

An Improved Geographically and Temporally Weighted Regression Model with a Novel Weight Matrix

Renrong Jiang

Shenzhen Planning and Land Development Research Center, Shenzhen

Hongxia Wang

Southeast University, Nanjing, China

Bo Huang

The Chinese University of Hong Kong, Hong Kong, China

Gao Guo

Xi'an University of Technology, Xi'an, China

Abstract

Geographically and temporally weighted regression (GTWR) has been developed to model both spatial and temporal non-stationarity in real estate market data. GTWR integrates both spatial and temporal information in the weight matrix to capture spatial and temporal heterogeneity, while the factor effects of the neighboring housing units (or zones) are totally ignored. On the other hand, a local linear fitting method (LLFM) that accounts for spatio-temporal heterogeneity in a regression context only considers distances in the factors space, ignoring the space-time locations of the neighbors and the relative space-time distance between neighbors. In this paper, we proposed a new weight function that combines the space-time distance and the distance in the factors space. A case study in Shenzhen showed that the proposed model with the new weight function performed better than the traditional GTWR or LLFM.

1. Introduction

Real estate prices are influenced by many factors, e.g., house age, size, and exterior and interior features. These factors interweave with location and time, making the prices more complex to model.

In general, there are two types of regression models to account for the effects of location and time, or spatial and temporal autocorrelations, on house prices. One is global models and the other is local models. Local models allow the consideration of varying influences of factors over space and/or time, i.e. spatial and/or temporal

non-stationarity. An example local model is the geographically and temporally weighted regression (GTWR) model (Huang *et al.*, 2010). In GTWR, any spatio-temporal non-stationarity in the relationship of interest is considered through a local estimation of model coefficients via a weight matrix. The weight matrix represents the different importance of each individual observation in the data set used to estimate the parameters at the focal location. In general, the observation data points ‘close’ to the focal point in the space-time coordinate system have a greater influence than the data located farther from it in the space-time coordinate system. However, GTWR just uses the distance of space-time, the mutual effects of the factors of the focal point and its neighbors are totally ignored. Thus, GTWR model may not be effective and plausible in some situations.

Another approach, the local linear fitting method (LLFM), has also been developed to deal with spatio-temporal data (Wang and Wang, 2009). Local polynomial fitting and particularly its special case - local linear fitting has become increasingly popular (see, e.g., Fan (1992) and Hallin *et al.* (2004)). The weight matrix in LLFM represents the different importance of each individual observation in the data set used to estimate the value of regression function. In general, for a given factor at the focal location, if the factor of its neighbor is similar to it, i.e., the distance between this factor and its neighbor’s factor is small; the neighbor is assigned a large weight by the weight function. In contrast, a neighbor’s factor dissimilar to the factor at the focal point is assigned a small weight, since it is far away from it in the distance of the factor space. Clearly, LLFM only considers the factor space distance in the weight matrix, and it does not consider space-time locations of the neighbors and the relative space-time distance between the focal point and its neighbors, which may also lower down the modeling accuracy.

From the above discussion, we can find that GTWR uses only the spatiotemporal distance but not the factor distance in its weight matrix. In contrast, LLFM utilizes only the factor distance but not the spatiotemporal distance in its weight matrix. Hence, in this paper, we propose an improved GTWR (IGTWR) equipped with a new weight function that combines the space-time distance with the factor distance to improve the performance of GTWR.

2. The new weight function

In view of the pros and cons of LLFM and GTWR, we propose an approach that combines these two methods. That is, similar to the product kernel method, we come

up with a new weight function which integrates the space-time distance with the factor distance:

$$\begin{aligned}
W(s_0, t_0) &= \text{diag}\{K_{10}, \dots, K_{n0}\} \\
K_{i0} &= K\left(\frac{s_i - s_0}{h_S}, \frac{t_i - t_0}{h_T}, \frac{X(s_i, t_i) - X(s_0, t_0)}{\mathbf{h}}\right) \\
&= K_1\left(\frac{s_i - s_0}{h_S}, \frac{t_i - t_0}{h_T}\right) \cdot K_2\left(\frac{X(s_i, t_i) - X(s_0, t_0)}{\mathbf{h}}\right) \quad (1)
\end{aligned}$$

where K_1 is related to the space-time distance and K_2 the factor distance.

Here we use multiplication with the following reasons:

1. When $s_i = s_0$ and $t_i = t_0$, $K_{i0} = K_2\left(\frac{X(s_i, t_i) - X(s_0, t_0)}{\mathbf{h}}\right)$ (Since $K_1(\mathbf{0}, 0) = 1$).
2. When $X(s_i, t_i) = X(s_0, t_0)$, $K_{i0} = K_1\left(\frac{s_i - s_0}{h_S}, \frac{t_i - t_0}{h_T}\right)$ (Since $K_2(\mathbf{0}) = 1$).

Remark: When $s_i = s_0$ and $t_i = t_0$, we just need to consider the difference of factors, thus at this point $K_{i0} = K_2\left(\frac{X(s_i, t_i) - X(s_0, t_0)}{\mathbf{h}}\right)$ is reasonable. Similarly, when $X(s_i, t_i) = X(s_0, t_0)$, $K_{i0} = K_1\left(\frac{s_i - s_0}{h_S}, \frac{t_i - t_0}{h_T}\right)$ is also reasonable (neither addition nor subtraction can satisfy these properties). In addition, K_{i0} is a decreasing function of the space-time distance and the factor distance.

3. Case study

To examine the applicability of our method, a case study was performed using the housing prices observed between 2001 and 2005 in the city of Shenzhen, China.

598 observations were available from the study area, which provided full information on the age, number of rooms, land price, density, and other variables. A recent selling price was taken as the dependent variable, standing as a proxy for the market value of the house. The explanatory variables comprised three groups which included a total of 10 variables: bus lines available (within 100 m) (X_1), number of bedrooms (X_2), building density (X_3), floor area (X_4), urban or suburban (X_5), distance to major road (X_6), distance to school (X_7), distance to central business district (CBD) (X_8), property management fee (X_9), and floor area ratio (FAR) (X_{10}). The house price of each zone was transformed to the unit house price per square meter, according to the total house price in the zone divided by the total floor area.

We applied the three models: LLM, GTWR, and IGTWR to the house price data set of Shenzhen. The results are shown in Tables 1-3. Because the output of local parameter estimates from the three models is voluminous, Tables 1, 2 and 3 only provide a five column summary of the distribution of each parameter to indicate the extent of its variability; the five parameters are minimum (Min), low quartile (LQ), median (Med), upper quartile (UQ), and maximum (Max) respectively.

Table 1. Parameter estimate summary using LLM

Parameter	Min	LQ	Med	UQ	Max
Intercept	0.1283	0.3174	0.4359	0.5064	0.7390
Bus lines available (X_1)	-0.1658	-0.0741	-0.0522	-0.0511	-0.0483
Number of bedrooms (X_2)	-0.0884	-0.0142	0.0991	0.1085	0.1909
Building density (X_3)	-0.1650	-0.0478	-0.0361	0.0392	0.1391
Floor area (X_4)	-0.0272	0.3482	0.4509	0.4602	0.5017
Urban or suburban (X_5)	0.0000	0.0000	0.0000	0.0000	0.0000
Dist. to major road (X_6)	0.0077	0.1492	0.3047	0.3135	0.3683
Dist. to school (X_7)	-0.1420	0.0269	0.1177	0.1274	0.1406
Dist. to CBD (X_8)	-0.0832	-0.0048	0.1361	0.1454	0.2695
Property management fee (X_9)	-0.0973	-0.0046	0.0082	0.0191	0.2884
FAR (X_{10})	-0.1750	0.0694	0.1106	0.1238	0.1471
Diagnostic information					
R^2				0.7152	
Residual standard error				0.2404	
Residual sum of squares				33.7598	
AIC				-1683.2	

Table 2. Parameter estimate summary using GTWR

Parameter	Min	LQ	Med	UQ	Max
Intercept	-0.0144	0.0127	0.0222	0.0484	0.1541
Bus lines available (X_1)	-0.1682	-0.0552	-0.0493	-0.0475	-0.0406
Number of bedrooms (X_2)	-0.0496	0.0150	0.0309	0.0445	0.1620
Building density (X_3)	-0.0854	0.0051	0.0164	0.0253	0.0627
Floor area (X_4)	0.2101	0.3629	0.4158	0.5003	0.6297
Urban or suburban (X_5)	0.0895	0.0948	0.1335	0.1424	0.2483
Dist. to major road (X_6)	-0.2033	0.1285	0.2284	0.2490	0.3904

Dist. to school (X_7)	-0.3352	0.0484	0.0738	0.0957	0.1526
Dist. to CBD(X_8)	-0.0704	-0.0308	-0.0154	-0.0006	0.4071
Property management fee (X_9)	-0.0235	0.0120	0.0528	0.0646	0.6787
FAR (X_{10})	-0.1026	0.0560	0.0933	0.1200	0.1719
Diagnostic information					
R^2				0.7561	
Residual standard error				0.2225	
Residual sum of squares				28.9202	
AIC				-1775.3	

Table 3. Parameter estimate summary using IGTWR

Parameter	Min	LQ	Med	UQ	Max
Intercept	0.0503	0.3245	0.4338	0.5042	0.9781
Bus lines available (X_1)	-0.2805	-0.0640	-0.0515	-0.0435	-0.0229
Number of bedrooms (X_2)	-0.3833	-0.0641	0.0503	0.1126	0.6358
Building density (X_3)	-0.6034	-0.0431	0.0048	0.0701	0.3207
Floor area (X_4)	-3.0121	0.2746	0.4116	0.4905	0.6739
Urban or suburban (X_5)	0.0000	0.0000	0.0000	0.0000	0.0000
Dist. to major road (X_6)	-0.5499	0.0772	0.1864	0.2949	0.9745
Dist. to school (X_7)	-0.4583	0.0257	0.0692	0.1318	0.2355
Dist. to CBD(X_8)	-2.0345	-0.0431	0.0089	0.1254	0.9100
Property management fee (X_9)	-0.23	0.0116	0.0424	0.1184	1.0744
FAR (X_{10})	-0.6584	0.0422	0.1001	0.1493	0.8657
Diagnostic information					
R^2				0.8148	
Residual standard error				0.1939	
Residual sum of squares				21.9612	
AIC				-1939.1	

Tables 1-3 show that the R^2 values of LLFM, GTWR, and IGTWR are 0.7152, 0.7561 and 0.8148, respectively. This indicates that IGTWR has achieved the best goodness-of-fit among the three models. In fact, for the other three model diagnostic indicators: Residual standard error, Residual sum of squares, and AIC, IGTWR has also obtained the best performance among the three models. We posit that this is because IGTWR not only considers the spatial and time effects as GTWR, but also the factor similarity effects as LLFM does in its weight function.

4. Conclusion

To overcome the drawbacks of GTWR and LLFM, this paper proposed a new weight function that combines the space-time distance and the factor distance. Intuitively, the closer of two observation points in space and time, the more similar the prices are; on the other hand, if two observations have similar factor values, the prices are also similar. The case study showed that our method performed better than GTWR and LLFM in terms of R^2 and other indicators. Moreover, we can prove that if the weight functions of LLFM and GTWR are the same, then the estimators of them are also identical.

References:

- Fan, J., 1992. Design-Adaptive Nonparametric Regression. *Journal of the American Statistical Association*, 87, 998-1004.
- Hallin, M., Lu, Z., and Tran, L.T., 2004. Local Linear Spatial Regression. *The Annals of Statistics*, 32, 2469-2500.
- 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.
- Wang, H., and Wang, J., 2009. Estimation of the Trend Function for Spatio-Temporal Models. *Journal of Nonparametric Statistics*, 21, 567-588.