumer behaviour, Lancaster (1966) has argued that consumers' utility is a function of product characteristics. Rosen (1974) subsequently reconsidered the relationship between observed prices and observed characteristics in the context of differentiated products and created the framework of hedonic analysis of product differentiation. Several other works such as Epple (1987) and Eckland et al. (2004) have built on Rosen's (1974) pioneering work and laid the empirical foundation for identification and estimation of hedonic regression models that are suitable for the economic analysis of real estate markets.

Building on this theoretical foundation, the existing empirical work in the real estate literature has employed either a linear or a nonlinear framework, and utilised a number of different property attributes and locational features. There are a handful of works that review the state of the art as portrayed in various hedonic housing studies. For example, Malpezzi (2002) provides a general review of the structural, locational and neighbourhood attributes, contract conditions, and time-specific attributes used in the existing literature. Vanderford et al. (2005) outline some structural, neighbourhood and geographical characteristics commonly used in various studies. In another review, Sirmans et al. (2005) investigate over 125 empirical works and discuss differences in the economic impact of certain characteristics. Despite the simplicity and intuitive appeal of hedonic regression modelling, it has been criticised on many grounds (Bitter et al., 2007; Lin & Mohan, 2011; McMillen, 2010). Since the theoretical framework of Rosen (1974) does not provide any functional form for the hedonic pricing method, such issues as misspecification and omitted variables have been highlighted as problematic (Can & Megbolugbe, 1997; Peterson & Flanagan, 2009).[5] Several works have pointed out the instability of the coefficients of hedonic estimation over time (Case et al., 2006). In an influential work, Dubin (1998) points out the lack of recognition of spatial effects in hedonic models.

An alternative that is now emerging in the literature – one that is well-suited for CAMA – is the utilisation of ML techniques that can learn from the valuation process and improve predictive accuracy (Borde et al., 2017; Mullainathan & Spiess, 2017; Perez-Rave et al., 2019). For example, Hausler et al. (2018) have applied an ML approach to capture textual sentiment relevant to US real estate markets and found evidence of a significant relationship between sentiment indicators and real estate market movements.; McCluskey et al. (2013) have analysed a number of geostatistical approaches relative to an artificial neural network model and found that in terms of methodology, the geographically weighted regression approach provides the best balance of performance and transparency. Perez-Rave et al. (2019) have employed an ML methodology for selecting variables, called 'incremental sample and resampling' (MINREM), for both inferential and predictive purposes. Overall, the existing applications of ML techniques fall under the category of automated valuation models to estimate market values for individual properties. The predictive accuracy of shrinkage-based ML methods for aggregated data from housing submarkets is not clearly known. Irrespective of the types of arguments identified in the literature, one of the challenging issues for any methodology is variable

selection and the reliability of a particular model or method. In other words, what guarantees that a given model will accurately predict housing prices for previously unseen data? The ML tools utilised in this paper address this question and give us a foundation from which we may, at a minimum, complement the existing evaluation procedures.

It is important to note that, although the focus on supervised ML in real estate analysis is fairly new, some of the tools and techniques of learning from cross-sectional data have been used in various existing studies.[6] For example, Dubin (1998) has analysed multiple listings (MLS) data for Baltimore, Maryland and divided 1,493 valid observations randomly into groups of estimation and prediction samples. In order to check out-of-sample predictions, Dubin (1998) randomly assigns 67% of the observations to the estimation sample and the remaining 33% of the observations to the prediction sample. In a series of important Goodman and Thibodeau (2003, 2007) randomly assign 90% of all transactions to the estimation sample and the remaining 10% to the prediction sample in order to evaluate the accuracy of hedonic price predictions.[7] Splitting the sample for predictive analysis is a strategy also used in Perez-Rave et al. (2019), who analyse existing home sale data in Columbia and employ their MINREM strategy using 70% of the total observations for the training sample and the remaining 30% for the validation sample. Perhaps the closest work to ours is by Bourassa et al. (2007), who examine 4,880 residential sales from Auckland, New Zealand, and compare the performance of simple OLS models with geostatistical and lattice models.[8] In order to compare the out-of-sample predictive ability of alternative methods, Bourassa et al. (2007) use 100 random samples consisting of 80% of all transactions, and for each 100 splits, generate out-of-sample predictions for the remaining 20% of data.

In our work, we form bases for comparison that correspond to those of the above-mentioned works. To the best of our knowledge, none of the existing papers has used regularisation methods in conjunction with learning techniques that can improve cross-sectional hedonic estimates of property values. We hope that this methodological innovation will complement some of the existing approaches and be further utilised in various models of performance measurement.

3. Data and methodology

In this section, we start with a brief description of our data and research design. We then provide a brief overview of our supervised ML tools.

3.1 Data and research design

This study uses data from the census tracts of California that captures the idea of neighbourhood. Three types of data are collected. The first set is based on the population and housing counts for all the census tracts from the 2010 census of population and housing summary file 1 (SF1). The second set utilises the population, income, and housing estimates for all the census tracts from the American Community Survey and 2010 census of population and housing summary file 3 (SF3). The third set includes several key pollution exposure indicators obtained from the Office of Environmental

Health Hazard Assessment of the California Environmental Protection Agency (CalEnviroScreen 2.0).

We obtain the following variables from the SF1: geographic codes, area, race/ethnicity, household type, housing tenure and average household size, and the following from the SF3: geographic codes, area, school enrolment, educational attainment, labour force, occupation, income, occupied housing units and housing unit value. We also utilise the following environmental variables from CalEnviroScreen 2.0: geographic codes, area, ozone concentration, diesel particulate matter (PM) emissions, drinking water contaminant index, and pesticide used in production-agriculture per square mile of the census tract. We acknowledge that we do not have some of the key information such as average housing size used for property appraisals. We assume that the information on the land area and local population proxies for average housing size. All files are combined using the internal points of each census tract and checked for consistency using county code, census tract code, and land area (in square miles) information. Altogether, the total number of reported census tracts in 2010 is 8,057. Of these, 153 census tracts contain missing values for the dependent variable and some predictors. After excluding all the tracts reporting missing values, our final data set consists of 7,904 census tracts.

A brief description of all the variables is given in Table 1. The list also includes various transformed variables covering a wide range of population, housing, socio-economic and environmental indicators. For reference, we also look at the mean and median value of owner-occupied housing units from 2000. It is important to note that any simple comparison of census tracts from different census years requires special care. Although census tract boundaries are delineated, due to population changes and physical changes in street patterns, etc., the number of census tracts covering the entire state is not the same in every census.[9] To make comparisons, information from standardised tracts using the Census Bureau's tract-to-tract relationship file can be utilised.[10]

Following the convention of a standard ML exercise, we start by randomly selecting a set of observations to create a training set (or estimation sample) and assign the remaining set of the available data into a testing set (or prediction sample). We construct nine such splits by randomly assigning 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of all the observations, respectively, into separate training sets. Consequently, for each of these splits, we generate out-of-sample predictions for the corresponding nine testing sets using a different set of observations. We use supervised learning methods to learn from the data in each of the training set and evaluate how well the ML method works for each corresponding testing set. Consequently, we calculate the sample predictions and summary statistics of prediction errors, and utilise those to evaluate predictive accuracy using the testing set. Since our main goal is to accurately predict housing prices for a new data set, the success of ML tools depends on the ability of each model to make accurate predictions based on various testing sets.

3.2 Hedonic regression analysis and parametric learning methods

Let us suppose that we have a collection of training data $D_{Training} = (y, X)$, where $y = (y_1, \ldots\ldots, y_n)'$ is an $(n \times 1)$ vector of observations on the log of median housing prices, X is an $(n \times p)$ matrix of covariates (or prediction or explanatory variables), and X has

Table 1. Definition of variables.

Variable Name	Definition
lattitue	Lattitude of the centroid of the census tract
longitude	Longitude of the centroid of the census tract
landarea	Total land area (in square miles)
total.pop	Total population
total.pop.sm	Total population per square mile
asian.pop	Asian population
black.pop	Black population
hispanic.pop	Hispanic population
white.pop	White population
pct.asian	Percentage of asian population
pct.black	Percentage of black population
pct.hispanic	Percentage of hispanic population
pct.white	Percentage of white population
pop.in.hh	Population household
bachelor.deg	Population with a Bachelor degree
grad.deg	Population with a Graduate or Professional degree
pct.bach.deg	Percentage of Bachelor degree
pct.grad.deg	Percentage of Graduate or Professional degree
pct.unemp	Unemployed as percent of civilian labour force
empl.civilian	Civilian 16 years and over in labour force
white.collar	Total number of white collar occupations
blue.collar	Total number of blue collar occupations
total.hhs	Total number of households
med.hh.inc	Median household income in past 12 months (2010 inflation-adjusted dollars)
total.hous.un	Total housing units
occ.hous.un	Total occupied housing units
own.occ.hous	Total owner occupied housing units
pct.own.occ	Percentage of owner occupied housing units
ozone	Amount of daily maximum 8 hour Ozone concentration over state standard
diesel	Diesel PM (Particulate Matter) emissions from on-road and non-road sources
water	Drinking water contaminant index for selected contaminants
pesticide	Total pounds of selected active pesticide ingredients used in production-agriculture per square mile in the census tract
med.value.2010	Median value of all owner-occupied housing units in 2010 (Dollars)
mean.value.2010	Mean value of all owner-occupied housing units in 2010 (Dollars)
med.value.2000	Median value of all owner-occupied housing units in 2000 (Dollars)
mean.value.2000	Mean value of all owner-occupied housing units in 2000 (Dollars)

Source: American Community Survey, Decennial Census of Population and Housing, the Office of Environmental Health Hazard Assessment of California Environmental Protection Agency, U.S. Census Bureau and authors' calculation.

full rank. Under a hedonic framework, X includes the set of relevant structural, neighbourhood, and environmental characteristics. Our objective is to build a model $f(X,\beta)$ using the training data to describe the nonlinear relationship between y and X with β as a $(p \times 1)$ vector of unknown parameters. Also, suppose that we want to verify the predictive accuracy of our model specification $f(X,\beta)$ using a testing dataset $D_{Testing} = (y^*, X^*)$, where y^* is an $(n \times 1)$ vector of observations on the dependent variable and X^* is an $(n \times p)$ matrix of covariates with full rank.

Given the training data $D_{Training}$, we can express the mean specification of a standard multiple regression model (MRM) by $y = X\beta + \varepsilon$, where it is assumed[11] that $E(\varepsilon) = 0$, and $E(\varepsilon\varepsilon') = \sigma^2 I_n$. Under the assumption of homoskedastic errors, the OLS estimate of the regression coefficient β is obtained by minimising the residual sum of squares (RSS). When the assumption of constant variance in the errors is violated, the OLS estimates continue to be unbiased and consistent, but an alternative estimate based on WLS or some variant of the generalised least squares (GLS) produces smaller variance and thus becomes more efficient. As with almost all hedonic regression models, the independence of the residuals cannot be assumed, as housing price regression residuals exhibit spatial dependence (Dubin, 1998). One common remedy is to capture absolute location and incorporate geographical coordinates or other spatial indicators as regressors (Bourassa et al., 2007; Fik et al., 2003; Xu, 2008).[12]

Since a comprehensive description of owner-occupied housing prices requires the presence of a large number of covariates, the selection of appropriate predictors can also be an important issue in estimating the MRM. For the training set, the presence of a large number of covariates may have a positive effect on the predictive accuracy of $\hat{y}$. For example, one can show that the expected value of the $TrainingError = (y - \hat{y}_{OLS})'(y - \hat{y}_{OLS})$ becomes a decreasing function of the number of covariates (see Stock & Watson, 2019, appendix 19.7 for details). The number of covariates, however, has a different effect on the predictive performance of the MRM using a testing set.[13] Given the testing data $D_{Testing}$, if we use the OLS predictions from MRM to verify the fit from $y^* = X\beta + \varepsilon^*, \varepsilon^* \sim N(0, \sigma^2 I)$, the expected value of the $TestingError = (y^* - \hat{y}_{OLS})'(y^* - \hat{y}_{OLS})$ increases as the number of covariates increases (Hastie et al., 2009; James et al., 2017).

As an alternative to the familiar least squares-based estimates, we consider three core learning methods that are based on penalised regressions: Ridge, LASSO, and Elastic Net. The key idea behind the penalised regressions, which are types of regularisation (or shrinking) methods, is to improve predictive accuracy by using estimators with smaller variance, which are not necessarily unbiased. Under the regularisation techniques, we minimise the RSS, as we do with OLS, but subject to a bound of the slope coefficients. For example, in the Ridge regression, we minimise the RSS subject to a bound on the L_2-norm (Euclidean distance) of the slope coefficients. In the LASSO regression, we minimise the RSS subject to an L_1-penalty (Manhattan distance) on the regression slope coefficients.[14] In contrast, in the Elastic Net regression, we minimise the RSS subject to a combination of the penalisations through the L_1- and L_2-norms. One of the features of a hybrid approach such as the Elastic Net is that, even when we have a group of correlated covariates, we use the information from the entire group in the model-building. As

continuous shrinkage methods, both LASSO and Elastic Net perform automatic variable selection and produce parsimonious final models.

In the Ridge regression (Hoerl & Kennard, 1988), the β coefficients are estimated by

$$\hat{\beta}_{Ridge} = \underset{\beta}{argmin}\left[(y - X\beta)'(y - X\beta) + \lambda\sum_{j=1}^{p}\beta_j^2\right] \tag{1}$$

where $\lambda \geq 0$ is a tuning or regularisation parameter (or hyper-parameter) that controls the amount of shrinkage.[15] Thus, for the training set, the prediction of y at the n observed data points under the Ridge regression is arrived at by $\hat{y}_{Ridge} = X\hat{\beta}_{Ridge} = X(X'X + \lambda I)^{-1}X'y$. Note that $\hat{\beta}_{Ridge}$ is still a linear estimator and is not invariant with respect to the scale of the predictors. Although the Ridge regression shrinks the coefficients towards zero, it will not select a subset of covariates, and thus meaningful variable selection remains an issue.

The LASSO regression (Tibshirani, 1996) is one of the alternative methods that can aid in selecting a small subset of predictors. Under LASSO, the MRM coefficients are estimated by

$$\hat{\beta}_{Lasso} = \underset{\beta}{argmin}\left[(y - X\beta)'(y - X\beta) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right] \tag{2}$$

where the tuning parameter λ is typically selected by the cross-validation methods described below. The algorithm for estimating λ is the same for (1) and (2). Finally, under the Elastic Net regression (Zou & Hastie, 2005), which combines both Ridge and LASSO, the β coefficients are estimated by

$$\hat{\beta}_{ElasticNet} = \underset{\beta}{argmin}\left[(y - X\beta)'(y - X\beta) + \lambda\left(\sum_{j=1}^{p}\left[(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right]\right)\right] \tag{3}$$

where the hyper-parameter α controls how much L_1- and L_2-norm are used. If $\lambda = 0$, there is no penalty term and $\hat{\beta}_{ElasticNet} = \hat{\beta}_{OLS}$. If $\alpha = 0$, $\hat{\beta}_{ElasticNet} = \hat{\beta}_{Ridge}$, and if $\alpha = 1$, $\hat{\beta}_{ElasticNet} = \hat{\beta}_{LASSO}$. Unlike the OLS-based estimation, all three ML techniques are continuous regularisation methods that shrink slope coefficients. However, the Ridge regression retains all the covariates, and both LASSO and Elastic Net perform automatic variable selection and produce parsimonious final models.

In addition to the above three ML techniques, the statistical learning literature has proposed regularised methods for enhancing predictive accuracy. These methods are based on variations of the penalised framework. One such well-known variation is Adaptive LASSO (ALASSO), suggested by Zou (2006), who argues that LASSO does not have the so-called oracle properties such as the identification of the right subset of variables and the optimal estimation rate. Under ALASSO, the MRM coefficients are estimated by

$$\hat{\beta}_{ALASSO} = \underset{\beta}{argmin}\left[(y - X\beta)'(y - X\beta) + \lambda\sum_{j=1}^{p}\hat{\omega}_j\left|\beta_j\right|\right] \tag{4}$$

where $\hat{\omega}_j = \frac{1}{(|\hat{\beta}_{Initial}|)^{\gamma}}$, $\hat{\beta}_{Initial} = \hat{\beta}_{OLS}$ and $\gamma > 0$. Since ALASSO uses the adaptive weights vector $\hat{\omega}_j$ and penalises each coefficient differently, it avoids penalising large coefficients.

In practice, we use two cross-validation (CV) methods to choose hyper-parameter values for the penalised regressions. In statistical learning, CV is a useful technique for assessing the expected test error and for choosing an appropriate regularisation (or tuning) parameter for penalised regressions. In essence, the CV enables us to test several combinations of λ and α to identify their optimal values, and the choice of λ and α depend on where the CV error (CVE) is minimised. There are various specialised algorithms available in software packages for estimating hyper-parameters. Three well-known methods are: maximum regularisation using the prediction of the λ that minimises the error of CV (λ-Min), maximum regularisation whose error of CV is only one standard error away from the minimum (λ-1SE), and a cross-model selection and averaging.

For the sake of clarity, here we briefly describe the K-fold CV, which is a simple and popular technique for choosing hyper-parameter values. First, we divide the observed training data $D_{Training} = (y, X)$ into K folds and compute the following CVE over a grid of λ values

$$CVE\left(\hat{\beta}_{\lambda}\right) = \frac{1}{K}\sum_{k=1}^{K} CV_k\left(\hat{\beta}_{\lambda}^{(-k)}\right)$$

where each $\hat{\beta}_{\lambda}^{(-k)}$ denotes the penalised regression coefficient (such as the Ridge, LASSO, etc.) for a tuning parameter value λ and is based on all samples except those in the k-th fold. The expression $CV_k\left(\hat{\beta}_{\lambda}^{(-k)}\right)$ indicates the test error for using $\hat{\beta}_{\lambda}^{(-k)}$ in the k-th fold. Under the usual error rule λ-Min, for each regularised method we choose the tuning parameter λ to minimise the cross-validation error

$$\hat{\lambda} = \underset{\lambda}{argmin}\, CVE\left(\hat{\beta}_{\lambda}\right).$$

Under the one-standard-error rule λ-1SE, we choose the tuning parameter λ for a model whose cross-validation error is within one standard error of the minimum of $CVE\left(\hat{\beta}_{\lambda}\right)$. In the empirical illustration of this paper, we implement both λ-Min and λ-1SE.

Finally, in order to compare the out-of-sample performance of predictions corresponding to each method, we utilise six alternative measures for summarising forecasting accuracy. We start with a collection of training $D_{Training} = (y, X)$ and testing $D_{Testing} = (y^*, X^*)$ datasets. Once we estimate the regression model using $D_{Training}$, we generate a set of predictions using X^* from the testing set for each census tract *i*. Since we are interested in out-of-sample accuracy, we look at the actual median housing value y_i^* from $D_{Testing}$ and compare it with the corresponding predicted value $\hat{y}_i^*$. For each methodology, we generate a separate set of predictions and utilise the following six measures for summarising forecasting accuracy:

$$Mean\ Squared\ Error = Average\left(\left(y_i^* - \hat{y}_i^*\right)^2\right)$$

,

$$Mean\ Absolute\ Error = Average\left(\left|y_i^* - \hat{y}_i^*\right|\right),$$

$$Median\ Abolute\ Error = Median\ \left(\left|y_i^* - \hat{y}_i^*\right|\right),$$

$$Median\ Abolute\ Percentage\ Error = \ Average\left(\frac{\left|y_i^* - \hat{y}_i^*\right|}{y_i^*}\right),$$

$$MinMax\ Accuracy = Average\left(\frac{Min\left(y_i^*, \hat{y}_i^*\right)}{Max\left(y_i^*, \hat{y}_i^*\right)}\right), \text{ and}$$

$$Theil's\ U\ statistic = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - \hat{y}_i^*\right)^2}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^{*2}} + \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i^{*2}}}\right)$$

The first four are common statistical loss functions commonly used in the literature for comparing predictions. The mean squared error (MSE) – the most widely known measure of accuracy – weighs larger errors more heavily than smaller errors and is sensitive to outliers. While the mean absolute error (MAE) can be somewhat sensitive to outliers, the median absolute error (MDAE) treats errors evenly according to their absolute value and is insensitive to outliers. The median absolute percentage error (MAPE) measures the size of the prediction error in percentage terms. The MinMax accuracy (MMA) measure provides an alternative indicator of prediction accuracy and suggests the degree to which the model's prediction is off. The lower the MMA measure, the worse the out-of-sample prediction. For a nearly perfect prediction, the measure will be approximately one. Theil's U-statistic is a relative accuracy measure. If the forecast is close to perfect, Theil's U-statistic should be close to zero, and if the model fails to predict or predicts no change, the value of Theil's U-statistic will be close to one.

4. Empirical analysis

In this section, we start with a brief description of our data and then provide an in-depth analysis of our findings.

4.1 Descriptive statistics and correlation

We start by analysing the sample pair that is split 50/50 between training and testing samples, each of which contains 3,952 census tracts. Table 2 presents the basic summary statistics such as the mean, the standard error of the mean, and the standard deviation of some of the key variables. The average median value of owner-occupied housing units is 478,650 USD for the training sample and 479,019 USD for the testing sample. Both average and median values of owner-occupied housing units in the 2010 census have doubled since the 2000 census. The average census tract in the training sample has an area of 20 square miles and a population of 4,634. The average census tract in the testing sample has an area of 19 square miles and a population of 4,708. The total population per square mile in the training sample is 8,218, which is comparable to that of the testing set.

Table 2. Summary Statistics for training and testing samples.

	Training Sample			Testing Sample		
Variable name	Mean	S.E(Mean)	Std. Dev	Mean	S.E(Mean)	Std. Dev
lattitue	35.48	0.033	2.083	35.53	0.033	2.09
longitude	−119.39	0.031	1.977	−119.43	0.032	1.98
landarea	20.09	2.57	161.97	18.90	2.58	162.71
total.pop	4634.28	30.42	1912.48	4708.43	31.35	1971.28
total.pop.sm	8218.6	147.22	9254.97	8285.41	141.07	8868.22
asian.pop	581.83	11.88	746.86	599.14	13.09	823.22
black.pop	261.96	6.690	420.55	271.29	7.192	452.15
hispanic.pop	1725.46	23.54	1480.09	1790.44	25.44	1599.79
white.pop	1889.97	23.49	1476.92	1868.29	22.28	1401.09
pct.asian	12.26	0.22	13.99	12.35	0.23	14.58
pct.black	5.82	0.15	9.18	5.80	0.15	9.40
pct.hispanic	36.44	0.41	25.93	36.63	0.42	26.61
pct.white	41.81	0.42	26.88	41.53	0.43	27.04
pop.in.hh	4553.25	29.49	1854.04	4628.25	30.98	1948.11
total.hous.un	1713.75	11.47	721.41	1730.30	12.10	761.18
bachelor.deg	568.68	6.971	438.21	568.58	7.152	449.59
grad.deg	323.32	5.497	345.58	318.32	5.444	342.25
pct.bach.deg	18.78	0.201	11.29	18.66	0.180	11.37
pct.grad.deg	10.76	0.179	10.01	10.52	0.157	9.91
pct.unemp	9.02	0.159	4.627	9.287	0.078	4.87
empl.civilian	2081.26	13.75	864.82	2107.71	14.48	910.83
white.collar	1288.25	11.38	715.77	1292.69	11.76	739.64
blue.collar	793.01	7.152	449.59	815.02	7.642	480.38
total.hhs	1576.46	10.34	650.39	1590.46	10.72	674.33
med.hh.inc	65,371.10	470.92	29,604.53	66,086.17	494.10	31,061.5
owner.occ.ho	897.57	8.656	544.15	902.03	8.549	537.44
pct.owner.occ	52.84	0.371	23.35	53.05	0.365	22.97
ozone	0.105	0.003	0.183	0.102	0.003	0.176
diesel	18.24	0.268	16.87	17.90	0.251	15.80
water	351.91	2.977	187.18	351.45	3.041	191.19
pesticide	312.84	41.70	262.16	265.24	41.47	260.72
med.value.2010	478,650	3565.33	224,134	479,019	3591.29	225,766
mean.value.2010 501,569	501569	3592.18	225822	501282	3607.80	226804
med.value.2000	231250	2479.63	155881	233327	2590.44	162847
mean.value.2000	248871	2632.19	165472	251143	2749.47	172845

The distributions of asian, black, hispanic, and white populations for the training and testing samples are also very similar. On average, the census tracts in the training set have a slightly higher percentage of individuals holding bachelor's degrees and graduate or professional degrees, although the differences are very small. The average unemployment rates as well as the distribution of blue- and white-collar workers in the training set are on par with the testing set. Compared to the testing set, the median household income in the training sample has a slightly lower mean and standard deviation. Although the average pesticide use in the training set is slightly higher than the testing set, the distribution of ozone concentration, diesel PM emission, and drinking water contamination index in the training set is comparable to that of the testing set.

Next, we summarise the cross-correlation structure of the set of potential predictors in Table 3. Consistent with our expectation, census tracts with higher median household income (med.hh.inc) tend to have a higher percentage of owner-occupied housing units (own.occ.hous), a greater share of population with college degrees (pct.bachelor.deg and pct.grad.deg), and a lower unemployment rate (pct.unemp). The median household

Table 3. Correlations between predictors from the training set.

	med.hh	tot.pop	own.oc	pct.own	pct.wh	pct.bl	pct.as	pct.his	pct.ba	pct.gr	pct.un	wh.coll	bl.coll	ozone	diesel	water	pestic
med.hh	1.0000																
tot.pop	−0.1508	1.0000															
own.oc	0.5431	−0.2415	1.0000														
pct.own	0.6658	−0.2813	0.7211	1.0000													
pct.wh	0.4958	−0.4329	0.3786	0.3032	1.0000												
pct.bl	−0.2490	0.1848	−0.1551	−0.1686	−0.3685	1.0000											
pct.as	0.2282	0.2370	−0.0064	−0.0524	−0.2357	−0.0492	1.0000										
pct.his	−0.5532	0.2597	−0.3379	−0.2829	−0.7873	0.0220	−0.2917	1.0000									
pct.ba	0.6832	0.0063	0.2544	0.2185	0.5809	−0.1808	0.3463	−0.7391	1.000 0								
pct.gr	0.6369	−0.0395	0.2186	0.2088	0.5482	−0.1539	0.2382	−0.6509	0.7646	1.0000							
pct.un	−0.5416	0.0240	−0.2111	−0.2616	−0.3434	0.2156	−0.1714	0.3698	−0.4975	−0.4384	1.0000						
wh.coll	0.5439	0.1579	0.6180	0.2715	0.3219	−0.1254	0.2457	−0.4326	0.5629	0.4454	−0.3805	1.0000					
bl.coll	−0.3930	0.2097	0.1303	−0.1627	−0.4668	0.0600	−0.1045	0.5259	−0.5179	−0.5917	0.2489	0.1794	1.0000				
ozone	−0.1226	−0.1084	0.0765	0.0794	−0.0686	−0.0035	−0.1917	0.1886	−0.2492	−0.1953	0.1793	−0.1152	0.1103	1.0000			
diesel	−0.2786	0.4641	−0.4058	−0.3840	−0.3548	0.1798	0.1535	0.2248	−0.0566	−0.0594	0.0724	−0.1163	0.0243	−0.0827	1.0000		
water	−0.1296	−0.2412	0.0039	0.0331	−0.1045	−0.1115	−0.1839	0.2737	−0.2695	−0.2343	0.1316	−0.1382	0.1445	0.4369	−0.0775	1.0000	
pestic	−0.0197	−0.0867	0.0153	0.0287	−0.0299	−0.0517	−0.0578	0.0865	−0.0690	−0.0528	0.0219	−0.0449	0.0361	−0.0445	−0.0627	0.0757	1.0000

Note: Here, med.hh = ln(med.hh.inc), tot.pop = ln(total.pop), own.oc = ln(own.occ.hous), pct.own = pct.owner.occ, pct.wh = pct.white, pct.bl = pct.black, pct.as = pct.asian, pct.his = pct.hisp, pct.ba = pct.bachelor.deg, pct.gr = pct.grad.deg, pct.un = pct.unemp, wh.coll = ln(white.collar), bl.coll = ln(blue.collar), pestic = pesticide as reported in previous tables.

income is positively correlated with white and asian population percentages (given by pct.white and pct.asian, respectively) as well as total number of white-collar occupations (white.collar). This evidence is supported by the fact that a large percentage of white and asian populations are college educated (with an average correlation of 0.56 and 0.29, respectively) and employed in white-collar occupations (with an estimated correlation of 0.32 and 0.25, respectively) that tend to have higher income potential. The unemployment rate (pct.unemp), which tends to be one of the key players on the demand side, has very high correlation with some of the attributes that influence median housing values. Higher percentages of the asian and black populations are located in census tracts characterised by higher diesel PM emission, but lower concentrations of ozone, water contamination and pesticide use. More polluted census tracts are characterised by a high percentage of hispanics and a large number of blue-collar occupations, as well as a low percentage of college graduates and small number of white-collar occupations.

4.2 Preliminary regressions

We now turn to our primary task, investigating how well the aforementioned set of predictors explains average housing prices in the census tracts of California. Thus, for the rest of this paper, we report the regression results where the dependent variable is the log of the median value of owner-occupied housing units (median housing value hereafter) in 2010. As is common in the literature, the theoretical framework of Rosen (1974) provides few restrictions on the form of the housing price function. To ensure the reliability of a particular model or method, in addition to typical ad hoc and goodness-of-fit criteria, we count on the out-of-sample predictive accuracy.

Before evaluating the impact of our set of predictors, in Figures 1a and 1b we get a first-hand glimpse of the distribution of median housing prices corresponding to the training sample observations. In Figure 1a, the darker shaded points reflect census tracts with lower median housing values, and the lighter points reflect higher-valued census tracts. The overall geographical distribution suggests that the higher-valued census tracts are located close to the big metropolitan areas and coastal areas, especially around San Francisco, Los Angeles and San Diego. The area between −121 and −120 longitude, which covers Napa and Sacramento, has higher housing values than the rest of central California. Almost all the lower-valued census tracts are located inland or away from densely populated areas.

In Figure 1b, we use a regression tree to illustrate the overall geographical distribution of overvalued and undervalued census tracts. The tree approach, which falls under the umbrella category of classification and regression trees (CART), is well known for modelling complicated relationships and is based on random samples of the data. We observe that the regression tree in Figure 1b recursively partitions the entire sample into smaller and smaller rectangles or sub-regions using the latitude and longitude of each census tract as two predictors. The figure shows that the partitioning correctly represents the overall geographical distribution of higher- and lower-valued census tracts in our data. It is evident that when both latitude and longitude matter, the tree captures the subtle variations in housing values in the relevant rectangles. Even though our results related to predictive accuracy are not directly related to the CART approach, one can use tools such as a random forest by averaging lots of trees to classify the data. Here we

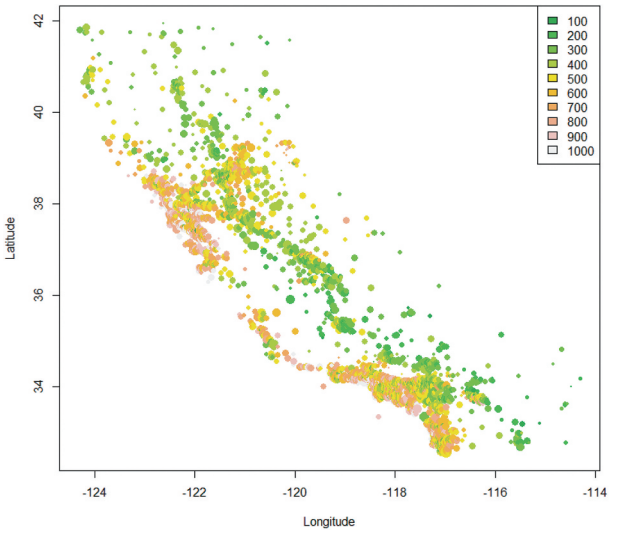

Figure 1a: Deciles of median house prices (in '000 of dollars) of all census tracts

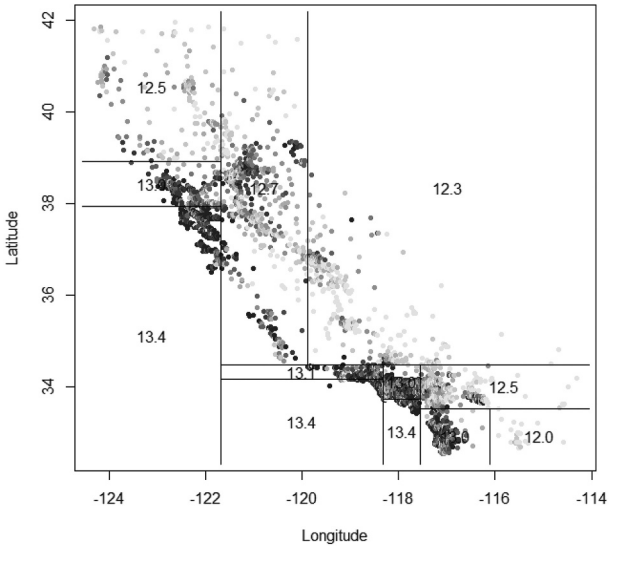

Figure 1b: Decomposition of log of median house prices of all census tracts using regression tree.

simply present the tree for the reader's reference and highlight its usefulness for partitioning using absolute locations with spatial coordinates.

For our empirical analysis of the California census tract data, we start by asking the question: what guarantees that the presumed MRM specification captures the true relationship among variables beyond the training data. For example, in a housing price equation with a log of median values of owner-occupied housing units as a dependent variable, in addition to quadratic and cubic polynomials of all predictors, one can incorporate higher degree polynomials in a model that closely fits the training set. If it

turns out that a specific model fits the training set reasonably well but performs poorly on the previously unseen test set, we have what is commonly known as overfitting.

In Table 4, we report the results of housing price regressions that include the impact of several socio-economic, housing and environmental variables for a training sample consisting of 50% of all census tracts. Here, we present four sets of estimated regression coefficients using both OLS and WLS. In column (1) and (5), we present the basic specification using socio-economic and housing variables, and in column (2) and (6), we add four environmental variables. In column (3) and (7), following the literature (Case et al., 2004; Dubin, 1998; Xu, 2008), we include spatial trend terms that are a polynomial function of the latitude and longitude of each census tract.[16] Finally, in column (4) and (8), we include all the predictors. The values of the estimated parameters with and without the presence of polynomial trends are in line with the existing empirical evidence. All the reported regressions suggest that the census tracts with higher percentages of educated population, who have bachelors or higher degrees, tend to have higher median housing prices.

In regressions where we modify the specification by including four pollution-exposure variables, $\bar{R}^2$ becomes slightly higher (e.g., $\bar{R}^2$ is 62.52% and 66.54% in (2) and (4) respectively). Ozone concentration and pesticide use both have a strong negative effect on median housing values. Diesel emissions have a positive effect in OLS but a negative effect in WLS regressions. To our surprise, water contamination has positive slope estimates in both OLS and WLS regressions – possibly suggesting that census tracts with higher median values may have already priced such pollution factors.

In columns where we modify the regression specification by including polynomial trends, all the slope estimates are similar to those without trends. One major difference is the estimated intercept – both in magnitude and in significance. For example, the estimate of the intercept is 4.73 in column (1) as opposed to 22.6 in column (3). Similar to Dubin (1998), we find that the presence of spatial trends makes the OLS intercept statistically insignificant. The OLS model without a spatial trend achieves a reasonable fit (e.g., $\bar{R}^2$ is 58.97% in (1)), but the polynomial function of the latitude and longitude of each census tract improves the overall performance (e.g., $\bar{R}^2$ is 65.86% in (3)).[17]

We now perform an exploratory analysis to understand the spatial distribution of our predicted median housing values. For this analysis, we plot the estimated residuals from our best OLS regression from Table 4 in Figure 2, along with positive-only and negative-only residuals. We postulate that a positive (negative) residual would suggest possible overvaluation (undervaluation) of the median housing value of each corresponding census tract. The first panel of Figure 3 suggests that the estimated residuals are not constant across the training set observations. The plot of the positive OLS residuals in the second panel suggests that the residuals tend to be disproportionately positive near the coasts and surrounding commuting areas, thus confirming overestimation in these areas. The plot of the negative OLS residuals does not show any clustering of higher values located either inland or along the coastal areas. While there is a high density of census tracts with positive OLS residuals around the big coastal cities and some related areas, the census tracts with negative residuals are scattered throughout the state.

Table 4. OLS and WLS regressions using training sample.

	OLS with and without trend				WLS with and without trend			
Coefficients	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	4.729***	4.214***	22.60	54.99	7.315***	4.228***	43.08***	37.34***
ln(med.hh.inc)	0.4555***	0.455***	0.3858***	0.3907***	0.3127***	0.1815***	0.3144***	0.3306***
ln(total.pop.sm)	0.0271***	0.0224***	0.0043	0.0042	0.0828***	0.0316***	0.0431***	0.0538***
ln(own.occ.ho)	−0.1577***	−0.1337***	−0.1182***	−0.1084***	−0.1716***	−0.1521***	−0.1328**	−0.1506***
pct.asian	0.0310***	0.0357***	0.0334***	0.0320***	0.0187***	0.0610***	0.0451***	0.0458***
pct.black	0.0315***	0.0365***	0.0330***	0.0317***	0.0018	0.0622***	0.0449***	0.0468***
pct.hispanic	0.0298***	0.0351***	0.0316***	0.0305***	0.0156***	0.0571***	0.0461***	0.0471***
pct.white	0.0307***	0.0360***	0.0336***	0.0325***	0.0178***	0.0608***	0.0489***	0.0495***
pct.bach.deg	0.0131***	0.0113***	0.0102***	0.0097***	0.0163***	0.0021***	0.0097***	0.0106***
pct.grad.deg	0.0105***	0.0098***	0.0063***	0.0064***	0.0107***	0.0188***	0.0005	0.0003
pct.unemp	−0.0108***	−0.0085***	−0.0062***	−0.0058***	−0.0236***	−0.0178*	−0.0030**	−0.0036**
ln(white.collar)	0.0560***	0.0619***	0.0751***	0.0701***	0.1889***	0.2343***	0.1440***	0.1468***
ln(blue.collar)	0.0692***	0.0545***	0.0258*	0.0269**	−0.04920	−0.0742***	−0.0467***	−0.0665***
ozone		−0.5779***		−0.2623***		−0.8856***		−0.0274***
diesel		0.0005*		0.0006*		−0.0007**		−0.0002**
water		0.0038		0.0089**		0.0004***		0.0086***
pecticide		−0.0002***		−0.0001***		−0.0001***		−0.0002**
lattitude			1.916*	0.2112*			5.228***	2.894**
longitude			1.032	1.612			8.843***	7.152***
latitude^2			0.0362***	0.0296***			0.0427***	0.0257***
longitude^2			0.0113*	0.0132			0.0485***	0.0366***
latt* long			0.0396***	0.0372***			0.0711***	0.0409***
N	3,952	3,952	3,952	3,952	3,952	3,952	3,952	3,952
p	12	16	17	21	12	16	17	21
R2	0.5910	0.6268	0.6601	0.6672	0.8722	0.8901	0.8954	0.9056
Adjusted R2	0.5897	0.6252	0.6586	0.6654	0.8630	0.8897	0.8947	0.9050

Note: Estimate of the parameters marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively. N is the sample size and p is the total number of regressors.

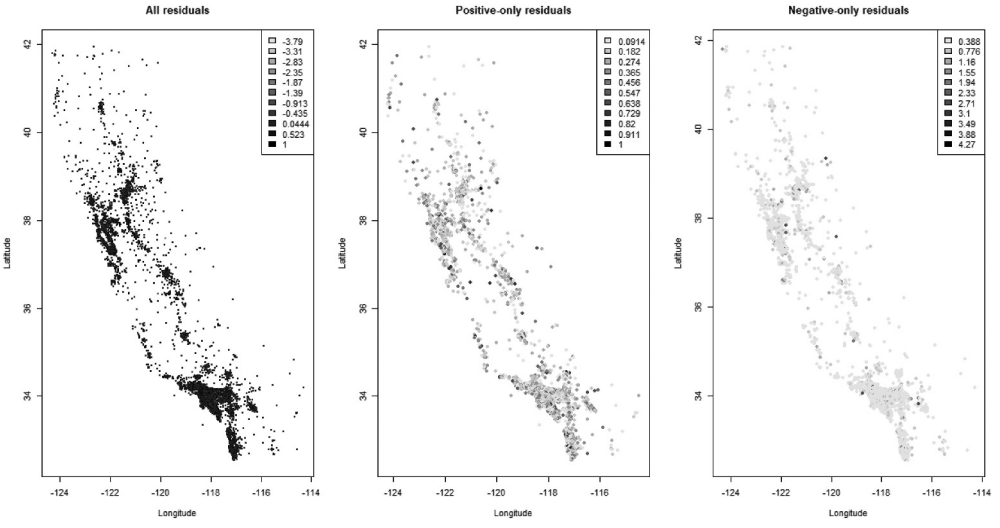

Figure 2. Distribution of the OLS residuals.

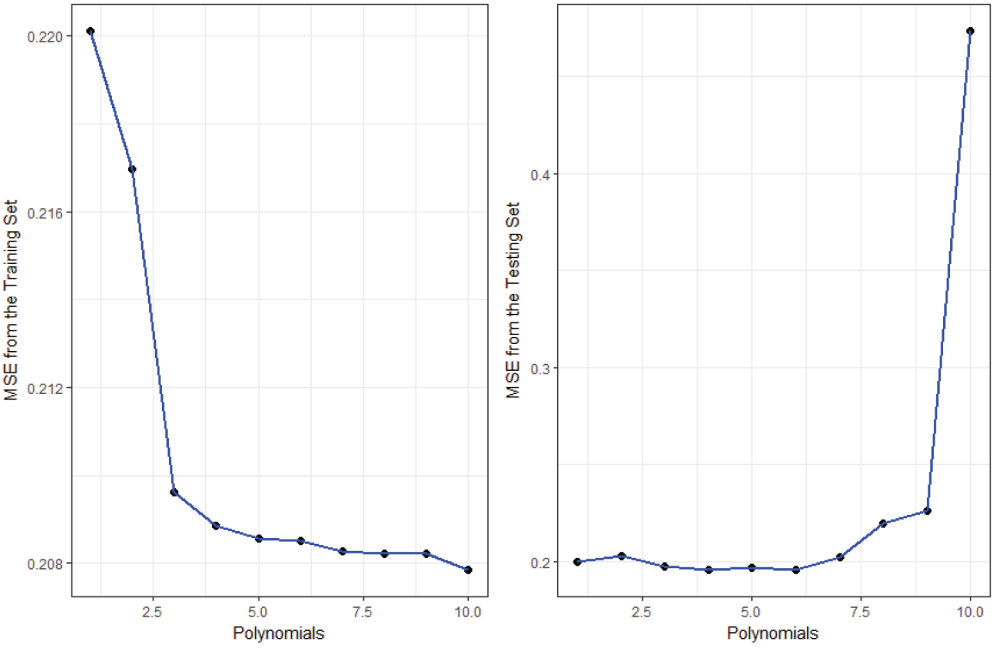

Figure 3. MSE of training set and testing set for polynomials 1 through 10.

For implementing a hedonic housing price regression, in addition to specification (4) of Table 4, one can incorporate higher degree polynomials for all the predictors. In Figure 3, we investigate whether a specific polynomial model fits the training set reasonably well, and that it also fits a previously unseen test set. The plot shows that the MSE of the training set decreases with each higher degree of polynomial, while the MSE of the

testing set grows with each higher degree of polynomial. Based on the testing set MSE alone, the specification (4) is comparable to all other polynomial regression functions.

4.3 Penalised regressions

In this subsection, we switch our focus to regressions based on penalised methods. Table 5 presents two sets of slope coefficients with each set corresponding to a different shrinkage-based regression. First, we present the Ridge regression estimates. This is followed by results from the LASSO and Elastic Net regressions. For each method, two types of estimate corresponding to minimum lambda (λ) are reported. The first is based on the usual error rule given by λ-min, and the second is based on the one-standard-error rule given by λ-1SE as described in section 3. The resulting values of λ (lambda) for each model are reported in the last row.

The first column of Table 5 shows the Ridge regression results based on λ-min, where we predict median housing values in each census tract using all the available predictors. Here, five out of sixteen socio-economic and environmental variables have negative slope coefficients. Compared to the best OLS model (i.e., column (4) in Table 4), most of the predictors have slope estimates that are lower in magnitude, and worth noting is that pct. black now has a negative slope. For the Ridge regression results based on λ-1SE, which is reported in the next column, two additional variables yield negative slope coefficients; they are pct.hispanic and ln(blue.collar). The next two sets of λ-min and λ-1SE show the LASSO and Elastic Net regressions. The LASSO regression under λ-min produces a set of slope coefficients that are similar in magnitude to the Ridge regression. The same is true for the Elastic Net regression under λ-min. Both the LASSO and Elastic Net regressions based on λ-1SE result in relatively sparse coefficients. For all three models, we allowed statistical packages to choose the best hyper-parameters. For Ridge and LASSO, the statistical package tests the range of possible values of λ and performs a cross-validation to find the optimal values for λ. For Elastic Net, on the other hand, we set up a grid of λ and α. The statistical package tests the range of possible λ and α values and performs a cross-validation to find the optimal values for λ and α jointly. We report that the optimal values of α in the Elastic Net regressions are 0.613 for λ-Min and 0.740 for λ-1SE.

Note that Table 5 does not have the usual information about statistical significance that readers typically expect to see from regression results. Performing the conventional significance tests of the penalised regression coefficients after the model selection is an issue not completely resolved in the literature. In the context of shrinkage-based regressions, if we select the most predictive variables and then perform statistical inference, we assume that these variables were selected independently of the data. Consequently, using a simple Gaussian approximation for creating confidence intervals will often fail. It has been mentioned in the literature that refitting the shrinkage-based regressions after performing model selection may lead to biased estimates that may cause a problem with the usual significance tests (Hastie et al., 2009). Tibshirani (1996) suggests a bootstrap-based procedure to estimate the coefficient's variance. Since our focus is predictive accuracy, a discussion of statistical inference using bootstrap-estimated variance is beyond the scope of this paper.

Furthermore, as is well known in the literature (see, e.g.,, Bajari et al., 2015; Mullainathan & Spiess, 2017), one has to be careful about interpreting estimated

Table 5. Penalised regression coefficients from the training set.

Variables	Ridge		LASSO		Elastic Net	
	λ-Min	λ-1SE	λ-Min	λ-1SE	λ-Min	λ-1SE
ln(med.hh.inc)	0.3334	0.2469	0.3811	0.3385	0.3759	0.3361
ln(total.pop.sm)	0.0131	0.0176	0.0022	0.0172	0.0026	0.0129
ln(own.occ.ho)	−0.0826	−0.0464	−0.1113	−0.0553	−0.1095	−0.0650
pct.asian	0.0001	0.0006	0.0304	.	0.0234	.
pct.black	−0.0005	−0.0011	0.0303	.	0.0233	−0.0004
pct.hispanic	0.0005	−0.0002	0.0291	.	0.0226	.
pct.white	0.0011	0.0005	0.0313	.	0.0246	.
pct.bach.deg	0.0094	0.0080	0.0098	0.0107	0.0098	0.0103
pct.grad.deg	0.0082	0.0079	0.0065	0.0084	0.0066	0.0083
pct.unemp	−0.0095	−0.0112	−0.0065	−0.0089	−0.0068	−0.0091
ln(white.collar)	0.0617	0.0674	0.0683	0.0273	0.0661	0.0428
ln(blue.collar)	0.0083	−0.0205	0.0305	.	0.0302	.
ozone	−0.4480	−0.4211	−0.3468	−0.4944	−0.3805	−0.4599
diesel	0.0006	0.0007	0.0005	4.706e-05	0.0004	0.0002
water	6.929e-05	−1.802e-06	0.0001	.	6.668e-05	.
pesticide	−4.888e-06	−2.181e-06	−8.314e-06	.	−8.163e-06	.
lattitude	−0.0436	−0.0234	−0.3674	.	−0.1778	−0.0338
longitude	−0.0519	−0.0235	−0.2341	−0.0294	−0.1136	−0.0310
latitude^2	−0.0006	−0.0003	0.0052	−0.0007	0.0002	−0.0005
longitude^2	0.0002	0.0001	.	.	0.0002	0.0001
latt* long	0.0002	0.0001	0.0015	.	0.0001	.
lambda	0.03798	0.1683	3.798e-05	0.0201	0.0001	0.0146

Note: For each method, two types of estimate of the parameters corresponding to minimum lambda are reported. First is based on the usual error rule and the second is based on the one standard error rule. The resulting values of lambda for each model is reported in the last row.

coefficients from the Ridge regression, because unlike those from OLS and WLS, the Ridge regression coefficients are biased. Instead, we should utilise regularisation paths, which can help to identify variables with highest predictive power for median housing values. In Figure 4, we display the regularisation paths for three methods. The Ridge regression is presented in the first panel, followed by the LASSO and Elastic Net regressions in the next two panels.

In Panel A, we plot the resulting regularisation paths for the top ten predictors from the Ridge regression. Here we allow λ to vary and investigate the resulting solution path. On the extreme right, when λ takes the smallest value, the Ridge coefficient estimate is similar to the OLS estimate. As λ increases from right to left, the coefficient for each of the predictors changes, and for large enough λ, all the coefficients shrink towards zero. Each panel also includes a 10-fold cross-validation curve illustrated by the thick dotted line, and upper and lower standard deviation curves along with λ sequences. Two vertical dotted lines indicate two λ's; the left line representing the rule that minimises the cross-validation error and the right line representing the one-standard-error rule. Altogether, the Ridge regularisation paths show that variables such as ln(med.hh.inc) and ozone are farthest away from zero for a significant portion of the plot. Thus, these two variables are the most influential in predicting median housing prices across the census tracts. Variables such as ln(white.collar) and ln(total.pop.sm), which are given by lg.whtcl and lh.ttl.p respectively in the plot, have coefficient values that are positive for the entire portion of the plot, whereas variables such as ln(own.occ.ho) and pct.unemp, which are given by lg.wnrcc and UnmplydP respectively in the plot, have negative coefficient values.

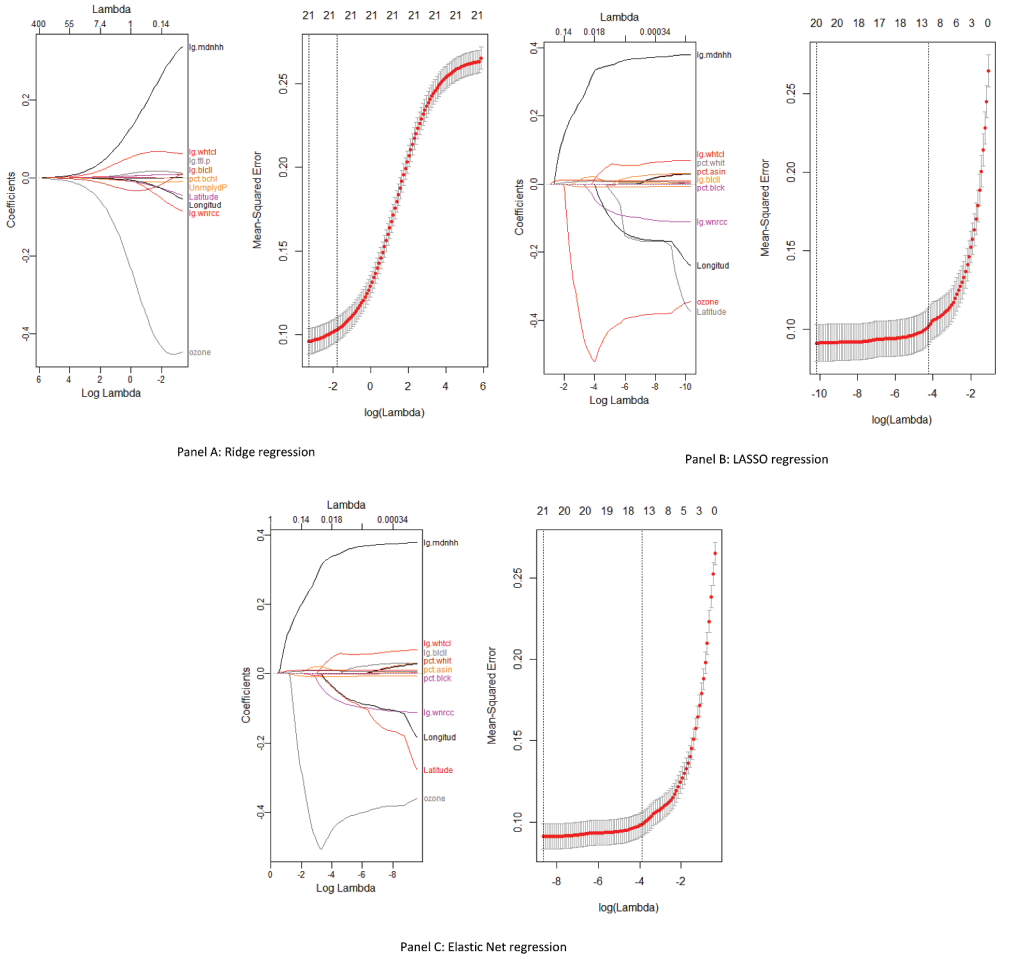

Figure 4. Solution paths and cross-validated MSE of three regularisation methods.

In Panel B, we show the LASSO regularisation paths for the top ten predictors in the model and the corresponding 10-fold cross-validation curve. In the regularisation paths, each line represents a different predictor and shows how the estimated slope coefficient changes for different values of λ. On the right, we have smaller values for λ, and on the left, we have higher values for λ. Unlike the Ridge regression, the LASSO regularisation paths do not vary smoothly. Furthermore, as is well known in the literature, the LASSO regression results in relatively sparse coefficients. Here again, under LASSO, both ln (med.hh.inc) and ozone continue to be the two most influential variables in predicting median housing prices across the census tracts. It is evident that both LASSO and Ridge allow the use of correlated predictors, but the coefficients of correlated predictors are similar in Ridge but not in LASSO. Furthermore, both Ridge and LASSO regression coefficients can move from positive to negative values as they shrunk towards zero. The LASO regularisation paths show that in addition to ln(white.collar), two other variables have estimated coefficient that are positive for a significant portion of the plot; they are

pct.white and pct.asian, given by pct.whit and pct.asin, respectively, in the plot. As λ increases from right to left, some of the coefficients shrink towards zero.

Finally, in Panel C, we show the regularisation paths and cross-validation curve associated with the Elastic Net regressions that provide a useful compromise between Ridge and Lasso. Here, the most important variables are median household income and ozone concentration, followed by latitude and longitude. The regularisation paths of other variables are very similar to what we already observe in the LASSO paths of Panel B. In sum, similar to LASSO, the regularisation paths of the Elastic Net regressions do not fail to provide information on which of the included factors are relatively more important and are helpful for visualising which of the variables have the greatest predictive power for the median house values.

Given the information presented so far, we are in a position to evaluate out-of-sample forecasting performance. In doing so, we investigate how well our presumed MRM specification (1) captures the true relationships among variables beyond the training data.

4.4 Out-of-sample predictions

In this subsection, we evaluate forecasting performance by examining summary statistics based on out-of-sample predictions, which are obtained using various methodologies including those outlined in the previous subsections. As a result, we are able to discuss the relative out-of-sample performance of the supervised ML tools compared to conventional least squares-based methods.

Table 6 presents summary statistics for the out-of-sample predictions where we evenly split the training and testing samples. To maintain consistency with earlier results, we report the out-of-sample prediction errors for OLS and WLS in conjunction with various regularised methods. Instead of focusing on only three specific hyper-parameter α's that control how much L_1- and L_2-norms are used in (6), we consider a range of eleven distinct values for α, starting from 0.0 and ending with 1.0 with an increment of 0.1. Thus, in addition to the Ridge, LASSO and Elastic Net regressions, we report the prediction accuracy of eight other regularised regressions that implement different combination of the L_1- and L_2-norms.

We observe that our best OLS model (given by specification (4) in Table 4) produces an MSE of 0.1019 and Theil's U-statistic of 0.3930. In contrast, the best WLS model (given by specification (8) in Table 4) produces an MSE of 0.1282 and Theil's U-statistic of 0.4935. The reported MSE's of the eleven regularised regressions range from 0.0922 to 0.0956. More specifically, the MSE of the regularised regression corresponding to Ridge ($\alpha = 0.0$) is 0.0956; for Elastic Net ($\alpha = 0.5$) it is 0.0924; and for LASSO ($\alpha = 1.0$) it is 0.0922. In terms of out-of-sample prediction measures, the penalised regression MAE's are quite close to each other and the OLS underperforms compared to the regularised regressions. Among all eleven regularised regressions, $\alpha = 1.0$ produces the second lowest MAE of 0.2060 and the lowest Theil's U-statistic of 0.3559. A close inspection of the table reveals that both Elastic net and LASSO produce very similar values for MDAE, MAPE and MMA. For a large number of α's, Theil's U-statistic is very close to 0.3560. Compared to OLS, the Ridge regression results in a 6.57% reduction in the sum of squared predicted errors. In contrast, when we utilise Elastic net and LASSO, the

Table 6. Performance of alternative methods for the testing set.

Summary Statistics	OLS	WLS	Ridge	Elastic Net									LASSO
	$\lambda = 0$	$\lambda = 0$	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$
Mean Squared Error	0.1019	0.1282	0.0956	0.0922	0.0922	0.0923	0.0924	0.0924	0.0924	0.0922	0.0925	0.0924	0.0922
Mean Absolute Error	0.2195	0.2410	0.2145	0.2071	0.2064	0.2062	0.2061	0.2061	0.2061	0.2061	0.2063	0.2062	0.2060
Median Absolute Error	0.1651	0.1772	0.1659	0.1560	0.1552	0.1554	0.1556	0.1557	0.1560	0.1554	0.1556	0.1558	0.1558
Median Absolute Percentage Error	0.0173	0.0283	0.0168	0.0163	0.0162	0.0162	0.0162	0.0162	0.0162	0.0162	0.0162	0.0162	0.0162
MinMax Accuracy	0.9631	0.9527	0.9835	0.9840	0.9841	0.9841	0.9841	0.9841	0.9841	0.9841	0.9842	0.9841	0.9841
Theil's U test statistic	0.3930	0.4935	0.3692	0.3560	0.3559	0.3561	0.3564	0.3566	0.3568	0.3559	0.3571	0.3566	0.3559

Note: The formulas for summarising forecasting accuracy are as described in the main text.

reductions in the sum of squared prediction errors become 10.35% and 10.15%, respectively. In sum, as we find, based on the out-of-sample prediction errors, the slight dominance of Elastic net and LASSO over the other methods persists. Therefore, from our set of forecasting results, we conclude that the incorporation of shrinkage does improve the out-of-sample performance of the housing valuation model.

4.5 Robustness of predictive performance

Despite the presented evidence, several questions regarding the robustness of our predictive accuracy remain unanswered. These issues, which are related to out-of-sample performance, extend well beyond the notion of variable selections, model specifications and underlying estimation methodologies.

First, one can argue that the conclusions we draw in previous subsections can be sensitive to the fact that we use a random sample by evenly splitting the data. What happens, however, when we compare the out-of-sample performance of alternative methods by using multiple random draws? In Figure 5, we report the outcome of such an experiment. Here we calculate MSE from alternative methods based on 100 random draws of training and testing samples. We find that both Elastic Net and LASSO produce the lowest estimates of test MSE, while the Ridge and OLS underperform. In sum, the reported box plots of simulated MSE over 100 samples confirm our conclusion from Table 6.

Second, a valid question regarding the results shown in Table 6 is whether the ratio of training to testing samples influences the outcome. Do we obtain the same conclusion when we draw different ratios of training to testing samples? To answer this question, in Table 7 we present summary statistics for various out-of-sample predictions obtained from nine different splits of our training and testing samples as outlined in the research design. For brevity, we focus on the MSE and MAE measures using a different seed, and

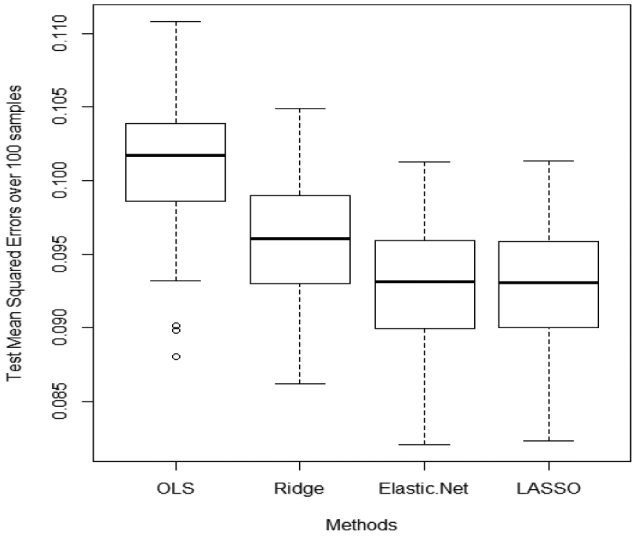

Figure 5. Box plots of simulated mean squared errors from alternative methods.

Table 7. Out-of-sample performance from alternative combination of training and testing samples.

Training/Testing sample percentages	Number of random draws	OLS		Ridge		Elastic Net		LASSO	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
90/10	1	0.1730	0.3349	0.1357	0.2336	0.1324	0.2268	0.1323	0.2268
	10	0.1537	0.3206	0.1015	0.2170	0.0979	0.2091	0.0979	0.2092
	100	0.1507	0.3187	0.0959	0.2144	0.0927	0.2074	0.0927	0.2074
	500	0.1413	0.3189	0.0964	0.2144	0.0932	0.2074	0.0932	0.2073
80/20	1	0.1642	0.3275	0.1184	0.2217	0.1166	0.2155	0.1166	0.2156
	10	0.1412	0.3199	0.0963	0.2140	0.0935	0.2071	0.0935	0.2071
	100	0.1421	0.3188	0.0957	0.2137	0.0926	0.2069	0.0925	0.2069
	500	0.1404	0.3182	0.0961	0.2142	0.0930	0.2072	0.0930	0.2072
70/30	1	0.1608	0.3272	0.1156	0.2206	0.1121	0.2123	0.1123	0.2125
	10	0.1526	0.3197	0.0978	0.2146	0.0950	0.2079	0.0951	0.2079
	100	0.1402	0.3184	0.0953	0.2137	0.0923	0.2069	0.0923	0.2070
	500	0.1409	0.3186	0.0960	0.2138	0.0929	0.2069	0.0929	0.2068
60/40	1	0.1525	0.3230	0.1071	0.2162	0.1034	0.2075	0.1035	0.2077
	10	0.1530	0.3200	0.0973	0.2151	0.0945	0.2083	0.0945	0.2083
	100	0.1418	0.3193	0.0964	0.2147	0.0934	0.2079	0.0934	0.2079
	500	0.1407	0.3187	0.0961	0.2142	0.0931	0.2075	0.0931	0.2075
50/50	1	0.1668	0.3193	0.1019	0.2131	0.0987	0.2050	0.0988	0.2053
	10	0.1522	0.3188	0.0967	0.2150	0.0937	0.2078	0.0937	0.2079
	100	0.1511	0.3188	0.0959	0.2145	0.0930	0.2078	0.0930	0.2078
	500	0.1511	0.3189	0.0960	0.2141	0.0930	0.2074	0.0930	0.2074
40/60	1	0.1464	0.2197	0.1018	0.2136	0.0981	0.2057	0.0982	0.2061
	10	0.1334	0.2195	0.0964	0.2144	0.0936	0.2082	0.0937	0.2084
	100	0.1312	0.2188	0.0963	0.2139	0.0933	0.2071	0.0933	0.2071
	500	0.1317	0.2192	0.0964	0.2144	0.0936	0.2079	0.0936	0.2079
30/70	1	0.1752	0.3203	0.1003	0.2138	0.0964	0.2052	0.0965	0.2057
	10	0.1713	0.3189	0.0970	0.2145	0.0941	0.2083	0.0941	0.2084
	100	0.1515	0.3195	0.0962	0.2143	0.0936	0.2081	0.0936	0.2082
	500	0.1417	0.3196	0.0965	0.2147	0.0939	0.2085	0.0939	0.2085
20/80	1	0.1660	0.3212	0.1002	0.2139	0.0969	0.2062	0.0970	0.2067
	10	0.1534	0.3209	0.0965	0.2145	0.0943	0.2081	0.0943	0.2081
	100	0.1527	0.3200	0.0969	0.2152	0.0946	0.2093	0.0945	0.2093
	500	0.1325	0.3203	0.0970	0.2154	0.0947	0.2098	0.0948	0.2098
10/90	1	0.1856	0.3206	0.0982	0.2123	0.0949	0.2047	0.0949	0.2050
	10	0.1844	0.3211	0.0988	0.2179	0.0967	0.2114	0.0966	0.2113
	100	0.1746	0.3229	0.0986	0.2175	0.0970	0.2126	0.0971	0.2126
	500	0.1546	0.3227	0.0985	0.2172	0.0970	0.2126	0.0970	0.2126

report the estimated out-of-sample accuracy measures for different splits. The reported estimates in each column are based on 1, 10, 100 and 500 random draws of training and testing samples. In all the Elastic Net regressions, we estimate the optimal values for λ and α jointly.

Let us first consider the MSE measures of Table 7. For a small number of draws, all the out-of-sample predictions display high MSE estimates, whereas for a large number of draws, the MSE estimates are low, with OLS being the highest and LASSO being the lowest regardless of the number of draws. For example, when we split the data into 90% training and 10% testing samples only once, the MSE's of OLS and LASSO predictions are 0.1730 and 0.1323, respectively. As we increase the number of random draws, the estimate of average MSE's of all four methods show a slight decrease for virtually all nine different sample splits. Among three regularised-based predictions, irrespective of any sample splits and random draws we consider, the Ridge regression always produces the highest MSE estimate, while the Elastic Net and LASSO regression MSE's are comparable and relatively lower. The lowest MSE's for both Elastic Net and LASSO predictions appear when we split the data into 90% training and 10% testing samples. The MAE results are in line with our findings from the MSE results. Altogether, the out-of-sample experiments of Table 7 convey that our main conclusions about supervised ML tools are not dependent on the data-generating process and are not random occurrences.

Third, another issue that can challenge our conclusion is the underlying 10-fold cross-validation used in all the penalised regressions. To investigate this assertion, we look into possible alternative regressions that use 25, 50, and 100-fold cross-validations, and find that the results are unchanged. As an example, in Figure 6, we report the MSE of the highest-performing Elastic Net regressions under λ-min using 100-fold cross-validation.

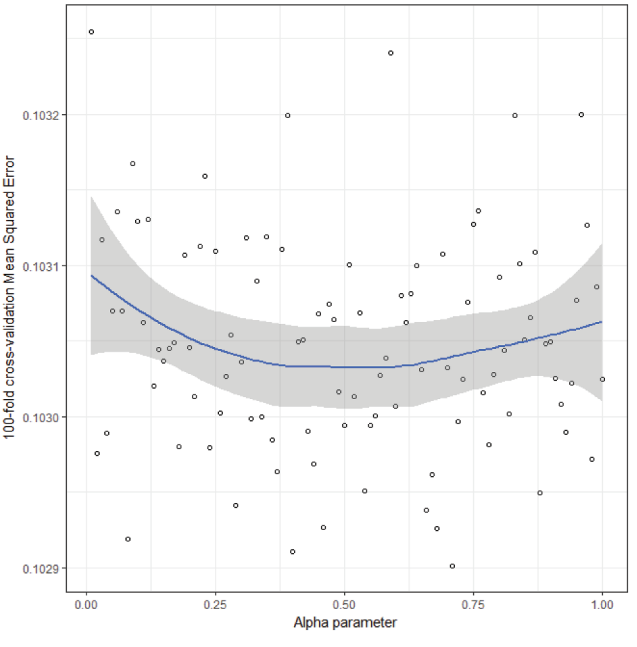

Figure 6. MSE of highest performing Elastic Net regressions as a function of alpha parameter.

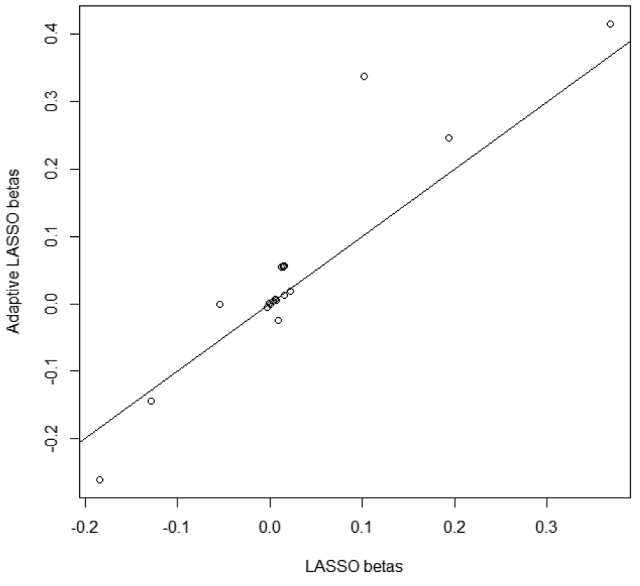

Figure 7. Slope coefficients of highest performing LASSO and Adaptive LASSO regressions.

We observe that the optimal value of α in Figure 6 is not the same as, but is comparable to, our reported Elastic Net regression results of Table 5. In Table 5, the optimal value of α in the Elastic Net regression is 0.613 under λ-min, whereas Figure 6 suggests that the mid-range α is better. Thus, the scatter plot of various MSE's with respect to the alpha parameter shows that the findings are not identical to, but are consistent with, what we have seen so far.

Finally, in order to improve predictive accuracy, it is essential that a researcher not only select a set of relevant variables and transform them without causing a high degree of collinearity but also that the chosen model is computationally robust and able to generalise to new data. Regularisation-based learning tools such as ALASSO offer such alternative options for model generalisability. Thus, we can evaluate the robustness of our highest performing penalised regressions and the resulting out-of-sample performance of MRM by ALASSO. As it turns out, when we estimate (4) and look at the out-of-sample predictive performance, the highest performing ALASSO regression offers little or no increase in predictive accuracy relative to the highest performing LASSO and Elastic Net regressions. There are however slight differences in the estimated regression coefficients. To visualise the differences in the shrinkage, we compare the slope coefficients of the highest performing LASSO and Adaptive LASSO regressions in Figure 7. We see that the ALASSO shrinks less for a large number (7 out of 21) of slope coefficients, while LASSO shrinks less for a small number (3 out of 21) of slope coefficients. The results are unaltered as we increase the number of random draws.

5. Conclusions

The purpose of this paper is to explore the practical implications of using machine learning (ML) techniques to predict owner-occupied housing values. We use data from

California census tracts and measure the influences of a wide set of socio-economic, locational, and environmental factors on median housing values using supervised ML tools. We argue that regularisation-based supervised ML tools provide reasonably precise out-of-sample predictions. Furthermore, as we find, the regularisation paths of the penalised regressions tell us which of the included factors are relatively more important. We also show that the resulting improvement in accuracy of out-of-sample predictions is not limited to the number of random draws or particular data splits.

Our work is not free from limitations. One can argue that our data source is not best-suited for mass appraisals that are instrumental for automated valuation models. It is conceivable that a researcher with access to granular data can achieve better precision from the combination of hedonic regressions and shrinkage-based ML tools. Our research design is also narrow in scope and can be extended in many directions. One possible avenue is to formulate a new theoretical framework and employ an alternative means of regularisation in the context of nonparametric regression (e.g., smoothing splines) that can be used in computer-assisted mass appraisal. Heuristically, one can also argue that an algorithm is not the only way to evaluate predictive performance in real estate analysis and ought certainly to be augmented with other techniques. Furthermore, if the training data does not represent the proper characteristics of the population, learning from the data might be limited in scope. It would be interesting to see whether ML techniques improve the prediction of housing prices for alternative types of aggregated data from housing submarkets such as those outlined by Goodman and Thibodeau (2003, 2007).

Notes

1. According to the guidelines of IAAO (2013, p. 5), mass appraisal requires complete and accurate data, effective valuation models, and proper resource management.
2. An incomplete list of works in this area includes Borst and McCluskey (2008, 2011), Bourassa et al. (2007), Dubin (1998), Goodman and Thibodeau (2003, 2007), Hausler, Ruscheinsky and Lang (2018), Lin and Mohan (2011), Páez (2005), Perez-Rave et al. (2019), Worzala et al. (1995), and Xu (2008).
3. Traditionally, hedonic regression models used for mass appraisal have employed a nonlinear MRA framework to explain the variability of housing prices at the disaggregated level.
4. The advent of cloud computing and the increasing availability of ML codes in R and Python languages have also created a favourable environment for large exploratory data analysis that is suitable for real estate analysis.
5. For example, Pavlov (2000) suggests that because of model misspsecifications and the influence of omitted variables, the implicit prices of housing attributes can be misleading.
6. Some recent work examines the time-series forecastability of housing prices under a data-rich environment. For example, Bork and Møller (2018) discuss how to reduce the dimension of a large set of predictor variables by using principal component analysis, partial least squares (PLS), and sparse PLS methods.
7. Goodman and Thibodeau (2003) utilise 28,561 single-family transactions from Dallas County and evaluate their predictive accuracy using three types of alternative housing submarket constructions: zip codes, census tracts, and hierarchical model. In contrast, Goodman and Thibodeau (2007) analyse 44,000 sales of single-family properties and examine two alternative procedures for delineating housing submarkets within the Dallas metropolitan-area market. The first procedure combines spatially adjacent census block groups and the second procedure allows spatial discontinuities.

8. Other related works that discuss various spatial statistical methods includes Case et al. (2004), who use a large sample of 50,000 transactions from Fairfax county, Virginia and compare out-of-sample prediction accuracy using a particular split of the data. Unlike the work of Bourassa et al. (2007), Case et al. (2004) explores out-of-sample predictive accuracy by using only one split of the data. There are complex methodological issues associated with penalised regressions in the presence of spatial dependence and thus a discussion of various spatial regression models using regularisation methods is beyond the current scope.
9. For example, the total number of census tracts increases from 5,732 in 1980 to 5,858 in 1990, and further to 7,049 in 2000. For many reasons, the number of census tracts during the pre-2000 years was considerably lower. In 1970 and 1980 the entire state of California had not yet been tracted. In 1990, in addition to the respective tracts from 1970 and 1980, data from enumeration districts (EDs) and census county divisions (CCDs) have also been incorporated.
10. However, as mentioned in the general Census guidelines, the converted 2000 tract data cannot be considered official U.S. Census Bureau data or California Department of Finance data. Even within 7,904 census tracts for which we have non-missing observations in 2010, 18 census tracts have $0 median values and 4 census tracts have $9,999 median values for the converted 2000 series.
11. Note that the matrix of prediction variables X may include an intercept term. In practice, if we centre X and y before computing the regression, the intercept becomes zero.
12. The specifications we use to capture geographical dependence are comparable to those of Case et al. (2004), Dubin (1998), and Xu (2008).
13. Apart from the predictive accuracy issue, another shortcoming of both OLS and WLS is the lack of interpretive ability. In a high-dimensional regression setting, where the number of predictors is large, we may want to implement a fitting procedure with a subset of important predictors. Interestingly, the above-mentioned issues have little to do with classical regression model assumptions.
14. In essence, the Ridge regression is a continuous shrinkage method that retains all the covariates but penalises large coefficients through the L2-norm. Unlike the Ridge regression, which keeps all the predictors in the presumed model, in the LASSO regression, the presence of multicollinearity results in the dropping of certain predictors while retaining others.
15. The first part of the loss function in Ridge regression is the same as the RSS of OLS. The second part of the loss function involves the regularisation of parameters because it penalises larger coefficient values. It is noticeable that as $\lambda \to 0$, $\hat{\beta}_{Ridge} \to \hat{\beta}_{OLS}$, and as $\lambda \to \infty$, $\hat{\beta}_{Ridge} \to 0$.
16. As mentioned by Xu (2008), incorporating absolute location using spatial coordinates in conjunction with the polynomial expansion approach can capture heterogeneity in housing attribute prices.
17. It is important to note that while the WLS regressions correct for issues such as cross-sectional heteroskedasticity, one has to be careful about interpreting the high value of $\bar{R}^2$ associated with such regressions, which may not guarantee the success of the underlying model's out-of-sample predictive capacity. We highlight this issue in a future subsection.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Dr. Prodosh E. Simlai is a Professor of Economics at the Department of Economics & Finance, Nistler College of Business and Public Administration, University of North Dakota, USA. Dr. Simlai received his M.S in Finance and Ph.D. in Economics from the University of Illinois at

Urbana-Champaign, USA. His research interests include financial markets, real estate, and applied econometrics. Dr. Simlai's work has appeared in both leading general interest and field journals including the Accounting Research Journal, Business Economics, Finance Research Letters, International Review of Financial Analysis, Journal of Asset Management, Journal of Derivatives and Hedge Funds, Journal of Real Estate Finance and Economics, Quarterly Review of Economics and Finance, Studies in Economics and Finance, and Research in Finance among others.

References

Athey, S. (2018). The impact of machine learning on economics. In A. K. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An Agenda* (pp. 507–547). University of Chicago Press.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–485. 10.1257/aer.p20151021.

Bitter, C., Mulligan, G. F., & Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, *9*(1), 7–27. 10.1007/s10109-006-0028-7.

Borde, S., Rane, A., Shende, G., & Shetty, S. (2017). Real estate investment advising using machine learning. *International Research Journal of Engineering and Technology*, *4*(3), 1821–1825.

Bork, L., & Møller, S. V. (2018). Housing price forecastability: A factor analysis. *Real Estate Analysis*, *46*(3), 582–611. 10.1111/1540-6229.12185.

Borst, R. A., & McCluskey, W. J. (2008). Using geographically weighted regression to detect housing submarkets: modelling large-scale spatial variations in value. *Journal of Property Tax Assessment and Administration*, *5*(1), 21–51.

Borst, R. A., & McCluskey, W. J. (2011). Detecting and validating residential housing submarkets: A geostatistical approach for use in mass appraisal. *International Journal of Housing Markets and Analysis*, *4*(3), 290–318. 10.1108/17538271111153040.

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

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, *14*(1/2), 203–222. 10.1023/A:1007744706720.

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

Case, B., Colwell, P. F., Leishman, C., & Watkins, C. (2006). The impact of environmental contamination on condo prices: A hybrid repeat-sale/hedonic approach. *Real Estate Economics*, *34*(1), 77–107. 10.1111/j.1540-6229.2006.00160.x.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–265. 10.1257/aer.p20171038.

Dubin, R. (1998). Predicting house prices using multiple listings data. *Journal of Real Estate Finance and Economics*, *17*(1), 35–59. 10.1023/A:1007751112669.

Eckland, I., Heckman, J., & Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, *112*(S1), 60–109. 10.1086/379947.

Epple, D. (1987). Hedonic prices and implicit markets: estimating demand and supply functions for differentiated products. *Journal of Political Economy*, *95*(1), 59–80. 10.1086/261441.

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modeling spatial variation in housing prices: A variable interaction approach. *Real Estate Economics*, *31*(4), 623–646. 10.1046/j.1080-8620.2003.00079.x.

Friedman, M. (1953). The methodology of positive economics. In Milton Friedman (Eds.),*Essays in positive economics* (pp. 3–43). Chicago: The University of Chicago Press.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, *12*(3), 181–201. 10.1016/S1051-1377(03)00031-7.
Goodman, A. C., & Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics*, *35*(2), 209–232. 10.1111/j.1540-6229.2007.00188.x.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edition). SpringerVerlag. 2009.
Hausler, J., Ruscheinsky, J., & Lang, M. (2018). News-based sentiment analysis in real estate: A machine learning approach. *Journal of Property Research*, *35*(4), 344–371. 10.1080/09599916.2018.1551923.
Hoerl, A., & Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. 10.1080/00401706.1970.10488634.
Hoerl, A., & Kennard, R. (1988). Ridge regression. *Encyclopedia of Statistical Sciences*, *8*, 129–136. New York: Wiley.
IAAO. (2013). *Standard on the mass appraisal of real property*. International Association of Assessing Officers.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. Springer. Ladd, Helen.
Kauko, T., & d'Amato, M. (2008). Introduction: Suitability issues in mass appraisal methodology. In T. Kauko & M. D. Amato (Eds.), *Mass appraisal methods: An international perspective for property valuers* (pp. 1–24). Wiley-Blackwell.
Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, *74*(2), 132–156. 10.1086/259131.
Lin, C. C., & Mohan, S. B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis*, *4*(3), 224–243. 10.1108/17538271111153013.
Malpezzi, S. (2002). Hedonic pricing models: A selective and applied review, In T. O'Sullivan and K. Gibb, *Housing Economics and Public Policy* (pp. 67–89). Oxford, UK: Blackwell Science Ltd.
Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. O. Sullivan & K. Gibb (Eds.), Housing Economics and Public Policy (pp. 67–89). Oxford, UK: Blackwell Science Ltd.
McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265. 10.1080/09599916.2013.781204.
McMillen, D. P. (2010). Issues in spatial data analysis. *Journal of Regional Science*, *50*(1), 119–141. 10.1111/j.1467-9787.2009.00656.x.
Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. 10.1257/jep.31.2.87.
Páez, A. (2005). Local analysis of spatial relationships: A comparison of GWR and the expansion method. In Gervasi O. et al. (Eds.), *Computational Science and Its Applications* – ICCSA 2005. Lecture Notes in Computer Science, vol 3482 (pp. 162–172). Berlin, Heidelberg: Springer.
Pavlov, A. D. (2000). Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Economics*, *28*(2), 249–283. 10.1111/1540-6229.00801.
Perez-Rave, J. I., Correa-Morales, J. C., & Gonzalez-Echavarria, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*(1), 59–96. 10.1080/09599916.2019.1587489.
Peterson, S., & Flanagan, A. B. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164. 10.1080/10835547.2009.12091245.
Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34–55. 10.1086/260169.
Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44. 10.1080/10835547.2005.12090154.
Stock, J. H., & Watson, M. W. (2019). *Introduction to Econometrics*. Pearson.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *58*(1), 267–288. 10.1111/j.2517-6161.1996.tb02080.x

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment and Finance*, *38*(3), 213–225. 10.1108/JPIF-12-2019-0157.

Vanderford, S. E., Mimura, Y., & Sweaney, A. L. (2005). A hedonic price comparison of manufactured and site-built homes in the non-MSA U.S. *Journal of Real Estate Research*, *27*(1), 83–104. 10.1080/10835547.2005.12091151

Varian, H. R. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. 10.1257/jep.28.2.3.

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

Xiao, Y. (2017). Hedonic housing price theory review. In *Urban Morphology and Housing Market* (pp. 11–40). Tongji University Press and Springer Nature Singapore Pte Ltd.

Xu, T. (2008). Heterogeneity in housing attribute prices: A study of the interaction behaviour between property specifics, location coordinates and buyers' characteristics. *International Journal of Housing Markets and Analysis*, *1*(2), 166–181. 10.1108/17538270810877781.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. 10.1198/016214506000000735.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(2), 301–320. 10.1111/j.1467-9868.2005.00503.x.