travel modelling group

tmg

MODELLING AVERAGE DWELLING VALUE OF TORONTO DISSEMINATION AREAS

Ethan Baron
Eric J. Miller, Ph.D.

April, 2021

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

UTTRI

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

Ethan would like to thank Prof. Eric Miller for the opportunity to pursue this project, enriching words of advice, and continued support along the way.

1 Introduction

This report outlines an attempt to model the average dwelling value of dissemination areas (DAs) in the City of Toronto. The report presents a description of the variables tested and their predicted feature importance from using a random forest regressor. Next, the development process of a linear regression model is described and results for the chosen model are discussed. Finally, several types of geographic models are estimated and discussed.

2 Data

See a table summarizing all the variables used in Appendix A.

2.1 Census Data (provided by Yu (Billy) Zhang)

DA-level data from the 2016 census is used, including average dwelling value including all dwelling types within the DA (the dependent variable), after-tax income, employment rate, housing unit age, number of rooms, and average number of people per room. Figure 1 shows the variations in the dependent variable across the city of Toronto. Figures 2 and 3 present histograms of the original dependent variable, and its log-transformed form, respectively. As indicated in Figure 2 the average dwelling value distribution is left skewed, with a long right-hand tail. Transforming the dwelling values through a logarithmic transformation produces a much more symmetric distribution (Figure 3), which is approximately normally distributed and so much more suitable as the dependent variable in the statistical modeling efforts presented in this report.

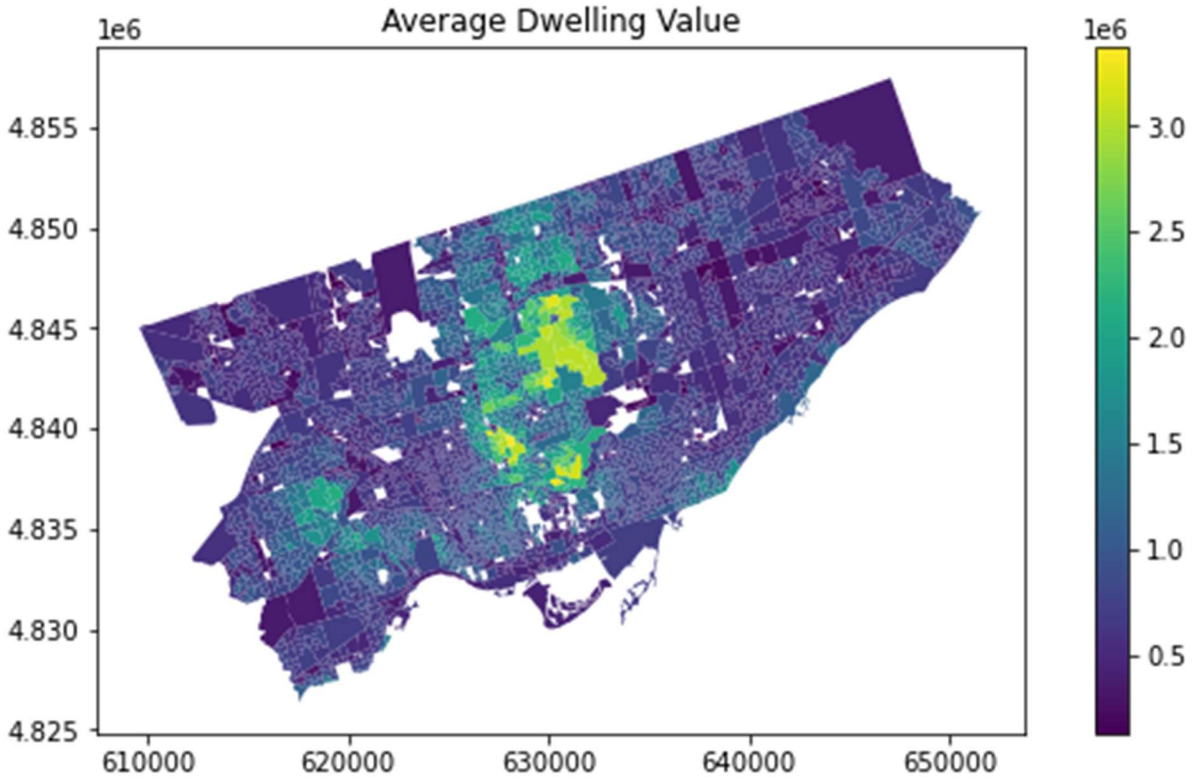

Figure 1: Map of Average Dwelling Value in Toronto DAs (Million $) from 2016 census. White = unavailable.

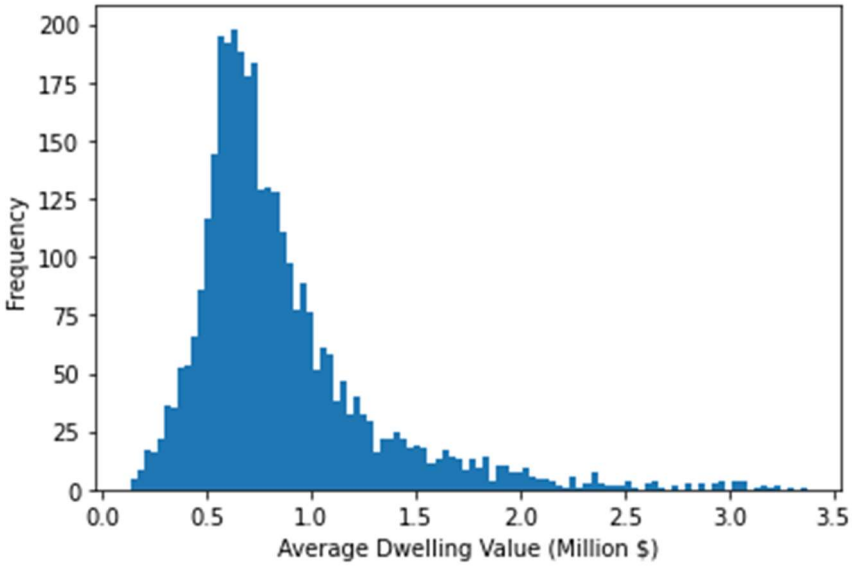

Figure 2: Histogram of Average Dwelling Values for Toronto DAs

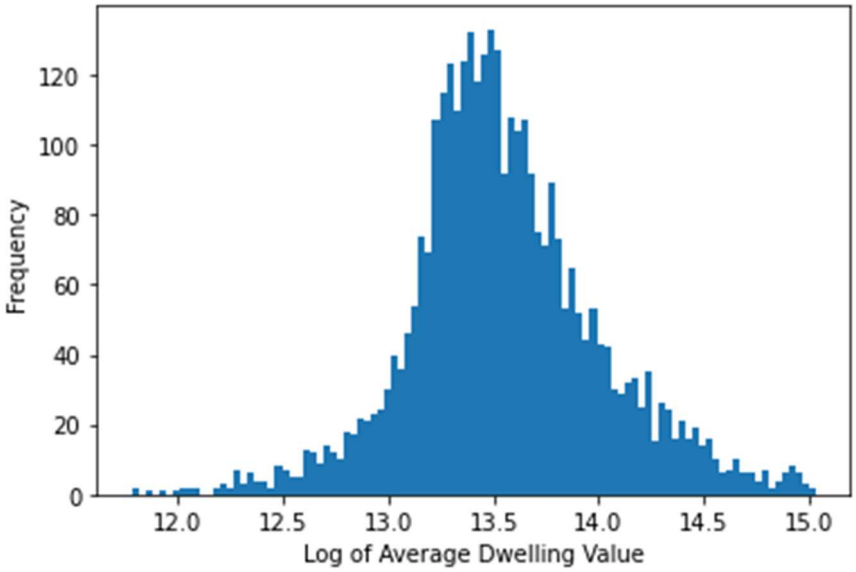

Figure 3: Histogram of Log-Transformed Average Dwelling Values for Toronto DAs

2.2 DMTI Land Use Data

This dataset (provided by Yu (Billy) Zhang, UTTRI PhD student) includes historical information on land use patterns within DA, including the proportion of the area used for residential purposes, commercial purposes, institutional purposes, industrial purposes, parks, open space, and water.

2.3 Accessibility Measures

Distances between the DA centroids and various categories of points of interest were calculated after projecting into UTM coordinates. Table 1 presents a summary of these categories and the data sources, while a box-and-whiskers plot of these accessibility variables is shown in Figure 4.

Table 1: Summary of Point of Interest Categories and Data Sources.

Category	Data Source	Note
TTC subway and GO rail stations	Coordinates provided by Reuben Briggs (Data Management Group)	Vaughan subway extension station coordinates obtained from Wikipedia
The Bay & King central business district (CBD)	Google Maps	43.64865 ° N, 79.38025 ° W
Highway exits	GTAModel road links and nodes provided by Yang (Luna) Xi (Travel Modelling Group)	Links filtered using the attached volume-delay functions
Community centres	Websites of Brampton, Durham Region, Mississauga, Toronto, and York Region	Community centre addresses geocoded using ArcGIS geocoder
Shopping centres, "major" shopping centres	Wikipedia's "List of largest shopping malls in Canada"	"Major" shopping centres have area of over 1 million square feet
Cinemas	DMTI points of interest dataset provided by Joven Chew (UofT undergraduate student)	NAICS code 51213
Bowling Alleys		NAICS code 71395
Bars		NAICS code 72241

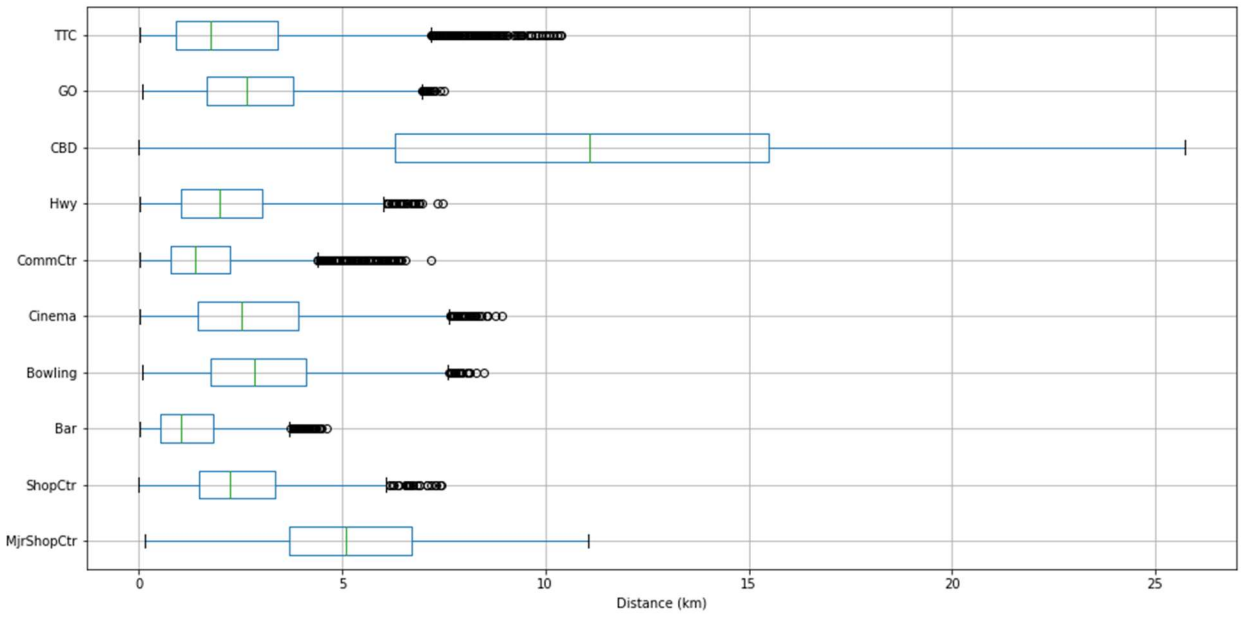

Figure 4: Box-and-Whiskers Plot for Distance Variables.

3 Random Forest Feature Importances

A random forest regressor is used to establish initial estimates of the relative importance of the different independent variables in predicting average dwelling value. The feature importance, permutation importance, and linear correlation coefficients for the variables are listed in Table 2. These are listed in decreasing order of permutation importance as this is a metric with lower bias towards high-cardinality variables than feature importance.

Table 2: Permutation Importances, Feature Importances and Correlations for Variables in Random Forest Regression

Feature	Permutation Importance	Feature Importance	Correlation
AftTaxInc	0.7535	0.5728	0.5975
GO_Dist	0.0934	0.1390	0.2582
EmplRate	0.0391	0.0678	0.1815
Bar_Dist	0.0371	0.0379	-0.0165
TTC_Dist	0.0323	0.0263	-0.2787
CBD_Dist	0.0223	0.0329	-0.2798
Bowling_Dist	0.0108	0.0174	-0.0052
MjrShopCtr_Dist	0.0099	0.0207	-0.0485
HAge	0.0079	0.0127	0.1768
P_RES	0.0063	0.0104	0.1793
ShopCtr_Dist	0.0053	0.0080	0.1184
Cinema_Dist	0.0047	0.0147	-0.1081
CommCtr_Dist	0.0025	0.0059	-0.2103
P_INST	0.0008	0.0049	-0.1676
Hwy_Dist	0.0007	0.0061	0.0114
RoomsPerH	0.0002	0.0016	-0.0292
P_COMM	0.0002	0.0006	-0.1021
P_WATER	0.0001	0.0005	0.0053
PplPerRoom	-0.0002	0.0020	-0.0029
P_IND	-0.0003	0.0011	-0.0295
P_PARK	-0.0007	0.0034	-0.0261
P_OPEN	-0.0021	0.0136	-0.0935

It is clear and unsurprising that after-tax income is the most useful variable and is highly correlated with average dwelling value for Toronto DAs. The land use indicators perform relatively poorly, and the accessibility variables show moderate usefulness, with the distance to the nearest GO train station seeming most promising.

4 Linear Regression Model

Linear regression models are commonly used as housing hedonic price models. Here, variables are iteratively added to the model in an order informed by the expected importance from Table 1. In adding after-tax income (AftTaxInc) and distance to downtown (CBD_Dist), the following three approaches were also tested and compared using log-likelihood ratio tests:

1. Polynomial fits of degrees $n \in [1, 8]$ using stats v3.6.2
2. Piece-wise linear regressions with number of intervals $n \in [1, 6]$ using (Muggeo 2021)
3. Smoothing splines with 10 degrees of freedom using stats v3.6.2

Log-transforming income was also tested but did not show any more promise than using the original variable.

For AftTaxInc, the simple linear fit provided an acceptable performance compared to higher-degree polynomial fits, piecewise linear regressions, and splines. For CBD_Dist, a piecewise linear regression is used. The chosen model is obtained by selecting variables with statistically significant estimated coefficients. It includes 13 dependent variables with linear relationships and three segments for the piecewise linear dependence on CBD_Dist. The model's results are listed in Tables 3 and 4. All coefficients included were significant at the 1% level. A plot of the piecewise linear dependence on CBD_Dist is shown in Figure 5.

Table 3: Estimated Coefficients for Linear Regression Model.U1.CBD_Dist and U2.CBD_Dist are "difference-in-slopes coefficients" with psi1=2.961 km and psi2 = 6.297 km.

Coefficient	Estimate
Intercept	12.6653
AftTaxInc ($100k)	2.1555
CBD_Dist (km)	0.2426
U1.CBD_Dist (km)	-0.3225
U2.CBD_Dist (km)	0.0767
EmplRate (%)	-0.0058
LU_P_COMM (proportion)	-0.4241
LU_P_INST (proportion)	-0.3611
LU_P_OPEN (proportion)	-0.4404
TTC_Dist (km)	-0.0237
GO_Dist (km)	0.0420
CommCtr_Dist (km)	-0.0492
Cinema_Dist (km)	0.0114
Bowling_Dist (km)	0.0126
Bar_Dist (km)	0.0423
MjrShopCtr_Dist (km)	0.0134

Table 4: Metrics for Linear Regression Model with After-tax Income.

# of observations	3390
# of degrees of freedom	18
Log-likelihood	-1136.39
Adjusted r-squared	0.448929
Mean squared error	0.11447
Residual sum of squares	366.12

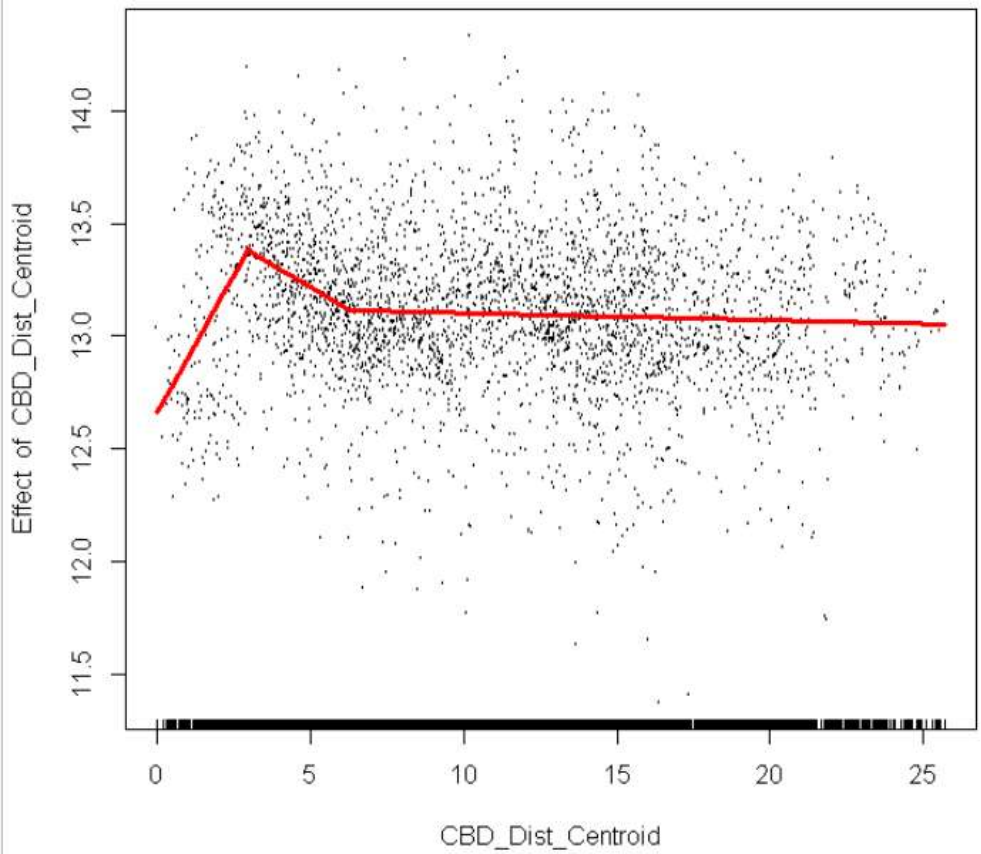

Figure 5: Plot of Piecewise Linear Dependence Estimated for CBD_Dist.

As seen in Figure 5, dwelling values in the immediate vicinity of the CBD tend to increase in value with distance from the CBD, possibly reflecting disamenities of living within the CBD itself or differences in dwelling type distributions. Dwelling values then fall beyond approximately 3km of the CBD, as expected from classic Bid Rent theory. However, the effect of distance from the CBD on dwelling value diminishes sharply after about 6.3km, perhaps as other measures of accessibility and centrality dominate.

- Unsurprisingly, the coefficient for after-tax income is positive and numerically strong. This confirms the expectations from Table 2.
- The negative sign for employment rate is somewhat surprising as it suggests that DAs with lower employment are home to higher-value dwellings. This might reflect patterns in family structure or life-cycle stage effects. Alternatively, the effect might be confounded by the relationship between employment rate and after-tax income.
- The land use parameters included all have negative coefficients. This might be explained by the hypothesis that DA's with almost exclusively residential land use might have higher average dwelling values, as these DAs tend to include detached houses rather than apartment buildings or condominiums.

- Proximity to TTC stations and commercial centres values increases average dwelling value, as evidenced by the negative distance parameters. However, proximity to other activity centres tested, including entertainment and major shopping centres, decreases average dwelling value. Presumably, this is due to other land use or neighbourhood factors.
- Prices generally rise as one moves away from GO stations. This might be reflecting their typical locations within rail corridors, which are generally unattractive residential locations.

5 Geographic Models

This section outlines preliminary results when applying several types of geographic extensions to the standard ordinary least squares (OLS) linear regression: spatial autoregressive (SAR), geographically weighted regression (GWR), and mixed geographically weighted regression (MGWR). These are estimated using the "mgwrsar" package in R (Geniaux & Martinetti, 2018a). Examples of how to use this package are provided in (Geniaux, 2018), and the complete theoretical explanation of these models can be found in (Geniaux & Martinetti, 2018b).

For these models, a weighting kernel (e.g. bisquare or gaussian) and bandwidth (which can be adaptive or fixed) must be specified. A fixed bandwidth specifies a particular distance which is held constant throughout the space. An adaptive bandwidth specifies the number of nearest neighbours and allows for the kernel to vary in size from point to point (this is helpful when there are changes in the density of points). Here, adaptive bandwidths are used, as recommended by (Pace et al, 1998).

5.1 Spatial Autoregressive Model Example

The spatial autoregressive (SAR) model expands on the ordinary least squares (OLS) model by including a term with the weighted observations for nearby points, with coefficient λ. Here, a basic SAR model is estimated with AftTaxInc as the only independent variable. An adaptive gaussian kernel with a bandwidth of 50 nearest neighbours is used for the weighting matrix. The results for this basic SAR model are shown in Table 5.

Table 5: Estimated Coefficients and Metrics for Simple Spatial Autoregressive Model.

Coefficient	**Estimate**
Intercept	5.949981
AftTaxInc ($1k)	0.01530183
lambda (weighed neighbours)	0.5248506
Metric	**Value**
# of observations	3390
# of degrees of freedom	3
Residual sum of squares	350.8468

5.2 Mixed Geographically Weighted Regression Model Example

The mixed geographically weighted regression (MGWR) model is a generalization of the geographically weighted regression (GWR) model. The GWR expands on the OLS method by allowing the coefficients to vary spatially, fitted by using a weighted average of nearby points. The model estimates coefficients at each point. It is possible to have multiple coefficients varying spatially, including the intercept, with different or equal bandwidths. The MGWR expands on GWR by allowing for the inclusion of fixed coefficients, as per the standard OLS.

Here, a basic MGWR model is estimated which includes a fixed coefficient for AftTaxInc and a free-to-vary-spatially intercept. Again, an adaptive gaussian kernel with a bandwidth of 50 nearest neighbours is used for the weights. The results are listed in Table 6. A map of the estimated geospatial variation in the intercept is shown in Figure 6. A few interesting observations include:

- Lower intercepts in the CBD itself are consistent with previous linear regression results.
- High-priced housing runs north from downtown along the Yonge St. corridor, including neighbourhoods such the Annex, Rosedale, the Bridle Path, Forest Hill, and Lawrence Park. This might reflect historical patterns in high-value dwelling in Toronto.
- Lower intercepts are observed in northwest Etobicoke and much of Scarborough, as might be expected.
- The historical so-called "industrial-U" following the railway lines from the north-west and north-east into downtown Toronto, broken by gentrified areas to west (Liberty Village, etc.) and east (Cabbagetown, Riverdale, the Beach) can be observed in terms of the lower (red) intercept values.

Table 6: Estimated Coefficients and Metrics for Simple Mixed Geographically Weighted Regression.

Coefficient	**Estimate**	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Intercept		12.684	12.833	12.944	12.971	13.104	13.386
AftTaxInc ($1k)	0.01815214						
Metric	**Value**						
# of observations	3390						
Residual sum of squares	343.4045						

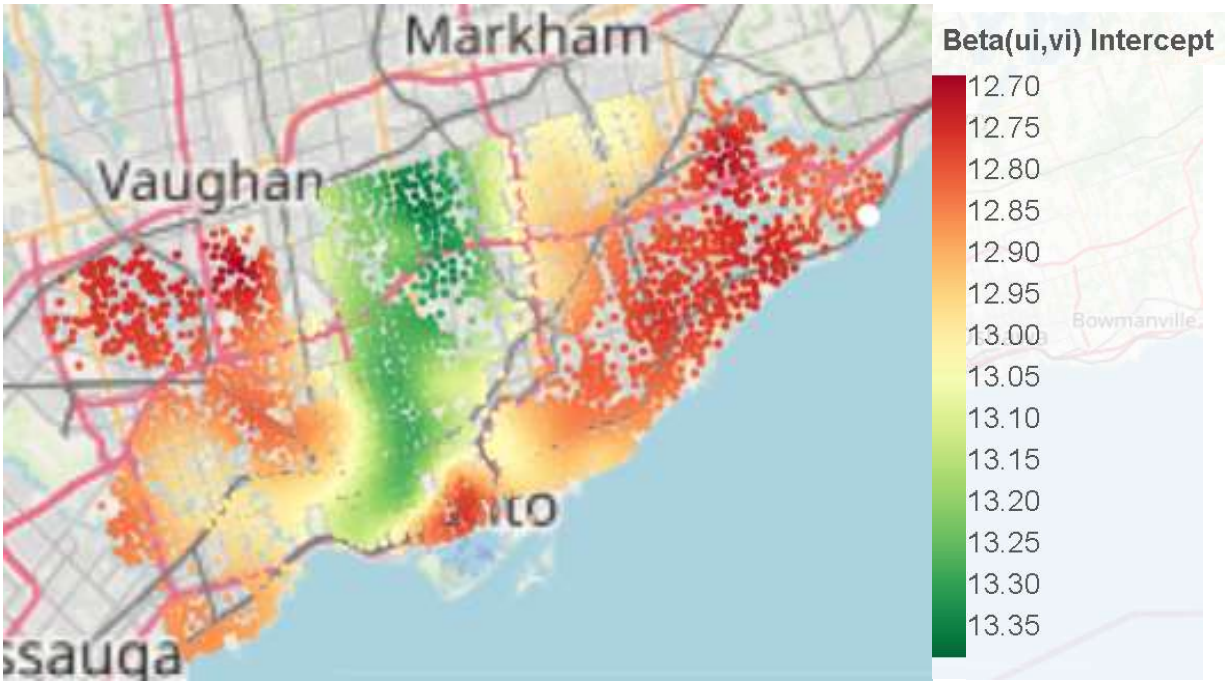

Figure 6: Map of Estimated Intercepts in Basic MGWR.

6 Discussion

The SAR model has surpassed the benchmark residual sum of squares from the linear regression model, despite using only three estimated parameters rather than 19. However, the predictive power of SAR depends on the reliable availability of information for nearby points. Hence, it is more useful as an explanatory model rather than a predictive one. In predictive settings, a temporal element can be included, as per (Pace et al, 1998).

Meanwhile, the MGWR model achieved slightly better results than the SAR but includes many more degrees of freedom, as it estimates local intercepts. Similar to the SAR, it is most useful as an explanatory tool to identify patterns in the data, such as how the effects of certain variables vary spatially. However, by estimating local coefficients, the risks of overfitting seem substantial, especially when multiple coefficients are allowed to vary spatially. Including a temporal aspect is also possible via the geographically and temporally weighted regression (GTWR) proposed by (Huang et al., 2010).

7 Future Work

Several potential areas of future work include:

- Use predicted DA dwelling values or similar modelling techniques to develop a property-level asking price model.
- Completing the development of MGWR and/or SAR models explaining average DA dwelling values in Toronto.
- Evaluating the ability of these geographic models to generalize usefully and accurately in forecasting applications, especially by incorporating temporal elements.

References

Geniaux, G. 2018. "Examples of basic uses of mgwrsar package." https://cran.r-project.org/web/packages/mgwrsar/vignettes/mgwrsar-basic_examples.html (accessed April 22, 2021).

Geniaux, G., & Martinetti, D. 2018a. "Package 'mgwrsar'." https://cran.r-project.org/web/packages/mgwrsar/mgwrsar.pdf (accessed April 22, 2021).

Geniaux, G., & Martinetti, D. 2018b. "A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models." *Regional Studies and Urban Economics*, 72, 74-85. https://doi.org/10.1016/j.regsciurbeco.2017.04.001 (accessed April 22, 2021).

Huang, B., Wu, B., and Barry, M. 2010. "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices." *International Journal of Geographical Information Science*, 24(3), 383-401, https://doi.org/10.1080/13658810802672469 (accessed April 22, 2021).

Muggeo, V. M. R. 2021. . "Package 'segmented'." https://cran.r-project.org/web/packages/segmented/segmented.pdf (accessed April 22, 2021).

Pace, R. K., Barry, R., Clapp, J. M., and Rodriquez, M. 1998. "Spatiotemporal autoregressive models of neighborhood effects." *The Journal of Real Estate Finance and Economics*, 17, 15-33. https://doi.org/10.1023/A:1007799028599 (accessed April 22, 2021).

8 Appendix A: Summary of Variables Used

Variable	Description	Source	Min	Mean	Max
AvgDwVal	Average dwelling value in DA ($)	2016 Census	130128	854365.86	3374128
AftTaxInc	Median after-tax income in DA ($)	2016 Census	12176	32135.94	89600
EmplRate	Employment rate in DA (%)	2016 Census	14.3	69.38	95.1
HAge	Mean housing unit age in DA (yrs.)	2016 Census	5.24	40.91	62.53
RoomsPerH	Average number of rooms per household in DA	2016 Census	0	6.62	11.8
PplPerRoom	Average number of people per room in DA	2016 Census	0	0.47	1.2
LU_P_RES	Proportion of DA land use residential	DMTI Land Use	0	0.763	1
LU_P_COMM	Proportion of DA land use commercial	DMTI Land Use	0	0.018	0.995
LU_P_INST	Proportion of DA land use institutional	DMTI Land Use	0	0.058	1
LU_P_IND	Proportion of DA land use industrial	DMTI Land Use	0	0.051	0.999
LU_P_PARK	Proportion of DA land use parks	DMTI Land Use	0	0.080	1
LU_P_OPEN	Proportion of DA land use open space	DMTI Land Use	0	0.025	0.997
LU_P_WATER	Proportion of DA land use water	DMTI Land Use	0	0.005	0.956
TTC_Dist	Distance from DA centroid to nearest TTC subway station (km)	See Table 1	0.028	2.418	10.397
GO_Dist	Distance from DA centroid to nearest GO rail station (km)	See Table 1	0.082	2.874	7.492
CBD_Dist	Distance from DA centroid to Bay & King CBD (km)	See Table 1	0	11.168	25.718
Hwy_Dist	Distance from DA centroid to nearest highway exit (km)	See Table 1	0.034	2.119	7.475
CommCtr_Dist	Distance from DA centroid to nearest community centre (km)	See Table 1	0.028	1.662	7.198

ShopCtr_Dist	Distance from DA centroid to nearest shopping centre (km)	See Table 1	0.006	2.469	7.429
MjrShopCtr_Dist	Distance from DA centroid to nearest major shopping centre (km)	See Table 1	0.141	5.186	11.040
Bar_Dist	Distance from DA centroid to nearest bar (km)	See Table 1	0.023	1.261	4.615
Bowling_Dist	Distance from DA centroid to nearest bowling centre (km)	See Table 1	0.078	3.094	8.481
Cinema_Dist	Distance from DA centroid to nearest cinema (km)	See Table 1	0.026	2.868	8.932