1-2019

A big data–based geographically weighted regression model for public housing prices: A case study in Singapore

Kai CAO
Singapore Management University, kaicao@smu.edu.sg

Mi DIAO

Bo WU

A Big Data–Based Geographically Weighted Regression Model for Public Housing Prices: A Case Study in Singapore

Kai Cao,* Mi Diao,† and Bo Wu‡

*Department of Geography, National University of Singapore
†Department of Real Estate, National University of Singapore
‡Department of Geography and Environment, Jiangxi Normal University

In this research, three hedonic pricing models, including an ordinary least squares (OLS) model, a Euclidean distance–based (ED-based) geographically weighted regression (GWR) model, and a travel time–based GWR model supported by a big data set of millions of smartcard transactions, have been developed to investigate the spatial variation of Housing Development Board (HDB) public housing resale prices in Singapore. The results help identify factors that could significantly affect public housing resale prices, including the age and the floor area of the housing units, the distance to the nearest park, the distance to the central business district (CBD), and the distance to the nearest Mass Rapid Transit (MRT) station. The comparison of the three models also explicitly shows that the two GWR models perform much better than the traditional linear hedonic regression model, given the identical variables and data used in the calibration. Furthermore, the travel time–based GWR model has better model fit compared to the ED-based GWR model in the case study. This study demonstrates the potential value of the big data–based GWR model in housing research. It could also be applied to other research fields such as public health and criminal justice. *Key Words: big data, GWR, Housing Development Board (HDB), hedonic pricing model, Singapore.*

本研究建立三个特徵价格模型来探讨新加坡建屋发展局（HDB）的公屋再销售价格之空间变异，这三个模型包括普通最小二乘（OLS）模型,根据欧式距离（ED）的地理加权迴归（GWR）模型，以及由一组包含百万笔智能卡交易的大数据集所支持的根据旅行时间的 GWR 模型。研究结果有助于指认可能显着影响公屋再销售价格的因素，包含屋龄及住房单位的面积,与最近公园的距离,与中央商业区（CBD）的距离，以及与最近的捷运站（MRT）之距离。三个模型的比较同时明白显示出，在校正中使用相同的变因与数据之下，两大GWR模型较传统线性特徵价格模型表现更佳。再者，在案例研究中，根据旅行时间的 GWR 相较于根据ED的 GWR 模型而言，具有更佳的模型契合度。本研究证实根据 GWR 模型的大数据在住宅研究中的潜在价值。该数据同时可应用于诸如公共健康与犯罪正义等其他研究领域。 *关键词：大数据, GWR, 建屋发展局 (HDB),特徵价格模型,新加坡。*

En esta investigación se han desarrollado tres modelos hedónicos de determinación del precio, que incluyen un modelo ordinario de mínimos cuadrados (MCO), el modelo de regresión geográficamente ponderada (GWR) basado en distancia euclidiana (basado en ED) y un modelo de GWR basado en tiempo de viaje apoyado en un conjunto de *big data* de millones de transacciones de tarjetas inteligentes, para investigar la variación espacial de los precios de reventa de vivienda de la Junta para el Desarrollo de la Vivienda (HDB) en Singapur. Los resultados ayudan a identificar los factores que podrían afectar significativamente los precios de reventa de vivienda, incluso la antigüedad y el área del piso de las unidades habitacionales, la distancia al parque más cercano, la distancia al distrito central de negocios (CBD) y la distancia a la estación más cercana del Transporte Masivo Rápido (MRT). La comparación de los tres modelos muestra también de manera explícita que los dos modelos de la GWR se desempeñan mucho mejor que el tradicional modelo de regresión linear hedónica, dados las idénticas variables y datos usados en la calibración. Además, el modelo de la GWR basado en tiempo de viaje encaja mucho mejor en su función al compararlo con el modelo de GWR del estudio de caso basado en ED. Este estudio demuestra el valor potencial del modelo de la GWR con base en big data para investigación de la vivienda. Podría también ser aplicado en otros campos de investigación tales como la salud pública y la justicia criminal. *Palabras clave: big data, GWR, Junta para el Desarrollo de la Vivienda (HDB), modelo hedónico de precios, Singapur.*

Annals of the American Association of Geographers, 109(1) 2019, pp. 173–186
Initial submission, November 2017; revised submission, January 2018; final acceptance, February 2018
Published by Taylor & Francis, LLC.

There has been a tremendous amount of interest in modeling housing prices among economists, planners, and policymakers in recent decades due to the significant role of properties in household wealth and national economy. The hedonic pricing model proposed by Rosen (1974) has been widely used by researchers to study housing prices (Blomquist and Worley 1981; Sheppard, 1999; Cebula 2009; Diao and Ferreira 2010; Geng et al. 2011; Shimizu 2014; Schläpfer et al. 2015; Diao, Qin, and Sing 2016; Chen and Li 2017; de Araujo and Cheng 2017; Diao, Fan, and Sing 2017; Diao, Zhu, and Zhu 2017; Lai et al. 2017; Wen, Xiao, and Zhang 2017). Hedonic pricing models treat a property as a composite good with a variety of attributes and provide an effective way to estimate the price of different housing characteristics, including structural attributes and locational amenities. In practice, hedonic pricing models are often calibrated using an ordinary least squares (OLS) estimator, with the assumption that the observations are independent and identically distributed, which means that there is no spatial autocorrelation among the observations. This assumption might not be satisfied in realistic cases, however, especially when spatial attributes are involved and spatial heterogeneity exists in geographic data sets. With the existence of spatial autocorrelation, the OLS estimation of hedonic pricing models could lead to biased, inconsistent, or inefficient results (Anselin 1998).

The advancement of spatial econometrics and spatial statistics (Cliff and Ord 1981; Upton and Fingleton 1985; Anselin and Griffith 1988) has led to many mathematical tools, such as spatial error/lag models and geographically weighted regression (GWR), to address spatial autocorrelation in hedonic pricing modeling, that have had tremendous success (Basu and Thibodeau 1998; Dubin 1998; Kelejian and Prucha 1998; Bowen, Mikelbank, and Prestegaard 2001; Fotheringham, Brunsdon, and Charlton 2002; Militino, Ugarte, and Garcia-Reinaldos 2004; Diao 2015). For instance, Yu, Wei, and Wu (2007) investigated the spatial dimension of housing market dynamics in Milwaukee by modeling the determinants of housing prices based on GWR models. Jim and Chen (2007) employed hedonic pricing models to study house-buyers' preference on outdoor environmental quality in Guangzhou, China. Debrezion, Pels, and Rietveld (2011) successfully employed a hedonic pricing model on three metropolitan areas in The Netherlands, including Amsterdam, Rotterdam, and Enschede, to analyze the contribution of railway accessibility to housing prices. Diao, Leonard, and Sing (2017) adopted a spatial difference-in-differences hedonic pricing model to investigate the impact of a new subway line on private housing prices in Singapore.

Most previous studies defined spatial relationships between observations based on Euclidean distance (ED), which is a good fit for geographic phenomena that only relate to ED, such as geological conditions or water pollution. For more complex housing price distribution, however, ED might not be able to well represent the spatial relationships between properties, whereas other cost distance measures, such as network distance and travel time, might perform better. Lu et al. (2014) proposed a GWR model with a non-ED metric based on both network distance and travel time and successfully conducted a case study on housing prices in London. How to define the cost distance between observations for better modeling of housing prices remains a challenge to researchers, though, given that most existing studies only considered network distance or estimated travel time from travel survey data sets. Travel surveys are the traditional data sources for transportation studies, which typically contain detailed information on all trips made by a few thousand respondents on the survey day, such as trip origin, trip destination, starting time, ending time, and mode of transport. The usefulness of travel surveys in generating a reliable travel time matrix between locations in the city is constrained by the small sample size and limited spatial and temporal coverage.

Emerging big data on individual mobility, such as transaction records of smartcards, along with big data analytical techniques, provide a unique opportunity for researchers to understand urban spatial structure and individual mobility patterns. In this research, we aim to explore the potential value of smartcard data sets in supporting spatial hedonic pricing modeling of housing prices using Singapore's public housing resale market as an example. We develop a travel time–based spatial weight matrix based on the movements of millions of travelers as revealed by their smartcard transactions in one week and integrate the spatial weight matrix into a GWR model of the resale prices of public housing units in Singapore. The research findings can help us understand the roles of various structural attributes and locational amenities in

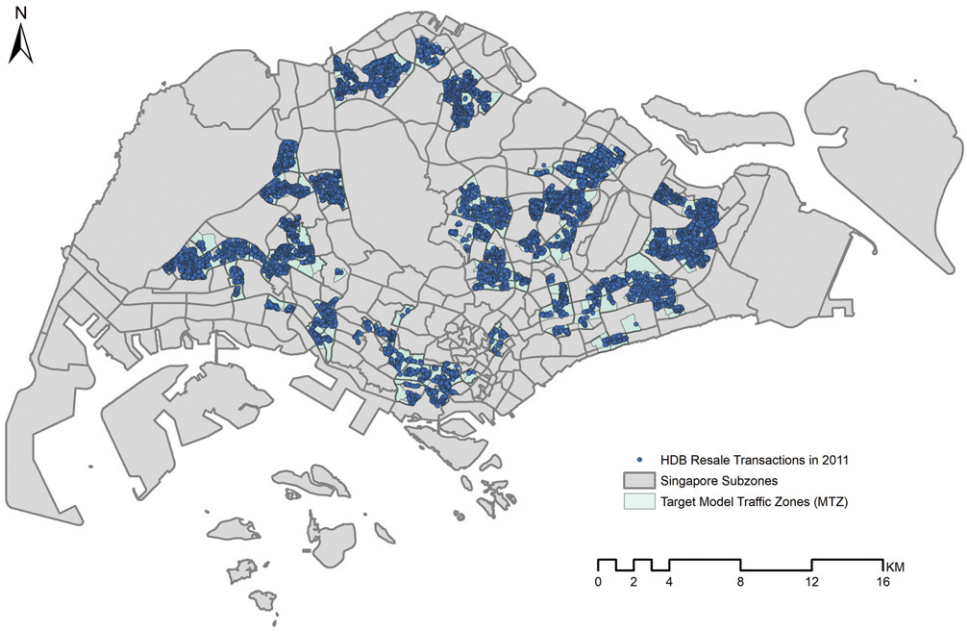

Figure 1. Study area with Housing Development Board resale flat transactions in 2011. HDB = Housing Development Board. (Color figure available online.)

explaining the spatial distribution of public housing prices in Singapore and how their effects vary over space. The results demonstrate that the GWR model with a travel time–based spatial weight matrix performs much better than a GWR model with an ED-based spatial weight matrix and OLS model in explaining the spatial patterns of housing prices.

The article is organized as follows. The next section covers the study area and the data sets, followed by an explanation of the methods employed in this research. After elaboration on the analyzed results as well as a comparison of the three models in the case study, the final section aims to summarize the research and present potential directions for future research as an extension of study.

Study Area and Data

Study Area

Singapore is a city-state in Southeast Asia renowned for its unique public housing system. Singapore has a two-tier housing market consisting of a dominant public housing market developed by the Housing Development Board (HDB) and a private housing market, which operates like any other laissez-faire markets. The HDB flats accommodate over 80 percent of Singapore's resident population. Newly built HDB flats are directly allocated to Singapore citizens who meet certain criteria at heavily subsidized prices. Singapore citizens and Singapore permanent residents are eligible to buy HDB flats on the resale market, but foreigners are not. Prices on the HDB resale flat market are determined by market forces. In this study, we perform spatial hedonic price analyses to understand the price patterns of the HDB resale flat market in Singapore.

Data

To study the spatial patterns of HDB resale flat prices in Singapore and explore the impact of different price determinants, HDB resale flat transaction data for 2011 were collected. The data set contains detailed information on 21,856 transactions, which include transaction price, transaction date, street address, postal code, and various attributes of properties, including floor area and floor level. Figure 1 illustrates the HDB block distribution based on subzones and model traffic zones (MTZs) in Singapore, which is also the primary research area and target in this study. In addition, it shows the 331 target MTZs, which is the spatial unit used by the Land Transport Authority of Singapore for travel demand modeling. Singapore consists of 1,040 MTZs, among

which 331 MTZs had HDB resale flat transactions during the study period.

The overall mean unit resale price according to the MTZs is S$4,559.74/m^2, and the standard deviation according to MTZ is around S$774.80/m^2. The mean unit prices of HDB resale flat transactions at MTZ level (331 MTZ) can be seen in Figure 2 with details.

In addition, based on literature and the specific context of Singapore, data sets for the potential determinants of housing prices have been collected, as shown in Table 1, covering both structural attributes such as average age and floor area of HDB units and locational attributes such as the distance to the nearest hospital, nearest shopping mall, nearest park, nearest prestigious primary school, central business district (CBD), nearest Mass Rapid Transit (MRT) station, and number of bus stops at the MTZ level.

Table 1. Definition of variables

Variables	Description of variables (at the model traffic zone level)
V_{age}	Average age of the apartments
V_{area}	Average floor area of the apartments
$V_{hopital}$	Distance to the nearest general hospital
V_{mall}	Distance to the nearest shopping mall
V_{park}	Distance to the nearest park
$V_{pri\text{-}school}$	Distance to the nearest prestigious primary school
V_{CBD}	Distance to central business district
V_{MRT}	Distance to the nearest MRT station
$V_{b\text{-}stops}$	Number of bus stops

Note: MRT = Mass Rapid Transit.

The two maps in Figure 3 illustrate the spatial distribution of general hospitals, parks, primary shopping malls, top primary schools, MRT stations, and bus stops in Singapore, which are the primary spatial data sets used in this study. After geoprocessing based on ArcGIS (Release 10.4, Esri, Redlands, CA, USA), the maps in Figure 4, including proximity maps of the locational amenities and thematic maps of structural attributes, were obtained for further analyses.

In addition, a unique big data set, the transaction records of all smartcards (i.e., EZ-Link cards) in Singapore in one week, were used in this research to derive a travel time–based matrix to describe the spatial relationship between HDB resale flat transactions. EZ-Link cards are accepted for payments in both the rail transit system and public buses and cover approximately 96 percent of public transportation trips in Singapore (Prakasam 2009). Commuters tap their cards on readers when they enter and leave rail transit stations and board and alight buses. Our data set includes the tap-in and tap-out information of approximately 4 million smartcards in one week, including boarding stop or station, alighting stop or station, trip starting time and ending time, and so on, including more than 30 million transactions or trips.

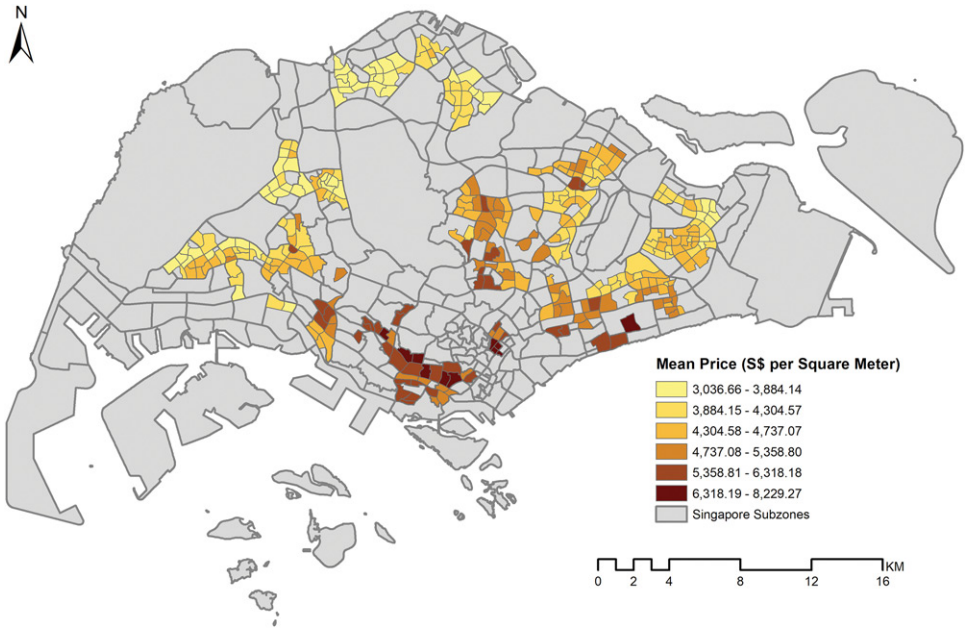

Figure 2. Mean unit price of Housing Development Board resale flat transactions at model traffic zone level.

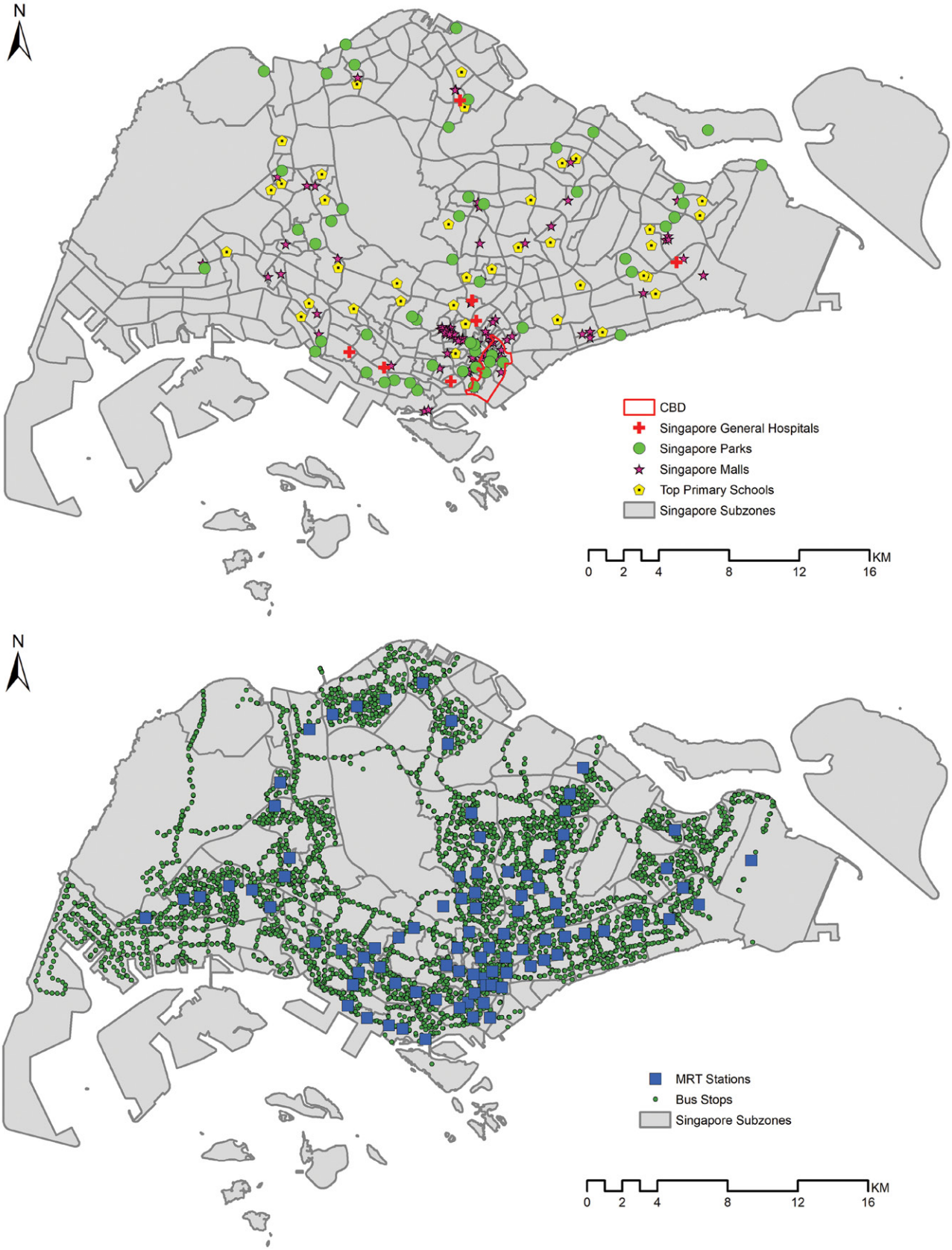

Figure 3. Illustrations of the spatial distribution of potential determinants of housing price in Singapore including general hospitals, parks, primary shopping malls, top primary schools, MRT stations, and bus stops. CBD = central business district; MRT = Mass Rapid Transit.

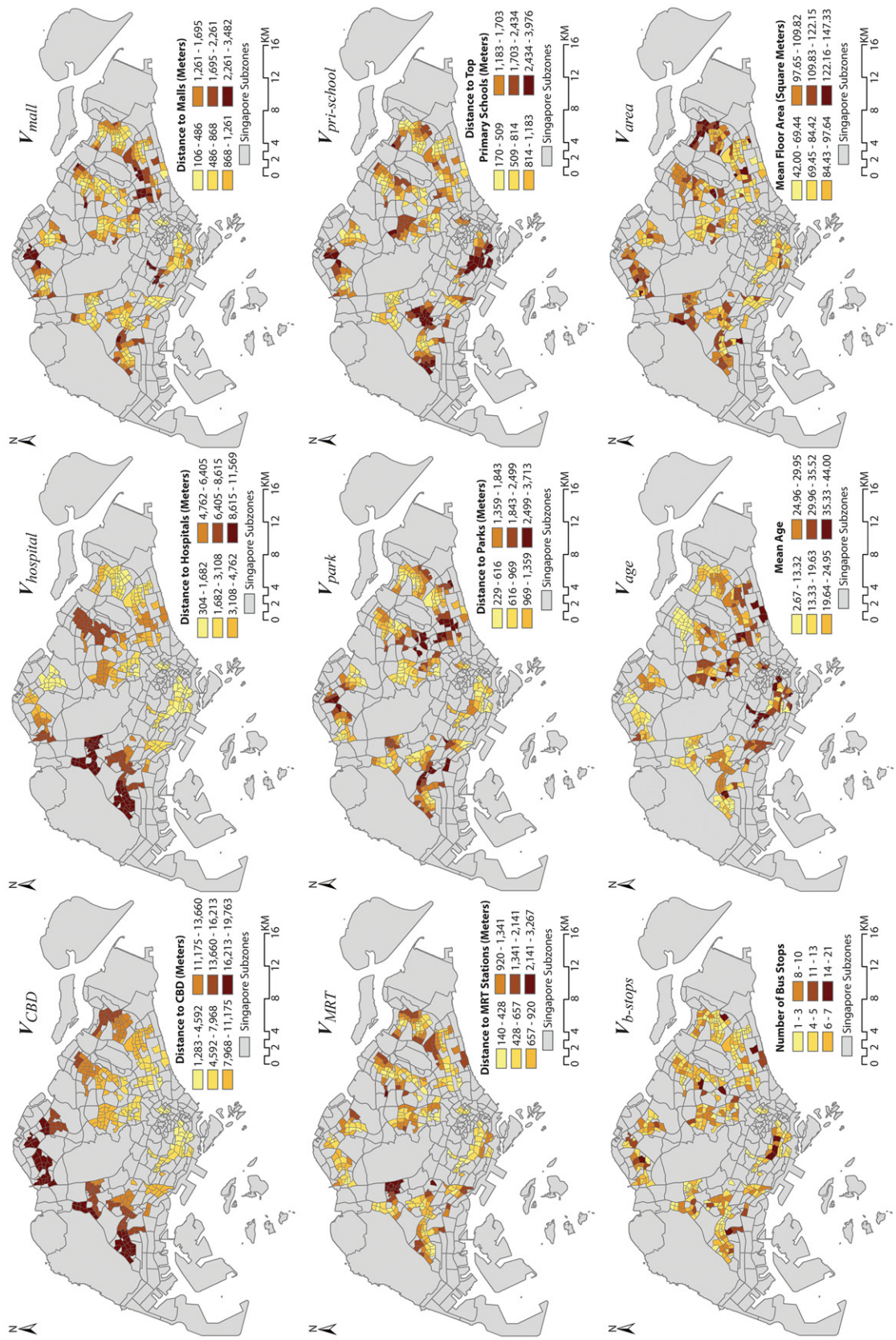

Figure 4. Proximity maps of the locational amenities and thematic maps of structural attributes at the model traffic zone level. CBD = central business district; MRT = Mass Rapid Transit.

Method

Hedonic Pricing Model

According to Lancaster's (1966) consumer behavior theory, the hedonic pricing model is based on the hypothesis that goods are valued on their attributes or characteristics. Because housing characteristics cannot be separated and are normally traded in bundles, real estate is normally treated as a heterogeneous good with a variety of attributes, including structural and locational or spatial attributes (Can 1992; Orford 2002). The structural attributes include the structure of the real estate, including floor area and age of the properties, whereas locational or spatial attributes are the externalities related to the properties' geographic locations, such as the proximity to various amenities.

A hedonic model can be defined as

$$P(RE) = f(S) + f(L) + \varepsilon,$$

where $P(RE)$ is a matrix containing sets of housing prices, $f(S)$ is a functional form of structural attributes, $f(L)$ is a functional form of locational or spatial attributes, and ε is the residual term. An extensive literature has emerged to study the willingness to pay for various factors using hedonic pricing models. An OLS estimation of the conventional hedonic pricing model could lead to biased and inconsistent or inefficient results, however, when there are spatial dependencies among the observations (Anselin 1998). Yu et al. (2007) discussed the status and the spatial effects in a hedonic pricing model with the support of spatial statistics, spatial econometrics, and geographic information systems (GIS), which have received increasing attention in past years.

Geographically Weighted Regression Model

A global regression model can only reflect global trends and might ignore some of the significant local (spatial) variations. To mitigate this issue, GWR and other spatial regression models can be employed. GWR is a nonstationary technique that models varying spatial relationships. Based on Fotheringham, Brunsdon, and Charlton (2002), a basic GWR model for the housing market can be specified as

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i,$$

where y_i refers to the transaction price of location i; x_{ik} refers to the kth attribute for location i; β_{i0} refers to the intercept parameter at location i; β_{ik} refers to the local regression coefficient for the kth independent variable at location i; and ε_i refers to the random error at location i.

Coefficients of GWR models can vary continuously over the research area, and a set of coefficients is supposed to be estimated at any location so that a coefficient surface can be plotted. GWR makes a point-based calibration around each regression point where nearer observations are supposed to have a higher impact on the local set of coefficients than observations further away (Fotheringham, Charlton, and Brunsdon 1998).

GWR is an intuitive and effective tool to explore spatially varying relationships in different fields (Páez, Farber, and Wheeler 2011). Numerous efforts have been made to improve its performance. For instance, a variety of kernel functions have been proposed (Brunsdon, Fotheringham, and Charlton 1996; Fotheringham, Charlton, and Brunsdon 1998) and different rules for the selection of an optimal bandwidth have been suggested (Fotheringham, Brunsdon, and Charlton 2002; Páez, Uchida, and Miyamoto 2002). In addition, different approaches to address local collinearity issues have been studied (Wheeler 2007, 2009).

A critical step in GWR models is determining the weights of the surrounding regressive points. In practice, GWR often employs ED to measure spatial proximity among observations (Charlton and Danforth 2007). Although ED-based methods perform well in capturing the spatial relationship, in certain simplified cases such as geological conditions and water pollution, they lack a solid foundation to represent the actual proximity in complex urban settings such as the real estate market. Of the few attempts to use a non-ED metric in GWR, Huang, Wu, and Barry (2010) proposed a spatiotemporal distance under an ellipsoidal coordinate system for the geographically and temporally weighted regression modeling. Lu et al. (2014) systematically summarized GWR models with a non-ED metric, including road network–distance and time–distance metrics, and demonstrated that GWR with a non-ED metric can improve the fitness of regression models and provide additional insights into the spatial patterns of property values. The indicator of travel time or other kinds of cost distances is apparently able to better represent the measure of spatial proximity in some

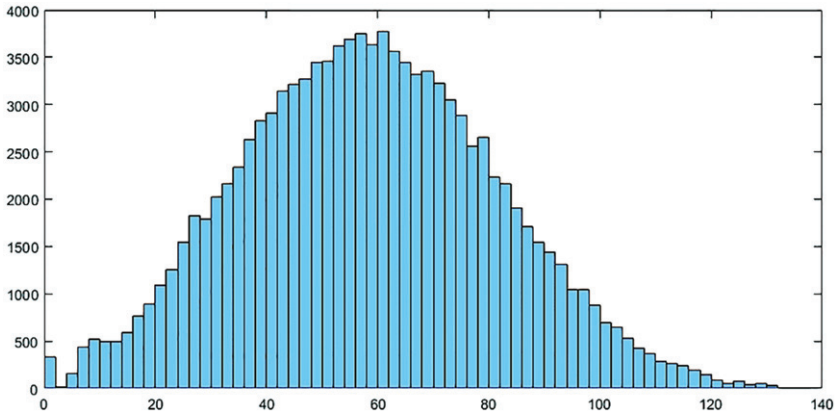

Figure 5. Histogram of travel time–based spatial matrix. (Color figure available online.)

social–spatial phenomena, such as distribution of housing prices and crimes, compared with some other physical–spatial phenomena, such as geological and ecological conditions and water and air pollution.

In this study, data on millions of smartcard transactions in Singapore are used to generate travel time–based distance metrics between HDB resale flat transactions. We employ OLS, traditional GWR with ED, and GWR with travel time–based distance metrics to carry out hedonic pricing analyses of HDB resale flat prices in Singapore; the performance of alternative model specifications is also compared.

Travel Time Matrix Based on Smartcard Transactions

In this study, we generate a travel time origin–destination (OD) matrix using the transaction records of millions of smartcards in Singapore to describe the spatial proximity among HDB resale flat transactions and integrate the travel time matrix into the GWR-based hedonic pricing models. The travel time OD matrix is generated at the MTZ level.

The boarding and alighting stops or stations as well as the travel time of each public transit trip are available from the transaction records of smartcards. We assign each public transit stop or station into MTZ using GIS tools. We then compute the average travel time for each MTZ OD pair to generate a travel time OD matrix at the MTZ level. For a few OD pairs without public transit service connection,[1] we use a linear estimator based on the ED matrix at the MTZ level (using the centroids of different MTZs) to estimate the travel time by public transportation between such pairs. After the linear estimation, a complete travel time matrix for the 331 MTZs with HDB resale flat transactions can be obtained for the travel time–based GWR modeling.

This travel time–based spatial matrix takes multiple factors such as the road infrastructure and congestion levels into account, representing a better measurement of the spatial proximity of HDB resale flat transactions than the ED-based matrix. The histogram of the travel time–based spatial matrix can be seen in Figure 5.

Analyses and Results

In this section, the results of our empirical analyses are presented, including an OLS-based hedonic pricing model, a GWR model with ED-based spatial weight matrix, and a GWR model with a travel time–based spatial weight matrix generated using smartcard transactions. The results and performance of the three models are compared within the context of this research.

OLS-Based Hedonic Pricing Model

We first calibrate an OLS-based hedonic pricing model of the mean HDB resale flat price at the MTZ level in Singapore, which includes all of the structural attributes and locational amenities discussed in the previous section as independent variables. Some independent variables are insignificantly correlated with the average price (per square foot) in each MTZ, including distance to the nearest hospital, nearest shopping mall, and nearest primary school and number of bus stops in the MTZ. Presumably, the relatively even resource allocation (e.g., education, health care) in Singapore makes each HDB town a self-sustained spatial unit to

Table 2. Estimation results of ordinary least squares–based hedonic pricing model

Coefficients							
	Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
Model	*B*	*SE*	Beta	*t*	Significance	Tolerance	VIF
Constant	8,401.896	245.123		34.276	0.000		
V_{age}	–29.421	3.751	–0.335	–7.845	0.000	0.507	1.972
V_{area}	–13.735	1.828	–0.322	–7.514	0.000	0.502	1.993
V_{park}	–0.082	0.033	–0.077	–2.506	0.013	0.966	1.035
V_{CBD}	–0.124	0.006	–0.725	–20.050	0.000	0.707	1.415
V_{MRT}	–0.289	0.044	–0.206	–6.604	0.000	0.950	1.053

Note: VIF = variance inflation factor; CBD central business district; MRT = Mass Rapid Transit.

Table 3. Calibrated parameters (mean values) comparison among the three models

	B (linear hedonic model)	*B* (ED-based GWR, mean)	*B* (travel time–based/ big data–based GWR, mean)
Bandwidth	Global	0.6119	1.1434
Constant	8,401.896	8,116.105	8,373.538
V_{age}	–29.421	–28.5762	–30.4519
V_{area}	–13.735	–10.7304	–11.4668
V_{park}	–0.082	–0.04884	–0.07148
V_{CBD}	–0.124	–0.1216	–0.12611
V_{MRT}	–0.289	–0.45164	–0.43662
Number of observations	331	331	331
Adjusted R_2	0.695	0.8873	0.9589

Note: ED = Euclidean distance; GWR = geographically weighted regression; CBD central business district; MRT = Mass Rapid Transit.

reduce excess travel. These findings are consistent with the spatial distribution patterns of local amenities as revealed in Figure 3.

The coefficients of all significant variables can be found in Table 2. Most coefficients have the expected signs. The standardized coefficients capture the magnitude of the effect on resale price. Structural attributes including age and floor area play an important role in affecting the HDB resale flat price in Singapore. One standard deviation change in age and floor area leads to a price change of 0.335 and 0.322 standard deviations, respectively, ranked second and third among all variables in the model. Distance to the CBD is the most important determinant of resale prices according the standardized coefficient. A reduction in distance to the CBD by one standard deviation could lead to a 0.725 standard deviation increase in resale prices. The other two location amenity variables, distance to the nearest park and distance to the nearest MRT station, also significantly affect HDB resale flat prices but not as much as the other three factors.

We also tested the variance inflation factor (VIF) of all independent variables and found that all VIFs are smaller than 2.0, which indicates that there is no significant multicollinearity among the explanatory variables. In terms of the performance of the overall model, the R^2 is 0.7 and the adjusted R^2 is 0.695 (Table 3), suggesting that the OLS model can explain approximately 70 percent of the variation in HDB resale flat price.

ED-Based GWR Model

To model the spatial variation of the HDB resale flat prices in Singapore, an ED-based GWR model is calibrated on the same set of significant independent variables, including V_{age}, V_{area}, V_{park}, V_{CBD}, and V_{MRT}. The analysis is carried out based on the Spatial-Econometrics Toolbox of Matlab. The overall adjusted R^2 can reach 0.8873, and the bandwidth is 0.6119. GWR is a data-driven local regression model, and the bandwidth defines the extent of different local models. The results reflect

that based on the same data and same variables, the ED-based GWR model can explain over 88 percent of the variation in HDB resale flat prices across all of Singapore in 2011 compared to the 70 percent that can be explained by the OLS-based hedonic pricing model. Apparently, the overall R^2 can explicitly demonstrate that the ED-based GWR model performs better than the linear hedonic regression model based on OLS. The mean values of the model parameters can be seen in Table 4.

Table 4. Calibrated parameters (mean values) of geographically weighted regression based on Euclidean distance

	B (GWR-based analysis, mean)
Constant	8,116.105
V_{age}	–28.5762
V_{area}	–10.7304
V_{park}	–0.04884
V_{CBD}	–0.1216
V_{MRT}	–0.45164

Note: CBD central business district; MRT = Mass Rapid Transit.

GWR Model with Travel Time–Based Spatial Matrix

In addition to the ED-based GWR, a GWR model based on travel time retrieved from smartcard transaction data has been calibrated using the same explanatory variables. The tap-in and tap-out information from more than 30 million EZ-Link card transactions in one week in 2011 was collected and cleaned to generate a travel time OD matrix at the MTZ level as the weights for the GWR analysis. The travel time between an OD pair with transit connection is calculated as the average travel time of all public transport trips between the two MTZs, including both directions. We further regress the observed average travel time between an MTZ pair on the ED between them and use the calibrated model to predict the travel time of a small set of MTZ OD pairs without public transportation connections (less than 10 percent). This approach leads to a complete 331×331 travel time weight matrix at the MTZ level based on the smartcard transaction data.

Table 3 reports the estimation results of the GWR model with travel time–based spatial weight matrix using the same variables as the OLS model and ED-based GWR model. The final results show that the adjusted R^2 can reach 0.9589, which is much higher than the 0.8873 of the ED-based GWR model and 0.695 based on the OLS-based hedonic regression model. In addition, the optimal bandwidth of the travel time–based GWR model can reach 1.1434, which is almost double the optimal bandwidth of the ED-based GWR model (i.e., 0.6119). The significant improvement in both bandwidth and goodness-of-fit statistics demonstrates that the travel time–based GWR can better explain the spatial variations in HDB resale flat prices compared to ED-based GWR. The optimal bandwidth in the ED-based GWR models is much smaller, suggesting that there exists a stronger heterogeneity in the study area. This might be due to the use of ED-based weights, where the isotropic-based measurement overestimated the heterogeneity. In contrast, the success of our GWR model with a travel time–based spatial weight matrix in modeling the spatial variation in housing resale prices in our case study demonstrates its potential to be applied to other hedonic housing pricing models in non-ED settings.

Table 3 compares the estimation results of the OLS model, ED-based GWR model, and travel time–based GWR model. For both the ED-based GWR model and travel time–based GWR model, the parameters are the mean values of the sets of parameters calibrated for individual local models. The estimated coefficients by different models have the same sign, indicating that the three models produce consistent results.

Unlike the OLS model that produces a global estimate of individual coefficients for the entire study area, the GWR model could reveal the spatial variations of estimated coefficients in the study area. In this study, we observe that despite the similarity in the mean values of the estimated coefficient across the three models, the spatial distribution of the estimated coefficients could differ significantly between the ED-based and travel time–based GWR models. As an example, Figure 6 plots the spatial patterns of coefficients of the variable V_{CBD} in the ED-based GWR model and the travel time–based GWR model, respectively. The clear difference between the two experiments presumably results from the different choices of spatial weight matrices. The map for the ED–based GWR model shows a polycentric urban structure. Some regional centers in the suburban area such as Woodlands, Tampines, and Punggol have similar coefficients of close to zero

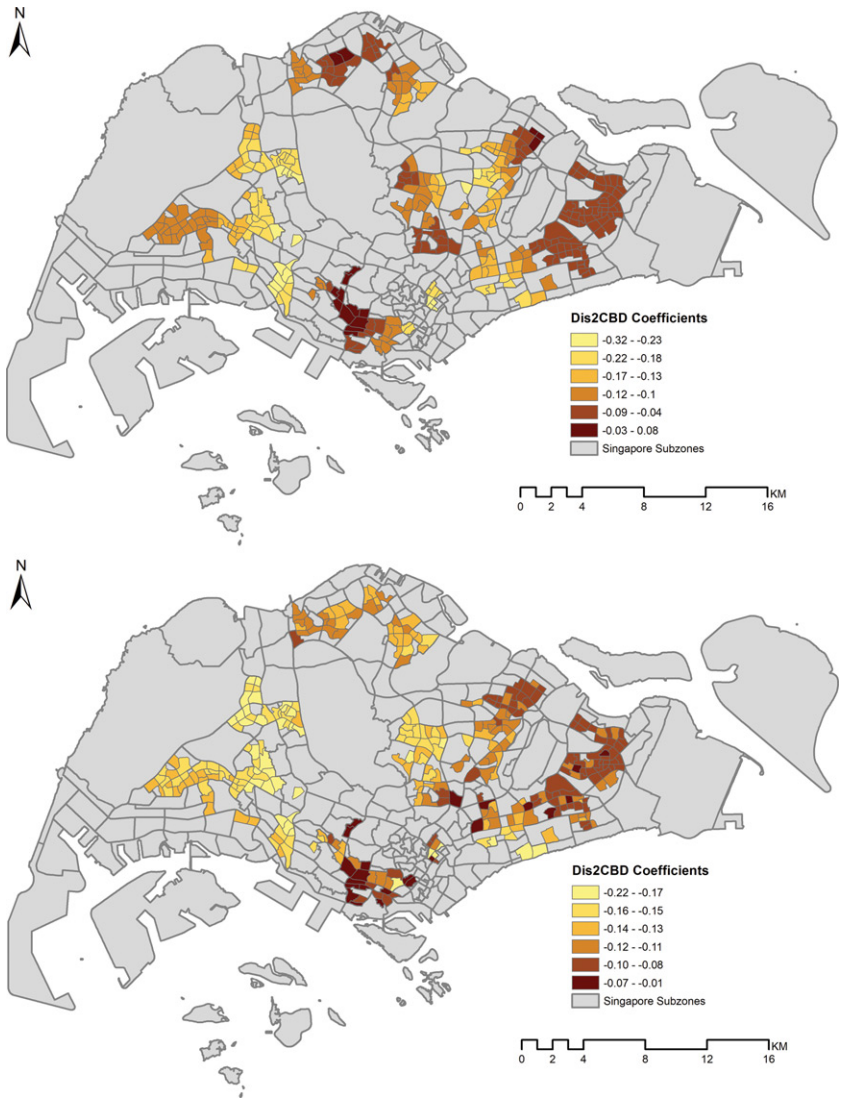

Figure 6. The coefficients (of V_{CBD}) distribution map based on Euclidean distance–based GWR model (top) and travel time–based GWR model with the support of big smartcard transaction data sets (bottom). CBD = central business district; GWR = geographically weighted regression.

as the city center. In contrast, the map for the travel-time-based GWR model shows a clear monocentric pattern, with the price gradient declining at a faster pace further away from the city center. This spatial pattern is consistent with the spatial structure of Singapore, a city where the CBD remains a major job center and property values in the urban center are much higher than that at the edge.

In summary, the ED-based GWR has improved the understanding of the locational variation in housing resale prices compared to a global linear hedonic regression model, which can only estimate one set of coefficients for all of these different districts. The spatial pattern revealed by the ED-based GWR model is still questionable, however, because ED cannot precisely represent the accessibility of these HDB blocks to the CBD area compared to the real travel time or travel distance. Furthermore, the big data set of smartcard transaction records has the advantage of characterizing accessibility over road network distance or travel time generated using traditional travel surveys, due to the rich and dynamic nature of big data. The R^2 achieved by the travel time–based GWR model explicitly demonstrates its value compared to the traditional ED-based GWR model.

Discussion and Conclusion

In this research, three hedonic pricing models, including an OLS model, an ED-based GWR model, and a travel time–based GWR model with the support of an extensive smartcard transaction data set, have been developed and applied to study the spatial or locational variation in HDB resale flat prices and the various determinants of HDB resale flat prices in Singapore in 2011.

First, an OLS-based hedonic regression model has been employed to analyze a set of commonly considered potential determinants of housing resale prices in Singapore in 2011, including the average age of the apartments, the average floor area of the apartments, distance to the nearest hospital, the distance to the nearest shopping mall, the distance to the nearest park, the distance to the nearest prestigious primary school, the distance to the CBD, the distance to the nearest MRT station, and the number of bus stops. The results demonstrate that among all factors, age of the apartments, floor areas of the apartments, distance to the nearest park, distance to the CBD, and distance to the nearest MRT station significantly affect the HDB resale flat prices based on the spatial data sets and HDB resale flat transaction data sets in Singapore in 2011. The distance to prestigious primary schools, distance to the nearest shopping mall, and distance to the nearest general hospital, which are supposed to significantly affect the housing resale prices in many other cities, are not significantly correlated with the HDB resale flat prices in Singapore in 2011. This might be due to the reasonable resource allocation (e.g., education, health care) in Singapore, and HDB flats are also subsidized housing (residents might not be willing to spend too much on the selection of better primary school education) in Singapore and nearly 80 percent of Singaporean families live in HDB apartments.

Furthermore, an ED-based GWR model and a travel time–based GWR model with the support of data from millions of smartcard transactions in one week have been employed to better study the locational variation in HDB housing resale prices in Singapore, with significant variables identical to those obtained from the linear hedonic regression model. A comparison between the linear hedonic regression model, ED-based GWR model, and travel time–based GWR model has also been conducted. The results explicitly show that both the GWR models perform much better than the traditional linear hedonic regression model using identical variables. In addition, the travel time–based GWR model can significantly improve performance compared to the ED-based GWR model in capturing the spatial variation of the HDB resale flat prices in Singapore in 2011.

All in all, it is evident that the GWR model based on travel time data retrieved from the big smartcard transaction data sets can be effectively used to support the study of the spatial variation of public housing resale prices in Singapore. In addition, the travel time–based/big data–based GWR model has shown its great potential for hedonic housing pricing modeling in other research areas or even other disciplines such as public health and criminal justice. Of course, there is still room for improvement. More in-depth analyses should also be conducted in the future on other submarkets, such as the private housing market and the rental market, to obtain a more complete picture of the housing market in Singapore.

Funding

Funding for this study was provided by the Singapore Ministry of Education (MOE) Academic Research Fund Tier 1 Grant (R-109-000-229-115).

Note

1. The main reason could be (1) the limit in the coverage of public transportation services, or (2) some of the stops or stations can serve MTZs even though they are physically located in one MTZ.

References

Anselin, L. 1998. GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research* 9 (1):113–33.

Anselin, L., and D. A. Griffith. 1988. Do spatial effects really matter in regression analysis? *Papers in Regional Science* 65 (1):11–34. doi: 10.1111/j.1435-5597.1988.tb01155.x.

Basu, S., and T. G. Thibodeau. 1998. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics* 17 (1):61–85. doi: 10.1023/A:1007703229507.

Blomquist, G., and L. Worley. 1981. Hedonic prices, demands for urban housing amenities, and benefit estimates. *Journal of Urban Economics* 9 (2):212–21. doi: 10.1016/0094-1190(81)90041-3.

Bowen, W. M., B. A. Mikelbank, and D. M. Prestegaard. 2001. Theoretical and empirical considerations regarding space in hedonic housing price model applications. *Growth and Change* 32 (4):466–90. doi: 10.1111/0017-4815.00171.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28 (4):281–98. doi: 10.1111/j.1538-4632.1996.tb00936.x.

Can, A. 1992. Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics* 22 (3):453–74. doi: 10.1016/0166-0462(92)90039-4.

Cebula, R. J. 2009. The hedonic pricing model applied to the housing market of the city of Savannah and its Savannah historic landmark district. *The Review of Regional Studies* 39 (1):9–22.

Charlton, J. P., and I. D. Danforth. 2007. Distinguishing addiction and high engagement in the context of online game playing. *Computers in Human Behavior* 23 (3):1531–48. doi: 10.1016/j.chb.2005.07.002.

Chen, W. Y., and X. Li. 2017. Cumulative impacts of polluted urban streams on property values: A 3-D spatial hedonic model at the micro-neighborhood level. *Landscape and Urban Planning* 162:1–12. doi: 10.1016/j.landurbplan.2017.01.012.

Cliff, A. D., and J. K. Ord. 1981. *Spatial processes: Methods and applications*. London: Pion.

de Araujo, P., and K. Cheng. 2017. Do preferences for amenities differ among home buyers? A hedonic price approach. *Review of Urban & Regional Development Studies* 29 (3):165–84.

Debrezion, G., E. Pels, and P. Rietveld. 2011. The impact of rail transport on real estate prices: An empirical analysis of the Dutch housing market. *Urban Studies* 48 (5):997–1015. doi: 10.1177/0042098010371395.

Diao, M. 2015. Selectivity, spatial autocorrelation and the valuation of transit accessibility. *Urban Studies* 52 (1):159–77. doi: 10.1177/0042098014523686.

Diao, M., Y. Fan, and T. F. Sing. 2017. A new Mass Rapid Transit (MRT) line construction and housing wealth: Evidence from the Circle Line. *Journal of Infrastructure, Policy and Development* 1 (1):64–89. doi:10.24294/jipd.v1i1.22

Diao, M., and J. Ferreira, Jr. 2010. Residential property values and the built environment: Empirical study in the Boston, Massachusetts, metropolitan area. *Transportation Research Record: Journal of the Transportation Research Board* 2174:138–47. doi: 10.3141/2174-18.

Diao, M., D. Leonard, and T. F. Sing. 2017. Spatial-difference-in-differences models for impact of new mass rapid transit line on private housing values. *Regional Science and Urban Economics* 67:64–77. doi: 10.1016/j.regsciurbeco.2017.08.006.

Diao, M., Y. Qin, and T. F. Sing. 2016. Negative externalities of rail noise and housing values: Evidence from the cessation of railway operations in Singapore. *Real Estate Economics* 44 (4):878–917. doi: 10.1111/1540-6229.12123.

Diao, M., Y. Zhu, and J. Zhu. 2017. Intra-city access to inter-city transport nodes: The implications of high-speed-rail station locations for the urban development of Chinese cities. *Urban Studies* 54 (10):2249–67. doi: 10.1177/0042098016646686.

Dubin, R. A. 1998. Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics* 17 (1):35–59. doi: 10.1023/A:1007751112669.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. New York: Wiley.

Fotheringham, A. S., M. E. Charlton, and C. Brunsdon. 1998. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A* 30 (11):1905–27. doi: 10.1068/a301905.

Geng, J., K. Cao, L. Yu, and Y. Tang. 2011. Geographically weighted regression model (GWR) based spatial analysis of house price in Shenzhen. In *2011 19th International Conference on Geoinformatics*, 1–5. Shanghai, China: IEEE. doi:10.1109/GeoInformatics.2011.5981032

Huang, B., B. Wu, and M. Barry. 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science* 24 (3):383–401. doi: 10.1080/13658810802672469.

Jim, C., and W. Y. Chen. 2007. Consumption preferences and environmental externalities: A hedonic analysis of the housing market in Guangzhou. *Geoforum* 38 (2):414–31. doi: 10.1016/j.geoforum.2006.10.002.

Kelejian, H. H., and I. R. Prucha. 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics* 17 (1):99–121. doi: 10.1023/A:1007707430416.

Lai, Y., X. Zheng, L. H. Choy, and J. Wang. 2017. Property rights and housing prices: An empirical study of small property rights housing in Shenzhen, China. *Land Use Policy* 68:429–37. doi: 10.1016/j.landusepol.2017.08.010.

Lancaster, K. J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74 (2):132–57. doi: 10.1086/259131.

Lu, B., M. Charlton, P. Harris, and A. S. Fotheringham. 2014. Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house price data. *International Journal of Geographical Information Science* 28 (4):660–81. doi: 10.1080/13658816.2013.865739.

Militino, A. F., M. D. Ugarte, and L. Garcia-Reinaldos. 2004. Alternative models for describing spatial dependence among dwelling selling prices. *The Journal of Real Estate Finance and Economics* 29 (2):193–209. doi: 10.1023/B:REAL.0000035310.20223.e9.

Orford, S. 2002. Valuing locational externalities: A GIS and multilevel modelling approach. *Environment and Planning B: Planning and Design* 29 (1):105–27. doi: 10.1068/b2780.

Páez, A., S. Farber, and D. Wheeler. 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A* 43 (12):2992–3010. doi: 10.1068/a44111.

Rosen, S. 1974. Hedonic price and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82:34–45.

Páez, A., T. Uchida, and K. Miyamoto. 2002. A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests. *Environment and Planning A* 34 (5):883–904. doi: 10.1068/a34133.

Prakasam, S. 2009. Evolution of e-payments in public transport—Singapore's experience. *JOURNEYS* 3:53–61.

Schläpfer, F., F. Waltert, L. Segura, and F. Kienast. 2015. Valuation of landscape amenities: A hedonic pricing analysis of housing rents in urban, suburban and periurban Switzerland. *Landscape and Urban Planning* 141:24–40. doi: 10.1016/j.landurbplan.2015.04.007.

Sheppard, S. 1999. Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics* 3:1595–1635. doi: 10.1016/S1574-0080(99)80010-8.

Shimizu, C. 2014. Estimation of Hedonic single-family house price function considering neighborhood effect variables. *Sustainability* 6 (5):2946–60. doi: 10.3390/su6052946.

Upton, G., and B. Fingleton. 1985. *Point pattern and quantitative data*. Vol. 1 of *Spatial data analysis by example*. Chichester, UK: Wiley.

Wen, H., Y. Xiao, and L. Zhang. 2017. Spatial effect of river landscape on housing price: An empirical study on the Grand Canal in Hangzhou, China. *Habitat International* 63:34–44. doi: 10.1016/j.habitatint.2017.03.007.

Wheeler, D. C. 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A* 39 (10):2464–81. doi: 10.1068/a38325.

———. 2009. Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environment and Planning A* 41 (3):722–42. doi: 10.1068/a40256.

Yu, D., Y. D. Wei, and C. Wu. 2007. Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning B: Planning and Design* 34 (6):1085–1102. doi: 10.1068/b32119.

KAI CAO is a Lecturer in the Department of Geography, National University of Singapore, 117570, Singapore. E-mail: geock@nus.edu.sg. His research interests include spatial simulation and optimization, urban studies and big data analytics, spatial planning, and spatially integrated social science.

MI DIAO is an Assistant Professor in the Department of Real Estate, National University of Singapore, 117566, Singapore. E-mail: rstdm@nus.edu.sg. His research interests include big data analytics, transportation, and urban and regional economics.

BO WU is a Professor in the Department of Geography and Environment at Jiangxi Normal University, Nanchang, Jiangxi Province, 330022, China. E-mail: wavelet778@sohu.com. His research interests include spatiotemporal data analysis, machine learning, and remote sensing image processing.