Ann Math Artif Intell (2017) 81:71–84
DOI 10.1007/s10472-016-9534-6

Reliable region predictions for automated valuation models

Anthony Bellotti[1]

Published online: 19 January 2017

Abstract Accurate property valuation is important for property purchasers, investors and for mortgage-providers to assess credit risk in the mortgage market. Automated valuation models (AVM) are being developed to provide cheap, objective valuations that allow dynamic updating of property values over the term of a mortgage. A useful feature of automated valuations is to provide a region of plausible price estimates for each individual property, rather than just a single point estimate. This would allow buyers and sellers to understand uncertainty on pricing individual properties and mortgage providers to include conservatism in their credit risk assessment. In this study, Conformal Predictors (CP) are used to provide such region predictions, whilst strictly controlling for predictive accuracy. We show how an AVM can be constructed using a CP, based on an underlying k-nearest neighbours approach. Time trend in property prices is dealt with by assuming a systematic effect over time and adjusting prices in the training data accordingly. The AVM is tested on a large data set of London property prices. Region predictions are shown to be reliable and the efficiency, ie region width, of property price predictions is investigated. In particular, a regression model is constructed to model the uncertainty in price prediction linked to property characteristics.

Keywords Automated valuation · Conformal predictor · Nearest neighbours

Mathematics Subject Classification (2010) 62P20

✉ Anthony Bellotti
a.bellotti@imperial.ac.uk

[1] Department of Mathematics, Imperial College London, South Kensington, London SW7 2AZ, UK

1 Introduction

Automated valuation models (AVMs) have been developed by both banks and specialist analytics companies, such as the members of the European AVM Alliance. They provide computer-generated valuations of property prices at an individual property level. Such valuations are essential to both buyers and sellers of properties and mortgage-providers who need to determine value of collateral on a mortgage. Traditionally, property valuations have been conducted by trained surveyors. However, if AVMs are developed with sufficient accuracy they may be cheaper, faster, more objective and more transparent than human surveyors. There is a parallel with the advent of credit scoring which automated lending decisions, traditionally made by bank managers. Because AVM estimates are cheap and fast, this would allow estate agents and mortgage-providers to dynamically update the value of a property. Through the lifetime of a mortgage, this would help the mortgage-provider assess risk and provide improved customer service facilites, such as re-mortgaging. There are many other possible uses of AVMs, such as portfolio valuation and fraud or negligence detection. They are already being used increasingly in property markets internationally; eg in the UK, they are applied in an estimated 30 % of mortgage originations [6, 7].

Typically, AVMs are developed as segmented models using a nearest neighbours approach, based on a rich data set of past house prices and variables such as:

- Property characteristics (eg size, number of rooms, garden, view from balcony);
- Local environment (eg schools, transportation, local services);
- Historic prices and economic conditions.

Statistically, the AVM problem is a regression problem and the goal of the AVMs is to produce the most accurate predictions of individual property prices when contrasted against actual purchase prices in the future. One of the earliest published models of house price is based on the Boston Housing data, published in 1978, linking house price to property features (number of rooms and age), neighbourhood features (such as social status and crime rate), accessibility and air pollution measures [8]. Using linear regression, this study found that, in particular, controlling for this wide range of variables, there remains a negative association between air pollution levels and house price.

Although AVMs primarily produce point estimates of price, it is also useful to have a measure of confidence in the estimates. In particular, an AVM that can output a *prediction interval* would be very useful for the following reasons.

1. They would allow us to understand the overall accuracy of the AVM. Indeed, particular property segments for which the AVM generates broad prediction intervals would suggest areas for which the AVM could be improved in future model development.
2. At the individual property level or segment level, they allow us to understand which properties are harder or easier to price. This could be practically useful since it would suggest difficult properties that should be followed-up with detailed manual surveying; whilst the properties for which the AVM generates a sufficiently narrow prediction interval, would not need to be followed-up.
3. They would enable mortgage-providers to take conservative lower bound estimates of future property prices.

For prediction intervals to work well, we need to ensure that they are reliable, in the sense that with a certain probability, the actual purchase price will be within the prediction interval. Traditional statistical methods based on classical and Bayesian approaches can be used. However, both frameworks rely on making strong underlying distributional assumptions about the data, to derive reliable prediction intervals. An alternative machine

learning methodology is the *conformal predictor* (CP). This algorithm will generate predictive intervals, or more generally, *region predictions* that are *guaranteed* to be reliable at any user-defined confidence level, in both a transductive and inductive setting [13]. The only distributional assumption they make is that the data is exchangeable, of which independent and identically distributed (iid) is a special case. The user of the CP can choose any confidence level. However, the consequence of higher confidence levels are broader region predictions. Hence an important second performance measure for AVMs outputting region predictions is the region size, or *inefficiency*, of the predictions. The narrower these are, the more useful, precise and efficient the AVM. An additional quality of CPs is that they are constructed as wrappers over existing machine learning or statistical point estimates. So, if an AVM has already been built that produces good point predictions of price, then the CP can be wrapped around it, to convert it into a reliable region predictor and the CP will draw on the power of the existing AVM to produce efficient predictions. CPs have been applied in many application domains, such as cancer diagnosis, biometrics, anomaly detection and network traffic classification [3] and have been developed for both classification and regression. CP for regression has been applied to the Boston Housing data [11], but the focus of that study was to show that regression CP performs well and reliably with respect to several different application areas. In contrast, the aim of this study is to investigate whether CPs are useful specifically in the domain of AVMs for prediction of individual property price. For that reason, the Boston Housing data is not quite appropriate since the outcome variable is the aggregate log median house price for metropolitan areas of Boston, rather than individual house price, and the sample size is small (506 examples). Instead we apply the method to a recent, large data set of individual property prices in London, as described below.

For property price prediction, there is a clear problem when estimating and predicting time trends in price movements. This is a problem for any AVM but especially so for CP since it violates the exchangeability assumption. In this study, this problem is addressed firstly by assuming a single systematic effect for price movements over time, secondly, adjusting prices in the model training data set by an estimate of price changes up to the time of the predicted property price(s), then, thirdly, allowing for uncertainty in forecast price changes by generating region predictions which are unions of region predictions for a range of possible price changes. This is explained in detail in Section 2.4.

The data used in this study are property transactions for London and its suburbs, derived from the Land Registry (UK), which is supplemented by geographical data on railway and tube stations and deprivation data. This is sufficient to demonstrate the use of CP with AVM as a proof-of-concept. However, without detailed property characteristics, which are not available publicly, the output of this study is not immediately of practical value. However, with the inclusion of such data, this study demonstrates that CP could produce reliable region predictions of property prices. In Section 2, the CP is described in more detail, along with the underlying weighted k-nearest neighbours algorithm that will be used as the basis of the model. Specific issues applying CP to AVM are also discussed. In Section 3 we describe the data and Section 4 presents results that show that the AVM-CP is reliable and investigates segments of inefficiency amongst the predictions. Finally, Section 5 presents some final conclusions and ideas for further work.

2 Methodology

Consider n examples. Let $\mathbf{x}_i$ denote a vector of predictor variables and y_i a real number outcome variable for each example $i \in \{1, \cdots, n\}$. For this paper the outcome is log of

price. Log price is used since the distribution of property prices is typically right skewed and the log allows property price movements over time to be expressed by additive terms, rather than % change.

2.1 *k*-nearest neighbours for regression

Typically the base predictive algorithm for AVM is k-nearest neighbours (k-NN) for regression. Define a distance metric between examples, d. For a new example given as $\mathbf{x}_{\text{new}}$, let $n(j)$ denote the index of the jth nearest neighbour to $\mathbf{x}_{\text{new}}$ from a pool of n training examples, based on the distances $d(\mathbf{x}_{\text{new}}, \mathbf{x}_i)$ for $i \in \{1, \cdots, n\}$. Then the predicted outcome for $\mathbf{x}_{\text{new}}$ is the weighted mean of outcomes of the k nearest neighbours:

$$\hat{y} = \frac{\sum_{j=1}^{k} w_j y_{n(j)}}{\sum_{j=1}^{k} w_j}. \tag{1}$$

Training examples with similar distance to the new example should naturally have similar weights and the further away a training example, the less weight it will have, hence a good choice of weight is

$$w_j = \exp\left(-\lambda d\left(\mathbf{x}_{\text{new}}, \mathbf{x}_{n(j)}\right)\right) \tag{2}$$

where $\lambda \geq 0$ is a user-defined decay parameter. For an examination of distance-weighted k-NN, see [10]. The usual Euclidean distance metric is used for this study:

$$d\left(\mathbf{x}, \mathbf{x}'\right) = \sqrt{||\mathbf{x} - \mathbf{x}'||_2}. \tag{3}$$

It can be shown that k-NN converges to no more than two times the optimal Bayes error rate as $n \to \infty$ [4]. However, k-NN is sensitive to the scaling of predictor variables and to achieve optimal solutions for small sample size, rescaling of the predictor variables is advisable prior to computation of distances. A simple approach is to normalize all predictor variables to have variance 1. However, in this study the method successfully applied by [9] is used: a linear regression model is developed using predictor variables $\mathbf{x}_i$ and outcome variable y_i for training data $i \in \{1, \cdots, n\}$, then the magnitude of coefficient estimates from this model are used to rescale predictor variables in k-NN.

Since this study uses a large data set of housing data, an efficient implementation of k-NN is required, using kd-trees and approximation methods [2]. In particular, the `RANN` package is used in the `R` statistical programming language.

2.2 Inductive conformal predictors for regression

Conformal predictors (CP) are a class of machine learning algorithms that can produce predictions in the form of *regions* which are sets of possible outcomes. For regression, the region will typically form a *prediction interval*. They have the important property that they are reliable in the sense that accuracy of prediction is precisely bound by a user-defined confidence level. CPs are usually constructed on the basis of existing machine learning algorithms that output point predictions without reliability. Although CPs were initially proposed for the online learning setting, for this study an inductive setting is required, hence reliability of CPs is expressed for the inductive conformal predictor (ICP) learning environment, as follows:-

1. Let examples 1 to $l - 1$ represent a proper training set, for some $l < n$;
2. Let examples l to n represents a calibration data set;
3. Let new examples $n + 1$ to $n + m$ represent a test set of size m.

Use the proper training set to construct a statistical model or decision rule D. Given a new example $\mathbf{x}$, $D(\mathbf{x})$ will output a point prediction of outcome. Then, for a given confidence level $1-\epsilon$, define the region prediction for each test example $n+h$ as

$$R_h = \left\{ y \in \mathbb{R} : \frac{|\{i \in \{l, \cdots, n\} : \alpha_i \geq \alpha_{n+h}\}| + 1}{n - l + 2} > \epsilon \right\} \tag{4}$$

where

$$\alpha_i = \begin{cases} A(D(\mathbf{x}_i), y_i) & \text{if } i \leq n, \\ A(D(\mathbf{x}_i), y) & \text{if } i > n \end{cases} \tag{5}$$

and A is a non-conformity measure (NCM) which expresses how strange the second argument is in relation to the prediction made by D.

Then, assuming that the data are exchangeable, the events that the true outcomes are in the regions, ie $y_{n+h} \in R_j$, are independent binomial events with probability

$$P(y_{n+h} \in R_j) \geq 1 - \epsilon. \tag{6}$$

It is this property that demonstrates the reliability of CP; in particular, *conservative validity*. See [13] for proof.

For this study, we use the following NCM:

$$A(\hat{y}, y) = \frac{|y - \hat{y}|}{\hat{s} + r} \tag{7}$$

where $\hat{s}$ estimates the standard deviation of the prediction made by D and $r \geq 0$ is a user-defined parameter, following [11]. The numerator measures the disparity between the prediction and proposed value y, but this is relative to the standard deviation of the prediction given as the denominator. This is because we want to measure high non-conformity when the proposed value is far from the typical range of the predicted value. The parameter r controls the importance of the standard deviation, with larger values of r indicating that the standard deviation is less useful in the NCM. It follows that

$$R_h = [\hat{y}_h - \gamma(\hat{s}_h + r), \hat{y}_h + \gamma(\hat{s}_h + r)] \tag{8}$$

where γ is the $(1-\epsilon)$-quantile of the empirical distribution of the NCMs on the calibration data set, ie $\alpha_l, \cdots, \alpha_n$ [11].

For this study, the NCM is based on k-NN so $\hat{y}_h$ is computed using formula (1) and $\hat{s}_h$ is computed by a corresponding weighted sample standard deviation

$$\hat{s} = \frac{\sum_{j=1}^{k} w_j (y_{n(j)} - \hat{y})^2}{\sum_{j=1}^{k} w_j}. \tag{9}$$

For AVM, this proposed NCM can be interpreted as follows: the further the proposed price y is from that predicted by k-NN, the higher its non-conformity measure, but this is relative to the range of prices offered by the k nearest neighbours. Informally, a prediction based on a large range of underlying prices would accept a larger range of proposed prices to be conformal.

2.3 Performance measures

For regression algorithms outputting point estimates, root mean square error (RMSE) is typically used as a performance measure. However, region predictions are different since error is indicated by whether the true outcome is in the region prediction or not. In particular, the performance of CP is measured for reliability and efficiency.

Reliability ensures that the CP is behaving according to the theory. This is necessary not only to test for coding errors but also to check the exchangeability assumption for the data. Accuracy on the test set is given by

$$\text{Acc} = \frac{1}{m}\sum_{h=1}^{m} I\left(y_{n+h} \in R_h\right) \tag{10}$$

Following (6), if Acc $\geq 1-\epsilon$ then the CP is exhibiting reliability. If Acc $< 1-\epsilon$, then we can perform a binomial test with null hypothesis that Acc follows a binomial distribution with probability at least $1-\epsilon$.

Efficiency measures how useful the prediction is. The narrower the region, the more precisely the prediction tells us about the true outcome. Hence region size is a good measure of *inefficiency*. In this study, the region is an interval $R_h = [a_h, b_h]$ for some a_h and b_h, in which case the inefficiency is given as $b_h - a_h$ and inefficiency on the test set is

$$\text{Ineff} = \frac{1}{m}\sum_{h=1}^{m} (b_h - a_h). \tag{11}$$

2.4 Handling price changes over time

Property prices are prone to strong movements over time and this needs to be factored into the analysis. Indeed, from 2013 to 2014, property prices in London generally increased by 11 %. In particular, CP relies on data being exchangeable and price changes over time will violate this assumption. This problem is dealt with by supposing a single systematic factor that is contributing equally to general property prices in London. This factor takes into account general economic and market conditions over time. Let p_t be price at time t for any particular property and $s(t, t')$ represent systematic price change from time t to t', as a fraction. Then price change is expressed as

$$p_t = p_{t'}(1 + s(t, t'))\epsilon \tag{12}$$

where $\epsilon > 0$ is an idyiosyncratic factor for the particular property. Working with log prices, the systematic term becomes the additive term

$$s^*(t, t') = \log\left(1 + s(t, t')\right) \tag{13}$$

which can be estimated from training data using linear regression. A linear term can be used to model s^* and tested to check whether non-linearity is required. Additionally, this systematic term can be extrapolated forward so that an estimate of $s^*(t, t^{(\text{test})})$ is computed, where $t^{(\text{test})}$ is the time at which prices in the test set need to be forecast. Then each log price in the training and calibration data is adjusted by an additive term $s^*(t_i, t^{(\text{test})})$ to create the outcome variable, where t_i is the time at which the price for example i was observed. Once all the data, training, calibration and test, have a common time reference for the log price outcome variable, the data can reasonably be assumed to be exchangeable.

The assumption of a single systematic effect is realistic since property price movements will generally have a common trend. Many property price movements will be the result of idiosyncratic factors, such as ageing of property and renovation. However, these idiosyncratic factors would be additional conditions we would expect the model to account for through the ϵ error term. On the other hand, there will be other systematic factors affecting price increases for different groups of properties: eg different localities will have somewhat different price movements, or different types of property (eg semi-detached) may be in more demand than others. Handling these factors would be the topic of an extended study,

using expert knowledge of property price movements. The use of a single environmental systematic factor mirrors the use of the one-factor model approach in credit risk [12].

The estimation of the systematic term over the period of the training data should be accurate since it is an in-time estimate. However, the estimation over the period from the end of the training data and the test data is less reliable. This period is likely to be long (ie several months) because of the delay in reporting property sale transactions. The source of this estimation can be based on past estimates of property price changes, but should be supplemented by expert jugement. However expert judgement can vary and with property price forecasting it usually does. Therefore, in this study, the approach taken is to allow multiple estimates of $s^*(t^{(\text{train})}, t^{(\text{test})})$ where $t^{(\text{train})}$ is the last date in the training data and from these construct bounds $l \leq s^*(t^{(\text{train})}, t^{(\text{test})}) \leq u$ that we believe with near certainty. Then consider all CP built with an additive adjustment on log prices y_i of $s^*(t_i, t^{(\text{train})}) + e$ for all $e \in [l, u]$. Using k-NN for regression as NCM, as defined in Section 2.2, this implies that $\hat{y}$ changes by an additive amount e, relative to the prediction given with data adjusted just to time $t^{(\text{train})}$, whilst $\hat{s}$ remains unchanged for all e. Therefore, we define the union

$$R_h^* := \bigcup_{e \in [l,u]} R_{h,e} = [\hat{y}_h + l - \gamma(\hat{s}_h + r), \hat{y}_h + u + \gamma(\hat{s}_h + r)] \tag{14}$$

where $R_{h,e}$ denotes region R_h from (8), constructed with price adjustment e. Since the true price change must be one of these regions, for adjustment e' say,

$$P(y_{n+h} \in R_h^*) \geq P(y_{n+h} \in R_{h,e'}) \geq 1 - \epsilon \tag{15}$$

from (6). Therefore R_h^* is a conservative region prediction and is used as the region prediction in this study.

2.5 Analysing region inefficiency

Each test example will have its own individual inefficiency. It would be useful to know the drivers of inefficient predictions, firstly, because this would give insight into which properties have unclear pricing (and which are more stable), and also would suggest segments of the data that require better modelling. This information can be used to direct future model development. This analysis can be conducted by taking individual inefficiency as given in Section 2.3, $b_h - a_h$, as the outcome variable, using linear regression against the predictor variables.

2.6 Experimental procedure

We follow this experimental procedure:

1. Extract a training/calibration data set `TC` and a test data set `Test` from the property data set, such that there is a realistic delay period between the two data sets.
2. Model price trend function s^* over period of `TC`.
3. Compute lower and upper bounds on s^* over period between `TC` and `Test`, based on minimum and maximum price changes within extended property price data.
4. Adjust prices in `TC` by the estimated s^* function as described in Section 2.4.
5. Use linear regression on `TC` to estimate coefficients for rescaling variables in the distance metric d.
6. Find optimal hyperparameters k and λ on `TC` using grid search, with RMSE as performance measure.

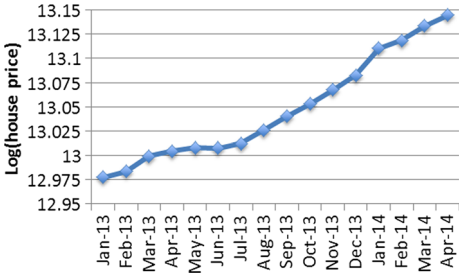

Fig. 1 London house prices; source: Acadata

7. Geographical location will be included in the distance metric d, but cannot be included in step 5, hence use grid search to find optimal rescaling, with RMSE as performance measure.
8. Randomly sample `TC` into separate training, calibration and test data sets. Use this with settings from steps 5 to 7 to perform a grid search to find optimal value of r in the NCM formula (7), with efficiency as performance measure and a fixed confidence level (0.9).
9. Randomly sample `TC` into separate training, `Train`, and calibration, `Cal`, data sets.
10. Run ICP with training and calibration data sets, `Train` and `Cal`, to extract region predictions on test set `Test`. Report final performance on `Test` and use linear regression to analyse inefficiency as described in Section 2.5.

3 Data

Residential property price data for England are made publicly available by the Land Registry in the UK and was originally sourced as part of a Kaggle competition to predict London house prices, along with locations of London tube and railway stations. This has been augmented by local deprivation data.

The Property Price data consists of details for property sales in and around London over the period 2009 to 2014. Along with the sale price and transaction date, other data is also provided about the property: whether new build, whether free- or leasehold and geographical information this includes grid reference and Lower layer Super Output Areas (LSOA) which are small regions defined in England by the Office of National Statistics. For this study we aim to predict prices in April 2014 (22,145 records). Knowing that there is up to a 3-month delay between completion of transaction and publication by Land Registry, we use a 3-month delay between `TC` and `Test`, hence our training/calibration data is taken from 2013 (302,978 records). The Land Registry data has been summarized by Acadata, LSL Property Services Ltd, and Fig. 1 gives a summary of log mean price over the period of interest, 2013 to early 2014.

The stations data set lists all London stations, along with whether they are tube or railway, along with geographical location. Station data is joined with property transaction data by finding the two nearest stations to each property, based on geographical location and using the nearest neighbours algorithm.

A rich source of deprivation data is publicly available as official statistics from `www.gov.uk` and is based on census information. This study uses the English indices of deprivation 2010 which measures relative levels of deprivation at LSOA level. The associations between levels of deprivation and property price will be used to develop improved

Table 1 Summary statistics for `TC` and `Test` data sets

Variable	*Training/calibration summary*	*Test data summary*
Log(price)	12.55 (0.616)	12.60 (0.632)
Property Type: Flat	97120 (32.1 %)	7216 (32.6 %)
Property Type: Semi-detached	65181 (21.5 %)	4644 (21.0 %)
Property Type: Terraced	83920 (27.7 %)	6187 (27.9 %)
Property Type: Detached	56718 (18.7 %)	4093 (18.5 %)
Leasehold	99596 (32.9 %)	7407 (33.5 %)
New build	29945 (9.88 %)	1028 (4.64 %)
log(distance from centre)	10.3 (0.95)	10.3 (0.94)
log(distance from 1st station)	6.99 (1.01)	7.01 (1.01)
log(distance from 2nd station - distance from 1st station + 1)	6.27 (1.75)	6.30 (1.74)
Station: Both tube and railway	38063 (12.6 %)	2699 (12.2 %)
Station: Tube	31195 (10.3 %)	2188 (9.88 %)
Station: Railway	233681 (77.1%)	17253 (77.9 %)
Income score	0.119 (0.0884)	0.121 (0.0889)
Employment score	0.0723 (0.0431)	0.0733 (0.0436)
HD score	-0.438 (0.802)	−0.426 (0.803)
Crime score	-0.00444 (0.729)	−0.00137 (0.730)
LIV score	21.5 (16.1)	21.3 (15.9)
Environment: Indoors score	18.5 (15.6)	18.4 (15.7)
Environment: Outdoors score	27.7 (23.4)	27.2 (22.9)
GB score	20.3 (18.6)	20.2 (18.6)
WB score	30.6 (23.8)	30.4 (23.6)
Education: Child score	15.9 (14.6)	16.3 (14.6)
Education: Skills score	11.8 (12.3)	12.1 (12.3)
IDACI score	0.186 (0.154)	0.188 (0.154)
IDAOPI score	0.177 (0.123)	0.178 (0.122)
Young population score	−1.51 (0.321)	−1.51 (0.316)
Old population score	−1.47 (0.382)	−1.46 (0.376)

For categorical variables, figures are: frequency (percentage). For continuous variables, figures are: mean (standard deviation)

predictive models of property price. However, it is important to understand that the deprivation measures cannot be used inversely to measure affluence [5]. The following deprivation scores are available: income, employment, health and disability (HD), education for children and skills for adults, barriers to housing and services with sub-domains wider barriers (WB) and geographical barriers (GB), crime, living environment score (LIV) with sub-domains for indoor and outdoor living (ie quality of housing and external environment, respectively). WB relates to household overcrowding, homelessness and access to owner-occupation. GB relates to distance to local services such as GP, shops, schools and Post Office. Additional to the general income score, separate scores for income deprivation affecting children (IDACI) and the older population (IDAOPI) are provided. In all cases, a higher value of the score

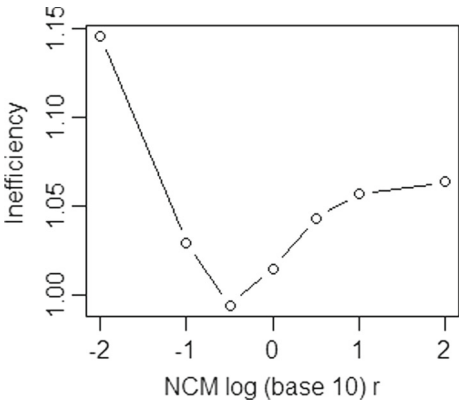

Fig. 2 Performance of ICP for different values of r in NCM, taking $k = 10$; demonstrates minimal inefficiency at $r = 10^{-0.5}$

indicates higher level of deprivation. For detailed information, see [5]. Population scores are also computed as log-odds of young (age 0–15) and old (60+) populations in each LSOA. There are just over 12000 LSOAs in the London property data and the deprivation data is linked to the property records using these LSOA. Hence, multiple properties will share the same deprivation information. There are 39 records in `TC` and 5 records in `Test` with missing deprivation data. Hence, these records are removed from the data prior to further analysis.

Although a rich source of environmental data has been included in this study, limited information about the individual properties is not publicly available. Characterictics such as property size, number of rooms, condition, garden and so on would typically be included in a commercial AVM and would be highly predictive of price. This is a limitation of this study. However, our goal is to demonstate the use of CPs for this application and to show proof-of-concept.

Table 1 shows summary statistics for variables used for modelling in the `TC` and `Test` data sets. The summaries for the two data sets are very similar. The only major difference is that the number of new builds is proportionally less in the `Test` set, compared to `TC`. Most variables are included in k-NN as they appear in the data. Additionally distance from the centre of London is also included (here, centre is defined as Buckingham Palace) and also geographic position given by OS Eastings and Northings. Since the association of distance

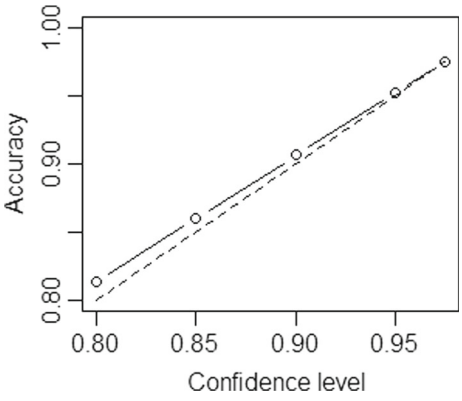

Fig. 3 Accuracy by Confidence level $1 - \epsilon$

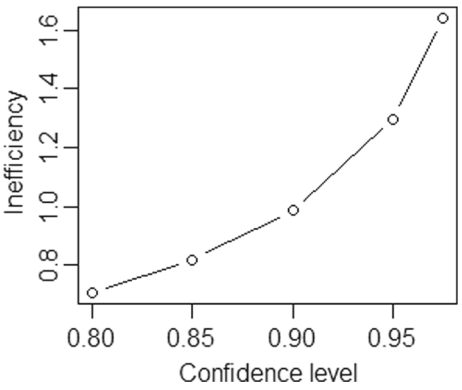

Fig. 4 Inefficiency by Confidence level $1 - \epsilon$

to the second nearest station to property price is likely to be related to distance of the first, the difference in these distances is included as a predictor variable.

4 Results

Ordinary least squares (OLS) regression is used to estimate the time trend of prices in data set TC, controlling for other variables. A linear trend $\hat{s}^*(t, t + 1)$ for all t is estimated (time measured in days), which translates into an annual price increase of 9.2 %. Non-linear effects were considered by including monthly indicator variables but these were not significant. This matches the price trends observed in Fig. 1 that shows approximately linear increase over 2013.

Lower and upper bounds on s^* over the period between training and test were constructed by finding the minimum and maximum 3-month price change over TC, then multiplying the maximum by a factor (1.52) representing the maximum 3-month price change over the period 2009 to 2013 , relative to the maximum in just 2013, using the Acadata time series. This ensures that a broad range of plausible price changes are accounted for and gives $l = -0.000536$ and $u = 0.0736$; ie price change between 0 % to 7.63 % over the 3-month period following training data.

Using grid search, following Section 2.6, hyperparameter values, $k = 10$, $\lambda = 4.65$ and $r = 10^{-0.5}$ were found. Performance for various values of (log) r are shown in Fig. 2. This result shows that including the standard deviation in the denominator for the NCM, (7), is important, although a non-zero value of constant r is required to dampen the effect of the standard deviation.

Contrasting k-NN with OLS regression for point estimates of log price, they give RMSE on the test set of 0.316 and 0.394, respectively, demonstrating that k-NN is the better algorithm for this valuation problem.

ICP is run and results are shown in Figs. 3 and 4. Figure 3 demonstrates that the ICP is conservatively valid with accuracy above the exact calibration line (shown as the dashed line) for all confidence levels. Figure 4 shows inefficiency increasing with increased confidence level, which is what would be expected since the cost of improved predictive accuracy is wider region size. As confidence level approaches 1, inefficiency rises sharply, towards infinity.

Focussing on the 90 % confidence level, test results give 0.906 accuracy and 0.985 inefficiency. If we do *not* allow for uncertainty in the change in log price estimate, as

Table 2 Model of region size (predictive inefficiency)

Predictor variable	*Coefficient estimate*	*Standard error*	*P-value*
Intercept	−0.672	0.0677	<0.0001
Log(price)	0.0975	0.00338	<0.0001
Property Type: Flat	0.0264	0.0119	0.0272
Property Type: Semi-detached	−0.0210	0.00451	<0.0001
Property Type: Terraced	−0.0138	0.00466	0.0031
Property Type: Detached *	0		
Leasehold	0.0869	0.0111	<0.0001
New build	−0.0654	0.00642	<0.0001
log(distance from centre)	0.0154	0.00319	<0.0001
log(distance from 1st station)	0.0101	0.00175	<0.0001
log(distance from 2nd station - distance from 1st station + 1)	0.00450	0.000918	<0.0001
Station: Both railway and tube	−0.00140	0.00478	0.769
Station: Tube	0.0147	0.00521	0.0047
Station: Railway *	0		
Income score	0.172	0.1000	0.0858
Employment score	0.577	0.0929	<0.0001
HD score	−0.0397	0.00362	<0.0001
Crime score	0.00437	0.00259	0.0912
LIV score	1.46	0.375	<0.0001
Environment: Indoors score	−0.973	0.250	<0.0001
Environment: Outdoors score	−0.486	0.125	0.0001
GB score	0.000569	9.16×10^{-5}	<0.0001
WB score	−0.000125	1.21×10^{-4}	0.303
Education: Child score	0.000647	0.000158	<0.0001
Education: Skills score	−0.00217	0.000191	<0.0001
IDACI score	-0.138	0.0361	0.0001
IDAOPI score	0.0311	0.0295	0.293
Young population score	−0.0267	0.00693	<0.0001
Old population score	0.0346	0.00373	<0.0001

* excluded category

given in (14), then the result is 0.888 accuracy and 0.910 inefficiency. The accuracy is well below the confidence level, fails the binomial test ($p < 0.0001$) and so is not reliable. This demonstrates the need to use some method to deal with uncertainty in price change. On the other hand, using the union of regions does not result in a large increase in inefficiency. The distribution of region size as a fraction of the true log price has the following characteristics: minimum=0.0425, 1st quartile=0.0663, median=0.0747, mean=0.0781, 3rd quartile=0.0865, maximum=0.219. This demonstrates that the distribution of region sizes (inefficiency) has a long tail, but most predictions have a reasonable size on the log scale;

Table 3 Example of a relatively efficient region prediction from the test data.

True property price:	337500 GBP and log(price)=12.73
Region prediction:	[190921, 347183] and log scale=[12.16, 12.76]
Main characteristics:	North London (NW10) freehold terrace, not new build with both tube and overground railway nearby. Has relatively high deprivation scores, especially for income.
Main characteristics of 10 nearest neighbours:	All North London (NW10 and NW2) properties: 3 terraces, 6 flats and 1 semi-detached, no new builds, serviced by tube (and two also by rail) . All with high or moderate deprivation scores, especially for income. Property price range from 204763 to 670957 GBP (log prices 12.23 to 13.42) after adjustment.

eg the median is less than 8 % of the true log price. The distribution of region size on the property price scale (ie by taking the exponential), as a fraction of true price, has the following characteristics: minimum=0.0667, 1st quartile=0.780, median=0.945, mean=1.10, 3rd quartile=1.22, maximum=22.7. This demonstrates that most predictions are not particularly practical; eg the median value suggests a region size which is almost as large as the price itself. However, this is not surprising since this study only uses environmental data about properties and not the individual property characteristics.

Table 2 shows linear regression coefficients for modelling region sizes in the test set on the predictor variables. This suggests many of the variables are drivers of inefficiency. In particular, price itself has a positive association with region size and new built properties demonstrate more price stability. Several of the environmental variables have an impact on inefficiency; eg higher income and employment deprivation is associated with predictive inefficiency. These results could help to refine the model for future development or identify particular types of property or localities that are difficult to value.

Table 3 shows an example of a region prediction taken from the test data. The diversity of the 10 nearest neighbours suggests that k-NN has selected geographical location and deprivation levels, especially for income, as the main predictive variables, for this case. The wide range of prices amongst the chosen 10 nearest neighbours is the explanation for the broad region prediction on the price scale.

5 Conclusion

In this paper a CP has been used to provide reliable region predictions for AVMs. In particular, the problem of time-dependency of property prices has been handled by assuming a systematic time effect which is estimated for training data. Uncertainty in this estimate is accounted for by outputting a conservative, broad, region prediction. This study shows that reliable region predictions can be generated, predicting property prices forward, using CP with weighted k-NN. Most efficient regions are predicted when a specific form of NCM is used which takes account of the standard deviation in the nearest neighbour's prices. The inefficiency of region predictions is analysed and, in particular, a linear regression is used to determine segments of properties for which the AVM produces more or less efficient regions. This information could be used in future AVM developement to refine the model.

This study is a proof-of-concept and the CP could readily be applied to commercial AVMs based on more detailed property data. We would expect this to perform well in terms of reliability and predictive efficiency. Future academic work in this area could focus on

building more efficient models by careful segmentation of the limited property data available, exploration of better use of the training data (eg using longer training periods with adaptive forgetting [1]), refining estimation of the property price trend, using expert knowledge of the housing market or considering property segments (eg leasehold separately to freehold, or segmenting by geographical location).

Acknowledgments Thank you to Prof. Vovk, Prof. Gammerman and members of the Department of Computer Science at Royal Holloway, University of London for their valuable suggestions and comments on this work.

References

1. Anagnostopoulos, C., Tasoulis, D., Adams, N., Hand, D.: Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. Stat. Anal. Data Mining **5**(2), 139–166 (2012)
2. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching. J. ACM **45**, 891–923 (1998)
3. Balasubramanian, V., Ho, S.S., Vovk, V. (eds.): Conformal prediction for reliable machine learning: Theory, Adaptations and applications. Morgan Kaufmann (2014)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inform. Theory **IT–13**(1), 21–27 (1967)
5. Department for Communities and Local Government: The English indices of deprivation 2010. Neighbourhoods Statistical Release (2011)
6. Downie, M.L., Robson, G.: Automated valuation models: an international perspective, London (2008)
7. European Mortgage Federation and European AVM Alliance: EMF/EAA joint paper on the use of automated valuation models in Europe (2016)
8. Harrison, D., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manag. **5**(1), 81–102 (1978)
9. Henley, W.E., Hand, D.J.: A k-nearest-neighbour classifier for assessing consumer credit risk. J. Royal Stat. Soc. Ser. D (Stat.) **45**(1), 77–95 (1996)
10. Macleod, J., Luk, A., Titterington, D.: A re-examination of the distance-weighted k-nearest neighbor classification rule. IEEE Trans. Syst. Man, Cybern. **17**(4), 689–696 (1987)
11. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. J. Artif. Intell. Res. **40**, 815–840 (2011)
12. Vasicek, O.A.: Distribution of loan portfolio value. RISK **15**(12), 160–162 (2002)
13. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer, US (2005)