Hedonic Regressions and the Decomposition of a House Price index into Land and Structure Components

W. Erwin Diewert, Jan de Haan and Rens Hendriks,[1] Revised November 6, 2011
Discussion Paper 11-01,
Department of Economics,
The University of British Columbia,
Vancouver, Canada, V6T 1Z1.
email: diewert@econ.ubc.ca

Abstract

The paper uses hedonic regression techniques in order to decompose the price of a house into land and structure components using readily available real estate sales data for a Dutch city. In order to get sensible results, it was useful to use a nonlinear regression model using data that covered multiple time periods. It also proved to be necessary to use some exogenous information on the rate of growth of construction costs in the Netherlands in order to get useful constant quality subindexes for the price of land and structures separately.

Key Words

House price indexes, land and structure components, time dummy hedonic regressions, Fisher ideal indexes.

Journal of Economic Literature Classification Numbers

C2, C23, C43, D12, E31, R21.

[1] Forthcoming in *Econometric Reviews*. A preliminary version of this paper was presented at the Economic Measurement Group Workshop, 2009, December 9-11, Crowne Plaza Hotel, Coogee Beach, Sydney, Australia. W. Erwin Diewert: Department of Economics, University of British Columbia, Vancouver B.C., Canada, V6T 1Z1 and the School of Economics, University of New South Wales, Sydney, Australia (e-mail: diewert@econ.ubc.ca); Jan de Haan, Statistics Netherlands (email: Jhhn@cbs.nl) and Rens Hendriks, Statistics Netherlands (email: r.hendriks@cbs.nl). The authors thank Christopher O'Donnell, Marc Francke, Ulrich Kohli, Alice Nakamura, Esmaralda Ramalho, Alicia Rambaldi, Mick Silver, Keith Woolford and two referees for helpful comments. The authors gratefully acknowledge the financial support from the Centre for Applied Economic Research at the University of New South Wales, the Australian Research Council (LP0347654 and LP0667655) and the Social Science and Humanities Research Council of Canada. None of the above individuals and institutions are responsible for the contents of this paper.

1. Introduction

For many purposes, it is useful to be able to decompose residential property values into a structures component and a land component. At the local government level, property tax rates are often different on the land and structures components of a property so it is necessary to have an accurate breakdown of the overall value of the property into these two components. At the national level, statistical agencies need to construct overall values of land and structures for the National Balance Sheets for the nation. If a user cost approach is applied to the valuation of Owner Occupied Housing services, it is necessary to have a decomposition of housing values into land and structures components since structures depreciate while land does not. Thus our goal in this paper is to use readily available multiple listing data on sales of residential properties and to decompose the sales price of each property into a land component and a structures component. We will use the data pertaining to the sales of detached houses in a small Dutch city for 22 quarters, starting in Quarter 1 in 2003 and running to the end of Quarter 2 in 2008. We utilize a *hedonic regression approach* to accomplish our decomposition but our approach is based on a cost oriented model which we call the *builder's approach* to modeling hedonic regressions in the housing context. A feature of our suggested approach is that it requires relatively little information on the characteristics of the houses that are in the data base: information on the plot area, the area of the structure, the age of the structure and the number of rooms in the house suffices to generate regression models that explain approximately 87% of the variation in the selling prices of the houses in the data base.

A more detailed outline of the contents of this paper follows.

In section 2, we will consider a very simple hedonic regression model where we use information on only three characteristics of the property: the lot size, the size of the structure and the (approximate) age of the structure. We run a separate hedonic regression for each quarter which leads to estimated prices for land and structures for each quarter. These estimated characteristics prices can then be converted into land and structures prices covering the 22 quarters of data in our sample. We postulate 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. Thus our approach to the valuation of a residential property is essentially a crude cost of production approach. Note that the overall value of the property is assumed to be the *sum* of these two components.

In section 3, we generalize the model explained in section 2 to allow for the observed fact that the per unit area price of a property tends to decline as the size of the lot increases (at least for large lots). We use a simple linear spline model with 2 break points. Again, a hedonic regression is run for each period and the results of these separate regressions were linked together to provide separate land and structures price indexes (along with an overall price index that combined these two components).

The models described in sections 2 and 3 were not very successful. The problem is due to multicollinearity and variability in the data and this volatility leads to a tendency for the

regression models to fit the outliers, leading to erratic estimates for the price of land and structures.

In section 4, in order to deal with the multicollinearity problem, we draw on *exogenous information* on new house building costs from the national statistical agency and assume that the price movements for new structures mirror the statistical agency movements in the price of new houses. We find that the use of exogenous information generates a very reasonable decomposition of house values into their structure and land components.

In section 5, we generalize the model in section 4 to include information on the number of rooms in the house as an additional price determining characteristic. The idea here is that a higher number of rooms in a house generally indicates that the quality of construction of the house will be higher. Our regression results support this hypothesis: the estimated increase in the price of a new structure per m^2 in Quarter 1 due to an additional room is about 2.7%.

We conclude this section by providing a brief literature review of methods used to provide a decomposition of the selling price of a dwelling unit into land and structures components. Basically, variations of *three methods* have been used:

- The vacant land method;
- The construction cost method and
- The hedonic regression method.

The first two methods utilize the following empirical relationship between the selling price of a property V, the value of the structure $p_S S$ and the value of the plot $p_L L$:

(1) $V = p_L L + p_S S$

where S is the floor space area of the structure, L is the area of the land that the structure sits on and p_S and p_L are the prices of a unit of S and L respectively. Typically, V, L and S will be available from real estate data on sales of houses so if either p_L or p_S can be determined somehow, then equation (1) will enable the other price to be determined.

The *vacant land method* for the determination of the price of land in (1) is described by Clapp (1979; 125) (1980; 256) and he noted that the method is frequently used by tax assessors and appraisers. The method works as follows: a price of land per unit area p_L is determined from the sales of "comparable" vacant land plots and then this price is applied to the comparable properties and equation (1) can then be used to solve for the structure price p_S. This method was used by Thorsnes (1997) and Bostic, Longhofer and Readfearn (2007).[2]

[2] The set of vacant lots can be augmented by properties which are sold and the associated structure is immediately demolished. Clapp (1980; 256) lists several reasons why the vacant land method is not likely to be very accurate.

The *construction cost method* uses an estimate for the per unit area construction cost p_S for the local area, which could be provided by a private company or a national statistical agency. Once p_S is known, equation (1) can be used to solve for the missing land price p_L. This method was used by Glaeser and Gyourko (2003), Gyourko and Saiz (2004) and Davis and Palumbo (2008) where the local construction cost data for U.S. cities was provided by the private company, R.S. Means. Davis and Heathcote (2007) used a variant of this method for the entire U.S. economy where Bureau of Economic Analysis estimates for both the price of structures p_S and the constant dollar quantity of housing structures S were used.[3]

A variant of the *hedonic regression method* is the method that will be used in this paper. Various versions of the method will be explained in sections 2-5. Some early papers that use a similar methodology include Clapp (1980), Palmquist (1984), Fleming and Nellis (1992) and Schwann (1998). Basically, land and structures are treated as characteristics in a hedonic regression model and marginal prices for land and structures for period t are generated as partial derivatives of the period t hedonic function and these marginal prices can be used to decompose the house value into land and structures components under certain conditions.

2. Model 1: A Simple Builder's Model

Hedonic regression models are frequently used to obtain constant quality price indexes for owner occupied housing.[4] Although there are many variants of the technique, the basic model regresses the logarithm of the sale price of the property on the price determining characteristics of the property and a time dummy variable is added for each period in the regression (except the base period). Once the estimation has been completed, these time dummy coefficients can be exponentiated and turned into an index.[5]

A residential property has a number of important price determining characteristics:

- The land area of the property (L);
- The livable floor space area of the structure (S);
- The age of the structure (A);
- The number of rooms in the structure (R);

[3] Muth (1971; 246) and Rosen (1978; 353-354) used the private company Boeckh building cost index for the various U.S. cities in their sample which determined p_S up to a multiplicative factor. The value of land and the price of land were determined by the U.S. Federal Housing Administration for their sample of U.S. properties. Then using equation (1), S was determined residually. The methods we will use in sections 4 and 5 below are close to the construction cost method but are not identical; we use only rates of change of construction costs, not their levels. Thus our suggested methods allow for local area quality adjustment factors for construction costs.

[4] For some recent literature, see Crone, Nakamura and Voith (2009), Diewert, Nakamura and Nakamura (2009), Gouriéroux and Laferrère (2009), Hill (2011), Hill, Melser and Syed (2009) and Hill (2011).

[5] An alternative approach to the time dummy hedonic method is to estimate separate hedonic regressions for both of the periods compared; this is called the hedonic imputation approach. See Haan (2008) (2009) and Diewert, Heravi and Silver (2009) for theoretical discussions and comparisons between these alternative approaches.

- The type of dwelling unit (detached, row, apartment);
- The type of construction (wood, brick, concrete);
- The location of the property.[6]

In our empirical work below, we will restrict our sample to sales of detached houses. We will not take into account the type of construction or the location variable since the house sales all take place in a small Dutch town and location should not be much of a price determining factor. However, we will use information on land area A, structure size in meters squared S, the age A of the structure and the number of rooms, R. We will find that hedonic regression models that use only the first three explanatory variables give rise to an R^2 that is in the range .87 to .88, which indicates that most of the variation in the data can be explained by using just these three variables.[7]

As noted in the introduction, for some purposes, it would be very useful to decompose the overall price of a property into *additive components* that reflected the value of the land that the structure sits on and the value of the structure. The primary purpose of the present paper is to determine whether a hedonic regression technique could provide such a decomposition.

Several researchers have suggested hedonic regression models that lead to *additive decompositions* of an overall property price into land and structures components.[8] We will now outline Diewert's (2007) justification for an additive decomposition.

If we momentarily think like a property developer who is planning to build 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 V_n^t in period t[9] and structure areas S_n^t and land areas L_n^t for n = 1,...,N(t). Assume that these prices are equal to the sum of the land and structure costs plus error terms ε_n^t which we assume are independently

[6] There are many other price determining characteristics that could be added to this list such as landscaping, the number of floors and rooms, type of heating system, air conditioning, swimming pools, views, the shape of the lot, etc. The distance of the property to various amenities such as schools and shops could also be added to the list of characteristics but if the location of the properties in the sample of sales is small enough, then it should not be necessary to add these characteristics. In our example, the Dutch town of "A" is small enough and homogeneous enough so that these neighbourhood effects can be neglected. In other cities or neighborhoods where geography creates important locational differences, our rather minimal basic model will probably not fit the data as well. Our simple builder's model will probably not work well for multiple unit structures where the height of the apartment becomes an important price determining characteristic.

[7] In section 5, we add the number of rooms as an additional explanatory variable.

[8] See Clapp (1980), Francke and Vos (2004), Gyourko and Saiz (2004), Bostic, Longhofer and Redfearn (2007), Davis and Heathcote (2007), Diewert (2007), Francke (2008), Koev and Santos Silva (2008), Statistics Portugal (2009), Diewert, Haan and Hendriks (2010) and Diewert (2010).

[9] Note that we have labeled these property prices as V_n^0 to emphasize that these are *values* of the property and we need to decompose these values into two price and two quantity components, where the components are land and structures.

normally distributed with zero means and constant variances.[10] This leads to the following hedonic regression model for period t where α^t and β^t are the parameters to be estimated in the regression:[11]

$$(1)\ V_n^t = \alpha^t L_n^t + \beta^t S_n^t + \varepsilon_n^t ; \qquad n = 1,...,N(t);\ t = 1,...,T.$$

Note that the two characteristics in our simple model are the quantities of land L_n^t and the quantities of structure S_n^t associated with the sale of 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 (1) for each period t in our sample.

The hedonic regression model defined by (1) is the simplest possible one but it applies only to new structures. But it is likely that a model that is similar to (1) applies to older structures as well. Older structures will be worth less than newer structures due to the depreciation (or deterioration due to aging effects) of the structure. Thus suppose in addition to information on the selling price of property n at time period t, V_n^t, the land area of the property L_n^t and the structure area S_n^t, we also have information on the age of the structure at time t, say A_n^t. Then if we assume a straight line depreciation model, a more realistic hedonic regression model than that defined by (1) above is the following *basic builder's model*:[12]

$$(2)\ V_n^t = \alpha^t L_n^t + \beta^t(1 - \delta^t A_n^t)S_n^t + \varepsilon_n^t ; \qquad n = 1,...,N(t);\ t = 1,...,T$$

where the parameter δ^t reflects the *net depreciation rate* as the structure ages one additional period. Thus if the age of the structure is measured in years, we would expect an annual δ^t to be between 0.5 and 1.5%.[13] Note that (2) is now a nonlinear regression

[10] We make the same stochastic assumptions for all of the regressions in this paper. For the models that are not linear in the parameters, we use maximum likelihood estimation.

[11] In order to obtain homoskedastic errors, it would be preferable to assume multiplicative errors in equation (1) since it is more likely that expensive properties have relatively large absolute errors compared to very inexpensive properties. However, following Koev and Santos Silva (2008), we think that it is preferable to work with the additive specification (1) since we are attempting to decompose the aggregate value of housing (in the sample of properties that sold during the period) into additive structures and land components and the additive error specification will facilitate this decomposition.

[12] Note that the model in this section is a *supply side model* as opposed to the *demand side model* of Muth (1971) and McMillen (2003). Basically, we are assuming identical suppliers of housing so that we are in Rosen's (1974; 44) Case (a) where the hedonic surface identifies the structure of supply. This assumption is justified for the case of newly built houses but we concede that it is less well justified for sales of existing homes. Our supply side model is also less likely to be applicable in the case of multiple unit structures where zoning restrictions and local geography lead to location specific land prices.

[13] 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. Since 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.

model whereas (1) was a simple linear regression model.[14] Both models (1) and (2) can be run period by period; it is not necessary to run one big regression covering all time periods in the data sample. The period t price of land will the estimated coefficient for the parameter α^t and the price of a unit of a newly built structure for period t will be the estimate for β^t. The period t quantity of land for property n is L_n^t and the period t quantity of structure for property n, expressed in equivalent units of a new structure, is $(1 - \delta^t A_n^t)S_n^t$ where S_n^t is the floor space area of property n in period t.

We implemented the above *Model 0* using real estate sales data on the sales of detached houses for a small city (population is around 60,000) in the Netherlands, City "A", for 22 quarters, starting in Q1 2003 and extending through Q2 in 2008 (so our T = 22). The data that we used can be described as follows:

- V_n^t is the selling price of property n in quarter t in units of 1,000 Euros where t = 1,...,22;
- L_n^t is the area of the plot for the sale of property n in quarter t in units of meters squared;[15]
- S_n^t is the living space area of the structure for the sale of property n in quarter t in units of meters squared;
- A_n^t is the (approximate) age in decades of the structure on property n in period t;[16]
- R_n^t is the number of rooms in structure n that was sold in period t.

It seems likely that the number of rooms in a structure will be roughly proportional to the area of the structure, so in our initial regressions in sections 3-5, we did not use the room variable R as an explanatory variable.[17]

Initially, there were 3543 observations in our 22 quarters of data on sales of detached houses in City "A" that were less than 50 years old when sold. However, there were some obvious outliers in the data. Thus we looked at the range of our V, L, S and R variables and deleted 54 range outliers. There were also two duplicate observations in Q1 for 2006 and these duplicates were also deleted. Thus we ended up with 3487 data points for the 22 quarters.[18] The sample means for the data with outliers excluded (standard deviations

[14] This formulation follows that of Diewert (2007) and Diewert, Haan and Hendriks (2010). It is a special case of Clapp's (1980; 258) model except that Clapp included a constant term.

[15] We chose units of measurement for V in order to scale the data to be small in magnitude so as to facilitate convergence for the nonlinear regressions. The statistical package used was Shazam (the nonlinear option).

[16] The original data were coded as follows: if the structure was built 1960-1970, the observation was assigned the dummy variable BP = 5; 1971-1980, BP=6; 1981-1990, BP=7; 1991-2000, BP=8. Our Age variable A was set equal to 8 – BP. Thus for a recently built structure n in quarter t, $A_n^t = 0$.

[17] In section 5 below, we did use the room variable as a quality adjustment variable.

[18] There were 3 observations where the selling price was less than 60,000 and 14 observations which sold for more than 550,000 Euros. There were no sales with L less than 70 m^2 and 25 sales where L exceeded 1500 m^2. There were no sales with S less than 50 and one observation where S exceeded 400 m^2. There were 13 sales where R was less than 2 and 3 sales where R exceeded 14. All of these observations were excluded. Some observations were excluded multiple times so that the total number of observations which

in brackets) were as follows: $\bar{V}$ = 182.26 (71.3), $\bar{L}$ = 258.06 (152.3), $\bar{S}$ = 126.56 (29.8), $\bar{A}$ = 1.8945 (1.23) and $\bar{R}$ = 4.730 (0.874). Thus the entire sample of houses sold at the average price of 182,260 Euros, the average plot size was 258.1 m^2, the average living space in the structure was 126.6 m^2 and the average age was approximately 18.9 years. The sample median price was 160,000 Euros.

The correlations between the various variables are also of interest. The correlation coefficients of the selling price V with L, S, A and R are .8014, .7919, –.3752 and .3790 respectively.[19] Thus the selling price V is fairly highly correlated with both land L and (unadjusted) structures S. The correlation between L and S is .6248 and thus there is the possibility of multicollinearity between these variables. Finally there is also a substantial positive correlation of .4746 between the structure area S and the number of rooms R.

Instead of running 22 quarterly regressions of the form (2), we combined the data using dummy variables and ran one big regression, which combined all 22 quarterly regressions into a single regression.[20] The R^2 for the resulting combined regression was .8729, which is quite good, considering we have only 3 explanatory variables (but 66 parameters to estimate). The resulting log likelihood was −16231.6. The *quality adjusted structures quantity in quarter t*, S^{t*}, is equal to the sum over the properties sold n in that quarter adjusted into new structure units; i.e., $S^{t*} \equiv \sum_{n \in N(t)} (1 - \delta^{t*} A_n^t) S_n^t$. The estimated decade net depreciation rates δ^{t*} were in the 6.4% to 13.7% range which is not unreasonable but the volatility in these rates is not consistent with our a priori expectation of a stable rate. We did not list our regression results because our estimated land and structures prices are not at all reasonable: the price of land sinks to a very low level in quarter 3 while the price of structures has a local peak in this quarter. In general, the land and constant quality structures prices are volatile in opposite directions, which is a sign of a severe multicollinearity problem.[21]

In an attempt to improve the results for the above Model 0, we assumed that the net depreciation rate was constant across quarters and so the model defined by (2) is replaced by the following *Model 1*:

$$(3)\ V_n^t = \alpha^t L_n^t + \beta^t (1 - \delta A_n^t) S_n^t + \varepsilon_n^t ; \qquad n = 1,...,N(t);\ t = 1,...,T$$

where the parameter δ reflects the sample *net depreciation rate* as the structure ages one additional decade but now it is assumed to be constant over the entire sample period.

---

were excluded was 54 (plus 2 more due to duplication in the data set). Exclusion of range outliers is important for the results.

[19] In order to illustrate the importance of deleting range outliers for all variables, the correlation coefficients of V with L, S, A and R for the original data set with 3543 observations was 0.33331, 0.80795, –0.34111 and 0.34291. Thus it is particularly important to delete land outliers.

[20] This one big regression generates the same parameter values as running the individual quarterly regressions but the advantage of the one big regression approach is that we can compare the log likelihood of the big regression with subsequent regressions.

[21] This period to period parameter instability problem was noted by Schwann (1998; 277) in his initial unconstrained model: "In addition, the unconstrained regression displays signs of multicollinearity. ... the attribute prices are nonsense in many of the periods, and there is poor temporal stability of these prices."

Thus the new builder's hedonic regression model has 45 unknown parameters to estimate as compared to the 66 parameters in the previous model defined by equations (2).

The R^2 for the resulting nonlinear regression model was .8703,[22] which is quite good, considering we have only 2 independent explanatory variables in each period. However, this is a drop in R^2 as compared to our previous model with variable depreciation rates where the R^2 was .8729. The log likelihood for the constant depreciation rate model was −16266.6, which is a decrease of 35.0 from the log likelihood of the previous model. This decrease in log likelihood seems to be a reasonable price to pay in order to obtain a stable estimate for the net depreciation rate. The estimated decade net depreciation rate is now $\delta^* = 0.10241$ or about 1% per year. The smallest t statistic for the parameters in this model was 11.9 for the parameter α^{1*}. The results for our new model (3) are summarized in Table 1 below. The estimated *quality adjusted structures quantity in quarter t*, S^{t*}, is equal to the sum over the properties sold n in that quarter, quality adjusted (for net depreciation) into new structure units; i.e.:

$$(4)\quad S^{t*} \equiv \sum_{n \in N(t)} (1 - \delta^* A_n^t) S_n^t ; \qquad t = 1,...,22$$

where δ^* is the estimated net depreciation rate for the entire sample period.

Table 1: Estimated Land Prices α^{t*}, Structure Prices β^{t*}, the Decade Depreciation Rate δ^*, Land Quantities L^t and Quality Adjusted Structures Quantities S^{t*}

Quarter	α^{t*}	β^{t*}	δ^*	L^t	S^{t*}
1	0.25162	0.97205	0.10241	35023	14677.2
2	0.30084	0.86961	0.10241	35412	14047.9
3	0.20130	1.07050	0.10241	39872	14680.1
4	0.26348	0.97486	0.10241	42449	16764.0
5	0.28792	0.95083	0.10241	37319	14787.8
6	0.24087	1.09845	0.10241	45611	16828.1
7	0.27564	1.02882	0.10241	33321	13234.3
8	0.23536	1.09186	0.10241	40395	17169.1
9	0.23548	1.10259	0.10241	38578	16680.0
10	0.30717	1.00917	0.10241	38246	15847.6
11	0.26523	1.14512	0.10241	39112	15831.3
12	0.22357	1.19693	0.10241	41288	16119.8
13	0.27415	1.09353	0.10241	43387	16873.5
14	0.24764	1.20932	0.10241	46132	19037.4
15	0.30056	1.11530	0.10241	39250	15889.7
16	0.26941	1.13981	0.10241	40102	15836.9
17	0.31121	1.08539	0.10241	39813	16234.7
18	0.23368	1.28996	0.10241	56992	20579.3
19	0.31558	1.10402	0.10241	35801	13661.4
20	0.27131	1.19228	0.10241	48031	19610.7
21	0.21835	1.29223	0.10241	37854	15344.4

[22] 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.

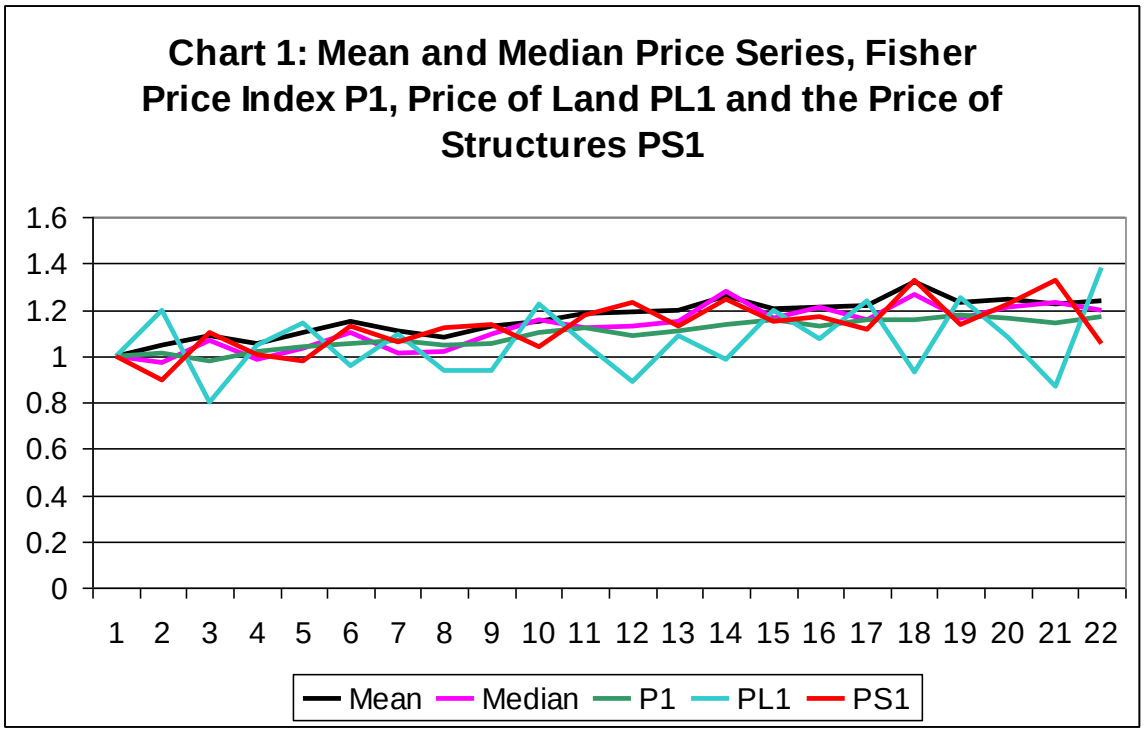
22	0.34704	1.02324	0.10241	45878	19645.7

It is of some interest to compare the above land and structures prices with the mean and median prices for houses in the sample for each quarter. These prices were normalized to equal 1 in quarter 1 and are listed as P_{Mean} and P_{Median} in Table 2 below. The land and structures prices in Table 1, α^{t*} and β^{t*}, were also normalized to equal 1 in quarter 1 and are listed as P_{L1} and P_{S1} in Table 2. Finally, we used the price data in Table 2, α^{t*} and β^{t*}, along with the corresponding quantity data, L^t and S^{t*}, in Table 1 in order to calculate a "constant quality" chained Fisher (1922) house price index, which is listed as P_1 in Table 2.

Table 2: Quarterly Mean, Median and Fisher Housing Prices P_1 and the Price of Land P_{L1} and Structures P_{S1}

Quarter	P_{Mean}	P_{Median}	P_1	P_{L1}	P_{S1}
1	1.00000	1.00000	1.00000	1.00000	1.00000
2	1.04916	0.97007	1.01150	1.19559	0.89461
3	1.08473	1.06796	0.97511	0.80001	1.10128
4	1.05544	0.98592	1.01626	1.04711	1.00289
5	1.10128	1.03521	1.03964	1.14425	0.97817
6	1.14688	1.10035	1.05462	0.95727	1.13004
7	1.10436	1.01408	1.06757	1.09546	1.05840
8	1.07874	1.02113	1.04559	0.93537	1.12326
9	1.12774	1.09155	1.05259	0.93584	1.13429
10	1.15032	1.15493	1.10079	1.22074	1.03819
11	1.18601	1.12148	1.12179	1.05409	1.17805
12	1.19096	1.12676	1.08897	0.88850	1.23134
13	1.19633	1.14789	1.10521	1.08951	1.12497
14	1.26120	1.28169	1.13606	0.98418	1.24409
15	1.20159	1.16197	1.15825	1.19450	1.14737
16	1.21170	1.21303	1.12513	1.07071	1.17258
17	1.21731	1.15493	1.15603	1.23682	1.11660
18	1.31762	1.26761	1.15751	0.92870	1.32705
19	1.22870	1.16056	1.17844	1.25419	1.13576
20	1.24592	1.20775	1.16364	1.07825	1.22656
21	1.22596	1.23239	1.14472	0.86778	1.32939
22	1.23604	1.19718	1.16987	1.37920	1.05266
Mean	1.1645	1.1234	1.0941	1.0617	1.1249

It can be seen that the mean and median series are rather volatile and differ substantially from P_1, the Fisher index that is compiled using the results of our builder's regression model (3) using the data on the price of land P_{L1} and quality adjusted structures P_{S1} and the associated quantities tabled in Table 2 above. The overall Fisher house price index P_1 is fairly smooth but its component prices P_{L1} and P_{S1} fluctuate violently. The price series listed in Table 2 are graphed in Chart 1.



It can be seen that the Mean and Median price series are on average substantially above the corresponding overall Fisher house price index P_1 and the series P_1 is much smoother.[23] It appears that the P_1 series provides satisfactory estimates for the overall price of houses. On the other hand, the component land and structure price series for P_1, P_{L1} and P_{S1}, are extremely volatile and hence are not very credible estimates for the underlying movements for the price of land and constant quality structures in the town of "A" over this period. It can be seen that when the price of land spikes up, the corresponding price of structures tends to spike downwards and vice versa. This erratic behavior in P_{L1} and P_{S1} is due to measurement errors in the quantity of land and the quantity of structures[24] along with a substantial correlation between the quantity of land and structures; i.e., we have a multicollinearity problem.

One possible problem with our highly simplified house price model is that our model makes no allowance for the fact that larger sized plots tend to sell for an average price that is below the price for medium and smaller sized plots. Thus in the following section, we will generalize the builder's model (3) to take into account this empirical regularity.

3. Model 2: The Builder's Model with Linear Splines on Lot Size

[23] We attribute the slower rate of growth in our hedonic index P_1 as compared to the Mean and Median indexes to the fact that new houses tend to get bigger over time. The Mean and Median indexes cannot take this quality improvement into account.

[24] The measurement errors here are include recording errors but also include errors due to our imperfect measurement of the quality of construction and the quality of the land; e.g., we are assuming that all locations in our sample have access to the same amenities and share the same geography and hence should face the same land price schedule but in fact, this will not be true.

In most countries, the reality is that large lots tend to sell at a lower price per unit area than smaller lots.[25] Thus in this section, we will assume that builders face a piecewise linear schedule of prices per unit land when they purchase a lot. This linear spline model will allow the price of large lots to drop as compared to smaller lots. We broke up our 3487 observations into 3 groups of property sales:

- Sales involving lot sizes less than 170 meters squared (Group S);
- Sales involving lot sizes between 170 and less than 270 meters squared (Group M) and
- Sales involving lot sizes greater than or equal to 270 meters squared (Group L).

The small lot size group had 1194 sales, the medium lot size group 1108 sales and the large lot size group had 1185 sales, so that the three groups were roughly equal in size. We define the sets of observations n which belong to Group S, M and L in period t to be $N_S(t)$, $N_M(t)$ and $N_L(t)$ respectively.

For an observation n in period t that was associated with a small lot size, our regression model was essentially the same as in (3) above; i.e., the following estimating equation was used:

$$(5)\ V_n^t = \alpha_S^t L_n^t + \beta^t(1 - \delta A_n^t)S_n^t + \varepsilon_n^t ; \qquad t = 1,...,22;\ n \in N_S(t)$$

where the unknown parameters to be estimated are α_S^t, β^t for t = 1,...,22 and δ. For an observation n in period t that was associated with a medium lot size, the following estimating equation was used:

$$(6)\ V_n^t = \alpha_S^t(170) + \alpha_M^t(L_n^t - 170) + \beta^t(1 - \delta A_n^t)S_n^t + \varepsilon_n^t ; \qquad t = 1,...,22;\ n \in N_M(t)$$

where we have added 22 new parameters to be estimated, the α_M^t for t = 1,...,22. Finally, for an observation n in period t that was associated with a large lot size, the following estimating equation was used:

$$(7)\ V_n^t = \alpha_S^t(170) + \alpha_M^t(270 - 170) + \alpha_L^t(L_n^t - 270) + \beta^t(1 - \delta A_n^t)S_n^t + \varepsilon_n^t ; \qquad t = 1,...,T;\ n \in N_L(t)$$

where we have added 22 new parameters to be estimated, the α_L^t for t = 1,...,22. Thus for small lots, the value of an extra marginal addition of land in quarter t is α_S^t, for medium lots, the value of an extra marginal addition of land in quarter t is α_M^t and for large lots, the value of an extra marginal addition of land in quarter t is α_L^t. These pricing schedules are joined together so that the cost of an extra unit of land increases with the size of the

[25] This empirical regularity was noted by Francke (2008; 168): "However, the assumption that the value is proportional to the lot size is not valid for large lot sizes. In practice, real estate agents often use a step function for the valuation of the lot, as shown in Figure 8.1. The first 300 m^2 counts for 100%, from 300 m^2 until 500 m^2 counts for 53% and so on." At first glance, it appears that Francke is using a step function to model the price schedule but in fact, he used linear splines in the same way as the present authors.

lot in a continuous fashion.[26] The above model can readily be put into a nonlinear regression format for each period using dummy variables to indicate whether an observation is in Group S, M or L. The nonlinear option in Shazam was used to estimate *Model 2* defined by (5)-(7) as one big regression.

The R^2 for this model was .8756, an increase over the previous two models (without splines) where the R^2 was .8729 (many depreciation rates) and .8703 (one depreciation rate). The new log likelihood was −16195.0, an increase of 71.6 from the previous model's log likelihood. The estimated decade depreciation rate was $\delta^* = 0.1019$ (0.00329).[27] The first period parameter values for the 3 marginal prices for land were $\alpha_S^{1*} = 0.2889$ (0.0497), $\alpha_M^{1*} = 0.3643$ (0.0566) and $\alpha_L^{1*} = 0.1895$ (0.319). Thus in quarter 1, the marginal cost per m^2 of small lots was estimated to be 288.9 Euros per m^2. For medium sized lots, the estimated marginal cost was 364.3 Euros/m^2. For large lots, the estimated marginal cost was 189.5 Euros/m^2. The first period parameter value for quality adjusted structures was $\beta^{1*} = 0.8829$ (0.0800) so that a square meter of new structure was valued at 882.9 Euros/m^2. All of the estimated coefficients were positive. The lowest t statistic for all of the 89 parameters was 2.79 (for α_S^8), so all of the estimated coefficients in this model were significantly different from zero. Our conclusion is that adding splines for the lot size gives us additional explanatory power.

Once the parameters for the model have been estimated, then in each quarter t, we can calculate the *predicted value of land for small, medium and large lot sales*, V_{LS}^t, V_{LM}^t and V_{LL}^t respectively, along with the associated *quantities of land*, L_{LS}^t, L_{LM}^t and L_{LL}^t as follows:

(8) $V_{LS}^t \equiv \sum_{n \in N_S(t)} \alpha_S^{t*} L_n^t$; t = 1,...,22;

(9) $V_{LM}^t \equiv \sum_{n \in N_M(t)} \alpha_S^{t*}[170] + \alpha_M^{t*}[L_n^t - 170]$; t = 1,...,22;

(10) $V_{LL}^t \equiv \sum_{n \in N_L(t)} \alpha_S^{t*}[170] + \alpha_M^{t*}[100] + \alpha_L^{t*}[L_n^t - 270]$; t = 1,...,22;

(11) $L_{LS}^t \equiv \sum_{n \in N_S(t)} L_n^t$; t = 1,...,22;

(12) $L_{LM}^t \equiv \sum_{n \in N_M(t)} L_n^t$; t = 1,...,22;

(13) $L_{LL}^t \equiv \sum_{n \in N_L(t)} L_n^t$ t = 1,...,22.

The *corresponding average quarterly prices*, P_{LS}^t, P_{LM}^t and P_{LL}^t, for the three types of lot are defined as the above values divided by the above quantities:

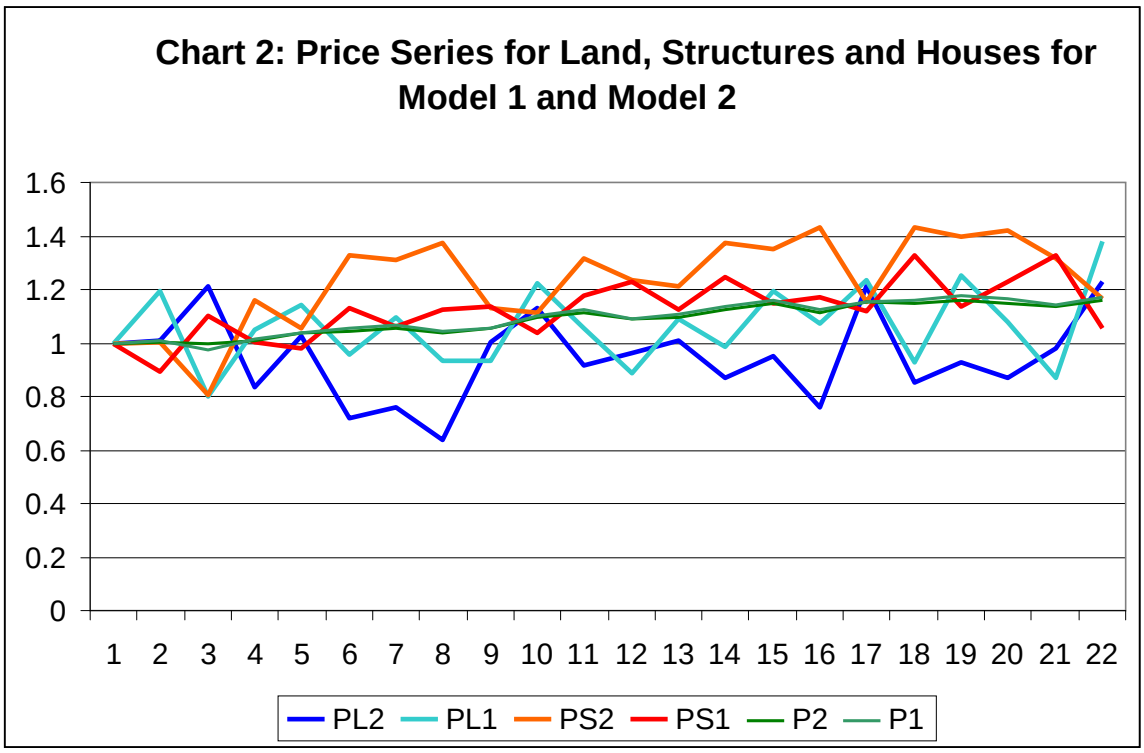
(14) $P_{LS}^t \equiv V_{LS}^t / L_{LS}^t$; $P_{LM}^t \equiv V_{LM}^t / L_{LM}^t$; $P_{LL}^t \equiv V_{LL}^t / L_{LL}^t$; t = 1,...,22.

[26] Thus if we graphed the total cost C of a lot as a function of the plot size L in period t, the resulting cost curve would be made up of three linear segments whose endpoints are joined. The first line segment starts at the origin and has the slope α_S^t, the second segment starts at L = 170 and runs to L = 270 and has the slope α_M^t and the final segment starts at L = 270 and has the slope α_L^t.

[27] Standard errors are in brackets.

The average land prices for small, medium and large lots defined by (14) and the corresponding quantities of land defined by (11)-(13) can be used to form a chained *Fisher land price index*, which we denote by P_{L2}. This index is plotted in Chart 2. As in the previous model, the estimated period t price for a square meter of quality adjusted structures is β^{t*} and the corresponding quantity of constant quality structures is $S^{t*} \equiv \sum_{n \in N(t)} (1 - \delta^* A_n^t) S_n^t$. The structures price and quantity series β^{t*} and S^{t*} were combined with the three land price and quantity series to form a chained *overall Fisher house price index* P_2 which is graphed in Chart 2. The *constant quality structures price index* P_{S2} (a normalization of the series $\beta^{1*},...,\beta^{22*}$) is also found in Chart 2.

In the following Chart, we will compare the price series P_{L2}, P_{S2} and P_2 generated by Model 2 with the price series P_{L1}, P_{S1} and P_1 that were generated by Model 1 in the previous section (which did not include splines on the size of the land area).



It can be seen that again there is a volatility problem with the price of land P_{L2} and the price of structures P_{S2} in our new builder's model with splines on land: when the price of land jumps up, the price of structures drops down and in fact, the offsetting jumps are now bigger than they were using the no splines model with a constant depreciation rate that was described at the end of the previous section. This offsetting volatility is again an indication of a severe multicollinearity problem. However, note that both models generate essentially the same overall house price index, which is quite smooth and looks reasonable; i.e., P_1 and P_2 can hardly be distinguished in Chart 2.

Due to the high correlation between the size of the structure and the size of the underlying plot and the measurement error in our land and quality adjusted structures series, it is going to be a difficult task to extract meaningful price and structure components out of information on house sales alone. Thus in the following section, we will add some *additional restrictions* on our basic model described in this section in attempts to obtain more meaningful land and structures price series.[28]

4. Model 3: The Use of Exogenous Information on New Construction Prices

Many countries have national or regional new construction price indexes available from the national statistical agency on a quarterly basis.[29] This is the case for the Netherlands.[30] Thus if we are willing to make the assumption that new construction costs for houses have the same rate of growth over the sample period across all cities in the Netherlands, the statistical agency information on construction costs can be used to eliminate the multicollinearity problems that we encountered in the previous sections.

Recall equations (5)-(7) in section 3 above. These equations are the estimating equations for Model 2. In the present section, the constant quality house price parameters, the β^t for t = 2,...,22 in (5)-(7), are replaced by the following numbers, which involve only the single unknown parameter β^1:

$$(15)\quad \beta^t = \beta^1 p^t ; \qquad t = 2,3,...,22$$

where p^t is the statistical agency estimated *construction cost price index* for the location under consideration and for the type of dwelling, where this series has been normalized to equal unity in quarter 1. This new regression *Model 3* is again defined by equations (5)-(7) except that the 22 unknown β^t parameters are now assumed to be defined by (15), so that only β^1 needs to be estimated for this new model.[31] Thus the number of parameters to be estimated in this new restricted model is 68 as compared to the Model 2 number, which was 89.

---

[28] Another approach to the volatility problem is to use a *smoothing method* in order to stabilize the volatile period to period characteristics prices. This approach dates back to Coulson (1992) and Schwann (1998) and more recent contributions include Francke and Vos (2004), Francke (2009) and Rambaldi, McAllister, Collins and Fletcher (2011). We have not pursued this approach because we feel that it is not an appropriate one for statistical agencies who have to produce non-revisable housing price indexes in real time. The use of smoothing methods is appropriate when the task is to produce historical series but smoothing methods do not work well in a real time context due to the inability of these methods to predict turning points in the series.

[29] As was seen in section 1, many countries have private companies that can provide timely construction price indexes for major cities in the country and this information could be used.

[30] From the Dutch Central Bureau of Statistics online source, Statline, we obtained a quarterly series for "New Dwelling Output Price Indices, Building Costs, 2005 = 100, Price Index: Building costs including VAT" for the last 14 quarters in our sample. Data from Statline for the first 8 quarters in our sample were also available but using the base year 2000 = 100. The older series was linked to the newer series and the resulting series was normalized to 1 in the first quarter. The resulting series is denoted by p^1 (=1), p^2,...,p^{22}.

[31] This type of hedonic model that makes use of construction price information is similar to that introduced by Diewert (2010).

Using the data for the town of "A", the estimated decade depreciation rate was $\delta^* = 0.1026$ (0.00448). The R^2 for this model was .8723, a drop from the previous Model 2 R-squared of .8756. The log likelihood was −16239.7, a substantial decrease of 44.7 over Model 2. The first period parameter values for the 3 marginal prices for land are $\alpha_S^{1*} = 0.1827$ (0.0256), $\alpha_M^{1*} = 0.3480$ (0.0640) and $\alpha_L^{1*} = 0.17064$ (0.0311). The first period parameter value for quality adjusted structures is $\beta^{1*} = 1.0735$ (0.0275) or 1073.5 Euros/m^2 which is substantially higher than the corresponding Model 1 and 2 estimates which were 972.1 and 882.9 Euros/m^2 respectively. Thus the imposition of a nationwide growth rate on the change in the price of quality adjusted structures for the town of "A" has had some effect on our previous estimates for the levels of land and structures prices.

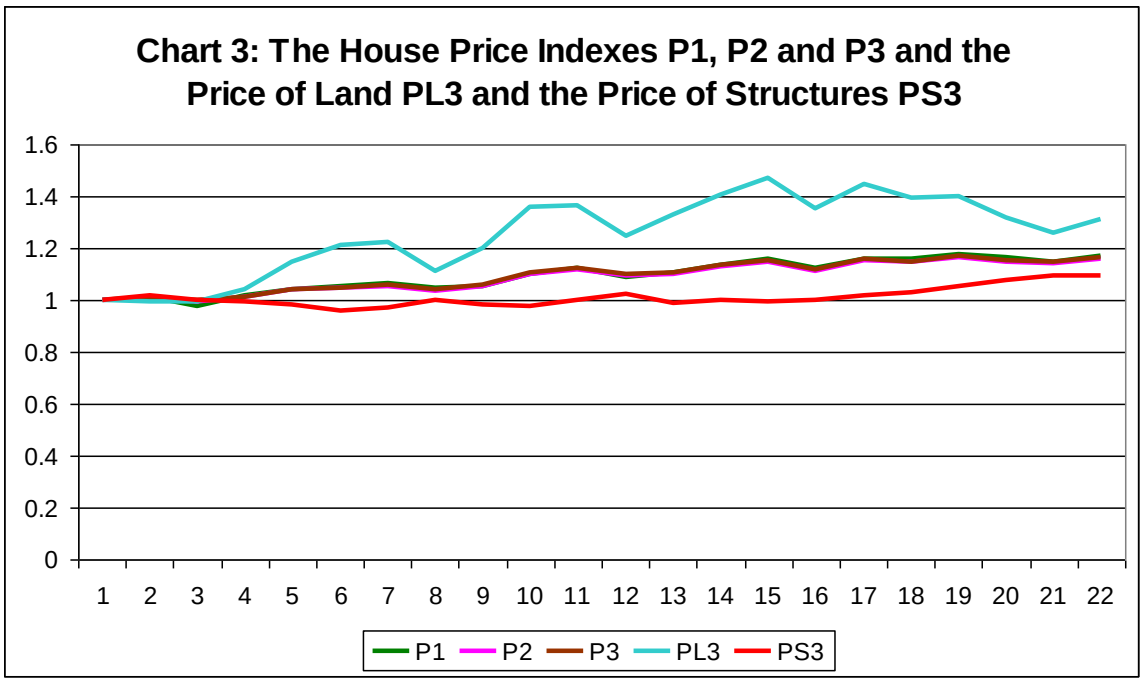
As usual, we used equations (8)-(14) in order to construct a chained Fisher index of land prices, which we denote by P_{L3}. This index is plotted in Chart 3 and listed in Table 3 below. As was the case for the previous two models, the estimated period t price for a square meter of quality adjusted structures is β^{t*} (which in turn is now equal to $\beta^{1*}p^t$) and the corresponding quantity of constant quality structures is $S^{t*} \equiv \sum_{n=1}^{N(t)} (1 - \delta^* A_n^t) S_n^t$. The structures price and quantity series β^{t*} and S^{t*} were combined with the three land price and quantity series to form a chained *overall Fisher house price index* P_3 which is graphed in Chart 3 and listed in Table 3. The *constant quality structures price index* P_{S3} (a normalization of the series $\beta^{1*},...,\beta^{22*}$) is also found in Chart 3 and Table 3. It should be noted that the quarter to quarter movements in P_{S3} coincided with the quarter to quarter movements in the Statistics Netherlands New Dwellings Building Cost Price Index.

Table 3: The Price of Land P_{L3}, the Price of Structures P_{S3} and the Overall House Price Index P_3 Generated by Model 3 with the Corresponding Quantities Q_{L3}, Q_{S3} and Q_3

Quarter	P_{L3}	P_{S3}	P_3	Q_{L3}	Q_{S3}	Q_3
1	1.00000	1.00000	1.00000	7446.9	15749.3	23196.2
2	0.99248	1.01613	1.00842	7602.4	15073.6	22671.1
3	0.99248	1.00000	0.99769	8622.7	15752.1	24366.2
4	1.04399	0.99194	1.01035	9172.6	17988.4	27138.6
5	1.14791	0.98387	1.04007	8057.7	15868.4	23904.1
6	1.20958	0.95968	1.04554	9898.8	18057.6	28026.7
7	1.22438	0.96774	1.05593	7200.3	14201.1	21364.1
8	1.11160	1.00000	1.04056	8659.1	18424.2	26956.4
9	1.20134	0.98387	1.05818	8285.6	17899.5	26048.9
10	1.35900	0.97690	1.10428	8221.2	17006.0	25161.9
11	1.36491	0.99881	1.12097	8406.4	16988.4	25373.0
12	1.24923	1.02271	1.09813	8842.9	17298.3	26169.8
13	1.33155	0.99084	1.10504	9338.7	18106.5	27488.3
14	1.40580	1.00080	1.13646	9931.0	20429.3	30275.2
15	1.47191	0.99582	1.15492	8436.9	17050.8	25454.5
16	1.35274	0.99881	1.11711	8633.4	16994.4	25649.0
17	1.44763	1.01773	1.16136	8566.5	17421.0	25944.6
18	1.39479	1.02769	1.14980	12262.7	22082.7	34613.1

19	1.40183	1.05159	1.16770	7709.2	14659.1	22456.3
20	1.32049	1.07449	1.15549	10337.4	21044.1	31382.3
21	1.25610	1.09540	1.14825	8141.9	16465.8	24614.7
22	1.31144	1.09540	1.16627	9853.6	21082.3	30881.3
Mean	1.2541	1.0114	1.0928	8801.3	17529	26324

It can be seen that the price of structures does not behave in a monotonic manner but after dipping 5% in quarter 6, it trends up to finish about 10% higher at the end of the sample period as compared to the beginning of the sample period. The variance of the land price series was much higher. The price of land peaked in Quarter 15, approximately 47% higher than the Quarter 1 level and then it generally trended downwards to finish 31% higher in Quarter 22. The results for this model look very reasonable since we expect the price of land to fluctuate much more than the price of structures.



Note to Jan: the basic data use to make up the above chart are in the following Table:

1	**1**	**1**	**1**	**1**
1.0115	**1.0056**	**1.00842**	**0.99248**	**1.01613**
0.97511	**0.99438**	**0.99769**	**0.99248**	**1**
1.01626	**1.01036**	**1.01035**	**1.04399**	**0.99194**
1.03964	**1.03968**	**1.04007**	**1.14791**	**0.98387**
1.05462	**1.04623**	**1.04554**	**1.20958**	**0.95968**
1.06757	**1.05326**	**1.05593**	**1.22438**	**0.96774**
1.04559	**1.03651**	**1.04056**	**1.1116**	**1**
1.05259	**1.05271**	**1.05818**	**1.20134**	**0.98387**
1.10079	**1.0976**	**1.10428**	**1.359**	**0.9769**
1.12179	**1.11485**	**1.12097**	**1.36491**	**0.99881**

1.08897	**1.09171**	**1.09813**	**1.24923**	**1.02271**
1.10521	**1.09851**	**1.10504**	**1.33155**	**0.99084**
1.13606	**1.12687**	**1.13646**	**1.4058**	**1.0008**
1.15825	**1.14963**	**1.15492**	**1.47191**	**0.99582**
1.12513	**1.11129**	**1.11711**	**1.35274**	**0.99881**
1.15603	**1.15468**	**1.16136**	**1.44763**	**1.01773**
1.15751	**1.1452**	**1.1498**	**1.39479**	**1.02769**
1.17844	**1.16228**	**1.1677**	**1.40183**	**1.05159**
1.16364	**1.14581**	**1.15549**	**1.32049**	**1.07449**
1.14472	**1.13869**	**1.14825**	**1.2561**	**1.0954**
1.16987	**1.15847**	**1.16627**	**1.31144**	**1.0954**

Chart 3 plots the price of land P_{L3} and structures P_{S3} for Model 3 along with the overall house price index generated by this model, P_3. We also plot the overall house price indexes generated by Models 1 and 2, P_1 and P_2, and compare these indexes with P_3. It can be seen that P_1, P_2 and P_3 can barely be distinguished as separate series in Chart 3.[32]

Although the present model seems satisfactory, in the following section, we explore how the model can be improved by using additional information on housing characteristics.

5. Model 4: The Use of Additional Characteristics Information

In the last two models, we made use of the fact that large lots are likely to have a lower price per meter squared than medium lots. By modeling this empirical regularity with the use of splines on the quantity of land, we were able to improve the fit of the regression. It is also likely that larger structures have a higher quality than small structures; i.e., larger houses are likely to use more expensive construction materials than smaller houses. Thus it seems likely that using the same type of spline setup, but on S rather than L, we could improve the fit in our regression model. However, a more parsimonious alternative to using spline techniques on structures is to use information on the number of rooms in the structure; i.e., as the number of rooms increases, we would expect the quality of the structure to increase so that the price per meter squared of a structure should increase as the number of rooms increases.[33] However, it should be noted that some housing experts believe that the price should decline as the structure size increases so the issue is not settled.[34]

[32] We ran a wide variety of hedonic regressions using the same price and characteristics data but different functional forms for the various regressions and found that they all fitted the overall price data fairly well and generated similar *overall* housing price indexes. However, these various models did not generate reasonable subcomponent land and structures price indexes.

[33] The correlation coefficient between the room variable R and the structure area S (not adjusted for depreciation) is 0.4746, somewhat lower than we anticipated.

[34] Palmquist (1984; 397) is one such expert: "It would be anticipated that the number of square feet of living space would not simply have a linear effect on price. As the number of square feet increases, construction costs do not increase proportionally since such items as wall area do not typically increase proportionally. Appraisers have long known that price per square foot varies with the size of the house." The empirical results of Coulson (1992; 77) on this issue indicate a great deal of volatility in price but for

Our regression *Model 4* is defined by equations (5)-(7) again except that the terms involving the quantity of structures, $\beta^t(1 - \delta A_n^t)S_n^t$ in each of the equations (5)-(7), are now replaced by the terms $\beta^1 p^t(1 - \delta A_n^t)(1 + \gamma R_n^t)S_n^t$ where β^1, δ and γ are parameters to be estimated, p^t is the Statistics Netherlands New Dwelling Construction Cost Price Index for quarter t described in the previous section, A_n^t is the age in decades of property n in quarter t, R_n^t is the number of rooms less 4 for property n in quarter t and S_n^t is the area of structure n in quarter t. Note that A_n^t is equal to 0 if property n sold in quarter t is a new house and that R_n^t is equal to 0 if property n sold in quarter t has 4 rooms. In order to identify the parameters β^1, δ and γ, we need the exogenous characteristics variables A_n^t and R_n^t to take on the value 0 for at least some observations (and the 0 values should not occur for exactly the same observations). Note that if γ equals 0, then the present model reduces to Model 3 in the previous section. Thus the present model has 69 parameters compared to the 68 parameters for Model 3. A priori, we expect the new parameter γ to be positive; i.e., as the number of rooms increases, we expect the price per m^2 of construction to also increase.

The R^2 for this model was .8736, an increase from the previous Model 3 R^2 of .8723. The log likelihood was −16222.6, a substantial increase of 17.1 over the previous Model 3 for the addition of only one new parameter, the room size parameter γ. The estimated decade depreciation rate was $\delta^* = 0.1089$ (0.00361). The first period parameter values for the 3 marginal prices for land were $\alpha_S^{1*} = 0.2207$ (0.0249), $\alpha_M^{1*} = 0.3465$ (0.0560) and $\alpha_L^{1*} = 0.1741$ (0.0307). The first period parameter value for quality adjusted structures was $\beta^{1*} = 1.0069$ (0.0212) or 1006.9 Euros/m^2. Note that this is the estimated construction cost for a new building (per meter squared) with four rooms in Quarter 1. Thus this new estimated Q1 building cost is not comparable to the Q1 building costs estimated by the previous model, since the earlier estimates applied to all houses irrespective of the number of rooms, which ranged from 2 to 14. The smallest t statistic was 4.64 for α_M^{3*} so that all parameters were significantly different from 0. The estimated number of rooms parameter was $\gamma^* = 0.02759$ (0.00493). Thus the estimated increase in the price of a new structure per m^2 in Quarter 1 due to an additional room is 0.02759/1.0069, which equals 2.74%. Thus the average premium in construction costs per m^2 in Quarter 1 of a 10 room house over a 2 room house is 2.74% times 8, which is 21.9% per m^2. This seems to be a reasonable quality premium.

As usual, we used equations (8)-(14) in order to construct a chained Fisher index of land prices, which we denote by P_{L5}. This index is plotted in Chart 4 and listed in Table 4 below. The estimated quarter t price for a square meter of quality adjusted structures for a four room house is $\beta^{t*} \equiv \beta^{1*} p^t$ and we use this price series as our constant quality price series for structures. The corresponding constant quality quarter t quantity of structures is

large structures, the price of structure per unit area trended up fairly strongly for his sample of U.S. properties.

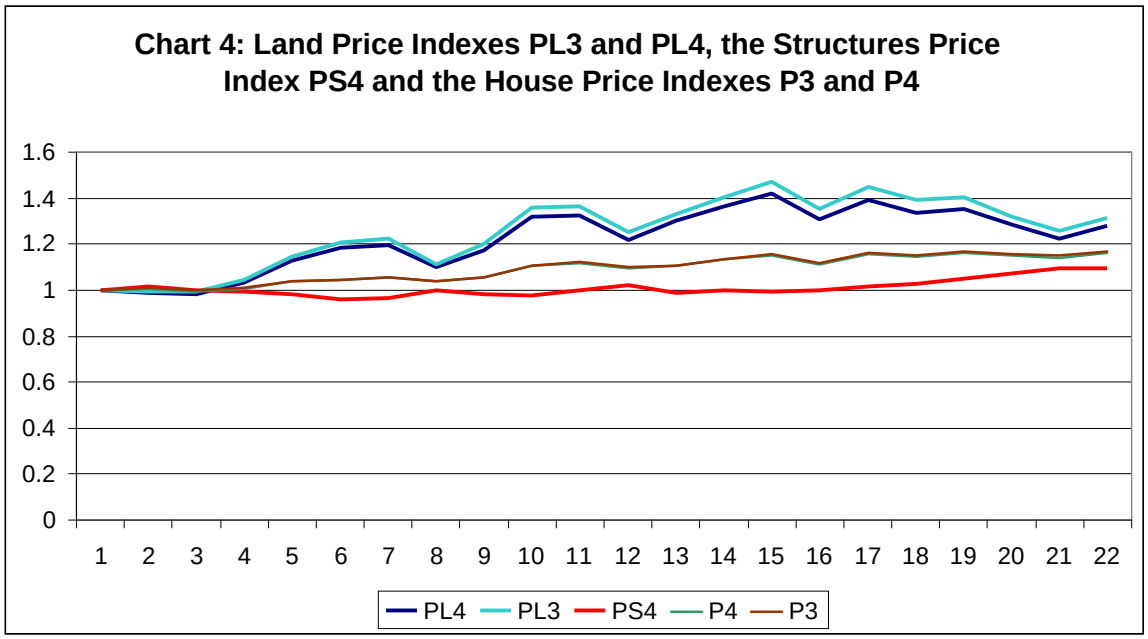
$S^{t*} \equiv \sum_{n=1}^{N(t)} (1 - \delta^* A_n^t)(1+\gamma^* R_n^t) S_n^t$.[35] The structures price and quantity series β^{t*} and S^{t*} were combined with the three land price and quantity series to form a chained *overall Fisher house price index* P_4 which is graphed in Chart 4 and listed in Table 4. The *constant quality structures price index* P_{S4} (a normalization of the series $\beta^{1*},...,\beta^{22*}$) is also found in Chart 4 and Table 4.

Table 4: The Price of Land P_{L4}, the Price of Structures P_{S4} and the Overall House Price Index P_4 Generated by Model 4 with the Corresponding Quantities Q_{L4}, Q_{S4} and Q_4

Quarter	P_{L4}	P_{S4}	P_4	Q_{L4}	Q_{S4}	Q_4
1	1.00000	1.00000	1.00000	8372.8	14816.2	23189.0
2	0.98919	1.01613	1.00626	8499.4	14218.0	22712.6
3	0.98251	1.00000	0.99362	9540.5	14929.4	24459.1
4	1.03180	0.99194	1.00760	10215.5	17005.8	27202.2
5	1.12890	0.98387	1.03902	8980.2	14954.4	23917.7
6	1.18484	0.95968	1.04555	10954.2	17004.3	28021.2
7	1.19793	0.96774	1.05555	8001.5	13397.9	21364.3
8	1.10152	1.00000	1.04067	9690.9	17363.8	26942.9
9	1.17454	0.98387	1.05632	9263.9	16952.2	26090.3
10	1.31868	0.97690	1.10370	9171.9	16053.3	25167.3
11	1.32326	0.99881	1.11928	9385.7	16035.6	25405.8
12	1.21947	1.02271	1.09563	9832.3	16368.2	26222.3
13	1.30263	0.99084	1.10718	10380.5	17003.3	27429.8
14	1.36153	1.00080	1.13530	11027.1	19376.0	30305.0
15	1.41932	0.99582	1.15332	9406.9	16109.7	25486.2
16	1.30854	0.99881	1.11409	9591.6	16114.3	25712.5
17	1.39053	1.01773	1.15633	9544.0	16562.4	26054.1
18	1.33811	1.02769	1.14266	13605.8	21006.0	34825.5
19	1.35373	1.05159	1.16328	8590.0	13876.1	22540.2
20	1.28629	1.07449	1.15240	11516.1	19960.1	31464.8
21	1.22226	1.0954	1.14219	9075.5	15670.2	24739.9
22	1.28276	1.0954	1.16410	11009.9	19980.8	30933.9
Mean	1.2236	1.0114	1.0906	9802.6	16580	26372

[35] Thus we are implicitly quality adjusting the quantities of houses with different room sizes into "standard" houses with four rooms using the quality adjustment factors $\gamma^* R_n^t$ for house n in quarter t. Thus we are forming a hedonic structures aggregate. Alternatively, instead of forming a quality adjusted aggregate, we could distinguish houses with differing number of rooms as separate types of housing and use index number theory to aggregate the 13 types of house into a structures aggregate. In this second interpretation, the quarter t structure price $\beta^{t*} = \beta^{1*} p^t$ applies to a new house with 4 rooms. The appropriate price (per m^2) for a new house with 5, 6, ..., 14 rooms would be $\beta^{1*} p^t(1+\gamma^*)$, $\beta^{1*} p^t(1+2\gamma^*)$, ... , $\beta^{1*} p^t(1+10\gamma^*)$ and the price for a new house with 2 and 3 rooms would be $\beta^{1*} p^t(1-2\gamma^*)$ and $\beta^{1*} p^t(1-\gamma^*)$. Thus in this second approach, we distinguish 13 types of house (according to their number of rooms) and calculate separate price and quantity series for all 13 types (adjusted for depreciation as well). However, if we then aggregate these series using Laspeyres, Paasche or Fisher indexes, we would find that the resulting aggregate structures price index would be proportional to the $\beta^{1*} p^t$ series. Thus the second method is equivalent to the first method.

It can be seen that the structures price series P_{S44} coincides with the structures price series P_{S3} for the previous model. This makes sense because both models impose the same rates of change on quality adjusted structures prices (equal to the Statistics Netherlands rates of change). Thus in Chart 4, we do not plot separately P_{S3} and P_{S4} since they are identical series.



From viewing Chart 4, it can be seen that our new model that allows for a quality adjustment for the construction of larger houses generates a somewhat different series for the price of land as compared to Model 3; i.e., P_{L4} lies below P_{L3} for Quarters 2-22. Note that the overall house price indexes, P_3 and P_4, are virtually identical[36]; i.e., they are difficult to distinguish in Chart 4.[37]

Recall that before running any regressions, we eliminated some outlier observations that had prices or characteristics which were either very large or very small relative to average prices and average amounts of characteristics. However, running the regressions associated with Models 1-4, there were additional outliers (i.e., observations with large error terms), which were not deleted. This non deletion of regression outliers could affect our estimated coefficients, particularly if the outliers are either mostly positive or mostly

[36] The correlation coefficient between P_3 and P_4 is .99942.

[37] If P_3 almost equals P_4 and P_{S3} is exactly equal to P_{S4}, one might ask how can P_{L3} and P_{L4} differ so much? The answer is that while the rates of growth in the *price* of constant quality structures is the same in Models 3 and 4, the addition of the quality adjustment for the number of rooms has changed the initial level (and rates of growth) for the constant quality *quantity* of structures. Using Model 3, the initial levels of land and constant quality structures were 7446.9 and 15749.3. Using Model 4, the initial levels of land and constant quality structures were 8372.8 and 14816.2. Thus going from Model 3 to 4, the value of Q1 land has increased about 12.4% and the value of structures has decreased to offset this increase. Since land prices increase more rapidly than structure prices and since the overall indexes P_3 and P_4 are virtually equal and the structures indexes P_{S3} and P_{S4} are exactly equal, it can be seen that these facts will imply that P_{L4} must grow more slowly than P_{L3}.

negative. To determine whether outliers are a problem with Model 4, we looked at the empirical distribution of the resulting error terms for this model. We constructed 10 error intervals: $e_n^t < -100$[38]; $-100 \leq e_n^t < -75$; $-75 \leq e_n^t < -50$; ... ; $75 \leq e_n^t < 100$; $100 \leq e_n^t$. The number of observations that fell into these 10 bins was as follows: 9, 10, 57, 333, 1358, 1297, 319, 64, 34 and 6. Thus the empirical distribution of error terms appears to be fairly symmetric with a relatively small number of very large in magnitude errors.

Our conclusion at this point is that Model 4 is a satisfactory hedonic housing regression model that decomposes house prices into sensible land and structures components. The quality adjustments to the quantity of structures for the age of the structure and for the number of rooms also seem to be reasonable. The overall fit of the model also seems to be satisfactory: an R^2 of .8736 for such a small number of characteristics is quite good.[39]

The builder's model that we developed here could be further modified to take into account additional characteristics but a certain amount of careful thought is required so that the effects of introducing additional characteristics reflect the realities of housing construction and locational effects.[40] These construction realities will determine the appropriate functional form for the hedonic regression.

6. Conclusion

A number of tentative conclusions can be drawn from this study:

- If we stratify housing sales by local area and type of housing and if we have data on the age of the dwelling unit, its land plot area (or share of the plot area in the case of multiple unit dwellings) and its floor space area, then a wide variety of hedonic regression models that use these variables seem to generate much the same *overall* house price indexes.
- It is much more difficult to obtain sensible land and structure price indexes by means of a hedonic regression. However, our builder's model, in conjunction with statistical agency information on the price movements of new dwelling units, generated satisfactory results for our data set.
- Adding the number of rooms in the dwelling unit as an explanatory variable in our hedonic regressions did improve the fit but did not change the indexes substantially.
- Splining land also improved the fit of our hedonic regressions and led to somewhat smoother land price indexes in our best builder's model.
- It is important to delete observations in the regressions which are range outliers.

[38] Thus if an observation belonged to this bin, the associated error term was less than −100,000 Euros; recall that we measure house prices in thousands of Euros when running our regressions.

[39] However, the Dutch data may not be representative of other data sets where there could be more heterogeneity due to geography or differences in the types of houses being built over time.

[40] In particular, the number of stories in the dwelling unit is likely to be a significant quality adjustment characteristic: a higher number of stories (holding structural area constant) is likely to lead to lower building costs due to shared floors and ceilings and less expenditures on roofing and insulation. A larger number of stories could also have a quality adjustment effect on the land component of the dwelling unit since a higher number of stories leads to more usable yard space.

Some topics for follow up research include the following:

- Can our method be generalized to deal with the sales of condominiums and apartment units with shared land and facilities?
- How exactly can other characteristics be used in more general versions of the builder's model?

References

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

Clapp, J.M. (1979), "The Substitution of Urban Land for Other Inputs", *Journal of Urban Economics* 6, 122-134.

Clapp, J.M. (1980), "The Elasticity of Substitution for Land: The Effects of Measurement Errors", *Journal of Urban Economics* 8, 255-263.

Coulson, N.E. (1992), "Semiparametric Estimates of the Marginal Price of Floorspace", *Journal of Real Estate Finance and Economics* 5, 73-83.

Court, A.T. (1939), "Hedonic Price Indexes with Automotive Examples", pp. 99-117 in *The Dynamics of Automobile Demand*, New York: General Motors Corporation.

Crone, T.M., L.I. Nakamura and R.P. Voith (2009), "Hedonic Estimates of the Cost of Housing Services: Rental and Owner Occupied Units", pp. 67-84 in *Price and Productivity Measurement, Volume 1: Housing*, W.E. Diewert, B.M. Balk, D. Fixler, K.J. Fox and A.O. Nakamura (eds.), Trafford Press.

Davis, M.A. and J. Heathcote (2007), "The Price and Quantity of Residential Land in the United States", *Journal of Monetary Economics* 54. 2595-2620.

Davis, M.A. and M.G. Palumbo (2008), "The Price of Residential Land in Large US Cities", *Journal of Urban Economics* 63, 352-384.

Diewert, W.E. (2003), "Hedonic Regressions: A Consumer Theory Approach", pp. 317-348 in *Scanner Data and Price Indexes*, R.C. Feenstra and M.D. Shapiro (eds.), Studies in Income and Wealth 64, Chicago: University of Chicago.

Diewert, W.E. (2007), "The Paris OECD-IMF Workshop on Real Estate Price Indexes: Conclusions and Future Directions", Discussion Paper 07-01, Department of Economics, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z1.

Diewert, W.E. (2010), "Alternative Approaches to Measuring House Price Inflation", Discussion Paper 10-10, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1.

Diewert, W.E., J. de Haan and R. Hendriks (2010), "The Decomposition of a House Price Index into Land and Structures: A Hedonic Regression Approach", Discussion Paper 10-01, Department of Economics, University of British Columbia, Vancouver, Canada, V6T1Z1.

Diewert, W.E., S. Heravi and M. Silver (2009), "Hedonic Imputation versus Time Dummy Hedonic Indexes", pp. 161-196 in *Price Index Concepts and Measurement*, W.E. Diewert, J.S. Greenlees and C.R. Hulten (eds.), Studies in Income and Wealth 70, Chicago: University of Chicago Press.

Diewert, W.E., A.O. Nakamura and L.I. Nakamura (2009), "The Housing Bubble and a New Approach to Accounting for Housing in A CPI", *Journal of Housing Economics* 18, 156-171.

Fleming, M.C. and J.G. Nellis (1992), "Development of Standardized Indices for Measuring House Price Inflation Incorporating Physical and Locational Characteristics", *Applied Economics* 24, 1067-1085.

Francke, M.K. (2008), "The Hierarchical Trend Model", pp. 164-180 in *Mass Appraisal Methods: An International Perspective for Property Valuers*, T. Kauko and M. Damato (eds.), Oxford: Wiley-Blackwell.

Francke, M.K. and G.A. Vos (2004), "The Hierarchical Trend Model for Property Valuation and Local Price Indices", *Journal of Real Estate Finance and Economics* 28:2/3, 179-208.

Glaeser, E.L. and J. Gyourko (2003), "The Impact of Building Restrictions on Housing Affordability", *Economic Policy Review* 9, 21-39.

Gouriéroux, C. and A. Laferrère (2009), "Managing Hedonic House Price Indexes: The French Experience", *Journal of Housing Economics* 18, 206-213.

Gyourko, J. and A. Saiz (2004), "Reinvestment in the Housing Stock: The Role of Construction Costs and the Supply Side", *Journal of Urban Economics* 55, 238-256.

Haan, J. de (2008), "Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Other Approaches", Centre for Applied Economic Research Working Paper 2008/01, Faculty of Economics and Commerce, University of New South Wales, Sydney, Australia.

Haan, J. de (2009), "Comment on Hedonic Imputation versus Time Dummy Hedonic Indexes", pp. 196-200 in *Price Index Concepts and Measurement*, W.E. Diewert, J.S. Greenlees and C.R. Hulten (eds.), Studies in Income and Wealth 70, Chicago: University of Chicago Press.

Haan, J. de and H. van der Grient (2011), "Eliminating Chain Drift in Price Indexes based on Scanner Data", *Journal of Econometrics* 161, 36-46.

Hill, R.J. (2011), "Hedonic Price Indexes for Housing", Statistics Directorate, Working Paper No 36, February 14, Paris: OECD.

Hill, R.J., D. Melser and I. Syed (2009), "Measuring a Boom and Bust: The Sydney Housing Market 2001-2006", *Journal of Housing Economics* 18, 193-205.

Ivancic, L., W.E. Diewert and K.J. Fox (2011), "Scanner Data, Time Aggregation and the Construction of Price Indexes", *Journal of Econometrics* 161, 24-35.

Koev, E. and J.M.C. Santos Silva (2008), "Hedonic Methods for Decomposing House Price Indices into Land and Structure Components", unpublished paper, Department of Economics, University of Essex, England, October.

McDonald, J.F. (1981), "Capital-Land Substitution in Urban Housing: A Survey of Empirical Estimates", *Journal of Urban Economics* 9, 190-211.

McMillen, D.P. (2003), "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.

Muellbauer, J. (1974), "Household Production Theory, Quality and the 'Hedonic Technique'", *American Economic Review* 64, 977-994.

Muth, R.F. (1971), "The Derived Demand for Urban Residential Land", *Urban Studies* 8, 243-254.

Palmquist, R.B. (1984), "Estimating the Demand for the Characteristics of Housing", *The Review of Economics and Statistics* 66:3, 394-404.

Rambaldi, A.N., R.R.J McAllister, K. Collins and C.S. Fletcher (2010), "Separating Land from Structure in Property Prices: A Case Study from Brisbane Australia", School of Economics, The University of Queensland, St. Lucia, Queensland 4072, Australia.

Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy* 82, 34-55.

Rosen, H.S. (1978), "Estimating Inter-City Differences in the Price of Housing Services", *Urban Studies* 15, 351-355.

Schwann, G.M. (1998), "A Real Estate Price Index for Thin Markets", *Journal of Real Estate Finance and Economics* 16:3, 269-287.

Shimizu, C., K.G. Nishimura and T. Watanabe (2010), "Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures", *Journal of Economics and Statistics* 230/6, 792-813.

Shimizu, C., H. Takatsuji, H. Ono and Nishimura (2010), "Structural and Temporal Changes in the Housing Market and Hedonic Housing Price Indices", *International Journal of Housing Markets and Analysis* 3:4, 351-368.

Statistics Portugal (Instituto Nacional de Estatistica) (2009), "Owner-Occupied Housing: Econometric Study and Model to Estimate Land Prices, Final Report", paper presented to the Eurostat Working Group on the Harmonization of Consumer Price Indices", March 26-27, Luxembourg: Eurostat.

Thorsnes, P. (1997), "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.