

Improving Prediction Models for Mass Assessment: A Data Stream Approach

Donghui Shi

Department of Computer Engineering
School of Electronics and Information Engineering
Anhui Jianzhu University
Hefei, China 230601
sdonghui@gmail.com

Jozef Zurada

Department of Computer
Information Systems, College of Business
University of Louisville
Louisville, KY 40292
jozef.zurada@louisville.edu

Jian Guan

Department of Computer Information Systems
College of Business, University of Louisville
Louisville, KY 40292
jeff.guan@louisville.edu

Alan S. Levitan

School of Accountancy, College of Business
University of Louisville
Louisville, KY 40292
alan.levitan@louisville.edu

Abstract

Mass appraisal is the process of valuing a large collection of properties within a city/municipality usually for tax purposes. The common methodology for mass appraisal is based on multiple regression though this methodology has been found to be deficient. Data mining methods have been proposed and tested as an alternative but the results are very mixed. This study introduces a new approach to building prediction models for assessing residential property values by treating past sales transactions as a data stream. The study used 110,525 sales transaction records from a municipality in the Midwest of the US. Our results show that a data stream based approach outperforms the traditional regression approach, thus showing its potential in improving the performance of prediction models for mass assessment.

1. Introduction

In the United States local governments provide a variety of services to its citizens. These services may include law enforcement, fire protection, public schools, public transportation, utilities, streets, sanitation, and many others taken for granted by the typical citizen. A major source of the revenue that pays for these services comes from property taxes on the local real estate property base. The property tax is an ad valorem tax, meaning "according to value." Therefore, the local government must have as accurate an estimate of the property value as possible to ensure fairness in property taxes. The term mass appraisal refers to the process of valuing a large collection of properties within a certain municipality. The common methodology for mass

appraisal is based on multiple regression though this methodology has been found to be deficient, suffering from problems such as nonlinearity, multicollinearity, and heteroscedasticity [1-3]. Data mining methods have been proposed and tested as an alternative but the results are very mixed [4-6].

Most prediction models are based on learning algorithms that use the entire training set in a batch mode. Such learning algorithms assume that the training data are independent, identically distributed, and are from a stationary distribution. For example, a popular and competitive regression model is a regression tree called M5 [7]. M5 builds multivariate trees using linear models at the leaves. Such a tree learns by recursively splitting the training space to select attributes that maximize the reduction of variance in the dependent variable. Though effective, these regression models are inadequate for sequentially generated data from time evolving distributions.

Data stream methods have different features than the traditional models [8]: (1) process a sample at a time; (2) use a limited amount of memory; (3) process samples at a limited time span; (4) predict a sample at any time. Up to now, many classification methods have been proposed for incremental learning using data streams. Fewer regression methods for data stream are found in the literature.

Data streaming addresses a key challenge in regression and machine learning based methods, concept drift in time-varying data streams. Existing techniques such as artificial neural network and multiple regression cannot cope with concept drifts, while data stream learning algorithms can solve those problems [9]. For example, Shaker and Ullermeier [10] proposed an instance-based learning algorithm for data streams (IBLSTREAMS) which can be used as

classification and regression models. It uses the nearest-neighborhood estimation principle for predicting the output variable.

The obstacle to high prediction accuracy for mass appraisal is the heterogeneous nature of real estate data [11, 12]. Clustering algorithms are usually used to segment the house submarkets to improve the prediction accuracy [13-15].

This study applies a new approach to mass assessment through incremental learning using data stream methodology. The results in the study point to better predictive performance of the data stream modeling approach over that of the traditional linear regression analysis commonly used in mass assessment. The study also explores the characteristics using data stream methods for the clustered data sets.

2. Data

The data in this study consist of 126,331 sales records from the Property Tax Assessment Office database in a Midwest city in the U.S. Table 1 lists the attributes in a typical sales record. 15,806 records were excluded from the data set as they contained missing and incomplete data values. The final data set contains features of 110,525 properties and 15 variables commonly used for tax assessment. It is noted that variables such as Land size, Year built, Square footage in the basement, Square footage on the floors, Garage size (number of cars), Number of baths, Year sold are numeric fields and the remaining variables are measured on the nominal scale. Table 1 presents a sample of sales transactions, whereas Table 1 shows the descriptive statistics of the entire data set.

Table 1. Sample sales transaction.

Attribute Name	Sample Record	Explanation
Sale price [\$](Dependent variable)	390000	Actual sale price
Year Built	1968	Year in which the property was built
Year Sold	2005	Year in which the property was sold
Square footage in the basement [Feet]	900	Square feet in basement
Square footage on the floors [Feet]	2931	Total square feet above basement
Land size	0.2135	Total lot size in acres
Fireplace	1	0=no fireplace; 1=fireplace is present
Garage size (number of cars)	2	Two-car garage (range 0-2)
Number of baths	4	0=substandard bath; 1=1 bath; 2=1 ½ baths; 3=2 baths; 4=2 ½ baths, etc. up to 6=more than 3 baths
Presence of central air	1	0=no central air; 1=central air is present
Lot type	1	1=up to one-fourth acre; 2=one-fourth to one-half acre; 3=one-half to 1 acre; 4=over 1 acre
Construction type	3	1=1 story; 2=1 ½ story; 3=2 story; 4=2 ½ story; 5=split-level; 6=bi-level; 7=condominium
Wall type	2	1=frame; 2=brick; 3=other
Basement type	1	0=none; 1=partial; 2=full
Basement code	1	0=none; 1=standard; 2=half standard; 3=walk-out
Garage type	3	0=none; 1=carport; 2=detached; 3=attached; 4=garage in basement; 5=built-in garage

Table 2. Descriptive statistics of the entire data set.

Variable	Mean	Std. Dev	Min	Max	Median
Land size	0.31	0.73	0	71.32	0.21
Year built	1964.08	28.35	1790	2016	1963
Square footage in the basement [Feet]	216.30	414.27	0	6100	0
Square footage on the floors [Feet]	1584.42	683.69	0	13575	1390
Garage size (number of cars)	1.19	0.86	0	2	1
Number of baths	2.65	1.51	0	6	3
Year sold	2008.28	6.01	1991	2016	2009
Sale price [\$]	148723	97322	500	3600000	127000

3. Methods

This section describes the prequential evaluation scheme and three data stream regression methods: Adaptive Model Rules (AMR), Perceptron Learning (PL), and RandomRules (RR) used in this study.

3.1. Prequential regression measurement

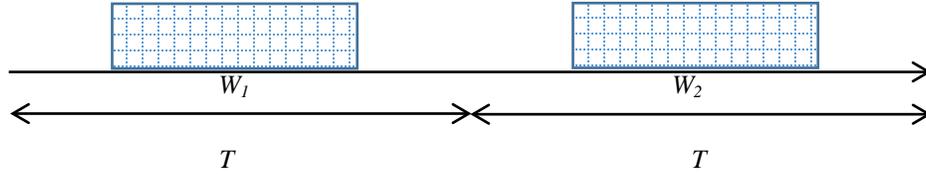


Figure 1. Sliding window W1 and W2 for regression performance evaluator.

3.2. Adaptive model rules (AMR)

Adaptive Model Rules (AMR) algorithm developed by Almeida etc. [16, 17] is a popular incremental data stream regression algorithm for rules-based learning. It uses a one-pass algorithm to build regression rule sets from a stream of input data, such as sales transactions. AMR can add and remove the rules as the data stream evolves. The form of the rule is the following [18]:

$$C \rightarrow M$$

In the above rule C represents the antecedent which is a conjunction of literals and M represents a model that can predict value a . The literal is a condition such as $A = a$, or $A \leq v$ or $A \geq v$, where A is a discrete attribute and a is one of its values, and A can also be continuous and v is a numerical value. M is a regression model. The AMR algorithm has three types of regression models: (1) the mean values of the target attribute; (2) a linear combination of the attributes; and (3) a choice between (1) and (2), resulting in a regression model with a lower mean absolute error according to the recent instances.

AMR has some different features from decision trees. For example, a decision tree model includes a set of exclusive and complete rules, whereas AMR uses a set of rules that are neither exclusive nor complete. The rules need not cover all instances and an instance may be covered by a set of rules. AMR supports a set of ordered or unordered rules. If the rules are ordered rules, the prediction result of an instance is that of the first rule. If the rules are unordered, all rules that cover an instance are used and the algorithm averages their predicting results. A critical feature of AMR is that it creates new rules, extends existing rules, and removes useless rules.

We used the prequential or interleaved test-then-train evolution evaluation scheme for the models. That is, a regression model on a data stream is evaluated by testing then training with each example in sequence. It is based on a sliding window, which measures the accuracy on the current sliding window of recent instances. Figure 1 shows the sliding windows W_1 and W_2 for the window-based regression performance. T is the sampling period and $f=1/T$ is the sampling frequency. The sliding window sizes are set to 50, 100 and 1000 in the study.

3.3. Perceptron learning

The Perceptron, proposed by Rosenblatt in 1957, is a linear classifier and one of the first methods for online learning [18]. It is a low computational cost algorithm. It can be taken as a regression method. Given a data stream of pairs (X, y_i) , where X is a case and y_i is its numeric output value. The target of the perceptron learning (PL) algorithm is to reduce the mean squared error (MSE) on these cases. To complete this task, the strategy is to move each weight in the weight vector W in the direction of the descending error gradient. The update rule is as follows:

$$W = W + \eta \sum_i (y_i - h_W(X))X;$$

where η is the learning rate. The function $h_W(X)$ is used to predict the output value, $h_W(X) = \sigma(W^T X)$; where $\sigma(x) = 1/(1 + e^{-x})$ is a soft-threshold function whose range is $[0,1]$. Weights can be updated at every single case, or taken in mini-batches for a single update step per mini-batch. The rule permits the fine-tuning of the trade-off between update time and adaptiveness.

3.4. RandomRules

RandomRules (RR) is an ensemble method using AMR as a base learner [9]. Domingos [19] proposed the bias-variance decomposition of the error of a learning algorithm that could improve its performance. Perturbing the set of examples used for training may improve regression models with a high-variance, and perturbing the set of attributes used for training may improve regression models with a low-variance. AMR

is a low-variance model, and it can be designed as the Random Forests algorithm. The ensemble method is called as Random AMR or RR.

The RR algorithm begins by initializing an ensemble F with k models f_m . The k models were constructed using the AMRules regression algorithm. When a rule R is created or expanded a subset of the data attributes with size d_0 , $1 < d_0 < d$, is randomly chosen. The next split decision for R considers only the attributes belonging to this subset. This procedure prevents the models from being correlated.

Every time a training example $(x; y)$ is available, the on-line error estimation of each model f_m is updated, and it is sent to each individual AMRules learner f_m for training. The on-line error of each f_m is estimated using a fading factor strategy. In order to perturb the training set for each model, we apply an on-line Bagging approach.

The prediction y of Random AMR is computed as a linear combination of the estimations produced by the models $f_m \in F$:

$$y = f(x) = \sum_{m=1}^k \theta_m f_m(x);$$

The weights θ_m can be computed using weighting functions. The most common approach is using a uniform weighting function, such that all the predictors have the same importance:

$$\theta_m = 1/k;$$

3.5. The framework of data stream methods

Figure 2 shows the framework for predicting real estate prices using data stream methods. In the study, the three data stream algorithms AMR, PL and RR, and traditional machine learning, Linear regression (LR) and M5 are used to predict sale prices for the whole dataset and the three clustered data sets. The framework includes two assessment steps: measurement for traditional machine learning and Prequential measurement for the data stream algorithms.

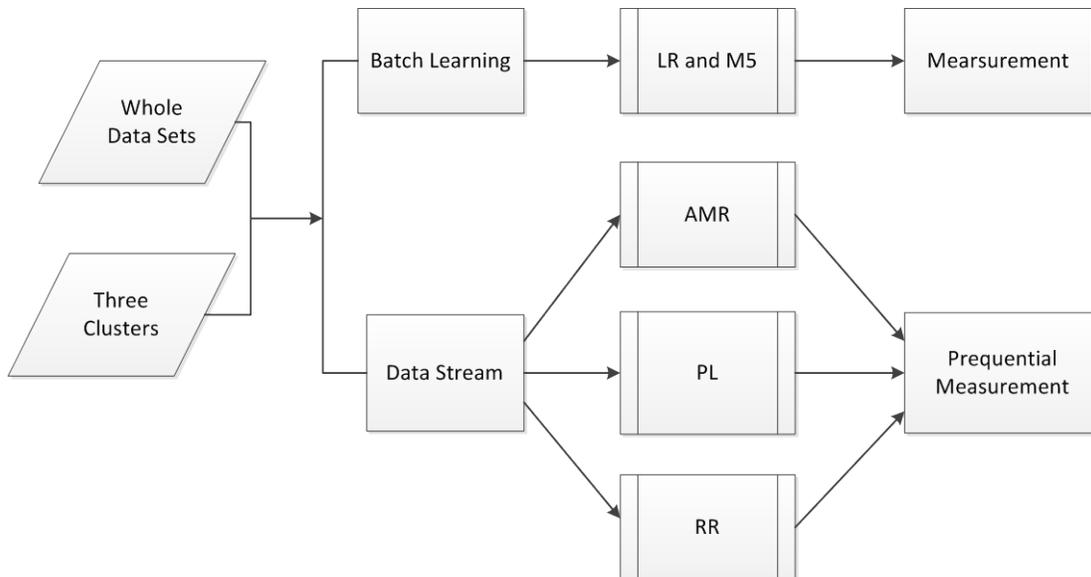


Figure 2. The framework of data stream methods for predicting real estate prices.

4. Results

Weka software was used in computer simulation (<https://www.cs.waikato.ac.nz/ml/weka/>) [20]. We created and evaluated models for the whole data set and three clusters of properties. Due to the very large data set containing 110,525 cases, only Linear regression (LR) and M5 models with 10-fold cross-validation is used because other data mining algorithms run very relatively slow. The performance of the models is evaluated by the Mean Absolute

Percentage Error (MAPE) expressed in [%], the Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE). The RMSE is the principal error measure and it is expressed in the same units as actual and predicted sale values, i.e. [\$]. The disadvantage of RMSE is that it tends to aggregate the effect of outliers. The MAE, also expressed in [\$], treats errors evenly according to their magnitude. The MAPE value measures the closeness of predicted sale prices to actual sale prices in the following 6 intervals: [0,5], (5,10], (10,15], (15,20], (20,25], and >25%. Columns 2

and 3 in Table 3 show the MAPE results for LR and M5, respectively. One can see that for LR the average MAPE is 38.1% and the MAPE results in the six different percentage ranges of MAPE. For example, the MAPE value =15.5% means that 15.5% of the properties predicted sale prices are within [0,5%] of their actual sale prices, whereas the MAPE value of 14.9% means that 14.9% of the properties predicted sale prices are within (5,10%] of their actual sale prices. Table 4 shows that the values of MAE and RMSE for LR are 29975 and 41648, respectively and those values for M5 are 26848 and 38450, respectively. Table 4 provides a basis for comparison of the LR and M5 models with the data stream models.

Table 3. MAPE results for LR, M5, AMR, PL and RR.

MAPE	LR	M5	AMR	PL	RR
≤ 5	15.5	17.5	15.8	15.4	16.4
(5,10]	14.9	16.5	15.0	14.7	15.8
(10,15]	13.3	14.2	13.4	13.2	13.7
(15,20]	11.6	11.6	11.4	11.3	11.3
(20,25]	9.1	9.1	9.3	9.3	9.1
>25	35.5	31.2	35.2	36.2	33.6
Total	100.0	100	100.0	100.0	100.0
Average	38.1	33.7	24.1	25.1	22.6

Table 4. MAE and RMSE results for LR and M5 with 10-fold cross-validation.

Methods	MAE	RMSE	Time(Sec)
LR	29975	41648	34.82
M5	26848	38450	867.25

Table 5. MAE and RMSE results for AMR, PL and RR for different window sizes.

Algorithms	AMR			PL			RR		
	50	100	1000	50	100	1000	50	100	1000
MAE	28491	29042	29583	29095	29752	30393	28695	29166	29799
RMSE	37013	38473	39943	37647	39314	41594	37332	38721	40312
Time(Sec)	18.41	18.38	18.47	1.03	1.03	1.23	173.15	172.60	171.91

The average MAPE as well as MAE and RMSE for AMR, PL, and RR are compared with LR using histograms in Figures 3, 4, and 4. Clearly, the three

We use data stream regression methods in Massive Online Analysis (MOA) platform [8] for the whole data set and the same data set clustered into three segments. The three data stream regression algorithms are AMR, PL and RR. Prequential Regression was used to measure the performance of the data stream models. Sampling frequency $f=500$ was used for the whole data set. We reported the MAPE, MAE and RMSE results for the different window sizes of 50, 100, and 1000. The average MAPE results for the three methods AMR, PL, and RR are 24.1%, 25.1%, and 22.6%, respectively (Table 3). They are less than the average of 38.1% for LR. The average MAPE value for RR =22.6% is the lowest. The MAPE values for the range [0,5%] for AMR, PL and RR are 15.8%, 15.4%, and 16.4%, respectively. MAPE for RR=16.4% is the highest, meaning that 16.4% of the properties predicted sale prices are within [0,5%] of the properties actual sale prices. This value is also higher than the MAPE value of 15.5% for LR.

Table 5 shows that the MAE values for AMR, PL and RR for the sliding window size 50 are 28491, 29095, and 28695; whereas the RMSE values for the three models are 37013, 37647, and 37332, respectively. With the increase of the window size, the values of MAE and RMSE tend to rise as well (Table 5). The results for AMR are the best. It is an obvious improvement over the results for LR. Although the MAE results for M5 are the best in all the algorithms, the averages of MAPE and RMSE results are less than those from the data steam approaches. The elapsed time for M5 is 867.25 seconds, which are far greater than those for LR, AMR, PL and RR. In the subsequent experiments, only LR was used for comparison with data stream methods.

data stream models outperform the LR model. The sliding window size for the data stream algorithms is set to 50.

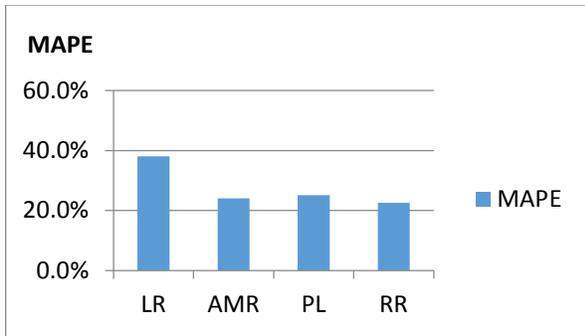


Figure 3. The average MAPE results for LR, AMR, PL and RR for window size of 50.

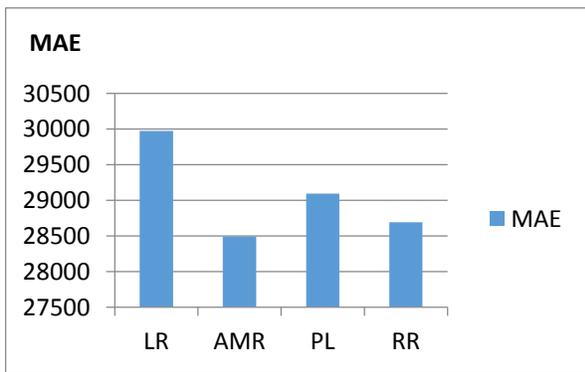


Figure 4. MAE results for LR, AMR, PL and RR for window size of 50.

Figures 6 and 7 show Prequential regression MAE results for AMR, PL, and RR for window sizes 50 and 1000, respectively. One can observe the error drifts as instances are processed. The curves of MAE for window size 50 tend to oscillate much more than the curves of MAE for window size 1000. In the two curves depicted in Figure 5, the values of MAE for the sampling points before instance 60000 are higher and fluctuate more than the sampling points after instance 60000. The MAE values for AMR are lower than those values for PL and RR in most sampling points. One can see from Figure 6 that the MAE values between 16000 and 32000 and the MAE values after 79000 are lower than those values for the other sampling points. The MAE value for PL at instance 15000 is the highest.

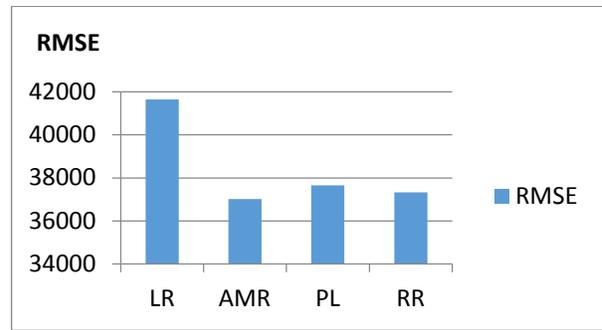


Figure 5. RMSE results for LR, AMR, PL, and RR for window size of 50.

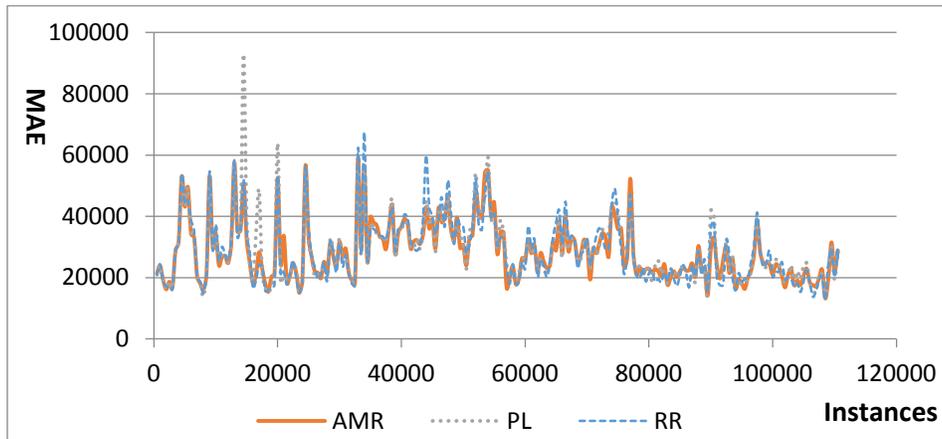


Figure 6. MAE Prequential results for AMR, PL, and RR for window size 50.

A common practice in a real estate market is to group properties into segments. Segmentation allows one to assess the property sale prices more accurately. Thus, we ran automatic K-means clustering algorithm to group the properties into three more homogeneous clusters. We normalized the attributes in the entire data set and then selected the following attributes: Land size,

Year built, Square footage in the basement, Square footage on the floors, and Number of baths as clustering fields. The three clusters were obtained; Cluster 1 with 40642 records (36.8%), Cluster 2 with 29514 (26.7%) records and Cluster 3 with 40369 records (36.5%).

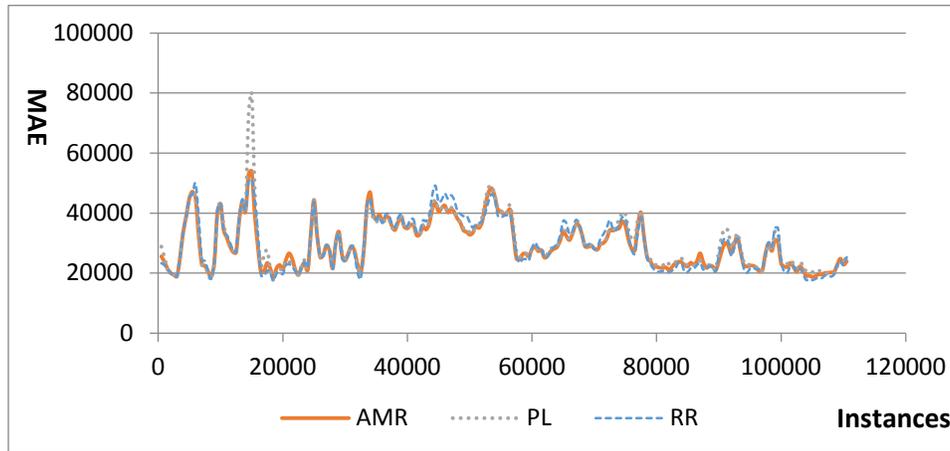


Figure 7. MAE Prequential results of AMR, PL, and RR for window size 1000.

Due to space constraints, we cannot present the full descriptive statistics for the three clusters. We only report that the mean Sale prices and the average Square footages on the floors are 139181 and 1493.6, 252329 and 2334.4, 82583.32 and 1127.6 for Clusters 1, 2, and 3, respectively. We define Cluster 1 as medium value properties, Cluster 2 as high value (more affluent) properties and Cluster 3 as low value (low-end) properties.

Table 6 shows MAE and RMSE results of LR for 10 folds for the three clusters. Table 7 shows MAPE results for LR, AMR, PL, and RR for the three clusters. The average MAE values for Clusters 1, 2, and 3 are 30499, 38366, and 24917, respectively. The average RMSE values for Clusters 1, 2, and 3 are 41354, 52685, and 34923, respectively. Table 7 shows that for LR the average MAPE values for Clusters 1, 2, and 3 are 39.7%, 16.4%, and 59.5%, respectively.

Table 6. MAE results of LR for 10 folds for the three clusters.

Methods	Cluster	MAE	RMSE
LR	Cluster 1	30499	41354
	Cluster 2	38366	52685
	Cluster 3	24917	34923

The same three data stream algorithms were also used to build and assess the models for the three clusters. Prequential Regression was used to measure the performance of the data stream models. Sampling frequency f was set to 200 for the clusters. Table 7 shows MAPE results for LR and the three data stream methods. The average MAPE values for LR, AMR, PL, and RR for Cluster 1 are 39.7, 22.6, 21.9, and 21.5, respectively. The average MAPE values for LR, AMR, PL, and RR for Cluster 2 are 16.4, 16.4, 16.9, and 16.6,

respectively. The average MAPE values for LR, AMR, PL, and RR for Cluster 3 are 59.5, 25.8, 26.2, and 25.2, respectively. Compared with LR, the MAPE results for the data stream algorithms for Clusters 1 and 3 representing medium value and low-end properties are significantly improved. There is no improvement for Cluster 2 representing high-end properties. The average MAPE value for RR for Cluster 1, that for AMR for Cluster 2 and that for RR for Cluster 3 are the best.

Table 8 shows MAE and RMSE results for the three data stream methods for three different window sizes. When the window size increases, the MAE and RMSE values increase. The average MAE values for AMR, PL and RR for Cluster 1 for window size 50 are 28648, 28943, and 28583, respectively. The average MAE values for AMR, PL and RR for Cluster 2 for window size 50 are 39409, 39572, and 41222, respectively. The average MAE values for AMR, PL and RR for Cluster 3 for window size 50 are 22132, 22425, and 22833, respectively. Compared with MAE results of LR (30499 for Cluster 1, 38366 for Cluster 2 and 24917 for Cluster 3), MAE results for data stream algorithms for Cluster 1 and Cluster 3 are improved. MAE results of data stream algorithms for Cluster 2 are comparable. RMSE results AMR for window size 50 for Cluster 1, those of PL for window size for Cluster 2 and those of AMR for window size 50 for Cluster 3 are the lowest, they are 37606, 51031 and 28442 respectively. Compared with RMSE results of LR (41354 for Cluster 1, 52685 for Cluster 2 and 34923 for Cluster 3). RMSE results of data stream algorithms for Cluster 1, Cluster 2 and Cluster 3 are all improved.

Table 7. MAPE results for LR, AMR, PL and RR for the three clusters.

Cluster Algorithms	Cluster1				Cluster2				Cluster3			
	LR	AMR	PL	RR	LR	AMR	PL	RR	LR	AMR	PL	RR
MAPE ≤ 5	16.2	16.8	16.8	17.2	21.0	21.1	20.3	20.0	11.6	11.9	12.2	12.8
(5,10]	15.6	15.9	15.4	16.3	20.1	19.6	19.3	19.0	11.2	11.7	11.3	12.0
(10,15]	13.9	14.1	14.1	14.2	16.6	16.6	16.3	16.5	10.8	10.8	10.8	11.3
(15,20]	11.8	11.4	11.5	11.3	13.1	12.7	13.0	13.1	9.8	10.2	9.9	9.8
(20,25]	9.2	9.4	9.3	9.3	9.1	9.5	9.7	9.9	8.7	9.0	8.8	8.5
>25	33.2	32.5	32.9	31.6	20.1	20.5	21.4	21.5	47.8	46.4	47.0	45.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Average	39.7	22.6	21.9	21.5	16.4	16.4	16.9	16.6	59.5	25.8	26.2	25.2

Table 8. MAE and RMSE results for AMR, PL and RR for the three clusters.

Algorithms	AMR			PL			RR			
	Window Size	50	100	1000	50	100	1000	50	100	1000
Cluster 1										
MAE		28648	28771	29021	28943	29065	29219	28583	28748	28788
RMSE		37606	37917	39350	38591	38757	40164	37387	37828	38631
Cluster 2										
MAE		39409	39148	40012	39572	39513	41074	41222	40726	41390
RMSE		51486	51031	54656	51442	51200	59380	53035	52542	54712
Cluster 3										
MAE		22132	22420	22294	22425	22633	22888	22833	22987	22857
RMSE		28442	29720	31869	28632	30040	35357	29113	30286	31568

Figure 8 shows the average MAPE results for LR, AMR, PL and RR for the three Clusters. One can see that MAPE results of the LR model in Clusters 1 and 3 are the worst compared with the three data stream algorithms. The MAPE results are comparable in Cluster 2. Figure 9 shows MAE results for LR, AMR, PL and RR for the three clusters. MAE results for the LR model in Cluster 1 and Cluster 3 are the worse than

those for the three data stream methods, and MAE result for the LR model for Cluster 2 is slightly better. Figure 10 shows RMSE results for LR, AMR, PL and RR in the three Clusters. RMSE results of the LR model for Clusters 1 and 3 are worse than those for the three data stream methods, and for Cluster 2 they are comparable.

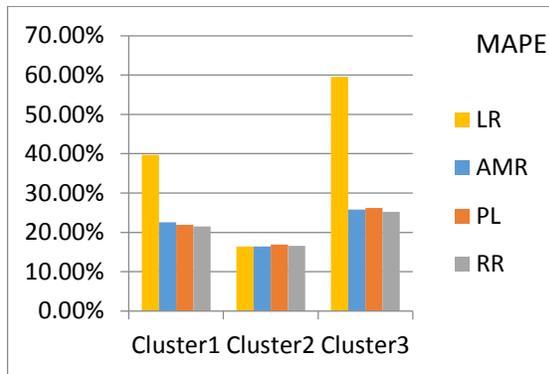


Figure 8. The average MAPE results for LR, AMR, PL and RR for the three clusters.

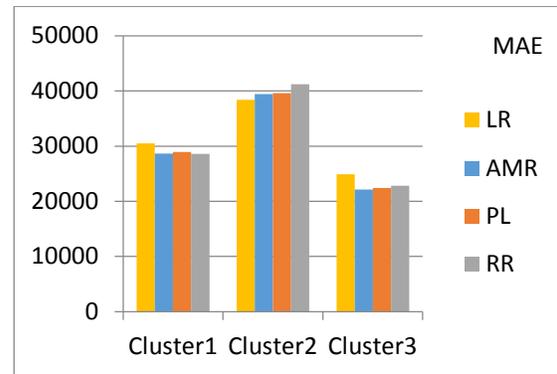


Figure 9. MAE results for LR, AMR, PL and RR for the three clusters.

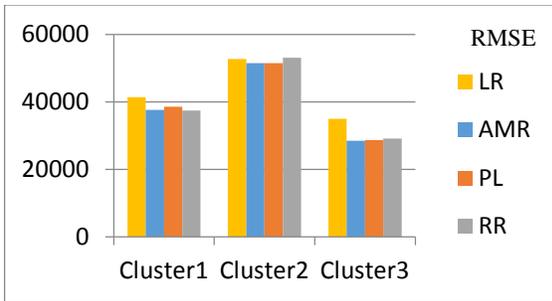


Figure 10. RMSE results for LR, AMR, PL and RR for the three clusters.

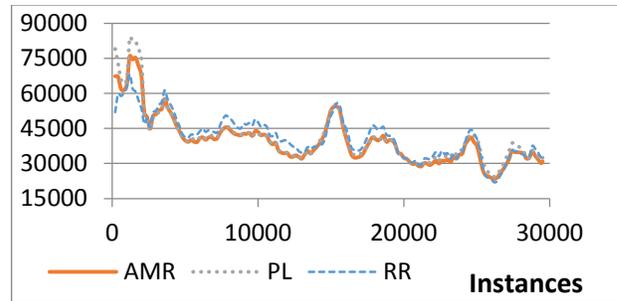


Figure 12. MAE Prequential results for AMR, PL and RR for window size 1000 for Cluster 2.

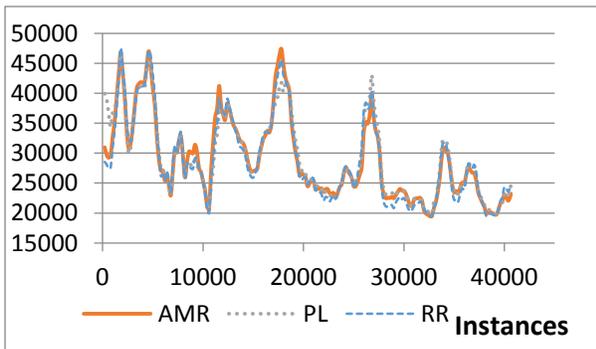


Figure 11. MAE Prequential results for AMR, PL and RR for window size 1000 and Cluster 1.

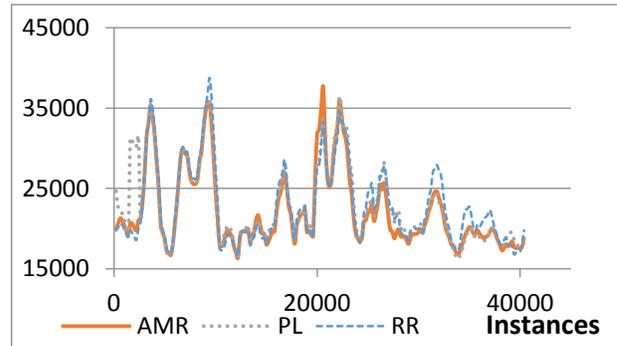


Figure 13. MAE Prequential results for AMR, PL and RR for window size 1000 for Cluster 3.

Figures 11, 12 and 13 show MAE Prequential Regression results of AMR, PL and RR for window size of 1000 for Clusters 1, 2, and 3. One can see the drift in MAE values. For Cluster 1, MAE values of RR are slightly better than those values for AMR and PL. For Clusters 2 and 3, MAE values of AMR are better than for Cluster 1. The three figures show that the curves for Cluster 2 are smoother than the curves for Clusters 1 and 3. It suggests that prediction of sale prices of more affluent properties grouped in Cluster 2 are more accurate than those sale prices for medium- and low-end properties grouped in Clusters 1 and 3. The MAE results also depict the decreasing trend as more instances are processed.

5. Conclusions

This paper investigates a novel approach to residential real estate price prediction in a mass assessment context. We used LR and the three data stream methods: AMR, PL and RR for a real estate data set including 110,525 records. The average MAPE, MAE and RMSE values of the three data stream methods are lower than those values for LR. The data stream methods have a better performance than LR. Among the three methods, AMR has the best performance. In addition, data stream methods allow

one to observe the drift of MAE and RMSE values. Different sliding window sizes 50, 100, and 1000 were used. The algorithms using a larger window size produced smoother curves of MAE, though the average MAE values are higher as the window size increases. The algorithm with a smaller sliding window size would generate more fluctuant curves of MAE.

We used K-means to divide the whole data set into three clusters representing medium value, high-end, and low-end properties. We created and evaluated the LR model and the three data stream methods for the three clusters. For medium and low value properties, the MAPE, MAE, and RMSE for LR are larger than those results for the data stream methods. Especially, MAPE values for the data stream methods are much lower than MAPE for LR. The data stream methods greatly improve prediction of house prices using clustered data sets. For high value properties, the average MAPE, MAE, RMSE values using the data stream methods are comparable to those of LR. Among the three data stream methods, in terms of the average of MAPE values, RR is the best for medium and low value properties, whereas AMR appears to be the best for high value properties. In terms of MAE and RMSE, AMR is the best for all the clusters. The Prequential results show that AMR has the best performance in

most sampling points. The data stream algorithms also enable one to observe the drifts in error values.

This novel data streaming approach produced better prediction results than the traditional multiple regression analysis commonly used in mass assessment. For processing larger datasets, the data streaming methods have obvious advantages in terms of prediction accuracy and running speed. We can observe the changing of models using data streaming. Given the

References

- [1].Antipov, E.A. and E.B. Pokryshevskaya, Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 2012. 39(2): p. 1772-1778.
- [2].Kilpatrick, J., Expert systems and mass appraisal. *Journal of Property Investment & Finance*, 2011. 29(4/5): p. 529-550.
- [3].Moore, J.W., Performance comparison of automated valuation models. *Journal of Property Tax Assessment and Administration*, 2006. 3(1): p. 43-59.
- [4].Guan, J., J. Zurada, and A.S. Levitan, An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment. *Journal of Real Estate Research*, 2008. 30(4): p. 395-420.
- [5].McCluskey, W., et al., The potential of artificial neural networks in mass appraisal: the case revisited. *Journal of Financial Management of Property and Construction*, 2012. 17(3): p. 274-292.
- [6].Zurada, J., A.S. Levitan, and J. Guan., A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, 2011. 33(3): p. 349-387.
- [7].Quinlan, J.R., *C4. 5: Programs for machine learning*. 2014: Elsevier.
- [8].Bifet, A., et al., *Massive Online Analysis Manual*, 2012.
- [9].Duarte, J. and J. Gama, Ensembles of Adaptive model rules from high-speed data streams, in *International Conference on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming MODELS and Applications*, 2014. p. 198-213.
- [10].Shaker, A. and E.H. Ullermeier, Iblstreams: a system for instance-based classification and regression on data streams. *Evolving Systems*, 2012: p. 235-249.
- [11].Goodman, A.C. and T.G. Thibodeau, The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics*, 2007. 35(2): p. 209-232.
- [12].Mark, J. and M.A. Goldberg, Multiple Regression Analysis and Mass Assessment: A Review. *The Appraisal Journal*, 1988. 56(1): p. 89.
- [13].Wilhelmsson, M., A method to derive housing sub-markets and reduce spatial dependency. *Property Management*, 2004. 22(4): p. 276-288.
- [14].Bourassa, S.C., F. Hamelin, M. Hoesli, and B.D. MacGregor, Defining housing submarkets. *Journal of Housing Economics*, 1999. 8(2): p. 160-183.
- [15].Shi, D., J. Guan, J. Zurada, and A.S. Levitan, An innovative clustering approach to market segmentation for improved price prediction. *Journal of International Technology and Information Management*, 2015. 24(1): p. 2.
- [16].Almeida, E., C.A. Ferreira, and J. Gama, Adaptive model rules from data streams, in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, 2013: Prague, Czech Republic. p. 480-492.
- [17].Almeida, E., C.A. Ferreira, and J. Gama, Learning model rules from high-speed data streams, in *The 3rd Workshop on Ubiquitous Data Mining Co-located with the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, 2013: Beijing, China. p. 10.
- [18].Bifet, A., et al., *Machine Learning for Data Streams with Practical Examples in MOA*. 2018: MIT Press.
- [19].Domingos, P. and G. Hulten, Mining high-speed data streams, in *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, ACM. p. 71-80.
- [20].Witten, I.H., F. Eibe, M. Hall and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016: Morgan Kaufmann.

importance of prediction accuracy in mass assessment the data streaming approach merits further examination. In the future, we plan to collect more data to test the performance of the methods, and do further research to understand concepts drifts in fluctuating home prices. It will also be meaningful to compare data streaming methods with traditional approaches such as time-series analysis.