Mass appraisal without statistical estimation: a simplified comparable sales approach based on a spatiotemporal matrix

DOI:

[10.1007/s00168-019-00959-2](https://doi.org/10.1007/s00168-019-00959-2)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Yousfi, S., Dubé, J., Legros, D., & Thanos, S. (2020). Mass appraisal without statistical estimation: a simplified comparable sales approach based on a spatiotemporal matrix. *Annals of Regional Science*, 64(2), 349–365. <https://doi.org/10.1007/s00168-019-00959-2>

Published in:

Annals of Regional Science

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Mass Appraisal without Statistical Estimation: A simplified Comparable Sales Approach based on a Spatio-temporal Matrix

Sonia Yousfi ¹, Jean Dubé ², Diègo Legros ³ and Sotirios Thanos ⁴

Abstract:

For mass appraisal in real estate, the hedonic pricing method (HPM) tends to be most commonly used by academic researchers and the comparable sales approach (CSA) is mostly preferred by professionals. This paper shows how CSA is a constrained version of a spatial autoregressive (SAR) model, which can be implemented by simple matrix calculations. The CSA takes into account information on individual characteristics identifying similar complex goods, spatial proximity reflecting similar spatial amenities and temporal constraints by only selecting past sales. Using US transaction data from Lucas County, Ohio, we compare CSA to a spatial HPM results, and conduct an out-of-sample exercise to gauge the prediction performance of the two approaches. The findings suggest that CSA is a very useful tool for mass appraisal, especially when the number of independent variables available is limited.

Keywords: Comparable sales approach; Hedonic pricing model; Spatial Econometrics; Spatio-temporal Weight Matrix; Out-of-sample prediction.

JEL Classification : C21; C40; R10; R15

This is author accepted manuscript version of the article. This article may be used for non-commercial purposes in accordance with Springer Nature Terms and Conditions for Self-Archiving. Please cite as:

Yousfi, S., Legros, D., Dubé, J., Thanos, S. (2019) Mass appraisal without statistical estimation: a simplified comparable sales approach based on a spatiotemporal matrix. *The Annals of Regional Science*, <https://doi.org/10.1007/s00168-019-00959-2>

¹ Université de Bourgogne, Laboratoire d'Economie de Dijon (LEDi), EA7467, 2 boulevard Gabriel, Dijon, France, 21066, E-mail: sonia.yousfi@u-bourgogne.fr

² Université Laval, 2325, rue des Bibliothèques, Québec, Canada, G1V 0A6, E-mail: jean.dube@esad.ulaval.ca

³ Université de Bourgogne, Laboratoire d'Economie de Dijon (LEDi), EA7467, 2 boulevard Gabriel, Dijon, France, 21066, E-mail: diego.legros@u-bourgogne.fr

⁴ University of Manchester, Humanities Bridgeford Street, Manchester, UK, M13 9PL, E-mail: sotirios.thanos@manchester.ac.uk

1. Introduction

Mass appraisal exercises are quite important not only for the real estate industry and finance sector to maximize profits and investment and minimize risk, but also for local and central government institutions to recover taxes, induce development, and address a number of economic and social issues tied to the housing stock value. There are three main methods for estimating the value houses: i) the comparable sales approach (CSA); ii) hedonic pricing method (HPM); and iii) the cost method (Pagourtzi et al., 2003).

HPM offers a solid theoretical framework for measuring the implicit prices of urban externalities and for mass appraisal (Des Rosiers et al., 2008). However, its implementation for prediction necessitates proficiency in statistical and modelling techniques to ensure an acceptable level of precision and stability of the value estimates. The CSA is an intuitive method that attempts to make inference on the final sale price of a given real estate good based on comparables that denote the information of past sale prices for similar goods (Ratcliffe, 1972). An a priori selection of comparables greatly simplifies the prediction process by eliminating the need for statistical analysis. The cost method calculates the full discounted cost of building a real estate good, taking into account obsolescence and depreciation.

While CSA and HPM greatly diverge on implementation, the two methods are intrinsically linked (Cavailhès, 2005; Huang, et al., 2017; Rodgers, 1994; Sheppard, 1997). The HPM puts emphasis on the individual characteristics of complex goods, treating the final price as a function of all the components' implicit prices, whereas CSA identifies goods with similar characteristics and uses their prices to estimate the price of the real estate good in question. CSA has been widely criticized on how the choice of the comparables is made, which in practice is often based on subjective criteria (French and Gabrielli 2004).

This paper does not focus on finding the most similar comparables in terms of structural characteristics, given the large literature on the topic (Krause and Kummerow, 2011; Isakson, 1986; Colwell et al, 1983; Ratcliffe 1972), but instead focuses on the spatial and temporal similarity between comparables that have identical housing attributes (Bidanset et al., 2018; Besner, 2002). Hence, the main aim is to provide a simple and intuitive method for implementing CSA, based on a spatio-temporal weight matrix specification (Dubé and Legros, 2013) for the selection of comparables or, in this case, spatio-temporal nearest neighbors. CSA is shown to be a constrained version of a simple spatial autoregressive (SAR) model extended to contain temporal connections. From a practical perspective, we explicitly link CSA with a matching estimator approach (Rosenbaum and Rubin, 1983). This facilitates the development of simple analysis methods that employ real estate transaction information. We provide an empirical illustration by employing a US data set from Lucas County, Ohio between 1993 and 1998. The empirical application compares the prediction power of the spatio-temporal nearest neighbors CSA to an "a-spatial" HPM. It is stressed that we do not follow Dubé et al., (2018), Thanos et al., (2016), or Pace et al (1998) in attempting to estimate a spatial econometric version of HPM that is assumed to be based on CSA. Instead, we specify CSA as a SAR taking into account temporal constraints on spatial relations, and compare its prediction accuracy to HPM.

The paper is structured as follows. The second section presents the conceptual framework, providing a clear and explicit illustration of HPM and CSA, and how CSA is a special case of SAR. The third section presents the methodology proposed to develop the CSA method based on simple mathematical constraints. The data used in statistical modeling are described in the fourth section. The fifth section illustrates and discusses the estimation and comparison results based on an out-of-sample exercise. Conclusions are drawn in section 6.

2. Conceptual framework

HPM, as first formally introduced by Rosen (1974), is widely used for predicting real estate prices. The Hedonic pricing model expresses the final sale price y_{it} of a complex good i sold at time t , typically in logarithmic form, as a function of its individual characteristics or amenities noted by x_{kit} . The heterogeneity of the individual goods is assumed to be expressed only by the differing quantity of each characteristic in each composite good.

In a linear regression context, the coefficients, β_k , related to the amenities reflect the implicit (or hedonic) price of the individual component as shown in equation 1.

$$\ln(y_{it}) = \alpha + d_{it}\delta_t + \sum_k x_{kit}\beta_k + \varepsilon_{it} \quad (1)$$

Where α is a constant term (the intercept), d_{it} is a dummy variable isolating the time period where transactions i occurs, δ_t is a coefficient for the nominal evolution of real estate prices, and ε_{it} is the error term.

The limits and criticisms of the HPM are well known (Pagourtzi et al. 2003; Dubé et al., 2014). The validity of the estimators and the model lies on three important assumptions: i) there is no unaccounted information influencing the price that is not included in the model; ii) the linear relation for the price is a good approximation of the unknown process and all non-linear relations are accounted for by mathematical transformations; and iii) the error term is independent and neither spatially nor temporally correlated. There is no empirical approach that ensures for all potentially influential variables on the price determination process to be included in the model. Since the work of Can (1992), it is well known that hedonic price equation residuals contain spatial patterns (Anselin and Lozano-Garcia, 2008), which signifies severe consequences on the validity of the estimators, depending on the source of these patterns (LeSage and Pace, 2009)

Colwell et al. (1983), followed by Pace and Gilley (1998), provide a connection between the HPM and CSA. Since the HPM equation holds for a house i sold at time t , it also holds for another transaction, c , that occurred before $(t-s)$. Hence, a number of comparables ($c = 1, \dots, C$) can be introduced in equation 1, weighted by w_c to express the sale price of good i as a function of the other goods c , as shown in equation 2.

$$\ln(y_{it}) = \alpha + d_{it}\delta_t + \sum_k x_{kit}\beta_k + \varepsilon_{it} + \sum_c w_c [\ln(y_{ct-s}) - [\alpha + d_{ct-s}\delta_c + \sum_k x_{kct-s}\beta_k + \varepsilon_{it-s}]] \quad (2)$$

Adding the last term on the right-hand side is mathematically neutral⁵. Equation 2 is rearranged to equation 3 to express the price equation based on the sale price of the comparable (neighbors) transactions.

$$\ln(y_{it}) = \sum_c w_c \ln(y_{ct-s}) + \alpha(1 - \sum_c w_c) + (d_{it}\delta_t - \sum_c w_c d_{ct-s}\delta_c) + (\sum_k x_{kit}\beta_k - \sum_c w_c \sum_k x_{kct-s}\beta_k) + (\varepsilon_{it} - \sum_c w_c \varepsilon_{it-s}) \quad (3)$$

In matrix notation, this is equivalent to the following equation:

$$y_{it} = W y_{ct-s} + \alpha(I - W)\mathbf{1} + (D_{it}\delta - W D_{ct-s}\delta) + (X_{kit}\beta - W X_{kct-s}\beta) + (\varepsilon_{it} - W \varepsilon_{it-s}) \quad (4)$$

⁵ The temporal ordering and arrow of time constraints to spatial matrices discussed in section 2.3 ensure that there are no endogeneities.

Where $\mathbf{W}$ is a weight matrix of dimension $(N_T \times N_T)$, $\mathbf{I}$ is the identity matrix of dimension $(N_T \times N)$, $\mathbf{D}$ is a matrix of dummy variables of dimension $(N_T \times (T-1))$ indicating the time of sale, $\mathbf{X}$ is a matrix of independent variables of dimension $(N_T \times K)$, $\boldsymbol{\varepsilon}$ is a vector of error term of dimension $(N_T \times 1)$, $\mathbf{1}$ is a vector that contains elements set to 1 of dimension $(N_T \times 1)$, α is a scalar, $\boldsymbol{\delta}$ is a vector of parameters of dimension $((T-1) \times 1)$, and $\boldsymbol{\beta}$ is another vector of parameters of dimension $(K \times 1)$. The specification takes a form familiar to spatial econometric models.

We proceed to assume that: i) comparable transactions are picked within the same time span ($d_i = \sum_c w_c d_{ct-s}$ or $\mathbf{D}_{it} = \mathbf{W}\mathbf{D}_{ct-s}$); ii) the characteristics of the comparables are identical ($x_{kit} = \sum_c w_c \sum_k x_{kct-s}$ or $\mathbf{X}_{kit} = \mathbf{W}\mathbf{X}_{kct-s}$); iii) the new error term ($\mathbf{u} = (\mathbf{I} - \mathbf{W})\boldsymbol{\varepsilon}$) is independent and identically distributed (iid); and iv) the weight matrices are row standardized ($\sum_c w_c = 1$). Under these assumptions, equation 4 is simplified to obtain the CSA specification in equation 5.

$$y_{it} = \mathbf{W}\mathbf{y}_{ct-s} + \mathbf{u}_{it} \quad (5)$$

There is a clear link to SAR with $\mathbf{W}\mathbf{y}_{ct-s}$ being the only independent variable with a coefficient constrained to be equal to one (LeSage and Pace, 2009; Dubé and Legros, 2014). However, the weight matrix $\mathbf{W}$ in equation 5 can neither be of full rank nor symmetric, since any element of the matrix, w_{ij} , takes a value different from zero, if and only if transaction j has occurred before transaction i ($t_j < t_i$). When transactions are chronologically ordered, $\mathbf{W}$ takes the form of a lower triangular matrix. This is a more realistic vision of how the spatial spillovers effect occur in real estate market (Pace et al., 1998; Dubé and Legros 2013).

In the professional practice, the weights, w_{ij} (the individual component of the $\mathbf{W}$ matrix), are such that each comparable, c , has a uniform and equal weight ($1/C$), with the number of comparable typically set to as low as 5 (Ratcliffe 1972). This weighting scheme can easily be extended to take into account spatial and temporal distances by adopting a spatio-temporal structure (Dubé et al., 2018; Thanos et al., 2016) that respects temporal constraints on spatial relations.

Both HPM and CSA allow for mass appraisal based on information from previous transactions. They offer a useful framework for predicting the price of a given real estate good, j , to be sold in a near future ($t + 1$), $\hat{y}_{jt+1}$. The HPM prediction exercise is based on the estimation of a set of vector of coefficients (δ_t and β_k), while the CSA prediction exercise is based on identifying the comparable (through the weights w_c - or w_{ij}). As the a-spatial HPM only examines the characteristics of complex real estate goods, it completely ignores the possibility of spatial spillover effects. However, the CSA accounts for spatial spillover effects as it is uniquely based on the prices of spatio-temporal neighbors, while also implicitly controlling for individual characteristics by selecting similar goods.

3. Methodological approach

The estimation and comparison of prediction performance between HPM and CSA is conducted by employing an out-of-sample one-step-ahead forecasting approach. For HPM, equation 1 is estimated for transactions in each time period to predict house prices for the subsequent time period. The procedure is repeated for all time periods ($t + 1, \dots, t + s, \dots, t + S$) until the last transactions. This iterative approach results in the coefficients changing over time, which points to a framework similar to CSA, ensuring consistency of comparisons.

The CSA prediction is based on identifying real estate goods with similar characteristics. The characteristics are decomposed in two distinct sets: i) one for discrete or categorical variables, such as number of rooms, bedrooms, bathrooms or garages; and ii) one for continuous variables, such as age, lot size, or living area. Figure 1 illustrates the criteria for identifying similar houses that also depend on the type of variables considered. The selection of a limited number of

comparables is based on spatial and temporal distance matrices, which are combined through a Hadamard matrix product to obtain a spatio-temporal weight matrix (Dubé and Legros, 2014; 2013). Combining both spatial and temporal restrictions allows us to limit the number of comparables or nearest neighbors. Drawing from professional practice, the number of comparables is typically low (Ratcliffe 1972). The individual weights are row-standardized and based on spatio-temporal distance.

Figure 1: The steps of identifying comparables and predicting prices

Figure 1 decomposes the different computational steps. Based on previous transactions, a list of comparable goods can be identified using the distance between observation i and j , d_{kij} , for a given characteristic, x_{ki} , with $k = 1, \dots, K$ (equation 6).

$$d_{kij} = \|x_{ki} - x_{kj}\| \quad (6)$$

Where $\|\dots\|$ is a general distance function. To simplify this exposition, the absolute value operator is used, which can readily be changed to other attribute distance functions, such as the Mahalanobis distance.

For discrete variables, a general indicator variable, c_{kij}^k , takes the value of one, if given characteristics of goods i and j are identical, and 0, otherwise. For the continuous variable, the difference can be set equal to a predefined critical value, θ_k (equation 7).

$$c_{kij} = \begin{cases} 1 & \text{if } d_{kij} = 0, \text{ when } x_k \text{ is a discrete variable,} \\ 1 & \text{if } d_{kij} < \theta_k, \text{ when } x_k \text{ is a continuous variable} \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

Combining all the different indicator variables, the list of comparable sales can be identified for unsold properties (equation 8).

$$c_{ij} = c_{1ij} \times c_{2ij} \times c_{3ij} \times \dots \times c_{Kij} \quad (8)$$

The expression can be translated into matrix notation. By noting $\mathbf{X}_k$ as a vector of k characteristic ($k = 1, \dots, K$) of dimensions $(N \times 1)$, and $\mathbf{1}$ a vector of elements set to one (1) of dimensions $(N_T \times 1)$, it is possible to build a matrix $\mathbf{C}_k$ of dimensions $(N_T \times N_T)$, which expresses the distance between all observations (equation 9). All $\mathbf{C}_k$ ($k = 1, \dots, K$) are pooled together by the use of a Hadamard product (term-by-term multiplication), $\odot$, to construct the matrix $\mathbf{C}$ of dimensions $(N_T \times N_T)$ in equation 10.

$$\mathbf{C}_k = (\mathbf{X}_k \times \mathbf{1}') - (\mathbf{1} \times \mathbf{X}_k') \quad (9)$$

$$\mathbf{C} = \mathbf{C}_1 \odot \mathbf{C}_2 \odot \dots \odot \mathbf{C}_K \quad (10)$$

Where the elements of the matrix $\mathbf{C}$, take a value of one, if observation j is identified as a comparable to observation i , and zero, otherwise.

The second step in Figure 1 also takes into account spatio-temporal proximity, which requires the estimation of two additional distance measures. The spatial distance between observations, d_{sij} , is calculated based on the individual geographical coordinates of transaction i , noted u_i and v_i in equation 11⁶.

$$d_{sij} = [(u_i - u_j)^2 + (v_i - v_j)^2]^{(1/2)} \quad (11)$$

The temporal distance, d_{tij} , between i and j is shown in equation 12.

$$d_{tij} = |t_i - t_j| \quad (12)$$

Where t_i and t_j is the time period respective transactions took place. Multiplying the individual spatial and temporal distances between comparable goods allows to retrieve a spatio-temporal weighted prices in equation 13. This approach can be used to isolate a predefined number of neighbors, who are assigned either a uniform or a spatio-temporal weight, depending on the selected specification. This approach takes into account both the first law of geography (Tobler, 1970), as well as the temporal constraints (Hägerstrand, 1989).

$$w_{ij} = f(d_{sij}) \times g(d_{tij}) \quad (13)$$

Where $f(\cdot)$ and $g(\cdot)$ signify functions that can be adjusted to the particular spatiotemporal context. These mathematical transformations can also be translated into matrix notation by noting $\mathbf{U}$ a vector of the X geographical coordinates, $\mathbf{V}$ a vector of Y geographical coordinates, and $\mathbf{\Gamma}$ a vector of sale date, all of dimension $(N \times 1)$. The matrix of spatial distances, $\mathbf{D}_s$ (equation 14), and the matrix of temporal distance, $\mathbf{D}_t$ (equation 15), can easily be computed for all observations.

$$\mathbf{D}_s = [(\mathbf{D}_X \odot \mathbf{D}_X) + (\mathbf{D}_Y \odot \mathbf{D}_Y)]^{(1/2)} \quad (14)$$

$$\mathbf{D}_t = (\mathbf{\Gamma} \times \mathbf{1}') - (\mathbf{1} \times \mathbf{\Gamma}') \quad (15)$$

Where $\mathbf{D}_X$ and $\mathbf{D}_Y$ are defined as follows.

$$\mathbf{D}_X = (\mathbf{U} \times \mathbf{1}') - (\mathbf{1} \times \mathbf{U}') \quad (16)$$

$$\mathbf{D}_Y = (\mathbf{V} \times \mathbf{1}') - (\mathbf{1} \times \mathbf{V}') \quad (17)$$

Combining the spatial weight matrix, $\mathbf{S} = f(\mathbf{D}_s)$ with the temporal weight matrix, $\mathbf{T} = g(\mathbf{D}_t)$ returns the spatio-temporal weight matrix, $\mathbf{W}$ (equation 18).

$$\mathbf{W} = \mathbf{S} \odot \mathbf{T} \quad (18)$$

When data are chronologically ordered, the spatio-temporal weight matrix takes a lower-triangular form, since only previous transactions can be used to predict the sale price of future transactions.

Proceeding to the third step in Figure 1, the matrices of comparable goods $\mathbf{C}$ and spatio-temporal weights $\mathbf{W}$ are combined in equation 19. $\mathbf{W}^*$ is then row-standardized and isolates the observations to be used in making inference (equation 20).

$$\mathbf{W}^* = \mathbf{C} \odot \mathbf{W} \quad (19)$$

$$\mathbf{y}_{it} = \mathbf{W}^* \mathbf{y}_{it-s} \quad (20)$$

Equivalently, $y_{it} = \sum_j [w_{ij}^* \times y_{jt-s}]$, with w_{ij}^* being an element of $\mathbf{W}^*$.

⁶ We use Euclidian distance, but other distance measurements can also be applied.

In the end, prediction is only possible, if and only if some individual elements of the rows of the matrix $\mathbf{W}^*$ are different from zero. Otherwise, the operation involves multiplications by 0, which makes the calculation impossible.

4. Data

The empirical investigation is based on transactions available for Lucas County (Ohio, USA) between 1993 and 1998 (LeSage and Pace, 2004; Dubé and Legros 2013).⁷ The dataset contains information on sale prices (in nominal US\$), the exact location of the house (geographical coordinates), the exact sale date and a number of property characteristics. For the empirical investigation, eight different variables are used: 3 are continuous; 4 are categorical variables; and 1 is a dummy variable (Table 1).

Table 1: Variable description

List of variables	Description
Discrete independent variables	
Number of rooms	Total number rooms in the house (1: < 5 rooms; 2: 6 rooms; 3: 7 rooms; 4: > 8)
Number of bedrooms	Total number of bedrooms (2: 2 or less; 3: 3 bedrooms; 4: more than 3)
Number of bathrooms	Number of bathrooms (1: 2 or less; 2: more than 2)
Garage	Dummy variable (1: with a garage; 0: no garage)
Number of stories	Number of stories (1: 1 story; 2: between 1 and 2; 3: more than 2)
Continuous independent variables	
Living area (in m ²)	Total living area in square meters
Lot size (in m ²)	Total lot size area in square meters
Age (in years)	Age, based on the construction year (transaction year - construction year)
Dependent variable	
Sale price (in \$)	Nominal sale price (in \$US)
Other key variables	
XMTM	X geographical coordinates (in meter) - MTM projection
YMTM	Y geographical coordinates (in meter) - MTM projection
Sale date	Exact sale (transaction) date:

This data is not used to demonstrate the superiority of one method over the other. The empirical strategy is mainly devoted to implement the proposed methodology. The comparison only serves as a benchmark for a basic specification to check whether this simple mass appraisal proposition offers any performance advantages to out-of-sample prediction.

A total of 25,357 transactions is available (Table 2). The data are split into two sub-samples. The first sub-sample includes all 1993-1997 transactions and is used as exogenous information to calibrate HPM and to provide the information on past transactions in CSA. The second sub-sample of 1998 transactions is used for out-of-sample inference. The first sub-sample contains 20,979 observations, while the second includes 4,378 observations.⁸ As the forecasting method is based on a one-step-ahead process, the in-sample size increases with time.

⁷ The database is available in the Matlab library

⁸ The database used for making inference (see the code in Appendix) is available upon request.

Table 2: Descriptive statistics

Variables	In sample (1993-1997)		Out-of-sample (1998)		Difference	
	Mean	s.d.	Mean	s.d.	t-test	sign.
Discrete						
Nb of rooms	2.1429	1.0371	2.0955	1.0137	0.0474	***
Nb of bedrooms	2.9796	0.6519	2.9459	0.6431	0.0337	***
Nb of bathrooms	1.2268	0.4188	1.1907	0.3929	0.0361	***
Garage	0.8645	0.3423	0.8527	0.3545	0.0118	**
Number of stories	1.8084	0.8900	1.7969	0.8858	0.0115	
Continuous†						
Living area (in m ²)	1,470	618	1,425	589	45.01	***
Lot size (in m ²)	13,508	29,261	12,490	27,343	1,017.48	**
Age (in years)	50	28	54	27	-4.30	***
Dependent variable						
Sale price (in \$)	77,901	58,739	84,370	63,596	-6,468.71	***
N	20,979		4,378			

Legend: † Continuous variables are used to perform gamma transformation.

The nominal mean sale price is 79,017.94 US\$ for the whole sample, but differences emerge between the two sub-samples (Table 2). This is related to the fact that sale price is expressed in nominal terms. For the other characteristics, the difference between sub-samples is small but statistically significant. Except for the number of stories, all the characteristics show different structures within the sub-samples.

4. Results and Discussion

One of the weakness of the CSA is that prediction is possible only if there is a comparable available to make a projection. Consequently, it is impossible to make an out-of-sample prediction for all transactions occurring in 1998, because some transactions are largely different from others. A total of 3,761 predicted sale prices is calculated and 550 transactions do not have any comparable, as shown in Table 3. Hence, 87.4% of the total out-of-sample observations (3,828 transactions) can be used to compare CSA and HPM. Furthermore, 844 transactions have less than five previously sold comparables, even though they might have similar characteristics to many other transactions. This results to 2,984 of the out-of-sample data having the required five spatio-temporal neighbors to comparables for prediction.

Table 3: Potential and final number of NNs for the out-of-sample data

Number of neighbors (comparable)	Obs.	Potential neighbors		
		Same characteristics		
		Mean	Min	Max
0	550	1.08	1	4
1	340	2.18	2	6
2	217	3.26	3	7
3	147	4.44	4	9
4	140	5.44	5	9
5	2,984	92.09	6	647
Total	4,378	63.55	1	647

The comparison between HPM and CSA lies on a limited set of observable characteristics. There are different tradeoffs for these methods in terms of the amount of attribute information. For CSA, the more characteristics are included to define the comparables, the less neighbors are available. For HPM, the more characteristics are included, the higher the explanatory power and prediction performance becomes. In situations with limited availability of housing information, the prediction performance of the a-spatial HPM necessarily suffers. For this reason, the analysis proposes a simple extension of the model based on the introduction of non-linear effect, while preserving the “linear in the parameter” condition. To do so, the model considers the level of the continuous variable, as well as the logarithmic transformation, x_{kit} and $\log(x_{kit})$. This is known as the gamma transformation (Des Rosiers et al., 1996, 2001), allows for non-linear impact without introducing too high correlation among the independent variables. Depending on the sign related to the variables, x_{kit} or $\log(x_{kit})$, this transformation provides interesting marginal effect (Figure 2).

Figure 2: Example of non-linear transformation for distance to CBD

Note: The parameters are set to: $|\beta_x| = 0.05$ and $|\beta_{\log(x)}| = 0.15$.

The comparison of the methods is based on the actual sale price (in log), $\ln(y_{it})$, and the predicted sale price (in log), $\ln(\hat{y}_{it})$. A total of seven different statistics is calculated: i) the mean square error (MSE), defined as $MSE = (1/N_t) \sum_t [\ln(y_{it}) - \ln(\hat{y}_{it})]^2$, where N_t is the total number of observations in the out-of-sample; ii) the root mean square error (RMSE), defined as $RMSE = (MSE)^{(1/2)}$; iii) the mean standard deviation (MSD), as defined by $MSD = (1/N_t) \sum_t [\ln(y_{it}) - \ln(\hat{y}_{it})]$; iv) the mean absolute deviation (MAD), as defined by $MAD = (1/N_t) \sum_t [|\ln(y_{it}) - \ln(\hat{y}_{it})|]$; v) the coefficient of dispersion (COD), as defined by $(1/N_t) \sum_t [(\ln(\hat{y}_{it}) - \ln(\hat{y}_{it})) / \ln(\hat{y}_{it})]$, where $\ln(\hat{y}_{it})$ is the median of $\ln(y_{it})$; vi) the pseudo- R^2 statistic, as defined by $\rho(y_{it}; \hat{y}_{it})$, with ρ indicating their correlation; and vii) a classification based on the method that provides the smallest gap between prediction and actual sale price.

The results in Table 4 suggest that compared to the typical HPM, CSA provides superior out-of-sample prediction performance, but similar performance for the HPM based on the gamma

transformation for the continuous variables. Even though the global performance of the models (Pseudo- R^2) is similar for both approaches, CSA provides predictions closer to the actual price.

The MSE, RMSE and MAD show CSA to perform better than the typical HPM, and worse than the gamma transformation HPM. This can be partly explained by the prediction pattern of both methods shown in Figure 3. Both methods underestimate the sale prices at the high end of the price spectrum, but HPM highly overestimates the prices at the lower end of the spectrum as well. In all cases, the gap for the prediction of high prices has a more pronounced impact in the HPM. This may largely explain why many empirical investigations of the HPM exclude from the analysis transactions that sold at high prices.

Table 4: Performance and out-of-sample statistics by method

Statistics (dependent variable: $\ln(y_{it})$)	Method		
	HPM		CSA
	Typical	gamma	
Number of best predictions	1,574	1,882	
Mean square error (MSE)	0.2303	0.1549	0.1868
Root mean square error (RMSE)	0.4799	0.3936	0.4322
Mean absolute deviation (MAD)	0.3450	0.2883	0.2965
Mean standard deviation (MSD)	-0.0208	-0.0185	0.0155
Coefficient of dispersion (COD)	0.0384	0.0468	0.0441
Pseudo- R^2	0.6144	0.7344	0.6814

Note: N = 3,828 transactions with at least one neighbor

Figure 3: CSA and HPM comparison of performance

Figure 3.1

Figure 3.2

Figure 3.3

The results in Figure 2 underline how the a-spatial HPM fails to precisely predict sale prices, especially at the high end of the price spectrum, under conditions of limited information reflected in the low number of independent variables in this model. This is to be expected, as in linear regression models, the “mean” price pattern contained by the limited explanatory information available, increasingly failing to produce accurate predictions as the prices move to the high end of the spectrum. The increasing prediction error may also be due to unaccounted non-linearities. Nevertheless, this exercise, with the very limited number of independent variables, does not necessarily show that HPM is inadequate, but that information constraints do have a serious impact on HPM predictions.

An important finding is that CSA proves to be a useful tool for predicting prices under information constraints, when only few characteristics are available. This is due to the implicit spatial and temporal structures captured in the comparables potentially accounting for unobservable characteristics with a spatial structure. A close link between the CSA method, SAR and spatio-temporal autoregressive (STAR) models is also underlined, which is a reason why CSA performs better than the a-spatial HPM. As underlined in recent studies (Dubé et al., 2018; Thanos et al., 2016), taking into account the temporal constraints of the spatial distribution of sales has a positive effect on the performance of the prediction power, with most of the spatial spillover components being related to previous transactions in spatial proximity.

There is an important drawback in that CSA may fail to identify an adequate number of comparables for some observations; thus, not all observations can be included in such analysis. This can only be partly ameliorated by relaxing some of the similarity constraints in defining comparables. There is also the option of introducing an adjustment based on attribute difference between comparables (see equation 4). However, this also means that CSA needs to deal with the main HPM disadvantage, the need to capture the implicit price of all characteristics from unbiased regression coefficients. This further illustrates the trade-off in CSA regarding the precision in identifying comparables, including the spatial and temporal dimensions that we argue should take center stage in the process.

7. Conclusions

This paper introduces a computationally simple and useful forecasting tool based on spatio-temporal nearest neighbors and a matching estimator approach (Rosenbaum and Rubin, 1983). We developed a SAR specification of the CSA with additional controls for the temporal relations and strict similarity constraints between comparables. Using US transaction data from Lucas County, Ohio, we conducted an out-of-sample exercise to compare the prediction accuracy between the CSA and an a-spatial HPM. The CSA based on the limited information of only a few independent variables outperforms the a-spatial HPM, as is demonstrated in the results section. This is potentially due to the implicit spatial and temporal patterns captured in the comparables, accounting for characteristics with a spatial structure. This also underlines a link between the CSA, SAR, and STAR (Thanos et al., 2016; Dubé et al., 2018) in terms of spatial or spatiotemporal structure, which may be at least partly the reason why CSA outperforms the a-spatial HPM.

This exercise is not an attempt to argue in favor of either CSA or HPM. There are indeed a number of papers (Dubé et al., 2018; Thanos et al., 2016; Pace and Gilley, 1998; Pace et al., 1998) estimating a spatial econometric version of an HPM that is assumed to be based on the CSA. Nevertheless, there are significant opportunities for further research, as the significant and distinct advantages of each of the two methods have not yet been fully integrated into a single framework.

From the perspective of the real estate industry, the CSA proposed in this paper can potentially provide a valuable tool, as does it not require statistical analysis, works well under informational constraints, and improves the comparable selection process by reducing arbitrariness and subjectivity. The HPM is one of the key methods in measuring the impact of urban externalities through real estate transactions. The use of CSA is also relevant to further research concerning policy evaluation, such as landscape policy (Devaux et al., 2017). CSA can also be potentially combined with other HPM based methods, such as the difference-in-differences (DID) approach, to evaluate the impact of extrinsic amenities on house prices.

References

- Anselin L, Lozano-Gracia N (2008) Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empir Econ* 34(1):5–34
- Besner C (2002) A spatial autoregressive specification with a comparable sales weighting scheme. *J Real Estate Res* 24(2):193–211
- Bidanset P, McCord M, Lombard JA, Davis P, McCluskey W (2018) Accounting for locational, temporal, and physical similarity of residential sales in mass appraisal modeling: introducing the development and application of geographically, temporally, and characteristically weighted regression (GTCWR). *J Prop Tax Assess Adm* 14(2):5–13
- Can A (1992) Specification and estimation of hedonic housing price models. *Reg Sci Urb Econ* 22(3):453–474
- Cavailhès J (2005) Le prix des attributs du logement. *Écon Stat* 381–382:91–123
- Colwell PF, Cannaday RE, Wu C (1983) The analytical foundations of adjustment grid methods. *Real Estate Econ* 11(1):11–29
- Des Rosiers F, Lagana A, Thériault M, Beaudoin M (1996) Shopping centers and house values: an empirical investigation. *J Prop Valuat Invest* 14(4):41–63
- Des Rosiers F, Lagana A, Thériault M (2001) Size and proximity effects of primary schools on surrounding house values. *J Prop Res* 18(2):1–20
- Des Rosiers F, Dubé J, Thériault M (2011) Hedonic price modelling: measuring urban externalities in Quebec. In: dans Thériault M, Des Rosiers F (eds) *Modelling urban dynamics: mobility, accessibility and real estate value*. ISTE-Wiley, London, pp 255–283
- Devaux N, Dubé J, Apparicio P (2017) Anticipation and post-construction impact of a metro extension on residential values: the case of Laval (Canada), 1995–2013. *J Transp Geogr* 62:8–19
- Dubé J, Legros D (2013) Dealing with spatial data pooled over time in statistical models. *Lett Spat Resour Sci* 6(1):1–18
- Dubé J, Legros D (2014) *Spatial econometrics using microdata*. GIS and territorial intelligence. Wiley, New York
- Dubé J, Legros D, Thériault M, Des Rosiers F (2014) A spatial difference-in-differences estimator to evaluate the effect of change in public mass transit systems on house prices. *Transp Res Part B Methodol* 64(C):24–40
- Dubé J, Legros D, Thanos S (2018) Past price “memory” in the housing market: testing the performance of different spatio-temporal specifications. *Spat Econ Anal* 13(1):118–138
- French N, Gabrielli L (2004) The uncertainty of valuation. *J Prop Invest Finance* 22(6):484–500
- Hagerstrand T (1989) Reflections on “what about people in regional science?”. *Pap Reg Sci* 66(1):1–6
- Huang Z, Chen R, Xu D, Zhou W (2017) Spatial and hedonic analysis of housing prices in Shanghai. *Habitat Int* 67:69–78
- Isakson H (1986) The nearest neighbors appraisal technique: an alternative to the adjustment grid methods. *Real Estate Econ* 14(2):274–286

- Krause A, Kummerow M (2011) An iterative approach to minimizing valuation errors using an automated comparable sales model. *J Prop Tax Assess Adm* 8:39–52
- LeSage JP, Pace RK (2004) Models for spatially dependent missing data. *J Real Estate Finance Econ* 29(2):233–254
- LeSage J, Pace KR (2009) Introduction to spatial econometrics. Chapman and Hall/CRC, London
- Pace RK, Gilley OW (1998) Generalizing the OLS and grid estimators. *Real Estate Econ* 26(2):331–347
- Pace RK, Barry R, Sirmans CF (1998) Spatial statistics and real estate. *J Real Estate Finance Econ* 17(1):5–13
- Pagourtzi E, Assimakopoulos V, Hatzichristos T, French N (2003) Real estate appraisal: a review of valuation methods. *J Prop Invest Finance* 21(4):383–401
- Ratcliffe R (1972) Valuation for real estate decisions. Democrat Press, Washington
- Rodgers T (1994) Property-to-property comparison. *Apprais J* 62(1):64–67
- Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *J Polit Econ* 82(1):34–55
- Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
- Sheppard S (1997) Chapter 8: Hedonic analysis of housing markets. In: Cheshire PC, Mills ES (eds) *Handbook of regional and urban economics*, vol 3. Elsevier, Amsterdam, pp 1595–163
- Thanos S, Dubé J, Legros D (2016) Putting time into space: the temporal coherence of spatial applications in the housing market. *Reg Sci Urb Econ* 58:78–88
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46:234–240

Appendix

```

clear all
clear mata          /*Clearing all possible values in Mata*/
use "LucasCountyDataFinal.dta", clear
/* Grouping low frequencies */
quietly replace garage = 1 if garage>=1 /* Garage presence */
quietly replace numtoilet = 1 if numtoilet>=1 /* having one toilet or more */
quietly replace rooms = 1 if rooms<=5 /* a house must have at least 5 rooms */
quietly replace rooms = 2 if rooms==6 /* house with 6 rooms */
quietly replace rooms = 3 if rooms==7 /* house with 7 rooms */
quietly replace rooms = 4 if rooms>=8 /* house with more than 7 rooms */
quietly replace numbath = 1 if numbath==0 /* house must have at least 1 bathroom */
quietly replace numbath = 2 if numbath>=2 /* having 2 bathrooms or more */
quietly replace numbed = 1 if numbed==0 /* a house must have at least 1 bedroom */
quietly replace numbed = 2 if numbed==1 /* having 2 bedrooms or less */
quietly replace numbed = 4 if numbed>=5 /* having 4 bedrooms or more */
quietly replace stories = 2 if stories>=2 & stories<=4 /* non-standard */
quietly replace stories = 3 if stories>=5 /* having 2 stories or more */

**Generate geographical variables in km
quietly generate xc = xmtm/1000
quietly generate yc = ymtm/1000
/*Building interest variables - in log*/
quietly generate age = year - yrconstruction
quietly replace age = 0 if age<0
quietly generate lny = log(saleprice)
quietly generate log_livearea = log(livearea + 1)
quietly generate log_lotsize = log(lotsize + 1)
quietly generate log_age = log(age + 1)
quietly generate outsample = year==1998
/*Creating a continuous temporal variable*/
gen time = mdy(month, day, year)
sort time          /*Chronologically ordering observation*/

global nkdc = 5 /*Number of discrete characteristics*/
global nkcc = 3 /*Number of continuous characteristics*/
global nnc = 5 /*Number of nearest neighbours*/
global start = $nkdc + 1
global end = $nkdc + $nkcc

mata:N = st_nobs()
/*Import vectors to build comparables*/
mata:x1 = st_data(., "rooms") /*Number of rooms*/
mata:x2 = st_data(., "numbed") /*Number of bedrooms*/
mata:x3 = st_data(., "numbath") /*Number of bathrooms */
mata:x4 = st_data(., "garage") /*Presence of a garage*/
mata:x5 = st_data(., "stories") /*Stories*/
mata:x6 = st_data(., "log_livearea") /*Living area*/
mata:x7 = st_data(., "log_lotsize") /*Lotsize*/
mata:x8 = st_data(., "log_age") /*Age of the dwelling*/

/*Identifying observations with the same/similar characteristics*/
**Discrete characteristics (same)
forvalues kd = 1/$nkdc {
/*Identify comparable goods*/
mata:C`kd' = (x`kd'*J(N,1,1)) - (J(N,1,1)*x`kd')
mata:C`kd' = (C`kd'==0) /*Same characteristics*/
}
**Continuous characteristics (similar)
forvalues kc = $start/$end {
mata:C`kc' = (x`kc'*J(N,1,1)) - (J(N,1,1)*x`kc')
mata:C`kc' = (abs(C`kc')<=.15) /*Similar characteristics*/
}

/*Identify comparables with matrix CO*/
mata:CO = C1:C2:C3:C4:C5:C6:C7:C8
mata:mata drop C1 C2 C3 C4 C5 C6 C7 C8

/*Import relevant spatial and temporal vectors*/
mata:XC = st_data(., "xc")

```

```

mata:YC = st_data(., "yc")
mata:T = st_data(., "time")
/*Building matrix spatial weights matrix*/
mata:DX = (XC*J(N,1,1)') - (J(N,1,1)*XC')
mata:DY = (YC*J(N,1,1)') - (J(N,1,1)*YC')
mata:DS = sqrt((DX:DX)+(DY:DY))
mata:S = exp(-DS/1000)
mata: mata drop DX DY DS
/*Building temporal weights matrix*/
mata:DT = (T*J(N,1,1)') - (J(N,1,1)*T')
mata:DT = lowertriangle(DT)
mata:T = exp(-DT/90) - (DT==0)
mata: mata drop DT
/*Building spatio-temporal weights matrix*/
mata:W = S:T
mata: mata drop S T
/*Building potential comparable matrix*/
mata:CW = CO:W
/*Calculating the number of "neighbours"*/
mata:NVC = rowsum(CO)
mata:NVW = rowsum(W:>0)
mata:NVT = rowsum(CW:>0)
mata: mata drop CO W
**Building the nearest neighbours matrix
mata:nn1 = rowmax(CW) /*Finding maximum weight (nearest neighbour)*/
mata:NN1 = (CW==nn1) /*Isolating the first nearest neighbour*/
mata:WT1 = CW - (CW*NN1) /*Adjusting spatio-temporal weights matrix*/
forvalues n = 2/$nn {
    global m = `n' - 1
    mata:nn`n' = rowmax(WT$m) /*Finding maximum weight*/
    mata:NN`n' = (WT$m==nn`n') + NN$m /*Isolating neighbour (k = n)*/
    mata:WT`n' = WT$m - (CW*(WT$m==nn`n')) /*Adjusting weights matrix*/
    mata: mata drop nn$m NN$m WT$m
}

/*Spatio-temporal nearest neighbors matrix - row-standardized*/
mata:WS = NN$nn:(CW)
mata:RS = rowsum(WS)
mata:WSS = WS/(RS + (RS==0))
mata:RSS = rowsum(WSS:>0)

/*Compute estimated price by CSA*/
mata:y = st_data(., "saleprice") /*Sale price (final) - $constant */
mata:yp = WSS*y /*Sale price (nominal) prediction*/

/*Export data on neighbours number*/
mata:st_addvar("long", "NbComparables") /* # of comparable goods*/
mata:st_addvar("long", "NbSTSales") /* # of (ST) neighbours*/
mata:st_addvar("long", "NbPotential") /* # of ST comparable*/
mata:st_addvar("long", "NbUsed") /* # of transactions used*/
mata:st_store(., ("NbComparables", "NbSTSales", "NbPotential", "NbUsed"), (NVC, NVW, NVT, RS
S)) /*Export results*/

/*Export prevision from CSA*/
mata:st_addvar("double", "ypred") /*Compute mean nominal price*/
mata:st_store(., ("ypred"), (yp)) /*Export results*/

```