

# **Residential Property Price Indexes: Spatial Coordinates versus Neighbourhood Dummy Variables**

**Erwin Diewert**

(University of British Columbia & UNSW)

**Chihiro Shimizu**

(The University of Tokyo & Nihon U)

Economic Measurement Group 2019,  
Centre for Applied Economic Research,  
School of Economics,  
University of New South Wales,  
Sydney, Australia,  
December 6, 2019.

# 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.
- **Repeat sales methodology: Bailey, Muth and Nourse (1963) .**
- **Hedonic regression model approach.**

# Standard Price Indices

- **Sample means or medians**

- Means are especially sensitive to **outliers**, but both are **sensitive to sample composition**

- **Hedonic price measurements**

- Attempt to estimate a “quality controlled” index
- ***Specification*** of the equation

- **Repeat Sales measurements**

- Both coefficients and explanatory variables must be constant over time.  
→ **Depreciation and renovation problems**
- Age adjusted Repeat Sales measurement (Wong, Chau, Karato and Shimizu(2018))
- Lose many observations – is the sample representative of the overall market?

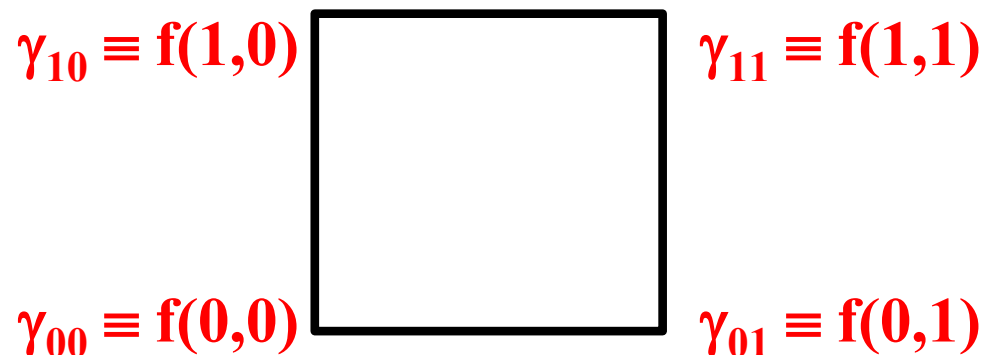
# The main question

- 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 neighbourhood 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.
- 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.

## 2. Bilinear Interpolation on the Unit Square

- Suppose that  $f(x,y)$  is a continuous function of two variables,  $x$  and  $y$ , where  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$ . Suppose that  $f$  takes on the values  $\gamma_{ij}$  at the corners of the unit square; i.e., we have:

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



- 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's Model (1989)

- Colwell (1998; 89) showed that the following **quadratic function** of  $x$  and  $y$ ,  $g(x,y)$ , satisfies these requirements:

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

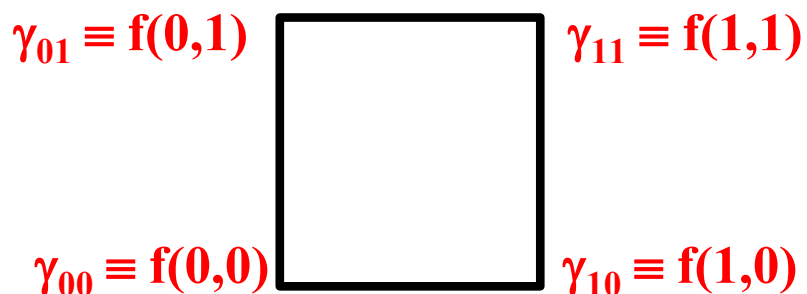
- Colwell (1998; 89) also showed that  $g(x,y)$  is a **weighted average of  $\gamma_{00}$ ,  $\gamma_{10}$ ,  $\gamma_{01}$  and  $\gamma_{11}$  for  $(x,y)$**  belonging to the unit square. In order to gain more insight into the properties of  $g(x,y)$ , rewrite  $g(x,y)$  as follows:

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

## Colwell's Model (1989)

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

- 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**; i.e., 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**.



### 3. Bilinear Spline Interpolation over a Grid

- In order to explain how Colwell's method works over a grid of squares, **we will explain his method for the case of a 3 by 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)  $X_{\min} \leq X \leq X_{\max} ; Y_{\min} \leq Y \leq Y_{\max}$

- 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**; i.e., define  $x$  and  $y$  as follows:

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



- Define the following 3 *dummy variable* (or **indicator**) *functions* of  $x$ :
  - (6)  $D_1(x) \equiv 1$  if  $0 \leq x < 1$ ;  $D_1(x) \equiv 0$  if  $x \geq 1$ ;  
 $D_2(x) \equiv 1$  if  $1 \leq x < 2$ ;  $D_2(x) \equiv 0$  if  $x < 1$  or  $x \geq 2$ ;  
 $D_3(x) \equiv 1$  if  $2 \leq x \leq 3$ ;  $D_3(x) \equiv 0$  if  $x < 2$ .
- Note that if  $0 \leq x \leq 3$ , then  $D_1(x) + D_2(x) + D_3(x) = 1$  so that the 3 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 3 *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)  $D_{ij}(x,y) \equiv D_i(x)D_j(y)$  ;  $i = 1,2,3$ ;  $j = 1,2,3$ .

- 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; i.e.,  

$$S_3 \equiv \{ (x,y) : 0 \leq x \leq 3; 0 \leq y \leq 3 \}.$$
- 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* of area 1 that corresponds to  $x$  and  $y$  such that  $D_{ij}(x,y) = 1$  as  $C_{ij}$  for  $i,j = 1,2,3$ . Thus the 3 cells in the grid of 9 cells that correspond to  $y$  values that satisfy  $0 \leq y < 1$  are  $C_{11}$ ,  $C_{21}$  and  $C_{31}$ . The 3 cells that correspond to  $y$  values such that  $1 \leq y < 2$  are  $C_{12}$ ,  $C_{22}$  and  $C_{32}$  and the 3 cells that correspond to  $y$  values such that  $2 \leq y \leq 3$  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*  $\gamma_{ij}$  of the function  $f(x,y)$  at the 16 vertices of the grid of unit area cells as follows:

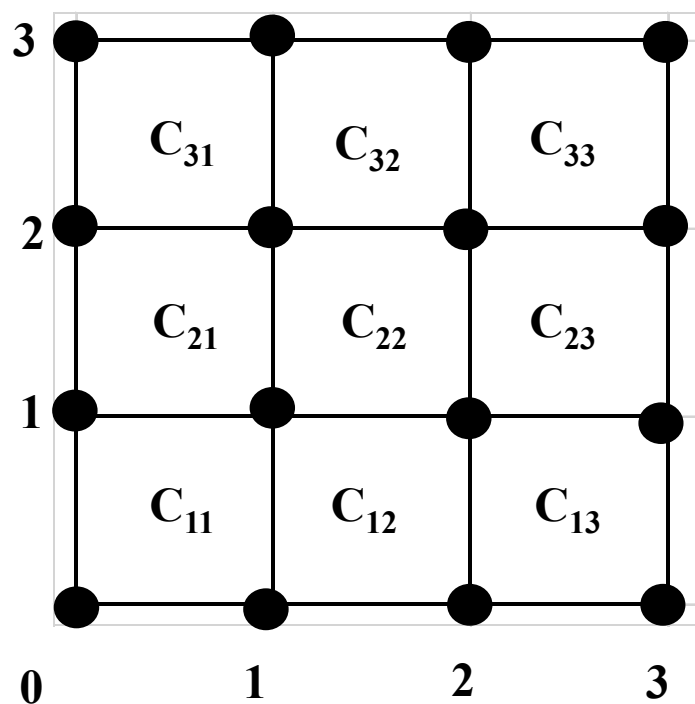
(8)  $\gamma_{ij} \equiv f(i,j)$  ;  $i = 0,1,2,3$ ;  $j = 0,1,2,3$ .

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

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

- 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; i.e., we have the following equalities for the 16 vertex points in  $S_3$ :

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



- 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 4 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 4 vertex point values  $\gamma_{i-1,j-1}$ ,  $\gamma_{i,j-1}$ ,  $\gamma_{i-1,j}$ ,  $\gamma_{i,j}$ .
- **Following Poirier (1976; 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  $n = 1, \dots, N$ . Suppose also that we can observe  $f(x_n, y_n)$  for  $n = 1, \dots, N$ . Finally, suppose that we can approximate the function  $f(x, y)$  by  $g_3(x, y)$  over  $S_3$ .
- Let  $\boldsymbol{\gamma} \equiv [\gamma_{00}, \gamma_{10}, \dots, \gamma_{33}]$  be the vector of the 16  $\gamma_{ij}$  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 which appear in the following linear regression model:

$$(11) \quad \mathbf{z}_n = \mathbf{g}_3(\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\gamma}) + \varepsilon_n ; \quad \mathbf{n} = 1, \dots, N.$$

- If we are willing to assume that the approximation errors  $\varepsilon_n$  are independently distributed with 0 means and constant variances, the unknown parameters  $\gamma_{ij}$  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 if it is continuous. [**Empty cells problem; boundary problems**]
- The **smoothing parameter** in this model is the size of the grid. It can be determined using leave one out cross validation.

## 4. Colwell's Nonparametric Method versus Penalized Least Squares (the method used by Hill and Scholz).

- Using the notation surrounding (11) above, a simplified version of this approach works as follows: find a function  $g(x,y)$  which is a solution to the following *penalized least squares minimization problem*:

$$(13) \min_g \sum_{n=1}^N [z_n - g(x_n, y_n)]^2 + \lambda \mathbf{J}(g)$$

- where it is assumed that  $g(x,y)$  is twice continuously differentiable and  $\mathbf{J}(g)$  is some function of the second order partial derivatives of  $g$  evaluated at the  $N$  observed  $(x_n, y_n)$ .
- It is difficult to explain how the penalized least squares approach works in the two dimensional case so we will explain how it works in the one dimensional case.



- The (unknown) continuous function of one variable is  $g(x)$ .
- We observe  $(z_n, x_n)$  for  $n = 1, \dots, N$ . **Suppose the  $x_n$  are equally spaced**. We assume  $z_n = g(x_n) + e_n$ , for  $n = 1, \dots, N$  where  $e_n$  is an error term with mean 0.
- Set  $s_n \equiv g(x_n)$  for  $n = 1, \dots, N$ .
- The one dimensional ***penalized least squares minimization problem*** is: choose  $s_1, s_2, \dots, s_N$  to solve the following unconstrained minimization problem:

$$(14) \min_{s_1, \dots, s_N} \{ \sum_{n=1}^N [z_n - s_n]^2 + \lambda \sum_{n=3}^N [\Delta^2 s_n]^2 \}$$

- where  $\lambda > 0$  is a positive tradeoff parameter and the ***first and second order finite differences*** of the  $s_n$  are defined as follows:

$$(15) \Delta s_n \equiv s_n - s_{n-1} ; \quad n = 2, 3, \dots, N;$$

$$(16) \Delta^2 s_n \equiv \Delta s_n - \Delta s_{n-1} ; \quad n = 3, 4, \dots, N.$$

- For a given  $\lambda$ , (14) can readily be solved using the first order conditions for the minimization problem and a bit of linear algebra. Denote the solution to (14) as the vector  $s(\lambda) \equiv [s_1(\lambda), \dots, s_N(\lambda)]$ . Denote the vector of observed  $z_n$  as  $z \equiv [z_1, \dots, z_N]$ . As  $\lambda$  tends to 0,  $s(\lambda)$  will tend to the observed vector  $z$ .
- As  $\lambda$  tends to plus infinity, the  $s_n(\lambda)$  will tend to a linear function of  $n$ ; i.e.,  $s_n(\lambda)$  will tend to  $\alpha + \beta n$  for  $n = 1, \dots, N$  for some  $\alpha$  and  $\beta$ . This nonparametric smoothing model was originally suggested by Henderson (1924; 30).
- Note that this smoothing method depends on the choice of  $\lambda$ . The method of cross validation can be used to choose  $\lambda$ ; see Silverman (1985; 5) for references to the literature.

- Our second highly simplified version or problem (13) is the following *penalized least squares minimization problem*: choose  $s_1, s_2, \dots, s_N$  to solve the following unconstrained minimization problem:

$$(17) \min_{s_1, \dots, s_N} \{ \sum_{n=1}^N [z_n - s_n]^2 + \lambda \sum_{n=3}^N [\Delta^3 s_n]^2 \}$$

- where  $\lambda > 0$  is again a positive tradeoff parameter between fit and the variability of the  $s_n$  and the third order finite differences of the  $s_n$ , are defined as follows:

$$(18) \Delta^3 s_n \equiv \Delta^2 s_n - \Delta^2 s_{n-1}; \quad n = 4, 5, \dots, N.$$

- Denote the solution to (17) as the vector  $s(\lambda) \equiv [s_1(\lambda), \dots, s_N(\lambda)]$ . As  $\lambda$  tends to 0,  $s(\lambda)$  will tend to the observed vector  $z$ . As  $\lambda$  tends to plus infinity, the  $s_n(\lambda)$  will tend to a quadratic function of  $n$ ; i.e.,  $s_n(\lambda)$  will tend to  $\alpha + \beta n + \gamma n^2$  for  $n = 1, \dots, N$  for some  $\alpha$  and  $\beta$ . This smoothing model was originally suggested by Whittaker (1923).

## Smoothing Technique

- A method for smoothing a discrete series  $z_n$  can be modeled as a mapping of the vector  $z \equiv [z_1, \dots, z_N]$  into a “smoothed” vector  $s \equiv [s_1, \dots, s_N]$ . Let  $F(z) \equiv [F_1(z), \dots, F_N(z)]$  be the vector valued smoothing function that transforms the “rough”  $z$  into the “smooth”  $s$  so that  $s \equiv F(z)$ . The function  $F(z)$  is a representation of the smoothing method.
- Diewert and Wales (2006; 107-110) developed a *test* or *axiomatic approach* to describe desirable properties of a smoothing method. We list their tests below along with two additional tests.

- ***Test 1; Sum Preserving Test.*** If  $s = F(z)$ , then  $1_N \cdot F(z) \equiv \sum_{n=1}^N F_n(z) = \sum_{n=1}^N z_n \equiv 1_N \cdot z$  where  $1_N$  is an  $N$  dimensional vector of ones. The test says that the sum of the values of the smoothed series should equal the sum of the values of the original series.
- ***Test 2; First Moment Preserving Test:*** If  $s = F(z)$ , then  $\sum_{n=1}^N ns_n = \sum_{n=1}^N nz_n$ . This test was suggested by Whittaker (1923; 68).
- ***Test 3; Identity Test:*** If  $z = k1_N$  where  $k$  is a scalar, then  $s = F(k1_N) = k1_N$ . Thus if the rough  $z$  is constant, then its smooth  $s$  reproduces this constant vector.

- ***Test 4; The Linear Trend Test:*** If  $z_n = \alpha + \beta n$  for  $n = 1, \dots, N$  where  $\alpha$  and  $\beta$  are constants, then  $F(z) = z$ .
- ***Test 5; The Quadratic Trend Test:*** If  $z_n = \alpha + \beta n + \gamma n^2$  for  $n = 1, \dots, N$  where  $\alpha$ ,  $\beta$  and  $\gamma$  are constants, then  $F(z) = z$ .
- ***Test 6; The Cubic Trend Test:*** If  $z_n = \alpha + \beta n + \gamma n^2 + \phi n^3$  for  $n = 1, \dots, N$  where  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\phi$  are constants, then  $F(z) = z$ .
- The last 3 tests were listed in Diewert and Wales (2006; 106). The following two tests were not listed in Diewert and Wales but they are obvious tests that are similar to Tests 4-6: if the rough is a smooth elementary function of one variable, then the smooth should be identical to the rough.

- ***Test 7; The Exponential Trend Test***: If  $z_n = \alpha e^n$  for  $n = 1, \dots, N$  where  $\alpha$  is a constant, then  $F(z) = z$ .
- ***Test 8; The Logarithmic Trend Test***: If  $z_n = \alpha \ln(n)$  for  $n = 1, \dots, N$  where  $\alpha$  is a constant, then  $F(z) = z$ .
- ***Test 9; The Diminishing Variation Test***:  $s = F(z)$  implies  $s \cdot s \leq z \cdot z$  or  $\sum_{n=1}^N s_n^2 \leq \sum_{n=1}^N z_n^2$ .
- If the smoothing method satisfies Tests 1 and 9, then the variance of the smooth cannot exceed the variance of the rough. Test 9 was proposed by Schoenberg (1946; 52).

- It can be seen that Tests 3-8 above are essentially due to Sprague. Diewert and Wales (2006; 109) argued that a consequence of the above quotation is the following test:
- ***Test 10; The Smoothing Invariance Test:***  $s = F(z)$  implies  $s = F(s)$  so that  $F[F(z)] = F(z)$ . Thus if we smooth the raw data  $z$  once and obtain the smooth  $s = F(z)$  and then if we smooth the resulting  $s$  and obtain  $F(s)$ , we find that the second round of smoothing just reproduces  $s$  so that  $F(s) = s$ . **Put another way, the smoothing method defined by the function  $F$  should produce a smooth series and so another round of smoothing should not change the smooth series produced by the initial use of  $F$ .**



- It can be shown that the Henderson smoothing method defined by the solution to (14) satisfies all of the above tests except Tests 6-8 and Test 10 and the Whittaker method defined by the solution to (17) satisfies all of the above tests except Tests 7, 8 and 10.
  - **The failure of a method to pass Test 10 is, in our view, is a serious problem with the method. Unfortunately, as noted by Diewert and Wales (2006; 109), most smoothing methods fail this test. For example, of the seven main types of nonparametric smoothing models listed by Buja, Hastie and Tibshirani (1989; 456-460): (i) running mean smoothers; (ii) bin smoothers; (iii) running line smoothers; (iv) polynomial regression; (v) cubic smoothing splines (the Henderson (1924) model); (vi) regression splines with fixed knots or break points; (vii) kernel smoothers; only methods (ii), (iv) and (vi) pass Test 10. The reason why these three smoothing methods pass Test 10 is that they are linear smoothers that are based on linear regression models.**

- Thus suppose that the rough  $z$  satisfies the linear regression model,  $z = X\beta + \varepsilon$  where  $X$  is an  $N$  by  $K$  matrix of exogenous variables of full rank  $K \leq N$  and  $\varepsilon$  is a vector of independently distributed error terms with means 0 and constant variances.
- Then the least squares estimator for  $\beta$  is  $b \equiv (X^T X)^{-1} X z$  and the predicted  $z$  vector is the smoothed vector  $s \equiv X b = X (X^T X)^{-1} X z = S z$  where  $S \equiv X (X^T X)^{-1} X$  is the *linear smoothing matrix* for this regression based smoothing method.
- Thus  $F(z) \equiv S z$  for this linear smoothing method. Any linear regression based smoothing method will satisfy Test 10 since  $F[F(z)] = S[Sz] = Sz = F(z)$  for this class of methods since  $SS = X(X^T X)^{-1} X X(X^T X)^{-1} X = X(X^T X)^{-1} X = S$ . **Methods (ii), (iv) and (vi) are all special cases of a linear regression based smoothing method.**
- Our Colwell based smoothing method is also a regression based smoothing method and so **it will satisfy the smoothing invariance test.**

## 5. The Tokyo Residential Property Sales Data

- 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. (Diewert and Shimizu (2015)).
- In addition, we had **8493 observations** on residential properties with ***no structure on the land plot***.
- Thus there was 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:
  - **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 squared;
  - **L** = **Lot area** in units of 100 meters squared;
  - **A** = Approximate **age** of the structure in years;
  - **NB** = **Number of bedrooms**;
  - **W** = **Width** of the lot in 1/10 meters;
  - **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; [Or we can use **Ward** or
  - **Y** = **Latitude** of the property; **Postal Code Dummy Variables**]
  - **PS** = **Construction cost** for a new structure in 100,000 Yen per meter squared.

**Table 1: Descriptive Statistics for the Variables.**

<b>Name</b>	<b>No. of Obs.</b>	<b>Mean</b>	<b>Std. Dev</b>	<b>Minimum</b>	<b>Maximum</b>
<b>V</b>	<b>14073</b>	<b>6.2491</b>	<b>2.9016</b>	<b>1.8</b>	<b>20</b>
<b>S</b>	<b>14073</b>	<b>0.43464</b>	<b>0.5828</b>	<b>0</b>	<b>2.4789</b>
<b>L</b>	<b>14073</b>	<b>1.0388</b>	<b>0.3986</b>	<b>0.5</b>	<b>2.4977</b>
<b>A</b>	<b>14073</b>	<b>5.8231</b>	<b>9.117</b>	<b>0</b>	<b>49.723</b>
<b>NB</b>	<b>14073</b>	<b>1.5669</b>	<b>2.0412</b>	<b>0</b>	<b>8</b>
<b>W</b>	<b>14073</b>	<b>46.828</b>	<b>12.541</b>	<b>25</b>	<b>90</b>
<b>TW</b>	<b>14073</b>	<b>9.3829</b>	<b>4.3155</b>	<b>1</b>	<b>29</b>
<b>TT</b>	<b>14073</b>	<b>31.244</b>	<b>7.3882</b>	<b>8</b>	<b>48</b>
<b>X</b>	<b>14073</b>	<b>139.67</b>	<b>0.0634</b>	<b>139.56</b>	<b>139.92</b>
<b>Y</b>	<b>14073</b>	<b>35.678</b>	<b>0.0559</b>	<b>35.543</b>	<b>35.816</b>
<b>P<sub>s</sub></b>	<b>14073</b>	<b>1.7733</b>	<b>0.0294</b>	<b>1.73</b>	<b>1.85</b>

## 6. 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 plus the value of the residential structure.
- This leads to the following *hedonic regression model* for period  $t$  where the  $\alpha_t$  and  $\beta_t$  are the parameters to be estimated in the regression:

$$(19) V_{tn} = \alpha_t L_{tn} + \beta_t S_{tn} + \varepsilon_{tn} ; t = 1, \dots, 44; n = 1, \dots, N(t).$$

- The hedonic regression model defined by (19) applies to **new structures**. But it is likely that a model that is similar to (19) 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 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 (19) above is the following ***basic builder's model***:

$$(20) \quad V_{tn} = \alpha_t L_{tn} + \beta_t (1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn} ;$$

$$t = 1, \dots, 44; n = 1, \dots, N(t)$$

where the parameter  $\delta$  reflects the ***net depreciation rate*** as the structure ages one additional period.

- Thus equations (20) above could be combined into one big regression and a single depreciation rate  $\delta$  could be estimated along with 44 land prices  $\alpha_t$  and 44 new structure prices  $\beta_t$  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 (20) due to the **multicollinearity** between lot size and structure size.
- Thus in order to deal with the multicollinearity problem, we draw on **exogenous information on new house building costs** from the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT).
- **(21)  $V_{tn} = \alpha_t L_{tn} + P_{St}(1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}$  ;  
 $t = 1, \dots, 44$ ;  $n = 1, \dots, N(t)$ .**



- Thus we have **14,073 degrees of freedom** to estimate 44 land price parameters  $\alpha_t$  and one annual geometric depreciation rate parameter  $\delta$ , a total of **45 parameters**.
- We estimated the nonlinear regression model defined by (21) 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, which is **not very satisfactory**. However, **there are no location variables in Model 0**.
- 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 by 4 Colwell spatial grid function**  $g_4(x,y)$  defined above in section 3.

## Model 1.

$$(22) \quad V_{tn} = \alpha_t \mathbf{g}_4(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma}) L_{tn} + P_{St}(1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn} ;$$

$$t = 1, \dots, 44; n = 1, \dots, N(t).$$

- Note that the  $\boldsymbol{\gamma}$  vector of parameters in  $\mathbf{g}_4(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$  consists of the **25 spatial grid parameters**  $\gamma_{ij}$  where  $i, j = 0, 1, 2, 3, 4$ .
- Thus equations (22) contain 44 unknown period  $t$  land price parameters  $\alpha_t$ , 25 unknown  $\gamma_{ij}$  spatial grid parameters and 1 depreciation rate parameter  $\delta$  for a total of **70 unknown parameters**.

- 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  $\alpha_t$  and for property  $n$  sold in period  $t$ , the corresponding constant quality quantity should be  $g_4(x_{tn}, y_{tn}, \gamma)L_{tn}$ . 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  $t$ .
- 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. But since 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  $\alpha_t$ .**

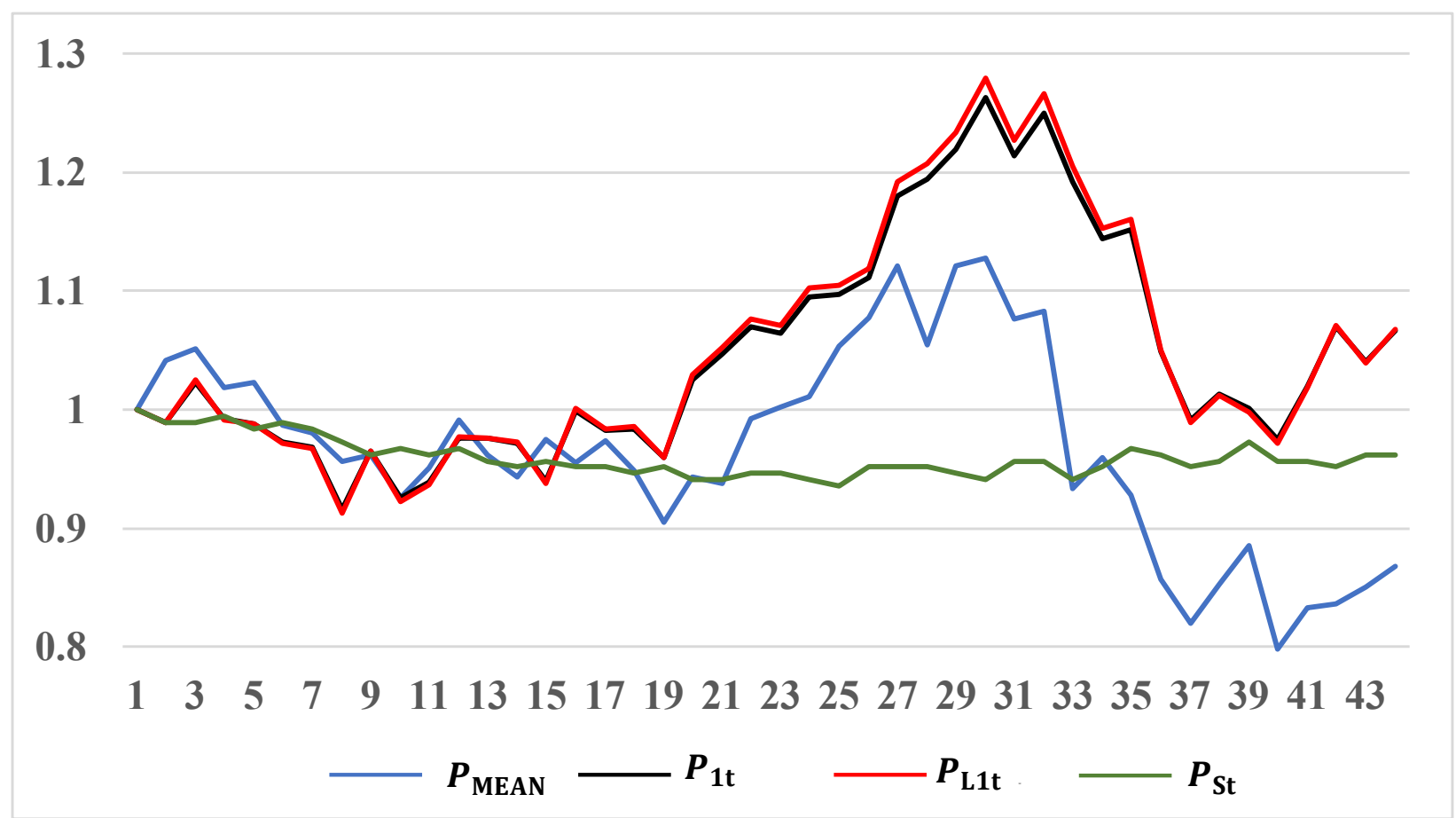
- Note that the above value decompositions of individual property prices sets the price of a square meter of land in quarter  $t$  equal to  $\alpha_t^*$ , the estimated parameter value for  $\alpha_t$  and sets the price of a square meter of structure equal to  $P_{St}$ , the official per meter structure cost for quarter  $t$ .
- These prices are assumed to be the same across 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  $t = 1, \dots, 44$ . The corresponding *aggregate constant quality quantities of land and structures* sold in period  $t$  are defined as follows:***

$$(23) \quad Q_{Lt} \equiv \sum_{n=1}^{N(t)} g_4(x_{tn}, y_{tn}, \gamma^*) L_{tn} ;$$

$$Q_{St} \equiv \sum_{n=1}^{N(t)} (1 - \delta^*)^{A(t,n)} S_{tn} ; \quad t = 1, \dots, 44.$$

- The prices  $\alpha_t^*$  and  $P_{St}$  and quantities  $Q_{Lt}$  and  $Q_{St}$  are used to form **chained Fisher overall property price indexes.**

# Chart 1 Mean Property Price Index and Model 1 Overall and Land Price Indexes and the Official Structure Price Index



## Model 2.

- For *Model 2*, which used  $\mathbf{g}_5(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$  in (22) in place of  $\mathbf{g}_4(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$ , the following cells in the **5 by 5** grid of cells had **no sales** over our sample period:  $C_{11}$ ,  $C_{41}$ ,  $C_{51}$  and  $C_{42}$ . This means that 3 height parameters could not be estimated 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  $\alpha_t$  could be identified. Thus Model 2 had  $36 - 3 = 33$   $\gamma_{ij}$  parameters, 43 land price parameters  $\alpha_t$  and 1 depreciation rate parameter  $\delta$  for a total of **77 parameters**.
- As the grid of squares becomes finer, some of the squares are over Tokyo Bay and so there are no sales for those squares. If these squares are not adjacent to a square which has sales, then the  $\gamma_{ij}$  parameters at the corners cannot be identified.

## Model 3.

- For **Model 3**, which used  $\mathbf{g}_6(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$  in (22) in place of  $\mathbf{g}_4(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$ , the following 5 cells in the **6 by 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 5 height parameters equal to 0 in order 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  $\alpha_t$  could be identified.
- Thus Model 3 had  $49 - 5 = 44$   $\gamma_{ij}$  parameters, **43** land price parameters  $\alpha_t$  and **1** depreciation rate parameter  $\delta$  for a total of **88 parameters**.

## Model 4.

- **Model 4** used  $\mathbf{g}_7(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$  in (22) in place of  $\mathbf{g}_4(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \boldsymbol{\gamma})$ . The following 9 cells in the **7 by 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 9 height parameters equal to 0 in order 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 the remaining land price parameters  $\alpha_t$  could be identified.
- Thus Model 4 had  $64 - 9 = \mathbf{55}$   $\gamma_{ij}$  parameters, **43** land price parameters  $\alpha_t$  and **1** depreciation rate parameter  $\delta$  for a total of **99 parameters**.



## Model 5.

- Finally, **Model 5** used  $g_8(x_{tn}, y_{tn}, \gamma)$  in (22) in place of  $g_4(x_{tn}, y_{tn}, \gamma)$ . The following 14 cells in the **8 by 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 4 corner cells were empty** along with many other boundary cells. Thus we set the following 14 height parameters equal to 0 in order 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  $\alpha_t$  could be identified.
- Thus Model 5 had  $91 - 14 = 77$   $\gamma_{ij}$  parameters, **43** land price parameters  $\alpha_t$  and **1** depreciation rate parameter  $\delta$  for a total of **111 parameters. We stopped adding cells at this point.**
- Note: **Model 5 did not fit as well as Model 4!**

## The Ward Dummy Model.

- An alternative to using spatial coordinates to measure the influence of location on property prices is to **use postal codes** or **neighbourhoods** as indicators of location.
- There are 23 Wards in Tokyo and each property in our sample belongs to one of these Wards. In order to take into account possible neighbourhood effects on the price of land, **we introduced *ward dummy variables*,  $D_{W,tn,j}$ , into the hedonic regression (20).**
- These 23 dummy variables are defined as follows: for  $t = 1, \dots, 44$ ;  $n = 1, \dots, N(t)$ ;  $j = 1, \dots, 23$ :

**(24)  $D_{W,tn,j} \equiv 1$  if observation  $n$  in period  $t$  is in Ward  $j$  of Tokyo;  
 $\equiv 0$  if observation  $n$  in period  $t$  is *not* in Ward  $j$  of Tokyo.**

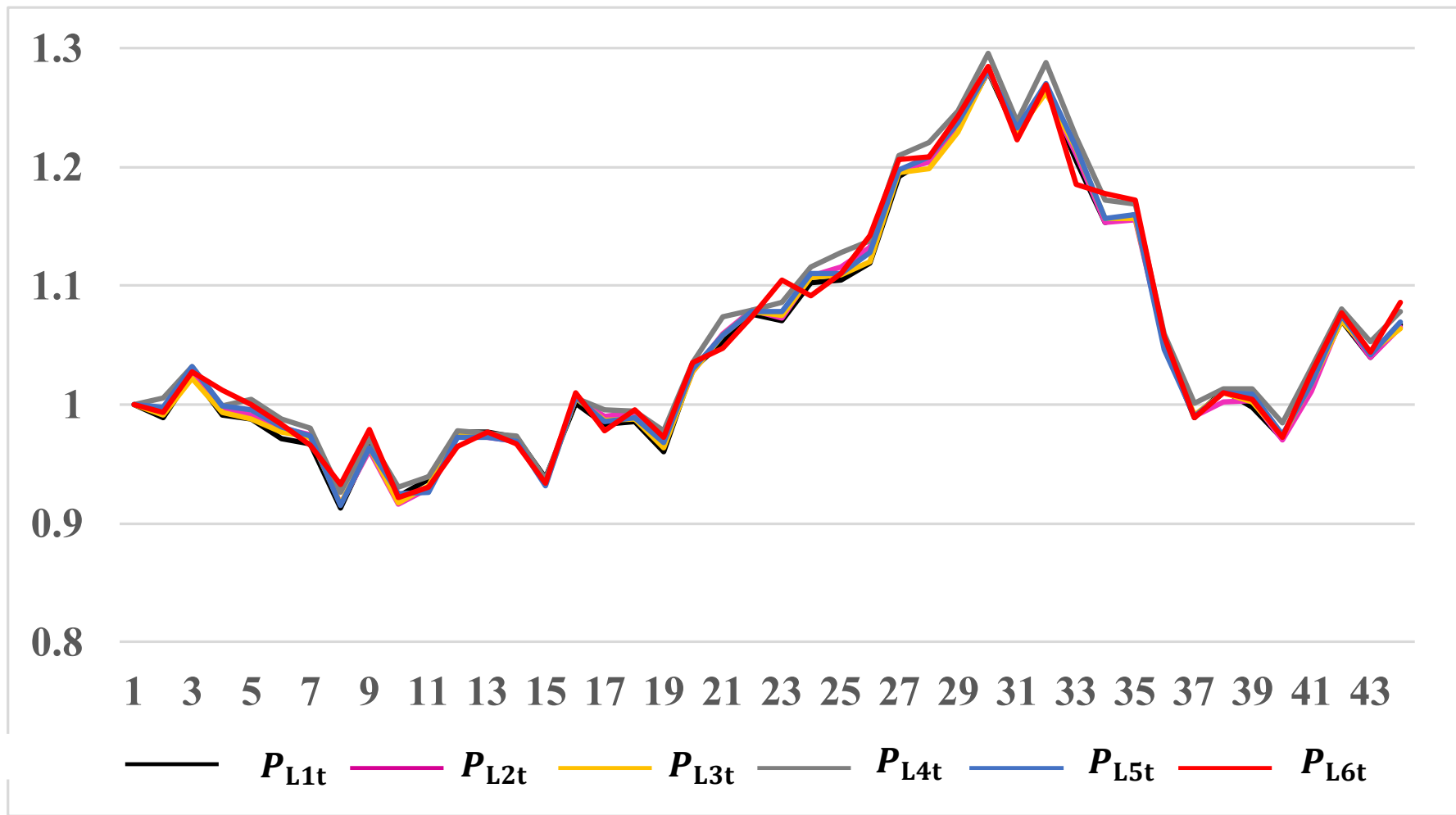
## Model 6.

- The new *Model 6* is defined by the following **nonlinear regression** model:

$$(25) V_{tn} = \alpha_t (\sum_{j=1}^{23} \omega_j D_{W,tn,j}) L_{tn} + P_{St} (1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn} ; \\ t = 1, \dots, 44; n = 1, \dots, N(t).$$

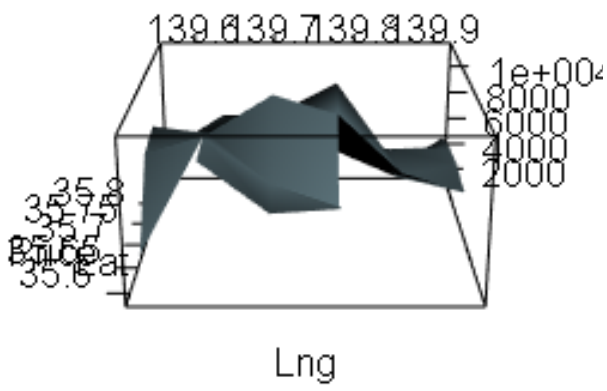
- Comparing the models defined by equations (20) and (25), it can be seen that we have added an additional **23 ward relative land value parameters,  $\omega_1, \dots, \omega_{23}$** , to the model defined by (20).
- However, looking at (25), it can be seen that the 44 land price time parameters (the  $\alpha_t$ ) and the 23 ward parameters (the  $\omega_j$ ) cannot all be identified. Thus we set  $\alpha_1$  equal to 1.
- We compare the land price series from the Ward Model 6 with the spatial Models 1-5 and find **practically no difference**.

## Chart 2 Land Prices for Models 1-6

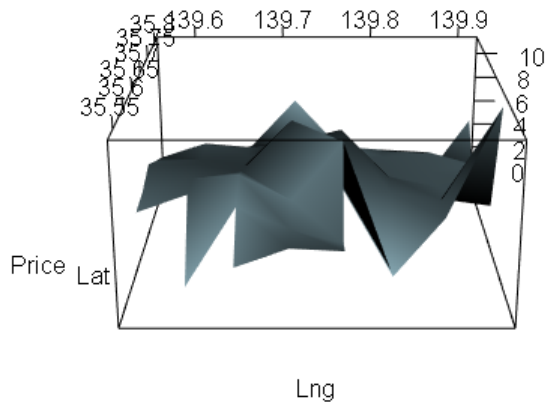


# Chart A-1. Land Price Distribution from Models 2, 4 and 5.

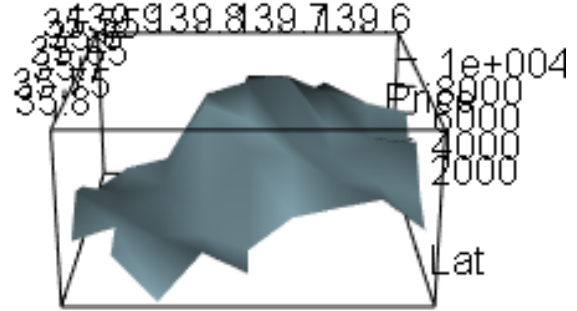
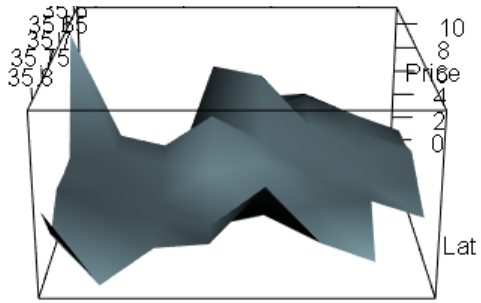
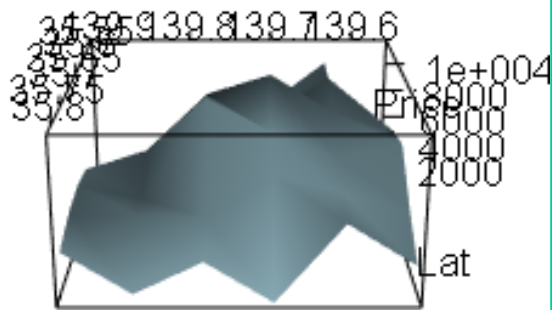
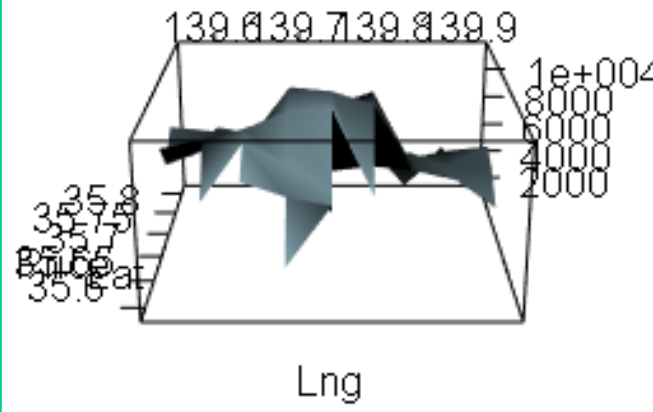
**5 × 5**



**7 × 7**

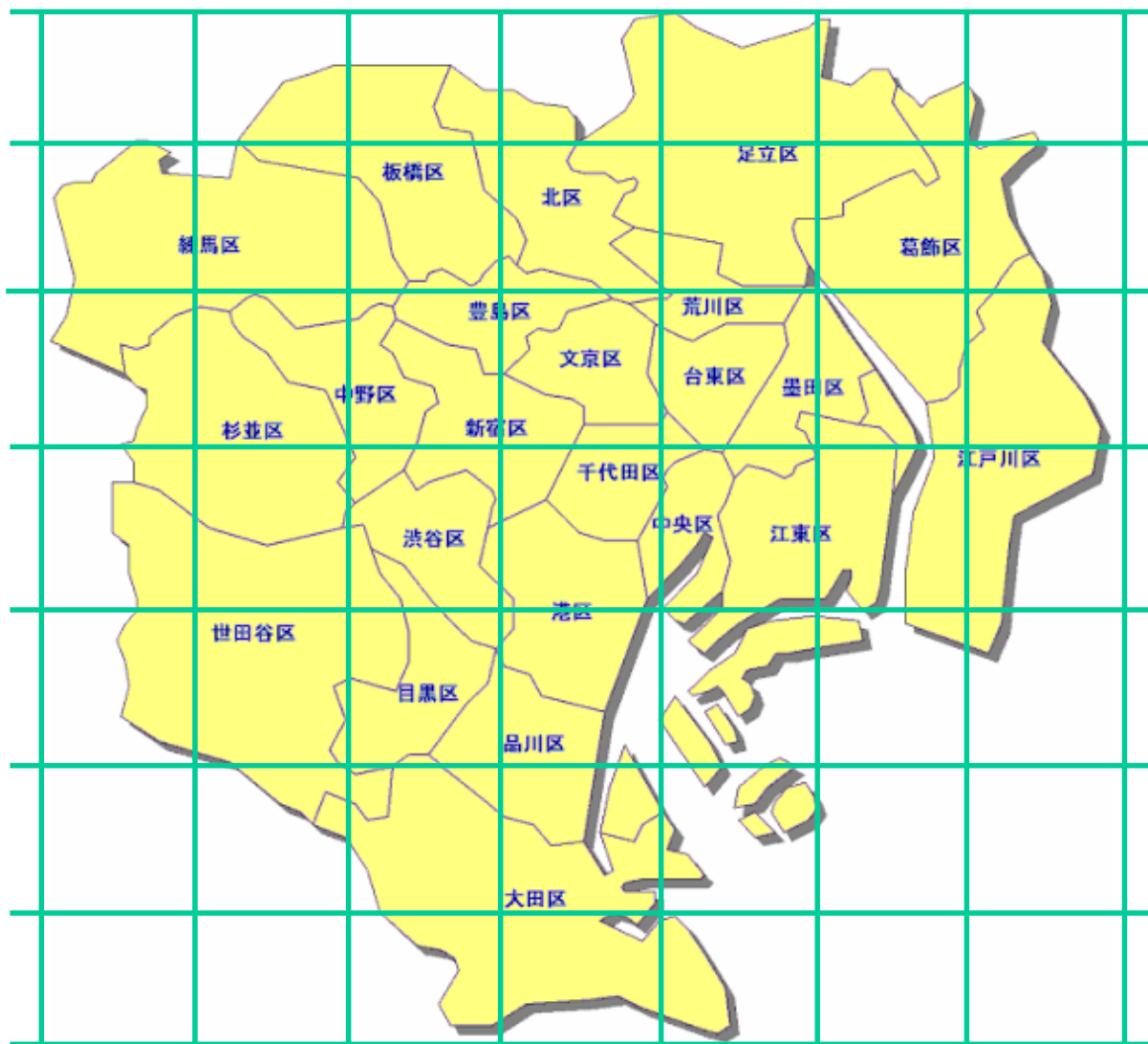


**8 × 8**



## Comparison in 6 Models.

- The 6 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 terms of its neighbourhood.
- **These are the most important residential property price determining characteristics in our view.** In the following section, we make use of additional information on housing characteristics and see if this extra information materially changes our estimated land price indexes.
- **We will use the spatial coordinate Model 4 as our starting point in the models which follow, since 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.**



## 7. Spatial Coordinate Models that Use Additional Information

- It is likely that property sales that have an older structure on the property will have a different land valuation than a nearby property of the same size that consists of cleared land, since demolition costs are not trivial.
- **Our *Model 7* takes this possibility into account.**
- Define the **land only dummy variable**  $D_{L,tn}$  as follows for  $t = 1, \dots, 44$  and  $n = 1, \dots, N(t)$ :  
**(26)  $D_{L,tn} \equiv 1$  if observation  $n$  in period  $t$  is a land only sale;  
 $\equiv 0$  otherwise.**



## Model 7.

- Define  $D_{S,tn} \equiv 1 - D_{L,tn}$  for  $t = 1, \dots, 44$ ;  $n = 1, \dots, N(t)$ . Thus if property  $n$  sold in period  $t$  has a structure on it,  $D_{S,tn}$  will equal 1. Model 7 estimates the following nonlinear regression:

$$(27) \quad V_{tn} = \alpha_t (\mathbf{D}_{S,tn} + \phi \mathbf{D}_{L,tn}) g_7(\mathbf{x}_{tn}, y_{tn}, \gamma) L_{tn} + P_{St} (1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn}; \quad t = 1, \dots, 44; n = 1, \dots, N(t).$$

- Thus **the parameter  $\phi$  gives the added premium to the property's land price (per meter squared) if the property has no structure on it.** The estimated  $\phi$  was  $\phi^* = 1.110$  ( $t = 153$ )
- We imposed the same restrictions on the  $\gamma_{ij}$  that were imposed in Model 4.
- The  $R^2$  for Model 7 was **0.8175** (the Model 4  $R^2$  was 0.8156).
- The final log likelihood for Model 7 was **128.75** points higher than the final log likelihood for Model 4 for adding one  $\phi$ .

## The size of the land plot:

- We group the observations into 4 groups, depending on the size of the land plot. The **cutoff sizes of land plot** are  $L_0$ ,  $L_1$ ,  $L_2$  and  $L_3$ .
- For each observation  $n$  in period  $t$ , we define the four **land dummy variables**,  $D_{L,tn,k}$ , for  $k = 1,2,3,4$  as follows:

(28)  $D_{L,tn,k} \equiv 1$  if observation  $tn$  has land area that belongs to group  $k$ ;

$\equiv 0$  if observation  $tn$  has land area that does not belong to group  $k$ .

- These dummy variables are used in the definition of the following **piecewise linear function** of  $L_{tn}$ ,  $f_L(L_{tn})$ , defined as follows:

$$(29) f_L(L_{tn}, \lambda) \equiv D_{L,tn,1} [\lambda_0 L_0 + \lambda_1 (L_{tn} - L_0)] + D_{L,tn,2} [\lambda_0 L_1 + \lambda_1 (L_1 - L_0) + \lambda_2 (L_{tn} - L_1)] + D_{L,tn,3} [\lambda_0 L_0 + \lambda_1 (L_1 - L_0) + \lambda_2 (L_2 - L_1) + \lambda_3 (L_{tn} - L_2)] + D_{L,tn,4} [\lambda_0 L_0 + \lambda_1 (L_1 - L_0) + \lambda_2 (L_2 - L_1) + \lambda_3 (L_3 - L_2) + \lambda_4 (L_{tn} - L_3)]$$

- where  $\lambda \equiv [\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4]$  and the  $\lambda_k$  are 5 unknown parameters and  $L_0 \equiv 0.5$ ,  $L_1 \equiv 1$ ,  $L_2 \equiv 1.5$  and  $L_3 \equiv 2$ .

## Model 8: Splines for the Land Plot Area

- Thus we are allowing the per meter squared price of land to vary as the size of the land plot increases. We expect the marginal price of land to decrease as lot size becomes very large. **Model 8** is the following nonlinear regression:

$$(30) \quad V_{tn} = \alpha_t (\mathbf{D}_{S,tn} + \phi \mathbf{D}_{L,tn}) \mathbf{g}_7(\mathbf{x}_{tn}, y_{tn}, \boldsymbol{\gamma}) \mathbf{f}_L(\mathbf{L}_{tn}, \boldsymbol{\lambda}) \\ + P_{St}(1 - \delta)^{A(t,n)} S_{tn} + \varepsilon_{tn} ; \quad t = 1, \dots, 44; \quad n = 1, \dots, N(t).$$

- where the function  $f_L$  is defined above by (29) and  $\varepsilon_{tn}$  is an error term. There are **43** unknown land price parameters  $\alpha_t$ , (we set  $\alpha_1 = 1$ ), **1** land only premium parameter  $\phi$ , **55** land price height parameters  $\gamma_{ij}$ , **4** marginal price of land parameters  $\lambda_k$  (we set  $\lambda_1 = 1$ ) and **1** depreciation rate  $\delta$  to estimate or **104 unknown parameters** in all.
- The  $R^2$  for Model 8 was **0.8222**, increase in LL was **328.27**.

## Model 9: Splines for the Structure Size.

- In our next model, we allow the per square meter price **of a square meter of structure to vary as the floor space of the structure increases**. The rationale for this model is that **bigger houses** are likely to be of **higher quality**.
- For each observation  $n$  in period  $t$ , we define the **3 structure dummy variables**,  $D_{S,tn,m}$ , for  $m = 1, 2, 3$  as follows:

(31)  $D_{S,tn,m} \equiv 1$  if observation  $tn$  has structure area that belongs to group  $m$ ;  
 $\equiv 0$  if observation  $tn$  has structure area that does not belong to group  $m$ .

- These dummy variables are used in the definition of the following piecewise linear function of  $S_{tn}$ ,  $f_S(S_{tn})$ , defined as follows:

$$(32) \mathbf{f}_S(\mathbf{S}_{tn}, \boldsymbol{\mu}) \equiv D_{S,tn,1} [\mu_0 S_0 + \mu_1 (S_{tn} - S_0)] + D_{S,tn,2} [\mu_0 S_1 + \mu_1 (S_1 - S_0) + \mu_2 (S_{tn} - S_1)] + D_{S,tn,3} [\mu_0 S_0 + \mu_1 (S_1 - S_0) + \mu_2 (S_2 - S_1) + \mu_3 (S_{tn} - S_2)].$$

## Model 9. Piecewise Linear Splines for Structure Size

- The **exogenous break points** are  $S_0 \equiv 0.5$ ,  $S_1 \equiv 1$  and  $S_2 \equiv 1.5$ .
- **Model 9** is the following nonlinear regression:

$$(33) \quad V_{tn} = \alpha_t (\mathbf{D}_{S,tn} + \phi \mathbf{D}_{L,tn}) g_7(x_{tn}, y_{tn}, \gamma) f_L(L_{tn}, \lambda) \\ + P_{St} (1 - \delta)^{A(t,n)} \mathbf{f}_S(\mathbf{S}_{tn}, \boldsymbol{\mu}) + \varepsilon_{tn}; \quad t = 1, \dots, 44; n = 1, \dots, N(t);$$

- where  $\boldsymbol{\mu} \equiv [\mu_0, \mu_1, \mu_2, \mu_3]$  and we set  $\mu_1 = 1$ .
- The function  $\mathbf{f}_L$  is defined above by (29), the function  $\mathbf{f}_S$  is defined by (32) and  $\varepsilon_{tn}$  is an error term. There are **43** unknown land price parameters  $\alpha_t$ , **1** land only premium parameter  $\phi$ , **55** land price height parameters  $\gamma_{ij}$ , **4** marginal price of land parameters  $\lambda_k$ , **3** marginal price of structure parameters  $\mu_m$  and **1** depreciation rate  $\delta$  to estimate or **107** unknown parameters to estimate.
- The  $R^2$  for Model 9 was **0.8256**, increase in LL was **136.32**.

## Adding the Subway Time Variables: TW and TT.

- Our next model, we make use of the **two subway variables**: **TW**, the walking time in minutes to the nearest subway station, and **TT**, the subway running time in minutes to the Tokyo central station.
- The sample minimum time for TW was 1 minute and the minimum time for TT was 8 minutes.
- Our next model allows the price of land to decrease as these two subway time variables increase.
- These variables have proven to be highly significant in other studies of Tokyo property prices.

## Model 10: Adding the Subway Time Variables

- Thus **Model 10** is the following nonlinear regression:

$$(34) V_{tn} = \alpha_t [D_{S,tn} + \phi D_{L,tn}] g_7(x_{tn}, y_{tn}, \gamma) f_L(L_{tn}, \lambda) \\ x[1 + \tau(TW_{tn} - 1)][1 + \rho(TT_{tn} - 8)] + P_{St}(1 - \delta)^{A(t,n)} f_S(S_{tn}, \mu) + \varepsilon_{tn} ; \\ t = 1, \dots, 44; n = 1, \dots, N(t)$$

- where the function  $f_L$  is defined above by (29), the function  $f_S$  is defined by (32),  $\tau$  is the percentage change in the price of land due to a one minute increase in walking time,  $\rho$  is the percentage change in the price of land due to a one minute increase in subway running time to Tokyo central station and  $\varepsilon_{tn}$  is an error term.
- There are **109** unknown parameters in Model 10.
- The  $R^2$  for Model 10 was **0.8383**, increase in LL was **531.13**.

## Adding the Number of Bedrooms

- In our next model, we introduce the **number of bedrooms**  $NB_{tn}$  as a property characteristic that can affect structure value if the property  $n$  in quarter  $t$  has a structure on it.
- For the properties in our sample, the number of bedrooms ranged from 2 to 8. Since there were relatively few observations with 6, 7 or 8 bedrooms, we grouped these last 3 categories into a single category.
- Define the **bedroom dummy variables**  $D_{NB,tn,i}$  for observation  $tn$  as follows for  $i = 2,3,4,5$ ;  $t = 1, \dots, 44$  and  $n = 1, \dots, N(t)$ :
 

**(35)  $D_{NB,tn,i} \equiv 1$  if observation  $tn$  has a structure on it with  $i$  bedrooms;  
 $\equiv 0$  elsewhere.**



## Model 11: Adding the Number of Bedrooms

- *Model 11* is the following nonlinear regression:

$$(36) V_{tn} = \alpha_t [\mathbf{D}_{S,tn} + \phi \mathbf{D}_{L,tn}] g_7(\mathbf{x}_{tn}, \mathbf{y}_{tn}, \gamma) f_L(\mathbf{L}_{tn}, \lambda) \\ \times [1 + \tau(\mathbf{T}W_{tn} - 1)] [1 + \rho(\mathbf{T}T_{tn} - 8)] \\ + P_{St}(1 - \delta)^{A(t,n)} f_S(\mathbf{S}_{tn}, \mu) [\sum_{i=2}^6 \kappa_i \mathbf{D}_{NB,tn,i}] + \varepsilon_{tn} ; \\ t = 1, \dots, 44; n = 1, \dots, N(t)$$

- where the all of the functions and parameters which appear in (36) were defined in the previous model except that we have now added 5 bedroom variables,  $\kappa_2, \kappa_3, \kappa_4, \kappa_5$  and  $\kappa_6$ .
- We make the same normalizations as we made in Model 10 and in addition, we set  $\kappa_2 = 1$ .
- Model 11 has a total **113** unknown parameters.
- The  $R^2$  for Model 11 was **0.8400**, increase in LL was **75.03**.

## Adding the Width of the Land Plot.

- The final additional variable that we introduced into our property nonlinear regression model was **the width of the land plot**,  $W_{tn}$  for property sale  $n$  in period  $t$ .
- Recall that  $W_{tn}$  is measured in 10ths of a meter and **the range of this property width variable was 25 to 90**.
- Other residential property hedonic regression models for Tokyo have shown that this variable is a very significant one: the greater is the lot width, the more valuable is the land plot.
- We assume that the width variable affects the land value component of property value and does not affect the structure value.
- We modeled the width variable as a single continuous variable rather than using splines or step functions on  $W_{tn}$ .

## Model 12: Adding the Property Width Variable

- **Model 12** is the following nonlinear regression:

$$(37) V_{tn} = \alpha_t [D_{S,tn} + \phi D_{L,tn}] g_7(x_{tn}, y_{tn}, \gamma) \\ \times f_L(L_{tn}, \lambda) [1 + \tau(TW_{tn} - 1)] [1 + \rho(TT_{tn} - 8)] [1 + \sigma(W_{tn} - 25)] \\ + P_{St}(1 - \delta)^{A(t,n)} f_S(S_{tn}, \mu) [\sum_{i=2}^6 \kappa_i D_{NB,tn,i}] + \varepsilon_{tn} ; \\ t = 1, \dots, 44; n = 1, \dots, N(t)$$

- where all of the functions and parameters which appear in (37) were defined in the previous model except  $\sigma$ .
- Thus we have added 1 additional unknown parameter to Model 11 so Model 12 has a total **114** unknown parameters.
- $\sigma^*$  was **0.00402** ( $t = 27.4$ ) so an extra meter of lot width adds about 4% to the per meter squared price of the land plot.
- The  $R^2$  for Model 11 was **0.8488**, increase in LL was **401.54**.

## The Problem of Negative Predicted Land Prices

- Although the fact that Model 12 generated 4 negative estimated  $\gamma_{ij}^*$  did not lead to any negative predicted prices for land for the properties in our sample, these negative estimates could lead to negative land prices for properties not in our sample.
- Hence, it may be useful to perform a final regression where we restrict the  $\gamma_{ij}$  to be nonnegative. **This can be done by replacing  $\gamma_{01}$ ,  $\gamma_{67}$ ,  $\gamma_{77}$  and  $\gamma_{52}$  in the function  $g_7(x_{tn}, y_{tn}, \gamma)$  by the squares of these parameters and then rerunning the model defined by (37).**
- ***Model 13*** is the resulting model.
- The reduction in LL for Model 13 over Model 12 was **1.19**.
- The  $R^2$  for Model 13 was **0.8488**, the same as for Model 12.

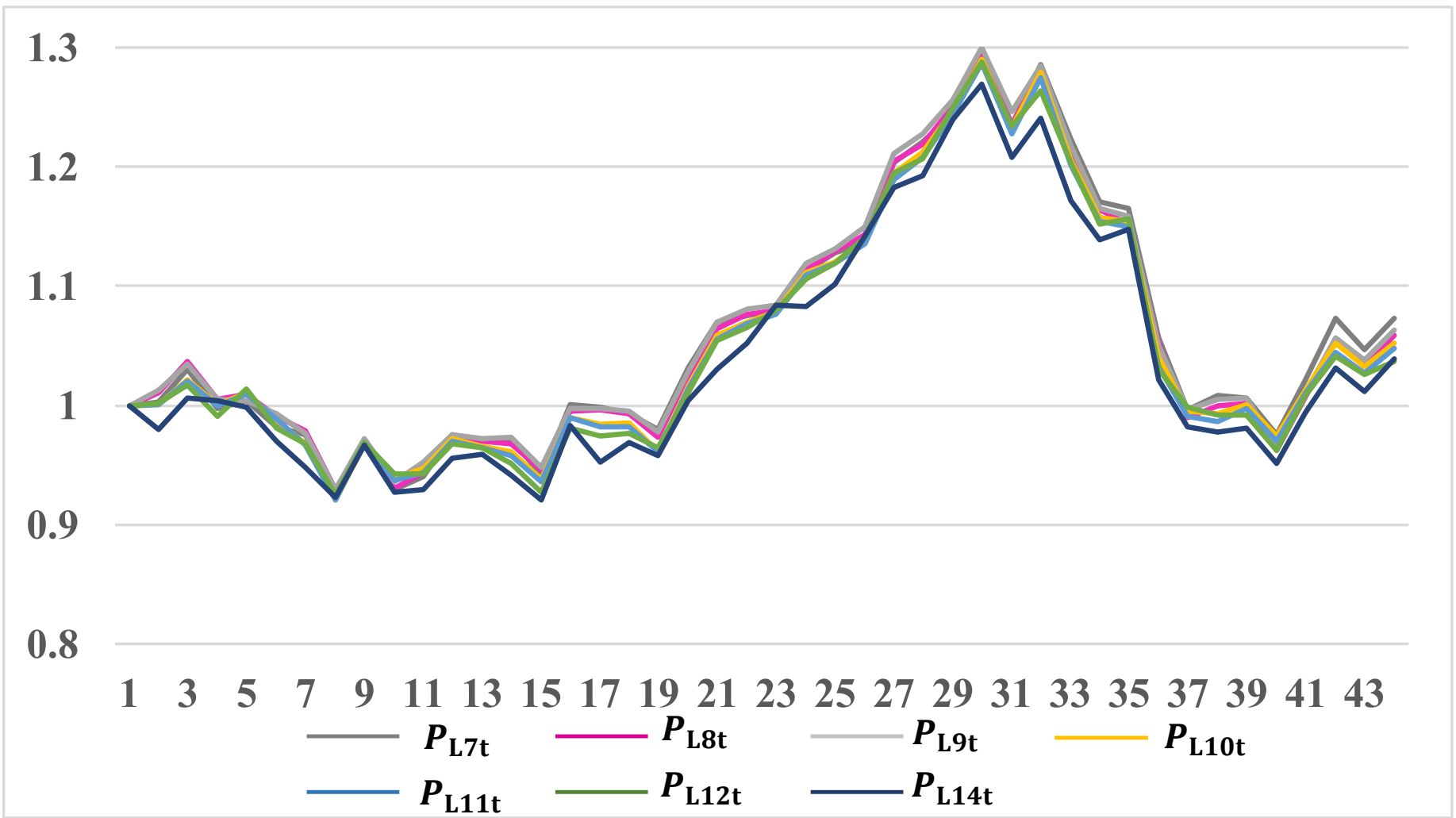
## Model 14: The Ward Dummy Variable Model Revisited

- Our final model in this section is a **Ward dummy variable model** that **adds more explanatory property characteristics** to the **Ward Dummy Model 6** defined by equations (25).
- **Model 14** is defined by the following nonlinear regression model:

$$(38) V_{tn} = \alpha_t [D_{S,tn} + \phi D_{L,tn}] [\sum_{j=1}^{23} \omega_j D_{W,tn,j}] \\ \times f_L(L_{tn}, \lambda) [1 + \tau(TW_{tn} - 1)] [1 + \rho(TT_{tn} - 8)] [1 + \sigma(W_{tn} - 25)] \\ + P_{St} (1 - \delta)^{A(t,n)} f_S(S_{tn}, \mu) [\sum_{i=2}^6 \kappa_i D_{NB,tn,i}] + \varepsilon_{tn} ;$$

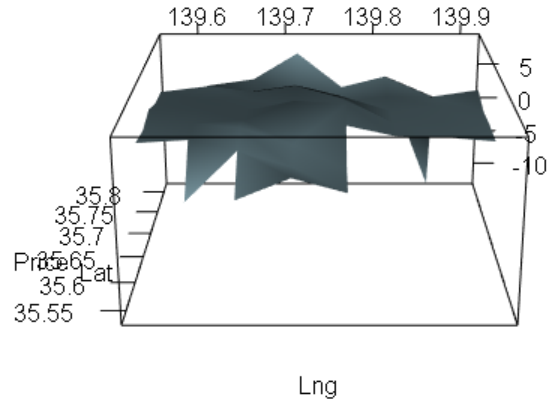
- Thus Model 14 is **basically the same** as Model 12 and 13 except that the Ward dummy variable terms,  $\sum_{j=1}^{23} \omega_j D_{W,tn,j}$ , replace the Colwell locational grid function,  $g_7(x_{tn}, y_{tn}, \gamma)$ .
- The  $R^2$  for Model 14 was **0.8300**, increase in LL over Model 6 was **478.6**.
- We compare land prices for Models 7-14 in the next slide.

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

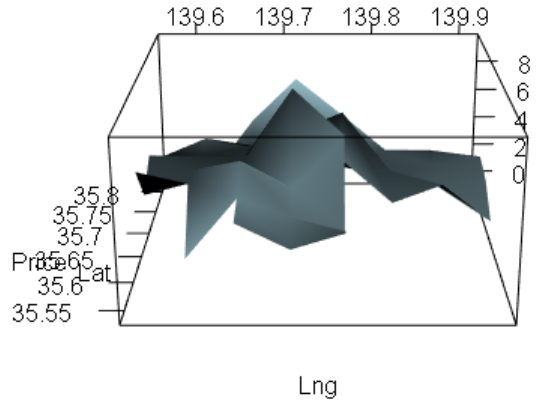


# Chart A-2. Spatial Price Distribution for 3 periods.

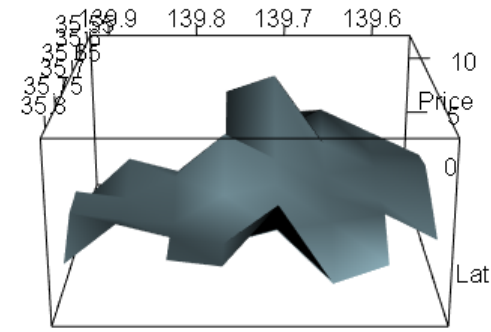
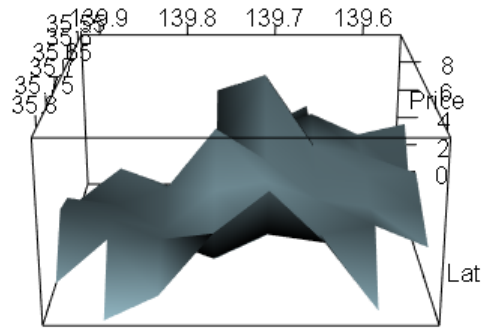
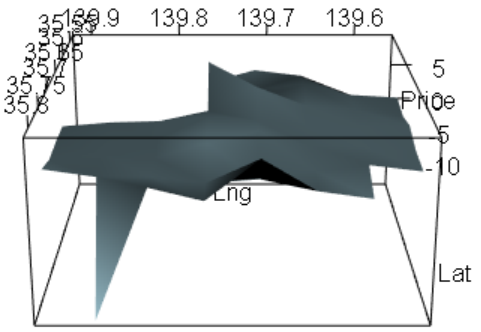
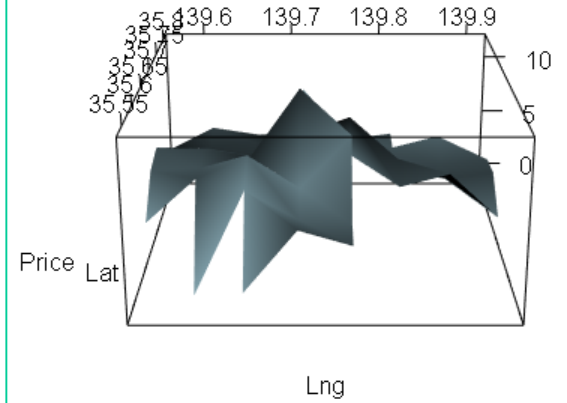
## I. 2000-2002



## II. 2004-2006

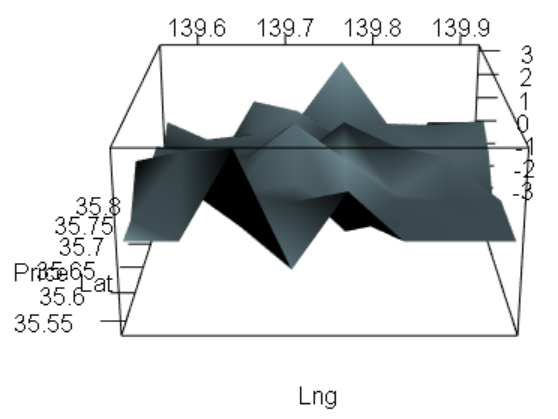


## III. 2008-2010

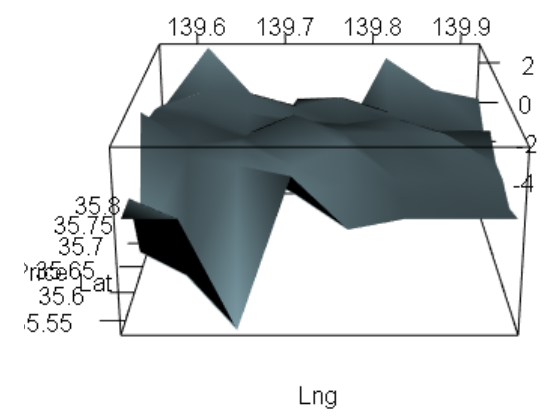


# Chart A-3. Spatial Price Dynamics in Tokyo Residential Market.

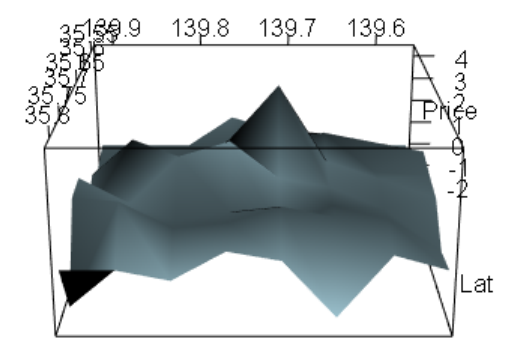
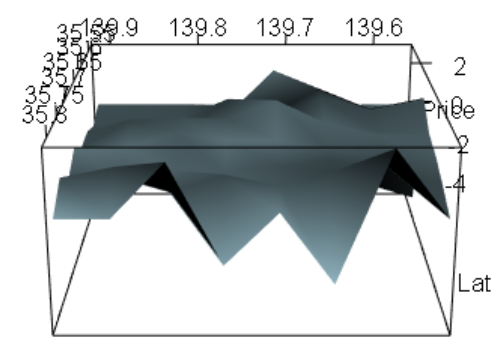
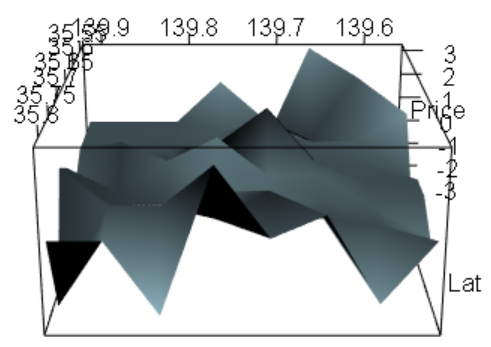
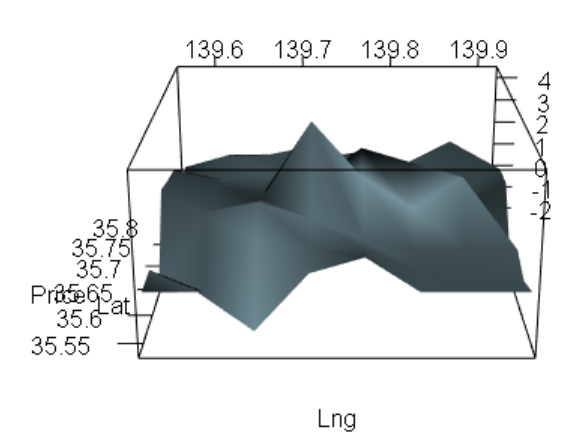
### I vs. II



### II vs. III



### I vs. III





## 8. Overall Residential Property Price Indexes

- There is one additional overall property price index that we calculate in this section and that is an index that is based on a **“traditional” hedonic property price regression** that uses the **logarithm of the selling price** as the dependent variable and has **time dummy variables**.

- Define the **kth time dummy variable**  $D_{T,tn,k}$  for property  $n$  sold in period  $t$  as follows:

$$(39) D_{T,tn,k} \equiv 1 \text{ if } t = k; D_{T,tn,k} \equiv 0 \text{ if } t \neq k.$$

- Our **best time dummy variable hedonic regression model** is the following *Model 15*:

$$(40) \ln V_{tn} = \sum_{k=2}^{44} \alpha_k D_{T,tn,k} + \sum_{j=1}^{23} \omega_j D_{W,tn,j} + \lambda \ln L_{tn} + \mu S_{tn} \\ + \delta A_{tn} + \tau TW_{tn} + \rho TT_{tn} + \sigma W_{tn} + \sum_{i=3}^6 \kappa_i D_{NB,tn,i} + \varepsilon_{tn}; \\ t = 1, \dots, 44; n = 1, \dots, N(t).$$

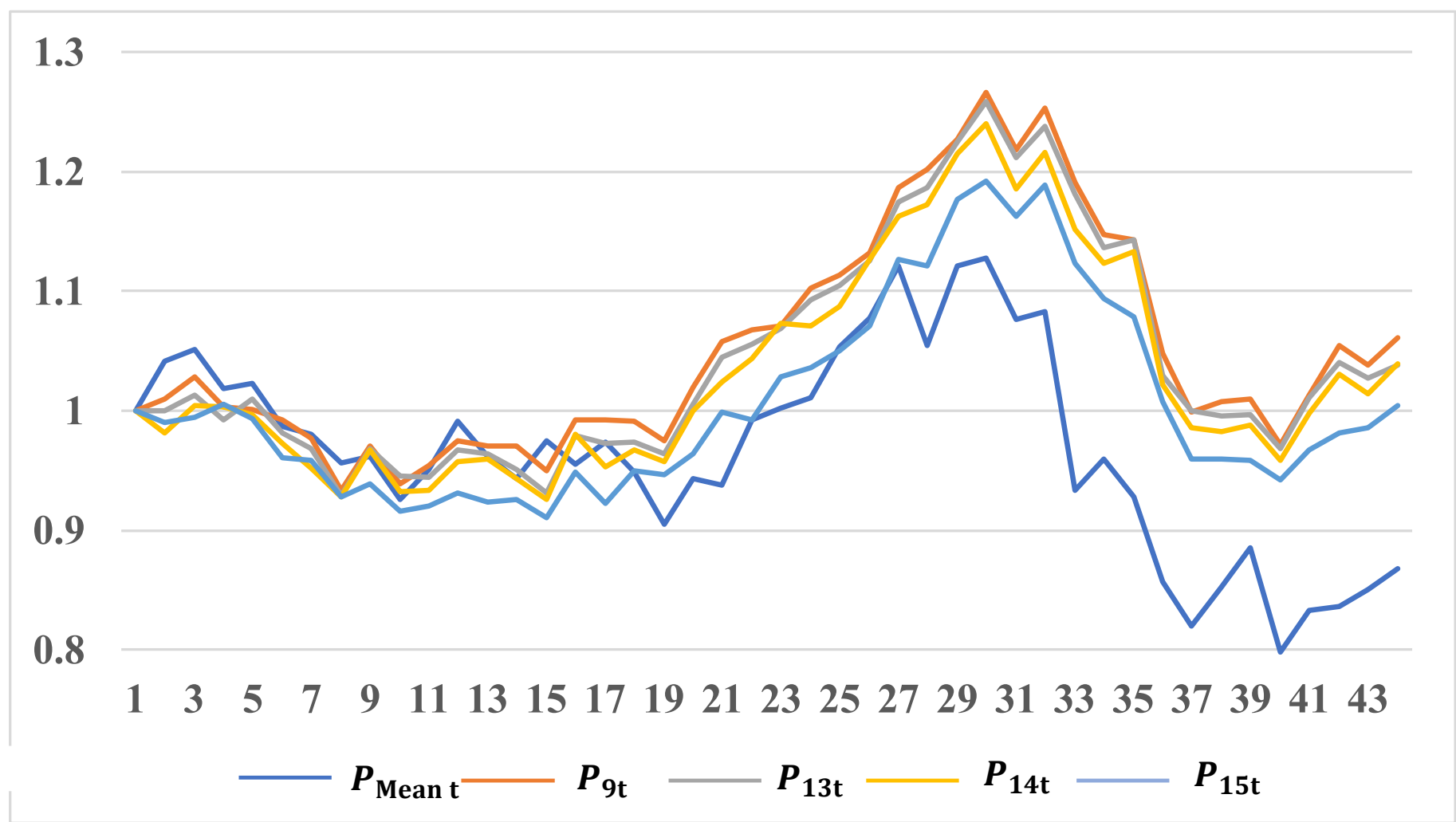
## Model 15: The Traditional Time Dummy Model

- $\ln V_{tn}$  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.
- 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 which was 204.99 points lower than our final linear regression defined by (40). 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.

## Model 15: The Traditional Time Dummy Model

- The log likelihood of this model **cannot be compared with other 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 (40).
- The  $R^2$  for Model 15 was **0.8323**. (Not bad!).
- We set  $\alpha_1^* = 0$ . The sequence of **overall property price indexes**  $P_{15t}$  generated by this model are the exponentials of the estimated  $\alpha_t^*$ ; i.e., define  $\equiv \exp[\alpha_t^*]$  for  $t = 1, \dots, 44$ .
- The next slide compares the **mean property price index**  $P_{\text{Mean } t}$ ,  $P_{9t}$  (based on Model 9, a minimal Colwell model),  $P_{13t}$  (our best Colwell spatial coordinates model),  $P_{14t}$  (our best Ward dummy variable model) and  $P_{15t}$  (our best log price time dummy hedonic model).

### Chart 4 Land Price Indexes for Models 7-12 and 14



- The mean index,  $P_{\text{Mean } t}$ , has a **large downward bias** as compared to the other 4 indexes which is due to its **neglect of age effects**. 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 (due to its imperfect specification of age effects)** but it is not large.
- The Model 9 property price index, a **Colwell spatial coordinates model that used only the 4 fundamental characteristics of a residential property** (land plot area, structure floor space area, the age of the structure and some locational variable) 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 in order 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 neighbourhood 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 utilize spatial coordinate information in order to construct a satisfactory overall residential property price index.**
- Diewert (2010) also observed a similar result.
- In addition to these **four fundamental variables**, we need an **exogenous building cost measure** in order to implement our basic models.

## 9. Conclusion

- Satisfactory residential land price indexes and overall residential property price indexes can be constructed using local neighbourhood 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 in order 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. In order 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 was 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% premium** over the land value of a neighbouring property of the same size but with a 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 section 4 of the paper, we pointed 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 in the previous section.
- Our emphasis in this paper 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 in order to obtain more accurate estimates of national (and sectoral) Total Factor Productivity growth but for the vast majority of countries, this information is simply not available.

## Extension of the Colwell Model to Higher Dimensions

- The Colwell method of interpolation over a unit square ( $x$  and  $y$  have the range  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$  and the interpolation function is

$$(A1) \ g(x,y) = \phi_{00}(1-x)(1-y) + \phi_{10}(x-0)(1-y) + \phi_{01}(1-x)(y-0) + \phi_{11}xy.$$

- For interpolation over a unit cube (3 dimensions), the interpolation function is

$$(A2) \ g(x,y,z)$$

$$= [\phi_{000}(1-x)(1-y) + \phi_{100}(x-0)(1-y) + \phi_{010}(1-x)(y-0) + \phi_{110}xy](1-z) \\ + [\phi_{001}(1-x)(1-y) + \phi_{101}(x-0)(1-y) + \phi_{011}(1-x)(y-0) + \phi_{111}xy]z.$$

- Thus we take a  $(1-z)$  times the  $xy$  square for  $z = 0$  plus  $z$  times the  $xy$  square for  $z = 1$  weighted average and get 8 terms because there are 8 vertices for a unit cube.
- There will be 16 terms for the hypercube in 4 dimensions, etc.

## Extension of the Colwell Model to Nonlinear Splines for Functions of Two Variables

- Suppose along each edge of the unit square, we fit a known continuous function of one variable instead of a linear function.
- Thus for  $y = 0$ , we have the known continuous function  $f(x,0)$  where this function of  $x$  is for example known quadratic or cubic function of  $x$ .
- For  $y = 1$ , we are given  $f(x,1)$ .
- For  $x = 0$ , we are given  $f(0,y)$ .
- For  $x = 1$ , we are given  $f(1,y)$ .
- Our problem now is to find a continuous interpolating function,  $g(x,y)$ , such that  $g$  takes on the above values for  $f$  along the 4 edges of the unit square; i.e., we want  $g$  to satisfy the following equations:

## Extension of the Colwell Model to Nonlinear Splines for Functions of Two Variables (continued)

$$(A3) \ g(x,0) = f(x,0);$$

$$(A4) \ g(x,1) = f(x,1);$$

$$(A5) \ g(0,y) = f(0,y);$$

$$(A6) \ g(1,y) = f(1,y).$$

- The following function  $g_1(x,y)$  defined as

$$(A7) \ g_1(x,y) \equiv (1-y)f(x,0) + yf(x,1)$$

- will satisfy (A3) and (A4) but not (A5) or (A6) in general.

- The following function  $g_2(x,y)$  defined as

$$(A8) \ g_2(x,y) \equiv (1-x)f(0,y) + xf(1,y)$$

- will satisfy (A5) and (A6) but not (A3) or (A4) in general.

- Define  $g_3(x,y)$  as the usual Colwell linear approximation to  $f$  which just uses the 4 vertex values of  $f(x,y)$ :

## Extension of the Colwell Model to Nonlinear Splines for Functions of Two Variables (continued)

(A9)  $g_3(x,y)$

$$\equiv (1-x)(1-y)f(0,0) + x(1-y)f(1,0) + y(1-x)f(0,1) + xyf(1,1).$$

- Define  $g(x,y)$  as the sum of  $g_1(x,y)$  and  $g_2(x,y)$  minus  $g_3(x,y)$ :

(A10)  $g(x,y) \equiv g_1(x,y) + g_2(x,y) - g_3(x,y)$ .

- Straightforward calculations show that the  $g(x,y)$  defined by (A10) satisfies equations (A3) to (A6). **It works!**
- It is of interest to calculate  $g(1/2,1/2)$ :

(A11)  $g(1/2,1/2)$

$$= 2[(1/4)f(0,1/2) + (1/4)f(1,1/2) + (1/4)f(1/2,0) + (1/4)f(1/2,1)] \\ - [(1/4)f(0,0) + (1/4)f(1,0) + (1/4)f(0,1) + (1/4)f(1,1)].$$

- Thus if one wants the approximation to  $f(x,y)$  to be differentiable, this can be done using the above algebra.



