Separating causes of spatial autocorrelation: An empirical analysis of the price formation process

Nikolaos Andreas Soultanidis[a], Kwong Wing Chau[b], Siu Kei Wong[c]

Abstract

Researchers deal with spatial dependence in urban transaction data due to econometrical reasons – to maintain model reliability. Few research is focused on the economic causes of the phenomenon. Researchers are unable to distinguish between econometrical causes (omitted variables) and economic ones (e.g. price formation). This work separates the price formation process from other causes by reweighting the spatial weight matrix according to the similarity of characteristics between observational pairs. This unique pattern captures only the spatial dependence effect that pertains to the price formation process. Based on more than 33,000 transactions in the heterogeneous residential area of Hong Kong, we find that spatial dependence structure is to a large degree linked to similarities in hedonic characteristics. This confirms the price formation – spatial dependence conjecture.

***Keywords*:** spatial autocorrelation; price formation; real estate; similarity

[a] Corresponding author. Department of Real Estate and Construction, University of Hong Kong, Pokfulam Road, Hong Kong; Tel: (852) 6920 6360; Email: nsoul@hku.hk
[b] Department of Real Estate and Construction, University of Hong Kong, Pokfulam Road, Hong Kong; Tel: (852) 2859 2128; Email: hrrbckw@hku.hk
[c] Department of Real Estate and Construction, University of Hong Kong, Pokfulam Road, Hong Kong; Tel: (852) 2859 1193; Email: skwongb@hku.hk

Introduction

Little research has been focused on the question why spatial dependence occurs in real estate models. As the focus of spatial economics has been on improving models, few have attempted to quantify the economic causes. Particularly, research has not been able to disentangle the two presumably most common reasons for this phenomenon: omitted spatial variables and the price formation process.
If researchers measure spatial dependence structures, what are the economic conclusions they can draw? This research tries to extract the price formation process from the bundle of possible causes for spatial dependence.

Spatial dependence data has been shown to have two major causes – firstly, it may stem from omitted variables that exhibit spatial structure. Significant spatial dependence may be an indication that the model is not specified well. The second reason that has been introduced by a number of researchers (Can & Megbolugbe, 1997; Wong, Yiu, & Chau, 2013) is the price formation process. Market participants have different sources of data - they may get their information from market wide aggregate information and local disaggregate transaction data. A heavy reliance on past local transaction data may be the other source of spatial dependence.

This paper attempts to disentangle above mentioned causes for spatial dependence in urban transaction data. Aware that the analysis can only be conducted by examining the price formation process rather than the unknown omitted variables, this paper uses the changes in the reliance on past local transaction data in order to pinpoint the amount of spatial dependence that is due to this process alone. We anticipate that the reliance on past local transaction data is likely to change across the data set. In other words, the valuation environment buyers and sellers face changes constantly; both the amount and quality of local price information vary across observations. This paper extends the literature on the quantitative determination of the price formation share of Wong et al. (2013), who focused mainly on the *amount* of local information (trading volume). Here, we examine spatial dependence in the context of different *quality* of local transaction information available. Borrowing from the concept of atypicality as firstly introduced by Haurin (1988), we introduce the measure of similarity to judge the qualitative aspects of the valuation environment for each possible local comparable.

As a result, this study clearly delineates between the share of spatial dependence that is caused by the price formation process, and other common causes.

This paper proceeds with a general introduction on the literature on spatial autocorrelation as it pertains to real estate prices. Further, the literature on atypicality is reviewed. We discuss how it may be adapted to describe the qualitative aspects of the price formation process thereby introducing the similarity measure. We proceed to present the models that are used for the empirical tests. Data and variables introduce the scope and depth of our data set; while results discuss the implications of our tests.

Literature Review

Real estate data are characterized by two spatial features, spatial heterogeneity and spatial dependence. In most cases spatial heterogeneity is thought to be the stronger phenomenon (Can, 1992). It describes a systematic variation of a spatial process in space.

The focus of this study is spatial dependence. The intuition for spatial dependence is best described by Tobler's (1979) first law of geography – 'Everything is related to everything else, but near things are more related than distant things'. In essence, positive spatial autocorrelation is detected if the analysis of elements in space yields that elements that are located close to each other are more likely to be similar than distant ones.

At first glance, the price formation process in the direct real estate markets seems to be a likely candidate to explain spatial dependence processes. However, research has shown that there are a multitude of other potential causes for spatial autocorrelation. The spatial autocorrelation that we measure may also appear due to model misspecifications, smaller scale spatial heterogeneity effects and other externalities.

The interest in spatial autocorrelation in the realm of urban economics is due to the fact that models that exhibit spatially autocorrelated residuals do not meet important model assumptions, the sampling variability around coefficient estimates are unreliable.

Together with the observation that many causes of spatial autocorrelation have a sampling/econometrical root, it does not surprise that most research is focused on devising

models to clear the phenomenon; rather than gain an understanding in the economic processes that may cause a part of it.

One early exception Can & Megbolugbe (1997) who in their work to incorporate the spatial dependence structure into house price indices, hint on the the importance of information search as one possible explanation.

One more recent paper by Wong et al. (2013) is the first to empirically test the impact of search on spatial dependence. They use the idea that past local trading volume changes the amount of information that was available at valuation. If the surrounding local transactions are considered a noisy signals of the true price (Quan & Quigley, 1989), a small sample size will decrease the confidence in that source of information and lead to a smaller spatial dependence. Their tests confirm that spatial dependence indeed varies according to the trading volume which strongly hints at the information search conjecture.

Despite the significant results of Wong et al. (2013), trading volume can be expected to proxy only parts of the spatial autocorrelation that is induced by the price formation process. This notion is further supported by the relatively small effect size on spatial dependence parameters that were scaled by the trading volume. In order to supply additional coverage of the price formation effects, it is worthwhile to introduce further dimensions.

During the price formation process, a multitude of factors may affect market participants' confidence for past local transactions. In addition to the quantity of available information in the vicinity, the quality of comparables is expected to change the degree of confidence and thus the spatial dependence. The quality of comparables will be a function of distance in (1) space, (2) hedonic characteristics and (3) time. If the distance in these three dimensions is large for a neighborhood relationship [i,j], where i denotes the observation that is to be valued by market participants - and j another past transaction in the data set, the transaction pair would be considered unsuitable to function as comparables. This notion is generally accepted in the measurement of spatial dependence as the structure of the dependence is frequently defined by an inverse spatial distance function. This study takes advantage of the fact that only price formation as a reason for spatial dependence would uniquely vary with the similarity in

characteristics between neighbors [i,j]. Thus this measure is able to delimit it from the omitted variable cause[1].

We introduce later the concept of similarity by modifying the related concept of atypicality, as firstly introduced by Haurin (1988). Atypicality describes the degree to which a piece of property tends to be more or less standard in comparison with others in the sample. The measure of atypicality is commonly used in context of research on Time-on- market (TOM). In this area, it is a common control factor (Allen, Faircloth, & Rutherford, 2005; Allen, Rutherford, & Thomson, 2009; Benefield, Cain, & Johnson, 2011; Haurin, 1988; Jud & Frew, 1990).

Haurin (1988) reasoned that an atypical property has a wider offer price distribution - buyer's do not value the property similarly - so that a seller is incentivized to wait longer under optimal stopping rules until she commits to a sale.

Similarity

Atypicality (Haurin, 1988) and similarity, as defined here, are related concepts. They differ in the importance of the reference [j] to which the characteristic distance of object [i] is judged against.

For this study each neighborhood relationship [i,j] carries valuable information, which leads us to two important departures from the concept of atypicality:

Firstly, similarity is described in form of a matrix, in which hedonic feature differences between [i,j] are aggregated by applying hedonic implicit prices[2].

Secondly, following Turnbull, Dombrow, & Sirmans (2006), the measure is only applied to a local context for observations that are within the spatial and temporal proximity [J]. In order to follow the conventions of spatial statistics, the inverse of each element is taken.

$$Sim_{ij} = \left(\sum_{h=1}^{m} \left| p_h \left(h_i - h_{j \epsilon J} \right) \right| \right)^{-1} \quad (1)$$

[1] The odds that an omitted variable could still be correlated with the atypicality factor is decreased dramatically.

[2] Which are derived from a first stage regression model.

Where h_i is the characteristic feature value for observation i; $h_{j \in J}$ the characteristic feature value for a neighboring observation; p_h is the implicit hedonic price from a first stage regression model which has m hedonic characteristics.

On an intuitive note, the measure expresses the similarity for each observation pair [i,j] in dollar terms. If, for example, an observation pair [i,j] is located in the same building that has for all apartments the same size, two of the three hedonic characteristics will be the same for our sample (building age, and saleable floor area). Conversely, they will likely differ in the floor level. In this case, the similarity measure will be the inverse of the coefficient of a first stage regression model for floor level multiplied by the difference in floor level between i and j. This renders the similarity between an observation pair in dollar terms, and thus facilitates a comparison across pairs.

The similarity values need to be distinguished from merely calculating the difference of projected values by the first stage regression. Although the building age[3] and floor size price effects may offset each other, the similarity value is not offset. A transaction pair [i,j] can have the same transaction price, but still very different hedonic characteristic that would render them unsuitable as comparables.

Using the implicit prices of a first stage regression to compute the similarity matrix in dollar terms has two advantages. Firstly, the scale of the different hedonic variables is addressed, so that changing between meters and square foot in the variable "saleable floor size" will not have any effect – as long as it is used consistently. Secondly, the importance between the hedonic variables at the computation of the similarity is approximated best by the first stage model.

The importance of similarity for the price formation process stems from the assumption that buyers/sellers are unable to compare prices of properties that are very dissimilar to each other. We assume that market participants in private residential transactions do not control for heterogeneity with a hedonic model, and that they are unable to obtain implicit prices. As such they are dependent on finding comparables that are closely related to the property that they are trying to value. This assumption seems to be in line with work of Case & Shiller (2003) who point out the naiveté of residential real estate actors.

[3] Generally negative coefficient in hedonic models.

Model Development

In a multi-period hedonic price model for real estate where P_{it} is the sale price of property i at time t; X_i is a vector of property i's characteristics; β is the implicit real price of the characteristics; τ_t is the market-wide price level at time t; and ε_{it} is an unobserved random element in each transaction with zero mean.

$$P_{it} = X_i\beta + \tau_t + \varepsilon_{it} \quad (2)$$

This model assumes that a property can be valued solely by its own characteristics and the prevailing implicit nominal prices.

Due to the nature of property markets, where trades happen infrequently in a decentralized manner and goods are heterogeneous, market participants look for local trading information in the recent past during the valuation process. This idea motivates the application of the spatial autoregressive model, that incorporates a spatial lag into the model. The spatial lag WP is a nx1 vector that resulted from multiplying a nxn inverse distance spatial weight matrix with a nx1 price vector. It captures the "indirect effects" stemming from nearby housing prices; ρ is a spatial autocorrelation parameter to be estimated, which indicates the degree to which prices of the subject houses can be explained by the average of house prices in their immediate vicinity.

$$P_{it} = W_{ij}^{dist} P_{j,t-k}\rho + X_i\beta + \tau_t + \varepsilon_{it} \quad (3)$$

A positive significant ρ is a necessary but insufficient condition to conclude that the data set shows dependence structure induced by the price formation process. As a significant ρ may also stem from a model misspecification by the researcher. Omitted variables are frequently detected by examining the nature of the spatial dependence by applying local statistics or visualizing it on the map.

Thus, in order to clearly distinguish between these two sources, we apply parallel to the case of Wong et al. (2013), a factor that alters the original weight matrix specification according to the similarity for each neighborhood relationship [i,j].

$$P_t = W^{dist}\mathrm{P}_{t-k}\rho + W^{sim}\circ W^{dist}\mathrm{P}_{t-k}\gamma + \mathrm{X}\beta + \tau_t + \varepsilon_t \quad (4)$$

Where P_t is an nx1 vector of sales prices. W^{sim} denotes a row-standardized weight of similarity relationships that forms together with W^{dist} a weight matrix that uniquely captures the price formation process with both inverse distance and inverse atypicality weights[4]. The parameter γ captures the part of the spatial dependence that can be attributed to the price formation process. Whereas ρ captures other combined spatial dependence effects.

The intuition behind this is as follows: If an important spatial variable such as nearby amenity in form of a park is excluded in the model, a spatial dependence structure is likely detected. In this case, the inverse distance spatial lag will be more capable to capture the variability in comparison to the newly introduced combined inverse distance/similarity structure. In fact, the new combined term will weaken the weight for certain observation pairs that are close in space, but dissimilar in their hedonic characteristics.

Although the capture of an omitted variable can never be fully excluded, by providing a rather complex algorithm, we can exclude the possibility that the omitted variable has only a spatial component. Such an omitted variable would need to possess also a component that correlates with the similarity of hedonic information available. This renders any omitted variable falsely captured by the combined inverse distance/similarity lag likely closely linked to the price formation process.

[4] Both weight matrices are combined by the Hadamard product (element wise matrix multiplication) as noted by this symbol °.

Data and Variables

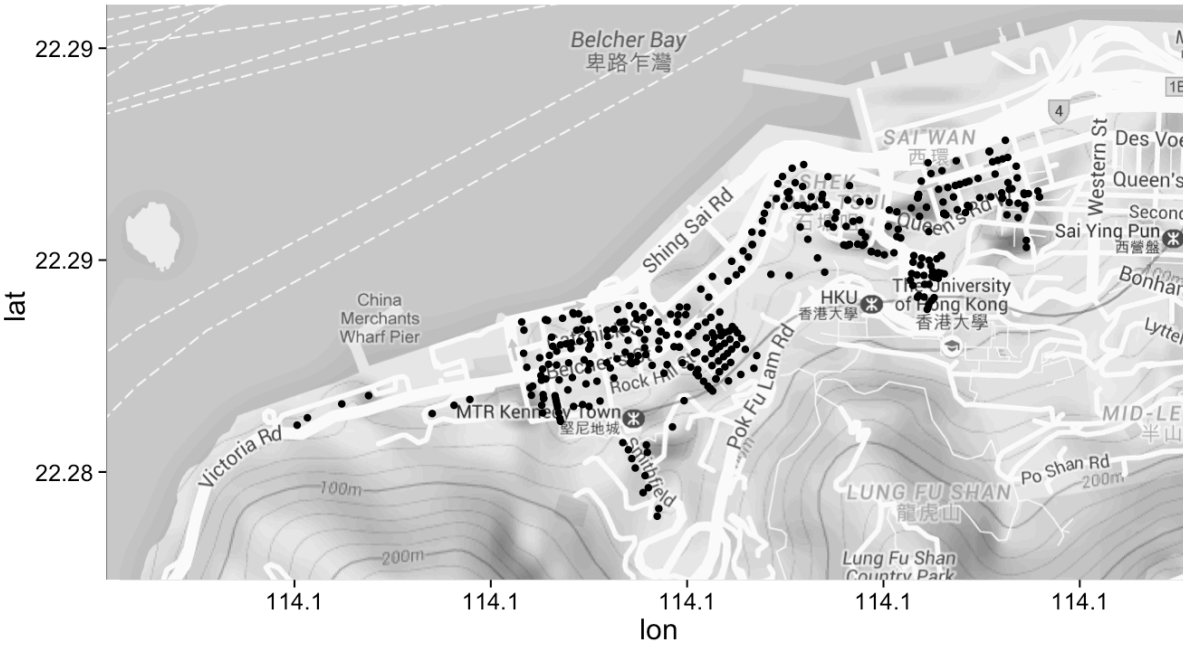
The above model was tested for a transaction data set of Kennedy Town (Hong Kong) between the years 1992 - 2016. The data set encompasses 33.762 transaction points. Kennedy Town is a densely populated part of Hong Kong that features an adequate amount of heterogeneity for the purpose of this study.

As the real estate markets in Hong Kong are liquid, the weight matrices were defined with smaller lag structures. A transaction pair [i,j] was considered neighboring if j was within a 90 day window in the past of the transaction date of i. Additionally, transactions that surpassed 200m were not considered neighboring. These strong limits rendered a median number of 61 neighbors with an interquartile range of 38-99 neighbors.

The weights of the standard spatial lag were determined by inverse distance, which was then row-standardized in order to reign in on the influence of each observation i on the spatial dependence effect.

The weights of the newly introduced inverse distance/similarity matrix were computed as follows: Firstly, the similarity matrix was constructed as noted in Eq.1, then the matrix was row-standardized before combined by element-wise matrix multiplication with the inverse distance matrix.

Figure 1 Spatial distribution of buildings in Kennedy Town, Hong Kong



Further variables, as introduced in Table 1 below, are the price in HK$M; the floor level, building age and saleable floor area. The fixed time effects are captured with a series of monthly time dummies.

Table 1 Descriptive statistics of our data

	Unit	Mean	St. Dev.	Min	Max
Price (P)	HK$M	3.29	3.28	0.32	46.26
Spatial lag ($W^{dist}P$)	HK$M	3.19	2.26	0	22.14
Scaled spatial lag (W^{dist}°$W^{sim}P$)	HK$M	3.19	2.34	0	30.28
Floor level (flr)	storey	18.19	13.04	1	61
Building age (age)	year	26.43	11.69	0.17	58.04
Saleable floor area (sfa)	ft2	475.8	212.1	128	2,032
					N = 33,762

Table 2 Regression results

	Dependent variable:		
	Prices Real Estate in HK Mio.		
	Simple	Spatial lag	Scaled spatial lag
	(1)	(2)	(3)
Spatial lag - distance		0.307***	0.035***
		(0.006)	(0.013)
Spatial lag - distance/similarity			0.277***
			(0.012)
Floor level	0.026***	0.023***	0.020***
	(0.001)	(0.001)	(0.001)
Building Age	-0.058***	-0.049***	-0.049***
	(0.001)	(0.001)	(0.001)
SFA	0.010***	0.009***	0.009***
	(0.00004)	(0.00005)	(0.00005)
Month transaction (exemplary for 280 other omitted levels)	0.044	0.019	0.017
	(0.175)	(0.167)	(0.166)
Constant	-1.142***	-1.283***	-1.243***
	(0.133)	(0.128)	(0.127)
Observations	33,762	33,762	33,762
R^2	0.827	0.841	0.844
Adjusted R^2	0.825	0.840	0.842
Residual Std. Error	1.369 (df = 33482)	1.311 (df = 33481)	1.301 (df = 33480)
F Statistic	573.300*** (df = 279; 33482)	634.100*** (df = 280; 33481)	643.300*** (df = 281; 33480)
Note:			*p<0.1; **p<0.05;** p<0.01

Results

Table 2 presents the estimated coefficient of each variable in Eqs. (2), (3), and (4), with its corresponding p-value in brackets. The fixed time effects are only hinted at to simplify presentation. The bottom of the table shows the R- squared and adjusted R-squared values of each model. The three models improved progressively in terms of adjusted R-squared and residual standard error.

Equation (2) was the traditional hedonic model. All coefficients were significant at the 1 % level with the expected signs. The adjusted R- squared value was 82.50 %, which was reasonable given the high heterogeneity of the observations in Kennedy Town.

Equation (3) was the hedonic model with one spatial autoregressive process and an autocorrelation structure imposed as inverse distance. We found that spatial effects were rather large and significant at the 1 % level: the inverse distance spatial lag had a coefficient of 0.307.

Equation (4) introduced one additional variable, the scaled inverse distance/similarity spatial lag, which captured a significant coefficient of 0.277. Whereas the general inverse distance spatial lag coefficient drops by a factor of 10 to 0.035. This indicates that the more complex spatial lag that traces not only distance, but also the similarity of neighbors - and thus their suitability as comparables - performs better.
Thus, the price formation process has a rather large share in explaining the spatial dependence structure of real estate prices. For this sample, omitted variables do not seem to play a large role.

Conclusion

The contributions of our study are as follows. Firstly, the study provides strong evidence for the price formation conjecture as one major cause of spatial dependence. By reweighting a traditional inverse distance spatial lag according to the hedonic similarity, omitted variables that follow simple spatial patterns are unlikely to be falsely captured. Thus the degree to which price formation causes spatial dependence can be computed more accurately.

Secondly, the weight matrix specification can be used as a control for further studies into the price formation process. Alternatively, it may be used to exclude problems with omitted variables without adding further variables into the model.

A better understanding of the economic causes for spatial dependence may improve model reliability and open up new research questions into small scale local effects that are not captured adequately by traditional hedonic models.

References

Allen, M. T., Faircloth, S., & Rutherford, R. C. (2005). The Impact of Range Pricing on Marketing Time and Transaction Price: A Better Mousetrap for the Existing Home Market? *Journal of Real Estate Finance and Economics*, *31*(1), 71–82.

Allen, M. T., Rutherford, R. C., & Thomson, T. A. (2009). Residential Asking Rents and Time on the Market. *Journal of Real Estate Finance and Economics*, *38*(4), 351–365.

Benefield, J. D., Cain, C. L., & Johnson, K. H. (2011). On the Relationship Between Property Price, Time-on-Market, and Photo Depictions in a Multiple Listing Service. *The Journal of Real Estate Finance and Economics*, *43*(3), 401–422.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, *22*(3), 453–474.

Can, A., & Megbolugbe, I. (1997). Spatial Dependence and House Price Index Construction. *The Journal of Real Estate Finance and Economics*, *14*(1-2), 203–222.

Case, K. E., & Shiller, R. J. (2003). Is There a Bubble in the Housing Market? *Brookings Papers on Economic Activity*, *2003*(2), 299–342.

Haurin, D. (1988). The Duration of Marketing Time of Residential Housing. *Real Estate Economics*, *16*(4), 396–410.

Jud, G. D., & Frew, J. (1990). Atypicality and the Natural Vacancy Rate Hypothesis. *Real Estate Economics*, *18*(3), 294–301.

Quan, D. C., & Quigley, J. M. (1989). Inferring an Investment Return Series for Real Estate from Observations on Sales. *Journal of the American Real Estate & Urban Economics Association*, *17*(2), 218–230.

Tobler, W. R. (1979). Cellular Geography. In S. Gale & G. Olsson (Eds.), *Philosophy in Geography* (pp. 379–386). Springer Netherlands.

Turnbull, G. K., Dombrow, J., & Sirmans, C. F. (2006). Big House, Little House: Relative Size and Value. *Real Estate Economics*, *34*(3), 439–456.

Wong, S. K., Yiu, C. Y., & Chau, K. W. (2013). Trading Volume-Induced Spatial Autocorrelation in Real Estate Prices. *The Journal of Real Estate Finance and Economics*, *46*(4), 596–608.