Reliable Region Predictions for Automated Valuation Models:  
*supplementary material for Ames Housing Data*  
Anthony Bellotti

**Abstract**

This short paper supplements the original paper on this topic with further results using the Ames Housing data set. We show that more efficient predictions are possible with a stronger set of environmental and property characteristics.

1 **Introduction**

In [1], results were given applying conformal predictors (CP) to London House Price data. Here we give supplementary results using the Ames Housing data set [2]. This is a rich data set of 2930 observed house sales in Ames, Iowa, USA, with 80 variables for each property including neighbourhood and physical property details such as property size, shape, quality, building year and materials and rooms. See [2] and supplementary documents for further details regarding the data set.

2 **Methodology**

The same methodology is applied as in [1]. In particular, the outcome variable is given as log-price. Figure 1 shows the distribution of log-price. However, since there is no discernible house price change in the data set over the four-year period of sales, the mechanism described in section 2.4 in [1] is not applied. Since there is no perceivable time effect in the data, for this study we apply 10-fold cross validation (CV) for evaluation. A calibration data set is extracted from the training set in each CV iteration.

Additionally, since the Ames Housing data contains many categorical variables, the use of $k$-NN is inappropriate since it is difficult to apply a distance metric on multivariate categorical data. Instead, we use random forest (RF). This is justified since RF can be interpreted as a local adaptive nearest neighbours method [3]. RF has been implemented for regression in the context of CP [4]. They use a similar non-conformity measure (NCM) as given in Equation (7) in [1]. This NCM is therefore used in this supplementary study with the standard deviation of predicted value $\hat{s}$ given by the standard deviation in predicted value given by the trees forming the RF. Intuitively, this makes sense since if the separate trees in the RF give similar predictions (hence relatively low standard deviation), then this implies greater certainty in the prediction and hence greater non-conformity for values that deviate from these predicted values.
In this experiment, the RF was fixed to 100 trees and 10 predictor variables randomly selected for each tree build. The calibration data set was fixed to a random sample of 700 from the CV training set. Different values for these parameters could have been investigated to improve performance, but this was not the focus of this study.

3 Results

The value of $r$ in Equation (7) of [1] is varied to see which gives the lowest inefficiency measure, using a confidence level 90%. Figure 2 shows the results. They are not as clear as the results gained in the first study, probably due to the relatively smaller sample size with the Ames Housing data, but indicate that a value $r = 0$ gives good results.

Figures 3 and 4 show accuracy and inefficiency, respectively. Figures 3 shows that the region predictions are well-calibrated with accuracy approximately equal to confidence level and the difference is not statistically significant (binomial test, $\alpha = 0.01$). Inefficiency is increasing with confidence level and rapidly as it increases towards 100%. This is as we expect and follows the pattern seen in [1].

Figure 5 shows a histogram of the ratio of prediction intervals (region sizes) at price scale (ie by taking the exponent of log-price prediction) to the true price. A lower ratio indicates a more precise prediction. There are two predictions with extremely high ratios, greater than 4, and these have been removed from the histogram to make it clearer. Figure 5 demonstrates a much improved efficiency compared to the original study: the mean of the distribution here is 0.568 compared to 1.05 reported in [1].
Figure 2: Performance of CP for different values of $r$ in 10-fold CV. Note that for $r = 0$, inefficiency was 0.557.

Figure 3: Conformal predictor 10-fold CV accuracy.
Figure 4: Conformal predictor 10-fold CV inefficiency.

Figure 5: Histogram of ratio of price prediction interval size to true price.
4 Conclusion

These results demonstrate that for a distinct data set from a different geographical location compared to [1], CP can be used to generate reliable predictions of property price. Additionally, this is done using the Random Forest algorithm. This study also achieves tighter region predictions, greater efficiency, due to the greater depth of data available for the individual properties. In this study, not a great deal of time was spent on optimizing the underlying pricing algorithm. If this were to be done, we would expect to achieve even better predictive efficiency whilst retaining reliability. Overall, this study supplements the original study by showing that CP is a potentially practical solution to providing reliable AVMs.

References


