ASCE Author Proofs

Important Notice to Authors

Attached is a PDF proof of your forthcoming article in **Journal of Urban Planning and Development**. The manuscript ID number is **UPENG-3144**.

No further publication processing will occur until we receive your response to this proof. Please return any corrections within 48 hours of receiving the download email. Your paper will be published in its final form upon receipt of these corrections. You will have no further opportunities to review your proof or to request changes after this stage.

Information and Instructions

- The graphics in your proof have been down-sampled to produce a more manageable file size and generally represent the online presentation. Higher resolution versions will appear in print.
- Proofread your article carefully, as responsibility for detecting errors lies with the author.
- Mark or cite all corrections on your proof copy only.
- Corrections should be completed within 48 hours after receipt of this message.
- If no errors are detected, **you are still required to log in, make note in the proof, and finalize the article to indicate that it is okay** to be published as is.
- You will receive a message confirming receipt of your corrections within 48 hours.

Questions and Comments to Address

The red numbers in the margins correspond to queries listed on the last page of your proof. Please address each of these queries when responding with your proof corrections.

Return your Proof Corrections

- Web: If you accessed this proof online, follow the instructions on the Web page to submit corrections.
- E-mail: Send corrections to ascejournals@novatechset.com. Include the manuscript ID UPENG-3144 in the subject line. Please do not provide a revised manuscript.

Please annotate and complete your proof review within 48 hours. Should this not be possible or should you encounter any problems or have further questions, please contact the Journal Production Manager at ascejournals@novatechset.com and include UPENG-3144 in the subject line.

ASCE Open Access

Authors may choose to publish their papers through ASCE Open Access, making the paper freely available to all readers via the ASCE Library website. ASCE Open Access papers will be published under the Creative Commons-Attribution Only (CC-BY) License. The fee for this service is $2,000, and must be paid prior to publication. If you indicate Yes, you will receive a follow-up message with payment instructions. If you indicate No, your paper will be published in the typical subscribed-access section of the Journal.

Selecting Yes does not commit you to publishing your article as Open Access. You will have the option to cancel the Open Access process later. If you are unsure, we recommend selecting Yes. If you select No now, your paper will be published online shortly after your proof corrections are received, and you will no longer have the ability to publish your article as Open Access.

Color Figures

Figures containing color will appear in color in the online journal. All figures will be grayscale in the printed journal unless you have agreed to pay the color figure surcharge and the relevant figure caption indicates “(Color)”. For figures that will be in color online but grayscale in print, please ensure that the text and captions do not describe the figures using colors.

If you have indicated that you will be printing color figures in color, you will receive a notification with a link to the payment system. Until payment is received or color printing is canceled, your article will not be published.

Reprints

If you would like to order reprints of your article, please visit https://www.asce.org/reprints.

Case Study

Spatial Autoregressive Analysis and Modeling of Housing Prices in City of Toronto

Yu Zhang[1]; Dachuan Zhang[2]; and Eric J. Miller[3]

Abstract: Previous housing price studies based on hedonic price modeling have mainly focused on applying various factors, including built environment variables in the analysis, without establishing a comprehensive theoretical framework as a basis for the model formulation. To address this gap, this study introduces a more systematic framework for decomposing housing prices into land prices as determined by built form, neighborhood socioeconomic characteristics and individual dwellings' physical conditions. Following this logic, this study experiments with the related variables through regression analysis, including consideration of spatial lags, as well as develops a housing price model using a random forests (RF) algorithm. A comprehensive time-series database of housing transaction data for the City of Toronto is used. Modeling results show that neighborhood socioeconomic factors contribute the most to the explanation of housing prices, while housing characteristics and accessibility measures are also significantly influential. The RF model achieves an overall accuracy of 85%, a relatively good performance in reproducing observed prices. The findings provide insights for planners concerning factors influencing housing prices and, hence, residential location decision-making. **DOI: 10.1061/(ASCE)UP.1943-5444.0000651**.

Author keywords: Housing price modeling; Geographical weighted regression (GWR); Random forests (RF) model.

Introduction

Housing market regulation and affordable housing provision have long been a key objective for government to improve residents' overall wellbeing and quality of life (Burt et al. 2001; Nguyen 2005). According to the Survey of Household Spending Report, shelter is the largest budget item for Canadian households, at 29.2% of the total consumption (Statistics Canada 2018). Housing prices in the Province of Ontario have been soaring since the beginning of the 21st century and have almost doubled from 2001 to 2016, whereas the average housing price in the City of Toronto in 2019 showed nearly a sevenfold increase since 2001. Housing is undoubtedly one of the highest-return investment products in the past 20 years, but it has also become increasingly unaffordable (Diamond and McQuade 2019; Massey and Rugh 2017; Tong et al. 2019). As a rigid demand product, housing price fluctuations greatly affect household spending and residents' quality of life. Many factors in the social context align with the spatial attributes affecting housing markets (Anderson et al. 1996; Habib and Miller 2008; Haider and Miller 2000). The relationship between these influential factors and housing price could provide the basis and logic for improved simulation models of housing prices. The objective of this research is to analyze the determination mechanisms of housing prices and provide market trend estimations and forecasting for planners and urban engineers to form proper policies and measures for regulating urban land use and housing markets (Chen et al. 2016).

In this study, we aim to build a housing price model that not only applies machine learning (ML) as a new and promising approach to housing price modeling, but also is developed based on a theoretical foundation concerning housing price determinants. Therefore, this research has two purposes: (1) to study the housing price determinants in mega-cities; and (2) to develop a microlevel housing price simulation model as a tool for short-run housing price modeling. Following a brief review of current research progress related to housing prices, the next section describes the details of the data and methods employed, including the conventional Hedonic Price Model and the random forests (RF) approach. This is followed by a description of an empirical study in City of Toronto. The penultimate section provides the results of housing price determinants from linear regression and the GWR model, and the performance of RF housing price model are interpreted and discussed. Conclusions and some policy implications for housing planning and housing market regulations are presented in the final section.

[1]Ph.D. Candidate, Dept. of Civil & Mineral Engineering, Univ. of Toronto (corresponding author). ORCID: https://orcid.org/0000-0001-5997-9805. Email: yyu.zhang@mail.utoronto.ca

[2]Ph.D. Candidate, School of Geography & Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-Sen Univ. ORCID: https://orcid.org/0000-0001-7378-9784. Email: zhangdchgis@foxmail.com

[3]Professor, Dept. of Civil & Mineral Engineering, Univ. of Toronto. Email: eric.miller@utoronto.ca

Note. This manuscript was submitted on May 15, 2020; approved on September 15, 2020No Epub Date. Discussion period open until 0, 0; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Urban Planning and Development***, © ASCE, ISSN 0733-9488.

Literature Review

Housing Price Models

Numerous quantitative models derived from urban economics have been developed since the 1960s (Mark and Goldberg 1984). In order to fully capture the determinants of housing prices, different approaches were applied, including hedonic housing price models, repeat sales models (RSM), hybrid approaches, and local quantile housing price models (Bourassa et al. 2006; Case et al. 1991; McMillen 2013; Morris et al. 2020; Rosen 1974).

Bailey et al. (1963) introduced the Repeated Housing Price model, which has been widely used for estimating housing market trends. It assumes that the individual housing price could be determined by its own transaction value and the overall variance in trends. This trend analysis approach assesses housing value by focusing on the historical transaction records instead of the housing itself, which ignores the potential impact from changes in urban-built form and surrounding land use. It is more commonly used in analyzing housing price volatility, even though this approach uses subsamples containing part of all transactions, which could be less representative. Researchers can control the hedonic housing price characteristics for multiple transactions and only focus on the changes due to time variation (Wallace and Meese 1997). A repeat sales estimator is subject to the sample data and is used under the assumption of time consistency. It is also assumed in RSM that the implicit attributes of housing itself remain the same over time. Without a fundamental inclusion of housing characteristics, RSM alone is less grounded in constructing housing price (Englund et al. 1999). A hybrid method using not only multiple transactions but also the information of each single sale was developed to overcome this shortcoming (Quigley 1994).

Local quantile housing price models allow the hedonic model to vary over space. Instead of using sample means, quantile regression focuses on the quantile points in the housing price distribution, which is more robust when applied to nonnormal distributed housing prices (Koenker and Hallock 2001). Zietz et al. (2008) used a quantile regression model to identify the different housing price determinants for high- and lower-priced houses. McMillen (2013) used quantile estimation to analyze the cross-sectional housing price variation. Local quantile regression can reveal the variation over space as well as the distribution of housing prices. It performs better for macrolevel housing price analysis than individual housing price simulation.

Hedonic housing price models are the most commonly used method in the literature and have been extensively explored. Within this approach, housing can be characterized as a bundle of services that fulfill consumers' needs, and housing prices are determined by the attributes of housing, constrained by the budget of utility-maximized 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). Housing price is constructed by decomposing housing into serval components that do not have individually observable market prices: physical condition, locational characteristics, surrounding neighborhood, and land use composition. Factors from structural, locational, neighborhood, and environmental aspects can also be included in the model (Kim et al. 2015). Socioeconomic factors and surrounding land use have also been taken into account in recent years, since the location value of housing plays a critical part in housing price. Several housing price studies have been conducted using this framework (Can 1992; Chau and Chin 2003; Goodman 1978, 1988). With changes in urban-built form over time, the factors in consideration gradually evolve from simply the physical condition of housing to the location, transport accessibility, diversity or the land use mix degree, and social environment (Levine 1998; Osland and Thorsen 2008; Wang et al. 2007). Investigating in depth into more detailed housing price determinants and exploring influential factors from the demand side could optimize current modeling of the housing market and facilitate housing planning.

Spatial Effect in Housing Price Analysis

As housing is fixed in space, spatial dependency of housing prices will exist among adjacent units. Spatial heterogeneity can affect the distribution of housing prices. The sale comparison approach in real-estate appraisal basically determines the housing value by comparing the transaction price of units that have similar locations and other characteristics (Clapp and Giaccotto 1992); in other words, housing prices will tend to be spatially autocorrelated. Therefore, housing units are prone to form a spatially aggregated cluster, which represent the "regional price" of a neighborhood. However, discrete administration boundaries cannot well represent continuous spatial lags in practices, and so spatially weighted regression is essential to account for such spatial lags in housing price models.

Geographical weighted regression (GWR) has been widely utilized in housing market research. Dubin et al. (1999) summarized the spatial autoregression method to solve the spatial residual dependency problem and to use fewer independent variables to improve model performance. Spatial lags of both dependent and independent variables were used. Can (1992) incorporated locational effects in the model specification and estimation of hedonic price models, and found that the incorporation of market segmentation, neighborhood, and adjacency effects should be considered to improve the model. In this study, census tracks were used as a proxy for neighborhoods, and demand was the only driving force of spatial heterogeneity regardless of the quality or physical features of an individual housing unit. Haider and Miller (2000) used a spatial autoregressive model to analyze the effect of proximity to transportation infrastructure on residential values. Bowen et al. (2001) studied the housing price determinants in Ohio with an extended hedonic price model to control for spatial dependency and heterogeneity. Bitter et al. (2007) analyzed housing attribute prices in Tucson, Arizona, comparing two approaches: spatial expansion and GWR. The marginal housing prices were examined and GWR was found to outperform the spatial expansion method in predictive accuracy. Huang et al. (2010) extended the GWR to include temporal variations (GTWR) in housing price variations and found that GTWR performs better than GWR and temporally weighted regression (TWR) models in housing price modeling in Calgary, Canada. Using a GWR specification, Cohen and Coughlin (2008) found that housing prices within an area disrupted by airport noise were about 20% less than in undisrupted neighborhoods. Cao et al. (2019) studied public housing prices in Singapore. Significant factors affecting public housing resale price were identified by applying three regression models and a travel time-based GWR model was found to yield the best fit. In summary, numerous studies have shown that spatial autoregressive models perform better in explaining housing prices than simpler regression models.

In this paper we adopt a hedonic housing price model as the basis to develop our model of housing prices, as well as using a spatial autoregressive model to reduce the spatial dependency error. The geographical proximity effect could partially explain the similarity of price due to externality effects and shared neighborhood characteristics. The assumption is that the relationship between housing price and independent factors can be better revealed after removing the spatial autocorrelation.

Housing Price Simulation

Simulation models to support public decision-making have been used since the 1980s. For example, Regional Economic Models, Inc. (REMI) has developed an economic–demographic simulation model that is widely are used in the United States for policy and general demographic simulation (Treyz 1995). Five components (output linkages; population and labor supply; labor and capital demand; market shares; and wage, price, and profit) define the model framework, which interact dynamically with each other. The simulation does not have housing price prediction as its objective, but rather supports public and private sector decision-making at a macro level. Landis (1994) developed a metropolitan simulation model, California Urban Future Model (CUF), which represents urban growth patterns and impacts of policies at different levels. Housing price was used as the input of the overall model to simulate the reaction of the system to different policy scenarios. Researchers also employed different statistical approaches to model housing price. Kouwenberg and Zwinkels (2014) used a smooth transition model and performed a simulation for the US housing market. Without much inclusion of housing characteristics, they based their estimation on rent and housing price index levels. Balcilar et al. (2015) show that a nonlinear model is necessary for housing price simulation in order to achieve reasonable predictive accuracy. In addition, numerous integrated transport–land use (ILUT) models exist that endogenously generate housing prices as part of a larger process of modeling land development and population and employment spatial distributions over time. Examples include, but are not limited to, UrbanSim (Waddell 2002), PECAS (Hunt 2003), MUSSA (Martinez 1996), MEPLAN (Echenique et al. 1990), TRANUS (De La Barra et al. 1984), and SILO (Ziemke et al. 2016), among others.

Going beyond conventional econometric models, in recent years many studies have applied ML algorithms to study housing markets, including support vector machines (SVM), artificial neural networks (ANN), and convolutional neural networks (CNN). Yan et al. (2007) used the TEI@I method (a systematic integration of artificial intelligence and traditional econometrical models) with the input of 114 indicators related to housing price from both macro- and microlevels, and supply and demand sides, to simulate commercial housing prices and evaluate macrolevel policies. Xie and Hu (2007) applied ANNs and SVMs to simulate the time series housing price index in Shanghai, and found that the ANN model and SVM model performed better in simulating long-term housing prices compared with a traditional ARIMA method. Gu et al. (2011) used genetic algorithms and support vector machines (G-SVMs) in housing price simulation and stated that G-SVM is the superior approach regarding the accuracy and robustness of the simulation compared with grid algorithm (GA) and SVM. Park and Bae (2015) developed a housing price prediction model to analyze housing price variations (trends in closing prices compared with list prices), applying the ML algorithms of C4.5, RIPPER, Naïve Bayesian, and AdaBoost. Factors under consideration include physical features, mortgage rates of individual housing units, and the public school rating of the located neighborhood. RIPPER was found to outperform other models in terms of accuracy and consistency. Oladunni and Sharma (2016) applied ML to traditional hedonic pricing theory, using support vector regression (SVR), K-nearest neighborhood (KNN) and principal component regression (PCR) as the learning algorithms in a case study of eight counties in Washington, DC. PCR was found to perform best in this application. Rafiei and Adeli (2016) developed a real-estate sale price estimation model for the supply side, which provides references for construction companies to forecast the housing market in their project management decision-making. The model used an integration of a restricted Boltzmann machine and nonmating genetic algorithm and optimized the input structure to reduce the dimensionality curse. Hu et al. (2019) studied the housing rental price variation with six ML algorithms, including RF, extra-tree regression (ETR), gradient-boosting regression (GBR), and identified the relative contribution of the determinants through social media datasets.

However, there are some shortcomings of these ML algorithms in simulating housing price. SVM is a nonlinear algorithm with strong adaptability, but it has low computational efficiency and it is difficult to generate a classifier with massive training datasets. ANN can address some of the above problems, but the internal mechanism of the training process is not clear and often generates overfitting results. It is time-consuming and difficult to parallelize. CNN can generate optimal validation accuracy with high efficiency, nonetheless it lacks convincing explanations concerning its implicit features. In contrast, the RF algorithm is one of the most suitable ML methods for minimizing the overfitting issue (Breiman 2001). It is considered an effective and universal algorithm that improves the ability of data regression/prediction during the model training process (Fernández-Delgado et al. 2014). RF algorithms are applied in urban studies including simulating urban growth (Kamusoko and Gamba 2015; Zhang et al. 2019), modeling land surface temperature (Yang et al. 2019), and mapping population distributions (Yao et al. 2017). Given this, this paper explores the application of the RF approach to housing price modeling.

Despite the many previous housing price modeling studies using both traditional and artificial intelligence methods, there's a research gap with respect to microlevel modeling of individual housing prices. After developing a framework for explaining house price determination, we implement this framework within GWR and ML housing price models for the City of Toronto.

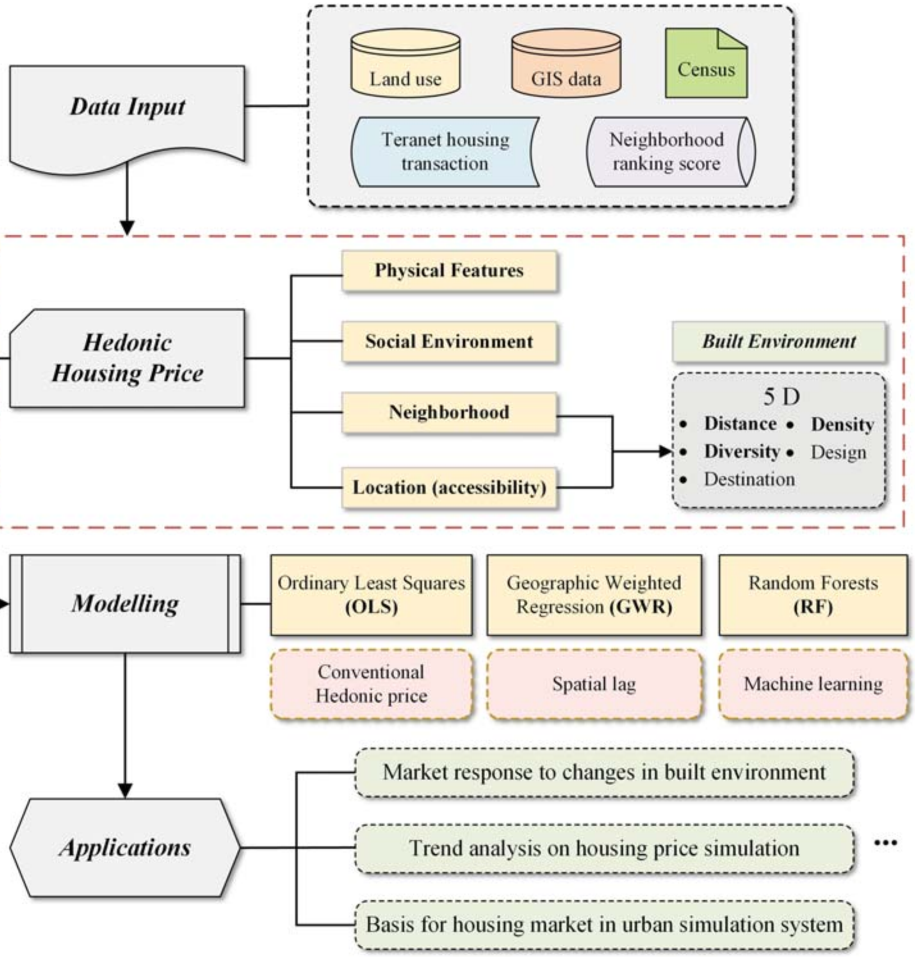
Methodology

The sales price of a house can be divided into two components: the value of the land upon which the house sits, and the value of the dwelling unit itself. The land price captures the surrounding built form and social environment. We incorporate Cervero's 5D model of built form and add the socioeconomic environment dimension. The price of the housing unit relates more to its physical condition and quality. Fig. 1 gives a summary of the workflow and research methods.

Constructing a Framework for Built Form and Social Environment

Cervero and Kockelman (1997) originally characterized built form in terms of three categories (Density, Diversity, Design): the "3D's." This typology was late and then extended to five dimensions with the addition of Distance to transit and Destination accessibility by Cervero et al. (2009). The 3/5D typology has been used in travel demand analysis for the past 20 years, in recognition that the built-form characteristics of trip origins and/or destinations (e.g., their land use, densities, design features) can affect not only trip generation, but also travel modes and routes (Cervero and Kockelman 1997). However, the concept has gradually spread to other applications, including housing analysis. Built environment provides the physical space for all human activities and has considerable influence on mobility, housing, and population distribution. Thirteen variables (e.g., population density, dissimilarity index, entropy, pedestrian and cycling provisions) were included in the initial 3D model and through factor analysis, two intuitive and interpretable factors were extracted, named as intensity, which captures the Density dimension, and walking quality, which captures the Design dimension. In order to further analyze the impact of built environment on walking and cycling, Cervero et al. (2009) extended the 3D model to 5D, adding distance to transit and destination accessibility into the framework.

We employ the 5D model in representing built environment to identify housing price determinants. From the utility perspective, the physical and locational elements mean that housing functions not only as a physical shelter, but also as an origin to get access to multiple activities. From the spatial competition perspective,



F1:1 **Fig. 1.** Analysis framework and the overall workflow.

the price of the housing should be the equilibrium price of the land lot among different competing land uses, plus the construction cost of the housing. In a word, built environment is the major part that determines the baseline price of a housing unit.

Following Cervero's 5D model, in this study we only applied three dimensions, Density, Diversity, and Distance, with an extra dimension representing the Socioeconomic Environment. Since the design of neighborhood (street design, pedestrian safety) and the destination dimension are more related to travel behavior and less to housing price, these two dimensions are not included in the analysis framework. For Density, the floor area ratio and population density of the neighborhood should be considered. For Diversity, how mixed the land use around the housing unit is considered, and the proportions of different land use types are included. For Distance, the locational characteristics of the housing can be represented by, for example, the distance to public facilities, distance to public transit, and distance to the CBD. For Socioeconomic Environment as the overall social perception of the built environment that the housing locates, the safety in the surrounding area, the average income of residents, and educational degree could be indicative. The overall framework of built environment is presented in Table 1.

Hedonic Housing Price and Variable Selection

The basic housing price function we apply in this study is the hedonic housing price model [Eq. (1)]. The housing price can be decomposed into the housing characteristics itself $\vec{s}$, the neighborhood quality $\vec{n}$ and surrounding environment $\vec{e}$ (Chau and Chin 2003; Malpezzi et al. 1998; Mok et al. 1995; Witte et al. 1979) as

$$P = f(\vec{s}, \vec{n}, \vec{e}) \tag{1}$$

With the foundation of hedonic housing price model, and the built form and socioeconomic environment framework, we built the housing price from two parts: the housing and the land. Housing is physically attached to a fixed location, which captures the price of the land. Therefore, characteristics of the surrounding environment set the basic price range and physical condition of the housing would differentiate the housing units in the same neighborhood.

The indicators included in the regression model are selected to represent the three aspects of housing price: physical characteristics, built form, and socioeconomic environment, as listed in Table 2. Housing characteristics should cover the physical features including the unit size, house age, number of rooms, garage area, number of bathrooms and kitchens, and maintenance condition. People may have different preferences over the design such as the direction of bedrooms or connection between each part of the housing, but the overall structural preference is more common (e.g., bigger housing should cost more).

Density can be represented by the population density and employment density. For the Distance dimension, locational characteristics can be represented by the physical travel distance or access to major transit station, major health and shopping centers.

Table 1. Dimensions of built and socioeconomic environment influencing housing price

Dimensions	Indicators
1. Density	
• Population density	Population density of the neighborhood.
• Employment density	Employment rate in the neighborhood; job density of the neighborhood; labor force/ job ratio of the neighborhood.
2. Diversity	
• Land use mix	Measurement of land use mix degree based on Entropy Index, Dissimilarity Index and Herfindahl-Hirschman Index.
• Intensity of different land use type	Proportion of commercial/retail/residential/ industrial/green land area on the site.
3. Distance	
• Centrality	Distance to the city center.
• Distance to public facilities	Distance to medical facilities, large shopping malls, major cultural facilities such as galleries and museums, and public schools.
• Accessibility to public transit	Distance to the subway station; distance to the bus stops; and accessibility measurement to the road network expressed as in the gravity model.
4. Socioeconomic environment	
• Safety	Crime rate of the neighborhood.
• Educational degree	Percentage of post-secondary education.
• Average income	Average income of the neighborhood.
• Community	Sense of belonging, and integration of different groups of people, whether the community is inclusive.
• Public service provision	Coverage and numbers of public hospitals, clinics, grocery stores, and police station.

We also employ distance to the city center to represent the centrality level, and distance to the metro stations and transit stops represents the accessibility to public transit. Average distance to large shopping centers and to medical facilities captures the accessibility to public facilities and services. In order to represent an overall accessibility of housing in each dissemination area (DA, the smallest geographic area defined in the Canadian census) to the road network, the accessibility computed using distance was calculated and included in the model. We calculated three indices to capture the Diversity dimension and land use mix degree based on the land use distribution of the study area: Entropy Index (ENT), Herfindahl-Hirschman Index (HHI) and Dissimilarity Index (DI) based on the following:

$$\mathrm{ENT} = \frac{\sum_{j=1}^{k} P_j * \ln(P_j)}{\ln(k)} \tag{2}$$

$$\mathrm{HHI} = \sum_{j=1}^{k} (100 * P_j)^2 \tag{3}$$

$$D = \frac{1}{2} \times \sum_{j=1}^{k} |R_j - S_j| \tag{4}$$

where P_j = the percentage of land area of land use type j on the site; k = the total count of land use types inside the DA; R_j = the percentage of land area of land use type j on the site compared to the total region; and S_j = the percentage of land area that is not land use type j on the site compared to the total region. The ENT range is from 0 to 1 with higher land use mix degree as it approaches 1; the HHI range is from 0 to $1/k$, and the higher the mix level, the lower the value (Ihlanfeldt 2007; Song and Knaap 2004).

Since the Socioeconomic dimension related to housing price is relatively hard to qualify and a large-scale disaggregated social survey is not the major goal of this study, we employed the neighborhood ranking scores as the indexes provided by *Toronto Life* (2018) (Note: the neighborhood ranking score is openly published in *TorontoLife National Magazine*, and calculated by UofT's Martin Prosperity Institute. The original report could be found at http://web.archive.org/web/20140925062941/http://martinprosperity.org:80/2013/10/08/insight-rankling-neighbourhoods/), which captures the perceived safety, sense of belonging and inclusiveness of the community, along with average income and education level for the demographic representation of the people living in the neighborhood.

Geographical Weighted Regression Model

GWR is a useful tool for reducing the spatial dependency of dependent variables by using the distance-weighted matrix in the regression. The GWR relaxes the assumption in ordinary regression that the dependent variable should be independent and identically distributed random variables. It is a local modeling approach that explicitly allows parameter estimates to vary over space (Bitter et al. 2007; Brunsdon et al. 1996, 2002; Farber and Páez 2007). Instead of simply using the stationary independent variables in the estimation, it estimates a separate model for each point and includes the distance-weighted observations as a "spatial lag" variable in estimating the price of this point. This method includes the comparison among each housing sales, which is appealing since it applies the "sales comparison" that frequently used by real-estate appraisers (Bitter et al. 2007), and can be represented as

$$y_i = a + \sum_{i}^{k} \beta X_i + \sum_{j}^{n} w_{ij} y_j + \varepsilon \tag{5}$$

where y_i = the housing price of point i as a function of the independent variables X_i and the housing price of other sales points weighted by the distance-decay function w_{ij}. In this study we used an adaptive bandwidth in the kernel density function in assigning weights.

Microsimulation: Housing Price Representation Based on Random Forest

Recently, many studies have applied ML algorithms to simulate housing markets. These prior studies mainly focused on methods of how to develop housing price simulations, with few explanations of the implicit driving factors of housing prices. Planners and practitioners are more concerned with the driving forces and functional mechanisms underlying housing market fluctuation. Reliable methods are still needed to explore and identify the dominant driving determinants.

A RF is a multiclassifier/regression combination model. Previous studies show that RF performs well in handling high data multicollinearity and dimensionality issues (Belgiu and Drăguţ 2016; Wyner et al. 2017). Beyond this, the RF is a theory of measurement through an out-of-bag (OOB) error estimation and bootstrapping sampling with replacement in model training, which theoretically generates a function of variable importance measures (VIMs) (Palczewska et al. 2014; Zhang et al. 2019). The statistics of VIMs can generate quantitative understandings on the importance

Table 2. Details and generalization of each variable

Categories	Variables	Abbr.	Description	Source
Housing price	Price	Price	The average Teranet housing transaction price for each DA in 2016	Teranet Housing Transaction Data
Distance	Distance to the city center	distcc	Distance from the geometric centroid of each DA to the Bay and King intersection, where most financial and stock institutes locate	Calculated in GIS based on Euclidean Distance
	Distance to the transit stops	disttrans	Distance from the geometric centroid of each DA to the nearest bus stop	
	Distance to the metro station	distsb	Distance from the geometric centroid of each DA to the nearest metro station	
	Distance to hospitals	med	Average distance from the geometric centroid of each DA to the ambulance station, hospital, nursing home, and other medical institutes	
	Distance to cultural center	cul	Average distance from the geometric centroid of each DA to the library, art gallery and museum, and exhibitions	
	Distance to school	sch	Average distance from the geometric centroid of each DA to the public and private primary school, secondary school, and universities	
	Distance to large shopping malls	shopcent	Average distance from the geometric centroid of each DA to the community, neighborhood, regional shopping center and cinema	
	Accessibility	access	Defined as a function that indicates the accessibility for residences locations relative to the road network, i.e., highways and main roads	
Diversity	ENT	ent	Land use mix index computed from the entropy index equation	Land use data from the Toronto City Planning Department
	HHI	hhi	Herfindahl–Hirschman Index (HHI)	
	Dissimilarity Index	di	Dissimilarity Index (DI)	
	Intensity of commercial, residential and greenland	intens_com, intens_res, intens_gre	Intensities or percentrage of different land use types as commercial, residential and greenland.	
Density	Population density	popdens	The total population divided by the area	2016 Census of Population
	Employment	emp	The number of employed labor force	
Housing characteristics	Number of rooms	nr	The average number of rooms in each unit	
	Crowded level	crowd	The percentage of housing units that contains shared room (number of persons per room greater than 1)	
	Housing maintenance condition	condi	The percentage of housing units that needs major repair (compared to minor repair)	
	House age	hage	The average house age (2016–built year)	
Socioeconomic Environment	Safety	safe	The number of crimes in each neighborhood	Toronto Life Neighborhood Ranking Score
	Housing affordability	housing	The cost of housing versus the income, appreciation and rate of home ownership	
	Diversity	diver	The percentage of visible minorities, people whose mother tongues are not French or English, and first- and second- generation	
	Community	commu	Voter turnout numbers, community space use per capita and how many people report a sense of community belonging	
	Health	health	The number of medical and mental health services per capita, the number of senior care service per senior, the number of people with family doctors and physical activity levels among residents for each neighborhood	
	Shopping	shop	The number of groceries, markets, and pharmacies per square kilometer	
	Education	edu	The number of schools per child, the number of daycares per child and the share of residents with postsecondary educations	
	Employment	empl	Employment and unemployment rates, the share of residents below the poverty line, the share of high-income and the share of employed residents	
	Income	income	The average annual income of each household	2016 Census of Population
	Education level	highedu	The percentage of residents with above high school level education	

of each determinant driving the dependent variable (i.e., the housing price) to change over time.

Therefore, we apply the RF algorithm in simulating Toronto housing prices. The RF-based simulation comprises two components: a training component and a simulating component. In the training component, the RF algorithm is trained and calibrated by using datasets containing housing price labels and various driving determinants. The VIM is generated during the model training procedure through estimating the OOB error. The well-trained and generated RF classifier is then used to simulate Toronto housing prices.

Empirical Study: City of Toronto

Study Area

We use City of Toronto as our study area (Fig. 2), the most populous city in Canada and the fourth largest city in North America. City of Toronto consists of 3,407 DAs in six districts. The population is 2.93 million in 2017 and the area is 630.2 km^2.

In the past few decades, Toronto has been among the fastest-growing large metropolitan areas in the high-income world and

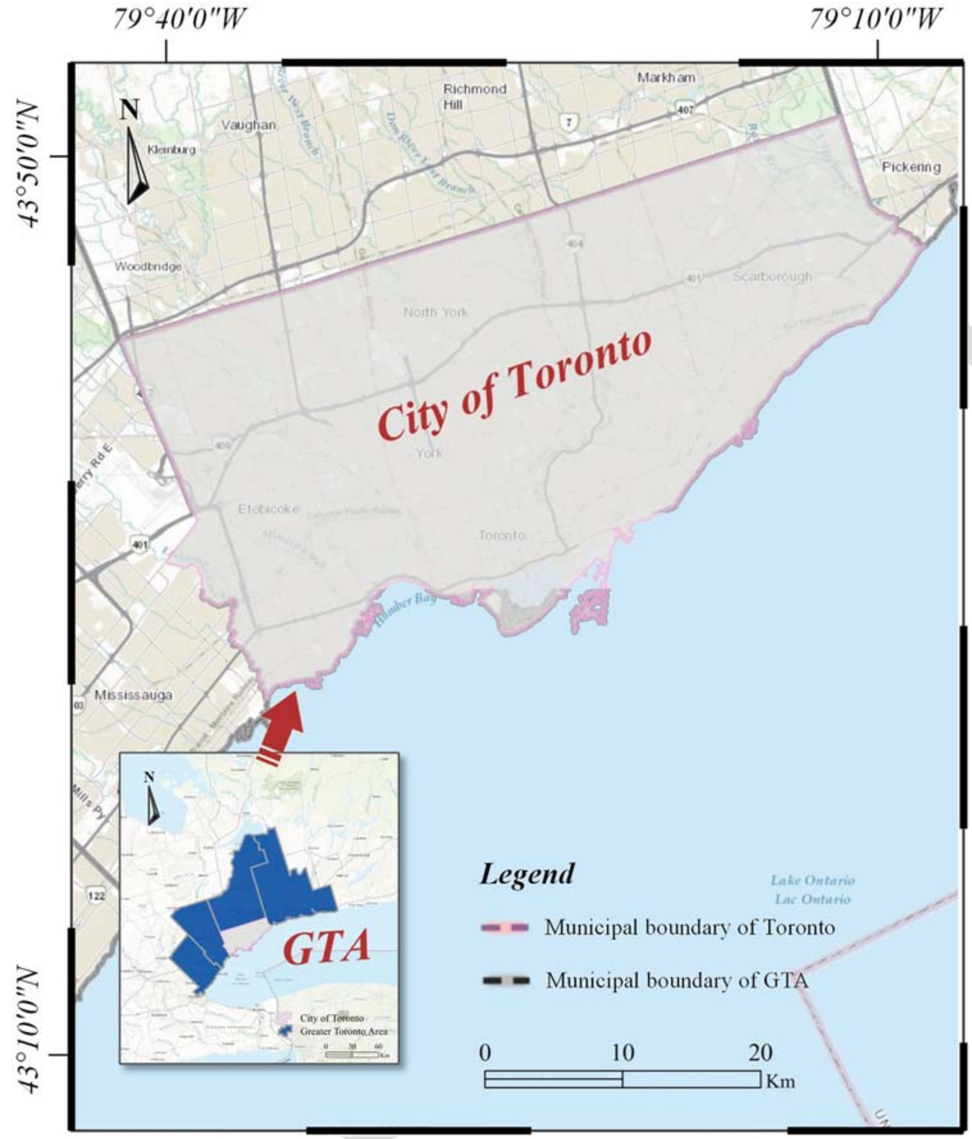
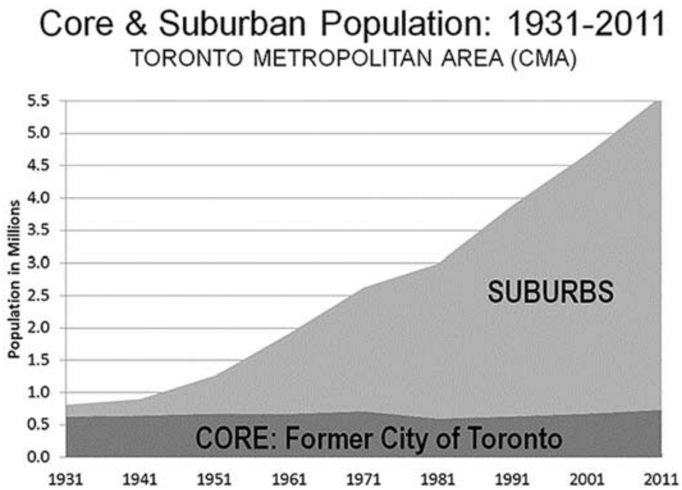


Fig. 2. Location map of the City of Toronto. (Map data from Esri, NASA, NGA, USGS, City of Toronto, Province of Ontario, Esri, HERE, Garmin, METI/NASA, USGS, EPA, NPS, USDA, NRCan, Parks Canada, York University, City of Brampton, City of Toronto, Ontario Base Map, Province of Ontario, Ontario MNR, Esri Canada, Esri, © OpenStreetMap contributors, HERE, Garmin, USGS, NGA, EPA, USDA, NPS, AAFC, NRCan.)

the principal commercial center in Canada. Like most megacities in North America, Toronto initially formed as a monocentric urban structure, with extensive suburban sprawl subsequently occurring post-WWII. Even though the downtown area is densely built with financial and commercial industries, the population growth has mainly occurred in suburban areas both within and adjacent to the traditional Toronto core, known as the Greater Toronto Area (GTA) (Fig. 3). Continuous growth has occurred throughout the metropolitan area, with the economy growing through ongoing investments (capital), immigrants (labor), and land development. As shown in Fig. 3, the most dense part of the City is in the central downtown area near Lake Ontario, with population densities declining in approximately concentric circles as one moves radially outwards from the central core area. Housing prices in the City have increased rapidly over the past 20 years. The average housing price reached 0.83 million CAD in 2017 and is now over 0.9 million CAD according to the Toronto Real Estate Board (TREB). Sales are increasing as well: around 113,040 units transacted in 2016 (Fig. 4). As a city of immigrants, the population inflow increased the housing demand, which raised housing prices, as well as induced further real-estate investment. The magnitude of population, housing market growth, and urban form make Toronto a good case study to analyze the determinants of housing prices in North American megacities.

Data Preparation

Housing prices and related indicators of built form and socioeconomic environment are needed to construct the models. A longitudinal dataset of housing sales data for the period 1986–2016 was obtained from Teranet Inc. This dataset contains the transaction price, date, and location of land sales in the Province of Ontario. This dataset provides us with the overall housing price distribution and fluctuations over a 20-year period. Owing to the high heterogeneity of housing transaction records at the individual parcel level, identification of the role of general housing price determinants might be difficult at the individual dwelling unit/parcel level. The average housing price at the DA level contains less randomness and fits better as the dependent variable in the regression models. Therefore, we aggregate the housing sale



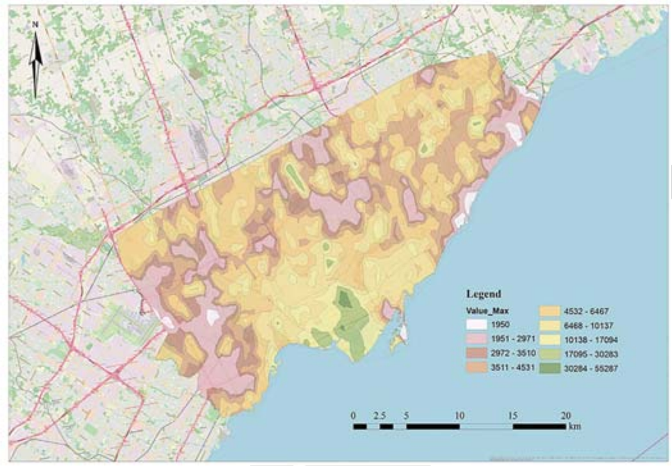


Fig. 3. The population growth and the population distribution in City of Toronto. (Data from Statistics Canada 2012.)

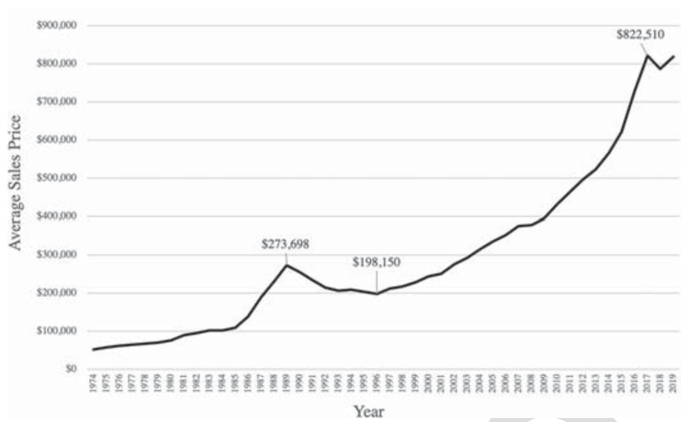


Fig. 4. Average housing sales price and sales amount in City of Toronto.

price data, the demographic characteristics and other housing-related variables to the DA level as our analysis unit. Since DAs are defined based on the population, we assumed that the aggregated results would not be affected by sample size in each DA and are representative of the relatively homogeneous demographic features inside each DA.

The focus of this study is on the determinants of housing price, rather than tracing trends in housing prices over years. Thus, we only model housing prices in 2016, leaving a longitudinal analysis of the full 1986–2016 time series Teranet dataset for future work. The Statistics Canada's Census Profile provides the basic demographic data, including the population, age, income, education, and employment distribution for each DA. We also use the Census of Housing characteristics (e.g., construction period, indoor amenities) to capture the physical condition of housing. In order to represent an overall perceived neighborhood condition (e.g., safety, entertainment, education, health, environment), we use the neighborhood ranking from *Toronto Life* (2018) due to the lack of official computed and commonly recognized neighborhood evaluation. The spatial variables were generated from a set of distance measurements (e.g., distance to the regional center, access to the public transit), calculated in ArcGIS. The points of interests (POIs) shapefiles were provided from the open data of Municipal Property Assessment Corporation (MPAC), which includes POIs in the cultural, medical, commercial, and education fields.

Results

Descriptive Analysis

The basic descriptive statistics of variables is listed in Table 3 and the spatial distributions of the variables are displayed in Fig. 5. Most of the sample data of the explanatory variables are moderately skewed (between −0.5 and 0.5). Before the regression, logarithmic and normalization transformations were performed to remove the skewness in the data. After removing the outliers, 3,264 records were used in the model. In general, housing units have good access to public facilities, with an average radius at around 100 m covering the basic facilities (education, retail, clinic, cultural center). Differences in access to transit stops and subway stations are larger, ranging from about 30 m to 6 km for bus stops, and 40 m to 13 km for metro stations, leaving the households in the uncovered areas with fewer options for travel modes. Even though proximity is observed to be better along the transit lines, the road network accessibility does not follow a clear decreasing pattern toward the peripheral area. The accessibility measure shows that suburban areas have good road access, which is the dominant travel mode for most suburban households. The population density demonstrates an obvious concentration in the downtown area, but this pattern does not show in employment. The number of employed labor force is almost evenly distributed in the entire region, which indicates a

Table 3. Descriptive statistics table of the dependent and independent variables

	Mean	Standard deviation	Median	Minimum	Maximum	Skew
UnitPrice (CAD/sq. meter)	3,073.49	1,211.75	2,718.73	1,024.22	7,166.67	0.85
distcc (m)	11,759.95	5,943.9	11,681.19	158.59	26,725.17	0.22
disttrans (m)	1,320.78	897.27	1,099.3	32.36	6,432.75	1.18
distsb (m)	3,154.67	2,553.53	2,389.28	38.37	13,209.84	1.55
med (m)	130	90	120	90	880	1.04
cul (m)	140	100	120	160	1,060	1.66
sch (m)	70	40	60	170	480	1.09
shopcent (m)	120	70	110	260	810	1.11
access	0.89	0.08	0.92	0.56	0.99	−1.49
safe	43.65	28.33	40.7	0.7	100	0.3
housing	51.9	27.75	52.1	0.7	100	−0.05
commu	51.97	28.12	52.9	0.7	100	−0.06
diver	51.26	29.34	50	0.7	100	−0.01
health	51.55	27.15	50.7	0.7	100	0
shop	48.48	28.59	48.6	0.7	100	0.05
edu	49.94	28.33	51.4	0.7	100	0
empl	50.69	27.81	50	0.7	100	0.03
income (CAD)	119,085.1	102,561.2	93,591	23,076	2,009,153	6.27
highedu (%)	0.49	0.13	0.49	0.13	0.95	0.1
popdens	7,802.57	7,944.75	5,780	47.3	93,700.8	3.99
emp	362.11	331.4	275	50	5,980	6.03
nr	6.02	1.48	6.1	2	11.4	0.09
crowd (%)	0.04	0.05	0	0	0.43	2.19
condi (%)	0.07	0.05	0.06	0	0.42	0.9
hage	43.97	9.97	46.65	3.31	56	−1.57
intens_com (%)	0.04	0.06	0.01	0	0.46	2.69
intens_res (%)	0.51	0.16	0.54	0	0.84	−1.21
intens_gre (%)	0.1	0.14	0.05	0	1	2.53
ent	0.61	0.22	0.65	0	1	−1.03
hhi	0.53	0.23	0.5	0.11	1	0.33

higher employment rate in the less populated area and coincides with the income and education degree distribution.

Fig. 6 shows the housing transaction price and population density over the entire city. From the transaction records, we find that housing price peaks in the midtown area and downtown area and declines as it approaches the edge between urban and suburban region. The urban center region has higher housing demand both from investors and home-owners, therefore the market segment is a diverse combination of the high and low income, tenants, owners, and investors, which form a highly heterogeneous resident group. The higher prices in the central region results from the fact that better access to public facilities leads to a higher land price, and, consequently, higher housing prices. The midtown area running north–south through the center of the city along Yonge Street is the most expensive residential area distributed with several wealthy enclaves. Housing in the East York are generally less pricy. Therefore, housing price in City of Toronto forms a modified monocentric pattern over space, with a peak in the urban core and midtown along the north–south Yonge Street axis, and gradually declines "horizontally" to the east and west.

Comparison of the price distribution, population density, and residential price is strongly correlated with density. North-middle and southwest parts of the city are characterized as expensive put less populous area; northwest and the entire eastern portions of the city have generally lower housing prices with moderate population density; and high-density central downtown area has mixed levels of housing prices.

We also examine the spatial autocorrelation in the housing unit price based on each transaction. The global Moran's I is 0.53, which indicates a high spatial dependency. Including a spatial lag into the model could generate better fitting results. The clustering pattern from local Moran's I is consistent with the preceding discussion about the zonal features of housing price distribution. The midtown area along Yonge Street and the downtown area are mostly the "High-High" region that indicates high-price zones. The west and eastern Toronto areas are "Low-Low" regions, indicating the low-price zones.

OLSQ Regression Results

While we expect spatial autocorrection to be important in explaining housing prices, we begin by estimating an ordinary least square (OLSQ) model as base against which other models can be compared. This model achieved an adjusted R^2 goodness of fit of 0.53 (Table 4). In the Distance dimension, all the variables significantly influence the housing price, except for the calculated road network accessibility index. The distribution of road network accessibility does not spatially differentiate much throughout the city. Distance to bus stops contributes very little in comparison with distance to subway stations. It is likely that residents, especially in suburban areas, favor the auto instead of attaching value to the proximity to a bus, even with wider coverage of bus stops. The factors that the model include in distance and density aspects did not show significant influence on housing price in terms of the magnitude of coefficients or significance.

However, almost all Socioeconomic Environment variables show the expected signs of coefficients and influence. This indicates that the conventional built form framework does not account for much in the housing price, whereas the social environment is strongly influential. The demographically diverse and safe communities covered by health care and educational provision are more valued, and a high education degree (percentage of residents with postsecondary certificates) shows a significant effect on housing price. The highly positive coefficient of the factor "high education"

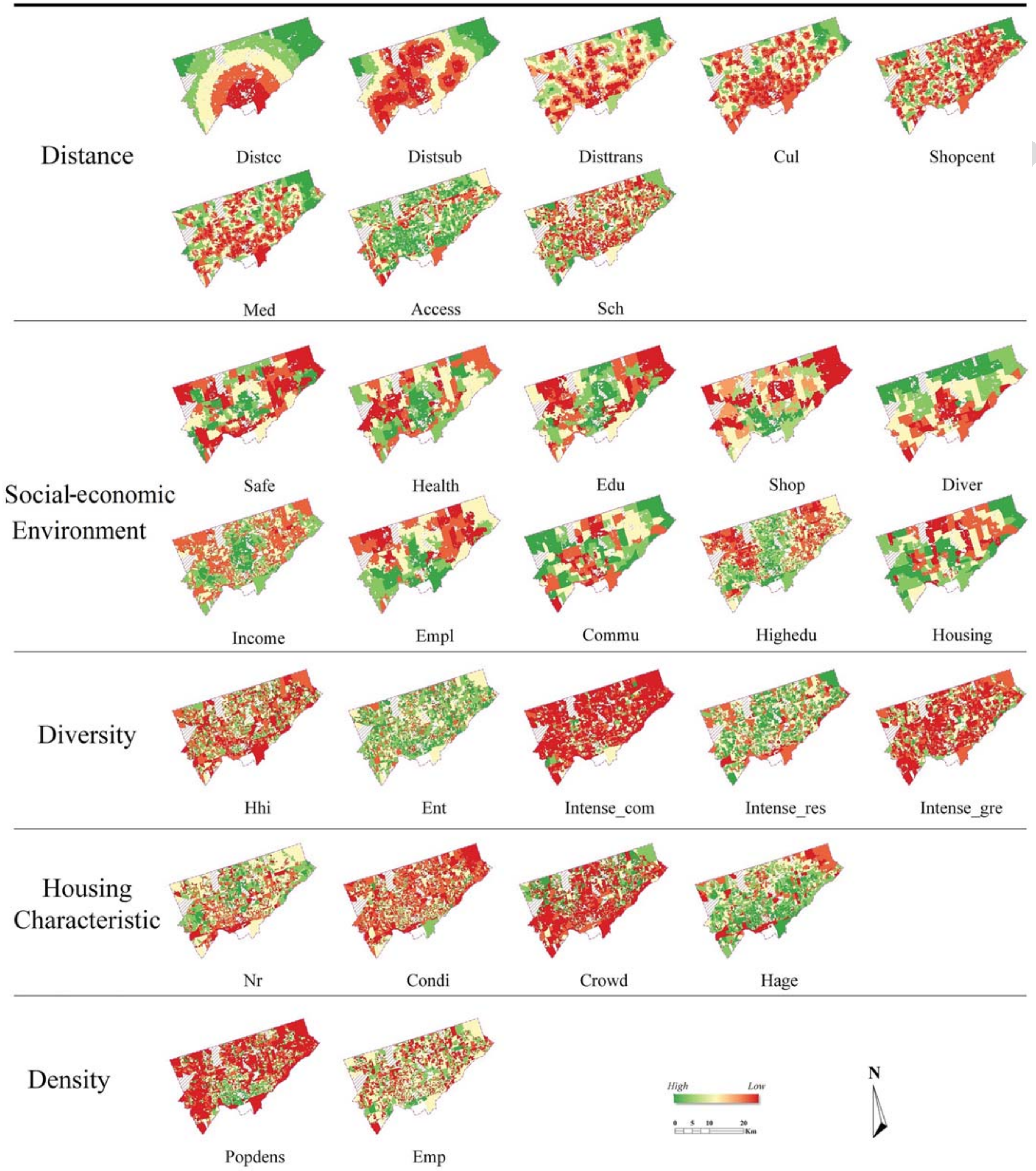


Fig. 5. Spatial distribution of the explanatory variables in the five aspects in the model.

does not imply a single direction causal relationship, and it cannot be interpreted as higher housing price results from better educated residents living here. It is clear that people with better income and higher education are supposed to have better budgets for housing consumption, and housing prices would be higher where they live, but it might not be true to interpret it as vice versa. The relationship between educated residents and housing price is a mutual causality: residents with higher education choose the housing based on the physical characteristics and desirable neighbors within the same social group, and the residential clustering of high education people with relatively same housing preference made the housing expensive in the neighborhood.

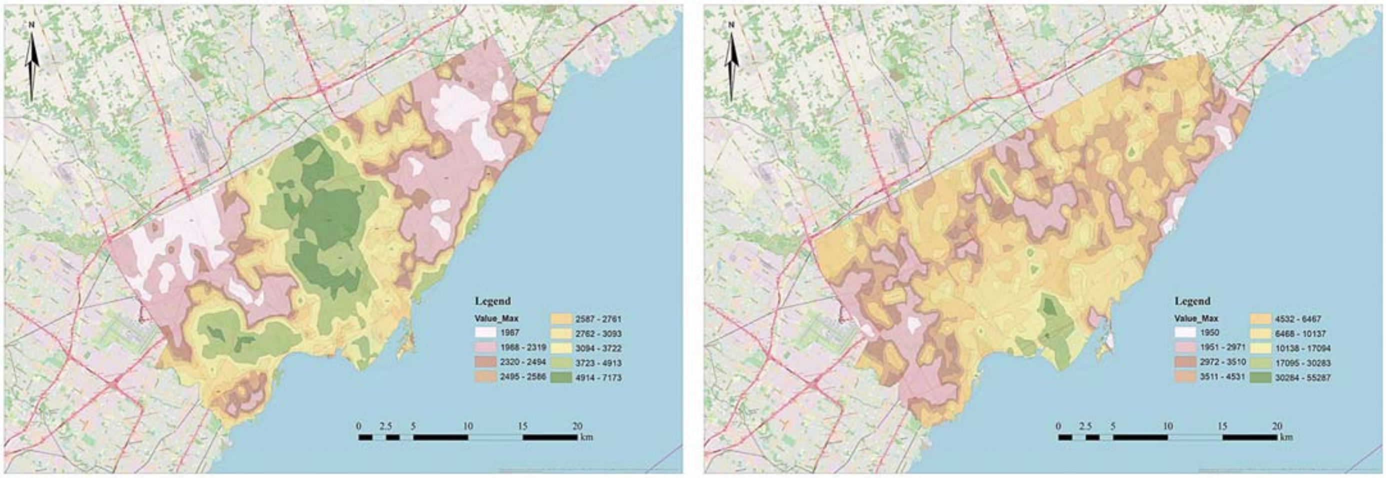


Fig. 6. Average housing unit price distribution and population density at City of Toronto.

Table 4. Regression results (OLSQ)

	Estimate	Std. Error	t	Sig.	F	Sig.	Expected sign	VIF
(Intercept)	0	0.012	0					
distcc	−0.125	0.033	−3.829	***	456.1	***	−	7.415
disttrans	0.001	0.016	0.046		125.2	***	−	1.800
distsb	−0.188	0.02	−9.277	***	535	***	−	2.853
med	−0.027	0.015	−1.795	.	66.97	***	−	1.608
cul	0.064	0.016	4.102	***	31.63	***	−	1.672
sch	0.018	0.013	1.41		4.33	*	−	1.185
shopcent	0.063	0.013	4.835	***	47.88	***	−	1.181
access	−0.018	0.013	−1.424		3.636	.	+	1.124
diver	0.089	0.025	3.587	***	454.8	***	+	4.244
housing	−0.081	0.018	−4.414	***	11.02	***	−	2.325
commu	−0.048	0.016	−2.905	**	78.73	***	+	1.886
safe	0.057	0.014	4.014	***	221.2	***	+	1.425
health	0.161	0.016	10.084	***	590.7	***	+	1.765
shop	−0.095	0.019	−4.972	***	9.743	**	+	2.523
edu	0.115	0.015	7.683	***	826.3	***	+	1.547
empl	0.198	0.023	8.595	***	1,062	***	+	3.706
income	0.158	0.017	9.072	***	968.5	***	+	2.103
highedu	0.22	0.017	12.871	***	1,331	***	+	2.026
popdens	0.038	0.017	2.207	*	3.258	.	+	2.042
emp	−0.068	0.016	−4.385	***	0.24		+	1.691
nr	0.15	0.022	6.77	***	146.8	***	+	3.415
crowd	0.037	0.015	2.383	*	270.4	***	−	1.637
condi	−0.006	0.013	−0.432		25.43	***	−	1.261
hage	−0.068	0.017	−3.875	***	1.045		−	2.116
ent	0.018	0.025	0.714		10.15	**	+	4.339
hhi	0.034	0.028	1.216		17.33	***	−	5.442
int_com	0.02	0.015	1.356		0.739		+	1.559
int_gre	0.046	0.017	1.168		12.18	***	+	1.617
int_res	0.018	0.015	2.743	**	44.21	***	+	1.917

Note: Multiple R^2: 0.5354, Adjusted R^2: 0.5312.

In the Diversity dimension, ENT and HHI are not significant in combination with other variables. The land use mix degree might not be valued in the same way under different circumstance among different group of people. Land use mix is valued in the low-income household with limited mobility, which provides better accessibility within walking distances. Yet for suburban areas without densely built commercial and business land use, relatively homogenous residential land use is valued for its serenity and safety preferred by some residents since they could afford a car. Their perceived utility gain outweigh the travel cost. Green land intensity and residential land intensity also significantly influence housing price, with the more green land distributed, the higher the housing price. Housing characteristics influence individual housing price as expected, as newly built housing with more rooms and less maintenance needs have higher housing price. The built form and social environments determine the basis for housing prices at the neighborhood level, and the individual housing price was differentiated based on physical characteristics.

Geographical Weighted Regression (GWR)

The spatial autocorrelated housing price could be better fitted with GWR and incorporation with proximity effects into the model could reduce the influence of spatial dependency. Owing to the high spatial autocorrelation indicated from Moran's I index, a geographical weighted regression was conducted, and the results are listed in Table 5. We use an AIC-minimized optimal bandwidth of 102 assuming the nearest 102 units in the neighborhood spatially correlated. The spatial kernel is set as adaptive bi-square. The adjusted R^2 improves to 0.79 implying that putting spatial relationship into consideration largely improves the goodness-of-fit of the model. The log-likelihood improves from the OLSQ value of −3,380 to −938 and the AIC reduces from 6,821 to 5,205, which indicates GWR as the better fitting model.

The coefficient summary (Table 6) also shows better result in the GWR model. For each sample, the GWR model generates a specific set of coefficients of each variable; in other words, the coefficient varies on each sample. Therefore, we look at the statistics of the coefficient to find the contribution of each variable. Compared with the result of OLSQ model, accessibility (Distance) still shows high influence on the housing price, the longer distance to the transportation network, the lower the price. Diversity and Density show slightly positive effects on housing price. Variables indicating physical characteristics show different results with OLSQ except for maintenance needs (condi). Housing with more rooms have a lower unit price, which coincides with the commonly found fact that condos (with smaller area) have higher unit price than detached houses. The GWR model takes into account the neighborhood effect, which can better manifest the contribution of the independent factors. The Socioeconomic aspect still explains a large part of the housing price in the GWR model. The coefficient of sense of community (commu) changes from negative to positive in the GWR model, which could better explain housing units in a friendly and connected neighborhood sold at a higher price.

The local R^2 is shown in Fig. 7. In general, the GWR model explains the housing price for each DA on average at about a 0.75 level, and housing prices of most areas are well fitted except for some regions in the downtown area. With higher heterogeneity, housing in the downtown area is differentiated based not only on the built form, but also on the more diverse socioeconomic environment. Owing to this heterogeneity, independent, district-specific GWR models might perform better than a citywide model, but this is left for future investigation.

Factors affecting housing price has been examined but limited to a diagnostic and explanation level. Further housing price simulation should take spatial and social heterogeneity into consideration in order to keep a homogeneous situation while applying the housing price modeling method. Before applying the ML method, we computed the confusion matrix and kappa index of GWR (Table 7) in order to make a comparison with the result of the random forests simulation, as discussed in the next section.

8 **Table 5.** Geographical weighted regression results

T5:1	Residual sum of squares	339.732
T5:2 9	Log-likelihood	−938.93
T5:3	AIC	5,205.233
T5:4	AICc	8,668.615
T5:5	BIC	15,338.26
T5:6	R^2	0.896
T5:7	Adj. R^2	0.788
T5:8	Adj. alpha (95%)	0.001

Table 6. Coefficient summary (GWR)

	Variable	Mean	STD	Min	Median	Max
T6:1	Variable	Mean	STD	Min	Median	Max
T6:2	(Intercept)	0.394	13.394	−325.567	−0.108	230.215
T6:3	distcc	−0.058	3.035	−14.711	−0.255	16.639
T6:4	disttrans	−0.008	0.721	−11.448	0.035	3.198
T6:5	distsb	−0.164	2.255	−8.875	−0.107	30.896
T6:6	med	0.01	0.3	−1.343	0.017	1.911
T6:7	cul	0.002	0.333	−1.102	−0.029	2.76
T6:8	sch	0.026	0.119	−0.382	0.022	0.546
T6:9	shopcent	0.002	0.191	−0.855	0.001	1.076
T6:10	access	−0.038	0.164	−0.919	−0.03	0.998
T6:11	safe	−0.051	5.055	−127.62	0.031	106.421
T6:12	housing	0.019	1.33	−20.615	0.093	29.437
T6:13	commu	0.024	0.915	−28.306	−0.047	10.277
T6:14	diver	0.021	2.748	−36.502	0.025	130.176
T6:15	health	0.245	12.793	−140.149	0.094	532.486
T6:16	shop	0.738	24.12	−158.453	0.037	1,160.827
T6:17	edu	0.349	16.939	−95.454	−0.03	811.274
T6:18	empl	1.187	32.137	−369.693	0.079	1,476.779
T6:19	income	0.306	0.535	−1.5	0.221	2.805
T6:20	highedu	−0.007	0.153	−0.713	−0.005	0.642
T6:21	popdens	−0.01	0.225	−0.977	−0.011	1.119
T6:22	emp	0.006	0.203	−1.108	0.001	1.249
T6:23	nr	−0.004	0.295	−1.357	0.034	0.784
T6:24	crowd	−0.01	0.182	−1.042	0.008	0.585
T6:25	condi	−0.024	0.109	−0.623	−0.02	0.485
T6:26	hage	0.001	0.21	−0.643	−0.01	0.749
T6:27	intens_com	0.027	0.156	−0.677	0.013	0.738
T6:28	intens_res	0.024	0.195	−0.738	0.005	0.877
T6:29	intens_gre	0.043	0.194	−0.733	0.024	0.958
T6:30	ent	0.024	0.228	−1.055	0.025	0.969
T6:31	hhi	0.06	0.251	−1.194	0.059	1.314

Random Forests Simulation

RF simulation was employed in modeling the housing price based on the variables analyzed in the previous section. In order to

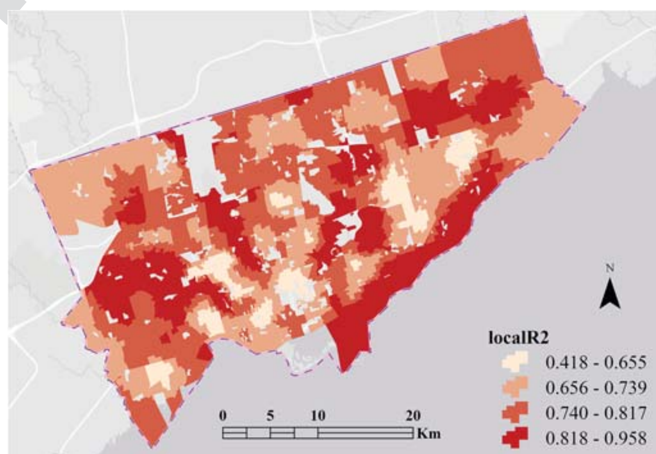


Fig. 7. Distribution of local R^2 from GWR model. F7:1

Table 7. Confusion matrix of GWR

	Reality (Unit: percent)						
GWR-regression (Unit: percent)	Very low	Low	Medium	High	Very high	Total	T7:2
Very low	**26.17**	2.73	1.43	0.51	0.12	30.96	T7:3
Low	3.08	**21.19**	2.73	1.07	0.99	29.07	T7:4
Medium	0.51	1.37	**13.71**	1.95	0.34	17.88	T7:5
High	0.02	0.33	1.69	**10.02**	0.86	12.92	T7:6
Very high	0.28	0.34	0.38	0.73	**7.44**	9.17	T7:7
Total	30.05	25.97	19.95	14.28	9.75	**100**	T7:8

Note: kappa = 0.721; OA = 0.785.

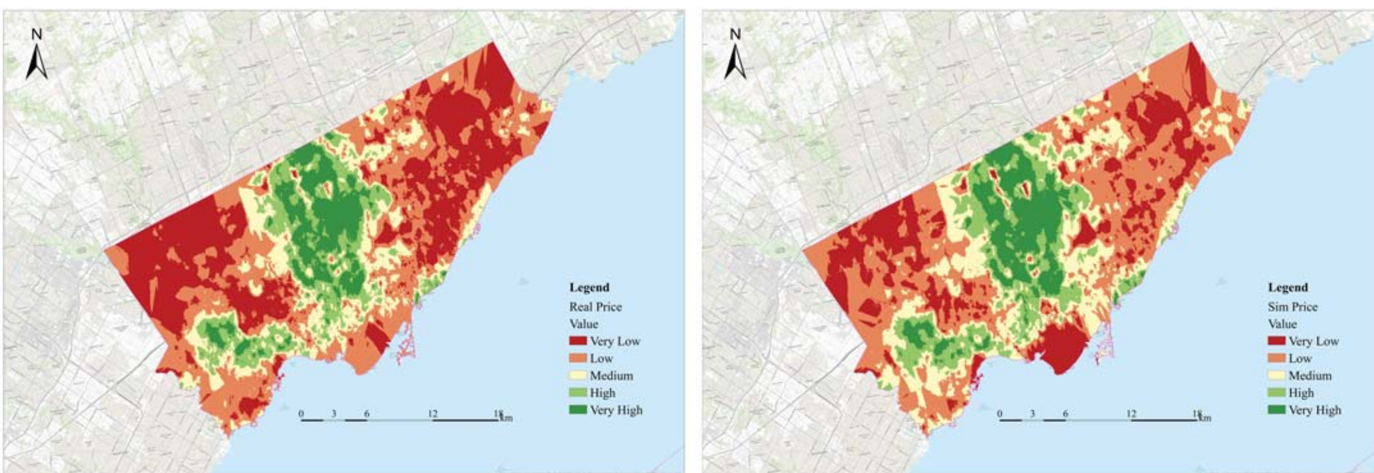
remove the impact of area difference, we project the dataset in ArcGIS and divided the entire region into 310,211 cells with 30 m × 30 m resolution. We divided the dataset into training data and validation data, and the shares were set to 40% and 60% respectively, to ensure the fitting accuracy and stability of this model. Eighty decision trees and 20% OOB data were established and we also cross-validated the model with bootstrap random sampling. The model achieved a kappa coefficient of 0.803, and an overall accuracy 0.849, which indicates a good predicting performance (see Table 8). The simulated housing price and real housing transaction price distribution are shown in Fig. 8. The simulated map follows the same distribution pattern as the real transaction one. As shown in Fig. 8, the RF algorithm tends to underpredict the housing price near the lakeshore region in the southern part, and overpredict the relatively very low-priced housing units in the western and eastern parts. As shown in the confusion matrix in Table 8, the RF model predicts better with "very low" and "very high" priced housing transactions, and not so accurately with medium-priced groups.

Table 8. Confusion matrix of random forest simulation

RF-simulation (Unit: percent)	Reality (Unit: percent) Very low	Low	Medium	High	Very high	Total
Very low	**27.59**	0.75	0.82	0.69	0.28	30.13
Low	2.92	**24.03**	1.25	0.36	0.43	28.99
Medium	0	0.67	**16.39**	1.35	0.64	19.05
High	0.19	0.56	1.74	**10.19**	0.69	13.38
Very high	0.01	0.26	0.39	1.05	**6.74**	8.45
Total	30.72	26.27	20.59	13.64	8.78	**100**

Note: kappa = 0.803; OA = 0.849.

Fig. 9 shows the contribution of each variable in predicting housing prices. Socioeconomic Environment and Distance dimensions have the highest contributions in the simulation. Household income, percentage of high education degree, overall health coverage, and housing affordability in the neighborhood each contribute higher than 4% in the simulation, which indicates the profound importance of the socioeconomic environment in predicting housing prices. In terms of the casual logic direction, the aggregation of the group of people with similar demographic characteristics is both a result from the "pulling force" of certain location and a self-reinforcement factor for more residents carrying similar background to gather there. Density does not necessarily contribute much to explaining housing prices. The coefficients of population density and employment density are not significantly different from zero as in the contribution matrix of RF estimation, and only make a difference in housing price in the downtown and midtown areas. The Distance dimension also makes a strong contribution, with about a 15% contribution from distance to the city center and 4% from distance to the subway station. Diversity shows limited



(a) (b)

Fig. 8. (a) The real housing transaction price; and (b) the simulated housing price from RF.

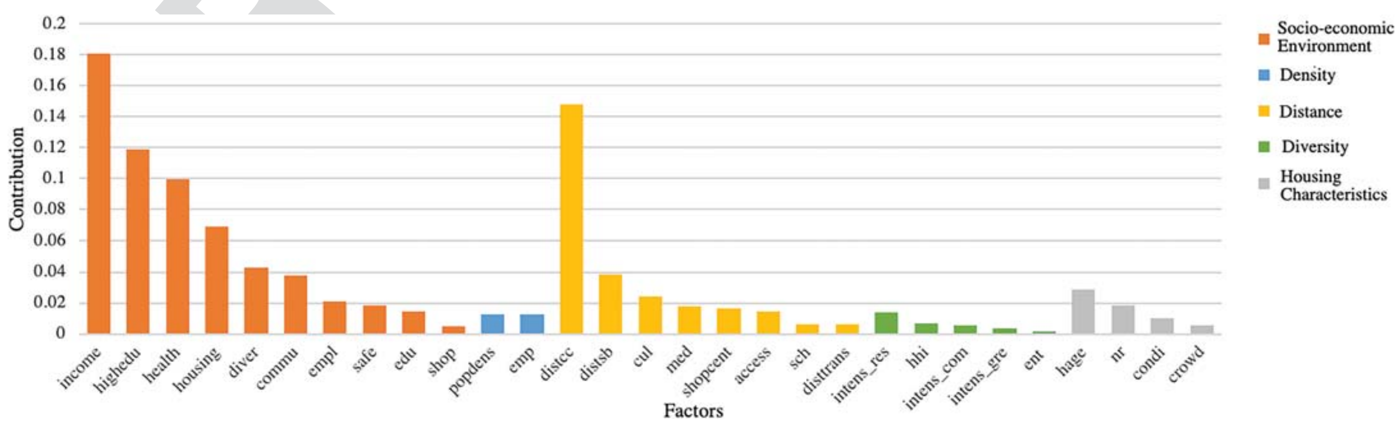


Fig. 9. Contribution of each variable in the five dimensions in predicting housing price.

impact on housing price, which coincides with the regression results. The influences of diversity under heterogeneous circumstance have separate paths and this aspect cannot be interpreted as unrelated to housing price. Housing characteristics also contributes around 3%, with housing age as the most influential factor.

In comparison with the GWR model (Tables 7 and 8), the RF model produces better kappa overall accuracy. In modeling the high-priced housing units, the two models have similar predictive performance. However, in the "very low," "low," and "medium" groups, the RF model has less percentage error and gives significantly better prediction than GWR. The RF model trained in this study presents a method to predict the housing price from the three parts, built form, socioeconomic environment, and the individual housing characteristics, with a simulation accuracy of around 85%, that could provide a reference for researchers and practitioners in housing price modeling.

Conclusion and Discussion

Housing price modeling has long been the focus of developers, housing planning administration departments, and the real-estate finance field. In this study, we construct housing price models based on a theoretical framework of built form, socioeconomic environment, and physical condition attributes. High spatial autocorrelation influences housing prices, and the externalities of housing should be taken into account in housing price modeling. Given this, a GWR model and a RF model were built to make the model more useful not only for diagnostic analysis, but also for explanation and simulation. Our study shows that the conventional 5D built environment is not the major contributor for housing price determination; rather the socioeconomic environment has much stronger explanatory power. In constructing the housing price, it is argued in this study that housing price consists of two components: the regional residential land price determined by the built form and socioeconomic environment, and the cost of the individual housing unit as differentiated by its physical features. In the case of the City of Toronto, housing price is primarily determined by the social environment and the distance or accessibility of the neighborhood, and the housing physical condition, especially the house age. The density and diversity of the surroundings show relatively little impact on housing prices.

We consider the housing price model developed in this paper to be applicable to other cities with relatively comparable population and economic characteristics to that of the City of Toronto. Built form, socioeconomic environment, and physical housing features should determine the fixed predictable part of housing price, while other factors, such as market regulations and special appreciation of individual housing units could also affect final transaction prices. Basic trend analysis and field investigation will facilitate model adjustment when applying it to other cases. The model could serve as a planning tool for estimating potential market response to the changes in built environment, simulating housing price variation and a logical basis for modeling housing markets in more comprehensive urban modeling systems.

There are several limitations to this study, including the following. The framework for housing prices was built from the demand side in this study, without comprehensive consideration of the supply side and the policy impact on the macro level, which is inconsistent with the real housing market with multiple interactions among different agents. Further research could investigate the formation mechanism of housing price as a result of the interplay process of multiple agents. Second, the land use mix index computed in this study through ENT and HHI did not show an expected significant influence on housing prices. It is assumed that the relationship between land use mix degree and housing price was not adequately captured by these measures, and that further studies should experiment more on diverse land use mix indices at different levels of land use type division. The framework and the model presented in this study could be employed as the basis of urban simulation including land use, housing, transportation, and human activities. The housing price volatility could be analyzed through examination of the available time series data to include not only the spatial lag, but also a temporal lag. This will be the next step in the research. With more advanced data collection methods currently available, housing price monitoring could be combined with residents' travel and daily activity behavior, which could help us better understand the function of housing in fulfilling residents' needs.

Data Availability Statement

The demographic census data, neighborhood scores, computation of accessibility and diversity indexes, some part of the locational data, and the regression model that support the findings of this study are available from the corresponding author upon reasonable request. The Teranet housing transaction data used during the study are proprietary or confidential in nature and may only be provided with restrictions.

References

Anderson, W. P., P. S. Kanaroglou, and E. J. Miller. 1996. "Urban form, energy and the environment: A review of issues, evidence and policy." *Urban Stud.* 33 (1): 7–35. https://doi.org/10.1080/00420989650012095.

Bailey, M. J., R. F. Muth, and H. O. Nourse. 1963. "A regression method for real estate price index construction." *J. Am. Stat. Assoc.* 58 (304): 933–942. https://doi.org/10.1080/01621459.1963.10480679.

Balcilar, M., R. Gupta, and S. M. Miller. 2015. "The out-of-sample forecasting performance of nonlinear models of regional housing prices in the US." *Appl. Econ.* 47 (22): 2259–2277. https://doi.org/10.1080/00036846.2015.1005814.

Belgiu, M., and L. Drăguţ. 2016. "Random forest in remote sensing: A review of applications and future directions." *ISPRS J. Photogramm. Remote Sens.* 114: 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Bitter, C., G. F. Mulligan, and S. Dall'erba. 2007. "Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method." *J. Geog. Syst.* 9 (1): 7–27. https://doi.org/10.1007/s10109-006-0028-7.

Bourassa, S. C., M. Hoesli, and J. Sun. 2006. "A simple alternative house price index method." *J. Housing Econ.* 15 (1): 80–97. https://doi.org/10.1016/j.jhe.2006.03.001.

Bowen, W. M., B. A. Mikelbank, and D. M. Prestegaard. 2001. "Theoretical and empirical considerations regarding space in hedonic housing price model applications." *Growth Change* 32 (4): 466–490. https://doi.org/10.1111/0017-4815.00171.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. 2002. "Geographically weighted summary statistics—a framework for localised exploratory data analysis." *Comput. Environ. Urban Syst.* 26 (6): 501–524. https://doi.org/10.1016/S0198-9715(01)00009-6.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. "Geographically weighted regression: A method for exploring spatial nonstationarity." *Geog. Anal.* 28 (4): 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Burt, M. R., L. Y. Aron, and E. Lee. 2001. *Helping America's homeless: Emergency shelter or affordable housing?* Washington, DC: The Urban Institute.

Can, A. 1992. "Specification and estimation of hedonic housing price models." *Reg. Sci. Urban Econ.* 22 (3): 453–474. https://doi.org/10.1016/0166-0462(92)90039-4.

Cao, K., M. Diao, and B. Wu. 2019. "A big data–based geographically weighted regression model for public housing prices: A case study in Singapore." *Ann. Am. Assoc. Geogr.* 109 (1): 173–186. https://doi.org/10.1080/24694452.2018.1470925.

Case, B., H. O. Pollakowski, and S. M. Wachter. 1991. "On choosing among house price index methodologies." *Real Estate Econ.* 19 (3): 286–307. https://doi.org/10.1111/1540-6229.00554.

Cervero, R., and K. Kockelman. 1997. "Travel demand and the 3Ds: Density, diversity, and design." *Transp. Res. Part D: Transp. Environ.* 2 (3): 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6.

Cervero, R., O. L. Sarmiento, E. Jacoby, L. F. Gomez, and A. Neiman. 2009. "Influences of built environments on walking and cycling: Lessons from Bogotá." *Int. J. Sustainable Transp.* 3 (4): 203–226. https://doi.org/10.1080/15568310802178314.

Chau, K. W., and T. L. Chin. 2003. "A critical review of literature on the hedonic price model." *Int. J. Hous. Sci. Appl.* 27 (2): 145–165.

Chen, Y., X. Liu, X. Li, Y. Liu, and X. Xu. 2016. "Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning." *Appl. Geogr.* 75: 200–212. https://doi.org/10.1016/j.apgeog.2016.08.011.

Clapp, J. M., and C. Giaccotto. 1992. "Estimating price indices for residential property: A comparison of repeat sales and assessed value methods." *J. Am. Stat. Assoc.* 87 (418): 300–306. https://doi.org/10.1080/01621459.1992.10475209.

Cohen, J. P., and C. C. Coughlin. 2008. "Spatial hedonic models of airport noise, proximity, and housing prices." *J. Reg. Sci.* 48 (5): 859–878. https://doi.org/10.1111/j.1467-9787.2008.00569.x.

De La Barra, T., B. Pérez, and N. Vera. 1984. "TRANUS-J: Putting large models into small computers." *Environ. Plann. B: Plann. Des.* 11 (1): 87–101. https://doi.org/10.1068/b110087.

Diamond, R., and T. McQuade. 2019. "Who wants affordable housing in their backyard? An equilibrium analysis of low-income property development." *J. Political Econ.* 127 (3): 1063–1117. https://doi.org/10.1086/701354.

Dubin, R., R. K. Pace, and T. G. Thibodeau. 1999. "Spatial autoregression techniques for real estate data." *J. Real Estate Lit.* 7 (1): 79–96. https://doi.org/10.1023/A:1008690521599.

Echenique, M. H., A. D. J. Flowerdew, J. D. Hunt, T. R. Mayo, I. J. Skidmore, and D. C. Simmonds. 1990. "The MEPLAN models of Bilbao, Leeds and Dortmund." *Transp. Rev.* 10 (4): 309–322. https://doi.org/10.1080/01441649008716764.

Englund, P., J. M. Quigley, and C. L. Redfearn. 1999. "The choice of methodology for computing housing price indexes: Comparisons of temporal aggregation and sample definition." *J. Real Estate Finance Econ.* 19 (2): 91–112. https://doi.org/10.1023/A:1007846404582.

Farber, S., and A. Páez. 2007. "A systematic investigation of cross-validation in GWR model estimation: Empirical analysis and Monte Carlo simulations." *J. Geog. Syst.* 9 (4): 371–396. https://doi.org/10.1007/s10109-007-0051-3.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim. 2014. "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.* 15 (1): 3133–3181.

Goodman, A. C. 1978. "Hedonic prices, price indices and housing markets." *J. Urban Econ.* 5 (4): 471–484. https://doi.org/10.1016/0094-1190(78)90004-9.

Goodman, A. C. 1988. "An econometric model of housing price, permanent income, tenure choice, and housing demand." *J. Urban Econ.* 23 (3): 327–353. https://doi.org/10.1016/0094-1190(88)90022-8.

Gu, J., M. Zhu, and L. Jiang. 2011. "Housing price forecasting based on genetic algorithm and support vector machine." *Expert Syst. Appl.* 38 (4): 3383–3386. https://doi.org/10.1016/j.eswa.2010.08.123.

Habib, M. A., and E. J. Miller. 2008. "Influence of transportation access and market dynamics on property values: Multilevel spatiotemporal models of housing price." *Transp. Res. Rec.* 2076 (1): 183–191. https://doi.org/10.3141/2076-20.

Haider, M., and E. J. Miller. 2000. "Effects of transportation infrastructure and location on residential real estate values: Application of spatial autoregressive techniques." *Transp. Res. Rec.* 1722 (1): 1–8. https://doi.org/10.3141/1722-01.

Hu, L., S. He, Z. Han, H. Xiao, S. Su, M. Weng, and Z. Cai. 2019. "Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies." *Land Use Policy* 82: 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030.

Huang, B., B. Wu, and M. Barry. 2010. "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices." *Int. J. Geog. Inf. Sci.* 24 (3): 383–401. https://doi.org/10.1080/13658810802672469.

Hunt, J. D. 2003. "Design and application of the pecas land use modelling system."

Ihlanfeldt, K. R. 2007. "The effect of land use regulation on housing and land prices." *J. Urban Econ.* 61 (3): 420–435. https://doi.org/10.1016/j.jue.2006.09.003.

Kamusoko, C., and J. Gamba. 2015. "Simulating urban growth using a random forest-cellular automata (RF-CA) model." *ISPRS Int. J. Geo-Inf.* 4 (2): 447–470. https://doi.org/10.3390/ijgi4020447.

Kim, H.-G., K.-C. Hung, and S. Y. Park. 2015. "Determinants of housing prices in Hong Kong: A Box-Cox quantile regression approach." *J. Real Estate Finance Econ.* 50 (2): 270–287. https://doi.org/10.1007/s11146-014-9456-1.

Koenker, R., and K. F. Hallock. 2001. "Quantile regression." *J. Econ. Perspect.* 15 (4): 143–156. https://doi.org/10.1257/jep.15.4.143.

Kouwenberg, R., and R. Zwinkels. 2014. "Forecasting the US housing market." *Int. J. Forecasting* 30 (3): 415–425. https://doi.org/10.1016/j.ijforecast.2013.12.010.

Landis, J. D. 1994. "The California Urban Futures Model: A new generation of metropolitan simulation models." *Environ. Plann. B: Plann. Des.* 21 (4): 399–420. https://doi.org/10.1068/b210399.

Levine, J. 1998. "Rethinking accessibility and jobs-housing balance." *J. Am. Plann. Assoc.* 64 (2): 133–149. https://doi.org/10.1080/01944369808975972.

Malpezzi, S., G. H. Chun, and R. K. Green. 1998. "New place-to-place housing price indexes for US Metropolitan Areas, and their determinants." *Real Estate Econ.* 26 (2): 235–274. https://doi.org/10.1111/1540-6229.00745.

Mark, J. H., and M. A. Goldberg. 1984. "Alternative housing price indices: An evaluation." *Real Estate Econ.* 12 (1): 30–49. https://doi.org/10.1111/1540-6229.00309.

Martinez, F. 1996. "MUSSA: Land use model for Santiago city." *Transp. Res. Rec.* 1552 (1): 126–134. https://doi.org/10.1177/0361198196155200118.

Mason, C., and J. M. Quigley. 1996. "Non-parametric hedonic housing prices." *Housing Stud.* 11 (3): 373–385. https://doi.org/10.1080/02673039608720863.

Massey, D. S., and J. S. Rugh. 2017. "Zoning, affordable housing, and segregation in US metropolitan areas." In Chap. 14 in *The fight for fair housing: Causes, consequences, and future implications of the 1968 federal fair housing act*, edited by G. D. Squires, 245–264. London: Routledge.

McMillen, D. 2013. "Local quantile house price indices." In *Univ. of Illinois, mimeo*, 1–41.

Mok, H. M., P. P. Chan, and Y.-S. Cho. 1995. "A hedonic price model for private properties in Hong Kong." *J. Real Estate Finance Econ.* 10 (1): 37–48. https://doi.org/10.1007/BF01099610.

Morris, A. C., H. R. Neill, and N. E. Coulson. 2020. "Housing supply elasticity, gasoline prices, and residential property values." *J. Housing Econ.* 48: 101669. https://doi.org/10.1016/j.jhe.2020.101669.

Nguyen, M. T. 2005. "Does affordable housing detrimentally affect property values? A review of the literature." *J. Plann. Lit.* 20 (1): 15–26. https://doi.org/10.1177/0885412205277069.

Oladunni, T., and S. Sharma. 2016. "Hedonic housing theory—A machine learning investigation." In *Proc., 2016 15th IEEE Int. Conf. on Machine Learning and Applications*, 522–527. New York: IEEE.

Osland, L., and I. Thorsen. 2008. "Effects on housing prices of urban attraction and labor-market accessibility." *Environ. Plann. A: Econ. Space* 40 (10): 2490–2509. https://doi.org/10.1068/a39305.

Palczewska, A., J. Palczewski, R. M. Robinson, and D. Neagu. 2014. "Interpreting random forest classification models using a feature contribution method." In *Integration of reusable systems*, T. Bouabana-Tebibel, and S. H. Rubin, 193–218. Cham, Switzerland: Springer.

Park, B., and J. K. Bae. 2015. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." *Expert Syst. Appl.* 42 (6): 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040.

Quigley, J. M. 1994. "A simple hybrid model for estimating real estate price indexes." *J. Housing Econ.* 4 (1): 1–12. https://doi.org/10.1006/jhec.1995.1001.

Rafiei, M. H., and H. Adeli. 2016. "A novel machine learning model for estimation of sale prices of real estate units." *J. Constr. Eng. Manage.* 142 (2): 04015066. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001047.

Rosen, S. 1974. "Hedonic prices and implicit markets: Product differentiation in pure competition." *J. Political Econ.* 82 (1): 34–55. https://doi.org/10.1086/260169.

Selim, H. 2009. "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network." *Expert Syst. Appl.* 36 (2): 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044.

Song, Y., and G.-J. Knaap. 2004. "Measuring the effects of mixed land uses on housing values." *Reg. Sci. Urban Econ.* 34 (6): 663–680. https://doi.org/10.1016/j.regsciurbeco.2004.02.003.

Statistics Canada. 2012. "Toronto, Ontario (Code 3520005) and Ontario (Code 35) (table). Census Profile." 2011 Census. Statistics Canada Catalogue no. 98-316-XWE. Ottawa. Released October 24, 2012. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E.

Statistics Canada. 2018. *Survey of household spending, 2017*. Ottawa: Statistics Canada.

Tong, D., Y. Zhang, I. MacLachlan, and G. Li. 2019. "Migrant housing choices from a social capital perspective: The case of Shenzhen, China." *Habitat Int.* 96: 102082. https://doi.org/10.1016/j.habitatint.2019.102082.

Toronto Life. 2018. "The ultimate neighbourhood rankings."

Treyz, G. I. 1995. "Policy analysis applications of REMI economic forecasting and simulation models." *Int. J. Public Admin.* 18 (1): 13–42. https://doi.org/10.1080/01900699508524997.

Waddell, P. 2002. "Urbansim: Modeling urban development for land use, transportation, and environmental planning." *J. Am. Plann. Assoc.* 68 (3): 297–314. https://doi.org/10.1080/01944360208976274.

Wallace, N. E., and R. A. Meese. 1997. "The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches." *J. Real Estate Finance Econ.* 14 (1–2): 51–73. https://doi.org/10.1023/A:1007715917198.

Wang, S., S. Zheng, and J. Feng. 2007. "Spatial accessibility of housing to public services and its impact on housing price: A case study of Beijing's inner city." *Prog. Geogr.* 26 (6): 78–85.

Witte, A. D., H. J. Sumka, and H. Erekson. 1979. "An estimate of a structural hedonic price model of the housing market: An application of Rosen's theory of implicit markets." *Econometrica* 47 (5): 1151–1173. https://doi.org/10.2307/1911956.

Wyner, A. J., M. Olson, J. Bleich, and D. Mease. 2017. "Explaining the success of adaboost and random forests as interpolating classifiers." *J. Mach. Learn. Res.* 18 (1): 1558–1590.

Xie, X., and G. Hu. 2007. "A comparison of Shanghai housing price index forecasting." In *Proc., 3rd Int. Conf. on Natural Computation*, 221–225. New York: IEEE.

Yan, Y., X. Wei, B. Hui, S. Yang, W. Zhang, Y. Hong, and S.-y. Wang. 2007. "Method for housing price forecasting based on TEI@I methodology." *Syst. Eng. Theory Pract.* 27 (7): 1–9. https://doi.org/10.1016/S1874-8651(08)60047-2.

Yang, C., Q. Zhan, Y. Lv, and H. Liu. 2019. "Downscaling land surface temperature using multiscale geographically weighted regression over heterogeneous landscapes in Wuhan, China." *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12 (12): 5213–5222. https://doi.org/10.1109/JSTARS.2019.2955551.

Yao, Y., X. Liu, X. Li, J. Zhang, Z. Liang, K. Mai, and Y. Zhang. 2017. "Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data." *Int. J. Geog. Inf. Sci.* 31 (6): 1220–1244.

Zhang, D., X. Liu, X. Wu, Y. Yao, X. Wu, and Y. Chen. 2019. "Multiple intra-urban land use simulations and driving factors analysis: A case study in Huicheng, China." *GISci. Remote Sens.* 56 (2): 282–308. https://doi.org/10.1080/15481603.2018.1507074.

Ziemke, D., K. Nagel, and R. Moeckel. 2016. "Towards an agent-based, integrated land-use transport modeling system." *Procedia Comput. Sci.* 83: 958–963. https://doi.org/10.1016/j.procs.2016.04.192.

Zietz, J., E. N. Zietz, and G. S. Sirmans. 2008. "Determinants of house prices: A quantile regression approach." *J. Real Estate Finance Econ.* 37 (4): 317–333. https://doi.org/10.1007/s11146-007-9053-7.

Queries

1. Please provide the ASCE Membership Grades for all authors who are members.

2. Please provide a city and postal code for all author affiliations.

3. There were two equations numbered "4" in this paper; hence, we have changed the repeated number to Eq. (5) and renumbered all subsequent equations accordingly. Please check all renumbering and update the citations in the text, if needed.

4. Please provide the name and location of the publisher of the proceedings for the reference "McMillen (2013)." If there is no publisher, please provide the name and location of the sponsor of the conference. For sponsors that are virtual groups (without a physical location), include the conference location instead of sponsor location and the URL for the group's website.

5. Are the source details for Fig 2 correct? The list is very long and several sources seem to be duplicated.

6. Please add a column header to the first column in Tables 3 and 4.

7. Please supply key to explain purpose of asterisks in Table 4.

8. Please Add a column header in Table 5.

9. Please confirm if the Table 5 format is okay here.