the review of income and wealth

Review of Income and Wealth
Series 68, Number 3, September 2022
DOI: 10.1111/roiw.12534

RESIDENTIAL PROPERTY PRICE INDEXES: SPATIAL COORDINATES VERSUS NEIGHBORHOOD DUMMY VARIABLES

BY W. ERWIN DIEWERT

University of British Columbia

AND

CHIHIRO SHIMIZU*

The University of Tokyo

This paper addresses the following question: can satisfactory residential property price indexes be constructed using hedonic regression techniques where location effects are modeled using local neighborhood dummy variables or is it necessary to use spatial coordinates to model location effects. Hill and Scholz (2018) addressed this question and found, using their hedonic regression model, that it was not necessary to use spatial coordinates to obtain satisfactory property price indexes for Sydney. However, their hedonic regression model did not estimate separate land and structure price indexes for residential properties. To construct national balance sheet estimates, it is necessary to have separate land and structure price indexes. The present paper addresses the Hill and Scholz question in the context of providing satisfactory residential *land* price indexes. The spatial coordinate model used in the present paper is a modification of Colwell's (1998) spatial interpolation method. The modification can be viewed as a general nonparametric method for estimating a function of two variables.

JEL Codes: C2, C14, C21, C23, C25, C43, R21

Keywords: residential property price indexes, System of National Accounts, balance sheets, methods of depreciation, land and structure price indexes, hedonic regressions, spatial coordinates, bilinear interpolation methods, nonparametric methods for fitting two-dimensional surfaces

1. INTRODUCTION

It is a difficult task to construct constant quality price indexes for residential (and commercial) properties. Properties with structures on them consist of two main components: the land component and the structure component. The problem is that each property has a unique location (which affects the price of the land component), and given the fact that the same property is not sold in every period, it is difficult to apply the usual matched model methodology when constructing constant quality price indexes. Bailey *et al*. (1963) developed the repeat sales methodology in an attempt to apply the matched model methodology to the problem

Note: The authors thank Jan de Haan, Alicia Rambaldi, Nigel Stapledon, Naohito Abe, and two referees for helpful comments and gratefully acknowledge financial support from the SSHRC of Canada and the JSPS (the Japan Society for the Promotion of Science) Grant-in-Aid for Scientific Research 20H00082.

*Correspondence to: Chihiro Shimizu, Center for Spatial Information Science (CSIS), The University of Tokyo 5-1-5, Kashiwanoha, Kashiwa, 277-8568, Japan (cshimizu@csis.u-tokyo.ac.jp).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

of constructing property price indexes, but this methodology ignores single sales of a property over the sample period. Thus, in particular, the sales of properties with new structures do not affect the resulting indexes, which could lead to biased indexes. Moreover, properties with structures on them do not retain the same quality over time due to structure depreciation or additions and renovations to the structure. Thus the matched model methodology for the construction of constant quality price indexes does not work well in the property price index context.

A possible solution to the above measurement problem is to use a hedonic regression model approach to the construction of property price indexes.[1] This approach regresses the sale price of a property (or the logarithm of the sale price) on various characteristics of the properties in the sample. An important price determining characteristic of a property is its location. The location of a property can be described by its neighborhood (a local government area or a postal code) or by its latitude and longitude (the spatial coordinates of the property). Most hedonic property regressions use the former approach for describing the location of a property, but in recent years, the availability of spatial coordinate information has grown. Colwell (1998) was an early pioneer in the use of spatial coordinate information in a property price regression and more recently, Hill and Scholz (2018) used spatial coordinates to model Sydney house prices.

The main question that this paper addresses is the following one: can satisfactory residential property price indexes be constructed using hedonic regression techniques where location effects are modeled using local neighborhood dummy variables or is it necessary to use spatial coordinates to model location effects? Hill and Scholz (2018) addressed this question and found that it was not necessary to use spatial coordinates to obtain satisfactory property price indexes for Sydney. However, their hedonic regression model did not estimate separate land and structure price indexes for residential properties. To construct national balance sheet estimates, it is necessary to have separate land and structure price indexes. The present paper addresses the Hill and Scholz question in the context of providing satisfactory residential *land* price indexes. The spatial coordinate model used in the present paper is a modification of Colwell's (1998) spatial interpolation method. The modification can be viewed as a general nonparametric method for estimating a function of two variables.

A basic building block in Colwell's method is a method of *bilinear interpolation* over a square that was developed in the mathematics literature. We explain this method in Section 2.

In Section 3, we explain how this bilinear method of interpolation over a square can be extended to a method of interpolation over a grid of squares. We then follow the example of Poirier (1976) and Colwell (1998) and convert the interpolation method into an econometric estimation model. The resulting method will be used in later sections to model the land price of a property as a function of its spatial coordinates.

In Appendix B, we compare Colwell's spatial coordinate model with the penalized least squares approach used by Hill and Scholz (2018) in their study of Sydney

[1]For expositions of the hedonic regression approach to the construction of constant quality price indexes, see de Haan and Diewert (2013) and Hill (2013).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

property prices. We note that the "computationally demanding and complex estimation procedures" nature of the Hill and Scholz methodology will deter national statistical agencies from adopting their approach.

Section 4 describes our data on sales of residential properties in Tokyo over the 44 quarters starting in the first quarter of 2000 and ending in the last quarter of 2010. We used the same data as we used in Diewert and Shimizu (2015a) on sales of residential houses in Tokyo, except the present study added additional data on sales of residential properties with no structures on the land plot.

Section 5 sets out the *builder's model* approach to hedonic property price regressions. This approach uses the property's sale price as the dependent variable and splits up property value into the sum of the land and structure components. This additive decomposition approach has a long history in the property hedonic regression literature but what is relatively recent is the use of an exogenous construction cost price series to value the structure component of the decomposition. It is this use of an exogenous index that allows us to decompose property value into plausible land and structure components.[2] This section uses both the nonparametric spatial coordinate approach due to Colwell as well as the neighborhood approach to model the influence of location on land prices. We look at the resulting land price indexes as we increase the size of the grid, and we find that there is little change in these land price indexes over a reasonable range of alternative grid sizes. Section 6 adds more characteristics to the model and again looks at how the resulting land prices change as we add more characteristics. The details of the various models that used additional characteristics can be found in Appendix C.

Section 7 compares the overall property price indexes generated by the important models explained in the previous sections (instead of comparing just the land price components of residential property sales). For comparison purposes, we also compared our best "model results with a traditional" hedonic property price hedonic regression which regresses the logarithm of property price on a linear combination of the property characteristics and time dummy variables.[3] This traditional approach does not generate reasonable subindexes for land and structures, but it can generate reasonable results for an overall property price index.

Section 8 concludes. Appendix A contains the results of selected regression models as well as the data underlying the charts in the main text.

2. Bilinear Interpolation on the Unit Square

Our task in this section is to explain how a particular method of bilinear interpolation works for functions of two variables defined on the unit square. This method of interpolation is a basic building block that can be used to construct a method for approximating a function of two variables that is defined over the unit square. Suppose that $f(x, y)$ is a continuous function of two variables, x and y,

[2]The basic idea of using an exogenous cost index can be found in Diewert (2010, pp. 33–35). See also Diewert *et al*. (2015).

[3]This traditional hedonic regression approach can be traced back to Court (1939).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

where $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Suppose that f takes on the values γ_{ij} at the corners of the unit square; that is, we have:

$$(1) \qquad \gamma_{00} \equiv f(0,0); \gamma_{10} \equiv f(1,0); \gamma_{01} \equiv f(0,1); \gamma_{11} \equiv f(1,1).$$

Assuming that we know (or can estimate) the heights of the function at the corners of the unit square, we look for an approximating continuous function that satisfies counterparts to equations (1) at the corners of the unit square and is a linear function along the four line segments that make up the boundary of the unit square. Colwell (1998, p. 89) showed that the following quadratic function of x and y, $g(x, y)$, satisfies these requirements:[4]

$$(2) \qquad g(x,y) \equiv \gamma_{00}(1-x)(1-y) + \gamma_{10}x(1-y) + \gamma_{01}(1-x)y + \gamma_{11}xy.$$

Colwell (1998, p. 89) also showed that $g(x, y)$ is a weighted average of $\gamma_{00}, \gamma_{10}, \gamma_{01}$, and γ_{11} for (x, y) belonging to the unit square.[5] To gain more insights into the properties of $g(x, y)$, rewrite $g(x, y)$ as follows:

$$(3) \qquad g(x,y) = \gamma_{00} + (\gamma_{10} - \gamma_{00})x + (\gamma_{01} - \gamma_{00})y + [(\gamma_{00} + \gamma_{11}) - (\gamma_{01} + \gamma_{10})]xy.$$

Thus if $\gamma_{00} + \gamma_{11} = \gamma_{01} + \gamma_{10}$, then $g(x, y)$ is a linear function over the unit square. However, if $\gamma_{00} + \gamma_{11} \neq \gamma_{01} + \gamma_{10}$, then $g(x, y)$ is a saddle function; that is, the determinant of the matrix of second-order partial derivatives of $g(x, y)$, $\nabla^2 g(x,y)$, is equal to $-[(\gamma_{00} + \gamma_{11}) - (\gamma_{01} + \gamma_{10})]^2 < 0$, and hence $\nabla^2 g(x,y)$ has one positive and one negative eigenvalue.

In the following section, we will follow the example of Colwell (1998, p. 91) and show how the function $g(x, y)$ defined over the unit square can be extended to define a continuous function over a grid of squares.

3. Bilinear Spline Interpolation over a Grid

To explain how Colwell's method works over a grid of squares, we will explain his method for the case of a 3×3 grid of squares. The method will be applied to the variables X and Y that are defined over a rectangular region in X, Y space. We assume that X and Y satisfy the following restrictions:

$$(4) \qquad X_{\min} \leq X \leq X_{\max}; Y_{\min} \leq Y \leq Y_{\max},$$

where $X_{\min} < X_{\max}$ and $Y_{\min} < Y_{\max}$. We translate and scale X and Y so that the range of the transformed X and Y, x and y, lie in the interval joining 0 and 3; that is, define x and y as follows:

$$(5) \qquad x \equiv 3(X - X_{\min})/(X_{\max} - X_{\min}); \; y \equiv 3(Y - Y_{\min})/(Y_{\max} - Y_{\min}).$$

[4]The function $g(x, y)$ defined by (2) is a special case of the bilinear function defined in matrix algebra textbooks such as Mirsky (1955, p. 353). Poirier (1976, p. 61) also defined the counterpart to (2) that is defined over a rectangle. The bilinear interpolation method defined by (2) is widely used in the engineering and applied mathematics literature; see Wikipedia (2019).

[5]It is straightforward to show that the sum of the nonnegative weights $(1-x)(1-y), x(1-y), (1-x)y$, and xy is equal to 1. Thus $g(x, y)$ will satisfy the inequalities $\min\{\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11}\} \leq g(x,y) \leq \max\{\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11}\}$.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Define the following three *dummy variable* (or indicator) *functions* of x:

$$(6)\qquad \begin{aligned} &D_1(x) \equiv 1 \text{ if } 0 \le x < 1;\ D_1(x) \equiv 0 \text{ if } x \ge 1;\\ &D_2(x) \equiv 1 \text{ if } 1 \le x < 2;\ D_2(x) \equiv 0 \text{ if } x < 1 \text{ or } x \ge 2;\\ &D_3(x) \equiv 1 \text{ if } 2 \le x \le 3;\ D_3(x) \equiv 0 \text{ if } x < 2. \end{aligned}$$

Note that if $0 \le x \le 3$, then $D_1(x) + D_2(x) + D_3(x) = 1$, so that the three dummy variable functions sum to 1 if x lies in the interval between 0 and 3.

The above definitions can be used to define the three *dummy variable functions* of y, $D_1(y)$, $D_2(y)$, and $D_3(y)$, where y replaces x in definitions (6). Finally, a set of $3 \times 3 = 9$ *bilateral dummy variable functions*, $D_{ij}(x, y)$, is defined as follows:

$$(7)\qquad D_{ij}(x,y) \equiv D_i(x)D_j(y);\quad i = 1,2,3;\ j = 1,2,3.$$

The domain of definition for the $D_{ij}(x, y)$ is the *square* S_3 in two-dimensional space with each side of length 3; that is, $S_3 \equiv \{(x, y): 0 \le x \le 3; 0 \le y \le 3\}$. Note that for any (x, y) belonging to S_3, we have $\sum_{i=1}^{3} \sum_{j=1}^{3} D_{ij}(x, y) = 1$. Thus the bilateral dummy variable functions $D_{ij}(x, y)$ will allocate any $(x, y) \in S_3$ to one of the nine unit square cells that make up S_3. Denote the *cell* that corresponds to x and y such that $D_{ij}(x, y) = 1$ as C_{ij} for $i, j = 1, 2, 3$. Thus the three cells in the grid of nine cells that correspond to y values that satisfy $0 \le y < 1$ are C_{11}, C_{21}, and C_{31}. The three cells that correspond to y values such that $1 \le y < 2$ are C_{12}, C_{22}, and C_{32} and the three cells that correspond to y values such that $2 \le y \le 3$ are C_{13}, C_{23}, and C_{33}.

Let $f(x, y)$ be the function defined over S_3 that we wish to approximate. Define the *heights* γ_{ij} of the function $f(x, y)$ at the 16 vertices of the grid of unit area cells as follows:

$$(8)\qquad \gamma_{ij} \equiv f(i,j);\quad i = 0,1,2,3;\ j = 0,1,2,3.$$

Define the Colwell (1998, pp. 91–92) *bilinear spline interpolating approximation* $g_3(x, y)$ to $f(x, y)$ for any $(x, y) \in S_3$ as follows:

$$(9)\qquad \begin{aligned} g_3(x,y) \equiv\ & D_{11}(x,y)[\gamma_{00}(1-x)(1-y)+\gamma_{10}(x-0)(1-y)+\gamma_{01}(1-x)(y-0)+\gamma_{11}xy]\\ &+D_{21}(x,y)[\gamma_{10}(2-x)(1-y)+\gamma_{20}(x-1)(1-y)+\gamma_{11}(2-x)(y-0)+\gamma_{21}xy]\\ &+D_{31}(x,y)[\gamma_{20}(3-x)(1-y)+\gamma_{30}(x-2)(1-y)+\gamma_{21}(3-x)(y-0)+\gamma_{31}xy]\\ &+D_{12}(x,y)[\gamma_{01}(1-x)(2-y)+\gamma_{11}(x-0)(2-y)+\gamma_{02}(1-x)(y-1)+\gamma_{12}xy]\\ &+D_{22}(x,y)[\gamma_{11}(2-x)(2-y)+\gamma_{21}(x-1)(2-y)+\gamma_{12}(2-x)(y-1)+\gamma_{22}xy]\\ &+D_{32}(x,y)[\gamma_{21}(3-x)(2-y)+\gamma_{31}(x-2)(2-y)+\gamma_{22}(3-x)(y-1)+\gamma_{32}xy]\\ &+D_{13}(x,y)[\gamma_{02}(1-x)(3-y)+\gamma_{12}(x-0)(3-y)+\gamma_{03}(1-x)(y-2)+\gamma_{13}xy]\\ &+D_{23}(x,y)[\gamma_{12}(2-x)(3-y)+\gamma_{22}(x-1)(3-y)+\gamma_{13}(2-x)(y-2)+\gamma_{23}xy]\\ &+D_{33}(x,y)[\gamma_{22}(3-x)(3-y)+\gamma_{32}(x-2)(3-y)+\gamma_{23}(3-x)(y-2)+\gamma_{33}xy]. \end{aligned}$$

It can be verified that $g_3(x, y)$ is a continuous function of x and y over S_3, and $g_3(x, y)$ is equal to the underlying function $f(x, y)$ when (x, y) is a vertex point of the grid; that is, we have the following equalities for the 16 vertex points in S_3:

$$(10)\qquad g_3(i,j) = \gamma_{ij} \equiv f(i,j);\quad i = 0,1,2,3;\ j = 0,1,2,3.$$

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

For each square of unit area in the grid, it can be seen that $g_3(x, y)$ behaves like the bilinear interpolating function $g(x, y)$ that was defined by (2) in the previous section. Thus if (x, y) belongs to the cell C_{ij} where i and j are equal to 1, 2 or 3, then $g_3(x, y)$ is bounded from below by the minimum of the four vertex point values $\gamma_{i-1,j-1}, \gamma_{i,j-1}, \gamma_{i-1,j}, \gamma_{i,j}$ and bounded from above by the maximum of the four vertex point values $\gamma_{i-1,j-1}, \gamma_{i,j-1}, \gamma_{i-1,j}, \gamma_{i,j}$.

Following Colwell (1998, p. 89), if we set $y = j$ where $j = 0, 1, 2$ or 3, then the resulting function of x, $g_3(x,j)$, is a *linear spline function in x* between 0 and 3; that is, $g_3(x,j)$ is a continuous, piecewise linear function of x that has three (joined) linear segments that can change their slopes at the break points $x = 1$ and $x = 2$. Similarly, if we set $x = i$ where $i = 0, 1, 2$, or 3, then the resulting function of y, $g_3(i, y)$, is also a *linear spline function in y* between 0 and 3. Thus we can view $g_3(x, y)$ as an interpolating function that merges these linear spline functions in the x and y directions into a consistent continuous function of two variables, where the interpolating function is equal to the function of interest at the 16 vertex points of the grid.

Following Poirier (1976, pp. 11–12) and Colwell (1998), we can move from the interpolation model defined by (9) to an econometric estimation model. Thus suppose that we can observe x and y for N observations, say (x_n, y_n) for $n = 1, \ldots, N$. Suppose also that we can observe $f(x_n, y_n)$ for $n = 1, \ldots, N$. Finally, suppose that we can approximate the function $f(x, y)$ by $g_3(x, y)$ over S_3. Let $\boldsymbol{\gamma} \equiv [\gamma_{00}, \gamma_{10}, \ldots, \gamma_{33}]$ be the vector of the 16 γ_{ij} which appear in (9) and rewrite $g_3(x, y)$ as $g_3(x, y, \boldsymbol{\gamma})$. Now view $\boldsymbol{\gamma}$ as a vector of parameters that appear in the following linear regression model:

$$z_n = g_3(x_n, y_n, \boldsymbol{\gamma}) + \varepsilon_n; \qquad n = 1, \ldots, N. \tag{11}$$

If we are willing to assume that the approximation errors ε_n are independently distributed with 0 means and constant variances, the unknown parameters γ_{ij} in (11) (which are the heights of the true" function $f(x, y)$ at the vertices in the grid) can be estimated by a least squares regression. It can be seen that this method for fitting a two-dimensional surface over a bounded set is essentially a nonparametric method. If the number of observations N is sufficiently large and the observations are more or less uniformly distributed over the grid, then we can make the grid finer and finer and obtain ever closer approximations to the true underlying function.[6]

To see how this nonparametric approach to the estimation of a surface could be applied in the context of sales of land plots in a geographical area, suppose that in a particular time period, we have information on the selling price of N land plots. Suppose that the selling price of land plot n is P_n and the area of the property is L_n square meters. Suppose also that we have data on the latitude and longitude of property n, X_n, and Y_n for $n = 1, \ldots, N$. Translate and scale these spatial coordinates into the variables x_n and y_n using definitions (4) and (5) above. We suppose that N

[6]If the dependent variable is observed with random errors, then the method for fitting the surface can also be regarded as a smoothing method. The smoothing parameter is the number of cells in the grid, k^2 (or k can be used as the smoothing parameter); the smaller the number of cells, the smoother will be the estimated $g_k(x, y)$ function. For a discussion of smoothing methods and alternative smoothing parameters, see Buja *et al.* (1989).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

is large enough and the observations are dispersed through all nine cells in the 3×3 geographical grid. An approximation to the *true land price surface* in the geographical area under consideration (which gives the price of land per meter squared as a function of the transformed spatial coordinates) can be generated by estimating the following linear regression model:

$$(12) \qquad P_n/L_n = g_3(x_n, y_n, \boldsymbol{\gamma}) + \varepsilon_n; \qquad n = 1, \ldots, N,$$

where the $g_3(x_n, y_n, \boldsymbol{\gamma})$ are defined by (9) for each (x_n, y_n) in the sample of observations. Thus estimates for the 16 unknown height parameters γ_{ij} in equations (12) can be obtained by solving a simple least squares minimization problem.

If observations are plentiful, then the grid can be made finer. Thus the 3×3 grid could be replaced by a $k \times k$ grid where k is an arbitrary positive integer. In this case, definitions (5) are replaced by $x \equiv k(X - X_{\min})/(X_{\max} - X_{\min})$ and $y \equiv k(Y - Y_{\min})/(Y_{\max} - Y_{\min})$. Definitions (6)–(9) can readily be modified to define the approximating function $g_k(x, y, \boldsymbol{\gamma})$ in place of $g_3(x, y, \boldsymbol{\gamma})$. Of course the new parameter vector $\boldsymbol{\gamma}$ in $g_k(x, y, \boldsymbol{\gamma})$ will have dimension $(k+1)^2$ in place of the parameter vector $\boldsymbol{\gamma}$ in $g_3(x, y, \boldsymbol{\gamma})$ which had dimension $4^2 = 16$. Thus Colwell (1998) realized that the well-known bilinear interpolation function $g(x, y)$ defined by (2) could be used as a basic building block in a powerful nonparametric method for approximating an arbitrary continuous function of two variables.[7]

However, Colwell did not exhibit the explicit representation for $g_3(x, y)$ defined by (9), so it is not clear exactly how he defined his linear regression model. Colwell (1998, p. 92) also made the following statement about his method of parameterization: As indicated earlier, one of the location variables must be omitted if perfect multicollinearity is to be avoided. Finally, it is not necessary to have data points within every section." Thus he seemed to suggest that one of the γ_{ij} on the right-hand side of (9) needed to be omitted to avoid perfect multicollinearity. However, such an omission would seem to destroy the flexibility of his method; that is, setting say $\gamma_{ij} = 0$ means that we would no longer have $g_k(i, j) = f(i, j)$. Moreover, as we shall see later in our empirical application of his method, problems can arise if some cells have no observations. Thus, although the spirit of his model is clear, the exact details on how to implement it are not spelled out in his paper.[8]

[7]Colwell (1998, p. 87) summarized his method as follows: A simple, non-parametric approach is needed—one that fits any function with the fewest possible restrictions. The purpose of this article is to describe a method for using a single, standard OLS regression to estimate a continuous price function in space that can approximate any shape. The cost of the method developed here is found in terms of degrees of freedom. It achieves flexibility by requiring large numbers of observations. Colwell (1998, p. 88) after noting that his approximating function was differentiable in the interior of each square in the grid but not necessarily at boundary points of each square offered the following view on the importance of continuity versus differentiability: This tradeoff of continuity for differentiability is worth accepting because continuity is compelling, whereas the worth of differentiability is dubious. Continuity of the price function is important because markets produce continuity. Discontinuities are destroyed by arbitrage. We agree with his assessment that differentiability of the approximating surface is not essential.

[8]Poirier (1976, pp. 59–62) developed an approach that is equivalent to our Colwell-based approach except that his parameterization of the approximating function is in terms of changes in levels rather than in the levels themselves. Thus his interaction terms are difficult to interpret. He also did not deal with the difficulties associated with empty cells.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

In Appendix B, we compare Colwell's method for estimating a surface with methods that are based on penalized least squares, which is the method used by Hill and Scholz (2018). Penalized least squares methods for fitting a surface have a computationally demanding and complex estimation procedures character to them, which means that national statistical agencies will not be able to use them as they are difficult to explain to the public. On the contrary, our adaptation of Colwell's method is very intuitive and can readily be explained to users.

4. The Tokyo Residential Property Sales Data

Our basic data set consists of quarterly data on V (the selling price of a residential property in Tokyo), L (the land area of the property in square meters), S (the floor space area of the structure if any on the land plot), A (the age in years of the structure if any on the land plot), the location of the property (specified in terms of longitude x and latitude y and in terms of the 23 Wards or local neighborhoods of Tokyo), and some additional characteristics to be explained below. These data were obtained from a weekly magazine, *Shukan Jutaku Joho* (Residential Information Weekly) published by Recruit Co., Ltd., one of the largest vendors of residential listings information in Japan. The Recruit data set covers the 23 special wards of Tokyo for the period 2000–2010, including the mini-bubble period in the middle of 2000s and its later collapse caused by the Great Recession. *Shukan Jutaku Joho* provides time series of housing prices from the week when it is first posted until the week it is removed due to its sale.[9] We only used the price in the final week because this can be safely regarded as sufficiently close to the contract price.[10]

After range deletions, there were a total of 5580 observations with structures on the property in our sample of sales of residential property sales in the Tokyo area over the 44 quarters covering 2000–2010.[11] In addition, we had 8493 observations on residential properties with no structure on the land plot.[12] Thus there were a total of 14,073 properties in our sample. The variables used in our regression analysis to follow and their units of measurement are as follows:

[9]There are two reasons for the listing of a unit being removed from the magazine: a successful deal or a withdrawal (i.e., the seller gives up looking for a buyer and thus withdraws the listing). We were allowed access to information regarding which the two reasons applied for individual cases, and we discarded those transactions where the seller withdrew the listing.

[10]Recruit Co., Ltd. provided us with information on contract prices for about 24 percent of all listings. Using this information, we were able to confirm that prices in the final week were almost always identical with the contract prices; see Shimizu *et al*. (2016).

[11]We deleted 9.2 percent of the observations with structures because they fell outside our range limits for the variables V, L, S, A, NB, and W. It is risky to estimate hedonic regression models over wide ranges when observations are sparse at the beginning and end of the range of each variable. The a priori range limits for these variables were as follows: $1.8 \le V \le 20$; $0.5 \le S \le 2.5$; $0.5 \le L \le 2.5$; $1 \le A \le 50$; $2 \le NB \le 8$; $25 \le W \le 90$. For properties with no structure, we set the corresponding S equal to 0.

[12]The large number of plots with no structures can be explained by the preference of Japanese buyers of residential properties to construct their own house. Thus sellers of residential properties that have a relatively old structure on the property tend to demolish the structure and sell the property as a land-only property.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

TABLE 1
Descriptive Statistics for the Variables

Name	No. of Obs.	Mean	Std. Dev	Min.	Max.
V	14,073	6.2491	2.9016	1.8	20
S	14,073	0.43464	0.5828	0	2.4789
L	14,073	1.0388	0.3986	0.5	2.4977
A	14,073	5.8231	9.117	0	49.723
NB	14,073	1.5669	2.0412	0	8
W	14,073	46.828	12.541	25	90
TW	14,073	9.3829	4.3155	1	29
TT	14,073	31.244	7.3882	8	48
X	14,073	139.67	0.0634	139.56	139.92
Y	14,073	35.678	0.0559	35.543	35.816
P_S	14,073	1.7733	0.0294	1.73	1.85

V = The value of the sale of the house in 10,000,000 Yen;
S = Structure area (floor space area) in units of 100 m^2;
L = Lot area in units of 100 m^2;
A = Approximate age of the structure in years;
NB = Number of bedrooms;
W = Width of the lot in 1/10 m;
TW = Walking time in minutes to the nearest subway station;
TT = Subway running time in minutes to the Tokyo station from the nearest station during the day (not early morning or night);
X = Longitude of the property;
Y = Latitude of the property;
P_S = Construction cost for a new structure in 100,000 Yen per m^2.

In addition, we have the address of each property and so we can allocate each property to one of the 23 Wards of Tokyo. This information was used to construct Ward dummy variables for each property in our sample. The basic descriptive statistics for the above variables are listed in Table 1.

Thus over the sample period, the sample average sale price was approximately 62.5 million Yen, the average structure space was 43.5 m^2 (but for properties with structures, the average was 110 m^2), the average lot size was 103.9 m^2, the average age of the structure was 5.8 years (for properties with a structure, the average age was 14.7 years), the average number of bedrooms in the properties that had structures was 3.95, the average lot width was 4.7 m, the average walking time to the nearest subway station was 9.4 min and the average subway traveling time from the nearest station to the Tokyo Central station was 31.2 min.

As is usual in property regressions using L and S as independent variables, we can expect multicollinearity problems in a simple linear regression of V on S and L.[13]

[13]See Diewert (2010) and Diewert *et al.* (2011, 2015) for evidence on this multicollinearity problem using Dutch data.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

To eliminate a possible multicollinearity problem between the lot size L and floor space area S for properties with a structure and to make our estimates of structure value consistent with structure value estimates in the Japanese national accounts, we will assume that the value of a new structure in any quarter is equal to a Residential Construction Cost Index per m^2 for Tokyo[14] (equal to P_{St} for quarter t) times the floor space area S of the structure.

5. The Basic Builder's Model Using Spatial Coordinates to Model Land Prices

The *builder's model* for valuing a residential property postulates that the value of a residential property is the sum of two components: the value of the land which the structure sits on and the value of the residential structure.

To justify the model, consider a property developer who builds a structure on a particular property. The total cost of the property after the structure is completed will be equal to the floor space area of the structure, say S square meters, times the building cost per square meter, β say, plus the cost of the land, which will be equal to the cost per square meter, α say, times the area of the land site, L. Now think of a sample of properties of the same general type, which have prices or values V_{tn} in period t[15] and structure areas S_{tn} and land areas L_{tn} for $n = 1, \ldots, N(t)$ where $N(t)$ is the number of observations in period t. Assume that these prices are equal to the sum of the land and structure costs plus error terms ε_{tn} which we assume are independently normally distributed with zero means and constant variances. This leads to the following *hedonic regression model* for period t where the α_t and β_t are the parameters to be estimated in the regression:[16]

$$V_{tn} = \alpha_t L_{tn} + \beta_t S_{tn} + \varepsilon_{tn}; \quad t = 1, \ldots, 44; \; n = 1, \ldots, N(t). \tag{13}$$

Note that the two characteristics in our simple model are the quantities of land L_{tn} and the quantities of structure floor space S_{tn} associated with property n in period t and the two *constant quality prices* in period t are the price of a square meter of land α_t and the price of a square meter of structure floor space β_t. Finally, note that separate linear regressions can be run of the form (13) for each period t in our sample.

The hedonic regression model defined by (13) applies to new structures. However, it is likely that a model that is similar to (13) applies to older structures as well. Older structures will be worth less than newer structures due to the depreciation of the structure. Assuming that we have information on the age of the

[14]This index was constructed by the Construction Price Research Association, which is now an independent agency but before 2012 was part of the Ministry of Land, Infrastructure, Transport, and Tourism (MLIT), a ministry of the Government of Japan. The quarterly values were constructed from the Monthly Residential Construction Cost index for Tokyo.

[15]The period index t runs from 1 to 44 where period 1 corresponds to Q1 of 2000 and period 44 corresponds to Q4 of 2010.

[16]Other papers that have suggested hedonic regression models that lead to additive decompositions of property values into land and structure components include Clapp (1980), Francke and Vos (2004), Gyourko and Saiz (2004), Bostic *et al.* (2007), Davis and Heathcote (2007), Francke (2008), Diewert (2008, 2010), Rambaldi *et al.* (2010), and Diewert *et al.* (2011) (2015).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

structure n at time t, say $A_{tn} = A(t, n)$ and assuming a geometric depreciation model, a more realistic hedonic regression model than that defined by (13) above is the following *basic builder's model*:[17]

$$(14) \qquad V_{tn} = \alpha_t L_{tn} + \beta_t (1-\delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}; \quad t = 1, \ldots, 44; \; n = 1, \ldots, N(t),$$

where the parameter δ reflects the *net depreciation rate* as the structure ages one additional period.[18] Note that (14) is now a nonlinear regression model whereas (13) was a simple linear regression model.

Note that the above model is a *supply side model* as opposed to the *demand side model* of Muth (1971) and McMillen (2003). Basically, we are assuming competitive suppliers of housing so that we are in Rosen's (1974, p. 44) Case (a), where the hedonic surface identifies the structure of supply. This assumption is justified for the case of newly built houses but it is less well justified for sales of existing homes.[19]

As was mentioned in the previous section, we have 14,073 observations on sales of houses in Tokyo over the 44 quarters in years 2000–2010. Thus equations (14) above could be combined into one big regression and a single depreciation rate δ could be estimated along with 44 land prices α_t and 44 new structure prices β_t, so that 89 parameters would have to be estimated. However, experience has shown that it is usually not possible to estimate sensible land and structure prices in a hedonic regression like that defined by (14) due to the multicollinearity between lot size and structure size.[20] Thus to deal with the multicollinearity problem, we draw on *exogenous information* on new house building costs from the Japanese MLIT. Thus if the sale of property n in period t has a new structure on it, we assume that the value of this new structure is equal to this measure of residential building costs p_{St} times the floor space area of the new structure, S_{tn}. We apply this same line of reasoning to property sales that have old structures on them as well. Thus our new builder's model replaces the parameter β_t which appears in equations (14) with the exogenous official price P_{St}. Our new model becomes the following:

$$(15) \qquad V_{tn} = \alpha_t L_{tn} + P_{St} (1-\delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}; \quad t = 1, \ldots, 44; \; n = 1, \ldots, N(t).$$

[17]This formulation follows that of Diewert (2008, 2010), Diewert *et al.* (2011, 2015), de Haan and Diewert (2013), and Diewert and Shimizu (2015a). It is a special case of Clapp's (1980, p. 258) hedonic regression model. For applications of the builder's model to condominium sales, see Diewert and Shimizu (2017a) and Burnett-Issacs *et al.* (2016).

[18]This estimate of depreciation is regarded as a *net depreciation rate* because it is equal to a true gross structure depreciation rate less an average renovations appreciation rate. As we do not have information on renovations and additions to a structure, our age variable will only pick up average gross depreciation less average real renovation expenditures. Note that we excluded sales of houses from our sample if the age of the structure exceeded 50 years when sold. Very old houses tend to have larger than normal renovation expenditures, and thus their inclusion can bias the estimates of the net depreciation rate for younger structures.

[19]Thorsnes (1997, p. 101) assumed that a related supply side model held instead of equation (14). He assumed that housing was produced by a CES production function $H(L, K) \equiv [\alpha L^\rho + \beta K^\rho]^{1/\rho}$ where K is structure quantity and $\rho \neq 0; \alpha > 0; \beta > 0$ and $\alpha + \beta = 1$. He assumed that property value V_n^t is equal to $p_t H(L_n^t, K_n^t)$, where p^t, ρ, α, and β are parameters to be estimated. However, our builder's model assumes that the production functions that produce structure space and that produce land are independent of each other.

[20]See Schwann (1998), Diewert (2010), and Diewert *et al.* (2011, 2015) on the multicollinearity problem.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Thus we have 14,073 degrees of freedom to estimate 44 land price parameters α_t and one annual geometric depreciation rate parameter δ, a total of 45 parameters. We estimated the nonlinear regression model defined by (15) for our Tokyo data set using the econometric programming package Shazam; see White (2004). The R^2 for the resulting preliminary nonlinear regression *Model 0* was only 0.5545 ,[21] which is not very satisfactory. However, there are no location variables in Model 0.

The value of a structure of the same type and age should not vary much from location to location. However, the price of land will definitely depend on the location of the property. Thus for our next model, we assume that the per meter price of land of a property is a function $f(x, y)$ of its spatial coordinates, x and y. Thus let x_{tn} and y_{tn} equal the normalized longitude and latitude of property n sold in period t. We will initially approximate the true land price surface $f(x, y)$ by the 4×4 Colwell spatial grid function $g_4(x, y)$ defined in Section 3. If X_{tn} and Y_{tn} are the raw longitude and latitude of property n sold in period t, then define the corresponding transformed spatial coordinates as $x_{tn} \equiv 4(X_{tn} - X_{\min})/(X_{\max} - X_{\min})$ and $y_{tn} \equiv 4(Y_{tn} - Y_{\min})/(Y_{\max} - Y_{\min})$ and define the Colwell approximation to $f(x_{tn}, y_{tn})$ as $g_4(x_{tn}, y_{tn})$ using the definitions in Section 3. *Model 1* is the following nonlinear regression model:

$$(16)\quad V_{tn} = \alpha_t g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma}) L_{tn} + P_{St}(1-\delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}; \quad t = 1, \ldots, 44;\ n = 1, \ldots, N(t).$$

Note that the $\boldsymbol{\gamma}$ vector of parameters in $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ consists of the 25 spatial grid parameters γ_{ij}, where $i, j = 0, 1, 2, 3, 4$. Thus equations (16) contain 44 unknown period t land price parameters α_t, 25 unknown γ_{ij} spatial grid parameters, and 1 depreciation rate parameter δ for a total of 70 unknown parameters. However, not all of these parameters can be estimated. If we multiply all components of $\boldsymbol{\gamma}$ by the positive number λ and divide all α_t by λ, it can be verified that the terms $\alpha_t g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma}) L_{tn}$ remain unchanged. Thus a normalization on the α_t and the γ_{ij} is required. We impose the normalization $\alpha_1 = 1$ which means that the sequence, $1, \alpha_2, \ldots, \alpha_{44}$, can be interpreted as an index of *residential land prices* for Tokyo for the 44 quarters in our sample, where the index is set equal to 1 in the first quarter of 2000.[22]

There are $4 \times 4 = 16$ cells C_{ij} in our grid of squares where C_{11} is the cell in the southwest corner of the grid, C_{41} is the southeast corner cell, C_{14} is the northwest corner cell, and C_{44} is the northeast corner cell. It turns out that cell C_{41} has no observed property sales over the entire sample period.[23] This means that γ_{44}, the value of land per meter squared at the southeast corner of the grid, cannot be

[21]All of the R^2 reported in this paper are equal to the square of the correlation coefficient between the dependent variable in the regression and the corresponding predicted variable. The estimated net annual geometric depreciation rate was $\delta = 10.49$ percent, with a T statistic of 23.3. This depreciation rate is too high to be believable. As we add more explanatory variables, we will obtain more reasonable depreciation rates.

[22]Note that the α_t shift the entire land price surface $g_4(x, y, \boldsymbol{\gamma})$ in a proportional manner over time. Thus all reasonable index numbers of the land price components of individual residential properties in Tokyo will be proportional to the estimated parameter sequence $1, \alpha_2^*, \ldots, \alpha_{44}^*$. This is perhaps a weakness of our model but given the nonparametric nature of our modeling of land prices, some simplifying assumptions had to be made to estimate all of the parameters in our model. In a real time setting, a rolling window approach would be used to implement our model, which would allow the height parameters to change over time; see Shimizu *et al*. (2010) for an example of this approach.

[23]This cell is defined as properties with normalized spatial coordinates (x, y) where x and y satisfy the restrictions $3 \leq x \leq 4$ and $0 \leq y \leq 1$.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

identified. Thus in addition to the normalization $\alpha = 1$, we set $\gamma_{44} = 0$ in equations (16). These normalizations will ensure that the nonlinear minimization problem associated with estimating Model 1 will have a unique solution. Thus Model 1 has 68 unknown parameters.

We used Shazam's nonlinear regression option to estimate the unknown parameters in (16). The R^2 for Model 1 turned out to be 0.7973, a huge jump from the R^2 for Model 0, which was only 0.5545. This large jump indicates the importance of including locational variables in a property regression. The log likelihood for Model 1 increased by 5524.50 points over the final log likelihood of Model 0 for adding 23 new location parameters. As Model 0 is a special case of Model 1, this is a highly significant increase in log likelihood. The estimated geometric depreciation rate from Model 1 was 6.33 percent per year (T statistic = 31.7), which is more reasonable than the Model 0 estimate of 10.49 percent.

We now address the problem of how exactly should the land, structure, and overall house price index be constructed? Our nonlinear regression model defined by (16) decomposes the period t value of property n into two terms: one which involves the land area L_{tn} of the property, $\alpha_t g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})L_{tn}$, and another term, $P_{St}(1-\delta)^{A(t,n)}S_{tn}$, which involves the structure area S_{tn} of the property. The first term can be regarded as an estimate of the land value of house n that was sold in quarter t, while the second term is an estimate of the structure value of the house (if $S_{tn} > 0$). Our problem now is how exactly should these two value terms be decomposed into *constant quality price and quantity components*? Our view is that a suitable constant quality land price index for all houses sold in period t should be α_t and for property n sold in period t, the corresponding constant quality quantity should be $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})L_{tn}$.[24] Turning to the decomposition of the structure value of property n sold in period t, $P_{St}(1-\delta)^{A(t,n)}S_{tn}$, into price and quantity components, we take P_{St} as the price and $(1-\delta)^{A(t,n)}S_{tn}$ as the corresponding quantity for property n sold in quarter t.

Note that the above model decompositions of individual property values into land and structure components sets the quality adjusted price of a square meter of land in quarter t equal to α_t^*, the estimated parameter value for α_t and sets the price of a square meter of structure equal to P_{St}, the official per meter structure cost for quarter t. These constant quality prices apply to all properties sold in period t and thus we can set the *aggregate* land and structure price for all residential properties sold in period t equal to P_{Lt} and P_{St} where $P_{Lt} \equiv \alpha_t^*$ for $t = 1, \ldots, 44$. The corresponding *aggregate* constant quality quantities of land and structures sold in period t are defined as follows:

$$(17) \quad Q_{Lt} \equiv \sum_{n=1}^{N(t)} g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma}^*)L_{tn}; \; Q_{St} \equiv \sum_{n=1}^{N(t)} (1-\delta^*)^{A(t,n)}S_{tn}; \quad t = 1, \ldots, 44,$$

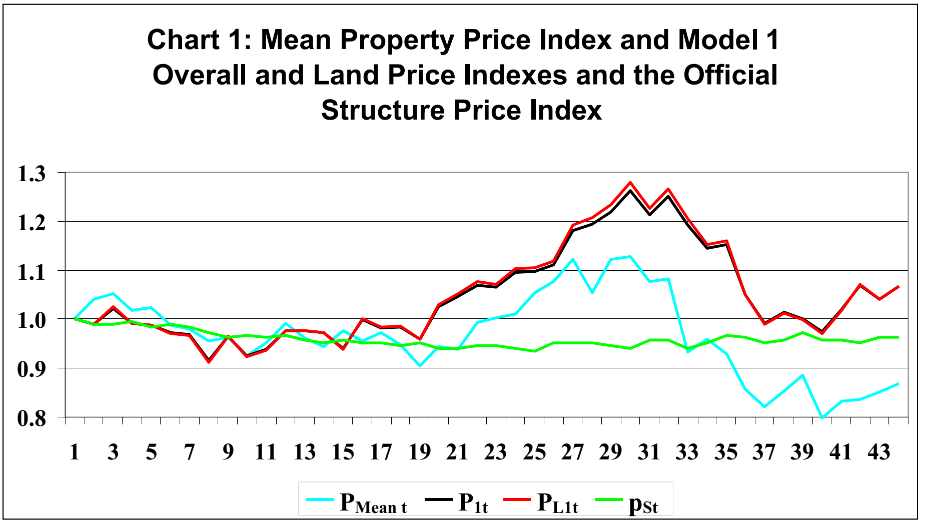
[24]An alternative way of viewing our land model is that land in each location indexed by the spatial coordinates x_n, y_n can be regarded as a distinct commodity with its own price and quantity. However, as our model forces all land prices in the same location to move proportionally over time, virtually all index number formulae will generate an overall land price series that is proportional to the α_t.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

where $\boldsymbol{\gamma}^* \equiv [\gamma^*_{00}, \ldots, \gamma^*_{44}]$ and δ^* are the estimated parameter values obtained by running the nonlinear regression model defined by (16).[25]

The price and quantity series for land and structures need to be aggregated into an overall Tokyo residential property sales price index. We use the Fisher (1922) ideal index to perform this aggregation. Thus define the *overall house price level for quarter t* for Model 1, P_t, as the chained Fisher price index of the land and structure series $\{P_{Lt}, P_{St}, Q_{Lt}, Q_{St}\}$. As these aggregate price and quantity series are generated by the Model 1 nonlinear regression model defined by equations (16), we relabel $Q_{Lt}, Q_{St}, P_t, P_{Lt}, P_{St}$, as $Q_{L1t}, Q_{S1t}, P_{1t}, P_{L1t}, P_{S1t}$ for $t = 1, \ldots, T$.[26]

The overall Model 1 house price index P_{1t} as well as the land and structure price indexes P_{L1t} and the normalized structure price index, $p_{St} \equiv P_{St}/P_{S1}$, for Tokyo over the 44 quarters in the years 2000–2010 are graphed in Chart 1.[27] We



[Correction added on 1st September 2022, after first online publication: Charts 1-4 have been included in this version.]

[25]We could use hedonic imputation or index number theory to form aggregate price and quantity indexes of land and structures but because our model makes the constant quality price of land and structures the same across all property sales in a quarter, our aggregation procedure can be viewed as an application of Hicks' Aggregation Theorem; that is, if the prices in a group of commodities vary in strict proportion over time, then the factor of proportionality can be taken as the price of the group, and the deflated group expenditures will obey the usual properties of a microeconomic commodity. Thus we have shown mathematically the very important principle, used extensively in the text, that if the prices of a group of goods change in the same proportion, that group of goods behaves just as if it were a single commodity. Hicks (1946, pp. 312–313).

[26]The Fisher chained index P_{1t} is defined as follows. For $t = 1$, define $P_{1t} \equiv 1$. For $t > 1$, define P_{1t} in terms of P_{1t-1} and P_{Ft} as $P_{1t} \equiv P_{1t-1}P_{Ft}$, where P_{Ft} is the quarter t Fisher chain link index. The chain link index for $t \geq 2$ is defined as $P_{Ft} \equiv [P_{LASt}P_{PAAt}]^{1/2}$, where the Laspeyres and Paasche chain link indexes are defined as $P_{LASt} \equiv [P_{L1t}Q_{L1t-1} + P_{S1t}Q_{S1t-1}]/[P_{L1t-1}Q_{L1t-1} + P_{S1t-1}Q_{S1t-1}]$ and $P_{PAAt} \equiv [P_{L1t}Q_{L1t} + P_{S1t}Q_{S1t}]/[P_{L1t-1}Q_{L1t} + P_{S1t-1}Q_{S1t}]$. Diewert (1976, 1992) showed that the Fisher formula had good justifications from both the perspectives of the economic and axiomatic approaches to index number theory.

[27]Define the normalized official structure price series as $p_{St} = P_{St}/P_{S1}$ for $t = 1, \ldots, 44$. This is the series that is plotted in Chart 1. It will not change as we introduce additional hedonic property regression models. We note that the official index $P_{St} = 18.5p_{St}$; that is, $P_{S1} = 18.5$.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

have also computed the quarterly mean selling price of properties traded in quarter t and then normalized this average property price series to start at 1 in Quarter 1 of 2000. This mean price series, $P_{\text{Mean } t}$, is also graphed in Chart 1.[28]

It can be seen that the official structure price series gradually trends downward over the sample period, which is not surprising as general deflation occurred in Japan during our sample period. Because there are so many land-only properties in our sample and because the value of structures is relatively small for properties that have structures on them, it can be seen that our estimated land price series, P_{L1t}, is relatively close to our Model 1 overall property price index, P_{1t}. It can also be seen that the average property price series, $P_{\text{Mean } t}$, has the same general shape as our overall property price index P_{1t}, but the average property price series lies well below our constant quality property price series by the end of the sample period. This is to be expected as the mean property price series does not consider depreciation of the structures for properties that have structures on them. However, the extent of the downward bias in the mean property price series by the end of the sample period is somewhat surprising.

Can we vary the number of cells in the spatial grid and explain more of the variation in residential property prices? We address this question in the next four hedonic regression models (Models 2–5) where we progressively increase the number of cells in the locational grid. Thus we will replace the land price approximating function $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ in (16) by $g_5(x_{tn}, y_{tn}, \boldsymbol{\gamma}), g_6(x_{tn}, y_{tn}, \boldsymbol{\gamma}), g_7(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ and $g_8(x_{tn}, y_{tn}, \boldsymbol{\gamma})$. The resulting Models 2–5 have 25, 36, 49, and 64 cells C_{ij} and 36, 49, 64, and 81 spatial land price height parameters γ_{ij}, respectively. Setting up the corresponding nonlinear regressions using (16) as a template is straightforward except that the existence of cells with no sample observations means that not all height parameters can be estimated.

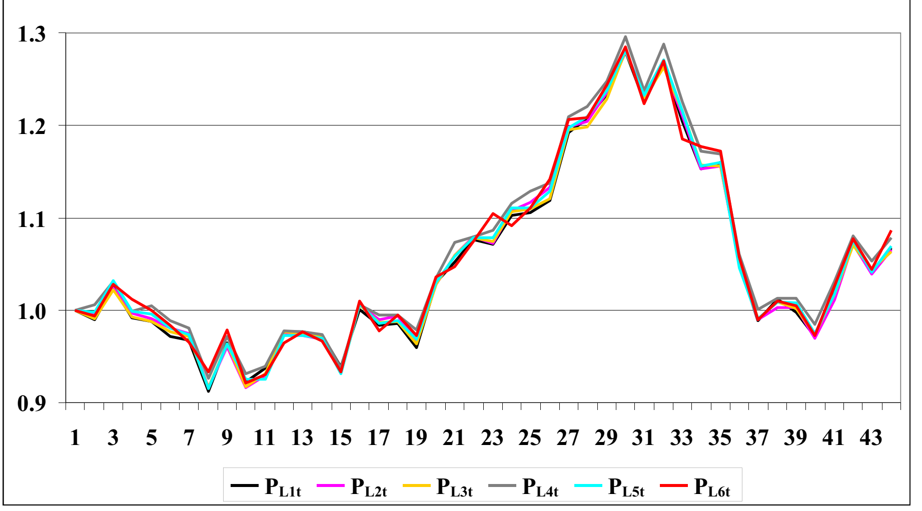
For *Model 2*, which used $g_5(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ in (16) in place of $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})$, the following cells in the 5×5 grid of cells had no sales over our sample period: C_{11}, C_{41}, C_{51}, and C_{42}. This means that three height parameters could not be estimated, and so we imposed the following restrictions on the parameters of Model 2: $\gamma_{00} = \gamma_{40} = \gamma_{50} = 0$. We also set $\alpha_1 = 1$ so that the remaining land price parameters α_t could be identified. Thus Model 2 had $36 - 3 = 33\gamma_{ij}$ parameters, 43 land price parameters α_t, and 1 depreciation rate parameter δ for a total of 77 parameters. The final log likelihood for Model 2 was 155.04 points higher than the final log likelihood for Model 1 for adding nine extra land price location parameters. The resulting R^2 was 0.8035 and the estimated geometric depreciation rate was $\delta^* = 6.29$ percent with a T statistic of 31.6. We expected that all of the estimated height parameters would be positive but two of them (γ_{51}^* and γ_{05}^*) turned out to be negative. However, the estimated land prices for each observation tn in our sample, $g_5(x_{tn}, y_{tn}, \boldsymbol{\gamma}^*)$, turned out to be positive for $t = 1, \ldots, 44$ and $n = 1, \ldots, N(t)$, and so we did not worry about these three negative γ_{ij}^* at this stage of our investigation.[29]

[28]The series $P_{\text{Mean}}, P_1, P_{L1}$, and p_S are also listed in Table A1 of Appendix A.

[29]The city of Tokyo is adjacent to the Pacific Ocean and so the boundaries of the city do not fit nicely into a rectangular grid (which we transformed into a square grid). Thus as the number of squares in the grid becomes larger, some squares at the boundaries of the grid will end up having no observations or very few observations. Thus suppose the observations in cell C_{11} are concentrated in the top north east corner of this cell. Then a better fit to the observed data in cell C_{11} may be obtained by setting γ_{00} equal to a negative number.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Chart 2: Land Prices for Models 1-6



The sequence of estimated α_t^* is our estimated land price series for Model 2, P_{L2t}, and this series is plotted in Chart 2 and is listed in Table A2 of Appendix A.

For *Model 3*, which used $g_6(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ in (16) in place of $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})$, the following five cells in the 6×6 grid of cells had no sales over our sample period: $C_{11}, C_{51}, C_{61}, C_{52}$, and C_{62}. Thus we set the following five height parameters equal to 0 to identify the remaining height parameters: $\gamma_{00} = \gamma_{50} = \gamma_{60} = \gamma_{51} = \gamma_{61} = 0$. We also set $\alpha_1 = 1$ so that the remaining land price parameters α_t could be identified. Thus Model 3 had $49 - 5 = 44\gamma_{ij}$ parameters, 43 land price parameters α_t, and 1 depreciation rate parameter δ for a total of 88 parameters. The final log likelihood for Model 3 was 82.43 points *lower* than the final log likelihood for Model 2 for adding 11 extra land price location parameters. Model 3 is not a special case of Model 2 so it can happen that moving to a larger number of squares in the grid does not improve the fit of the model. The problem is that there are likely to be discrete neighborhood land price effects and our relatively course partition of the city into squares does not adequately capture these discrete neighborhood effects. The resulting R^2 for Model 3 was 0.8014 (less than the Model 2 R^2 of 0.8035) and the estimated geometric depreciation rate was $\delta^* = 6.25$ percent with a T statistic of 31.8. There were five negative estimates for the land price height parameters: $\gamma_{01}^*, \gamma_{10}^*, \gamma_{41}^*, \gamma_{06}^*$, and γ_{56}^*. However, the estimated land prices $g_6(x_{tn}, y_{tn}, \boldsymbol{\gamma}^*)$ turned out to be positive for each observation in our sample. The sequence of estimated α_t^* is our estimated land price series for Model 3, P_{L3t}, and this series is plotted in Chart 2 and is listed in Table A2 of Appendix A.

Model 4 used $g_7(x_{tn}, y_{tn}, \boldsymbol{\gamma})$ in (16) in place of $g_4(x_{tn}, y_{tn}, \boldsymbol{\gamma})$. The following nine cells in the 7×7 grid of cells had no sales over our sample period: $C_{11}, C_{21}, C_{51}, C_{61}, C_{71}, C_{52}, C_{62}, C_{72}$, and C_{17}. Thus we set the following nine height parameters equal to 0 to identify the remaining height parameters: $\gamma_{00} = \gamma_{10} = \gamma_{50} = \gamma_{60} = \gamma_{70} = \gamma_{51} = \gamma_{61} = \gamma_{71} = \gamma_{07} = 0$. We also set $\alpha_1 = 1$ so that

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

the remaining land price parameters α_t could be identified. Thus Model 4 had $64 - 9 = 55\gamma_{ij}$ parameters, 43 land price parameters α_t, and 1 depreciation rate parameter δ for a total of 99 parameters. The final log likelihood for Model 4 was 501.88 points higher than the final log likelihood for Model 3 for adding 11 extra land price location parameters. The resulting R^2 for Model 4 was 0.8156 and the estimated geometric depreciation rate was $\delta^* = 5.99$ percent with a T statistic of 31.9. There were three negative estimates for the land price height parameters: $\gamma_{01}^*, \gamma_{67}^*$ and γ_{77}^*. As usual, the estimated land prices, $g_7(x_{tn}, y_{tn}, \gamma^*)$ for $t = 1, \ldots, T$ and $n = 1, \ldots, N(t)$, turned out to be positive for each observation in our sample. The sequence of estimated α_t^* is our estimated land price series for Model 4, P_{L4t}, and this series is plotted in Chart 2 and is listed in Table A2 of Appendix A.

Finally, *Model 5* used $g_8(x_{tn}, y_{tn}, \gamma)$ in (16) in place of $g_4(x_{tn}, y_{tn}, \gamma)$. The following 14 cells in the 8×8 grid of cells had no sales over our sample period: $C_{11}, C_{12}, C_{21}, C_{18}, C_{61}, C_{62}, C_{63}, C_{71}, C_{72}, C_{73}, C_{81}, C_{82}, C_{83}$, and C_{88}. All four corner cells were empty along with many other boundary cells. Thus we set the following 14 height parameters equal to 0 to identify the remaining height parameters: $\gamma_{00} = \gamma_{10} = \gamma_{01} = \gamma_{60} = \gamma_{61} = \gamma_{62} = \gamma_{70} = \gamma_{71} = \gamma_{72} = \gamma_{80} = \gamma_{81} = \gamma_{82} = \gamma_{88} = 0$. We also set $\alpha_1 = 1$ so that the remaining land price parameters α_t could be identified. Thus Model 5 had $91 - 14 = 77\gamma_{ij}$ parameters, 43 land price parameters α_t, and 1 depreciation rate parameter δ for a total of 111 parameters. The final log likelihood for Model 5 was 249.72 points *lower* than the final log likelihood for Model 4 for adding 12 extra land price location parameters. The resulting R^2 for Model 5 was 0.8086 (compared to 0.8156 for Model 4) and the estimated geometric depreciation rate was $\delta^* = 6.18$ percent with a T statistic of 31.5. There were five negative estimates for the land price height parameters: $\gamma_{02}^*, \gamma_{11}^*, \gamma_{50}^*, \gamma_{51}^*$ and γ_{52}^*. As usual, the estimated land prices, the $g_8(x_{tn}, y_{tn}, \gamma^*)$, turned out to be positive for each observation in our sample. The sequence of estimated α_t^* is our estimated land price series of index numbers for Model 5, P_{L5t}, and this series is plotted in Chart 2 and is listed in Table A2 of Appendix A.

At this point, we decided to stop the process of increasing the number of height parameters. It is clear that our best model up to this point was Model 4 (because the R^2 for Model 5 fell below the R^2 for Model 4). Thus increasing k does not necessarily improve the fit of the model.

One of the main purposes of this paper is to see if the use of spatial coordinates in a residential hedonic property value regression can lead to more accurate estimates for a property price index and for a land price subindex for residential properties than can be obtained using just postal codes or other neighborhood locational variables. Hill and Scholz (2018) made this comparison for residential property price indexes but not for the land price component of their overall property price index as their methodological approach did not allow for separate land and structure subindexes.

An alternative to using spatial coordinates to measure the influence of location on property prices is to use postal codes or neighborhoods as indicators of location. There are 23 Wards in Tokyo and each property in our sample belongs to one of these Wards. To consider the possible neighborhood effects on the price of land, we introduced *ward dummy variables*, $D_{W,tn,j}$, into the hedonic

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

regression (15). These 23 dummy variables are defined as follows: for $t = 1, \ldots, 44; n = 1, \ldots, N(t); j = 1, \ldots, 23$:[30]

$$(18) \quad D_{W,tn,j} \equiv \begin{cases} 1 \text{ if observation } n \text{ in period } t \text{ is in Ward } j \text{ of Tokyo;} \\ 0 \text{ if observation } n \text{ in period } t \text{ is not in Ward } j \text{ of Tokyo.} \end{cases}$$

We now modify the model defined by (15) to allow the *level* of land prices to differ across the 23 Wards of Tokyo. The new *Model 6* is defined by the following nonlinear regression model:

$$V_{tn} = \alpha_t\left(\sum\nolimits_{j=1}^{23} \omega_j D_{W,tn,j}\right)L_{tn} + P_{St}(1-\delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}; \quad t = 1, \ldots, 44; n = 1, \ldots, N(t).$$

(19)

Comparing the models defined by equations (15) and (19), it can be seen that we have added an additional 23 *ward relative land value parameters*, $\omega_1, \ldots, \omega_{23}$, to the model defined by (15). However, looking at (19), it can be seen that the 44 land price time parameters (the α_t) and the 23 ward parameters (the ω_j) cannot all be identified. Thus we need to impose at least one identifying normalization on these parameters. We chose the following normalization $\alpha_1 = 1$. Thus equations (19) contain 43 unknown period t land price parameters α_t, 23 Ward relative land price parameters, the ω_j, which replace the 25 unknown γ_{ij} spatial grid parameters in (16), and 1 depreciation rate parameter δ for a total of 67 unknown parameters. Thus this *Ward dummy variable hedonic regression* (Model 6) has roughly the same number of parameters as our spatial coordinate Model 1, which had 68 unknown parameters.

The final log likelihood for Model 6 was −24,318.67, a gain of 5045.90 over the final log likelihood of Model 0 defined by equations (15). The R^2 for Model 6 was 0.7853. The final log likelihood for Model 1 was −23,840.07 and the R^2 was 0.7993. Thus the spatial coordinates Model 1 fit the data better than the dummy variable Model 6. Both models had roughly the same number of parameters. An important question for our purposes is: how different are the resulting land price indexes generated by these two models? As usual, the sequence of estimated α_t^* is our estimated land price series for Model 6, P_{L6t}, and this series is plotted in Chart 2 and is listed in Table A2 of Appendix A.[31]

It can be seen that all six models produce much the same land price indexes.[32] Because our best fitting model was Model 4, P_{L4t} is our preferred land price series.

[30]The number of observations in each Ward in our sample was as follows: 3, 5, 195, 429, 348, 28, 62, 94, 453, 1260, 1114, 3434, 382, 701, 2121, 274, 107, 76, 432, 1679, 361, 212, 303. Thus Wards 1 and 2 had very few observations.

[31]Define the aggregate constant quality amounts of residential land and structures sold in period t by $Q_{Lt} \equiv \sum_{n=1}^{N(t)}\left(\sum_{j=1}^{23}\omega_j^* D_{W,tn,j}\right)L_{tn}$ and $Q_{St} \equiv \sum_{n=1}^{N(t)}(1-\delta^*)^{A(t,n)} S_{tn}$ for $t = 1, \ldots, 44$. The overall period t property price index for Model 6, P_{6t}, is defined as the chained Fisher price index using the above Q_{Lt} and Q_{St} as the period t quantity series and $P_{L6t} \equiv \alpha_t^*$ and the official structure prices P_{St} as the period t price series when constructing the Fisher index chain links.

[32]Because the structure component of overall property prices is relatively small compared to the land component and because the structure price index is the same across all six models, the overall property price indexes generated by Models 1–6 are all very similar.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Note that the Ward dummy variable model land price index, P_{L6t}, is fairly close to our preferred series.

6. The Builder's Model Using Additional Information

The above six models make use of information on land plot size, structure floor space, the age of the structure (if the property has a structure), and its location, either in terms of spatial coordinates or in terms of its neighborhood.[33] These are the most important residential property price determining characteristics in our view. In Appendix C, we made use of additional information on housing characteristics and we checked to see if this extra information materially changes our estimated land price indexes.[34] We used the spatial coordinate Model 4 as our starting point in these models, as it was the best fitting model studied in this section. This model used the Colwell nonparametric model for modeling the land price surface with the $7 \times 7 = 49$ cell grid.

A brief summary of the estimated land price series that resulted from the use of additional information on property characteristics follows. *Model 7* introduced dummy variables for the existence of a structure on the property. We found that Japanese purchasers of properties were willing to pay a premium of about 11 percent for a lot that had no structure on it. *Model 8* introduced piecewise linear splines on the lot size, while *Model 9* added splines on the structure size. *Model 10* introduced two subway access and travel time to the city core variables. *Model 11* added the number of bedrooms to the list of explanatory variables as an additional quality determining characteristic for the structure on the property. *Model 12* introduced lot width as an additional characteristic which affected the land value component of property value. Model 12 ended up with an R^2 of 0.8488, which means this spatial coordinates model, with all the extra bells and whistles imbedded in it, explained the data fairly well. This model generated three negative land price parameters. All nonparametric methods of surface fitting can have problems fitting the boundaries of the region where data are available and our suggested method also suffers from this boundary fitting problem.[35] The negative estimated parameters can be eliminated in our method by replacing them with squared parameters in our nonlinear regression. We did this replacement for *Model 13*, which is our final Colwell type regression. The R^2 for this model was the same as for Model 12 but there was a loss of 1.2 log likelihood points. The land price series that were generated by Models 12 and 13 were virtually identical; see Table A3 in Appendix A. The estimated annual geometric depreciation rate for Model 13 was $\delta^* = 0.0417$. *Model 14* is the same as Models 12 and 13 except that Ward dummy variable terms replaced the Colwell locational grid function, $g_7(x_{tn}, y_{tn}, \boldsymbol{\gamma})$, for each observation *tn*.

[33]We also require an exogenous building cost per square meter to get realistic land and structure subindexes.

[34]We were also interested in determining whether the extra information will change our estimates of structure depreciation rates.

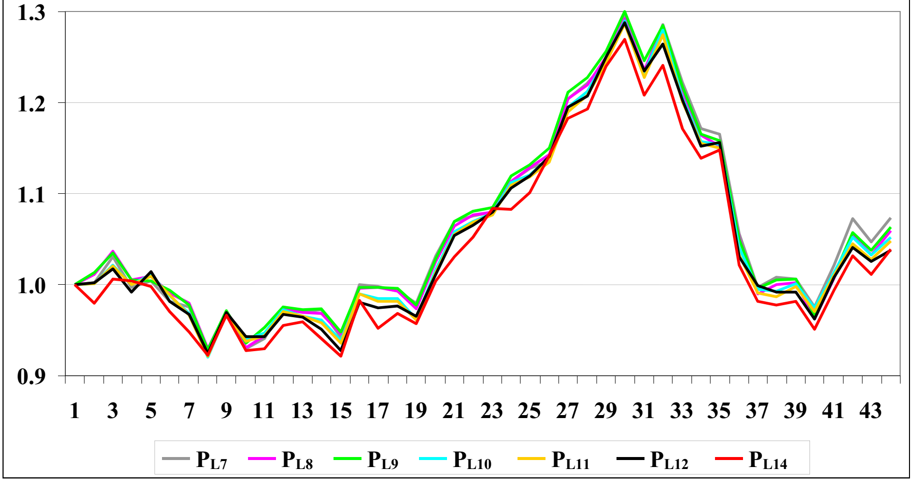
[35]At the boundaries of the data region, nonparametric methods will tend to fit the error terms.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

The final log likelihood for Model 14 was 478.6 log likelihood points higher than the final log likelihood of Model 6, which also used Ward dummy variables but did not use the extra characteristic information. However, the log likelihood of Model 14 was 827.0 points below the final log likelihood of Model 13, our best Colwell spatial coordinate model. The R^2 for Model 14 was 0.8300 which was below the R^2 for Model 13, which was 0.8488. The estimated parameters for Model 14 are listed in Table A6 in Appendix A. The sequence of land price indexes is the series of estimated coefficients, the α_t^*. This series is labeled as P_{L14t} and is listed in Table A3 of Appendix A.

Models 7–14 are explained in more detail in Appendix C. The land price indexes that these models generated are plotted in Chart 3.

Chart 3: Land Price Indexes for Models 7-12 and 14



Our conclusion at this point is that the neighborhood dummy variable models do not fit the data quite as well as a spatial coordinate model but the two types of model generate much the same land prices and hence overall residential property price indexes.[36] Looking at Chart 3, it can be seen that Model 14, the model that used Ward dummy variables to consider location effects on the price of land, produced the lowest measure of residential land price inflation in Tokyo. Our best spatial coordinate models, Models 12 and 13,[37] had the next lowest measure of land price inflation. The land price indexes generated by Models 7–11 are marginally above the Model 13 and 14 indexes.

In the following section, we compute the overall residential property price indexes that are generated by Models 1–14, and we compare the resulting indexes with a traditional log price time dummy property price index.

[36]Hill and Scholz (2018) came to the same conclusion for Sydney overall residential property price indexes.

[37]We did not plot the land price index for Model 13 as it could not be distinguished from the Model 12 index.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

7. Overall Residential Property Price Indexes

Models 1–14 in the previous two sections all have the same general structure in that property value is decomposed into the sum of land value plus structure value plus an error term. For example, using Model 6, the predicted value of property n in quarter t, V_{tn}, is equal to the predicted land value, $\alpha_t(\sum_{j=1}^{23}\omega_j^* D_{W,tn,j})L_{tn} \equiv V_{Ltn}$, plus predicted structure value, $P_{St}(1-\delta^*)^{A(t,n)}S_{tn} \equiv V_{Stn}$. Thus quarter t total predicted land value is $V_{Lt} \equiv \sum_{n=1}^{N(t)} V_{Ltn}$ and quarter t total predicted structure value is $V_{St} \equiv \sum_{n=1}^{N(t)} V_{Stn}$. The period t price of land for Models 1–14, P_{Lt}, is always α_t^* and the corresponding period t price of a structure is always $p_{St} \equiv P_{St}/P_{S1}$ for $t = 1, \ldots, 44$ where P_{St} is the official structure cost per m^2 of structure. For all models, define the corresponding period t aggregate quantity of land and structure as $Q_{Lt} \equiv V_{Lt}/P_{Lt}$ and $Q_{St} \equiv V_{St}/p_{St}$ for $t = 1, \ldots, 44$. Thus the basic price and quantity data for each model are $(P_{Lt}, p_{St}, Q_{Lt}, Q_{St})$ for $t = 1, \ldots, 44$. The overall property price indexes for Models 1–14 are calculated as Fisher (1922) chained indexes using the price and quantity data for land and structures that has just been defined. Label the resulting *overall property price indexes* for quarter t for Models 7–14 as $P_{7t}, P_{8t}, P_{9t}, P_{10t}, P_{11t}, P_{12t}, P_{13t}$, and P_{14t}. These series are listed in Table A7 in Appendix A. As was the case with the corresponding land price indexes, there these overall property price indexes approximate each other fairly closely.

There is one additional overall property price index that we calculate in this section, which is an index that is based on a "traditional" hedonic property price regression that uses the logarithm of price as the dependent variable and has time dummy variables.[38] Define the kth time dummy variable $D_{T,tn,k}$ for property n sold in period t as follows: for $t = 1, \ldots, 44;\ n = 1, \ldots, N(t);\ k = 2, 3, \ldots, 44$:

$$D_{T,tn,k} \equiv 1 \text{ if } t = k;\ D_{T,tn,k} \equiv 0 \text{ if } t \neq k. \tag{20}$$

Our best time dummy variable hedonic regression model[39] is the following *Model 15*:

$$\begin{aligned}\ln V_{tn} &= \sum_{k=2}^{44}\alpha_k D_{T,tn,k} + \sum_{j=1}^{23}\omega_j D_{W,tn,j} + \lambda \ln L_{tn} + \mu S_{tn} + \delta A_{tn} + \tau TW_{tn} + \rho TT_{tn} \\ &\quad + \sigma W_{tn} + \sum_{i=3}^{6}\kappa_i D_{NB,tn,i} + \varepsilon_{tn}; \qquad t = 1, \ldots, 44;\ n = 1, \ldots, N(t),\end{aligned} \tag{21}$$

where $\ln V_{tn}$ and $\ln L_{tn}$ denote the natural logarithms of property value V_{tn} and property lot size L_{tn} respectively, the $D_{T,tn,k}$ are time dummy variables, the $D_{W,tn,j}$ are Ward dummy variables, S_{tn} is the floor space area of the property (if there was no structure on the property n in period t, $S_{tn} \equiv 0$), TW_{tn}, and TT_{tn} are the subway time variables, W_{tn} is the lot width variable, A_{tn} is the age of the structure on property n sold in period t ($A_{tn} \equiv 0$ if the property had no structure), and the $D_{NB,tn,i}$ are

[38]This type of model does not generate reasonable separate land and structure subindexes; see Diewert *et al.* (2017, pp. 24–25) for an explanation of this assertion.

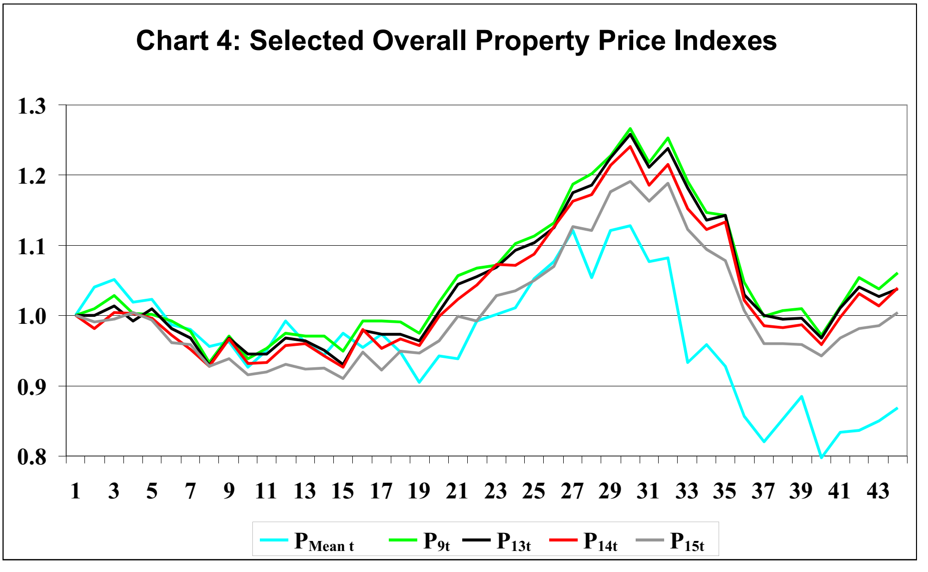
[39]We ran an initial linear regression using L_{tn} as an independent variable in place of $\ln L_{tn}$. However, this regression had a log likelihood that was 204.99 points lower than our final linear regression defined by (21). The R^2 for this preliminary regression was 0.8274. Note that we could not use $\ln S_{tn}$ as an independent variable because many observations had no structure on them and hence S_{tn} is equal to 0 for these properties, and thus we could not take the logarithm of 0.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

the bedroom dummy variables. The log likelihood of this model cannot be compared to the log likelihood of the previous models because the dependent variable is now the logarithm of the property price instead of the property price. There are 75 unknown parameters in the model defined by equations (21). The R^2 for Model 15 was 0.8323. Set $\alpha_1^* = 0$ and denote the estimated α_2 to α_{44} by $\alpha_2^*, \alpha_3^*, \ldots, \alpha_{44}^*$. The sequence of overall property price indexes P_{15t} generated by this model are the exponentials of the α_t^*; that is, define $P_{15t} \equiv \exp[\alpha_t^*]$ for $t = 1, \ldots, 44$. This series is listed in Table A7 of Appendix A.

Chart 4 compares several of the overall residential property prices that are defined above: the mean property price index $P_{\text{Mean } t}$ that appeared in Chart 1, P_{9t} (this is based on Model 9 which did not use information on the subway variables, the number of bedrooms, and the lot width variable), Model 13 (P_{13t}: our best Colwell spatial coordinates model), Model 14 (P_{14t}: our best Ward dummy variable model), and Model 15 (P_{15t}: our best traditional log price time dummy hedonic regression model that used all of our property price characteristics except the spatial coordinates).

Several points emerge from a study of Chart 4:



- The mean index, $P_{\text{Mean } t}$, has a large downward bias as compared to the other four indexes which is due to its neglect of structure depreciation. However, the movements in this index are similar to the movements in the other indexes.
- The property price index P_{15t} generated by a traditional log price time dummy hedonic regression model has a downward bias but it is not large.[40]

[40]Diewert (2010) also observed a similar result.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

- The Model 9 property price index, a Colwell spatial coordinates model that used only the four fundamental characteristics of a residential property (land plot area, structure floor space area, the age of the structure, and some locational variable)[41] generated an overall property price index P_{9t} that is quite close to our best Colwell spatial model, Model 14, which generated the overall property price index P_{14t}. Thus it is probably not necessary for national statistical agencies to collect a great deal of information on housing characteristics to produce a decent overall property price index (as well as decent land and structure subindexes).
- The Model 14 property price index, P_{14t}, that used local neighborhood information about properties instead of spatial coordinate information turned out to be fairly close to our best Colwell spatial index, P_{13t}. Thus following the advice of Hill and Scholz (2018), it is probably not necessary to use the spatial coordinate information to construct a satisfactory overall residential property price index.

8. Conclusion

Here are the main points that emerge from our paper:

- Satisfactory residential land price indexes and overall residential property price indexes can be constructed using local neighborhood dummy variables as explanatory variables in residential property regression models. It is not necessary to use spatial coordinates to model location effects on property prices.
- However, the use of spatial coordinates to model location effects does lead to better fitting regression models.
- The most important housing characteristics information that is needed to construct satisfactory residential land and overall property price indexes is information on lot size, floor space area of the property structure (if there is a structure on the property), the age of the structure, and some information on the location of the property. To obtain a satisfactory land price index, our method requires the use of exogenous information on residential construction costs.
- However, additional information on the characteristics of the property will improve the fit of our hedonic regressions, but the effects of the additional information on the resulting land and structure price indexes were minimal for our application to Tokyo residential property price indexes.
- Having land-only sales of residential properties should help improve the accuracy of the land price index that is generated by a property regression model. However, for our Japanese data, we found that the value of the land component of a land-only property earned a 10–15 percent premium over the land value of a neighboring property of the same size but with a

[41]In addition to these four fundamental variables, we need an exogenous building cost measure to implement our basic models.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

structure on the property. We attribute this premium to the costs of demolishing an older structure.

- Our models that used spatial coordinates to account for locational effects on the value of land used Colwell's nonparametric method for fitting a surface. This nonparametric method is much easier to implement than the penalized least squares approach used by Hill and Scholz (2018) to model locational effects on property prices. In Appendix B of the paper, we point out some of the theoretical advantages of Colwell's method.
- The potential bias in using property price indexes that are based on taking mean or median averages of property prices in a period can be very large. Typically, these methods will have a downward bias due to their neglect of structure depreciation.
- A traditional log price time dummy hedonic regression model that has structure age as an explanatory variable will typically reduce the bias that is inherent in an index based on taking averages of property prices. For our Tokyo data, we found that the traditional hedonic regression model led to an index which had a small downward bias (see Chart 4).

It should be noted that if a national statistical agency were to apply the regression models that were explained in this paper, they would not just run a regression using the entire sample data. A rolling window approach would be used: a window length of say 12–16 quarters would be chosen and as the data for each new quarter was processed, the movements in the index over the last two quarters in the sample would be used to update the last published index value; see Shimizu *et al*. (2010) for an application of this rolling window approach.

Our emphasis in this paper (and in other papers[42]) has been to develop reliable methods for the construction of the land component of residential property price indexes. This task is important for national statistical agencies because the balance sheet accounts in the System of National Accounts requires estimates for the price and volume of land used in production and consumption. In particular, this information is required to obtain more accurate estimates of national (and sectoral) Total Factor Productivity growth[43] but for the vast majority of countries, this information is simply not available. We hope that the methods explained in the present paper will be of use to national statistical agencies in developing improved estimates for the price and volume of land used in production and consumption. An additional benefit of our suggested method for obtaining land price indexes is that it also generates evidence-based estimates of structure depreciation rates.

Our suggested method for fitting a surface defined over a grid of squares can be extended in several ways:

- The grid of squares can be replaced by a grid of rectangles[44];

[42]See Diewert and Shimizu (2015a, 2015b, 2017a, 2017b, 2019) and Diewert *et al*. (2016).

[43]See Jorgenson and Griliches (1967, 1972) who developed the methodology used by national and international statistical agencies to measure TFP growth or Multifactor Productivity growth.

[44]This was already suggested by Poirier(1976).

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

- If data are sparse in a particular square, then a block of four adjacent squares can be combined into a larger single square such that the larger square has a sufficient number of observations to enable estimation; and
- The basic method can be extended to functions defined over three or more variables.

On the last point listed above, we note that the key to Colwell's method is defining a function that interpolates the data in a continuous manner over a unit square and is linear on the edges of the square. Recall that the Colwell interpolation function $g(x, y)$ defined over the unit square was defined as follows:

$$g(x,y) \equiv \gamma_{00}(1-x)(1-y)+\gamma_{10}x(1-y)+\gamma_{01}(1-x)y+\gamma_{11}xy. \quad (22)$$

For interpolation over a unit cube, the interpolation function is defined as follows:

$$\begin{aligned} g(x,y,z) &\equiv [\gamma_{000}(1-x)(1-y)+\gamma_{100}(x-0)(1-y)+\gamma_{010}(1-x)(y-0)+\gamma_{110}xy](1-z) \\ &\quad +[\gamma_{001}(1-x)(1-y)+\gamma_{101}(x-0)(1-y)+\gamma_{011}(1-x)(y-0)+\gamma_{111}xy]z \\ &= \gamma_{000}(1-x)(1-y)(1-z)+\gamma_{100}(x-0)(1-y)(1-z)+\gamma_{010}(1-x)(y-0)(1-z) \\ &\quad +\gamma_{110}xy(1-z)+\gamma_{001}(1-x)(1-y)(z-0)+\gamma_{101}(x-0)(1-y)(z-0) \\ &\quad +\gamma_{011}(1-x)(y-0)(z-0)+\gamma_{111}xyz. \end{aligned} \quad (23)$$

Thus we take $(1-z)$ times the x and y square for $z=0$ plus z times the x and y square for $z=1$ to obtain an overall weighted average of the heights at the vertices of the cube. There will be eight γ_{ijk} height parameters to be estimated because there are eight vertices for a unit cube. The eight weights sum to 1; therefore, for all x, y, z belonging to the unit cube, we have:

$$\min\{\gamma_{ijk}: i=0,1; j=0,1; k=0,1\} \leq g(x,y,z) \leq \max\{\gamma_{ijk}: i=0,1; j=0,1; k=0,1\}. \quad (24)$$

Thus the interpolation method is likely to smooth the actual $f(x, y, z)$ to some extent. There will be $16 = 2^4$ parameters for the hypercube in four dimensions and so on. The algebra linking the cubes into a linear regression will take some patience but it can be done.

References

Bailey, M. J., R. F. Muth, and H. O. Nourse, "A Regression Method for Real Estate Price Construction," *Journal of the American Statistical Association*, 58, 933–42, 1963.

Bostic, R. W., S. D. Longhofer, and C. L. Readfearn, "Land Leverage: Decomposing Home Price Dynamics," *Real Estate Economics*, 35(2), 183–2008, 2007.

Buja, A., T. Hastie, and R. Tibshirani, "Linear Smoothers and Additive Models," *Annals of Statistics*, 18, 454–509, 1989.

Burnett-Issacs, K., N. Huang, and W. E. Diewert, "Developing Land and Structure Price Indexes for Ottawa Condominium Apartments." *Discussion Paper 16-09*. Vancouver School of Economics, University of British Columbia, Vancouver, B.C., Canada, 2016.

Clapp, J. M., "The Elasticity of Substitution for Land: The Effects of Measurement Errors," *Journal of Urban Economics*, 8, 255–63, 1980.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Colwell, P. F., "A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices," *Journal of Real Estate Finance and Economics*, 17(1), 87–97, 1998.
Court, A. T., "Hedonic Price Indexes with Automotive Examples." *The Dynamics of Automobile Demand*. General Motors Corporation, New York, 98–117, 1939.
Davis, M. A., and J. Heathcote, "The Price and Quantity of Residential Land in the United States," *Journal of Monetary Economics*, 54, 2595–620, 2007.
J. de Haan, and W. E. Diewert (eds), *Residential Property Price Indices Handbook*. Eurostat, Luxembourg, 2013.
Diewert, W. E., "Exact and Superlative Index Numbers," *Journal of Econometrics*, 4, 114–145, 1976.
______, "Fisher Ideal Output, Input and Productivity Indexes Revisited," *Journal of Productivity Analysis*, 3, 211–48, 1992.
______, "The Paris OECD-IMF Workshop on Real Estate Price Indexes: Conclusions and Further Directions." *Proceedings From the OECD Workshop on Productivity Measurement and Analysis*. OECD, Paris, 11–36, 2008.
______, "Alternative Approaches to Measuring House Price Inflation." *Discussion Paper 10-10*. 2010. Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1.
Diewert, W. E., J. de Haan, and R. Hendriks, "The Decomposition of a House Price Index into Land and Structures Components: A Hedonic Regression Approach," *The Valuation Journal*, 6(2011), 58–106, 2011.
______, "Hedonic Regressions and the Decomposition of a House Price index into Land and Structure Components," *Econometric Reviews*, 34(1–2), 106–26, 2015.
Diewert, W. E., N. Huang, and K. Burnett-Isaacs, "Alternative Approaches for Resale Housing Price Indexes." *Discussion Paper 17-05*. 2017. Vancouver School of Economics, The University of British Columbia, Vancouver, Canada, V6T 1L4.
Diewert, W. E., and C. Shimizu, "" Residential Property Price Indices for Tokyo," *Macroeconomic Dynamics*, 19, 1659–714, 2015a.
______, "A Conceptual Framework for Commercial Property Price Indexes," *Journal of Statistical Science and Application*, 3(9-10), 131–52, 2015b.
______, "Hedonic Regression Models for Tokyo Condominium Sales," *Regional Science and Urban Economics*, 60, 300–15, 2017a.
______, "Alternative Approaches to Commercial Property Price Indexes for Tokyo," *Review of Income and Wealth*, 63, 492–519, 2017b.
______, "Alternative Land Price Indexes for Commercial Properties in Tokyo," *Review of Income and Wealth*, 2019. forthcoming.
Fisher, I., *The Making of Index Numbers*. Houghton-Mifflin, Boston, 1922.
Francke, M. K., "The Hierarchical Trend Model." in Kauko, T., and M. Damato (eds), *Mass Appraisal Methods: An International Perspective for Property Valuers*. Wiley-Blackwell, Oxford, 164–80, 2008.
Francke, M. K., and G. A. Vos, "The Hierarchical Trend Model for Property Valuation and Local Price Indices," *Journal of Real Estate Finance and Economics*, 28(2–3), 179–208, 2004.
Gyourko, J., and A. Saiz, "Reinvestment in the Housing Stock: The Role of Construction Costs and the Supply Side," *Journal of Urban Economics*, 55, 238–56, 2004.
Hicks, J. R., *Value and Capital*. 2nd ed. Clarendon Press, Oxford, 1946.
Hill, R. J., "Hedonic Price Indexes for Housing: A Survey, Evaluation and Taxonomy"," *Journal of Economic Surveys*, 27, 879–914, 2013.
Hill, R. J., and M. Scholz, "Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines," *Review of Income and Wealth*, 64(4), 737–56, 2018.
Hodrick, R. J., and E. C. Prescott, "Post-War U.S. Business Cycles: An Empirical Investigation." *Discussion Paper 451*. Center for Mathematical Studies in Economics and Management Science, Northwestern University, 1980.
Jorgenson, D. W., and Z. Griliches, "The Explanation of Productivity Change," *Review of Economic Studies*, 34, 249–83, 1967.
______, "Issues of Growth Accounting: A Reply to Edward F. Denison," *Survey of Current Business*, 55(5), 1972. part II, 65–94.
McMillen, D. P., "The Return of Centralization to Chicago: Using Repeat Sales to Identify Changes in House Price Distance Gradients," *Regional Science and Urban Economics*, 33, 287–304, 2003.
Mirsky, L., *An Introduction to Linear Algebra*. Clarendon Press, Oxford, 1955.
Muth, R. F., "The Derived Demand for Urban Residential Land," *Urban Studies*, 8, 243–54, 1971.
Poirier, D. J., *The Econometrics of Structural Change*. North-Holland Publishing Company, Amsterdam, 1976.
Rambaldi, A. N., R. R. J. McAllister, K. Collins, and C. S. Fletcher, *Separating Land from Structure in Property Prices: A Case Study from Brisbane Australia*. 2010. School of Economics, The University of Queensland, St. Lucia, Queensland 4072, Australia.

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth

Rosen, S., "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34–55, 1974.
Schwann, G. M., "A Real Estate Price Index for Thin Markets," *Journal of Real Estate Finance and Economics*, 16(3), 269–87, 1998.
Shimizu, C., K. G. Nishimura, and T. Watanabe, "Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures," *Journal of Economics and Statistics*, 230(6), 792–813, 2010.
______, "House Prices at Different Stages of Buying/Selling Process"," *Regional Science and Urban Economics*, 59, 37–53, 2016.
Thorsnes, P., "Consistent Estimates of the Elasticity of Substitution between Land and Non-Land Inputs in the Production of Housing," *Journal of Urban Economics*, 42, 98–108, 1997.
White, K. J., *Shazam: User's Reference Manual, Version 10*. Northwest Econometrics Ltd, Vancouver, Canada, 2004.
Wikipedia, *Bilinear Interpolation*. 2019. downloaded December 31, 2019.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site:

© 2021 The Authors. *Review of Income and Wealth* published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth