Longitudinal modelling of housing prices with machine learning and temporal regression

Yu Zhang, Arnab Rahman and Eric Miller
Department of Civil and Mineral Engineering, University of Toronto, Toronto, Ontario, Canada

Received 27 February 2022
Revised 25 March 2022
Accepted 30 March 2022

Abstract

Purpose – The purpose of this paper is to model housing price temporal variations and to predict price trends within the context of land use–transportation interactions using machine learning methods based on longitudinal observation of housing transaction prices.

Design/methodology/approach – This paper examines three machine learning algorithms (linear regression machine learning (ML), random forest and decision trees) applied to housing price trends from 2001 to 2016 in the Greater Toronto and Hamilton Area, with particular interests in the role of accessibility in modelling housing price. It compares the performance of the ML algorithms with traditional temporal lagged regression models.

Findings – The empirical results show that the ML algorithms achieve good accuracy (R^2 of 0.873 after cross-validation), and the temporal regression produces competitive results (R^2 of 0.876). Temporal lag effects are found to play a key role in housing price modelling, along with physical conditions and socio-economic factors. Differences in accessibility effects on housing prices differ by mode and activity type.

Originality/value – Housing prices have been extensively modelled through hedonic-based spatio-temporal regression and ML approaches. However, the mutually dependent relationship between transportation and land use makes price determination a complex process, and the comparison of different longitudinal analysis methods is rarely considered. The finding presents the longitudinal dynamics of housing market variation to housing planners.

Keywords Housing price modelling, Machine learning, Temporal lagged regression, Longitudinal analysis, North America, Housing market analysis

Paper type Case report

1. Introduction

The purpose of this paper is to model housing price temporal variations and to predict price trends within the context of land use–transportation interactions using machine learning methods based on longitudinal observation of housing transaction prices. Housing is fundamental to the security and well-being of households. Housing ownership is a major source of personal wealth, and the housing market is a major component of regional and national economies (Miles, 1994; Muellbauer and Murphy, 2008). Housing supply and prices play a major role in the determination of travel patterns and so are fundamental to transportation planning concerns as well. Planners implement policies aiming to regulate the real estate market, provide affordable housing and curb speculation and market bubbles (Barker, 2008; Jenkins *et al.*, 2006; Oxley, 2004). Analysis and modelling of housing market trends in a longitudinal manner would provide considerable insight to planners, academic researchers and practitioners.

Location or accessibility is the key factor in housing decisions (Hu and Wang, 2019; Levine, 1998; Quigley, 1985; Rodriguez and Rogers, 2014). Residential units with higher

International Journal of Housing Markets and Analysis

1753-8270
DOI 10.1108/IJHMA-02-2022-0033

accessibilities to all types of activities are valued by home buyers and builders, which increases housing prices. In turn, increased resident population increases transportation demand, which influences the location of new transport facilities (Banister, 2001; Black, 2018; Morris *et al.*, 1979). The positive feedback loop between land use and transport makes housing price trends an emergent outcome of complex interactive processes (Farooq and Miller, 2012; Rosenfield *et al.*, 2013). Therefore, this paper also gives particular interests to the spatio-temporal dynamics of housing prices and accessibility, with an objective to reemphasize the role of accessibility in forming the housing price and to provide an innovative perspective in regulating housing market and promoting sustainable urban development.

In this study, the major research questions are:

RQ1. Choice of model formulation to best capture regional housing price dynamics; and

RQ2. Investigation of accessibility dynamics effects on housing prices.

Accessibility to various activities by different modes is measured, as well as the number of neighbouring places of interest (POI), and the paper examines the impact of these factors on housing price. Ordinary least square (OLSQ) regression, temporal lagged regression and three machine learning (ML) algorithms are tested. The paper is organized as follows: Section 2 summarizes the literature on the factors and methods used in housing price modelling; Sections 3 and 4 describe the study area and the data used in the empirical study; Section 5 discusses the econometric and ML methods tested; Section 6 presents the estimation results for all models and compares the performance of the ML and econometric models; conclusions and policy implications are discussed in Section 7.

2. Literature review

Housing price has been extensively modelled using several methods, most of which follow the hedonic price framework (Rosen, 1974). Within this approach, housing can be characterized as a bundle of services that fulfil consumers' needs, and housing prices are determined by the attributes of housing, constrained by the budget of utility-maximizing consumers (Chau and Chin, 2003; Mason and Quigley, 1996; Mok *et al.*, 1995; Rosen, 1974). Housing price is, therefore, regarded as the explicit representation of the composite value of a dwelling unit's attributes (Rosen, 1974; Selim, 2009) and the value of the land upon which the housing units are situated. Lieske *et al.* (2019) model the impact of transportation infrastructure on housing price through a hedonic price model and find that urban design characteristics such as street connectivity and road density significantly influence property prices. Another stream of research analyzes housing prices from a macroscopic view, considering the influence of macroeconomic attributes such as gross disposable income, employment rate and gross domestic product within long-run demand–supply interactions (Al-Masum and Lee, 2019; Apergis and Rezitis, 2003; Hossain and Latif, 2009; Leung *et al.*, 2006; Sari *et al.*, 2007). Several findings indicate that hedonic price is still the mainstream in forming the housing price.

Three major methods are used in modelling housing prices, vector or temporal autoregression, spatial weighted regression and ML algorithms. Usman *et al.* (2020) review the modelling of housing prices in different market segments and argue that the hedonic price model exhibits aggregation bias due to the lack of coefficient spatial variation. While the spatial weighted regressions could account for the spatial dependencies in housing market modelling, Páez *et al.* (2008) found that market segmentation is more effective than several modelling techniques, including moving windows regression (MWR), geographic

weighted regression (GWR) and moving windows Kriging (MWK). Local housing market regulation policies could be evaluated through spatial regression such as mixed geographic weighted regression (MGWR) (Crespo and Grêt-Regamey, 2013). Some recent papers use the spatial regression to model the housing price (Soltani *et al.*, 2021; Zhang *et al.*, 2021) and the impact of transportation on property price (Lieske *et al.*, 2021). Temporal regression better captures the price variation and changes in the effect of each determinant over time. The endogeneity of land availability, interest rate, housing supply and demand over time could be modelled through vector correction model (Kenny, 1999). Jadevicius and Huston (2015) use auto-regressive integrated moving average (ARIMA) to predict the current housing prices as a function of historical prices (and other factors) and find the temporal autoregressive model to be useful to assess broad market price variations. Al-Masum and Lee (2019) apply a temporal autoregressive model to investigate the long-term relationship between housing prices and market fundamentals and find that Sydney housing prices can be explained by macroeconomic fundamentals.

ML algorithms are becoming widely used for housing price prediction. ML algorithms predict housing prices without explicit modelling equations that represent the relationship between the dependent and independent variables, but instead build the model "through experience", i.e. forecasting from the sample data, and improving the model performance through mathematical optimization (Koza *et al.*, 1996). Kauko (2010) apply neural network modelling to the housing market of Helsinki, Finland, and identifies the housing market segmentation by classification within the neural network. Park and Bae (2015) experiment with four different ML algorithms with data including housing physical features, mortgage rates and school rating in Virginia, to predict a binary dependent variable of whether the transaction price is higher than the listing price or not. Chen *et al.* (2017) use a support vector machine (SVM) to forecast housing market dynamics. The empirical results for Taipei City show that the model achieved high predictive accuracy. The support vectors are selected from the stepwise multi-regression approach in the first step, and then applied in SVM in the second step for the prediction. Phan (2018) uses the combination of stepwise linear regression and SVM in predicting Melbourne housing prices. Truong *et al.* (2020) present an empirical study of the housing prices in Beijing and compare random forest, extreme gradient boosting (XGBoost), light gradient booting machine (LightGBM), hybrid regression and stacked generalization modelling techniques.

3. Study area

The study area is the Greater Toronto and Hamilton Area (GTHA) (Figure 1), which is the largest metropolitan area in Canada. The GTHA has 9,183 dissemination areas (DAs). The GTHA area is 8,244 km^2, and its population was 7.36 million in 2020 (projected to be 8.6 million by 2031) (Statistics Canada *Population estimates, July 1, by census metropolitan area and census agglomeration, 2016 boundaries*, 2020).

In recent decades, the GTHA has been among the fastest-growing large metropolitan areas in the high-income world and became the principal commercial centre in Canada. The old City of Toronto provides a strong and dense central core for the region, including being the financial capital of Canada. But, like most North American cities, considerable suburbanization has occurred, starting post-Second World War, leading to the current large urban metropolitan region. The development of new cities in the GTHA brought more economic potential into this metropolitan area, which boosts the economy with more investments (capital), immigrants (labour) and land development. GTHA housing prices have increased considerably in the past 20 years. The 2016 average unit housing price is almost three times that of 2001 (Figure 2). The population inflow increased the housing demand, which raised housing prices, and induced further real estate investment as a result.

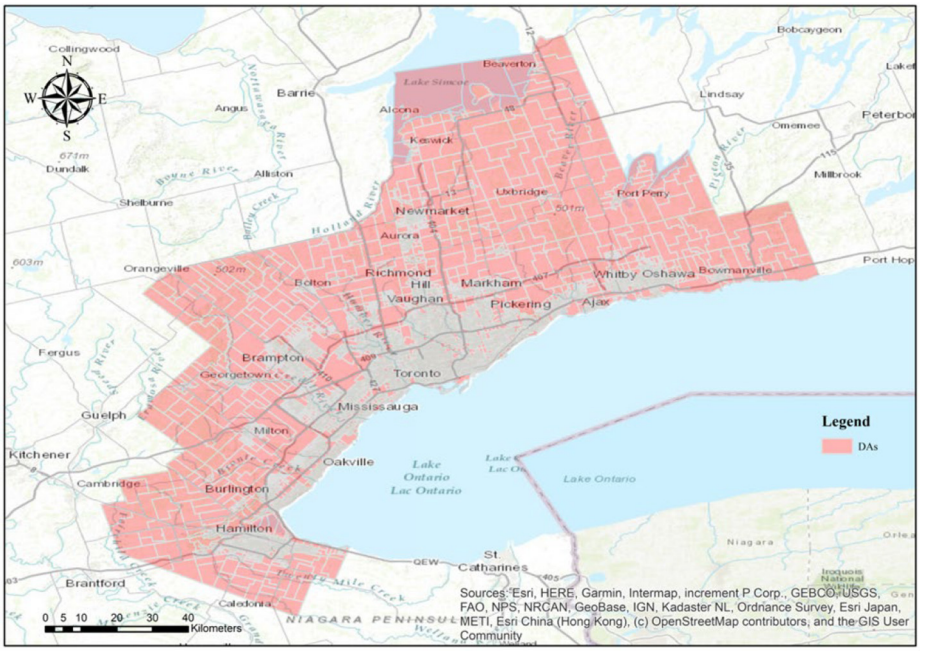

Figure 1.
The GTHA

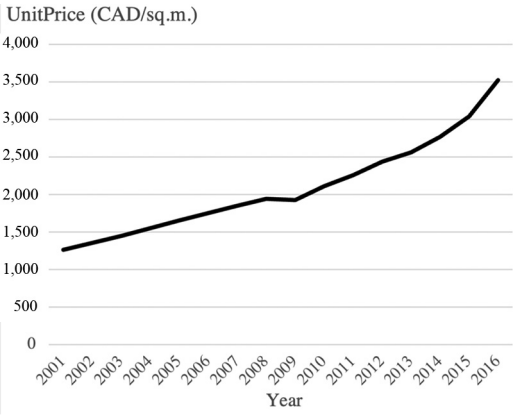

Figure 2.
Average GTHA unit housing prices, 2001 to 2016

Source: Author generated from the Teranet Housing Price Dataset

4. Data

Following the framework of Rosen (1974), the housing price models developed in this paper incorporate several factors in four aspects: built form, location, neighbourhood socio-economic characteristics and housing physical condition. To examine the effects of accessibility on housing prices, the accessibility to jobs and people, by transit and car, is measured. POI represents the places for multiple activities, including restaurants, shopping

centres, health-care facilities, schools, etc. Proximity to these activities is assumed to affect housing price positively. The variables and corresponding indicators are listed in Table 1. Housing transaction prices are represented by the average housing transaction price per sq. m. in each DA. DAs are the smallest geographic unit for which Canadian census data are routinely available. DA populations typically range from 400 to 700 people. Data provided by *Teranet Inc.* track the housing transactions from 1986 to 2016 in the GTHA, providing this study with a longitudinal record of housing price.

The spatial distribution of housing prices changes is shown in Figure 3, which maps GTHA housing unit price is mapped. In general, the suburban regions are the fastest growing, and the downtown area even displaying pockets of price decreases. From the distribution, it is seen that the real estate market "hot spots" have extended from several centres (downtown Toronto and some other subcentres) to a wider range across the suburban region. The middle-price housing units spread and cover more suburban regions over time. The spatial pattern of the housing price evolution clearly demonstrates the urban sprawl process through the past two decades in the GTHA. The suburban price increases might be explained by the improvement in accessibility throughout the suburban area. In the early 21st century, regions in the GTHA were not as connected as they currently are in terms of transportation and economy. With the development and extension of the transportation network and increasingly connected economy of different districts in the GTHA, the region became more connected as one entity, and the higher housing price locations became more continuous and extended alongside the transportation network. This trend is particularly easily seen from 2011 to 2016, as shown in the left bottom map of Figure 3, in which the major housing price increases occur in the midtown or suburban areas that are well connected to major regional highways.

Housing price follows a left-skewed distribution, as shown in Figure 4, while the peak of the density curve (median) moves rightwards, indicating increasing prices over time. Another observation is that the variance of the curve becomes larger, and the peak density drops considerably over time, which implies smaller differences in prices for all housing units. This coincides with the Figure 3 spatial trends. In comparison, even though the distribution of housing prices follows the same pattern with the income distribution, housing price experiences drastic growth while income remains almost the same.

In terms of built form, several indicators representing the land use composition of each DA are used. Variables include percentage of residential land, industrial land, park and recreational land and open area, to represent the surrounding land use. It is expected that the DAs with a higher degree of mixed land use will have lower housing prices, and higher percentage of residential, park and recreational land use, lower percentage of industrial land will increase the overall housing price in the DAs. The land use percentage indicators are calculated by *DMTI Spatial*.

To examine the effect of accessibility on housing price, several measures are used. First, standard gravity accessibilities to people and jobs are computed using equations (1) and (2) with impedance function based on average travel times by mode as follows:

$$L_i^t = \sum_{j=1}^{n} E_j^t \times e^{-\beta C_{ij}^t} \tag{1}$$

$$M_i^t = \sum_{j=1}^{n} E_j^t \times e^{-\beta F_{ij}^t} \tag{2}$$

where L_i^t and M_i^t indicate the accessibility of zone i at time t by transit and by road networks, respectively, E_j is the number of people or jobs in zone j, β is the distance

Table 1. Description of variables

Aspect	Var	Description	Source
	Year	Categorical variable indicating year 2001, 2006, 2011, 2016	
	UnitPrice	Calculated from the Teranet housing transaction point records (in CAD) aggregated to DA level. Area of condos are assumed as 80 sq. m and other housing structure type as 220 sq. m. Outliers are removed before average taken under each DA and after unit price of each data point computed	Teranet
Physical condition	Hage	Average house age (years)	Census
	Nr	Number of rooms	
	Nper	Number of persons per room	
	maj_repair	Percentage of structures requiring major repairs	
Socio-economic characteristics	high_sch	Number of high school graduates in the resident DA labour force	
	ps_certdip	Number of post-secondary certificate or diploma holders in the resident DA labour force	
	ps_deg	Number of post-secondary bachelors or above degree holders in the resident DA labour force	
	Avg_Hhinc	Average DA household income (CAD)	
	dweld	DA dwelling density (number of dwelling units per sq.km)	
	emp	DA percentage of employed labour force in the total labour force	
Built-form	pres	Percentage area of DA = residential	DMTI
	pind	Percentage area of DA = industrial	
	ppark	Percentage area of DA = park and recreational	
	popen	Percentage area of DA = open area	
Location	eat4_x	Number of eating/drinking establishments within 400 m buffer of DA centroid	
	eat8_x	Number of eating/drinking establishments within 400–800 m buffer of DA centroid	
	grc4_x	Number of grocery stores within 400 m buffer of DA centroid	
	grc8_x	Number of grocery stores within 400 m—800 m buffer of DA centroid	
	hcr4_x	Number of health care facilities within 400 m buffer of DA centroid	
	hcr8_x	Number of health care facilities within 400 m–800 m buffer of DA centroid	
	prk4_x	Number of public park/protected green area within 400 m buffer of DA centroid	
	prk8_x	Number of public park/protected green area within 400–800 m buffer of DA centroid	
	rel4_x	Number of religious facilities within 400 m buffer of DA centroid	
	rel8_x	Number of religious facilities within 400–800 m buffer of DA centroid	
	sch4_x	Number of primary/secondary schools within 400 m buffer of DA centroid	
	sch8_x	Number of primary/secondary schools within 400–800 m buffer of DA centroid	

(*continued*)

Aspect	Var	Description	Source
	smt4_x	Number of supermarkets within 400 m buffer of DA centroid	
	smt8_x	Number of supermarkets within 400–800 m buffer of DA centroid	
	chd4_x	Number of childcare facilities within 400 m buffer of DA centroid	
	chd8_x	Number of childcare facilities within 400–800 m buffer of DA centroid	
	cul4_x	Number of cultural facility (museum, art gallery, cinema, library) within 400 m buffer of DA centroid	
	cul8_x	Number of cultural facility (museum, art gallery, cinema, library) within 400–800 m buffer of DA centroid	
	ACAR	Population accessibility by car	Computed from the
	ATRAN	Population accessibility by transit	University of Toronto
	JACAR	Job accessibility by car	Transportation
	JATRAN	Job accessibility by transit	Research Institute
	RD_DEN	Road line density (km/sq. km)	EMME network
	TR_DEN	Transit line density (km/sq. km)	model
	TDEN	Trip density (all trips originating from the given zone/area of zone)	

Table 1.

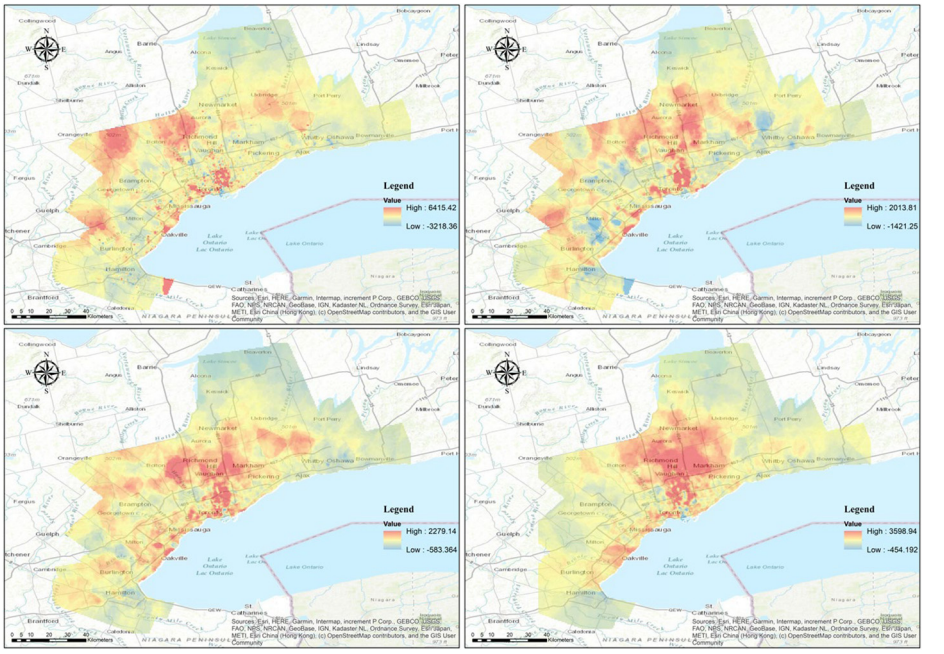

Figure 3.
Housing price variation (increments or decreases) from 2001 to 2016 in the GTHA

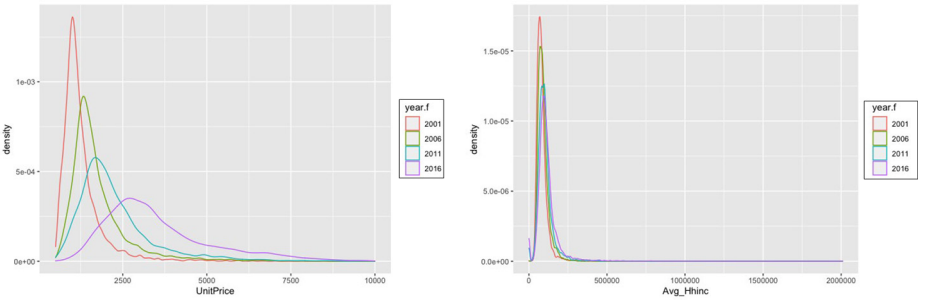

Figure 4.
Housing transaction prices and income distribution in the GTHA

decay parameter, $C_{i,j}$ and $F_{i,j}$ are the travel times between zone i and zone j by transit and road networks, respectively. The higher the value of L_i^t and M_i^t, the better accessibility could be acquired by zone i at time t. The distance decay parameter β was set to 0.05, as the classical first-order estimate of β is 1/(average travel time), and the trip-weighted travel time by auto and transit combined in the GTHA was approximately 20 min over the study period. Travel times for auto and transit were calculated from the transportation tomorrow survey (TTS), one of the largest and longest running household travel surveys that is conducted every five years, covering approximately 5% of the GTHA households at every survey wave and the travel behaviour of people greater than ten years of age. The O-D travel time matrices were calculated between traffic analysis zones (TAZs) using EMME model, assuming zero

congestion conditions as well as under peak AM conditions. The accessibility measurements are calculated by TAZs and converted to the census dissemination areas by area-weighted interpolation in ArcGIS. Second, several isochrone measures are included, representing local access to daily activities (restaurants, grocery stores, health-care facilities, etc.) using the number of POI within the buffered area (0–400 and 400–800 m) of the centroid of each DA. Finally, neighbourhood road density, transit line density and trip density, calculated by TAZs and interpolated to the DA level, are also included as potential explanatory variables.

For housing physical conditions, data from the Canadian Census are used and include variables of house age, size (number of rooms), crowdedness (number of persons per room), conditions on maintenance and repairs in the model. Newer housing units with less need for repairs are expected to be pricier.

In terms of socio-economic characteristics, educational degree, income, employment rate and dwelling density are the assumed influential factors, all of which are obtained from the Census.

Prior to modelling, the base data were prepared and cleaned. The original data set from *Teranet Inc.* was spatially joined with DA shapefiles and land use types within the GTHA. A new data set was then created by extracting the DA, total price, x-y coordinates and residential land use types for the years of 1996, 2001, 2006, 2011 and 2016. While Teranet sales data are available for every year, Census data are only available for these years, and so the time-series analysis worked with five-year time steps. The average price of each condominium unit was formulated by constricting the land use type to condominiums and getting the average price of transactions within the same x-y coordinates. Unit prices for all residential land use types were then formulated by dividing prices by areas. Housing units with total price below CAD$50,000 and above AD$5,000,000, or units with unit price less than AD$500 and above CAD$10,000 CAD were identified as outliers and removed from the data set. The unit prices were then aggregated and averaged per DA. Census and DMTI land use variables for the years of 2001, 2006, 2011 and 2016 were prepared separately based on DAs. These Census and DMTI variables were then joined to the data set based on the relevant DA and year to create a combined data set with DA, year, unit price and the land use variables. Null values were then removed from the data set to prepare it for ML algorithm implementation. A one-step time lag was then implemented into the data with the unit price of the analysis year before: e.g. for the year of 2001, the unit prices of 1996 were included in the data set.

5. Method

5.1 Ordinary least square regression and temporal lagged regression models

Before undertaking ML modelling, two linear regression models were first estimated to find the importance of each factor and to provide a comparison with the ML results. The first model, OLSQ regression, was formulated in equation (3) as follows:

$$Y_{i,t} = \alpha + \sum_{k=1}^{n} \beta_k X_{i,k} + \epsilon_i \tag{3}$$

where $Y_{i,t}$ is the average housing transaction price in DA i in year t, $X_{i,k}$ is the k^{th} explanatory variable for DA i (variables include housing physical condition, locational

characteristics, built form and socio-economic features), β_k is the parameter associated the k^{th} explanatory variable and ϵ_i is a normally distributed error term.

The second model tested in a temporal lagged regressive model, which follows the basic structure of the OLSQ, except for additional time lag variables among the explanatory variables. Stationary time series can be modelled through moving average (MA), auto-regressive process (AR), mixed auto-regressive moving average (ARMA), while non-stationary time series can be modelled using several possible transformations (logarithmic or other non-linear transformation) or ARIMA models (Ahn and Reinsel, 1990; Harrison *et al.*, 2003; Hsiao, 1982). A lagged dependent variable was added to represent the temporal correlations in the model as follows:

$$Y_{i,t} = \alpha + \gamma Y_{i,t-1} + \sum_{k=1}^{n} \beta_k X_{i,k} + \epsilon_i \tag{4}$$

where Y_{t-1} is the average housing transaction price of the DA in time period $t-1$, which should have an influence on the current housing price. Further time lags such as $t-2$ and $t-3$, are not included in the model as the effects of previous years are assumed to be captured in Y_{t-1}. In this study, since the influential factors such as land use and the socio-economic factors change slowly over time, and Census data are only available in five-year intervals, a time lag of five years is used, e.g. 2011 prices are used to model 2016 prices, and 2006 prices are used to model 2011 prices.

5.2 Machine learning algorithms

The ML-ready data set was used to develop three supervised ML models. Supervised ML uses data with known labels (observed dependent variables), as opposed to unsupervised ML, which uses data with no labels (Kotsiantis *et al.*, 2007). Linear regression, random forest regression and decision tree regression models, with and without implementing time lags, initially using default parameters. In total, 80% of the data set was used for training purposes, while the remaining 20% was used for testing purposes.

The linear regression ML algorithm fits a linear model and predicts values within a continuous range rather than categorical values, and has been used widely in price modelling (Goodfellow *et al.*, 2016; Kavitha *et al.*, 2016). The decision tree ML algorithm is used to visually represent decisions and decision-making. The decision tree model consists of a tree with an arbitrary number of nodes, and branches that are connected to each node. Decisions are made by performing tests at each branch and proceeding to a corresponding node related to the result of the performed test. This is done until the terminal node is reached (Kitts, 1997), which represents the outcome of the model. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it is also widely used in machine learning (Dietterich and Kong, 1995; Navada *et al.*, 2011; Somvanshi *et al.*, 2016). A disadvantage of a decision tree model is that the likelihood of overfitting to the data tends to increase as the size and complexity of the tree grows (Al-Akhras *et al.*, 2021). The decision tree model, however, is advantageous, in the sense that it can be used for both classification and regression problems. The random forest ML algorithm (Breiman, 2001; Breiman *et al.*, 2011) builds a "forest" from an ensemble of decision trees while monitoring the strength of individual trees, correlation between the trees, errors and variable importance (Breiman, 2001). The random forest model adds an additional layer of randomness through bootstrap aggregating (bagging) by forcing each split to consider only a randomly chosen subset of

candidate predictors, instead of the full set. Furthermore, the generalization error of the model, which is a measure of the accuracy of the model of predicting correct values from unseen data, converges to a limit as more decision trees are added to the forest. This eliminates overfitting after a certain threshold, which is an advantage of random forest models over decision tree models (Breiman, 2001).Furthermore, similar to decision trees, a significant advantage of the random forest approach is that it can be used for both classification and regression problems, which form the majority of current machine learning systems (Cutler *et al.*, 2012).

5.3 Performance evaluation

Model performance is assessed using four standard goodness-of-fit measures: R^2, mean absolute error (MAE), mean square error (MSE), root mean square deviation (RMSD) and the goodness-of-fit (R^2):

$$MAE = \frac{\sum_{i=1}^{m} |Y_i - \hat{Y_i}|}{m} \tag{5}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m} \left(Y_i - \hat{Y}_i\right)^2 \tag{6}$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^{m} \left(Y_i - \hat{Y}_i\right)^2}{m}} \tag{7}$$

$$R^2 = 1 - \sum_{i=1}^{m} \frac{\left(Y_i - \hat{Y}_i\right)^2}{\left(Y_i - \overline{Y}\right)^2} \tag{8}$$

where Y_i stands for the observed value, $\hat{Y}_i$ stands for the model predicted value and $\overline{Y}$ stands of the average of the total observed value.

6. Housing price modelling

6.1 Ordinary least squares regression and temporal lagged regression models

To identify the importance of the factors and temporal effect in determining housing price, an OLSQ regression model and a temporal lagged regression model were estimated. Parameter estimation results are shown in Table 2.

6.1.1 Physical conditions. In general, variables describing housing physical conditions, built form and accessibility all show significant influence in determining housing prices. The age of the housing units shows a slightly negative influence. Housing size shows significant positive impact on price, while crowding (number of persons per room) was not significant. Figure 5 shows that in 2001 and 2006, housing price mainly increased in the downtown area and regional subcentres, the centrality of which attracts residents even though housing units are generally smaller or more crowded in these areas. Nevertheless, in 2011 and 2016, housing prices increase in suburban areas, especially in the middle part of the GTHA alongside highways (e.g. Vaughan and Richmond Hill), as both the

Table 2. OLSQ regression and temporal lagged regression model results

	Modelling with temporal lag effect						OLSQ Model						
Variables	Estimate	SE	Pr(>|t|)	VIF	TOL	Klein	Estimate	SE	*t* value	Pr(>|t|)	VIF	TOL	Klein
(Intercept)	1.72600	0.032	54.30***				5.97000	0.025	236.50	0.00***			
log(Lag)	0.72320	0.004	163.36***	3.75	0.27	0							
year.f2006	0.17660	0.005	35.74***	3.52	0.28	0	0.36360	0.007	54.89	0.00***	3.33	0.30	0
year.f2011	0.13410	0.009	15.60***	11.32	0.09	1	0.41880	0.012	36.12	0.00***	10.86	0.09	1
year.f2016	0.41820	0.009	44.39***	13.77	0.07	1	0.85360	0.012	68.56	0.00***	12.67	0.08	1
hage	−0.00020	0.000	−1.68	2.00	0.50	0	−0.00266	0.000	−16.17	0.00***	1.97	0.51	0
nr	0.03239	0.001	22.46***	3.18	0.31	0	0.08865	0.002	45.96	0.00***	3.00	0.33	0
nper	−0.00431	0.020	−0.22	11.35	0.09	1	0.06576	0.027	2.42	0.02*	11.34	0.09	1
maj_repair	0.00013	0.000	1.66	1.54	0.65	0	0.00115	0.000	10.57	0.00***	1.53	0.65	0
high_sch	−0.00010	0.000	−3.94***	6.73	0.15	0	−0.00013	0.000	−3.62	0.00***	6.73	0.15	1
ps_certdip	0.00035	0.000	16.16***	5.24	0.19	0	0.00084	0.000	28.20	0.00***	5.14	0.19	1
ps_deg	0.00027	0.000	17.26***	5.55	0.18	0	0.00084	0.000	39.53	0.00***	5.29	0.19	1
Avg_Hhinc	0.45900	0.056	8.18***	2.03	0.49	0	3.49000	0.073	47.81	0.00***	1.81	0.55	0
dweld	−0.31330	0.048	−6.53***	2.16	0.46	0	−0.65690	0.066	−9.96	0.00***	2.16	0.46	0
emp	−0.00022	0.000	−12.10***	18.02	0.06	1	−0.00068	0.000	−27.92	0.00***	17.57	0.06	1
pres	0.00012	0.000	1.54	4.42	0.23	0	0.00022	0.000	2.04	0.04*	4.42	0.23	1
pind	−0.00014	0.000	−1.34	1.85	0.54	0	−0.00047	0.000	−3.18	0.00**	1.85	0.54	0
ppark	−0.00002	0.000	−0.18	1.83	0.55	0	0.00048	0.000	2.73	0.01**	1.83	0.55	0
popen	0.00054	0.000	5.71***	3.46	0.29	0	0.00163	0.000	12.45	0.00***	3.45	0.29	0
eat4_x	−0.00058	0.000	−2.28*	6.92	0.14	0	−0.00086	0.000	−2.45	0.01*	6.92	0.14	1
eat8_x	−0.00002	0.000	−0.20	11.40	0.09	1	0.00044	0.000	3.26	0.00**	11.39	0.09	1
grc4_x	0.00053	0.001	0.56	4.16	0.24	0	0.00048	0.001	0.37	0.71	4.16	0.24	0
grc8_x	−0.00055	0.000	−1.24	8.12	0.12	1	−0.00491	0.001	−7.98	0.00***	8.09	0.12	1
hcr4_x	0.00054	0.000	2.21*	1.90	0.53	0	0.00107	0.000	3.20	0.00**	1.90	0.53	0
hcr8_x	0.00036	0.000	3.38***	2.76	0.36	0	0.00152	0.000	10.50	0.00***	2.74	0.36	0
prk4_x	−0.00074	0.001	−0.49	2.70	0.37	0	−0.00373	0.002	−1.82	0.07	2.70	0.37	0
prk8_x	0.00083	0.001	1.17	3.50	0.29	0	0.00211	0.001	2.16	0.03*	3.49	0.29	0
rel4_x	0.00179	0.001	2.20*	2.80	0.36	0	0.00581	0.001	5.20	0.00***	2.80	0.36	0
rel8_x	−0.00053	0.000	−1.29	6.08	0.16	0	−0.00321	0.001	−5.71	0.00***	6.07	0.16	1
sch4_x	−0.00331	0.001	−4.02***	1.80	0.56	0	−0.00641	0.001	−5.65	0.00***	1.80	0.56	0
sch8_x	0.00028	0.000	0.68	2.92	0.34	0	−0.00095	0.001	−1.67	0.09	2.92	0.34	0
smt4_x	0.00211	0.001	1.72	3.10	0.32	0	0.00639	0.002	3.80	0.00***	3.10	0.32	0

(*continued*)

Table 2.

	Modelling with temporal lag effect						OLSQ Model						
Variables	Estimate	SE	Pr(>|t|)	VIF	TOL	Klein	Estimate	SE	*t* value	Pr(>|t|)	VIF	TOL	Klein
smt8_x	−0.00176	0.001	−3.29**	5.12	0.20	0	−0.00326	0.001	−4.43	0.00***	5.11	0.20	1
chd4_x	0.00682	0.002	4.33***	1.33	0.75	0	0.01723	0.002	7.95	0.00***	1.33	0.75	0
chd8_x	−0.00031	0.000	−0.80	2.98	0.34	0	0.00118	0.001	2.18	0.03*	2.98	0.34	0
ACAR	0.92600	0.069	13.44***	229.43	0.00	1	1.62200	0.095	17.12	0.00***	228.55	0.00	1
ATRAN	−0.61940	0.054	−11.38***	95.20	0.01	1	−1.37500	0.075	−18.41	0.00***	94.51	0.01	1
JACAR	−0.85570	0.083	−10.33***	278.61	0.00	1	−1.14900	0.114	−10.07	0.00***	278.48	0.00	1
JATRAN	1.05500	0.073	14.50***	134.10	0.01	1	2.33200	0.100	23.39	0.00***	132.56	0.01	1
RD_DEN	−0.05935	0.025	−2.39*	1.72	0.58	0	−0.33680	0.034	−9.86	0.00***	1.71	0.58	0
TR_DEN	0.20200	0.021	9.77***	2.77	0.36	0	0.36700	0.028	12.90	0.00***	2.76	0.36	0
TDEN	−0.16250	0.026	−6.25***	2.33	0.43	0	−0.37260	0.036	−10.43	0.00***	2.33	0.43	0

Notes: *R2*: 0.8766, adjusted *R*: 0.8765 *R*: 0.7659, adjusted *R2*: 0.7656 ***, **, *. indicate significance of 0.001, 0.01, 0.05, 0.1 respectively. 1 in Klein test indicate collinearity is detected by the test, 0 otherwise

transportation network and transit lines improved and extended. The switch indicates that more households prefer the less dense suburban area well connected to the transportation network. DAs with higher percentage of housing units that need major repair have higher housing prices, which contradicts with our expectations that maintenance decreases the housing value but is likely be due to gentrification effects in these older neighbourhoods.

6.1.2 Accessibility and housing price. The coefficients of accessibility show interesting phenomena, in that accessibility by mode has different effects by activity type. The expectation of the coefficients of accessibility indexes are positive, as higher accessibility would always increase the attractiveness of the location, and hence raise the housing price; however, the results indicate that housing units that have better access to jobs by transit, or better access to all activities by cars have higher prices. The overall accessibility by cars increases housing price, as shown in Table 2, while public transit accessibility has a negative effect, indicating that the housing units located at the regions with higher auto accessibility, but lower transit accessibility, have higher price. This phenomenon flips with respect to job accessibility, i.e. places with higher job accessibility by transit are pricier. Transit line density also has significant positive correlation with housing prices. Trip density is negatively related to housing price, as Figure 5 shows that the central region with dense trip flow like downtown area did not increase much in the housing price. Not all of the accessibilities to multiple facilities in the surrounding show significant influences, among which number of health-care facilities within 800 m, number of primary and secondary schools within 400 m and number of childcare facilities within 400 m show significant positive influence on housing price.

6.1.3 Socio-economic factors. Neighbourhood socio-economic variables show significant influence in determining housing prices. Average household income shows positive effect on housing price, as expected, and DAs with higher number of residents with a post-secondary degree also have higher housing prices (possibly an indication of neighbourhood "status"). The coefficients of employment rate are significantly negative, which might result from a collinearity problem. A Farrar–Glauber test (F-test) was conducted for the location of the multicollinearity, and the variance inflation factor (VIF) and Klein factor both show that employment rate might be non-significant due to multicollinearity. Built form (composition of land use of each DA) shows significant influence in the regression without the time lag variable, which reflects that the land use changes slowly over the years, and the time lag variable could obscure some of the land use composition effect. The percentages of residential land and open area show significant positive signs and percentage of industrial land affects the housing price negatively, as expected.

6.1.4 Temporal lag in housing price. Addition of a time lag variable (the average DA price in the previous time step) significantly improves the performance of the model, for

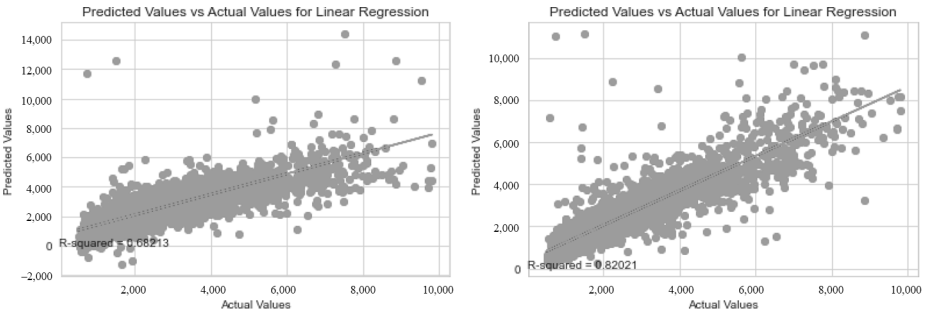

Figure 5.
Linear regression performance (predicted vs actual values)

which the R^2 increases from 0.76 to 0.87. Housing price was well modelled through the estimation based on the factors from the four aspects but becomes less predictable in more recent years. As more investors entered the housing market in the past five to ten years, investment and speculation demand increased, the rational of which is not limited to the land use, location, physical condition factors considered in the model. Speculative behaviour can be explained by the expectation that housing values will further appreciate in the future (Fox and Tulip, 2014). In addition, real estate development projects target different submarkets, including condominiums in the city centre and houses in suburban areas, which adds to the heterogeneity in residential development. The increasing complexity over time from both the demand and supply side increases the uncertainty in modelling housing prices. Modelling with a time lag largely improved the goodness of fit, indicating that the housing prices are largely following the trend of the previous year, and the time lag might result from the market response in investment demand, as investors rely heavily on the historical housing price records in their decision-making. This also indicates that housing price determination is a complex process and is driven by factors in addition to the explanatory variables used in this model, with the more implicit dynamics contained in the lag variable.

The interrelations and mutual dependence among residential construction, commercial property and transportation development through the accessibility factors further increase the uncertainty, which cannot be modelled through the OLSQ model. In the next section, several ML algorithms are tested to compare their performance with the regression models.

6.2 Machine learning

Several iterations of performance tuning were undertaken to optimize the solutions for each model. two methods were considered for tuning: grid search and random search. Grid search is an exhaustive method in which a grid of hyperparameter values is set up to test each combination of hyperparameter values, and random search is a method in which the same grid is set up to test random combinations of hyperparameter values (Worcester, 2019). Random search was chosen as the method to tune the model as it is computationally less expensive.

As shown in Table 3, random forest regression outperforms other algorithms and achieves the lowest MAE, MSE and RMSD, as well as the highest R^2. After adding the lagged variable into the training model, the random forest regression, decision trees regression and linear regression models all reach R^2 values of above 0.82, with random forest regression reaching an R^2 value of around 0.873. This demonstrates that ML models can serve as a tool for accurate housing price prediction compared to regular linear regression models.

Index	Linear regression		Random forest regression		Decision trees regression	
	a*	b	A	b	a	b
MAE	524.212	353.043	377.159	295.265	478.403	344.113
MSE	658,962.775	372,711.048	384,542.570	262,473.987	622,515.290	371,883.019
RMSD	811.765	610.501	620.115	512.322	788.996	609.822
R^2	0.682	0.820	0.815	0.873	0.708	0.824
Adjusted R^2	0.680	0.819	0.813	0.872	0.706	0.823

Note: *a, b indicates without and with time lag, respectively

Table 3. Prediction results of linear regression, random forest regression, and decision trees regression

To test for data overfitting, k-folds cross-validation was applied on the time-lagged variants of the three models with ten folds. K-folds cross-validation trains a model using K-1 of the folds as the training data and applies the resulting model on the remaining data to compute performance scores (Ojala and Garriga, 2010). Through this method, the mean ten-fold R^2, mean ten-fold MAE, mean ten-fold MSE and the mean ten-fold RMSD were computed and compared to the initial MAE, MSE, RMSD and R^2 that were generated after making predictions on the test data.

After applying k-folds cross-validation, it was found that there was a degree of overfitting in all of the models, which likely occurred due to noisiness in the data and a high number of variables. To reduce overfitting, feature selection was attempted to lower the total number of variables. Two different approaches to feature selection were considered: the filter method and the wrapper method. Filter methods are independent of the ML model used, whereas wrapper methods use the chosen ML model to choose optimal features (Maldonado and Weber, 2009). To maintain consistency in the chosen variables for each model, the filter method of computing Pearson's correlation coefficients for features and choosing the k best features was attempted. The number of features were iteratively reduced using this method to find the optimal ten-fold MAE, ten-fold MSE, ten-fold RMSD and ten-fold R^2 while minimizing overfitting for each model. Ultimately, feature selection was not found to reduce overfitting significantly or increase model performance, and as such, was not used in any of the final models (Tables 3 and 4).

6.2.1 Linear regression machine learning model. The ML model that applies linear regression improves significantly (Figure 5) after including the lagged variable, significantly reducing the RMSD and increasing the adjusted R^2. The coefficients of the linear regression model are outlined in Table 5.

6.2.2 Decision trees regression machine learning model. The decision trees regression model performs similarly to the linear regression model and performs slightly better than the linear regression model in terms of RMSD. The hyperparameters were tuned using random search, with the optimal *max_depth* being 10. The predicted values against actual values are plotted in Figure 6.

The two decision tree regression models (with and without lagged variables) give the same prediction of housing prices for several different actual values, which compromises the accuracy of this method. As the likelihood of overfitting to the data positively correlates with an increase in the size and complexity of the decision tree (Al-Akhras *et al.*, 2021), decision trees can be prone to overfitting, which can undermine model validity. As such, it is found that random forest regression is a better choice in housing pricing modelling. The generalization error of a random forest model converges to a limit as more decision trees are added to the forest, which limits overall model overfitting to a certain threshold (Breiman, 2001).

Table 4.
Mean ten-fold cross-validation results of linear regression, random forest regression and decision trees regression

Index	Linear regression		Random forest regression		Decision trees regression	
	a	b	a	b	a	b
MAE	562.089	379.132	482.876	358.116	636.677	413.130
MSE	722,895.134	396,408.382	540,855.351	338,970.365	1,088,499.222	554,525.510
RMSD	797.413	595.027	689.334	546.789	941.566	677.435
R^2	0.466	0.693	0.606	0.750	0.284	0.611

Variable	Coefficient	Coefficient with lagged variable	Variable	Coefficient	Coefficient with lagged variable
lag		0.907	prk4_x	−14.735	−7.608
hage	−4.336	−0.514	prk8_x	17.942	11.19
nr	175.555	83.694	rel4_x	20.756	9.021
nper	514.171	120.317	rel8_x	−0.87	2.571
maj_repair	3.916	1.055	sch4_x	−16.403	−9.959
high_sch	0.076	−0.214	sch8_x	1.145	4.132
ps_certdip	1.039	0.623	smt4_x	18.388	5.555
ps_deg	2.217	0.915	smt8_x	−6.291	−4.342
Avg_HHinc	0.007	0.001	chd4_x	27.625	19.312
dweld	−0.038	−0.023	chd8_x	−0.139	−3.312
emp	−1.532	−0.577	ACAR	0.001	0.001
pres	0.979	0.778	ATRAN	−0.005	−0.003
pind	0.189	0.382	JACAR	−0.001	−0.001
ppark	0.673	−0.11	JATRAN	0.012	0.006
popen	3.349	1.582	RD_DEN	−8.057	−2.982
eat4_x	−3.95	−2.288	TR_DEN	66.961	49.683
eat8_x	2.633	1.061	TDEN	−0.009	−0.004
grc4_x	−0.574	1.735	year_2001	−623.191	−324.018
grc8_x	−23.69	−7.476	year_2006	−96.477	−58.363
hcr4_x	4.858	1.893	year_2011	−211.217	−207.802
hcr8_x	4.277	1.204	year_2016	930.885	590.183

Table 5.
Coefficients for linear regression ML

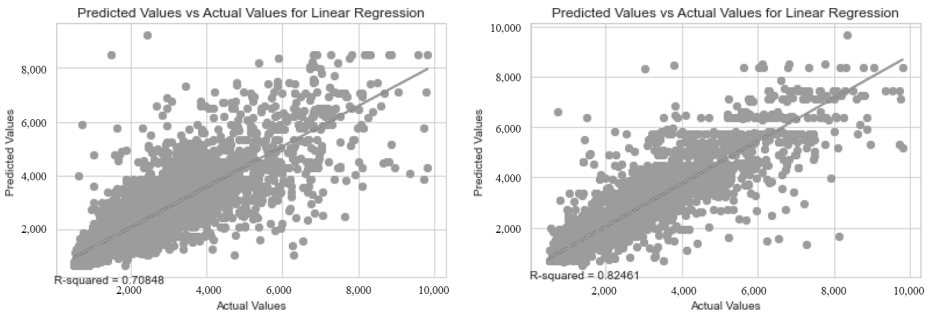

Figure 6.
Decision trees regression performance

In the rank of the feature importance, time lag is once again shown to be the most effective factor when included, whether the year is 2016, the average household income and population accessibility by car are the most effective factors when the lagged variable is not included. All the feature importance values for the decision trees regression model are outlined in Table 6.

The performance of the three ML algorithms in predicting housing prices suggests that temporal effects play an important role. Random forest regression outperforms the other two algorithms and achieves an adjusted R^2 of 0.872. Given the relatively similar performance of the two approaches, it is suggested that the temporal lagged regression approach may be preferred, as it provides greater insights into the significance of explanatory variables, which is important in supporting housing market policy analysis.

Variable	Feature importance	With lagged variable	Variable	Feature importance	With lagged variable
lag		0.878	prk4_x	0.001	0.000
hage	0.008	0.005	prk8_x	0.004	0.001
nr	0.024	0.006	rel4_x	0.001	0.001
nper	0.007	0.001	rel8_x	0.000	0.001
maj_repair	0.002	0.000	sch4_x	0.001	0.002
high_sch	0.002	0.001	sch8_x	0.002	0.001
ps_certdip	0.005	0.002	smt4_x	0.001	0.001
ps_deg	0.008	0.003	smt8_x	0.003	0.001
Avg_HHinc	0.323	0.015	chd4_x	0.001	0.000
dweld	0.014	0.002	chd8_x	0.001	0.001
emp	0.003	0.001	ACAR	0.123	0.006
pres	0.008	0.004	ATRAN	0.005	0.005
pind	0.001	0.001	JACAR	0.052	0.007
ppark	0.002	0.000	JATRAN	0.021	0.002
popen	0.002	0.001	RD_DEN	0.011	0.002
eat4_x	0.000	0.001	TR_DEN	0.005	0.001
eat8_x	0.002	0.001	TDEN	0.004	0.001
grc4_x	0.001	0.000	year_2001	0.010	0.001
grc8_x	0.002	0.001	year_2006	0.000	0.001
hcr4_x	0.002	0.001	year_2011	0.007	0.000
hcr8_x	0.003	0.002	year_2016	0.327	0.037

Table 6.
Feature importance values for decision trees regression

6.2.3 Random forest regression machine learning model. The random forest regression ML model initially showed promising results with default parameters. To further optimize performance, the hyperparameters within the model were tuned. The tuned hyperparameters were the "number of decision trees per forest" (*n_estimators),* "minimum number of samples required to split an internal node" (*min_samples_split*), "minimum number of samples required to be at a leaf node" (*min_samples_leaf*), "number of features to consider when looking for the best split" (*max_features*), "maximum depth of the tree" (*max_depth*) and "whether bootstrap samples are used when building trees" (*bootstrap*) (scikit-learn developers, 2020). The random search method was used to discover the optimal hyperparameters, which were found to be 800 for *n_estimators*, 2 for *min_samples_split*, 2 for *min_samples_leaf*, square root for *max_features*, 50 for *max_depth* and false for *bootstrap*. The results after tuning the random forest regression model are shown in Figure 7.

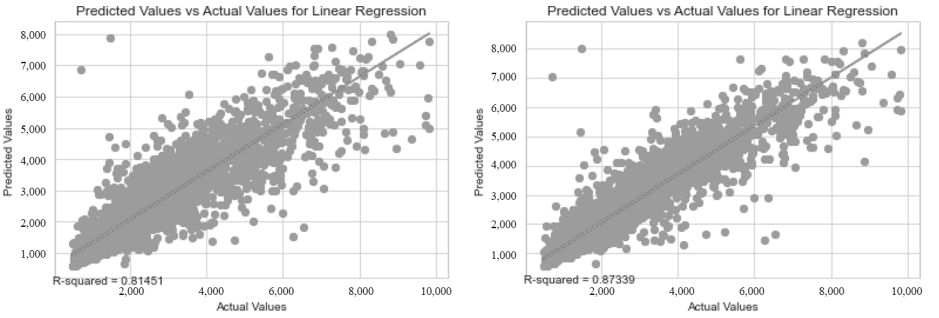

Figure 7.
Random forest regression performance

Tuning the model increases the model fit, especially after adding the time-lagged variable. In the rank of the feature importance, it is seen that the time lag is the most effective factor when included, and that the average household income, whether the year is 2016 and the number of post-secondary bachelors or above degree holders in the labour force are the most effective factors when the lagged variable is not included. All the feature importance values for the random forest regression model are listed in Table 7.

7. Discussion

Housing price has been extensively researched in terms of behavioural mechanisms, determinants and influential factors, including the effects of temporal variation in both the long and short term. The transportation and land use interaction process includes multiple dynamics over time and space, which is more of a black box rather than a simple combination of a set of factors. In consequence, housing price is difficult to predict precisely through the conventional hedonic-based framework. ML as a new modelling technique emerging over recent years uses mathematic optimization algorithms in decision-making, including application in predicting housing prices. Planners and practitioners should recognize the complex determination process of housing price when analyzing the housing market condition and apply proper modelling method. Following the hedonic housing price framework, this study compares the performance of temporal lagged models and three ML algorithms in modelling housing prices in the GTHA for 2001, 2006, 2011 and 2016. The results show that temporal lagged models achieve an R^2 of around 0.87, which is competitive with ML models. In addition, among the ML models, random forest (RF) is found to have the best predictive performance.

In all the models tested, accessibilities show significant influence on housing price. Housing planners should pay close heed to the accessibility provided by the transportation system, especially to jobs by transit, when implementing affordable housing programmes to

Variable	Feature importance	With lagged variable	Variable	Feature importance	With lagged variable
lag		0.335	prk4_x	0.006	0.005
hage	0.014	0.009	prk8_x	0.014	0.010
nr	0.046	0.026	rel4_x	0.003	0.002
nper	0.073	0.053	rel8_x	0.007	0.004
maj_repair	0.005	0.003	sch4_x	0.003	0.002
high_sch	0.013	0.008	sch8_x	0.006	0.004
ps_certdip	0.020	0.012	smt4_x	0.002	0.002
ps_deg	0.077	0.051	smt8_x	0.007	0.004
Avg_HHinc	0.170	0.115	chd4_x	0.002	0.001
dweld	0.018	0.011	chd8_x	0.017	0.012
emp	0.011	0.007	ACAR	0.069	0.045
pres	0.011	0.007	ATRAN	0.029	0.018
pind	0.003	0.002	JACAR	0.065	0.038
ppark	0.005	0.003	JATRAN	0.037	0.025
popen	0.008	0.005	RD_DEN	0.012	0.007
eat4_x	0.005	0.003	TR_DEN	0.010	0.006
eat8_x	0.010	0.007	TDEN	0.013	0.008
grc4_x	0.003	0.002	year_2001	0.028	0.022
grc8_x	0.007	0.005	year_2006	0.008	0.005
hcr4_x	0.004	0.003	year_2011	0.023	0.015
hcr8_x	0.010	0.006	year_2016	0.125	0.093

Table 7. Feature importance values for random forest regression

prevent the mismatches between the locational preference of targeted groups and project location. The positive effect of accessibility to people by car on housing price indicates that less dense areas with attractive environments that are well connected to the road network are still preferred by home buyers, which might encourage continuing urban sprawl. To keep a balance between jobs and housing, residential projects within walkable or transit-accessible distance to job centres should be encouraged, and transit-oriented communities in suburbs could be considered. The negative relationship between accessibility to people by transit and housing price indicates that the current transit coverage might not be comparable to that of the road network. People living in suburban areas rely heavily on autos, whose preference might not be easy to alter.

This study has several limitations. Due to data availability, the study only analyzes prices in Canadian Census years – 2001, 2006, 2011 and 2016, instead of building a fully temporal autoregression on each year from 2001 to 2016. The time interval of five years may be too long to fully capture temporal effects in price variation. Only three ML algorithms are tested in this study. More empirical studies with multiple modelling techniques should be done to extend the temporal modelling of housing price and to better understand the temporal dynamics of transportation and housing development.

References

Ahn, S.K. and Reinsel, G.C. (1990), "Estimation for partially nonstationary multivariate autoregressive models", *Journal of the American Statistical Association*, Vol. 85 No. 411, pp. 813-823, doi: 10.1080/01621459.1990.10474945.

Al-Akhras, M., El Hindi, K., Habib, M. and Shawar, B.A. (2021), "Instance reduction for avoiding overfitting in decision trees", *Journal of Intelligent Systems*, Vol. 30 No. 1, pp. 438-459.

Al-Masum, M.A. and Lee, C.L. (2019), "Modelling housing prices and market fundamentals: evidence from the Sydney housing market", *International Journal of Housing Markets and Analysis*, Vol. 12 No. 4.

Apergis, N. and Rezitis, A. (2003), "Housing prices and macroeconomic factors in Greece: prospects within the EMU", *Applied Economics Letters*, Vol. 10 No. 9, pp. 561-565.

Banister, D. (2001), *Transport Planning*, Emerald Group Publishing Limited. Bingley.

Barker, K. (2008), "Planning policy, planning practice, and housing supply", *Oxford Review of Economic Policy*, Vol. 24 No. 1, pp. 34-49.

Black, J. (2018), *Urban Transport Planning: Theory and Practice*, Vol. 4, Routledge, New York, NY.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32.

Breiman, L. Cutler, A. Liaw, A. and Wiener, M. (2011), "Package random Forest", Software, available at: http://stat-www.berkeley.edu/users/breiman/RandomForests

Chau, K.W. and Chin, T.L. (2003), "A critical review of literature on the hedonic price model", *International Journal for Housing Science and Its Applications*, Vol. 27 No. 2, pp. 145-165.

Chen, J.-H., Ong, C.F., Zheng, L. and Hsu, S.-C. (2017), "Forecasting spatial dynamics of the housing market using support vector machine", *International Journal of Strategic Property Management*, Vol. 21 No. 3, pp. 273-283.

Crespo, R. and Grêt-Regamey, A. (2013), "Local hedonic house-price modelling for urban planners: advantages of using local regression techniques", *Environment and Planning B: Planning and Design*, Vol. 40 No. 4, pp. 664-682.

Cutler, A., Cutler, D.R. and Stevens, J.R. (2012), "Random forests", *Ensemble Machine Learning*, Springer, New York, NY pp. 157-175.

Dietterich, T.G. and Kong, E.B. (1995), "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms".

Farooq, B. and Miller, E.J. (2012), "Towards integrated land use and transportation: a dynamic disequilibrium based microsimulation framework for built space markets", *Transportation Research Part A: Policy and Practice*, Vol. 46 No. 7, pp. 1030-1053.

Fox, R. and Tulip, P. (2014), "Is housing overvalued?", available at: https://ssrn.com/abstract=2498294 or doi: 10.2139/ssrn.2498294.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), "Machine learning basics", *Deep Learning*, Vol. 1, pp. 98-164.

Harrison, L., Penny, W.D. and Friston, K. (2003), "Multivariate autoregressive modeling of fMRI time series", *NeuroImage*, Vol. 19 No. 4, pp. 1477-1491.

Hossain, B. and Latif, E. (2009), "Determinants of housing price volatility in Canada: a dynamic analysis", *Applied Economics*, Vol. 41 No. 27, pp. 3521-3531.

Hsiao, C. (1982), "Autoregressive modeling and causal ordering of economic variables", *Journal of Economic Dynamics and Control*, Vol. 4, pp. 243-259.

Hu, L. and Wang, L. (2019), "Housing location choices of the poor: does access to jobs matter?", *Housing Studies*, Vol. 34 No. 10, pp. 1721-1745.

Jadevicius, A. and Huston, S. (2015), "ARIMA modelling of Lithuanian house price index", *International Journal of Housing Markets and Analysis*, Vol. 8 No. 1.

Jenkins, P., Smith, H. and Wang, Y.P. (2006), *Planning and Housing in the Rapidly Urbanising World*, Routledge. New York, NY.

Kauko, T. (2010), "Value stability in local real estate markets", *International Journal of Strategic Property Management*, Vol. 14 No. 3, pp. 191-199.

Kavitha, S., Varuna, S. and Ramya, R. (2016), "A comparative analysis on linear regression and support vector regression", Paper presented at the 2016 Online International Conference on Green Engineering and Technologies (IC-GET).

Kenny, G. (1999), "Modelling the demand and supply sides of the housing market: evidence from Ireland", *Economic Modelling*, Vol. 16 No. 3, pp. 389-409.

Kitts, B. (1997), "Regression trees".

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. (2007), "Supervised machine learning: a review of classification techniques", *Emerging Artificial Intelligence Applications in Computer Engineering*, Vol. 160 No. 1, pp. 3-24.

Koza, J.R., Bennett, F.H., Andre, D. and Keane, M.A. (1996), "Automated design of both the topology and sizing of analog electrical circuits using genetic programming", *Artificial Intelligence in Design'96*, Springer, New York, NY pp. 151-170.

Leung, C.K.Y., Leong, Y.C.F. and Wong, S.K. (2006), "Housing price dispersion: an empirical investigation", *The Journal of Real Estate Finance and Economics*, Vol. 32 No. 3, pp. 357-385.

Levine, J. (1998), "Rethinking accessibility and jobs-housing balance", *Journal of the American Planning Association*, Vol. 64 No. 2, pp. 133-149.

Lieske, S.N., van den Nouwelant, R., Han, J.H. and Pettit, C. (2019), "A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices", *Urban Studies*, Vol. 58 No. 1.

Lieske, S.N., van den Nouwelant, R., Han, J.H. and Pettit, C. (2021), "A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices", *Urban Studies*, Vol. 58 No. 1, pp. 182-202.

Maldonado, S. and Weber, R. (2009), "A wrapper method for feature selection using support vector machines", *Information Sciences*, Vol. 179 No. 13, pp. 2208-2217.

Mason, C. and Quigley, J.M. (1996), "Non-parametric hedonic housing prices", *Housing Studies*, Vol. 11 No. 3, pp. 373-385.

Miles, D. (1994), *Housing, Financial Markets and the Wider Economy*, Wiley. New York, NY.

Mok, H.M., Chan, P.P. and Cho, Y.-S. (1995), "A hedonic price model for private properties in Hong Kong", *The Journal of Real Estate Finance and Economics*, Vol. 10 No. 1, pp. 37-48.

Morris, J.M., Dumble, P.L. and Wigan, M.R. (1979), "Accessibility indicators for transport planning", *Transportation Research Part A: General*, Vol. 13 No. 2, pp. 91-109.

Muellbauer, J. and Murphy, A. (2008), "Housing markets and the economy: the assessment", *Oxford Review of Economic Policy*, Vol. 24 No. 1, pp. 1-33.

Navada, A., Ansari, A.N., Patil, S. and Sonkamble, B.A. (2011), "*Overview of use of decision tree algorithms in machine learning*", Paper presented at the 2011 IEEE control and system graduate research colloquium.

Ojala, M. and Garriga, G.C. (2010), "Permutation tests for studying classifier performance", *Journal of Machine Learning Research*, Vol. 11 No. 6.

Oxley, M. (2004), *Economics, Planning and Housing*, Springer. New York, NY.

Páez, A., Long, F. and Farber, S. (2008), "Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques", *Urban Studies*, Vol. 45 No. 8, pp. 1565-1581.

Park, B. and Bae, J.K. (2015), "Using machine learning algorithms for housing price prediction: the case of Fairfax county, Virginia housing data", *Expert Systems with Applications*, Vol. 42 No. 6, pp. 2928-2934.

Phan, T.D. (2018), "Housing price prediction using machine learning algorithms: the case of Melbourne city, Australia", Paper presented at the 2018 International Conference on Machine Learning and Data Engineering (iCMLDE).

Population estimates, July 1, by census metropolitan area and census agglomeration, 2016 boundaries (2020), Available at: www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710013501

Quigley, J.M. (1985), "Consumer choice of dwelling, neighborhood and public services", *Regional Science and Urban Economics*, Vol. 15 No. 1, pp. 41-63.

Rodriguez, D.A. and Rogers, J. (2014), "Can housing and accessibility information influence residential location choice and travel behavior? An experimental study", *Environment and Planning B: planning and Design*, Vol. 41 No. 3, pp. 534-550.

Rosen, S. (1974), "Hedonic prices and implicit markets: product differentiation in pure competition", *Journal of Political Economy*, Vol. 82 No. 1, pp. 34-55.

Rosenfield, A., Chingcuanco, F. and Miller, E.J. (2013), "Agent-based housing market microsimulation for integrated land use, transportation, environment model system", *Procedia Computer Science*, Vol. 19, pp. 841-846.

Sari, R., Ewing, B.T. and Aydin, B. (2007), "Macroeconomic variables and the housing market in Turkey", *Emerging Markets Finance and Trade*, Vol. 43 No. 5, pp. 5-19.

Selim, H. (2009), "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2843-2852.

Soltani, A., Pettit, C.J., Heydari, M. and Aghaei, F. (2021), "Housing price variations using spatio-temporal data mining techniques", *Journal of Housing and the Built Environment*, Vol. 36 No. 3, pp. 1199-1227.

Somvanshi, M., Chavan, P., Tambade, S. and Shinde, S. (2016), "A review of machine learning techniques using decision tree and support vector machine", Paper presented at the 2016 International Conference on Computing Communication Control and automation (ICCUBEA).

Truong, Q., Nguyen, M., Dang, H. and Mei, B. (2020), "Housing price prediction via improved machine learning techniques", *Procedia Computer Science*, Vol. 174, pp. 433-442.

Usman, H., Lizam, M. and Adekunle, M.U. (2020), "Property price modelling, market segmentation and submarket classifications: a review", *Real Estate Management and Valuation*, Vol. 28 No. 3, pp. 24-35.

Worcester, P. (2019), "A comparison of grid search and randomized search using scikit learn".

Zhang, Y., Zhang, D. and Miller, E.J. (2021), "Spatial autoregressive analysis and modeling of housing prices in city of Toronto", *Journal of Urban Planning and Development*, Vol. 147 No. 1.

Further reading

License (2020), "3.2.4.3.1. Sklearn.ensemble.Random Forest classifier", available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Corresponding author

Yu Zhang can be contacted at: yyu.zhang@mail.utoronto.ca