28TH ANNUAL PACIFIC RIM REAL ESTATE SOCIETY CONFERENCE
ONLINE VIRTUAL 19TH JANUARY 2022

MACHINE LEARNING APPROACH TO IDENTIFY NON-LOCAL PREMIUM IN THE HOUSING MARKET

Ka Shing Cheung[1]*, Julian TszKin Chan[2], Sijie Li[3] and Chung Yim Yiu[1]

[1] Department of Property, The University of Auckland, 12 Grafton Road, Auckland 1142, New Zealand
[2] Bates White Economic Consulting, 2001 K Street NW, North Building, Suite 500, Washington, DC 20006
[3] Freddie Mac, 8200 Jones Branch Drive, McLean, VA, 22102, United States

ABSTRACT

The machine learning approach has expanded the frontier of housing studies. This paper applies a novel machine learning algorithm with the latest technique in natural language processing for classifying local versus non-local home buyers to test the information asymmetry hypothesis. While the efficient markets hypothesis postulates that the "law of one price" should hold, shreds of empirical evidence suggest that non-local property buyers usually pay a premium for comparable residential properties relative to their local counterparts. However, most previous studies rely on indirect information to classify non-local buyers and ignore non-local sellers. This study develops a machine learning algorithm to identify non-local buyers and sellers from a large-scale housing transaction dataset from Hong Kong. Using the repeat-sales method that avoids omitted variable biases, non-local buyers (sellers) are found to buy (sell) at a higher (lower) price than their local counterparts.

Keywords: unsupervised machine learning, natural language process, non-local buyers; anchoring biases, information asymmetry, repeat-sales estimates

JEL classification: D80, L85, M31, R39

* Corresponding author (william.cheung@auckland.ac.nz)

1. INTRODUCTION

Traditionally, hedonic pricing analysis considers the implicit price of property qualities, including but not limited to property attributes, neighbourhood characteristics, time, and locational effects in a competitive market (Rosen 1974). However, the analysis usually overlooks the effect of market participants, not until the emergence of behavioural economics in the 1990s (Camerer and Loewenstein 2003). This strand of studies is mainly based on theories such as information and search costs, bargaining power, and anchorage bias. Most of these studies are in relation to the effect of real estate agents and buyers' behaviours (Zumpano et al. 1996; Elder et al. 1999; Clauretie and Thistle 2007; Ihlanfeldt and Mayock 2012; Edelstein and Qian 2014), with limited studies focusing on the behaviours of sellers. Sun and Ong (2014) is one of the exceptions, but they examine the effects of transacted prices on sellers' asking prices rather than the sellers' behaviours on prices. All these studies require the availability of explicit information of buyers and sellers, which is rare. This study exploits a machine learning algorithm to extract non-local buyers' and sellers' information from raw housing transaction data to conduct an empirical study.

While the efficient markets hypothesis postulates that the "law of one price" should hold, shreds of empirical evidence suggest that non-local property buyers usually pay a premium for comparable residential properties relative to their local counterparts. Many propositions attempt to rationalise such a price premium. Two plausible theories explain such non-locals price premium for home purchases, namely: information asymmetry and anchoring biases. If the non-local premium is due to asymmetric information, the premium should be inversely related to the length of stay of the buyers/sellers before the transaction. The premium associated with non-local buyers is expected to be mirrored to a discount associated with non-local sellers but to a less extent, given that the non-local sellers must, at the least, gain some experience in their previous searches (Garmaise and Moskowitz 2004; Harding et al. 2003a, 2003b; Ihlanfeldt and Mayock 2012; Turnbull and Sirmans 1993). However, suppose the premium is due to anchoring biases; the price premium may fluctuate and, in some cases, may even switch from a premium to a discount, depending on the returns of alternative in-vestments (as the anchor) and the subject asset. Additionally, anchoring biases should be applicable to both buyers and sellers. For the anchoring effect, evidence suggests that homebuyers moving from more expensive housing markets tend to have upward biased perceptions about local housing markets and overpay on average (Ihlanfeldt and Mayock 2012; Zhou et al. 2015), but the evidence is inconclusive (Lambson et al. 2004).

Many studies have focused only on the premium paid by non-local buyers and have paid little attention to non-local sellers. Most literature in this area emphasises the information asymmetry hypothesis. Scholars use various measures to define the "distant buyer" and thereby examine the effects of local knowledge and search costs on property prices (Ihlanfeldt and Mayock 2012; Neo et al. 2008; Lambson et al. 2004; Clauretie and Thistle 2007; Zhou et al. 2015). However, the evidence is mixed. Many of these studies were criticised for the small sample sizes of non-local buyers and/or for failing to control for the property-specific and location-specific characteristics (Turnbull and Sirmans 1993; Watkins 1998). Worse still, conclusions from earlier observed effects have, in many cases, been based on inappropriate statistical comparisons with con-founding factors, such as the use of different payment methods of the buyer (and seller) groups (Wright and Yanotti 2019). The non-locals may choose to pay for cash purchases at a discount because sellers usually prefer to cash deals, as getting a mortgage could take time for non-locals. Sometimes, it is not guaranteed that the mortgage application of a non-local will go through. Nevertheless, cash purchases give the non-locals bargaining power relative to the locals who need to apply for mortgages. The impacts of such confounding factors on the market can render conflicting results.

In this study, we apply a standard search model to predict the non-local home purchase premium. The model demonstrates that the non-local premium is consistent with theories of both information asymmetry and anchoring biases. Housing transaction data usually does not contain information on whether buyers and sellers are local or non-local. Previous studies have used different geographical approaches to identify non-local buyers, such as home addresses or mobile phone number codes, which validity is debatable. This paper develops a machine learning approach to identify non-local buyers from their names. Non-local buyers are those individuals who move into a housing market from out of town and likely be at an information disadvantage compared with local buyers who already reside in the market and observe unique market conditions over a long time. In this study, we defined local homebuyers (sellers) by their implied length of residence because most people born in Hong Kong or became permanent residents of Hong Kong before July 1, 1997 would have their English names Romanised using Hong Kong unique romanisation naming system. Due to the former colony's history, Hong Kong has a different romanisation system of Chinese names. For example, Romanised surname Chan and Chen refer to the same Chinese surname. The father could have the

Romanised surname Chen if he were born in Mainland China, while the son has the Romanised surname Chan if the son was born in Hong Kong. Many Romanised surnames have indicated that a person was born in Hong Kong or became a permanent resident before July 1997. We regard this group of people as locals in this study.

Comparing non-local buyers with non-local sellers, confounding factors, including settlement method, can be controlled, as the payment issue will affect only buyers and not sellers. If an empirical test can investigate the premium (or discount) for both non-local buyers and sellers, then the alternative hypotheses can be critically differentiated. Specifically, to provide a critical test to examine whether non-local buyers and sellers pay a premium or a discount, we apply both the hedonic pricing model and the repeat-sales approach to a large dataset that includes all residential transactions in Hong Kong between January 2010 and September 2015. Instead of using geographical measures to define non-locals, we identify the non-locals, buyers and sellers, across different regions using subtle differences in the feature of the Chinese name Romanisation. This is a strength of this paper. Previous studies on this topic usually defined local buyers using their addresses or mobile phone numbers etc. Such definition did not take into account their length of stay in a city. Furthermore, only transactions before 2015 are used in our empirical tests to preclude most Mainland immigrants who become permanent residents and are not liable to the non-local transaction taxes.2 The impact on price from the second generation of Mainland immigrants who could purchase properties before 2015 without being subject to the non-locals transaction taxes can be regarded as negligible, if any.

The paper is organised as follows. Section 2 describes the Machine Learning Algorithm for classifying names of locals and non-locals. Section 3 outlines the empirical evidence used to examine the price differentials of properties purchased/sold by non-local buyers and sellers. Section 4 concludes.

2. MACHINE LEARNING FOR CLASSIFYING NAMES OF BUYERS AND SELLERS

Machine learning algorithms are relatively new in real estate research, and most of the attempts are on valuation (Pace and Hayunga 2020), i.e., using machine learning methods to identify a complex relationship between the outcome variable (housing price) and the predictors (characteristics of the house). Others have used machine learning methods to find new information for predictors. For example, Shen and Ross (2021) used a machine learning approach to quantify the value of "soft" information from unstructured real estate property descriptions. In this study, we applied machine learning methods to extract new information from transaction records, i.e., the ethnicity of property buyers and sellers. Different from Humphreys et al. (2019), in which they applied binomial and multinomial name classifiers to categorise Chinese and non-Chinese (mainly on Korean) buyers in the U.S. housing market, we used a novel natural language processing (NLP) machine learning tool based on the Gated Recurrent Units (GRU; Cho et al. 2014), a variant of the recurrent neural network (RNN), to classify the ethnicity of buyers and sellers into locals and non-locals, i.e., among Mainland and Hong Kong Chinese. The differences in their Romanised names are much more complicated and subtle to classify accurately. Indeed, the motivation of applying the GRU to perform this classification task is due to the difficulty in differentiating ethnicity based on their names, not just in Chinese but also in many other languages. This study considers such subtle differences in the romanisation feature of Chinese names of different ethnic groups to develop the machine learning algorithm directly applicable to other languages.

Every Mandarin or Cantonese syllable can be spelt with one initial followed by one final. Romanisation of Chinese characters is using the Latin alphabet to transliterate Chinese characters. These Romanised Latin alphabets in Cantonese (used in Hong Kong) and Mandarin (used in Mainland China) essentially follow a distinct pattern in their *positioning* and *sequencing*. On the one hand, in terms of positioning, when a surname starts with "ng" such as "Ngai" (倪; in China as "Wei"), it will very likely be a Romanised character of Cantonese. Nevertheless, both Romanized Cantonese and Mandarin characters can end with "ng"; thus, positioning a specific combination of alphabets will allow us to better classify the name of local Hong Kong Chinese from the non-local Mainland Chinese. On the other hand, the sequence of those Romanised alphabets also follows a pattern. Take another typical Chinese surname as an example. "Wong" and "Wang" both represent the Chinese surname "王" If the initial "W" follows suit with a final "ong", it will likely be a Romanised Cantonese surname, whereas the initial "W" ends with a final "ang" it is more likely a Mandarin surname. As such, the distributional vectors or word embeddings will capture the characteristics of the neighbours of a group of these alphabets. This approach of identifying non-local buyers and sellers provides another advantage in controlling unobservable differences due to cultural and ethnic differences.

One might argue why researchers should use machine learning rather than creating a rule-based program on the differences between Cantonese and Mandarin Romanization to automate the classification process. Indeed, a machine learning approach possesses three distinct advantages that rule-based automation cannot achieve.

First, the proposed machine learning algorithm can be applied to multiple languages. As we will further discuss, the algorithm exploits the position and sequence of alphabets; it does not necessarily require researchers to be proficient in a specific language. In genealogy, studying the subtle difference in family names can help properly evaluate genealogical evidence (Haley 1983). The advantage of using the machine learning method is that it does not require the researcher to understand the surname etymology as long as the surname etymology exists and sufficient examples are available for the algorithm to identify the etymology. For example, many English surnames are also with underlying patterns that hint at their family origins. Surnames such as Oswald, Cobbald are more likely to be British names, whereas Aames, Deloria tend to be American names.

Second, a machine learning approach can identify patterns beyond the differences between the Romanisation of Cantonese and Mandarin. A rule-based approach could classify whether a single Chinese character originated from Hong Kong, Mainland Chinese, or both. However, the rule-based approach does not consider the likelihood of a character to be a name and whether the sequence of characters could form a name. For example, the Romanisation of "Chan Mei" can be "陳薇" in Hong Kong Chinese, but in Mainland Chinese, it represents "产妹", i.e., identical Romanisation but different Chinese writing characters. A rule-based method cannot determine whether "Chan Mei" is a name of Hong Kong or Mainland Chinese. Nevertheless, the proposed machine learning algorithm can classify "Chan Mei" as a name of Hong Kong people rather than Mainland Chinese; by identifying "Chan" as a very likely surname commonly used in Hong Kong but very unlikely a surname in Mainland China.

Third, developing a machine learning method is more cost-effective. Using a rule-based approach, researchers need to specify all the Romanisation rules of Cantonese and Mandarin, which is difficult and costly, given the complexity of the Chinese languages and the naming convention. The proposed machine learning algorithm does not require researchers to understand every single difference in the Romanisation rules between Cantonese and Mandarin. Using the machine learning approach, researchers only need to provide examples of Cantonese and Mandarin names to the machine learning algorithm for the training purpose. The algorithm will learn and identify the hidden rules based on the examples provided. Instead of studying the differences between Cantonese and Mandarin Romanization, researchers only need to make sure the input examples in the training dataset are accurate[5] and to verify that the prediction is precise. The machine learning method allows researchers who may not have to thoroughly acquire the language to classify names into local and non-local.

To begin with, we draw a 10% sample from a list of Romanised names and classify them into one of the three categories: Hong Kong Chinse, Mainland Chinese, and others. This sample data will be used as an example for the algorithm to classification names. The classification algorithm starts with "tokenisation," a standard data pre-processing technique that converts the non-numeric information into a numeric format. The process tries to convert a sequence of characters into a sequence of integers. Each of the 26 alphabet, space, and other symbols is presented by an integer. Each digit is analogous to an alphabet in a Romanised Chinese name (i.e., " " to 0, "A" to 1, "B" to 2, etc.); after all, the difference does not matter to the machine. The machine learning model takes the tokenised data as input and classifies these tokenised names through three major layers in sequence inside the model. The first layer is the word embedding layer (Mikolov et al., 2013), which estimates a nominal value of the input data. In NLP, word embedding is a widely used technique that reduces data dimensionality and maps the character (or word) vectors of real numbers in a vector space. The method reduces the dimensionality of texts to that of the vector space.

The second layer consists of three other sub-layers of Recurrent Neural Networks (RNN). RNNs specialise in handling texts and other sequential data. The networks capture the autocorrelations and patterns in the sequences of characters. Salehinejad et al. (2017) present a survey of the literature and recent advancements of RNNs. This study implements a variant of RNN—Gated Recurrent Units (GRU) (Cho et al. 2014) to classify the names. GRU decides what information should be passed down to the next step to generate the hidden variables and outputs. It improves upon RNN methods by aiming to solve the vanishing gradient problem

recognised in the literature. The GRU performs better than LSTM (Long Short-Term Memory) when sequences are short because it has fewer parameters and less memory than LSTM (Cho et al., 2014).

The third layer is a Multilayer Perceptron (MLP), a standard layer in neural networks. This layer will classify the information from the GRU into four named categories and implement a Dilution (also known as dropout) procedure to reduce overfitting and improve out-of-sample performance. Panel A of Table 1 shows the structure and the hyperparameters of the machine learning model. We randomly split the data into training (80%), validation (10%), and testing (10%) samples. We use the training sample to estimate the parameters in the above model and the hyperparameters such as the number of neurons, layers of GRU and MLP, the percentage of the dropout node in the Dropout layer using the validation sample. Then we calculate the accuracy of the model using the test to report the overfitted performance. Panel B of Table 1 further shows the model prediction accuracy between the three samples close to 99%. This suggests that our model is not subject to an overfitting problem.

Table 1. Structure, hyperparameters and performance of the model.

Panel A—Hyperparameters of the Machine Learning Model			
Sequence	**Layer**	**Hyperparameter**	**Value**
1st layer	Word Embedding	Max length	50
		Number of embeddings	30
2nd layer	GRU	Number of layers	3
		Number of neurons	30
3rd layer	Activation	Functional form	tanh
4th layer	Dropout	Probability of dropout	20%
5th layer	MLP	Number of layers	2
		Number of neurons	10
		Activation	Sigmoid
Panel B—Performance of the Machine Learning Model			
Training Sample	**Validation Sample**		**Testing Sample**
99.14%	98.94%		99.00%

3. EMPIRICS: NON-LOCAL BUYER PREMIUM AND SELLER DISCOUNT

The data in this study are based on residential property transactions in Hong Kong between 2010 and 2015. This period circumvents the shocks from the Global Financial Crisis in 2008. It excludes the drastic effect of implementing a flat rate double stamp duty of 15% on all residential properties since November 2016. The number of valid observations is 93,726 for 69 months, a considerable sample size from an international perspective. Our dataset provides the sale prices of each transacted housing unit and detailed information about house locations, housing attributes, and, more importantly, the buyers' and sellers' names. Table 2 shows the schema and summary statistics of all the variables used.

Table 2. Summary statistics of variables for the hedonic price model.

Variable	Description	Mean/Count	S.D.	Min.	Max.
P	Sales Price (in HK$ Million)	4.01	3.57	0.10	236
AGE	Building Age (in years)	20.40	10.40	−3.25	58
FLR	Floor Level (in Storey)	16.95	12.15	0.00	86
GFA	Gross Floor Area (in sq ft)	655.49	242.2	134	6315
U_RATIO	Utility Ratio = Saleable to Gross Floor Area (in sq ft)	0.78	0.06	0.32	0.99
BW	Bay Window Area (in sq ft)	15.31	15.73	0.00	250
LEASE	Remaining land lease period (in years)	111.47	223.55	12	890
PRESALE	Pre-sale Dummies	238	-	0	1
MLS	Mainland Seller (1, or 0 otherwise)	5265	-	0	1
MLB	Mainland Buyer (1, or 0 otherwise)	7632	-	0	1
Direction Dummies		8			
Time Dummies		69	2010M1–2015M9		
District Dummies		59	Districts designated by EPRC		

Source: Economic Property Research Centre (EPRC).

In this study, given Hong Kong is statutorily required to use a real estate agent to engage in housing transactions, the effects of a real estate agent on buyers versus sellers' premium/discount would be eliminated.

The first focus of our empirical tests is on the asymmetric information hypothesis.

Hypothesis 1 (H1) – Asymmetric information hypothesis: *Ceteris paribus, the higher the search cost the non-local buyers incur and the higher reservation prices they have; thus, the sooner non-local buyers stop searching and pay higher prices than their low-search-cost local counterparts.*

The second focus of our empirical tests is on the anchorage bias hypothesis.

Hypothesis 2 (H2) – Anchorage bias hypothesis: *Ceteris paribus*, *non-local buyers/sellers who are more time-constrained andrely more on (i.e., anchored in) an external price distribution that is usually higher/lower than local buyers believe, thus they pay a higher/lower reservation price compared to their less time-constrained local counterparts.*

So far, there have been very few empirical studies on the housing premium/discount of non-local buyers/ sellers. To achieve this, we will apply the novel machine-learning algorithm described in the previous section to differentiate non-local from local buyers and sellers based on their ethnicity, indicating how long they have stayed in the city. More details about empirics will be discussed in the ensuing sections.

Hedonic Pricing Model Analysis and Results

To examine purchase prices of non-local relative to local buyers, many previous studies applied a standard hedonic price model (Equation (1)) and included a dummy variable indicating whether the buyer is new to the housing transaction area (Lambson et al. 2004; Ihlanfeldt and Mayock 2012; Zhou et al. 2015). Following the hedonic methodology (Equation (1)), we run a model with an additional dummy variable of non-local buyers (MLB) as Equation (2) plus non-local sellers (MLS) as Equation (3):

$$ln(P_{int}) = \alpha + \sum_{s=1}^{14} \gamma_s S_{is} + \sum_{n=1}^{59} \delta_n N_{in} + \sum_{t=1}^{69} \theta_t T_{it} + \varepsilon_{int} \dots \quad (1)$$

$$ln(P_{int}) = \alpha + \beta_1 MLB_i + \sum_{s=1}^{14} \gamma_s S_{is} + \sum_{n=1}^{59} \delta_n N_{in} + \sum_{t=1}^{69} \theta_t T_{it} + \varepsilon_{int} \dots \quad (2)$$

$$ln(P_{int}) = \alpha + \beta_1 MLB_i + \beta_2 MLS_i + \sum_{s=1}^{14} \gamma_s S_{is} + \sum_{n=1}^{59} \delta_n N_{in} + \sum_{t=1}^{69} \theta_t T_{it} + \varepsilon_{int} \dots \quad (3)$$

where P_{int} is the transaction price of residential property i in neighbourhood n sold at month t. $\gamma_s, \delta_n, \theta_t$ are the implicit prices of structural quality, neighbourhood quality, and time effects. β_1 and β_2 measure any premium and discount associated with non-local to local buyers (MLB; i.e., Mainland to Hong Kong Chinese buyers) and non-local to local sellers (MLS; i.e., Mainland to Hong Kong Chinese sellers). Moreover, we include 14 variables of structural quality S_i, including building age (AGE), floor level (FLR), floor area (GFA), bay window area (BW), utility ratio (U_RATIO), etc. In Hong Kong, as pre-sales (i.e., purchase before completion) is common in the first-hand market, we further control the age effect of pre-sales (Yiu 2009). In addition, given that all land in Hong Kong is leasehold, the remaining land lease period is thus controlled. Neighbourhood fixed effects are captured by *N*, which is defined as 59 districts dummies, a practice commonly used by the real estate industry in Hong Kong. We also include the time effects *T*, 69 monthly time dummies from January 2010 to September 2015. To avoid exact collinearity, one neighbourhood dummy, the first-period time dummy, and the direction east (D_E) are omitted as the base case.

Table 3 presents the results of these models. From the results in column (2), the Mainland buyers (MLB) are buying at a significantly higher price than the Hong Kong local buyers, while the Mainland sellers (MLS) are selling at a significantly lower price than the Hong Kong local sellers for an identical housing. The price premium for non-local buyers is about 4.9%. This estimated premium paid by non-local buyers is consistent with the related literature, ranging between the 0.3% premium estimated by Ihlanfeldt and Mayock (2012) and the 5.5% premium estimated by Lambson et al. (2004). To the best of our knowledge, this is the first study

that confirms a discount offered by non-local sellers, and such non-local sellers discount is at 1.0%, ceteris paribus.

Table 3. The results of the hedonic price model.

	Equation (1) Baseline	Equation (2)	Equation (3)
Dep. Var.	The Logarithm of Sales Prices *ln(P)*		
MLB	-	0.049	0.049
		(0.003) ***	(0.003) ***
MLS	-	-	−0.010
			(0.004) ***
|AGE| × PRESALE	0.135	−0.128	−0.124
	(0.100)	(0.100)	(0.100)
AGE × (1 − PRESALE)	−0.012	−0.012	−0.012
	(0.000) ***	(0.000) ***	(0.000) ***
FLR	0.003	0.003	0.003
	(0.000) ***	(0.000) ***	(0.000) ***
GFA	0.001	0.001	0.001
	(0.000) ***	(0.000) ***	(0.000) ***
U_RATIO	1.680	1.681	1.681
	(0.020) ***	(0.020) ***	(0.020) ***
BW	0.003	0.003	0.003
	(0.000) ***	(0.000) ***	(0.000) ***
LEASE	0.000	0.000	0.000
	(0.000) ***	(0.000) ***	(0.000) ***
Constant	−0.862	−0.865	−0.865
	(0.016) ***	(0.016) ***	(0.016) ***
Direction Fixed Effect	Included (8 Directions)		
Time Fixed Effect	Included (2010M1–2015M9)		
Neighbourhood Fixed Effect	Included (59 Subdistricts)		
Observations:	93,726	93,726	93,726
R-squared:	0.851	0.852	0.852

Notes: The dependent variable *ln*(*P*) is the logarithm of the transacted house prices in Hong Kong dollars, and *** mean that the coefficient is significant at the 1% levels. Figures in the parentheses are the standard errors.

Repeat-Sales Method as a Robustness Check

One may argue that the premium or discount could be attributable to the specification error of the hedonic model. Therefore, the repeat-sales method is applied to serve as a robustness check. Equation (4) shows the repeat-sales model, which can be considered as the subtraction of Equation (2) of the *first* transaction from the *second* transaction of the same housing unit; hence differencing out all the structural and Neighbourhood quality variables, with the time dummy variables D_{jt} redefined as follows.

$$ln(P_{jt2}/P_{jt1}) = \beta(MLB_{t2} - MLS_{t1}) + \sum_{t=1}^{T} \alpha_t D_{jt} + \varepsilon_{jt} \ \ldots \quad (4)$$

$$ln(P_{jt2}/P_{jt1}) = \beta(MLB_{t2} - MLS_{t2}) + \sum_{t=1}^{T} \alpha_t D_{jt} + \varepsilon_{jt} \ \ldots \quad (5)$$

Model (4) is a typical repeat-sales model incorporating a series of time dummy variables, D_{jt} with coefficients α_t, where *t* ranges from period 0 to T (i.e., the period covered by the sample). For a particular pair of transactions, D_{jt} takes the value −1 when *t* is the time of a previous sale of housing *j*, +1 when *t* is the time of the repeat-sale, and 0 when there are no sales of housing *j* at time *t*. It is worth noting that D_{j0} was normalised to zero. Given the buyer in the first sale must be the seller in the second sale, so MLB_{t1} in Equation (4) can be replaced by MLS_{t2}, as shown in Equation (5).

Specifically, if $(MLB_{t2} - MLS_{t2}) = +1$, it represents a non-local buyer engages with a local seller in the second sale; while if $(MLB_{t2} - MLS_{t2}) = -1$, it represents a local buyer engages with a non-local seller in

the second sale, and 0 otherwise.[11] To test for these two different effects, we fit the hedonic model that introduces separate terms for $(MLB_{t2} - MLS_{t2}) = +1$ and $(MLB_{t2} - MLS_{t2}) = -1$, such that:

$$ln(P_{jt2}/P_{jt1}) = \beta_3(MLB_{t2} - MLS_{t2})^{+} + \beta_4(MLB_{t2} - MLS_{t2})^{-} + \sum_{t=1}^{T} \alpha_t D_{jt} + \varepsilon_{jt} \ldots \quad (6)$$

where X^+ is the second sale in which a non-local buyer engages with a local seller, 0 otherwise; and X^- is the second sale with a local buyer engaging with a non-local seller.

Table 4 reports the results of Equations (4)–(6). The results reinforce the findings for the hedonic price model in Equation (3) by identifying that non-local buyers/sellers are buying/selling at a price higher/lower from/to local buyers/sellers. The signs of the coefficients are consistent with that of Equation (3). The significance of the coefficient can be improved by converting from a monthly dummy to yearly dummy specifications (from Model (6) and (7)).

Table 4. Results of the repeat-sales models of Equations (4)–(6).

Dep. Var:	$ln(\boldsymbol{P_{jt2}/P_{jt1}})$		
Variable	Model (4)	Model (5)	Model (6)
$MLB_{t2} - MLS_{t1}$	0.0283 (0.0037) ***		
$(MLB_{t2} - MLS_{t1})^{+}$		0.0067 (0.0053)	0.0025 (0.0055) ***
$(MLB_{t2} - MLS_{t1})^{-}$		−0.0499 (0.0053) ***	−0.0381 (0.0055) ***
Time Fixed Effects	Yes (2010M1–2015M9)		Yes (2010–2015)
Observations	54,794	54,794	54,794
R-squared	0.2288	0.2302	0.1741

Notes: The dependent variable $ln(\boldsymbol{P_{jt2}/P_{jt1}})$ is the logarithm of the difference in transacted house prices of two repeated sales in Hong Kong dollars, and *** mean that the coefficient is significant at 1% level. Figures in the parentheses are the standard errors.

4. CONCLUSIONS

The contribution of this paper is twofold. In terms of theoretical contribution, this is the first study to argue that other than a 4.9% non-local buyers' premium, a 1.0% non-local sellers' discount could simultaneously exist. In terms of empirical contribution, this study develops a novel machine learning algorithm with natural language processing to identify the non-local Mainland Chinese from the local Hong Kong Chinese in a residential property transaction database based on the romanisation feature of Chinese names. This approach of identifying non-local buyers and sellers provides another advantage in controlling unobservable differences due to cultural and ethnic differences. Indeed, machine learning algorithms are relatively new in real estate research. So far, most applications are merely focusing on mass appraisals or improving specific predictive analytics. To the best of our knowledge, using a machine-learning approach, along with hedonic and repeat sales methods to test anchoring and asymmetric information theories in the real estate market, is new, if not novel, in terms of methodology.

One important application of machine learning is to directly test theories that are inherently about predictability. For empirical researchers, theory and data-driven analysis have always coexisted. While many estimations are based on top-down, theory-driven, and deductive reasoning, machine learning adopted a bottom-up, data-driven, and inductive reasoning approach to let the data speak themselves more clearly than ever. In fact, these two approaches need not be in conflict (Mullainathan and Spiess 2017). This study aims to serve as a convincing demonstration as such. In analysing the data, machine learning could help manage multiple outcomes and estimate heterogeneous treatment effects. This real estate study presents a new way of using machine learning that gives its place in the econometric toolkit. It is imperative to know that machine learning provides new tools that eventually increase research scope and solve more new challenging problems. To facilitate researchers applying our developed algorithm on their projects, the source code and a user manual are uploaded to Github. We believe these findings and machine-learning applications will substantially impact academic research by opening up new research directions.

References

Akerlof, George A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84: 488–500.

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse. 1963. A Regression Model for Real Estate Price Index Construction. *Journal of the American Statistical Association* 58: 933–42.

Bucchianeri, Grace W., and Julia A. Minson. 2013. A homeowner's dilemma: Anchoring in residential real estate transactions. *Journal of Economic Behavior & Organization* 89: 76–92. https://doi.org/10.1016/j.jebo.2013.01.010.

Camerer, Colin F., and George Loewenstein. 2003. Behavioural economics: past, present, future. In *Advances in Behavioral Economics*. Princeton: Princeton University Press.

Chang, Chuang-Chang, Ching-Hsiang Chao, and Jin-Huei Yeh. 2016. The role of buy-side anchoring bias: Evidence from the real estate market. *Pacific-Basin Finance Journal* 38: 34–54.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* arXiv:1406.1078.

Clauretie, Terrence M., and Paul D. Thistle. 2007. The Effect of Time-on-Market and Location on Search Costs and Anchoring: The Case of Single-Family Properties. *The Journal of Real Estate Finance and Economics* 35: 181–96. https://doi.org/10.1007/s11146-007-9034-x.

Edelstein, Robert, and Wenlan Qian. 2014. Short-Term Buyers and Housing Market Dynamics. *The Journal of Real Estate Finance and Economics* 49: 654–89. https://doi.org/10.1007/s11146-012-9395-7.

Elder, Harold W., Leonard V. Zumpano, and Edward A. Baryla. 1999. Buyer Search Intensity and the Role of the Residential Real Estate Broker. *The Journal of Real Estate Finance and Economics* 18: 351–68. https://doi.org/10.1023/A:1007737102125

Garmaise, Mark J., and Tobias J. Moskowitz. 2004. Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies* 17: 405–37.

Haley, Alex. 1983. *Ethnic Genealogy: A Research Guide*. Santa Barbara: ABC-CLIO.

Harding, John P., John R. Knight, and C. F. Sirmans. 2003a. Estimating bargaining effects in hedonic models: Evidence from the housing market. *Real Estate Economics* 31: 601–22.

Harding, John P., John R. Knight, and C. F. Sirmans. 2003b. Estimating bargaining power in the market for existing homes. *The Review of Economics and Statistics* 85: 178–88.

Humphreys, Brad R., Adam Nowak, and Yang Zhou. 2019. Superstition and real estate prices: transaction-level evidence from the US housing market. *Applied Economics* 51: 2818–41.

Ihlanfeldt, Keith, and Tom Mayock. 2009. Price Discrimination in the Housing Market. *Journal of Urban Economics* 66: 125–40.

Ihlanfeldt, Keith, and Tom Mayock. 2012. Information, Search, and House Prices: Revisited. *Journal of Real Estate Finance and Economics* 44: 90–115.

Lambson, Val E., Grant R. McQueen, and Barrett A. Slade. 2004. Do Out-of-State Buyers Pay More for Real Estate? An Examination of Anchoring-Indexed Bias and Search Costs. *Real Estate Economics* 32: 85–126.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv* arXiv:1301.3781.

Mullainathan, Sendhil, and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31: 87–106.

Neo, Poh Har, Seow Eng Ong, and Yong Tu. 2008. Buyer exuberance and price premium. *Urban Studies* 45: 331–45.

Rosen, Sherwin. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82: 34–55.

Salehinejad, Hojjat, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. *arXiv* arXiv:1801.01078.

Shen, Lily, and Stephen Ross. 2021. Information value of property description: A machine learning approach. *Journal of Urban Economics* 121: 103299. https://doi.org/10.1016/j.jue.2020.103299.

South China Morning Post. 2017. Plunging Chinese Rental Yields Point to Property Bubbles in Major Cities. *South China Morning Post*, July 18. Available online: https://www.scmp.com/business/china-business/article/2103116/plunging-chinese-rental-yields-point-property-bubbles-major (accessed on 2 September 2021).

Sun, Hua, and Seow Eng Ong. 2014. Bidding Heterogeneity, Signaling Effect and its Implications on House Seller's Pricing Strategy. *The Journal of Real Estate Finance and Economics* 49: 568–97. https://doi.org/10.1007/s11146-013-9409-0.

Turnbull, Geoffrey K., and Casey F. Sirmans. 1993. Information, Search, and House Prices. *Regional Science and Urban Economics* 23: 545–57.

Watkins, Craig. 1998. Are New Entrants to the Residential Property Market Informationally Disadvantaged? *Journal of Property Research* 15: 57–70.

Wong, Kai On, Osmar R. Zaïane, Faith G. Davis, and Yutaka Yasui. 2020. A machine learning approach to predict ethnicity using personal name and census location in Canada. *PLoS ONE* 15: e0241239.

Wright, Danika, and María B. Yanotti. 2019. Home advantage: The preference for local residential real estate investment. *Pacific-Basin Finance Journal* 57: 101167.

Yiu, Chung Yim. 2009. Disentanglement of Age, Time, and Vintage Effects on Housing Price by Forward Contracts. *Journal of Real Estate Literature* 17: 273–91.

Zhou, Xiaorong, Karen Gibler, and Velma Zahirovic-Herbert. 2015. Asymmetric Buyer Information Influence on Price in a Homogenous Housing Market. *Urban Studies* 52: 891–905.

Zumpano, Leonard V., Harold W. Elder, and Edward A. Baryla. 1996. Buying a house and the decision to use a real estate broker. *The Journal of Real Estate Finance and Economics* 13: 169–81. https://doi.org/10.1007/BF00154054.