The Hierarchical Trend Model for Real Estate Valuation and Price Indices

by

Marc K. Francke (mkfrancke@wxs.nl) and G.A.Vos (gjavos@fee.uva.nl)

Vrije Universiteit Amsterdam,

University of Amsterdam

Abstract

This article presents a Hierarchical Trend Model for selling prices of houses, addressing two main problems: the spatial and temporal dependence of selling prices. In this model cluster-level trends, a general trend, and specific characteristics play a role. In this set-up every cluster, a combination of neighbourhood and house type, has his own price development.

This dynamic hedonic model is used for real estate valuation, and for determining local price indices. Two applications are provided, one for the Breda region, and one for the Amsterdam region. For these regions price indices based on weighted median selling prices are compared with the quality adjusted price indices from the Hierarchical Trend Model. It is shown for both housing market regions that the hedonic approach produces price indices that are more accurate, detailed, and more up-to-date.

1 INTRODUCTION

This article concerns the modeling of selling prices of houses. A hedonic price model is presented, developed for the determination of real estate taxes in Amsterdam. Besides the size, and the location of the house, in a time of rapid price movements the selling date is an important characteristic to explain selling prices. A Hierarchical Trend Model (HTM) is presented, addressing two main problems: the spatial and temporal dependence of selling prices. In this hedonic price model cluster-level trends, a general trend, and specific characteristics play a role. Together with the influence of the specific characteristics these trends are estimated within the HTM. The cluster-level trends are specified as deviations from the general trend. The general, and cluster-level trends are modelled as stochastic trends, so no a priori structure is assumed. The clusters, or market segments can be defined by for example neighborhoods, and dwelling types. In this set-up it is possible that every market segment has a different price development.

Sections 2 - 5 concern model specification. In section 2 the functional form of the dependent variable is motivated. In section 3 the choice of the functional form of some of the explanatory variables, like lot and house size, is discussed. Section 4 adds the spatial model component. In section 5 the final HTM is provided, including the temporal component.

This model can be used for real estate valuation. Given the characteristics of the house the model is able to produce values for all time points in the time period considered. At this moment this model is operational in Amsterdam for the determination of real estate taxes.

Another application is provided by the determination of price indices. Changes in the levels of selling prices can be caused by changes in the underlying characteristics of the houses. For this reason selling price levels in one period cannot be directly compared with selling price levels in another period, but the levels must be adjusted for differences in house characteristics. It is shown that the estimated trends from the HTM provide the correct measurement of house price movements, and levels over time.

In this article the HTM is estimated on two datasets over the period 1985-1999, both from the Dutch Broker Organization (NVM). The first dataset contains selling prices for the Breda region, the second database selling prices for the Amsterdam region[1]. Section 6 provides a brief description of both datasets.

Section 7 provides some model results for the Breda region. In section 8 price indices are shown for both the Breda and Amsterdam region. The indices produced by the HTM are compared with indices from an often-used method that simply consists of taking averages, or medians for every cluster. It is shown that the hedonic approach produces price indices that are more accurate, detailed, and more up-to-date. For both methods standard deviations are compared for price indices for the region as a whole as well as per neighborhood, and house type within the region, both on a monthly, quarterly, and a yearly basis. Section 9 concludes with the key results.

2 DEPENDENT VARIABLE

In this section a motivation is provided for the specification of the dependent variable. The dependent variable is the selling price, or a transformation of the selling price. Examples of transformations are the square root, and the natural logarithm of the selling price. The Box-Cox method is often used as a guideline to choose a specific transformation, see for example Halvorsen and Pollawski (1979). Let Y_i denote the selling price of sale i for $i = 1, \ldots, n$. The Box-Cox transformation is given by

$$Y_i(\theta) = \frac{(Y_i)^\theta - 1}{\theta}.$$

For $\theta = 1$, the dependent variable is the selling price, for $\theta = \frac{1}{2}$, the dependent variable is the square root of the selling price, and for $\theta \to 0$, the dependent variable is the natural logarithm of the selling price. In general θ is unknown, and along with the coefficients of

[1] This is another database as is used for the determination of real estate taxes. The last dabase consists of all selling prices of Amsterdam only. The NVM database consists of selling prices for the Amsterdam region, including a few other cities. The NVM has a market share of approximately 60%.

the explanatory variables, it needs to be estimated.

We did not use the Box-Cox analysis to choose a transformation. We use the natural logarithm of the selling price as dependent variable. The reason for this that we assume that variables for neighborhoods and trends, work in a multiplicative way on the size of the house, see section 3. Another reason for this is that our goal is to minimize the relative standard deviation. This can also be done in the more general cases of a Box-Cox transformation, but that is more complex to evaluate. An additional assumption of the natural logarithm is that the error terms have a lognormal density, that can be checked by evaluating the residuals.

Let y_i denote the natural logarithm of Y_i, so $y_i = \ln Y_i$. It is assumed that all selling prices $Y_i > 0$. The model is given by

$$y = X\beta + \varepsilon, \tag{1}$$

and $\varepsilon \sim N(0, \sigma^2 I_n)$.

Because of the logarithmic specification of the dependent variable, the standard deviation can be interpreted as a relative standard deviation. Let e denote the vector of residuals, so $e = y - X\widehat{\beta}$, with $\widehat{\beta}$ the Ordinary Least Squares (OLS) estimator of β. Let M_i denote the model value, so $M_i = \exp(X_i\widehat{\beta})$, then

$$y_i - X_i\widehat{\beta} = \ln Y_i - \ln M_i = \ln(1 + \frac{Y_i - M_i}{M_i}) \approx \frac{Y_i - M_i}{M_i},$$

due to the fact that $\ln(1 + \varepsilon) \simeq \varepsilon$, for small ε. If the residuals are not too big, they can be interpreted as relative errors, so the standard deviation from the residuals can be interpreted as a relative standard deviation.

In the standard linear model the residual sum of squares, and hence the standard deviation is minimized. This means that in the logarithmic specification of the dependent variable the relative errors $(Y - M)/M$ are approximately minimized. If the dependent variable is the selling price, the absolute errors $(Y - M)$ are minimized. In the first case an error of £10.000 on a selling price of £100.000 has a greater impact on the standard deviation than an error of £10.000 on a selling price of £1.000.000. In the last case both errors have the same impact on the standard deviation of the residuals.

In (1), $E\left[y_i - X_i\widehat{\beta}|y,\sigma^2\right] = 0$. The variance is given by $var(y_i - X_i\widehat{\beta}|y,\sigma^2) = \tau_i$, with $\tau_i = X_i var(\widehat{\beta})X_i' + \sigma^2$, and $var(\widehat{\beta}) = \sigma^2(X'X)^{-1}$. The exponent of the model residuals are of interest. The expectation, and variance are given by[2]

$$\begin{aligned} E\left[\exp\left\{y_i - X_i\widehat{\beta}\right\}|y,\sigma^2\right] &= \exp\left(\tau_i/2\right), \\ var\left(\exp\left\{y_i - X_i\widehat{\beta}\right\}|y,\sigma^2\right) &= \exp\tau_i\left(\exp\tau_i - 1\right), \end{aligned}$$

Even in the case that $var(\widehat{\beta}) = 0$, the expectation is greater than 1, because in general σ^2 is not zero. For example, if $\sigma = 0.15$, $\sigma^2 = 0.0225$, and $\exp\left\{\sigma^2/2\right\} \simeq 1.01$. So, a standard deviation of 0.15 leads to over valuation of about 1 percent. So, in order to obtain an expected value of 1, for the ratio between actual and model value all model values can be corrected by a factor $\exp\left(-\tau_i/2\right)$.

3 MULTIPLICATIVE/ADDITIVE MODEL

3.1 Model specification

The internal floorspace is an important characteristic for explaining the selling price of a dwelling. If the natural log of the selling price is used as dependent variable, and the natural log of the internal floorspace (x_1) as independent variable, then

$$y = \beta_1 \ln x_1 + Z\delta + \varepsilon,$$

with z the other explanatory variables. In this specification an increase of x_1 by 1 percent, will result in an increase of Y of approximately β_1 percent. It is expected that $\beta_1 < 1$, so the value will be less than proportional with the internal floorspace.

[2] In general, if $y \sim N(\mu,\sigma^2)$ and $Y = \exp(y)$, then $E[Y] = \exp\left(\mu + \sigma^2/2\right)$, and $var(Y) = \exp(2\mu + \sigma^2)\left(\exp\sigma^2 - 1\right)$. So, in y the expectation and mode coincide, in Y the mode is smaller than the expectation.

Another important characteristic is the lot size (x_2). If the natural log of the lot size is added as independent variable, then

$$Y = {x_1}^{\beta_1} {x_2}^{\beta_2} \exp(Z\delta + \varepsilon).$$

So, in this example (a power of) the internal floorspace is multiplied by (a power of) the lot size. This is an undesirable feature of this model: we expect the mutual influence of floorspace and lot size to be additive, rather than multiplicative. For that reason we change the model specification. Let the $k \times 1$ vector β denote the coefficients of the additive variables $X = \begin{bmatrix} x_1 & \cdots & x_k \end{bmatrix}$, and Z all other variables, then

$$y = \alpha \ln(X\beta) + Z\delta + \varepsilon. \tag{2}$$

Note that the coefficient β_1 for x_1 is 1, because otherwise in case that in Z a constant is included, the model is not identified. If we take the exponent for this model, we get

$$Y = (X\beta)^{\alpha} \exp(Z\delta + \varepsilon), \tag{3}$$

so this model is additive in X.

The variables Z influence as a factor both the lot size and internal floorspace, as is shown from (3). This is a desirable feature for variables concerning time trends, and the influence of the neighborhood. For variables like the age of the building, and the maintenance this specification is undesirable, because these variables in practice only influence the value of the floorspace, and in the model specification (2) also the lot size.

3.2 Estimation

The model (2) cannot be estimated by OLS, because it is nonlinear in β. It is quite easy to linearize (2) by using the approximation $\ln(1+\varepsilon) \simeq \varepsilon$, for small ε. Define $x(j) = \sum_{i=1}^{k} x_{ij}\beta_i$, and $x^*(j) = \sum_{i=1}^{k} x_{ij}\beta_i^*$ for some β^*, with $\beta_1^* = 1$. So the index j denotes observation j. We can write $\alpha \ln x(j)$ as

$$\begin{aligned}\alpha \ln x(j) &= \alpha \left[\ln x^*(j) + \ln \left(1 + \sum_{i=2}^{k} \frac{\beta_i - \beta_i^*}{x^*(j)} x_{ij} \right) \right] \\ &\simeq \alpha \left[\ln x^*(j) + \sum_{i=2}^{k} \frac{\beta_i - \beta_i^*}{x^*(j)} x_{ij} \right] \\ &= \alpha \left(\ln x^*(j) - \frac{x^*(j) - x_{1j}}{x^*(j)} \right) + \sum_{i=2}^{k} \alpha \beta_i \frac{x_{ij}}{x^*(j)}.\end{aligned}$$

So, model (2) can be approximated by

$$y_j = \alpha \left(\ln x^*(j) - \frac{x^*(j) - x_1}{x^*(j)} \right) + \sum_{i=2}^{k} \theta_i \frac{x_{ij}}{x^*(j)} + (Z\delta)_j + \varepsilon_j, \tag{4}$$

with $\theta_i = \alpha\beta_i$, for $i = 2, \ldots, k$. This model can be estimated as follows:

1. Choose some β^* such that $|\beta_i - \beta_i^*|$ is small,
2. Calculate x^*,
3. Estimate (4) by OLS, this provides estimates $\widehat{\alpha}$, and $\widehat{\theta}_i$, so $\widehat{\beta}_i = \widehat{\theta}_i / \widehat{\alpha}$.
4. Substitute β^* with $\widehat{\beta}$, and repeat $1 - 3$, until $|\beta_i - \beta_i^*| \simeq 0$.

In general this process will converge quickly. A more general approach is provided by Gauss-Newton regression, see for example Davidson en MacKinnon (1993). Consider the model $y = x(\beta) + \varepsilon$, with $x(\beta)$ some nonlinear function in β. Let $\dot{x}(\beta) = \partial x(\beta)/\partial\beta$. The first order Taylor expansion of this model around β^* is provided by

$$\begin{aligned} y &\simeq x(\beta^*) + \dot{x}(\beta^*)(\beta - \beta^*) + \varepsilon, \\ y - x(\beta^*) &= \dot{x}(\beta^*) b + \varepsilon \end{aligned} \tag{5}$$

For observation j in (2) without $Z\delta$, $\dot{x}_j(\beta)$ is provided by

$$\dot{x}_j(\beta)' = \begin{pmatrix} \ln(X\beta) \\ \alpha\beta_2 (\sum_i x_{ij}\beta_i)^{-1} x_{2j} \\ \vdots \\ \alpha\beta_k (\sum_i x_{ij}\beta_i)^{-1} x_{kj} \end{pmatrix}.$$

(5) can be estimated by OLS. The estimate of b must equal 0. The OLS must be done recursively, and using gradients and Hessians can speed up convergence (see Chapter 6.8 in Davidson and MacKinnon (1993)). Stop criteria are based on t-statistics.

3.3 Linearization of the model value

Real estate agents and valuers are often interested in prices per square or cubic meters of floorspace and lot size. It is possible to linearize (2) in X, so the model value M_j can be written as $M_j = \sum_{i=1}^{k} x_{ij}\phi_{ij}$. From (2) it follows that

$$\begin{aligned} M_j &= c\left(\sum_{i=1}^{k} x_{ij}\beta_i\right)^{\alpha-1}\left(\sum_{i=1}^{k} x_{ij}\beta_i\right) \\ &= \sum_{i=1}^{k}\left[\frac{c\beta_i}{\left(\sum_{i=1}^{k} x_{ij}\beta_i\right)^{1-\alpha}}\right] x_{ij} \\ &= \sum_{i=1}^{k} x_{ij}\phi_{ij}, \\ \phi_{ij} &= \frac{\beta_i}{\sum_{i=1}^{k} x_{ij}\beta_i} M_j, \end{aligned}$$

with $c = \exp(z_j\delta)$, with z_j row j of Z.

4 INCORPORATION OF SPACE

One of the most important influences on housing prices is location. Two dwellings with the same characteristics, but in different neighborhoods, could have very different market values. This can be described by varying constants over neighborhoods, so dummy variables have to be introduced. Another example that shows the influence of location is, that the influence of a characteristic varies over neighborhoods. For example, the influence of a garage on the value of a dwelling is much bigger in the city center than in a suburb. In the model specification this means that the coefficients β vary over location. A last example that shows the influence of location on values is the simple fact that the value of a house is dependent on the values of adjacent houses. In model specification this means that a specific spatial covariance structure is assumed.

From spatial econometrics two notions are known, spatial heterogeneity and spatial dependence, see for example Anselin (1988). Spatial heterogeneity can be described as follows: functional forms and parameters vary with location and are not homogeneous throughout the data set. And spatial dependence: the variation is a function of distance.

Spatial models for housing prices can be specified on an individual level (observation) and on a cluster level, for example neighborhood, or city level. Spatial models on an individual level are complex to evaluate. Examples of such spatial models are provided by for example Can (1992), and Dubin (1992,1998). An example of a spatio-temporal model is given in Pace et all (1998). Those models are not considered here.

Another choice is to specify the spatial component on a cluster level: spatial heterogeneity and spatial dependence on a cluster level. A drawback of this approach is that there might be undesirable discontinuities on borders, and it requires knowledge of the spatial structure, which might be different from available administrative clusters.

In the next subsection some clusterlevel models with spatial heterogeneity are described. In the second subsection a model containing spatial dependence is provided. In both subsections a fixed effects model, a random effects model, and a hierarchical model are provided.

4.1 Spatial heterogeneity

4.1.1 Varying constant

A simple cluster level model for housing prices is a fixed effects model

$$y_{ij} = \alpha_j + x_{ij}\beta + \varepsilon_{ij}, \tag{6}$$

with $j = 1, \dots, B$, $i = 1, \dots, n_j$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$[3]. The index j denotes the clusters, and i the individual observations, n_j denotes the number of observations in cluster j, y_{ij}

[3] Except where otherwise specified, the ε_{ij}'s are uncorrelated.

the natural logarithm of the selling price of the ith observation in cluster j, and x_{ij} the row vector of explanatory variables. This model is sometimes called the least squares dummy variable model (LSDV) (see for example Greene, 1993). In this model the constants α_j vary over the clusters, say neighborhoods. It is assumed that the α_j's are unknown fixed quantities. It can be shown that the least square estimator of α_j can be given by

$$\widehat{\alpha}_j = \overline{y}_{.j} - \overline{x}_{.j}\widehat{\beta},$$

with $\overline{y}_{.j}$, and $\overline{x}_{.j}$ given by $\sum_{i=1}^{n_j} y_{ij}/n_j$, and $\sum_{i=1}^{n_j} x_{ij}/n_j$ respectively. $\widehat{\beta}$ is the least squares estimator of β given by $(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y}$, with $\widetilde{x}_{ij}$, and $\widetilde{y}_{ij}$ deviations from the cluster means, defined by $\widetilde{x}_{ij} = x_{ij} - \overline{x}_{.j}$, and $\widetilde{y}_{ij} = y_{ij} - \overline{y}_{.j}$. The variance of the estimator $\widehat{\alpha}_j$ is given by

$$var(\widehat{\alpha}_j - \alpha_j) = \overline{x}_{.j} var(\beta - \widehat{\beta})\overline{x}'_{.j} + \sigma^2/n_j.$$

So, apart from $\widehat{\beta}$ the estimator $\widehat{\alpha}_j$ only depends on the observations in cluster j. If n_j is small, then $\widehat{\alpha}_j$ is very sensitive to outliers. This can also be deduced from $var(\widehat{\alpha}_j - \alpha_j)$, which in general is dominated by σ^2/n_j, and for small n_j this variance is relatively big. For $n_j = 1$, the variance $var(\widehat{\alpha}_j - \alpha_j)$ equals the variance for an individual observation, σ^2.

A drawback of this specification is that nothing can be said about clusters that are not included in the sample, for example clusters without selling prices. If α_j is specified as a random effects this problem is overcome.

The random effects model is given by

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta + \varepsilon_{ij}, \\ \alpha_j &\sim N(\mu, \tau\sigma^2). \end{aligned}$$

In this model α_j is a random effect, drawn from a normal distribution. It can be shown that in case of $n_j = n$, and conditional on β, the estimator of α is given by

$$\begin{aligned} \widehat{\alpha}_j &= (1-\omega^{-1})(\overline{y}_{.j} - \overline{x}_{.j}\beta) + \omega^{-1}(\overline{y} - \overline{x}\beta), \\ \omega &= n\tau + 1, \end{aligned}$$

with $\overline{y}$, and $\overline{x}$ overall means. So, the estimator $\widehat{\alpha}_j$ is a weighted average of the cluster mean and the overall mean. The weight ω^{-1} is small for large n, so for large n the estimator $\widehat{\alpha}_j$ is approximately $\overline{y}_{.j}$. On the other hand for small n the estimator $\widehat{\alpha}_j$ is approximately $\overline{y}$. So $\widehat{\alpha}_j$ is a shrinkage estimator. For $\tau \to \infty$ the estimator of α coincides with the estimator in the fixed effects model.

The random effects model is a special case of a hierarchical model. A hierarchical model is given by

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta + \varepsilon_{ij}, \\ \alpha_j &= z_j\delta + \eta_j, \\ \eta &\sim N(0, \pi\sigma^2). \end{aligned}$$

The row vector z_j is a vector of explanatory variables on cluster level. The second equation provides an explanation for the differences in α_j. For example, differences in α_j are explained by cluster level variables as distance to city center, crime rate, etc. If z_j is 1, the hierarchical model coincides with the random effects model.

Given β, and δ, and $n_j = n$, the estimator $\widehat{\alpha}_j$ is given by

$$\begin{aligned} \widehat{\alpha}_j &= (1-\omega^{-1})(\overline{y}_{.j} - \overline{x}_{.j}\beta - z_j\delta) + \omega^{-1}(\overline{y} - \overline{x}\beta - \overline{z}\delta), \\ \omega &= n\tau + 1. \end{aligned}$$

So the estimator $\widehat{\alpha}_j$ is also a shrinkage estimator. The estimation of β, and δ is not treated here. For a thorough treatment of hierarchical, or multilevel models we refer to Bryk (1992), Goldstein (1995), Longford (1993), and O'Hagan (1994).

4.1.2 Varying coefficients explanatory variables

A fixed effects model for varying coefficients per cluster is provided by

$$y_{ij} = x_{ij}\beta_j + \varepsilon_{ij}.$$

The estimates for β_j are the same as the OLS estimates from the separate models $y_j = x_j\beta_j + \varepsilon_j$. The only difference is the estimation of the variance σ^2. In the model it is assumed that the variance is constant over the clusters. A disadvantage of the fixed effects approach is that β_j is poorly estimated if only a few observations are available in cluster j. The estimate does not exists or will have large variances if the number of observations is small relative to k, the number of explanatory variables.

This problem is overcome if it is assumed that β_j are random effects. The random effects model is provided by

$$\begin{aligned} y_{ij} &= x_{ij}\beta_j + \varepsilon_{ij}, \\ \beta_j &= \delta + \eta_j, \end{aligned}$$

and $\eta \sim N(0, \tau\sigma^2 I_B)$. Like in the case of varying constants also in the case of varying coefficients for all explanatory variables, the estimates of β_j are shrinkage estimator. The estimate of β_j can be seen as a weighted average of the overall estimate of δ and the fixed effects cluster estimate of β_j.

The hierarchical model is an extension of the random effect model. In the second equation the constant 1 for δ is replaced by a vector z_j containing cluster level variables, so

$$\begin{aligned} y_{ij} &= x_{ij}\beta_j + \varepsilon_{ij}, \\ \beta_j &= z_j\delta + \eta_j. \end{aligned}$$

This kind of models are described by Can (1992), and Orford (1999)

4.2 Spatial dependence

The modelling of spatial dependence concerns merely the specification of the error term. In this subsection an example of spatial dependence is provided at a cluster level.

Specify a matrix W with elements w_{ij} as

$$w_{ij} = \begin{cases} 1 & \text{cluster } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} .$$

The spatial autocorrelation matrix Ω can be modelled as

$$\Omega = (I_B - \rho W)'(I_B - \rho W).$$

This variance matrix can be used in a random effect model or hierarchical model. For $\rho = 0$, no spatial dependence is apparent. The hierarchical model is given by

$$\begin{aligned} y_{ij} &= \alpha_j + x_{ij}\beta + \varepsilon_{ij}, \\ \alpha_j &= z_j'\delta + \eta_j, \\ \eta &\sim N(0, \sigma^2 \pi \Omega). \end{aligned}$$

5 INCORPORATION OF TIME

5.1 Hierarchical trends

Let the vector y_t represent the logarithms of the selling prices of houses at time t. We denote the length of y_t by p_t, and the k-th observation in y_t by y_{kt} $(k = 1, \ldots, p_t)$. First, we assume that all prices follow a common trend, which we can write as

$$y_t = \mathbf{i}\mu_t + \epsilon_t, \tag{7}$$

where $\mathbf{i}$ is a p_t-vector of ones, and $\epsilon_t \sim N(0, \sigma^2 I)$, with I a $p_t \times p_t$ identity matrix. Note that we have suppressed the time dependency of $\mathbf{i}$ and I in the notation. μ_t is a scalar stochastic trend process.

A simple example of a stochastic process is the random walk

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{8}$$

with a given μ_1. The disturbances ϵ and η are assumed to be independent. We can also view as a generalization of the model $y_t \sim N(\mu, \sigma_\epsilon^2)$ in which the global level μ is allowed to change over time. The model is called the *random walk plus noise* model, or the *local level* model.

Note that in case explanatory variables are added, for $\sigma_\eta^2 \to \infty$, equations (7) and (8) provide to the LSDV model.

We can give a random walk some direction by adding a drift parameter, which changes (8) into $\mu_{t+1} = \beta + \mu_t + \eta_t$. Notice that if we let $\sigma_\eta^2 \to 0$, this reduces to a straight line with slope β. The random walk with drift can be further generalized by allowing β to vary over time:

$$\mu_{t+1} = \beta_t + \mu_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{9}$$

$$\beta_{t+1} = \beta_t + \zeta_t, \qquad \zeta_t \sim N(0, \sigma_\zeta^2) \tag{10}$$

with known μ_1, β_1, and independent η and ζ. Equation (7) with these specifications for the trend is called the *local trend* model. The trend μ becomes smoother when we decrease σ_η^2; in the limiting case of $\sigma_\eta^2 = 0$, μ_t is said to follow an *integrated random walk*, since its first difference follows a random walk.

Suppose we have a method to categorize houses into L different types. We can include a dummy matrix D_t for house types as regressors in the model. Each row in the $p_t \times L$ matrix D_t has a one in the l-th column and zeros elsewhere, if the house is of type $l = 1, \ldots, L$. Writing λ for the regression parameter vector for house types,

$$y_t = \mathbf{i}\mu_t + D_t\lambda + \epsilon_t. \tag{11}$$

In this model, the relative price differences between house types stays constant through time. If we expect the prices to grow at different rates, we could allow λ to vary over time. The elements of the vector λ_t can be modeled as trends in a similar fashion as the common trend μ_t. The specification for the house type trends are typically less elaborate than for the common trend; we will model them as simple independent random walks, with a common variance level.

We can see immediately that if both λ and μ are constant, an unrestricted specification like (11) leads to the dummy trap. In the general time-varying case, there is also an identification problem if we try to extract both a general trend and a trend for house types from the data. We can solve this by imposing the restriction $\mu_1 = 0$. With this restriction, the level of μ_t indicates the general price increase relative to the first time period. A trend for a specific house type is obtained as the sum of μ_t and the element of λ_t of corresponding to the house type.[4]

Of course, there is no need to restrict this approach to house types; any qualitative independent variable can be treated in this way. We will refer to these as clusters. An obvious example is a variable which indicates the neighborhood in which the house is located. Note that if we model two classifications simultaneously (e.g., house types and neighborhoods), additional restrictions are required.

We model the vector of log house prices with an extended version of (11), the Hierarchical Trend Model (HTM):

$$y_t = \mathbf{i}\mu_t + D_t\lambda_t + \dot{D}_t\theta_t + \ddot{D}_t\phi + X_t\beta + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2 I) \tag{12}$$

We specify the general trend as an integrated AR(1) process with drift:

$$\mu_{t+1} = \nu_t + \mu_t \tag{13}$$

$$\nu_{t+1} = \kappa + \varpi\nu_t + \eta_t, \qquad \eta_t \sim N(0, q_1\sigma^2). \tag{14}$$

The vector λ_t contains trend levels for house types at time t, while the vector θ_t contains

[4] An alternative solution is to drop the common trend from the model. Without a common trend, the correlations between the house type trends will have to be specified through the disturbance variance matrix.

trend levels for neighborhoods. The matrices D and $\dot{D}$ contain ones and zeros such that they select the appropriate house type and neighborhood for the observation. For now, we assume random walks for these trends:

$$\begin{aligned} \lambda_{t+1} &= \lambda_t + \varsigma_t, \qquad \varsigma_t \sim N(0, q_2\sigma^2 I) \quad (15) \\ \theta_{t+1} &= \theta_t + \omega_t, \qquad \omega_t \sim N(0, q_3\sigma^2 I), \quad (16) \end{aligned}$$

where the identity matrices I have the appropriate dimensions.

Each neighborhood is divided in a number of subneighborhoods, for which we assume separate levels. We collect all levels in a vector ϕ, and use a selection matrix $\ddot{D}$ to assign the appropriate subneighborhood level to the observations. We can treat the levels as fixed or random effects. An example model for random effects would be $\phi \sim N(0, q_4\sigma^2 I)$.

Finally, we add a number of explanatory variables X_t with fixed parameters. We will keep the basic form of the model linear, so a specification like $(\ln x_t'\beta)^\alpha$ will be approximated by an iterative procedure as described in section 3. Note that β is kept constant over time and over clusters. So,an obvious generalization would be to vary β over time and clusters, for example $\beta_{t+1} = \beta_t + \varrho_t$, with $\varrho_t \sim N(0, q_5\sigma^2 I)$.

In the method of time series modeling we described, observations are assumed to be aggregates of unobserved parts with some interpretation, such as trend, and cycle. Each part can be modeled further with as much detail as desired. These models are known in the literature as *Structural time series*, or *Unobserved components* models. For a thorough treatment we refer to Harvey (1989) and West and Harrison (1997), who discuss these models as examples of *State space* or *Dynamic linear* models. In the State space form, the unobserved components can be estimated with the *Kalman Filter* algorithm.

To put the model into State space form, we stack the variables μ_t, ν_t, κ, and the vectors $\lambda_t, \theta_t, \phi, \beta$ in the state vector α_t. The measurement equation is simply

$$y_t = Z_t\alpha_t + \epsilon_t = [\ \mathbf{i} \quad 0 \quad 0 \quad D \quad \dot{D} \quad \ddot{D} \quad X_t\ ]\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2 I). \quad (17)$$

In the transition equation $\alpha_{t+1} = T_t\alpha_t + \xi_t$, the transition matrix T_t is a time independent block diagonal matrix, with

$$\begin{bmatrix} 1 & 1 \\ 0 & \varpi \end{bmatrix}$$

on the upper block, and I on the lower block. The zero-mean Normal transition disturbance ξ_t has a diagonal variance matrix, with

$$\sigma^2[\begin{array}{cccccccc} 0 & q_1 & 0 & q_2 \ldots q_2 & q_3 \ldots q_3 & q_4 \ldots q_4 & 0 \ldots 0 \end{array}]$$

on the diagonal.

5.2 Estimation issues

We already mentioned an identification problem in specifying trends on different levels. In the general model (12), we will also set the initial general trend level μ_0 at zero. For the slope component ν_t we have specified a stationary AR process. The initial value ν_0 follows from the unconditional AR mean $\sigma^2 q_1/(1-\varpi^2)$.

For the neighborhood trends, we will model the initial levels θ_0 explicitly, as

$$\theta_0|\pi \sim N(V\pi, \sigma^2\Psi), \tag{18}$$

where V contains explanatory variables for neighborhood value levels. Examples are crime rate, and distance from the city center. The variance matrix Ψ is specified as a spatial autocorrelation matrix, as described in section 4. With a scaling factor, we can use the same matrix to model correlations in the neighborhood trend disturbances; in equation (16), $\omega_t \sim N(0, q_3\sigma^2\Psi)$. These more elaborate specifications are especially valuable if in some neighborhoods few observations are available. With this model we can also value houses in neighborhoods without any selling price data.

We already mentioned that the subneighborhood levels ϕ_t can be modeled as fixed or random. If we model them as fixed, we have an identification problem in each neighborhood, comparable to specifying a full set of dummies besides a general level. This can be solved by omitting a subneighborhood in every neighborhood (for example, the first one). The identification problem does not arise when the levels are modeled as random effects.

A final identification issue results from the fact that we have two complete classification for the houses: neighborhoods and house types. This can be solved by set the initial level of some house type at zero. The general trend μ_t is interpreted as general with regard to neighborhoods for houses of this type. The vector $\mathbf{i}\mu_t + D_t\lambda_t$ provides the trends for all house types, general with regard to neighborhoods.

We already mentioned that models in state space format can be estimated by the Kalman filter. The Kalman filter assumes the first and second moment of the initial state to be known. In general this is not true, (a part of) the initial state is diffuse. This leads to an initialization problem which can be solved by the diffuse Kalman filter. The recursions for the diffuse Kalman filter are provided in for example De Jong (1987), and Koopman (1992,1997).

In Francke and De Vos (2000), it is shown how a hierarchical trend model with explanatory variables can be computed efficiently. First, we calculate the means per neighborhood $\bar{y}_1, \ldots, \bar{y}_T$, and the deviations from these means $\tilde{y}_1, \ldots, \tilde{y}_T$. The length of vector $\bar{y}_t$ is the number of different neighborhoods for which we have observations at time t, while $\tilde{y}_t = y_t - \dot{D}_t\bar{y}_t$ has the same dimension as y_t. Likewise, we calculate means and deviations from means for the explanatory variables. The coefficients of the explanatory variables are time- and neighborhood invariant, and can be computed by applying OLS on the stacked deviation from mean vectors and matrices $\tilde{y} = [\begin{array}{ccc} \tilde{y}_1' & \ldots & \tilde{y}_T' \end{array}]'$ and $\tilde{X} = [\begin{array}{ccc} \tilde{X}_1' & \ldots & \tilde{X}_T' \end{array}]'$. Subsequently, the Kalman Filter is run with the mean data $\bar{y}_t, \bar{X}_t$, and with the OLS estimates as initial mean and variance of the explanatory variables in the state. The likelihood is obtained as the product of the OLS likelihood and the Kalman Filter likelihood.

6 APPLICATIONS

6.1 Data description

The HTM-model as described in the previous section, is applied on two different datasets. The first dataset contains selling prices for houses in the Amsterdam region, an urban district with a relatively high proportion of apartments. The second database contains selling prices of the Breda region, a rural district with one middle-sized city Breda with about 160,000 inhabitants. The Breda region has a relatively high proportion of single-family houses.

The two databases were established by the National Association of Property Brokers (NVM). They have several merits from the point of view of this study. First, the size of the database is large because the number of transactions registered by the NVM is on average more than 60% of all transactions registered by the Land Registry (Kadaster). Sample sizes of these magnitudes undoubtedly provide an adequate foundation for measuring house price changes at a regional level. Secondly, the data concerning house characteristics are more extensive than anything available in this area and this again helps to improve the reliability of the statistical analysis. The available information about house characteristics is summarized below:

1. Purchased price: date of selling, asking price, condition on sales
2. Location: address (street, number, postal code)
3. Housing characteristics:

(a) House type: detached, semi-detached, terraced, apartment (with sub-classification)

(b) Tenure: freehold, land leasehold condition

(c) Garage: type of garage

(d) Heating type

(e) House size: area in m3

(f) Plot size: total size in m2

(g) Garden: length and position of garden

(h) Space: number of rooms, kitchen, bathroom, type of living room

(i) Age: year of construction

(j) Physical condition: interior maintenance, exterior maintenance

(k) Marketing period

(l) Listed building

As indicated above, the data refer to transactions at the selling agreement as opposed to the notarial act stage. This means that the price information is more up-to-date as an indicator of price movements because of the time lags that occur between the price negotiation stage and the ultimate completion of the transaction at the notarial act - a lag that may extend over several months.

In the next two subsections both databases will be described.

6.2 Amsterdam region

A special database was designed to accommodate the various measurement problems associated with house prices. The database covers house 44,780 purchase transactions of existing dwellings in the Amsterdam region from January 1985 until July 1999. This market area is built up of four municipalities: Amsterdam, Amstelveen, Diemen and Ouder-Amstel.

Transactions without proper postal code were excluded from the database. For the segmentation in Amstelveen the year of construction was one of the crucial criteria. These restrictions concerning house type, postal code and year of construction resulted in 42616 usable transactions from the original 44780. To be able to correct for differences in quality of the houses it is necessary to take into account housing characteristics like house size, lot size, year of construction, etc. This leads to extra demands on the data resulting in

Type	Number
Terraced	5613
Semi-detached	2562
Detached	595
Apartment	22678
Total	31448

Table 1: Number of relevant transactions per house type, Amsterdam region

31448 usable transactions. Especially in the earlier years less than half of the database could be used.

The sales volume per year doubled during the period 1985-1999Q2. The fraction of sales in the municipality of Amsterdam alone is more than 75% of all transactions, most of them being apartments. The NVM database registers ten house types which are assembled into four categories, namely: detached, semi-detached, terraced, and apartment. The distribution of transactions to house type can be seen in Table 1. Clearly the fraction of apartments is dominant and the influence of detached houses on the total price development is small.

The selling prices are a priori divided in different segments, depending on neighborhood, and house type. The Amsterdam data were ordered according to the existing division in neighborhoods (Onderzoek en Statistiek, 1996). These neighborhoods can be recognized by their postal codes so that the NVM database can be ordered accordingly. From these about 350 neighborhoods we constructed 10 different sub-regional districts which generated relatively homogeneous groupings of neighborhoods with respect to house price development. As a classification of house types we use the one given in table 1. This resulted in a 40-segments classification produced from four property types and ten sub-regional districts. We refer to Francke and Vos (2000) for a more extensive treatment of the segmentation.

On the basis of these transactions a model was constructed, as explained in the previous section, in which for each transaction a price was estimated which was compared with the actual purchase price. When the actual price differed more than 80% (about 4 times the standard deviation of the model) from the value calculated with the model, transactions were excluded (229 transactions (0,7%)) because they were considered unreliable. This resulted in a final database with 31219 transactions over the period 1985-1999Q2, a loss of 30,3% (13561 transactions) compared to the original database.

6.3 Breda region

The Breda database contains 25,644 transactions covering the period January 1985 until October 1999. The number of NVM transactions in the Breda region is relatively high, about 65% of the total number of transactions. The Breda region contains selling prices of different municipalities: Baarle-Nassau, Breda, Chaam, Dongen, Dussen, Geertruidenberg, Gilze en Rijen, 's Gravenmoer, Made en Drimmelen, Nieuw-Ginniken, Oosterhout, Prinsenbeek, Raamsdonk, Teteringen en Waspik.

To be able to correct for differences in quality of the houses it is necessary to take into account housing characteristics like house size, lot size, year of construction, postal code, etc. This leads to extra demands on the data resulting in 21,175 usable transactions. Especially in the earlier years less than half of the database could be used.

The sales volume per year doubled during the period 1985-1999Q2. The fraction of sales in the municipality of Breda alone is more than 45% of all transactions. The NVM database registers ten house types which are assembled into four categories, namely: detached, semi-detached, terraced, and apartment. The distribution of transactions to house type can be seen in Table 2. Clearly the fraction of single-family homes is dominant.

The selling prices are a priori divided in different segments, depending on neighbor-

Type	Number
Terraced	7275
Semi-detached	8591
Detached	3460
Apartment	1849
Total	21175

Table 2: Number of relevant transactions per house type, Breda region

hood, and house type. We distinguished 4 different sub-regional districts which generated relatively homogeneous groupings of neighborhoods with respect to house price development. As a classification of house types we use the one given in table 2. This resulted in a 16-segments classification produced from four property types and four sub-regional districts.

On the basis of these transactions a model was constructed, as explained in the previous section, in which for each transaction a price was estimated which was compared with the actual purchase price. When the actual price differed more than 60% (about 4 times the standard deviation of the model) from the value calculated with the model, transactions were excluded (50 transactions) because they were considered unreliable. This resulted in a final database with 21125 transactions over the period 1985-1999 October, a loss of 17,6% (4519 transactions) compared to the original database.

7 MODEL RESULTS

The model of selling prices in the Breda region is specified as described in equation (12). One general trend (μ_t) is specified as a random walk with drift, 4 district trends (θ_t), and finally 4 house type trends (λ_t), both as random walks. This results in 16 different trends. The districts are divided in 73 subneighborhoods (ϕ) (postal area), for which we assume separate levels, modelled as random effects.

The model contains 50 coefficients of explanatory variables, and 5 variances to be estimated. The definitions of the variables are provided in table 10 and 11 in the appendix. The additive variables, as explained in section 2, are specified as

$$\beta_1 \ln(\text{HouseSize800}+\beta_2\text{HouseSizeRest}+\beta_3\text{PlotSize500}+\beta_4\text{PlotSizeRest}$$
$$+\beta_5\text{GarageDetached}+\beta_6\text{GarageAnnex}+\beta_7\text{GarageBuiltIn}).$$

The estimation results are shown in the tables 3 and 4. In the appendix more results are shown: table 14 contains the subneighborhood levels (between brackets the number of observations per neighborhood), table 13 contains the coefficients for the different housing types, and table 12 provides the coefficients for the interior and exterior maintenance.

All coefficients have the good sign. An increase of the House size by 10% leads to an increase of the value by approximately $0.673 \times 10\% \simeq 7\%$. The coefficient for a detached garage is somewhat lower than the other garage coefficients. Maybe this is due to the fact that the detached garages are more in the rural areas then in the city areas. A listed building is about 15% more expensive than a "normal" house. The linear drift has a coefficient of 0.0066, indicating an average yearly price rice of $12 \times 0.0066 \simeq 8\%$ over the whole period. The coefficients for interior and exterior maintenance show differences of respectively 0.26 (30%) and 0.23 (26%) between perfect and bad maintenance.

The standard deviations are provided in table 4. The standard deviation of the measurement equation σ is 0.1262, which can be interpreted as a standard deviation of about 13%, due to the log specification. So, 66 percent of the residuals is within one time the standard deviation. The standard deviations for the random walks (μ), the general trend, the district trends (θ), and the house type trends (λ), are small compared with the standard deviation of the measurement equation. The general price deviation per year, apart from the drift, has a standard deviation of $\sqrt{12} \times 0.0074 \simeq 2,6\%$.The standard deviation of the random effects for the subneighborhoods is about 10%. This means that a

Variable	Coefficient	Standard Deviation	T-value
HouseSize800	0.673	0.0057	117.58
HouseSizeRest	0.883	0.0423	20.89
PlotSize500	0.901	0.0219	41.11
PlotSizeRest	0.085	0.0033	25.68
GarageDetached	45.82	2.3922	19.15
GarageAnnex	70.07	3.1811	22.03
GarageBuiltIn	55.35	4.3825	12.63
NRooms	0.0134	0.00104	12.84
Age1900	-0.1745	0.0089	-19.51
Age1920	-0.2280	0.0064	-35.42
Age1945	-0.1817	0.0046	-39.90
Age	-0.0052	0.00012	-41.81
Listed	0.1385	0.0197	7.02
Term	-0.0020	0.0003	-7.50
SalesConditions	-0.0019	0.0132	-0.14
LivingRoom1	0.0342	0.0027	12.77
LivingRoom2	0.0214	0.0073	2.94
LivingRoom3	0.0251	0.0052	4.84
LivingRoom4	0.0067	0.0026	2.59
LivingRoom5	0.0075	0.0051	1.49
Time in months (κ)	0.0066	0.0006	10.73

Table 3: Estimation results Breda region

	estimate
σ	0.1262
$\sigma\sqrt{q_1}$ (μ)	0.0074
$\sigma\sqrt{q_2}$ (θ)	0.0060
$\sigma\sqrt{q_3}$ (λ)	0.0024
$\sigma\sqrt{q_4}$ (ϕ)	0.0983

Table 4: Estimation results variances Breda region

	estimate
σ	0.1824
$\sigma\sqrt{q_1}$ (μ)	0.0116
$\sigma\sqrt{q_2}$ (θ)	0.0110
$\sigma\sqrt{q_3}$ (λ)	0.0630
$\sigma\sqrt{q_4}$ (ϕ)	0.1312

Table 5: Estimation results variances Amsterdam region

subneighborhood level is in 66 percent within -10% and +10% from the district level.

Figure 1 gives the general trend and figure 2 an example of the trend for a specific district and house type as a deviation from the general trend. The y axis is in logarithms, so an increase of .1 means that the increase in selling prices is approximately 10%. The dashed lines indicate the 95% confidence intervals and the points the average standardized selling prices, corrected for individual characteristics and the general trend, see equation (12)

In table 7 the estimates of the variance for the Amsterdam region are shown. The standard deviation for the measurement equation σ is 18%, about 6% more than in the Breda region.

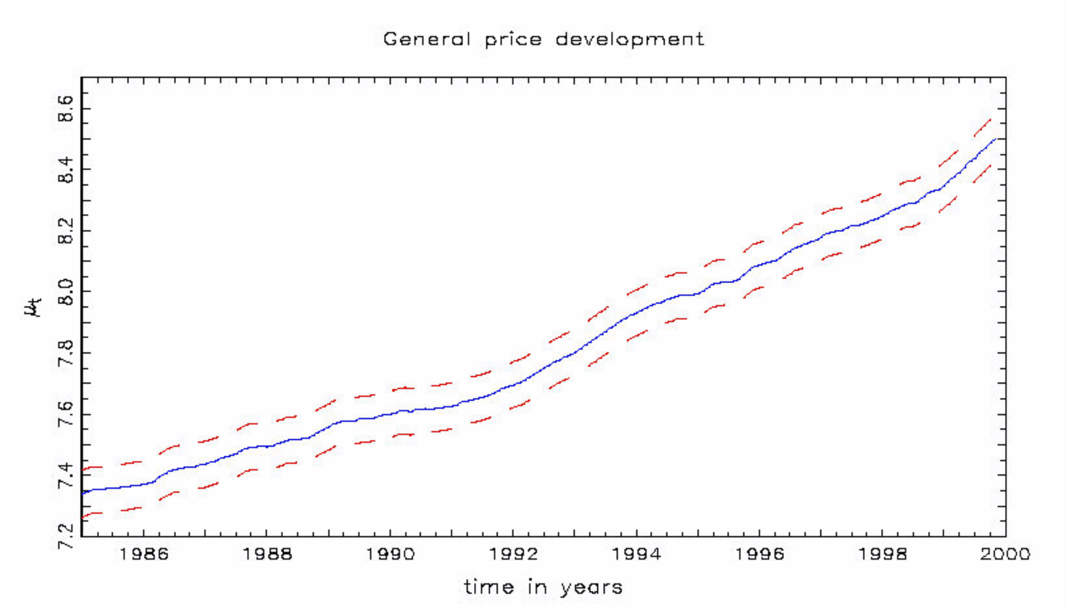

Figure 1: General trend for the Breda region.

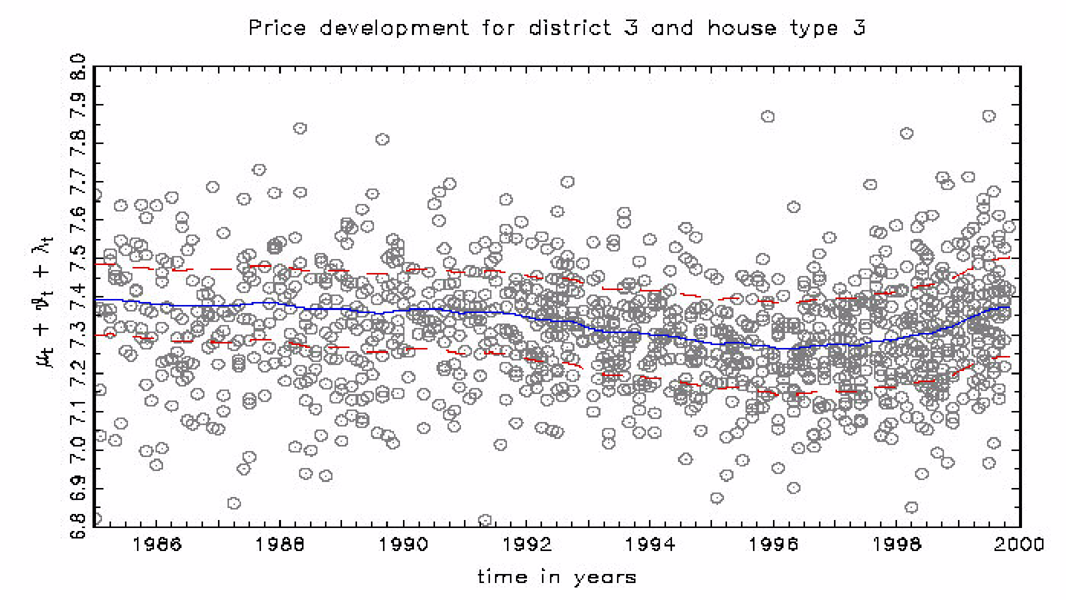

Figure 2: A specific cluster trend for the Breda region.

8 PRICE INDICES

In this section we compare price indices obtained by the HTM (the fixed-sample hedonic indices) with price indices based on a weighted average of median selling prices (simple-weighted indices). In the first subsection we describe these methods, in the next subsection we compare the results of both methods for the Amsterdam and Breda region. The last subsection deals with the reliability of the price indices.

8.1 The simple-weighted and the fixed-sample hedonic index

In this method for every market segment the median selling price is calculated in period t and $t+1$. Next a weighted average of the segment medians is calculated with as weights the relative number of sales in the segment, for both periods. The relative difference between the two weighted averages provides the price index. The weights are not fixed but are presented by the relative number of sales in each separate period (rolling basis).

In formula, with $i = 1, \dots, B$, the market segments and t the period,

$M_{i,t}$ the median selling price in market segment i and period t,

M_t the weighted median selling price in period t,

$n_{i,t}$ the number of sales in market segment i and period t,

n_t the number of sales in period t.

Then

$$\begin{aligned} n_t &= n_{1,t} + \cdots + n_{B,t}, \\ M_t &= (n_{1,t} \times M_{1,t} + \cdots + n_{B,t} \times M_{B,t})/n_t. \end{aligned}$$

So the relative price movement equals $(M_{t+1}/M_t - 1) \times 100\%$.

In the fixed-sample hedonic index the market segment price movements are constructed from the HTM model, as described in section 5. Fixed weights are used to obtain a general price index from the segment price movements. The fixed weights in time are taken as the relative number of selling prices per market segment over a long reference period

(1985 - 1999Q2). The model simultaneously provides trends for different districts and house types on a monthly basis. From these trends it is quite easy to construct price indices on a monthly, quarterly, or yearly basis. In the model corrections are also made for differences in structural and locational characteristics in order to be able to compare the selling prices.

The differences between both methods are quite obvious. The fixed-sample hedonic index corrects for differences in characteristics of the houses, and the simple-weighted does not correct for any difference. The fixed-sample hedonic index uses fixed weights per period, the simple-weighted index has time varying weights. The fixed-sample hedonic index takes into account the selling date (on a monthly basis), the simple-weighted index has an implicit assumption of equally spaced selling dates. The simple-weighted index compares the "middle" of one period with the "middle" of the next period. So the simple-weighted index has a time lack of half a period. The fixed-sample hedonic method is capable of producing up-to-date indices, having a lag of at most half a month.

8.2 Price indices for the Amsterdam and Breda region

In this subsection price indices are shown for the Amsterdam and Breda region for the simple weighted and the fixed-weighted hedonic method.

The fixed-weighted hedonic index shows that for the Amsterdam region the price development varies over house type and district, and for the Breda region the price development varies merely over house type.

Figure 3 shows the price changes for the Amsterdam region over the period 1985 - 1999Q2. It is notable that differences between the methods are substantial, for instance up to 16 percent points in 1989 despite the fact that there were more than 1000 transactions per year. So, it seems to be very important to correct for differences in characteristics of the sold houses.

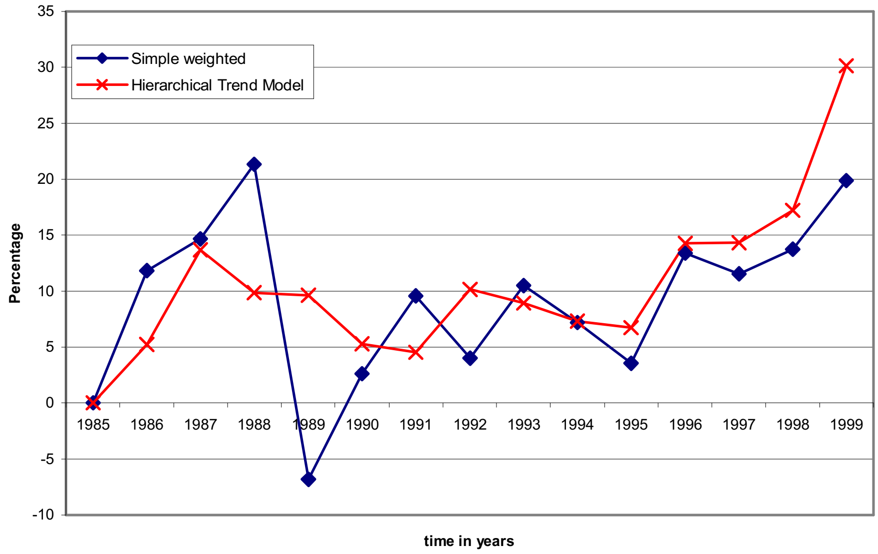

Figure 3: Price change for Amsterdam region.

For the Breda region similar results are shown in table 6.

For small market segments the differences are even more apparent. Table 7 shows for the price changes for apartments in a district in the Breda region for both methods. It seems to be that the simple-weighted index method is not a reliable method for small market segments.

8.3 Reliability

In the last subsection it is shown that for small market segments the simple-weighted method does not produce reliable indices. The reliability depends merely on the number of observations and the heterogeneity of the sold houses. The number of observations is dependent of the number of clusters, and the time period. The more cluster are distinguished, and the shorter the time period considered, the less observations are available. In this subsection standard deviations are provided for the price changes in the Breda region.

	Simple-weighted		fixed-sample hedonic	
	Change	Cumulative	Change	Cumulative
1985				
1986	3.9	3.9	5.4	5.4
1987	7.3	11.4	6.1	11.9
1988	3.3	15.1	5.1	17.6
1989	7.2	23.4	6.3	25.0
1990	0.6	24.2	3.2	29.0
1991	4.9	30.3	4.9	35.3
1992	9.3	42.4	9.4	48.1
1993	10.8	57.8	12.7	67.0
1994	7.9	70.2	10.8	85.0
1995	9.0	85.6	7.2	98.3
1996	9.4	102.9	9.7	117.6
1997	7.3	117.8	8.5	136.1
1998	7.8	134.7	8.5	156.3
1999 Oct.	13.4	166.0	13.4	190.7

Table 6: Price change in percentage per year for Breda region

Year	Number of observations	Simple-weighted	fixed-sample hedonic
1985	21		
1986	23	-0.9	7.1
1987	18	-5.4	6.0
1988	22	27.2	4.8
1989	29	-3.1	6.5
1990	37	3.0	1.9
1991	39	7.9	3.9
1992	47	8.2	8.8
1993	53	17.8	12.5
1994	67	4.9	10.8
1995	91	9.2	6.7
1996	100	13.3	9.3
1997	115	1.9	8.8
1998	129	9.8	8.9
1999 Oct.	106	15.7	13.5

Table 7: Price changes in percentages for small market segment in Breda region

	Simple-weighted			Fixed-sample hedonic		
	Year	Quarter	Month	Year	Quarter	Month
Region	0.6%	1.2%	2.1%	0.36%	0.6%	0.9%
District and House type	2.5%	5%	8.7%	0.85%	1.2%	1.4%

Table 8: Table Standard deviation for simple-weighted and fixed-sample hedonic index Breda region

The price changes are produced on a monthly, quarterly, and yearly basis, for the region as a whole, and for an "average" district and house type. Table 8 shows the standard deviations for both methods.

The differences between the two methods are striking. The standard deviation for the fixed-sample hedonic method is 2 till 7 times smaller than for the simple-weighted method. If a yearly regional price change of 10% is computed by the simple-weighted index the 95% confidence interval is provided by [8.8%;11.2%]. If it is computed by the fixed-sample hedonic index this confidence interval is [9.3%;10.7%]. For a monthly submarket price change these intervals are [-7.4%;27.4%], and [7.2%;12.8%], respectively.

It can be concluded that the fixed-sample hedonic index is far more accurate. It is possible to obtain reliable indices on a more detailed level, for small time periods, so it is more up-to-date.

Of course the most recent price changes can be estimated less reliable than a price change some periods before. Table 9 shows the standard deviations of the monthly price changes for the Breda region at the end of the time period considered. In this example the standard deviation rises from 0.60% till 0.77%.

9 CONCLUSIONS

This article presented a dynamic hedonic price model for selling prices of houses. The model considered is a hierarchical trend model with general and cluster trends. The clus-

Period 1999	Price Change	Standard deviation
1/1 - 1/2	1.70%	0.62%
1/2 - 1/3	1.71%	0.61%
1/3 - 1/4	1.23%	0.60%
1/4 - 1/5	1.52%	0.60%
1/5 - 1/6	1.84%	0.60%
1/6 - 1/7	1.19%	0.61%
1/7 - 1/8	2.04%	0.62%
1/8 - 1/9	1.41%	0.63%
1/9 - 1/10	1.70%	0.66%
1/10 - 31/10	1.08%	0.77%

Table 9: Price Change Breda region for the fixed-sample hedonic index

ters are constructed by location and house type. This model can be seen as an extension of a dummy variable model, with time varying constants for the different clusters. For the general trend a random walk with drift is assumed, for the cluster trends random walks. The coefficients of the explanatory variables are kept constant over time, location, and house type. These kind of dynamic models, even with varying coefficients, can be put in state space format, so they can be estimated by the (diffuse) Kalman filter.

Model results are shown for the regions Breda and Amsterdam. It is shown that an estimate of the value of a house can be produced with an average standard deviation of 18% for the Amsterdam region, and 13% for the Breda region.

Fixed-sample hedonic indices were constructed from the trends of the hierarchical trend model. These indices were compared with simple weighted indices. The question was which method provides the most adequate price change for standardized houses of constant quality, thereby measuring price changes in the market due to market forces only.

The findings of this research are summarized in the following.

For the house price indices of both the regional market and the inner-regional market (small market segments), using the hedonic method more up-to-date, detailed, and reliable results were obtained when studying yearly price developments than while using the simple-weighted method.

When small market segments with few transactions are concerned the use of the hedonic method seems to be the only accurate price index construction method, especially when indices have to be produced on a monthly or quarterly basis with even less transactions per period. On a monthly or quarterly basis the simple-weighted method will produce more unreliable results because of the small number of transactions.

References

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic.

Bryk, A.S., and S.W. Raudenbush (1992), *Hierarchical linear models. Applications and Data Analysis Method*, Sage Publications, Newbury Park, CA.

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

Cropper M.L., L.B. Deck, and K.E. McConnell (1988), "On choice of functional form for hedonic price functions," *Review of Economics and Statistics*, 70.4, 668-675.

de Jong, P. (1991), "The diffuse Kalman filter," *The Annals of Statistics*, 2, 1073-1083.

de Jong, P. (1991), "Stable algorithms for the state space model," *Journal of time series analysis*, vol 12, No. 2, 143-157.

de Jong, P. (1994) and S. Chu-Chun-Lin, "Fast likelihood evaluation and prediction for nonstationary state space models," *Biometrika*, vol. 81, No. 1, 133-142.

Davidson, R. and J.G. MacKinnon (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

Dinan, T.M. and J.A. Miranowski (1989), "Estimating the implicit price of energy efficiency. Improvements in the residential housing market: a hedonic approach," *Journal of Urban Economics*, 25, 52-67.

Dubin, R.A. (1991), "Spatial Autocorrelation and Neighborhood Quality," *Regional Science and Urban Economics*, Vol. 22, 433-452.

Dubin, R.A. (1998), "Predicting House Prices Using Multiple Listings Data, " - *Journal of Real Estate and Economics*, Vol. 17:1, 35-59.

Dunn, R., P.A. Longley, and N. Wrigley (1987), "Graphical procedures for identifying functional form in binary discrete choice models - a case study of revealed tenure choice," *Regional Science and Urban Economics*, 17.1, 151-171.

Francke, M.K. and A.F. de Vos (2000), "Efficient Computation of Hierarchical Trends," *Journal of Business & Economic Statistics*, vol. 18, No. 1, 51-57.

Francke, M.K. and G. A. Vos (2000), "Standardised Price Indices for the Regional Housing Market: A Comparison between the Fixed-Sample Index and the Hedonic Index,", Paper presented at the 7th ERES International Real Estate Conference, Bordeaux.

Goldstein, H. (1995), *Kendall's Advanced Theory of Statistics, Volume 3, Multilevel Statistical Models*, Second Edition, London, Arnold.

Greene, W.H.G. (1993), *Econometric Analysis*, Second Edition, New York, Maxmillan Publishing Company.

Halverson, R. and H. Pollawski (1982), "Choice of the functional form for hedonic price equations," *Journal of Urban Economics*, 10, 37-49.

Harvey, A.C. (1989), *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Johnson, R.C. and D.L. Kaserman (1983), "Housing market capitalization of energy-saving durable good investments," *Economic Inquiry*, 21.3, 487-500.

Jones, K. and N. Bullen (1993), "A multilevel analysis of the variations in domestic property prices: Southern England 1980-1987," *Urban Studies*, 30.8,1409-1426.

Koopman, S.J. (1992), "*Diagnostic checking and Intra-daily effects in Time series models,*"Amsterdam: Thesis Publishers.

Koopman, S.J. (1997), "Exact Initial Kalman Filtering and Smoothing for Nonstationary Time Series Models," *Journal of the American Statistical Association*, vol. 92, No. 440, Theory and Methods, 1630-1638.

Linnenman, P. (1980), "Some empirical results on the nature of the hedonic price function for urban housing markets," *Journal of Urban Economics*, 8, 47-68.

Longford, N.T. (1993), *Random Coefficients Models*, Oxford, Clarendon Press.

Longley, P.A. and R. Dunn (1988), "Graphical assessment of housing market models," *Urban Studies*, 25.1, 21-33.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian inference*, London, Arnold.

Orford, S. (1999), *Valuing the Built Environment: GIS and House Price Analysis*, Ashgate Publishing Company, Brookfield, VT, U.S.A.

Ohlsfeldt, R.L. (1988), "Valuing location in an Urban Housing Market," *Proceedings of the 3rd international conference on geocomputation, Bristol, UK, 17-19 September 1998*, GeoComputation CD-ROM, ISBN 0-9533477-0-2.

Pace, R.K., R. Barry, J.M. Clapp, and M. Rodriquez (1998), "Spatiotemporal autoregressive Models of Neighborhood Effects, "*Journal of Real Estate Finance and Economics*, 17:1, 15-33.

Palmquist, R.B. (1984), "Estimating the demand for the characteristics of housing," *Review of Economics and Statistics*, 66, 394-404.

Quigley, J.M. (1982), "Non-linear budget constraints and consumer demand - an application to public programs for residential housing," *Journal of Urban Economics*, 12.2, 177-201.

West, M. and J. Harrisson (1997), *Bayesian Forecasting and Dynamic Models*, Second Edition, Springer-Verlag, New York.

A Appendix

Variable	Definition
HouseSize800	the minimum of the house size in cubic meters, and 800
HouseSizeRest	the maximum of the house size in cubic meters - 800, and 0
PlotSize500	the minimum of the lot size in square meters, and 500
PlotSizeRest	the maximum of the lot size in cubic meters - 500, and 0
GarageDetached	1 if detached garage, 0 otherwise
GarageAnnex	1 if annex garage, 0 otherwise
GarageBuiltIn	1 if built-in garage, 0 otherwise
NRooms	number of rooms
Age1900	1 if year of construction $<$ 1900, 0 otherwise
Age1920	1 if 1900 $\leq$ year of construction $<$ 1920, 0 otherwise
Age1945	1 if 1920 $\leq$ year of construction $<$ 1945, 0 otherwise
Age	if year of construction $\geq$ 1945, selling year - year of construcion, 0 otherwise
Listed	1 if listed building, 0 otherwise
Term	Sellingperiod in days
SalesConditions	1 of no legal charges, 0 otherwise
Time in months	selling date in months from 1 January 1985
MI	interior maintenance, -1 Unkown,1 Perfect,2 Good,3 Reasonable,4 Moderate,5 Bad
ME	exterior maintenance
LivingRoom	Type of living Room, -1 Unknown, 1 L-Room, 2 T-Room, 3 Z-Room, 4 Through Room, 5 Room en suite

Table 10: Variable definitions Breda region

House Type	Description
10	Simple house
11	Middenstandswoning
12	Manor house
13	Residence
14	Countryhouse
15	Country Estate
16	Bungalow
17	Bungalow with a patio
18	Semi-bungalow
19	Split level house
20	Meanderhouse
21	Groundfloor flat
22	Upstairs flat
23	Groundfloor flat or Upstairs flat
24	House with a porche
25	Canalside house
26	Maisonette
27	Flat for the elderly
28	Flat with elevator
29	Flat without elevator
30	House with surgery, etc.
31	Drive-in home
32	Converted cottage/farmhouse

Table 11: Definition House types

Variable	Coefficient	Standard Deviation	T-value
MI1	0.0907	0.0062	14.54
MI2	0.0526	0.0041	12.70
MI4	0.0468	0.0091	-5.13
MI5	-0.1760	0.0219	-8.05
ME1	0.0593	0.0065	9.06
ME2	0.0467	0.0044	10.67
ME4	-0.0887	0.0096	-9.20
ME5	-0.1717	0.0233	-7.37

Table 12: Estimation results Maintenance Breda region

Variable	Coefficient	Standard Deviation	T-value
HouseType10	-0.058	0.0037	-15.81
HouseType12	0.107	0.0032	33.29
HouseType13	0.236	0.0064	37.09
HouseType14	0.2690	0.0100	26.89
HouseType15	0.1441	0.0496	2.90
HouseType16	0.1757	0.0083	21.28
HouseType17	0.1979	0.0102	19.46
HouseType18	0.1657	0.0065	25.52
HouseType19	0.0191	0.0226	0.84
HouseType20	0.0223	0.1267	0.176
HouseType30	0.0567	0.0199	2.84
HouseType31	-0.0799	0.0119	-6.72
HouseType32	0.1068	0.0113	9.47
HouseType21	-0.1391	0.0238	-5.84
HouseType22	-0.1355	0.0219	-6.18
HouseType23	-0.2732	0.0294	-9.28
HouseType24	-0.0155	0.0235	-0.66
HouseType26	-0.1147	0.0152	-7.56
HouseType27	-0.2113	0.0289	-7.32
HouseType28	-0.0107	0.0089	-1.19
HouseType29	-0.0878	0.0090	9.71

Table 13: Estimation results House type Breda region

SubN	Level	T-value	SubN	Level	T-value	SubN	Level	T-value
4835	0.068	2.74 (764)	4841	0.018	0.53 (844)	5121	0.002	0.08 (819)
4836	0.050	1.38 (19)	4854	0.040	1.18 (404)	5122	0.002	0.08 (198)
4837	0.118	4.60 (226)	48142	-0.084	-2.22 (36)	5124	-0.030	-1.10 (40)
4847	-0.041	-1.64 (517)	4271	-0.095	-4.18 (113)	5165	-0.073	-3.32 (151)
4851	0.061	2.44 (379)	4273	-0.057	-2.64 (175)	48141	0.003	0.13 (402)
4856	-0.114	-2.39 (7)	4822	-0.021	-1.02 (577)	4812	0.060	2.57 (492)
4858	0.055	0.90 (3)	4823	0.041	1.98 (392)	4815	0.042	1.77 (255)
4859	0.071	0.90 (1)	4824	-0.021	-1.02 (684)	4816	0.040	1.77 (179)
4902	-0.115	-4.64 (805)	4901	0.100	4.97 (952)	4825	-0.005	-0.10 (6)
48181	0.018	0.71 (298)	4907	0.064	3.21 (1448)	4826	-0.052	-2.21 (492)
48182	0.109	4.26 (266)	4921	0.073	3.52 (338)	4827	-0.038	-1.53 (168)
48183	0.005	0.18 (207)	4924	0.040	1.31 (27)	4849	0.039	1.61 (169)
48184	0.138	4.06 (24)	4931	0.061	3.00 (492)	4855	0.068	2.35 (43)
48185	0.087	3.01 (59)	4941	0.034	1.65 (509)	4861	0.045	1.77 (298)
48191	0.116	4.42 (144)	4942	0.013	0.61 (331)	4903	-0.075	-2.23 (22)
48192	0.058	2.26 (240)	4944	-0.024	-1.01 (97)	4905	0.068	2.64 (88)
49040	-0.579	-10.32 (4)	5101	0.009	0.41 (265)	4906	0.032	1.08 (40)
49041	-0.104	-4.20 (743)	5102	0.007	0.34 (407)	4908	0.071	2.99 (287)
4811	0.068	2.03 (460)	5103	0.030	1.43 (366)	4909	-0.006	-0.21 (67)
4813	-0.068	-2.01 (395)	5104	0.019	0.91 (312)	4911	0.044	1.58 (50)
4817	-0.051	-1.54 (1074)	5105	-0.031	-0.70 (8)	5111	-0.023	-0.89 (93)
4834	0.063	1.91 (1004)	5106	-0.050	-1.62 (25)	5113	-0.137	-3.94 (19)
4838	-0.001	-0.02 (57)	5107	-0.093	-1.20 (1)	5114	-0.154	-2.53 (3)
4839	0.015	0.41 (44)	5109	-0.002	-0.09 (146)	5125	-0.028	-0.76 (16)
						5126	0.004	0.18 (239)

Table 14: Estimation results Subneighbourhood levels Breda region