6-2022

Taxi travel time based Geographically Weighted Regression Model (GWR) for modeling public housing prices in Singapore

Yi’an WANG

Fangyi CAI

Shih-Fen CHENG
Singapore Management University, sfcheng@smu.edu.sg

Bo WU

Kai CAO

Citation

WANG, Yi’an; CAI, Fangyi; CHENG, Shih-Fen; WU, Bo; and CAO, Kai. Taxi travel time based Geographically Weighted Regression Model (GWR) for modeling public housing prices in Singapore. (2022). *Proceedings of the 29th International Conference on Geoinformatics, Beijing, China, 2022 June 15 - 18*.
Available at: https://ink.library.smu.edu.sg/sis_research/7708

Taxi travel time based Geographically Weighted Regression Model (GWR) for modeling public housing prices in Singapore

Yi an WANG[1,2,3]; Fangyi CAI ; Shih-Fen CHENG[5]; Bo WU[6]; Kai CAO [1,2,3]*

1. School of Geographic Sciences, East China Normal University, Shanghai, China
2. Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China
3. Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai, China
4. Department of Geography, National University of Singapore, Singapore
5. School of Computing and Information Systems, Singapore Management University, Singapore
6. School of Geography and Environment, Jiangxi Normal University, Nanchang, Jiangxi, China

2022 29th International Conference on Geoinformatics | 979-8-3503-0988-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/GEOINFORMATICS57846.2022.9963833

Abstract—**In this research, a taxi travel time based Geographically Weighted Regression model (GWR) is proposed and utilized to model the public housing price in the case study of Singapore. In addition, a comparison between the proposed taxi data driven GWR and other models, such as ordinary least squares model (OLS), GWR model based on Euclidean distance and GWR model based on public transport travel time, have also been carried out. Results indicates that taxi travel time based GWR model has better fitting performance than the OLS model, and slightly better than the Euclidean distance-based GWR model, however, it is not as good as the GWR model based on public transport travel time according to the metric of Adjusted R^2. These experiments indicate that the public transport travel time may has a major part to play in modeling the public housing resale prices compared to taxi travel time or driving time, and both the taxi travel time and public transport travel time can better explain the public housing resale prices in Singapore compared to Euclidean distance in the GWR modeling.**

Keywords—hedonic model; GWR; public housing prices; taxi travel time

I. Introduction

The analyses of housing prices from spatial and/or temporal perspective have received a lot of attention in the past decades [1-5]. Many researchers have employed spatial analysis approaches, such as hotspot analysis and kernel density estimation [6], to model housing prices. Driven by the technological innovation, more data sources like spatio-temporal big data [7] and remote sensing images [8], and methods, such as machine learning, and deep learning [9], have been introduced to help explore and explain the spatio-temporal variation of housing prices in different research areas.

The most commonly used model in housing price modeling related studies is the hedonic model [10]. Initially, researchers often utilized the ordinary least squares model (OLS) to reveal the change of house prices. In addition to these socio-economic indicators that are affecting the housing prices, such as government policies [11,12], economy and population dynamic changes [13]; micro-scale level spatial and non-spatial characteristics have also been widely considered while modeling housing prices, such as the distances to bus stops [14], distances to CBD area [15], floor level [16], plot ratio [17], age of the property [18].

As a global model, OLS is not able to well reflect the spatial non-stationarity, which is understood as part of the nature of the spatial variation of housing prices in general. Researchers started to apply geographically weighted regression model (GWR) with different forms to help better model the house prices in a variety of researches [19]. Most existing studies based on GWR models used Euclidean distance matrix to measure the spatial patterns of house prices across different regions, while some scholars [20] argued that non-Euclidean distance, including road network distance and travel time based on road speed limit, performs better than Euclidean distance in estimating and explain the housing prices variation. The widely used smartcards in commuting and the big volume of smartcard transactions records enriched the approaches to capture people's actual travel time. In addition to help better define the spatial independent variables in global hedonic models, it can also help to improve the design of the GWR model. For example, Cao et al [7] have successfully utilized the transaction records of Singapore's transportation smartcards, namely EZ-link card, to generate the spatial weight matrix based on public transport travel time to help improve the performance of GWR model. However, there is still lack of studies exploring how big data could contribute to the modeling of housing prices. With no doubt, big data and big data analytics bring tremendous opportunities to help better address various issues in the era of the big data.

In this study, a Taxi travel time based GWR model is proposed and utilized to model the public housing price in Singapore. Moreover, a comparison between the proposed taxi data driven GWR and other models, such as OLS model, general GWR model based on Euclidean distance and GWR model based on public transport travel time considering the same independent variables, will also be performed. Both spatial and non-spatial variables will be considered in the experiment. The overall structure of this article includes five sections. Apart from Section I, which is introducing the research background and relevant studies, Section II introduces the data sets and the research area, then Section III explains the methods and models that are employed in this research, followed by an analysis and comparison of the results obtained based on different models in Section IV. The final Section V summarizes this research and discusses the limitations as well as the future directions of this research.

II. Research Area and Datasets

A. Research area

Singapore is a highly populated city state in Asia. By 2020, 5.686 million people in Singapore reside in a total of around 728.6 square kilometers area. The types of real estate in Singapore can be divided into the public housing and private housing in general. The public housing is developed and managed by the Singapore Housing Development Board (HDB), which accommodates more than 80% of the total Singapore population. The private housing in Singapore includes condominium apartments and landed houses, which are in general more expensive and just serving the minority of the overall population. In this research, given these characteristics of public housing in Singapore and the data availability, the resale market prices of HDB flats are investigated in this study.

B. Datasets and Preprocessing

1)Singapore public housing transaction records

The HDB flats transaction records collected in this research include 218,560 public housing transactions happened in 2011. The dataset contains detailed information of each flat such as street address, flat age, flat area, and floor level as well as transaction time and price. The HDB buildings were geocoded to obtain the geographic coordinates. And the transaction data were also aggregated into 331 model traffic zones (MTZ) to support the further research. Figure 1 indicates the spatial distribution of HDB flats at the level of MTZs in Singapore. These MTZs in dark red are MTZs with HDB blocks, while the other MTZs in grey have no HDB blocks inside.

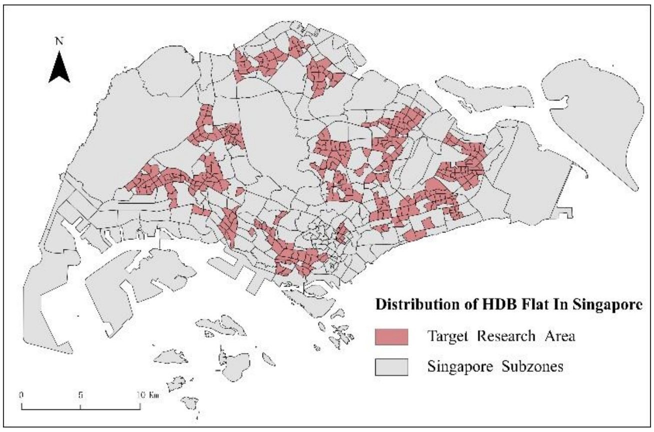

Fig 1. Distribution of HDB Flat Transaction Data

2)Singapore taxi travel dataset

The taxi data set contains 3.2 million taxi operation records collected in seven consecutive days from June 15 to June 22, 2011. Each record consists of unique ID of taxi , start time, end time, travel distance, cost, duration, longitude and latitude under the WGS84 coordinate system. Records that have extremely long travel time were eliminated from the dataset to ensure the accuracy of aggregated results. All the datasets have also been anonymized according to the private policy. In addition, an origin-destination (OD) time matrix is generated based on the calculated time between each pair of MTZs. Linear interpolation approach was used to estimate the travel time of those MTZ pairs without direct travel records.

3)Singapore facility dataset

To support the modeling of public housing prices, several facilities datasets were collected as the potential spatial influence factors of housing prices [21], including parks, major hospitals, top tier primary schools, CBD area, MRT stations. The distances between these facilities and MTZs were calculated to be utilized in the model as spatial variables. In addition, the bus stop number in each MTZ was also counted as a variable.

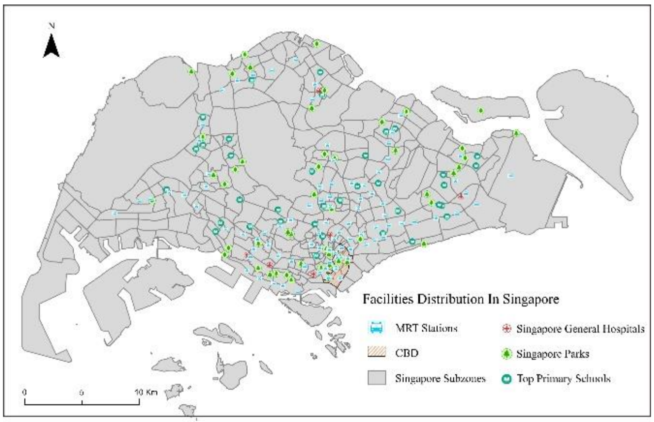

Fig 3. The Facilities distribution in Singapore

III. Methodology

A. Hedonic Model

Hedonic model is designed to investigate the roles of characteristics of house in determining the housing prices [22]. In this study, we divide these features into two groups, namely structural or non-spatial features, and locational or spatial features. A hedonic model can be defined as formula (1):

$$P = F(x_1, x_2, x_3, \cdots\cdots, x_n) \quad （1）$$

where P represents the housing price, $x_1, x_2, x_3, \cdots\cdots, x_n$ represents the housing characteristics. F represents the fitting function, which can be a linear function or a nonlinear function.

B. GWR model based on taxi travel time

As a local regression model, GWR considers the non-stationarity of various spatial variables. It has been successfully applied in many studies from a wide range of fields, such as epidemiology [23] population, environmental pollution [24], housing price modeling [7]. The conventional GWR model can be expressed by formula (2) [25]:

$$y_i = \sum_{k=1}^{m} \beta_k(u_i v_i) x_{ik} + \beta_0(u_i v_i) + \varepsilon_i \quad (2)$$

Where y_i represents the predicted house price of the location i , $x_{ik} (k = 1,2 \dots m)$ represents the value of independent variables at the location i, $(u_i v_i)$ represents the coordinate of location i , $\beta_k(u_i v_i)$ represents the coefficient of these independent variables at location i, $\beta_0(u_i v_i)$ represents the intercept of the regression formula. In addition, ε_i refers to the error term in the model.

Generally, GWR Model relies on Euclidean distance to measure the spatial proximity between different sub regions. Spatial weight matrix is the key component within GWR model that may significantly affect the effectiveness of the model. The default approach is to generate Euclidean distance based spatial weight matrix. With the continuous development, research on GWR model based on non-Euclidean distance spatial weight matrix started to be explored. For instance, Lu et al. [20] utilized Manhattan distance to construct the spatial weight matrix, which promoted the fitting performance of GWR model greatly [26]. Cao et al [7] innovatively proposed and utilized a public transport travel time based GWR model based on millions of smartcard transactions to model the resale prices of HDB flats in Singapore.

In this research, a GWR model based on taxi travel time is developed. Millions of taxi travel records are utilized to extract travel time between different MTZ pairs. In order to analyze the records at MTZ level, the records are mapped to the MTZs according to the coordinates of the origin and destination points. Then the mean travel time is calculated between each MTZ pair to construct the travel time matrix. Nevertheless, some of the MTZ pairs don't have valid taxi travel records, hence, a linear interpolation approach has also been utilized to estimate travel time in these MTZ pairs instead. Since the data is anonymized, the record loses exact location after being mapped to the MTZ, which might bring a huge error to the accuracy of the records with short travel distances, therefore these records where the origin and the destination are within the same MTZ are also excluded to ensure the reliability of the aggregated travel time matrix. Moreover, for short trips via taxis or driving, the boarding and alighting time as well as the waiting time due to the traffic conditions may impose significant uncertainties to the overall travel time, and given that people may not be very sensitive to the taxi travel time or driving time for short trips while buying flats or houses, hence, we have removed the trips that were shorter than a certain distance to avoid the possible uncertainty caused, instead, linear interpolated travel time is utilized.

To optimize the performance of the model, the parameters have also been calibrated. First, to determine the threshold value of duration to be substituted, 5 km is employed as the basis and 0.1 km as the step size to search for the best threshold value. Result indicates that 5.6 km is the best calibrated distance. In addition, the bandwidth of GWR Model is also optimized according to AICc criterion.

IV. Analyses and Results

A. *Analyzed results using the GWR model based on taxi travel time*

In the experiment, these independent variables considered include mean age of HDB flat in each MTZ, Mean Floor level in each MTZ, distance from centroid point to CBD, distance to the nearest MRT station or park. After running the taxi travel time based GWR model, the result is shown in Table 1.

TABLE I. Estimation results based on the Taxi Travel Time Based GWR model

Variable	**Taxi Travel Time Based GWR Result**
	coefficients （mean value）
$Intercept$	0.853201
V_{Age}	-0.24518
V_{Floor}	-0.26759
V_{Park_dis}	-0.02521
V_{CBD_dis}	-0.42351
V_{MRT_dis}	-0.22554
$Bandwidth$	0.7873
$Adjusted\ R^2$	0.889

It is noted that non-spatial or structural attributes affect the resale prices a lot, the mean coefficients of mean age and mean floor are around -0.24 and -0.26, respectively. The spatial or locational attributes also play an important role, especially the distance to the CBD area and the nearest subway station, i.e., MRT station. The mean coefficients are around -0.42 and -0.22, respectively. In addition, the $Adjusted\ R^2$ of the model has reached 0.889, showing that 89% of the unit price change of HDB flat can be explained by this model. Result indicates that the GWR hedonic model based on taxi travel time is capable of effectively help explain and estimate the house prices in the case study.

B. *Comparision between the GWR model based on taxi travel time other other peer models*

To evaluate the performance of this proposed GWR model based on taxi travel time for modeling house prices, a comparison between the GWR model based on taxi travel time and other representative models, such as OLS regression model, the GWR model based on Euclidean distance and the GWR model based on public transport travel time [7], has been conducted according to the same independent variables. The comparison results can been see in Table II.

TABLE II. Performance Comparison of different models

Standardized coefficients	**Model Comparison**			
	Taxi Travel Time Based GWR	***OLS***	***ED Based GWR***	***PT Travel Time Based GWR***
$Bandwidth$	0.7873	Global	0.6119	1.1434
$Adjusted\ R^2$	0.889	0.6950	0.8873	0.9589

All these models perform well in the experiments, the most commonly used OLS model performs the worst compared to all the other models in terms of the Adjusted R^2, but it still has reached 0.695. In comparison, the GWR model using Euclidean distance has reached more than 0.887, while the GWR model using public transport travel time and the GWR model using taxi travel time have reached 0.9589 and 0.889, respectively. It is noted that the travel time based GWR models outperform Euclidean distance based GWR model, and GWR models perform better than the mostly used OLS model. It is not surprising and aligns with our understanding on how the spatial distribution of house prices can be modeled and explained. As a global model, the OLS model can only establish one model for the entire research area, however, as one of its characteristics, the spatial variation in the price of resale housing also has spatial non-stationarity. Hence, the local models, such as GWR models might be able to perform better in general when considering constructing various local models instead. In addition, it also echoes our common sense that the performance of the GWR models based on travel time is more superior than the GWR model based on Euclidean distance given that people may have more direct sensing of travel time while considering how convenient the flat they plan to buy really is, and how much they really would like to pay for it. Apart from the Adjusted R^2, the values of the bandwidth for the travel time based GWR models are also larger than Euclidean distance based GWR model in this case study.

It is also noted that the performance of the GWR model using taxi travel time is not as good as the performance of the GWR model using public transport travel time in this case study, which is a bit different from our hypothesis that the GWR model using taxi travel time might be better than or at least similar as the GWR model using pubic travel time in modeling housing prices in this case study. According to the result we have obtained, we believe that it is attributed to the characteristics of prices of HDB flats in Singapore we are studying. As we have mentioned, the HDB flats in Singapore accounts for the majority of the housing market in Singapore, but as the government subsidized housing, there are some restrictions on the income level of the owners and transactions. And in addition, the public housing in Singapore, namely HDB blocks, are initiated by the government and mostly very close to these public transport facilities, including MRT stations and bus stops. Meanwhile, the private housing in Singapore, such as condominium apartments and landed houses, are purely market driven and occupied by relatively high-income residents. Hence, the resale prices of public housing, i.e., HDB flats, could better be explained by the public transport travel time as what we have found out in the experiments. Furthermore, it can also demonstrate that the government of Singapore did a good job in providing very convenient and good quality public housing to the population residing in HDB blocks in Singapore. It might be highly possible that the GWR model based on taxi travel time might perform better while modeling private housing prices in Singapore.

V. Conclusion and discussion

In this research, a GWR model using taxi travel time has been proposed and utilized to model the HDB flats resale prices in the Singapore. In addition, a comparison between the proposed GWR model using taxi travel time and other models, i.e., OLS model, GWR model using Euclidean distance and GWR model using public transport travel time, have also been conducted. The experiments and comparison results demonstrate that the GWR model using taxi travel time performs superior than the OLS model, and slightly better than the GWR model using Euclidean distance as well. Nevertheless, it is worse than GWR model using the public transport travel time in this case study according to the metrics, i.e., Adjusted R^2 and Bandwidth. These experiments indicate that the public transport travel time may be more effective in the modeling of the public housing resale prices compared to taxi travel time or driving time, and both the taxi travel time and public transport travel time can better explain the public housing resale prices in the research area compared to Euclidean distance in the GWR modeling.

This research has demonstrated the capability of our proposed travel time based GWR model in terms of modeling the public housing resale prices in Singapore. Moreover, this research has also explained and discussed why the proposed model might not be as good as the public transport travel time based GWR model according to their performance comparison in the case study. We also hope this research could further promote the efforts and discussion from more scholars on exploring how big data can help contribute to the improvement of different spatial models, apart from help more comprehensively and in-depth reveal different spatial or geographic phenomena from different facets. On the other hand, there is still room for improvement in this research. For example, the independent variables considered in the modeling could be more comprehensive if collected datasets allow; the taxi travel time is aggregated information based on MTZs that might bring some uncertainty as well to the experiments results, smaller scale would be able to bring forward more details and findings in this research. Of course, we would definitely like to improve our research in the future on these aspects. In addition, given that the taxi travel time based GWR model is outperformed by the public transport travel time based GWR model in the modeling of HDB flat resale prices in Singapore, we believe that the situation might be different while modeling private housing prices in Singapore, which might be more significantly affected by the convenience brought by the shorter taxi travel time or driving time. It will also be another direction of our research in the future.

Acknowledgment

We are very grateful to all the reviewers of this paper for their time and efforts on this paper. This research is supported by the research grants of 13901-41211D-22002 and 13901-120215-11131.

References

1. Lake, I.R.; Lovett, A.A.; Bateman, I.J.; Day, B. Using GIS and large-scale digital data to implement hedonic pricing studies. *International Journal of Geographical Information Science* **2000**, *14*, 521-541, doi:10.1080/136588100415729.
2. Yu, S.M.; Han, S.S.; Chai, C.H.J.E.; Planning, P.B.; Design. Modeling the value of view in high-rise apartments: A 3D GIS approach. **2007**, *34*, 139-153.
3. Kuburić-Subotica, M.; list, G.Ć.-B.J.g. The Application of Intelligent Techniques for Massreal Estate Appraisal. **2012**.
4. Brankovic, S.J.G.L. Real Estate Mass Appraisal in the Real Estate Cadastre and GIS Environment. **2013**, *67*, 119-134.
5. Calven, V.D.B. A statistical model for valuation of residential property in the Nelson Mandela Metropolitan area. **2012**.
6. Wang, S.; Wang, Y.; Lin, X.; Zhang, H.o. Spatial differentiation patterns and influencing mechanism of housing prices in China: Based on data of 2872 counties. *Acta Geographica Sinica* **2016**, *71*, 1329-1342.
7. Cao, K.; Diao, M.; Wu, B. A Big Data-Based Geographically Weighted Regression Model for Public Housing Prices: A Case Study in Singapore.

Ann Am Assoc Geogr **2019**, *109*, 173-186, doi:10.1080/24694452.2018.1470925.

8. Lu, Z.Y.; Im, J.; Quackenbush, L.J.; Yoo, S. Remote Sensing-based House Value Estimation Using an Optimized Regional Regression Model. *Photogrammetric Engineering and Remote Sensing* **2013**, *79*, 809-820, doi:10.14358/pers.79.9.809.
9. Yao, Y.; Zhang, J.B.; Hong, Y.; Liang, H.L.; He, J.L. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *T Gis* **2018**, *22*, 561-581, doi:10.1111/tgis.12330.
10. Jayantha, W.M.; Oladinrin, O.T. Bibliometric analysis of hedonic price model using CiteSpace. *Int J Hous Mark Anal* **2019**, *13*, 357-371, doi:10.1108/ijhma-04-2019-0044.
11. Phang, S.Y.; Wong, W.K. Government policies and private housing prices in Singapore. *Urban Stud* **1997**, *34*, 1819-1829, doi:10.1080/0042098975268.
12. Sing, T.F.; Tsai, I.C.; Chen, M.C. Price dynamics in public and private housing markets in Singapore. *Journal of Housing Economics* **2006**, *15*, 305-320, doi:10.1016/j.jhe.2006.09.006.
13. Wang, L.P.; Chan, F.F.; Wang, Y.L.; Chang, Q.; Ieee. Predicting Public Housing Prices Using Delayed Neural Networks. In Proceedings of the IEEE Region 10 Conference (TENCON), Singapore, Nov 22-25, 2016; pp. 3589-3592.
14. Yang, L.C.; Zhou, J.P.; Shyr, O.F.; Huo, D. Does bus accessibility affect property prices? *Cities* **2019**, *84*, 56-65, doi:10.1016/j.cities.2018.07.005.
15. Ibeas, A.; Cordera, R.; dell'Olio, L.; Coppola, P.; Dominguez, A. Modelling transport and real-estate values interactions in urban systems. *J Transp Geogr* **2012**, *24*, 370-382, doi:10.1016/j.jtrangeo.2012.04.012.
16. Xiao, Y.; Hui, E.C.M.; Wen, H.Z. Effects of floor level and landscape proximity on housing price: A hedonic analysis in Hangzhou, China. *Habitat Int* **2019**, *87*, 11-26, doi:10.1016/j.habitatint.2019.03.008.
17. He, C.; Wang, Z.; Guo, H.; Sheng, H.; Zhou, R.; Yang, Y. Driving Forces Analysis for Residential Housing Price in Beijing. In *International Conference on Ecological Informatics and Ecosystem Conservation*, Yang, Z., Chen, B., Eds.; Procedia Environmental Sciences; 2010; Volume 2, pp. 925-936.
18. Clapp, J.M.; Giaccotto, C. Residential hedonic models: A rational expectations approach to age effects. *Journal of Urban Economics* **1998**, *44*, 415-437, doi:10.1006/juec.1997.2076.
19. Bidanset, P.E.; Lombard, J.R.; Davis, P.; Mccord, M.; Mccluskey, W.J.J.S.I.P. Further Evaluating the Impact of Kernel and Bandwidth Specifications of Geographically Weighted Regression on the Equity and Uniformity of Mass Appraisal Models. **2017**.
20. Lu, B.; Charlton, M.; Harris, P.; Fotheringham, A.S. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science* **2014**, *28*, 660-681, doi:10.1080/13658816.2013.865739.
21. Eboy, O.V.; Samat, N. Modeling property rating valuation using Geographical Weighted Regression (GWR) and Spatial Regression Model (SRM): the case of Kota Kinabalu, Sabah. **2015**.
22. Aladwan, Z.; Ahamad, M.S.S. HEDONIC PRICING MODEL FOR REAL PROPERTY VALUATION VIA GIS - A REVIEW. *Civ Environ Eng Rep* **2019**, *29*, 34-47, doi:10.2478/ceer-2019-0022.
23. Liu, Y.; Jiang, S.; Liu, Y.; Wang, R.; Li, X.; Yuan, Z.; Wang, L.; Xue, F. Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis. *Int J Health Geogr* **2011**, *10*, doi:10.1186/1476-072x-10-50.
24. Zou, B.; Fang, X.; Feng, H.; Zhou, X. Simplicity versus accuracy for estimation of the PM2.5 concentration: a comparison between LUR and GWR methods across time scales. *J Spat Sci* **2021**, *66*, 279-297, doi:10.1080/14498596.2019.1624203.
25. Lu, B.; Ge, Y.; Qin, K.; Zheng, J. A Review on Geographically Weighted Regression. *Geomatics and Information Science of Wuhan University* **2020**, *45*, 1356-1366.
26. Lu, B.B.; Charlton, M.; Harris, P.; Ieee. Geographically Weighted Regression Using a Non-Euclidean Distance Metric with Simulation Data. In Proceedings of the 1st International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Shanghai, PEOPLES R CHINA, Aug 02-04, 2012; pp. 267-270.