EzModel: An Interactive Tool to Model HDB Resale Prices using Mixed Geographically Weighted Regression

Daniel CHIN
School of Information Systems
Singapore Management University
Singapore
daniel.chin.2017@business.smu.edu.sg

Patrick LIM
School of Information Systems
Singapore Management University
Singapore
patrick.lim.2016@sis.smu.edu.sg

Jianrong SHI
School of Information Systems
Singapore Management University
Singapore
jrshi.2016@sis.smu.edu.sg

Guided by Associate Professor of Information Systems: Dr. Kam Tin Seong

ABSTRACT

In recent decades, modeling house prices has become a hot topic among economists, planners, and policymakers due to the significant role of properties in household wealth and national economy. As many existing hedonic pricing models fail to take into account the effect of local spatial variations on the housing prices, there has been an increasing interest in using spatial econometrics, specifically geographically weighted regression (GWR) to more accurately model housing prices. However, with many different existing GWR models, it is not very easy for the casual users to use one for analysis. Addressing this need, we designed and developed EzModel, a dynamic and interactive GWR modelling tool to help economists, planners, policymakers to explore and analyse how variations of features in the local surroundings, such as number of MRT/LRT stations around a HDB unit, may affect housing prices. As there may be certain factors that should be more accurately modelled as global factors, our tool also allows users to easily switch to a mixed geographically weighted regression (MGWR) model to model both local and global variables. The potential of EzModel is demonstrated through the use of robust GWR modeling options as well as interactive isoline maps to analyse the effects of local spatial variations on Singapore's Housing Development Board (HDB) Flat resale prices.

KEYWORDS

Singapore HDB Resale Prices, Amenities, Geospatial Analytics, Geographically Weighted Regression

1 INTRODUCTION

In the past few decades, housing markets have boomed with rapid urbanisation and population growth. Both housing prices and housing transactions have continued to witness steady growth in many parts of the world. With the increasingly significant role of the housing market on household wealth and economic prosperity, modelling housing prices has become a key subject of interest among economists, planners, and policymakers.

Currently, there are many existing housing pricing models that make use of traditional hedonic regressions which are linear in nature and fail to consider the potential spatial effects on housing prices. This has led to increasing interest in using spatial econometrics, specifically GWR to more accurately model housing prices.

Yet, with many different existing GWR models, it is not very easy for the economists, planners and policymakers to use one for modeling and analysis. Even though some sophisticated GIS tools come with GWR modeling, they are not always readily available as 'off-the-shelf' modeling products They are also not as robust, with numerous customisations required for specific GWR modelling. This may prove both costly and time-consuming.

Also, while the use of GWR models for modeling housing prices is not entirely new, currently, most of them tend to use the basic model only which assumes all variables as localised variables. This lacks the flexibility to model global variables at the same time, which is possible under the mixed GWR model.

Hence, to address these needs, open-source tools such as R Shiny could be used to easily build GWR models to analyse spatial variation effects on housing prices. It can also provide users with the flexibility to choose between the basic and mixed GWR model for modelling housing prices in one application.

To demonstrate the utility of this, our project aims to provide economists, planners, and policymakers with an open-source, interactive, GWR modelling tool to help analyse how spatial variation among the features in local surroundings affect housing prices. To increase the robustness of the tool, the platform also provides users with the option of using a mixed GWR model on top of the basic model, and users could compare which model is able to more accurately account for the spatial variation effects on housing prices using the variables they selected. For users who want to model housing prices with a more diverse set of spatial variables, the platform also comes with the option for users to easily upload their own datasets and define their own variables for GWR modeling.

We have selected Singapore housing market as our case study for this project as Singapore is known for being one of the most geographically constrained city in the world and is also ranked as the second-most expensive housing market in the world.

To limit the scope of the project, we have decided to focus on Singapore's resale flat market, since more than 80% of Singapore's population live in public housing apartments managed by the Housing Development Board, commonly referred to as HDB flats.

This paper details our research and development efforts to design and implement a web-based GWR modelling tool for analysing the significance of various local features in explaining HDB resale prices. It consists of nine sections. Section 1 provides a general context of the housing market and discusses the issues and motivation behind our project. Section 2 describes some of the existing literature that serves as reference and learning examples for our team's project. Section 3 discusses how housing and local features data is collected and prepared.. Next, in Section 4, we detail our methodology, development tools and system architecture used to develop the platform. Section 5 provides the description of our application's functionalities. In Section 6, we discuss some of the interesting findings obtained from the modelling tool. Section 7 documents the feedback obtained from industry experts at the town hall showcase. In Section 8, we discuss possible future works for the project and lastly, the paper concludes by highlighting the techniques used used in the application.

2 LITERATURE REVIEW

A study conducted in 2017 discusses the weakness of Ordinary Least Square (OLS) regression approach to model housing prices in the context of Shanghai (Huang, Chen, Xu, Zhou, 2017). It discusses one of the OLS assumptions that the error terms is normally distributed and independent, which is actually not the case for housing as apartments in the same area, given a similar area, tend to hover around the similar prices due to the same amenities around the estates. The paper goes on by computing the spatial autocorrelation statistics and ultimately, using the total floor area, GDP per capita, distance to downtown areas and male-female ratio as the independent variables in the GWR model that they proposed.

Lu, Harries, Charlton and Brunsdon, in their article that is published in 2014 explained the use of weighing matrix clearly and applied it to their GWR implementation of Dublin 2004 voters turnout, alongside principal component analysis (PCA) for dimensional reduction on independent variables such as age group, education, social class, migration and public

housing. They then went on and also used a semi-parametric model, which is also a MGWR model, treating some variables as global and some as local to explain the variations of voters (Lu, Harries, Charlton and Brunsdon, 2014).

3 DATA COLLECTION & PREPARATION

To perform GWR modelling on our application, we would require 2 main types of data:

1. HDB Flat Resale Prices Data
2. Data on the features which can be selected by the user as independent variables for the GWR model

The HDB Flat Resale Prices dataset is obtained from data.gov.sg[1]. The data is then converted from CSV to Shapefile after geocoding HDB Addresses using OneMap's[2] Search API, which allowed us to derive a set of X and Y coordinates when provided with the address.

The HDB Resale dataset was then processed to include individual columns for the Year and Month the transaction took place. This was to facilitate the filtering of records to specific time periods based on user inputs. A new column Storey Median was also calculated from the dataset's Storey Range column. For example, a record with storey range "04 to 06" resulted in a Storey Median value of 5. This allowed for a new numerical variable that could be introduced into resulting regression models.

Although EzModel allows users to upload their own data to be used as independent variables for the GWR models, EzModel also provides some preloaded features data to be used as potential independent variables for modeling housing prices, as follows:

1. Sports facilities locations
2. MRT/LRT stations locations
3. Park locations
4. Preschool locations
5. Primary/Secondary school locations
6. Shopping mall locations
7. Food centre locations.

Food Centre data was obtained by scraping websites such as Food Republic, Kopitiam and Koufu for their actual location of the food centre and geocoding them. Whereas Shopping Mall data was obtained from an online list of Singapore's shopping malls that was subsequently geocoded as well. The remaining datasets had location data already prepared, and were obtained from data.gov.sg.

The map of Singapore and its regions are defined in the Shapefile that is obtained from data.gov.sg. This shows Singapore's planning subzones and its defined boundaries. The central business district (CBD) is defined as Raffles Place Park, where it is indicated as a point coordinate (1.2841836, 103.8515103).

4 METHODS

4.1 Application Architecture

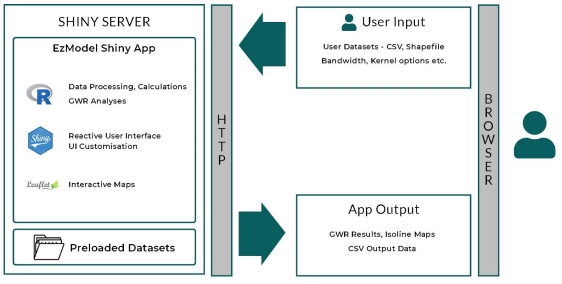


Figure 1: Application Architecture Diagram

The application was developed using the R Shiny web application framework that is based on the R programming language. R Shiny is an easy way to build interactive dashboard applications. In the backend, the CSV and SHP data are used for geocoding, projection conversion, Geographically Weighted Regression and Mixed Geographically Weighted Regression.

The R Shiny application runs on a Shiny server, currently hosted by Shinyapps.io. Data

[1] Data.gov.sg is a web portal that is maintained by the Singapore government that provide publicly available datasets for free

[2] OneMap is the authoritative national map of Singapore with the most detailed and timely updated information developed by the Singapore Land Authority

mentioned in Section 3.1 are stored on the server and loaded by the application each time a user accesses the app. The mapping features of the app also makes calls to OpenStreetMap to generate the interactive maps that are displayed to the user.

4.2 Application Overview

4.2.1 R Packages

The following R Packages are used in the development of the EzModel Application:

shiny	shinydashboard	shinyWidgets
shinyjs	shinycssloaders	shinythemes
shinyBS	tidyverse	sp
maps	maptools	gstat
rgeos	sf	raster
rgdal	heatmaply	lattice
tmap	tmaptools	classInt
sdep	grid	gridExtra
geofacet	ggmap	dendextend
leaflet	DT	GWmodel
nngeo	corrplot	rlang

4.2.2 Algorithms

4.2.2.1 Geographically Weighted Regression

EzModel uses the GWR model, a local statistical technique that takes into account spatial nonstationarity in terms of the coefficients of each variable for each observation in the resulting regression model. This technique, which is based off and in accordance of Tobler's first law of geography (Waters, 2017), results in an equation as such:

$$y_i = \beta(u_i, v_i) + \sum_k \beta_k (u_i, v_i) x_{ik} + \varepsilon_i$$

Where:

y_i is the dependent variable of price at location i

u_i, v_i is the coordinates of the i-th point in space

Figure 2: GWR Equation

The intercepts and the coefficients in the formula are varied according to the location of the observation and the surrounding observations in the spatial context.

Thus, there are a few parameters that have to be calibrated before running the GWR. Firstly, as different observations will be assigned different weights depending on the location with respect to each point, a weighting kernel function has to be decided upon to determine the allocation of weights to each observation according to distance. This is in contrast with the global function, which gives equal weightage to all observations. Such kernel functions include:

1. Gaussian
2. Exponential
3. Box-car
4. Bi-square
5. Tri-cube.

The functions can be categorised into two main types: Continuous and Discontinuous. Continuous functions include the Gaussian and Exponential kernels, where weightage decreases gradually as distance increases. Even beyond the determined bandwidth, observations are still assigned a weightage, although the weightage is very small. Whereas discontinuous functions include the Box-car, Bi-square and Tri-cube kernels, whereby observations' weightages are reduced to zero once distance between observation and the center-point exceeds the specified bandwidth.

Global Model	$w_{ij} = 1$
Gaussian	$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right)$
Exponential	$w_{ij} = \exp\left(-\frac{\lvert d_{ij}\rvert}{b}\right)$
Box-car	$w_{ij} = \begin{cases} 1 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Bi-square	$w_{ij} = \begin{cases} (1-(d_{ij}/b)^2)^2 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Tri-cube	$w_{ij} = \begin{cases} (1-(\lvert d_{ij}\rvert/b)^3)^3 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$

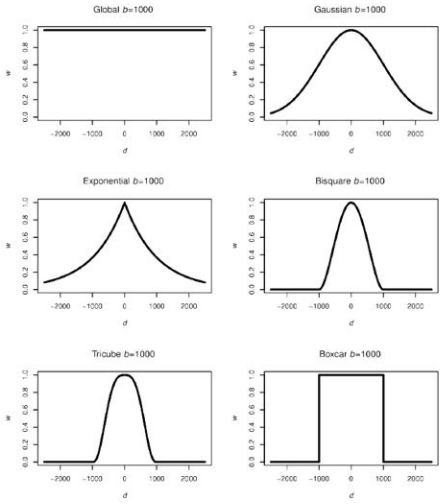
Where:

w_{ij} is the j-th element diagonal of the matrix of geographical weights $W(u_i, v_i)$

d_{ij} is the distance between observation i and j

b is the bandwidth

Figure 3: Weighing Calibration Functions



Where:

b is the bandwidth of 1000
w is the weight
d is distance between 2 observations

Figure 4: Plots of Kernel Functions

Secondly, another parameter that has to be calibrated for the GWR model would be the weighting scheme. In essence, there are two main weighting schemes: Fixed and Adaptive. This is largely tied in with the third parameter to be customised: bandwidth. In a fixed weighting scheme, the same bandwidth is applied to all observations when applying the weighting kernel function. This, however, might cause issues whereby there are lesser observations taken into account in areas where data points are sparse, and more points included in areas where observations are dense. This is where an adaptive weighting scheme applies, in which bandwidth is adjusted according to the context of each observation, for example, to a predetermined k-nearest neighbours. Thus where data points are sparse, bandwidth increases, and where data points are dense, bandwidth is reduced.

Lastly, the method to determining bandwidth also has to be calibrated for the model. Aside from the user entering a pre-defined bandwidth, there are two other possible methods. Firstly, the Least Cross-Validation (CV) score method helps determine a bandwidth based on minimizing squared errors. The other method would be using the Least Akaike Information Criterion (AIC) method that takes into account different degrees of freedom for varying models from the different observations.

Due to the fact that the use of different kernel functions, weighting schemes, as well as bandwidth determination methods will affect the overall GWR model output, we want to give users the ability to calibrate their model based on these parameters based on what they wish to explore, or based on what they deem is most appropriate for the variables selected.

Due to the nature of GWR model, all of the variables that the user has specified will be included in the GWR model, regardless of whether it is specified as global or local. The choice of selecting global or local only matters in the MGWR model, which will be discussed in further detail in the next subsection.

Constructing the model will yield all the predicted price (ŷ), intercept estimate, as well as coefficient estimates, critical values and p-values for all the variables that are selected into the model. Additionally, diagnostic information such as R-square, adjusted R-square and two versions of the Akaike Information Criterion (AIC) and Corrected Akaike Information Criterion (AICc) are obtained. AIC is a relative measure of goodness of fit which penalises the number of variables in a model and also information loss at the same time, allowing for comparison between models, where a lower AIC score is reflective of a better model. AICc is simply a variation of AIC that corrects for small sample size (Burnham and Anderson, 2004).

4.2.2.2 Mixed Geographically Weighted Regression

Another model that is computed by EzModel is the MGWR model. The MGWR model, as suggested by its name, allows for a mix of both analysis variables that will be regressed according to the geographic weights of the observations around it, as well as variables in which coefficients estimates derived from a global regression will be kept constant

throughout all observations and resulting models. Its formula is as follows:

$$\sum_{j=1,k_a} a_j x_{ij}(a) + \sum_{l=1,k_b} b_i(u_i, v_i) x_{il}(b) + \varepsilon_i$$

Where:

$\{a_1 \dots a_{ka}\}$ are the k_a global coefficients

$\{b_i(u_i, v_i) \dots b_{k_b}(u_i, v_i)\}$ are the k_b local coefficients

Figure 5: MGWR Equation

In the context of modeling housing prices, the coefficient for which the floor range of a flat affects its resale price might be deemed to be/approximately constant throughout observations. Hence, this Floor Range variable could be selected to be a variable in which its coefficient estimate would be globally applied to all the resulting MGWR models. (.e.g. floor area and remaining lease left).

Users can experiment in creating an optimal model by selecting independent variables in which they want the coefficient estimates to be kept global, while leaving the other variables to be run against the GWR.

Similar to GWR, the MGWR model yields the intercept estimate, coefficient estimates for the variables as well as AICc, allowing users to compare with the GWR model.

4.2.2.3 Isoline Mapping via Inverse Distance Weighted Interpolation

Rather than merely plotting the results of the user-customised model in a point map form, coloured by R-squared values of the individual regression models around each point, we wish to convey more information. This information is in the form of highlighting regions in which a certain coefficient estimate is greater in scale than other regions. For example, resale prices around a certain HDB town or subzone could be more affected by the number of primary schools around the flats, compared to other regions.

Hence, to convey such information to users, we will adopt the use of an isoline map to show regions of high/low coefficients of a user-specified variable. Through meshing of the Singapore region that is to be displayed, missing values are obtained by interpolating the individual points' coefficient estimates of the regression model's variables via an inverse distance weighting technique, a surface containing the interpolated data across the entire map area can be layered onto the output display.

5 APPLICATION DESCRIPTION

This section will highlight some of the application features with relevant screenshots of the EzModel application at different stages of a user's journey.

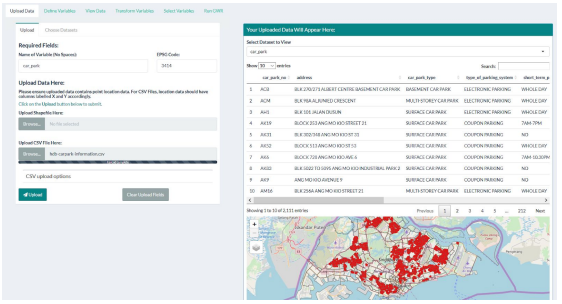


Figure 6: Uploading and Specifying of Data

This page shows an example of how data can be uploaded onto the application. HDB carparks is used as an example of a CSV file, which contains X and Y coordinates of the locations. Upon naming the attributes and assigning the EPSG code, a data table and a plot of the coordinate points are shown on the map, allowing users to quickly visualise his uploaded data and check if the data is indeed the one that he intend to enter into the model. To the right of the "Upload" contains another tab where users can select the existing preloaded datasets that he/she can use. It is in the format checkboxes such that users can select multiple attributes that they feel are in interest to them.

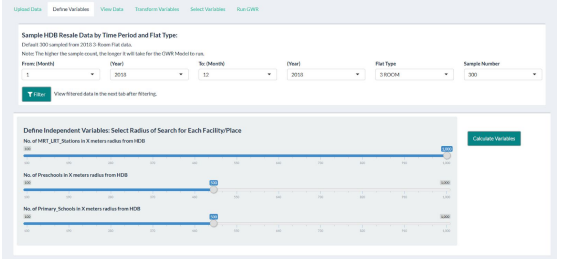


Figure 7: Defining of Radius and Sampling Data

After the user selects the variables that he/she want to include in the model, he can switch over to the “Define Variables” tab to decide on the HDB dataset that he will use. Options are provided for the user to filter to within a time frame, type of HDB flat and the decision for the size of the sample. The time frame can be adjusted to the granularity of months of each year. Sampling is included as it significantly cuts down processing time while still maintaining accuracy.

The sliders on bottom are displayed in accordance to the variables that the user chooses. Users can set their own range for the radius of the amenities/facility that is is from the HDB flats. This is made available as each user may have a different preference for the proximity of the amenities. After defining the radius, the user then can select the “Calculate Variables” button to see the computed values of the variables appended to the data table on “View Data” tab.

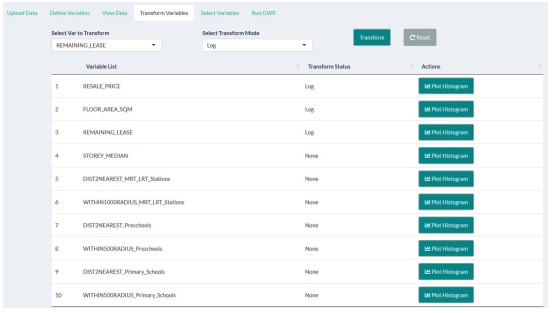


Figure 8: Data Transformation

This tab allows users to look at their distribution of the data that they selected and do data transformation according to the distribution of the histogram. Three different transformation options are provided for them under the dropdown bar “Select Transformation Mode”.

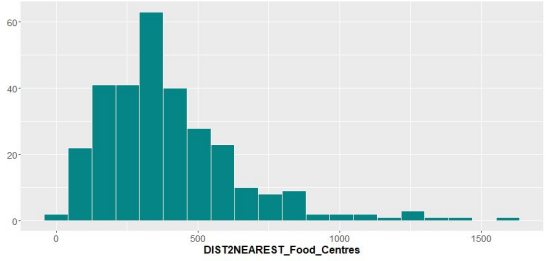


Figure 9: Histogram Plot of a Skewed Variable

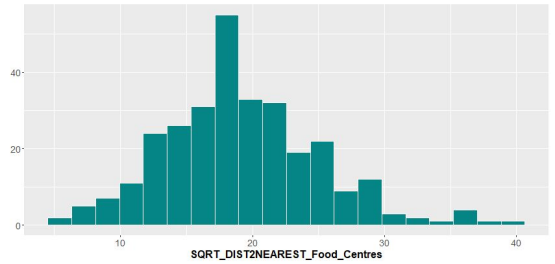


Figure 10: Histogram Plot of a Transformed Variable using Square Root Function

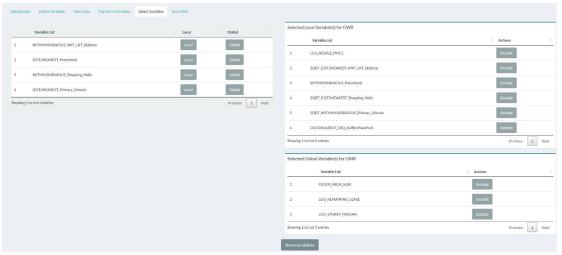


Figure 11: Global-Local Selection for Variables

This page is is reached after completing the transformation, users now can decide on the variables they want to add as global and as local. The difference between a global and a local variables is that global variables have the same degree of variability in the whole of Singapore, examples of such attributes are floor area and lease remaining. These attributes affect the price around the same way no matter when the user is. Local variables, on the other hand, varies to a large degree according to the different points on the map. Number of primary schools and preschools are examples of such local variables. Users should note that global or local selection will only affect the MGWR model as the GWR model classifies all of the variables selected as local. After users bin them into the two categories, they are also able to look through the correlation plot by pressing the button below and decide if any variables should be excluded due to high correlation as it will adversely affect the regression models later.

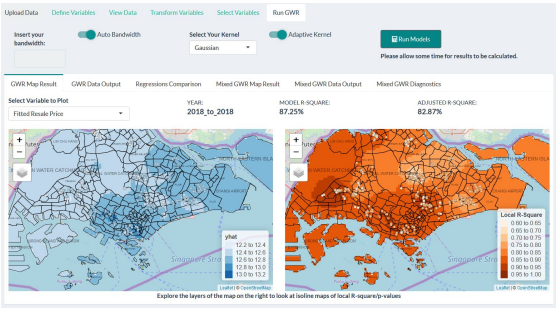


Figure 12: Global-Local Selection for Variables

The final step before running GWR is specifying the bandwidth and the kernel used. The bandwidth can be set manually by the user by “sliding the “auto bandwidth” slider and key in their own bandwidth in the input box below. Kernel can be changed using the dropdown box too. The GWR and MGWR model will then be generated and run, which might take a few while depending on the sample size that was decided previously.

The result on the first sub-tab plots of the local coefficient, p values and R-squares of the different variables that can be selected and viewed separately. This is the GWR result.

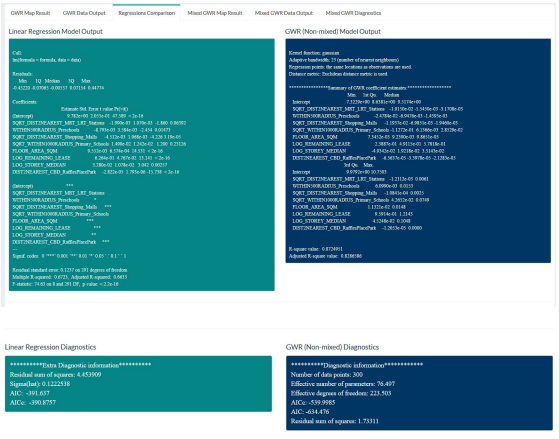


Figure 13: MLR and GWR Results Comparison

Tab 3 of “Run GWR” shows the report generated for MLR and GWR, MLR assumes all variables are global and a coefficient estimate for each variable is obtained as well as its t-statistics and the p-value. The overall R square is displayed also. As MLR assumes local all variables, a range of coefficients estimates are obtained and shown in its quartiles, together with the overall R-square.

More importantly, users are able to see the AICc score of the GWR in the additional diagnostics and can use it to compare with the MGWR result that is discussed below.

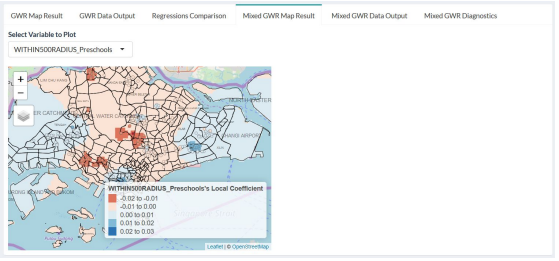


Figure 14: MGWR plots

The next sub-tab shows the MGWR plots of the different variables.

```
**********Summary of mixed GWR coefficient estimates:**********
Estimated global variables :
                FLOOR_AREA_SQM LOG_REMAINING_LEASE
Estimated global coefficients:       0.00945            0.72347
                LOG_STOREY_MEDIAN
Estimated global coefficients:          0.0298
Estimated GWR variables :
                                  Min.     1st Qu.      Median
Intercept                    8.6090e+00  9.0332e+00  9.1967e+00
SQRT_DIST2NEAREST_MRT_LRT_Stations -1.4187e-02 -4.5668e-03 -2.3745e-03
WITHIN500RADIUS_Preschools         -2.0795e-02 -8.7596e-03 -1.2665e-03
SQRT_DIST2NEAREST_Shopping_Malls   -1.2648e-02 -6.7129e-03 -1.7028e-03
SQRT_WITHIN1000RADIUS_Primary_Schools -5.4287e-02  1.8410e-02  3.1977e-02
DIST2NEAREST_CBD_RafflesPlacePark  -4.5697e-05 -3.0316e-05 -2.4806e-05
                                  3rd Qu.     Max.
Intercept                    9.5841e+00 9.7412
SQRT_DIST2NEAREST_MRT_LRT_Stations -6.3004e-04 0.0062
WITHIN500RADIUS_Preschools          7.4905e-03 0.0221
SQRT_DIST2NEAREST_Shopping_Malls    2.4170e-04 0.0038
SQRT_WITHIN1000RADIUS_Primary_Schools 4.6985e-02 0.0789
DIST2NEAREST_CBD_RafflesPlacePark  -1.8392e-05 0.0000
**********Diagnostic information**********
Effective D.F.:  43.65
Corrected AIC:  -500.1
Residual sum of squares:  2.334

**********
Program stops at: 2019-04-12 15:29:36
```

Figure 15: MGWR Diagnostics

The last sub-tab shows the summary of MGWR evaluation. The AICc is contained in under the diagnostic information and users can use this score to compare with the one obtained earlier from GWR. Users might want to switch over to MGWR model instead, if the AIC of MGWR is lower than that of the GWR.

6 RESULTS AND ANALYSIS

The team used some of the preloaded datasets to conduct analysis. The variables that are used are:

Global Variables	Local Variables
Floor_Area_SQM	Preschools
Storey_Median	MRT_LRT_Stations
Remaining_Lease	Primary_School
	Shopping_Malls
	CBD_RafflesPlace

The radius for primary schools in the vicinity of resale HDB units was increased to 1000m. Transformation was also performed on selected sets of the variables.

Transformation	
Log	**Sqrt**
Resale_Price	Dist2Nearest_MRT_LRT_Stations
Storey_Median	Dist2Nearest_Shopping_Malls
Remaining_Lease	Within1000Radius_Primary_Schools
No Transformation	
Within500Radius_Preschools	
Floor_Area_SQM	
Dist2Nearest_CBD_RafflesPlacePark	

Figure 16: Transformation of Variables

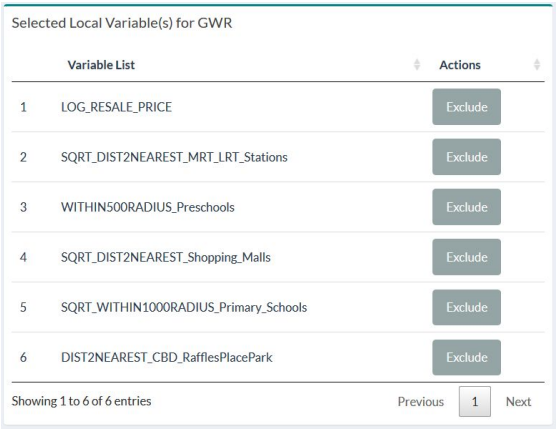


Figure 17: List of Local Variables

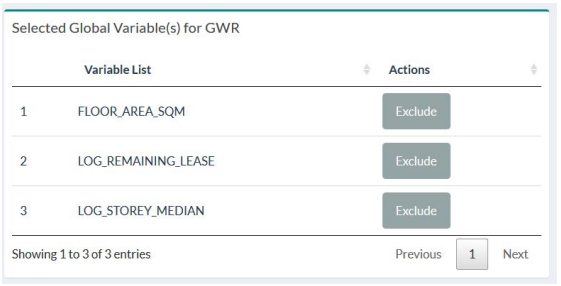


Figure 18: List of Global Variables

The kernel function used was set as the default Gaussian and auto bandwidth was used. The results obtained was as follow:

MLR	GWR	MGWR
R-square: 67.13% **Adj R-square:** 66.6% **AICc:** -702.764 All variables significant at 90% CL	**R-square**: 88.67% **Adj R-square:** 84.32% **AICc:** -964.8	**AICc:** -982.9

Figure 19: Model Comparison

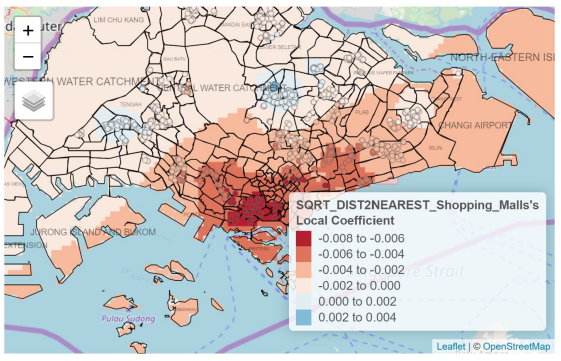


Figure 20: (GWR) Plot of Local Coefficient of Sqrt Distance to Nearest Shopping Mall Variable

The GWR reveals that in areas nearer to CBD and the East, the housing prices are actually more sensitive to the distance to nearest shopping mall (Fig. 20). There is a greater decrease in price as the distance to the nearest shopping mall increase from a HDB resale flat. This possible suggests that buyers looking for flats within these regions view

convenient access to shopping as higher priority when making decisions. Given the density of shopping options within these regions, especially in the central areas, it is no surprise that buyers looking at these locations are likely to be more interested in the shopping options aspect when buying a flat, and hence willing to pay greater premiums for for closer and more convenient access to suit their preferences.

It can be further noted that there are some regions that present a positive correlation between resale prices and distance to nearest shopping mall. Possible reasons for this include buyers looking at these areas show less interest towards having convenient shopping options. Another likely possibility that buyers do not see distance to shopping malls as a disadvantage could be due to affluence whereby they are more likely to use ways of transport such as driving, to get to shopping malls. In such a case, being close to a shopping mall is not such a great incentive as compared to individuals who have a preference for walkable distances to their shopping options.

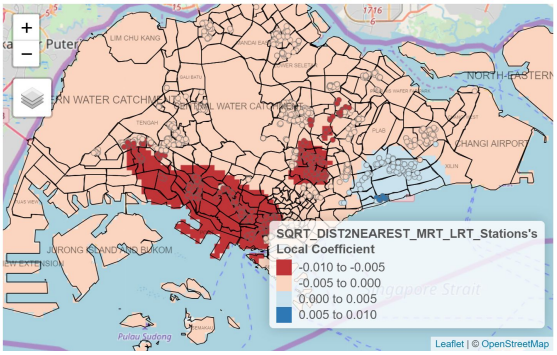


Figure 21: (GWR) Plot of Local Coefficient of Sqrt Distance to Nearest MRT/LRT Variable

Looking at a different variable, Distance to Nearest MRT/LRT Station, a different distribution of coefficient estimates can be observed. Especially noticeable is the south-west region, where resale flat buyers are more sensitive to a flat's distance to the nearest MRT station. This presents different possible insights, depending on perspective. Firstly, this could be an indication of a high demand for access to the MRT stations along that area, which stretches from stations such as Clementi, up till Tanjong Pagar/Raffles Place. Given the heavy peak-hour demands of the East-West line and its direct access to the bustling city centre where many offices are located at, it is not surprising that having access to the stops along this MRT line will bring about greater convenience to buyers of houses near these stations. Another perspective could be that access to these MRT stations via vehicular transport options may not be convenient. For example, overcrowded buses to/from MRT stations during peak hours could be a factor that incentivise commuters to walk instead. As such, buyers are more willing to pay greater premiums for nearer distances to MRT stations in these areas.

However, in Fig. 21, in the eastern regions around Bedok, it can also be observed that there is positive correlation between resale price and distance to nearest MRT station. A reason for this could be that there are few MRT transport options within that region, and buyers who are looking for properties in this region do not see accessibility to MRT transport as an important factor. Instead, transport via buses could suffice. The points located near the East Coast Park region, near to the coast, could be evidence that connectivity to the MRT network is not a priority for the buyers of those resale flats.

Comparing the AICc values of the models, it is shown that GWR is far more superior in performance as compared to the MLR and MGWR was also slightly better than the GWR due to the rendering of variables that does not have too much spatial variations.

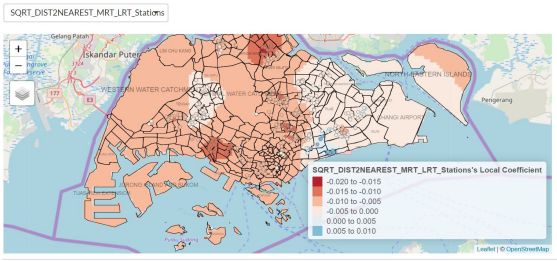


Figure 22: MGWR Model on 3-Room Resale HDB

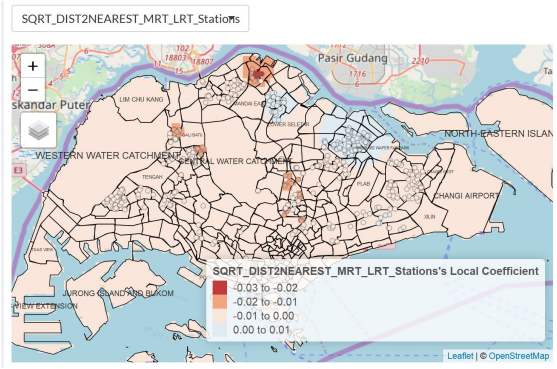


Figure 23: MGWR Model on 5-Room Resale HDB

Using the same attributes for the MGWR model on 3-room resale HDB and 5-room resale HDB data yields interesting results (Fig. 22 and 23). The 5-room HDB resale flats observes a smaller range for the coefficient estimates as compared to the 3-room HDB resale flats. One possible explanation is that 5-room HDB flats occupants and buyers are more wealthy in general and it might be possible that a larger proportion of them have their own family vehicles, making them not be too affect if the nearest MRT/LRT stations are more further away as compared to 3-room HDB flat buyers and occupants.

7 DISCUSSION

EzModel's application was showcased at Singapore Land Authority (SLA) geospatial industry centre to some of the industry experts and interested parties on the 8th April 2019. The common consensus among the visitors are that the tool is rather helpful for policy planners and economists as it provides a great degree of freedom for the these group of user, who are likely to have some of their own data, to upload these data points and visualise how they might affect resale value of HDB flat.

Some notable comments were the preloaded data was helpful as these are many of the existing variables that policy planners are working with already. Some of the users found the transformation functions that EzModel has to be very practical as they would not have to use other data visualisation tool to plot a histogram or a box plot to look at the skewness of the attributes. A small number of users pointed out that in the "Select Variables" tab, a better way would be allowing users to do a multi-select/checkbox options to allow users to input the variables as local or global instead of adding the variables to the two groups one by one, which will in fact, speed up the process.

The correlation plot in the same tab was also critiqued on by some of the visitors as the attribute name was mashed into the middle of the correlation plots. This confused some of them as a number of them took some time in comprehending the plot. One of the visitors commented that he is used to seeing the variable names to be at the edges of the plot instead. The team took note of this and discussed that this can be fixed easily with limiting the correlation matrix plot to maybe on the top triangle by specifying the 'type' parameter.

Some of the visitors were also particularly interested in "elite" primary schools. These are the primary schools that usually see higher perceived academic performances and linked to higher quality of education as compared the others. Housing prices are usually higher in within 1km of these schools[3]. To this comment, the team agreed that adding an additional layer of filter to sieve out schools that match these criteria would value-add and add an extra depth to the analysis of resale housing prices. However, this can also be done by the user themselves. Just uploading a CSV file containing the coordinates of these schools and choosing this variable will suffice.

8 FUTURE WORK

While GWR and MGWR models can be used to a great degree in looking into what are the attributes that affect HDB resale prices, time could also be a factor in the price modelling. It is understandable that housing prices do not change significantly in the short run, but if we are looking at a span of 10 years or more, time would very much affect prices. Over a long period of time, infrastructure and amenities will be gradually added to an area which will increase or even decrease its prices to a certain degree. In the past, resentment and

[3]https://www.straitstimes.com/opinion/how-school-proximity-affects-house-prices-in-singapore

complaints of newly added amenities have been observed before, and these are usually attributed to newly added infrastructures such as nursing homes and workers' dormitories[4].

Singapore, being a developed country, still sees some of its public infrastructure, such as newer MRT lines being developed rapidly. At the same time, we are also facing the problems that any developed nation is facing - aging population. With the need for more eldercare facilities such as nursing homes in the already limited land area, SLA has to plan carefully where to locate these facilities. Plans for such facilities, both the MRT stations and the nursing homes are examples of new infrastructure that will affect housing prices. As such, if we are to look in the long run, it is vital to include a temporal element within the data and its analysis. A geographical and temporal weighted regression (GTWR) can be explored to account not only for spatial changes, but also spatial changes across time.

9 CONCLUSION

EzModel demonstrates how GWR and MGWR can be implemented to solve regression problems that has a spatial element to the variables. For housing prices in this case, it has generally led to greater accuracy, resulting in closely predicted prices and more in depth details of the what are the spatial factors that can explain these variations and differences.

GWR and MGWR techniques can also be extended to other analysis and used more widely. Some other possible applications of GWR and MGWR would be in the healthcare industry, helping pathologist and epidemiologist to monitor spread of diseases and infections with reference to geographical factors such as climate and other environmental characteristics.

[4] Unhappiness resulting from newer infrastructure built in the vacinity of existing living https://www.straitstimes.com/singapore/housing/unhappiness-over-sengkang-temple-with-columbarium-6-other-cases-of-residents

ACKNOWLEDGMENTS

The authors would like to thank Professor Kam Tin Seong from Singapore Management University for his kind guidance, support and patience in the process of working on this project. The advice and feedback that were brought up by him during consultations played a pivotal role in the ideation and implementation stages of the product prototype.

REFERENCES

[1] Zezhou, Huang & Ruishan, Chen & Di, Xu & Wei, Zhou. (2017). Spatial and hedonic analysis of housing prices in Shanghai. DOI: https://doi.org/10.1016/j.habitatint.2017.07.002.

[2] Binbin Lu, Paul Harris, Martin Charlton & Chris Brunsdon (2014) The GWmodelR package: further topics for exploring spatial heterogeneity using geographically weighted models,Geo-spatial Information Science, 17:2, 85-101, DOI: 10.1080/10095020.2014.917453.

[3] Waters, Nigel. (2017). Tobler's First Law of Geography. DOI: 10.1002/9781118786352.wbieg1011.

[4] Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 33(2), 261–304. https://doi.org/10.1177/0049124104268644