Nordic Journal of Surveying and Real Estate Research 3:1 (2006) 20–57
submitted on 2 March 2005
revised on 10 November 2005
accepted on 21 June 2006

Forecastability of Land Prices: A Local Modelling Approach

Marko Hannonen
Department of Surveying, Helsinki University of Technology, Espoo
marko.hannonen@hut.fi

Abstract. *This paper considers in depth the predictive accuracy of several different static and dynamic local hedonic models of land prices with adaptive bandwidth selection learning. All local models' post-sample predictions were compared to their proper corresponding global analogues with two data sets differing both in terms of quality and quantity of observations. The article tries to simultaneously combine the concepts of nonlinearity and nonstationary into same hedonic model specification by using local regression and error correction mechanisms. It turns out that local estimation of land prices results in a significant improvement in predictive accuracy over the global hedonic models, but only with the second data set, which is the high-quality one having been manually pre-checked for errors. The local estimation of equilibrium correction mechanisms resulted in a predictive failure, which is inherently caused by some profound problems with error correction processes, and thus with co-integrated processes, when modelling the temporal dimension of land prices.*

Keywords: *local regression, nonlinearity, nonstationarity, error correction.*

1 Introduction

Much of the aim of applied hedonic analysis is to produce a reasonable approximation to the generally unknown mean response function. The primary implication of the theoretical literature concerning hedonic prices in the real estate markets is that hedonic relationships are expected to be highly nonlinear due to their locational uniqueness that induces spational heterogeneity of regression surfaces (Wallace, 1996; McMillen and Thorsnes, 2003) that cannot be, in general, specified a priori (Anglin and Gencay, 1996). *Nonlinearity* indicates locally changing degrees of curvature in the hedonic function with non-constant characteristic values. Along with the spatial variation, an important source of variability observed in property and hedonic prices is temporal evolution as evidenced by trends and cycles, among others, present in the data (Schulz, 2003, p. 58). Temporal variation usually induces instability to hedonic functions over time, i.e. data-generating processes are time-varying, which

causes series of observations on property prices and attributes to exhibit *nonstationarity*[1].

Nonlinearity and nonstationarity of hedonic relationships caused by the spatio-temporal variation are fundamental features that characterise processes in the real estate markets imposing serious threats to empirical validity of hedonic models that in current practise are predominantly used. The complex question of validity underlying hedonic model specification can be divided into three subproblems that involve determining (Pace, 1993 & 1995; Wallace, 1996):

(1) the relevant set of response and attribute variables;
(2) the appropriate functional form between these variables;
(3) the adequate error distribution for inference.

Nonlinearity and nonstationarity are principal components of spatio-temporal structure, which govern choices in all subproblems (1)–(3), albeit their effect is most pronounced in the selection of an appropriate functional form. Economic theory and past experience usually provide useful a priori information of what variables should enter the model structure that substantially reduce the threat of omitted variable bias. Phenomena in real estate markets are, however, strongly dependent on the particular submarket, time period and property type and, as a consequence, the selection of proper set of dependent and conditioning variables is partially an empirical question, too.

Economic theory or previous experience rarely provides any specific, valid guidance on the choice of an appropriate functional form of the hedonic model (Pace, 1993 & 1995; Anglin and Gencay, 1996; Gencay and Yang, 1996). A prespecified functional form is, however, the fundamental assumption underlying the use of theory-laden parametric models; a poor choice imposes artificial structure on data and significantly invalidates results of the subsequent analysis[2]. In contrast, nonparametric techniques are data-driven, flexible approaches that can learn much of the genuine structure from available facts and, therefore, allow greatly reduced attention to the question of which functional form ought to be used.

An ideal statistical method for real estate valuation would possess (Pace, 1993):

(1) low specification error;

---

[1] Nonstationarity denotes the general sense of processes whose first two moments (conditional expectation and the variance of its error distribution) are not constant over time. The dynamic nature of data-generating processes is attributable to changes in economic environments, technological progress, political shifts, cultural movements, etc. A more detailed account of the fundamental sources of change in economics can be found, e.g. (Day, 1992).

[2] In the parametric modelling context, a common solution to the problem of selecting an appropriate functional form is to consider a set of parametric functions with the objective of finding a model structure that matches the evidence in most measurable respects. However, there is no clear evidence that this practice will be successful in avoiding functional form mis-specification (Anglin and Gencay, 1996; Hannonen, 2005). Specification searches can be highly time-consuming and the intrinsic power of these specification tests is somewhat questionable.

(2) robustness against outlying observations;
(3) superior post-sample predictive accuracy;
(4) known statistical properties.

Local regression techniques can significantly reduce the mis-specification error by letting the data determine the appropriate functional relationship between the response and a set of attributes. Locally weighted regression adapts locally to changing curvature in the hedonic surface by giving more weight to nearby observations (McMillen, 1996) and, therefore, can account for complex nonlinear and nonstationary patterns. The local adaptation property, which is achieved by *parametric localization* (Cleveland and Loader, 1996), makes it a highly attractive tool for estimating spatio-temporally non-homogeneous hedonic functions. Furthermore, any specific assumption underlying the error distribution can be relaxed and, in most cases, derived directly from the evidence e.g. by resampling techniques.

Data on land values are imperfect and scarce, which generates difficult problems with conventional parametric approaches. In particular, extreme points, influential and outlying observations, which might represent erroneous data or otherwise reflect unusual market conditions such as non-arm's length transactions, can seriously undermine the performance of a parametric estimator. The results of locally weighted regression can be robustified in a straightforward manner by a scheme, which is a variant of M-estimation. This simple regulation robuste typically offers enough protection against unusual or aberrant observations. Moreover, the use of local regression techniques results in highly parsimonious representations for the hedonic equation, if the set of regressors is kept reasonably low-dimensional and the training of model structure is implemented wisely by locally adaptive bandwidth selection rules. The parsimony of representations enables local regression to be used with moderate (or even with small) data sets.[3]

Predictive accuracy is perhaps the single most important operational criterion in the evaluation of performance of a chosen model. The success of hedonic model-based forecast depends on (see, Hendry, 1997):

(1) the existence of structure;
(2) whether such structure is informative about the future;
(3) the proposed method capturing the structure;
(4) the exclusion of irregularities that swamp the structure.

The aspects in (1)–(2) are characteristics of the economic system and the last two of the chosen forecasting method. When structure is understood as a systematic relation between the entity to be forecast and the available information, the conditions in (1)–(4) are sufficient for forecastability.

The objective of this study is to focus on the *forecastability of local regression*[4] when quality-adjusted or conditional land prices of undeveloped

[3] The standard statistical theory and the associated techniques underlying global least squares, likelihood, etc. fitting carries over in a natural manner to local regression (Loader, 1999) so that the criterion (4) about known statistical properties of estimators is fulfilled.

[4] The fundamental idea underlying the hedonic approach is the hedonic hypothesis that implies property prices, which are appearances of the corresponding property values,

sites are analysed in the framework of hedonic pricing theory. The paper tries to combine the important concepts of nonlinearity and nonstationary, which appear to lack any unified treatment in current hedonic usage, by simultaneously encapsulating them into model structure using local regression and error correction processes. Two municipality-specific data sets differing in both quality and quantity of observations from the Finnish land markets are analysed. In post-sample predictive testing several commonly used localised econometric models (double-log model, translog model, cubic expansion model and different error correction models) are analysed in depth and their predictions are compared with associated global alternatives. Any specific hypothesis testing of hedonic prices, mean response function, etc. as well as problematic issues relating to average derivate estimation are excluded from the study[5].

2 Overview of Nonparametric Local Modelling Studies in Real Estate

Hedonic modelling approaches can be classified by their *flexibility* in discovering conditional structure to three categories (see, e.g. Pace, 1995):

(1) parametric approaches;
(2) semiparametric approaches;
(3) nonparametric approaches.

Parametric models that represent *data modelling culture* (Breiman, 2001) have formed the conventional dogma of hedonic pricing methods in real estate studies, where prespecified global models are estimated by means of ordinarily least squares or some modification thereof. Benefits of parametric approaches undeniably include: simplicity, interpretability, parsimony and comprehensive statistical theory. The fundamental obstacle, however, underlying the general use of parametric models is their inflexibility, i.e. inability to learn genuine structure about the hedonic relationship from the evidence in such decision-making settings, where theoretically unknown nonlinearity or nonstationarity is expected. This is the typical case when the effects of variables representing location and time are considered (McMillen and Thorsnes, 2003). The

---

should be modelled statistically *conditional* on a small set of fundamental characteristics rather than analysed directly in terms of a large number of similar properties and their unconditional prices. The relevancy of this statistical conditioning is analogous to testing whether a hedonic-based model can predict property prices accurately in terms of some predefined criteria such as mean-squared forecast error. The predictive failure implies that the conditioning is irrelevant and, specifically, unpredictability of property prices, if their conditional mean equals its unconditional mean. Predictability is necessary for forecastability; the latter concept is more stringent requiring knowledge about the parametric class used in the conditioning. In this study concepts of forecastability and predictability are, however, used interchangeably.

[5] Unlike the term nonparametric would suggest, nonparametric estimators are "over-parametric" in a sense that they generate an infinite number of hedonic prices for each attribute depending on the values of that attribute. In practice, for a particular characteristic a single representative hedonic price is often of direct interest and, consequently, some kind of average derivative is usually needed (see McMillen and Thorsnes, 2003).

conventional result is that even the best parametric model tends to impose restrictions that substantially reduce the explanatory and predictive power of hedonic equation (Pace, 1993 & 1995; Anglin and Gencay, 1996; inter alia). Unless the theory-laden parametric model coincides with the data-generating process (this is a strong assumption), profound mis-specification errors may result imposing serious threats to their empirical validity.

Semiparametric and nonparametric approaches are representative of *algorithmic modelling culture* (Breiman, 2001) that emphasise aspects of learning the complex structure from the available facts and adaptability to the features underlying the data. They are particularly suitable for many hedonic modelling situations, where incomplete knowledge prevents the exact a priori specification of nonlinear or nonstationary components of functional form. Semiparametric estimators are, more precisely, an intermediate strategy between theory-laden and data-driven estimators that have restricted learning ability, i.e. semiparametric estimators can approximate functions only within some prespecified classes. Their practical relevance is mainly in balancing the dual goals of low specification error and high efficiency (Pace, 1995; Anglin and Gencay, 1996) and in enchaining the interpretability of results. Nonparametric estimators are by their nature highly flexible and, thus, capable of approximating very general classes of functions (e.g. smooth functions, square integrable functions) that neither require any restrictive, unwarranted prespecification of the functional form of the mean response function nor any specific error distribution assumption. This renders nonparametric estimators to be powerful data-driven tools, albeit highly sensitive to the problem of undersmoothing or overfitting, if local estimation is implemented unduly.

In the past fifteen years nonparametric and semiparametric approaches have significantly increased their popularity in different areas of modelling real estate phenomena. This section reviews some key results of the most commonly used nonparametric fitting methods in the real estate markets, which are often embedded within the semiparametric framework[6]. In real estate research, local polynomial modelling approaches have been popular: (1) the *Nadaraya-Watson estimator* (Pace, 1993 & 1995; Anglin and Gencay, 1996; Gencay and Yang, 1996; Maxam and Lacour-Little, 2001; McMillen and Thorsnes, 2003; inter alia) and (2) *locally weighted least squares* (Meese and Wallace, 1991 & 1997; Wallace, 1996; McMillen, 1996; Case et al., 2004; inter alia) represent typical choices that are relatively simple to use, yet effective in their inferences. These techniques are closely related; the Nadaraya-Watson estimator is, in fact, a special case of local polynomial fitting, where the local basis functions consist of the simplest polynomials, i.e. local constant functions (see Fan and Gijbels, 1996). The review focuses on the results of the local polynomial modelling approaches in the real estate markets.

[6] Common nonparametric procedures form the core of many semiparametric estimators, i.e. they are widely used in estimating the flexible part of a hedonic specification in partially linear models, in conditionally parametric models, etc.

Meese and Wallace (1991 & 1997) and Wallace (1996)[7] investigated locally weighted least squares in the problem of constructing residential house price indices. Meese and Wallace (1991) used large sample data over the period of 1970–1988 on housing sale prices and their characteristics for Alameda County with 136,782 observations and San Francisco County with 44,686 observations. They conducted different diagnostic tests and reported a generally good performance of local regression with an unusual choice of nearest neighbour bandwidth of one, which corresponds to a globally weighted least squares estimator. In particular, (locally) weighted regression "generated price index series with more reliable short-run dynamics than parametric approaches (double-log or translog model)". In a later, highly similar study (Meese and Wallace, 1997) provided strong evidence against the maintained assumptions underlying repeated sales method (no sample selection bias and constancy of implicit prices) by using locally weighted least squares. They used a nearest neighbour bandwidth choice and suggested that "a 30% window size provides the best balance between bias and sampling variability". Wallace (1996) applied locally weighted regression to the residential housing price index formulation for fifteen municipalities in Alameda County over the period of 1970 to first quarter of 1995 with 171,131 enclosed transactions. A nearest neighbour bandwidth of 0.33 was used and, in sum, the results indicated that "the nonparametric hedonic-based indexes provide better controls for the effect of quality evolution on price movements than alternative methods".

McMillen (1996) analysed locally weighted regression in modelling land values in Chicago using two different data sets from 1836 to 1990. The first data set spanned from 1830 to late 1920s with cross-sections of observations on price and its locational attributes over six different years with an average of approximately 120 observations per section. The second data set consisted of data for the years 1960–1990 with cross-sections of observations on price and its locational attributes over four different years with an average of approximately 710 observations per section. Two parametric models were estimated: a simple monocentric model and a more flexible spatial expansion model. These fits were compared to local linear regression estimates, which locally estimated the spatial expansion model. Nearest neighbour bandwidths of approximately 0.4 and 0.3 were applied, respectively, for the years 1836–1928 and the years 1960–1990. The results suggested that "the bulk of the improvement over simple negative exponential model estimates comes from additional explanatory variables in the spatial expansion, rather than from the increased nonlinearity of locally weighted regression estimates". However, as the simple monocentric model was rejected in the favour of a more complex spatial expansion model, local linear regression estimates, in turn, rejected both parametric models. McMillen also demonstrated that *local regression is useful for both prediction as well as testing hypotheses in land markets*. McMillen summarised: "Locally weighted regression is a useful tool for spatial modelling. Nonlinearity is handled directly and simply."

---

[7] The documentation about the polynomial order used in these studies is unclear.

More recently (Clapp, 2003) considered a partial linear model in the estimation of local house price indices. The hedonic model was specified in two parts: (1) ordinary least squares regression was performed to hold structural characteristics constant and, consequently, the usual implicit prices of housing characteristics were obtained; (2) a local linear smoother was used to calculate the location values flexibly as a function of northing and easting, which obviated the need for defining neighbourhood boundaries. The data set with 5,713 observations on all single-family transactions in six towns in Massachusetts from January, 1990 to April, 1999 was analysed. The estimates, which used locally adaptive bandwidth selection procedures, yielded "substantial, significant and plausible spatial patterns in location values". In particular, the mean squared prediction error was reduced between 5 and 11 percent by the chosen local regression model when compared to the results of the ordinary least squares estimator.

Clapp (2004) analysed a partially linear model for estimating local house price indices. In the spirit of the (Clapp, 2003), the log of housing sales prices were regressed on a vector of housing characteristics and three dimensional vector of northing, easting and time (annual dummy variables). The consistent two-step algorithm due to (Robinson, 1988; Stock, 1989) was used for estimation. In estimating the influence of the space-time cube, a local polynomial approximation with cross-validation criterion for local bandwidth selection was used. The data set with 49,511 transactions of singly-family properties sold in Fairfax county (Virginia) over the period from the first quarter of 1972 through the second quarter of 1991 was analysed and, in summary, Clapp concluded that "the LRM (local regression model) reduces out-of-sample mean squared error by 11% when compared to the OLS (ordinary least squares). Also the utility of the LRM in conducting explorative analysis was strongly supported by the data.

Highly similar to (Clapp, 2003) is the investigation due to (Clapp, Kim and Gelfand, 2002) who applied the partial linear modelling approach; local regression was used to estimate the effect of two-dimensional vector of northing and easting on house prices. A data set covering 2,338 observations on single-family transactions with the associated characteristics from January, 1996 to April, 1999 in the area of Massachusetts was analysed using a partially Bayesian approach in order to capture the uncertainty associated with inferences. The results of post-sample predictive testing are similar to (Case, 2003 & 2004): out-of-sample mean squared error was reduced by 9%.

Pace (1993) investigated the Nadaraya-Watson kernel estimator under different variable transformations (the Box-Cox transformation) and across data subsets differing in quality. Two sets of data were analysed; the first data set consisted only of 32 observations on rental rates with four attributes in the area of Michigan. The second one consisted of 506 observations of housing prices with thirteen attributes relating to different aspects of pollution effects (on housing prices). Over the both data sets, the Nadaraya-Watson kernel estimator outperformed ordinary least squares estimator in post-sample predictive testing. The results also suggested that this nonparametric estimator was more robust

than the ordinary least squares estimator. Pace (1995) compared parametric (ordinary least squares), semi-parametric (index model) and nonparametric (the Nadaraya-Watson) estimators with price data on single-family dwellings and four attributes over a six-month period in 1986, which consisted of 379 observations. Transformed (log-price) and untransformed (price) responses were considered. Results indicated that all three estimators yielded estimates with plausible magnitudes and signs. However, the index model estimator exceeded or equalised the other estimators' performance in terms of frequency of ex-ante sign violations. The Nadaraya-Watson kernel estimator outperformed or matched the ordinary least squares.

Anglin and Gencay (1996) considered a semiparametric estimating technique based on the Nadaraya-Watson kernel estimator and a two-stage procedure due to (Robinson, 1988; Stock, 1989) for separating the pieces of information between linear and nonlinear part of the hedonic model. A global cross-validation procedure was used to establish the proper bandwidth choice. A data set of 546 observations on residential housing sale prices with the associated characteristics during the three-month period in 1987 in the county of Windsor and Essex was analysed. They concluded that their semiparametric model significantly outperformed the best parametric alternative in post-sample prediction testing, and that *the common specification tests used in parametric modelling may not be as powerful as desired*. The same methodology as in (Anglin and Gencay, 1996) was used by (Gencay and Yang, 1996) to study residential housing prices. They produced similar results; in particular, they suggested that the chosen semiparametric models provided much smaller mean squared prediction error than the para-metric alternatives (Box-Cox and Wooldridge transformations).

More recently (Maxam and LaCour-Little, 2001) used the Nadaraya-Watson kernel estimator for estimating mortgage loan prepayments. Their nonparametric technique was shown to exhibit superior out-of-sample predictive ability when compared to both proportional-hazards and proprietary-practitioner models. They also documented that their estimation procedure led to parsimonious representations for the phenomena. The same methodology was applied in (LaCour-Little et al., 2002) with similar conclusions. In particular, the Nadaraya-Watson estimator was shown to be superior to conventional methods in detecting nonlinearities and multivariate interactions, and that they provided the best out-of-sample predictions among the candidate models. McMillen and Thorsnes (2003) considered a highly important methodological issue of average derivate estimation in situations, where time can be treated as a nearly continuous variable. They built on the idea of Robinson's (1988) semiparametric estimator and used the Nararaya-Watson kernel estimator as a part of a three-stage procedure to derive time-varying coefficients measuring the effect of a copper smelter on house prices in Tacoma, Washington. These effects were analysed over a long period, 1977–1999 with a sample consisting of 11,365 observations on house price and the related attributes. In summary, their

semiparametric estimates provided a more plausible description of the average effect of a copper smelter on house prices than discrete-time dummy variables.[8]

3 Local Regression

Local regression is a *nonparametric method* of fitting surfaces to data by smoothing in which parametric functions are locally estimated in the attribute space using weighted least squares (Cleveland and Devlin, 1988; Cleveland and Grosse, 1991; Cleveland and Loader, 1996). It is a *flexible*, *data-driven approach* letting "the data speak for themselves" in order to determine the most appropriate form of hedonic function. *Parametric localization* is the fundamental feature underlying local regression: the fit at a given point is the value of a parametric function fitted only to those observations in a neighbourhood of that point. The basic principle is to give more weight to nearby observations and lesser weight to those that are further away.

Local regression is most useful when the hedonic function is nonlinear and the form of this nonlinearity is unknown a priori. In land valuation studies the correct functional form is rarely known in advance, although the regression surface is expected to be spatially inhomogeneous with varying degrees of curvature (McDonald & Bowman, 1979; McMillen, 1996; inter alia). An obvious way to proceed is to use flexible functional forms that can account for complicated spatial, as well as temporal[9], patterns in land values and to infer the form of the hedonic relationship from the available evidence. The main objective in local regression is to smooth the data sufficiently, so that non-systematic variability is ignored, while maintaining as much genuine structure as necessary to produce unbiased results (McMillen, 1996). This culminates in the well-known *bias-variance trade-off*: the bias can be reduced only at the expense of variance, and vice versa.

When using local regression, one has to specify certain properties of the regression surface and error terms, i.e. to make some assumptions about them. The data-generating process is here assumed to be of form:

$$p_i = f(\mathbf{x}_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

[8] There are also some comprehensive land values studies conducted in the Finnish land markets using parametrised hedonic models that are of direct relevancy to this study. These include the investigations of Leväinen (1991), Heinonen (1993) and Hiltunen (2003).

[9] An assumption that the effect of time can be captured by some measurable variable is made in the modelling of temporal dimension of land prices. In particular, the dynamics are expected to be adequately encapsulated by an error correction term.

where $\mathbf{x}_i$ is a $d+1$ vector of quantitative and qualitative attribute variables[10]. ε_i are independent normal[11] random variables with mean 0 and variance σ^2. f is assumed to be a smooth function of the attributes that is well approximated in a certain neighbourhood of $\mathbf{x}$ by a function from a parametric class. Specifically, the chosen local function is here a representation of the class of cubic, quadratic or linear polynomials: the chosen parametric class for static hedonic models corresponds to local double-log model (local linear model), local translog model (local quadratic model) or local cubic expansion model (local cubic approximation); for dynamic local error correction models linear, quadratic and cubic approximations are tested *in two different scenarios*: (1) error correction term is constructed as a linear combination of the original attributes and (2) error correction component is estimated flexibly as a nonlinear transformation of the original characteristics.

3.1 Mathematical Formulation

The local regression problem can be formalized by using locally weighted least squares (e.g. Ruppert and Wand, 1994; Loader, 2004):

$$\textit{Minimize} \sum_{i=1}^{n} \mathbf{W}_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x})\big(p_i - \langle \boldsymbol{\theta}, \mathbf{F}(\mathbf{x}_i - \mathbf{x})\rangle \big)^2 \tag{2}$$

where $\boldsymbol{\theta}$ is the $d+1$ vector of unknown coefficients and $\mathbf{F}(\cdot)$ is a vector of basis polynomials. $\mathbf{W}_{\mathbf{H}}$ is a multivariate *weight function* and $\mathbf{H}^{1/2}$ is a *bandwidth matrix*. The local least squares estimate of the unknown regression function $f(\mathbf{x})$ is then[12]:

$$\hat{f}(\mathbf{x}) = \mathbf{e}_1'\big(\mathbf{X}'\mathbf{W}\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{W}\mathbf{p} \tag{3}$$

For local cubic regression $\mathbf{e}_1$ is a $\left\{1+d+\tfrac{1}{2}d(d+1)+\tfrac{1}{6}d(d+1)(d+1)\right\}\times 1$ vector having 1 in the first entry and all other entries 0. For local quadratic and linear model the dimension of $\mathbf{e}_1$ is, respectively, $\left\{1+d+\tfrac{1}{2}d(d+1)\right\}\times 1$ and

---

[10] There is no consensus view on how to model the effects of indicator variables in nonparametric or semi-parametric models. According to (McMillen and Thorsnes, 2003): "Some studies account for them by adding a parametric component to the equation, whereas others restrict the analysis to a subset of the sample in which all dummy variables take on a particular value. Still other studies ignore the problem altogether by analyzing all data without regard to the dummy variables."

[11] These conventional error assumptions can be relaxed just as in the parametric counterpart.

[12] Assuming, as usual, that $\mathbf{X}'\mathbf{W}\mathbf{X}$ is non-singular.

$\{1+d\}\times 1$. $\mathbf{p}=[\,p_1,\ldots,p_n\,]'$ is a vector of observed land prices and the *data matrix* $\mathbf{X}$ for the local cubic model is:

$$\mathbf{X}=\begin{bmatrix} 1 & (\mathbf{x}_1-\mathbf{x})' & vec'\{(\mathbf{x}_1-\mathbf{x})(\mathbf{x}_1-\mathbf{x})'\} & (\mathbf{x}_{(1)}\otimes\mathbf{x}_{(1)}\otimes\mathbf{x}_{(1)})' \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (\mathbf{x}_n-\mathbf{x})' & vec'\{(\mathbf{x}_n-\mathbf{x})(\mathbf{x}_n-\mathbf{x})'\} & (\mathbf{x}_{(n)}\otimes\mathbf{x}_{(n)}\otimes\mathbf{x}_{(n)})' \end{bmatrix} \quad (4)$$

where $\mathbf{x}_{(i)}=(\mathbf{x}_i-\mathbf{x})$ and the *vec*-operator stacks the columns of $(\mathbf{x}_i-\mathbf{x})(\mathbf{x}_i-\mathbf{x})'$, each below the previous, but with entries above main diagonal omitted; $\otimes$ is the Knonecker (tensor) product. The local linear and quadratic model uses only the first two and three, respectively, columns of the data matrix. The weight matrix is $\mathbf{W}=diag\{\mathbf{W_H}(\mathbf{x}_1-\mathbf{x}),\ldots,\mathbf{W_H}(\mathbf{x}_n-\mathbf{x})\}$ $\equiv diag\{w_1(\mathbf{x}),\ldots,w_n(\mathbf{x})\}$.

5 Advantages of Local Regression

Local regression has several benefits that, *when combined*, make it an attractive tool for modelling complex land value patterns. These include (Hastie and Loader, 1993; Fan and Gijbels, 1995; Fan and Gijbels, 1996; Cleveland and Loader, 1996):

1. Attractive minimax properties: the asymptotic minimax efficiency is 100% for commonly used orders among all linear estimators and only marginal loss beyond this class.
2. Adaptation to various types of designs such as random designs, fixed designs, strongly clustered and nearly uniform designs. It does not require specific smoothness or regulatory conditions.
3. Easy to understand, interpret and implement. It can utilise the existing statistical theory of least squares regression in a natural way.
4. Adaptation to boundary bias problem and biases in regions of high curvature.
5. Adaptation to derivate estimation and spatially inhomogeneous surfaces.
6. Fast computational methods exist for multivariate smoothing that can also be tailored to different distributional assumptions. Generalises directly to local likelihood estimation framework.
7. Enables straightforward derivation of response adaptive methods for bandwidth and polynomial order selection.
8. Has exhibited strong empirical performance in explaining and predicting phenomena in different fields of study.

6 Components of Local Regression

There are four basic components that have to be specified in practice when using local regression. These are (Cleveland and Loader, 1996; Loader, 1999):

– bandwidth;

– weight function;
– local parametric family;
– fitting criterion.

6.1 Bandwidth

The *bandwidth* or smoothing parameter has a critical influence on the performance of local fitting as it significantly controls the model complexity, i.e. the amount of smoothing being performed. Overly small bandwidths lead to undersmoothed estimates that overfit to technical details of the data, while unacceptably large bandwidths oversmooth the data set ignoring relevant systematic features of the dependence of response on the attribute variables. The optimal bandwidth is the one that strikes a balance between bias and variance elements. In real estate valuation practice a natural interpretation can be given to the bandwidth parameter, which sets the number of effective comparables used in the modelling (Pace, 1993).

There are different ways to proceed in specifying the bandwidth parameter: Nearest-neighbour bandwidths are widely applied due to their relative simplicity. In real estate markets typical choices of nearest-neighbour parameters have been from 0.3 to 0.4 (McMillen, 1996; Wallace, 1996; inter alia). Global cross-validation (Allen, 1974) or generalized cross-validation (Craven and Wahba, 1979) scores, which are motivated by prediction error assessment, are also commonly utilised in bandwidth selection problems. Cross-validation based bandwidth selection rules typically lead to significantly lower values for smoothing parameters (see e.g. Thorsnes and McMillen, 1998) and are often susceptible to the overfitting problem. Indeed, in many instances, the nearest-neighbour bandwidths or global cross validation methods fail to provide a satisfactory fit, and as a result, a different approach to bandwidth selection is often needed.

In this study locally adaptive bandwidth selection procedures in the spirit of (Clapp, 2003 & 2004) are used to control model complexity and, in particular, to choose a local smoothing parameter. To derive locally adaptive bandwidths, *the local generalized CP criterion* is used as a benchmark measure of local goodness of fit. The local generalized CP criterion (with variance penalty α) is defined (Cleveland and Devlin, 1988; Cleveland & Loader, 1996):

$$LoCP(\hat{f}) = \frac{1}{tr(\mathbf{W})}\sum_{i=1}^{n}\frac{w_i(\mathbf{x})}{\sigma_i^2}\left(p_i - \hat{f}(\mathbf{x}_i)\right)^2 - 1 + \alpha\frac{\upsilon(\hat{f})}{tr(\mathbf{W})} \; . \qquad (5)$$

If $\alpha = 2$, this gives an unbiased estimate of the sum of the squared error over the design points, and:

$$\upsilon(\hat{f}) = tr\left(\left(\mathbf{X}'\mathbf{W}\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^2\mathbf{V}\mathbf{X}\right) \qquad (6)$$

is the localised fitted degrees of freedom with $\mathbf{V}$ being a diagonal matrix with entries $1/\sigma_i^2$.

6.2 Weight Function

The form of a *weight function* has much less dramatic influence on the bias-variance dilemma (see e.g. Thorsnes and McMillen, 1998). Typically, the weight function is chosen to be continuous, symmetric, peaked at 0 and supported on [–1, 1], which assures that the most weight is given to observations close to the fitting point $\mathbf{x}$. Here the standard *tricube function* is applied that determines the weight for the observation $(p_i, \mathbf{x}_i)$ to be:

$$w_i(\mathbf{x}) = \left(1 - \left(\frac{\rho(\mathbf{x}, \mathbf{x}_i)}{h_i} \right)^3 \right)^3, \tag{7}$$

where $\rho(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{d} \left(\frac{x_j - x_{ij}}{s_j} \right)^2}$ is the scaled Euclidean distance between $\mathbf{x}$ and $\mathbf{x}_i$. Specifically, $s_j \equiv 1$ as the observations in this paper are measured on the logarithmic scale and further scaling is therefore unnecessary.

6.3 Local Parametric Family

The underlying principle of local regression is that the unknown, smooth regression function can be well approximated locally by a member of a parametric class, frequently taken to be polynomials of a certain degree. A high polynomial degree can always provide a less biased approximation to underlying functional relationship but only at the expense of increased variability in the estimate. Therefore, *like the bandwidth choice, the degree of the local polynomial affects significantly the bias-variance dilemma and, to some extent, their relative influences are confounded*.

Asymptotic theory (see e.g. Ruppert and Wand, 1994) suggests that polynomials of odd degree beat those of even degree. Indeed, local linear approximation can almost invariably be preferred to local constant fitting (Nadaraya-Watson kernel estimator), but for local quadratic and cubic fitting the asymptotics perform very poorly (Cleveland and Loader, 1996). The most natural way to proceed in deciding the appropriate polynomial degree is to utilize available knowledge on global parametric fitting[13]. These considerations (e.g. Hannonen, 2005) suggest using *linear, quadratic or cubic approximations*. The exact preference between the different polynomial orders is a priori somewhat uncertain, albeit local quadratic and cubic approximations seem to be

[13] It might sometimes be useful to consider adaptive methods that can locally select the degree of polynomial, but these issues are not considered here.

preferable as they are expected to reproduce better the peaks and valleys present in the data and, on the other hand, the use of quadratics and cubic polynomials relies on less stringent economic considerations. However, e.g. Clapp (2003) documents that the use of linear approximation resulted a somewhat better out-of-sample MSE than the local constant or quadratic fitting.

6.4 Fitting Criterion

Virtually any global fitting criterion can be localized and, therefore, local regression can proceed with the same rich collection of distributional assumptions as in the global case. In this study, error terms are assumed to be independent normal random variables with mean 0 and variance σ^2, which leads conveniently to the least squares criterion. A major problem in practice with the least squares criterion is its high sensitivity to outlying observations, which are very common in studies of land values. In this study the idea of *iterative downweighting* (Cleveland, 1979) is used for robustifying the results of local regression by implementing a variant of local M-estimation that uses the input of locally weighted least squares regression. It is easy to implement and usually gives enough protection against extreme observations.

7 Computation of Local Regression Model

The computational method for estimating the fit at points of evaluation is based on a damped Newton-Raphson algorithm (Loader, 1999, pp. 209–211; Loader, 2004):

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k + \frac{1}{2^j}\mathbf{J}^{-1}\mathbf{X}'\mathbf{W}\mathbf{q}\,. \tag{8}$$

$\tilde{\boldsymbol{\theta}}_{k+1}$ is an estimate of the parameter vector $\boldsymbol{\theta}$ at the $k+1$ iteration. Here j is selected to be the smallest non-negative integer that results in an increase of the local log-likelihood at every step, and $\mathbf{q}$ is the usual score function. Furthermore, the Jacobian matrix can be expressed as $\mathbf{J} = \mathbf{D}^{½}\mathbf{P}'\boldsymbol{\Sigma}\,\mathbf{P}\mathbf{D}^{½}$, where $\mathbf{D}$ is the matrix of diagonal elements of $\mathbf{J} = \mathbf{X}'\mathbf{W}\mathbf{V}\mathbf{X}$ with $\mathbf{P}$ and $\boldsymbol{\Sigma}$ representing, respectively, the eigenvectors and eigenvalues of $\mathbf{D}^{½}\mathbf{J}\,\mathbf{D}^{½}$. $\mathbf{V}$ is the diagonal observed information matrix. Since the sample sizes and the dimensions of attribute spaces are small in this study, direct evaluation of fit is potentially feasible (although not applied) using the recursion of (8), which would mean that separate weighted least squares regression is estimated for each data point, with more influence given to nearby observations.

Direct evaluation is, however, computationally infeasible for larger data sets[14] and for increased dimensions of attribute space; a general algorithm is

[14] Also visualisation of regression surfaces, variance functions etc. demands that a separate, reduced number of fitting points are selected.

needed to perform the selection of evaluation points, where the fit is subsequently estimated. Tree-based structures are popular; in particular, the k-d-trees due to (Friedman, Bentley and Finkel, 1977) or growing adaptive trees (see, inter alia, Loader, 1999, pp. 212–218). Different interpolation schemes (e.g. blending functions) can be used to define the fit elsewhere.

8 Prediction, Hypothesis Testing and Calculation of Hedonic Prices

Usually, the main objective of nonparametric modelling is to predict the value of the response conditional on a set of observed attribute variables rather than test any particular hypothesis. The *predicted value of response* is calculated simply:

$$\overline{p}_j = \mathbf{x}'_j \hat{f}(\mathbf{x}_j) = \mathbf{x}'_j \left(\sum_{i=1}^{n} w_i(\mathbf{x})\, \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i=1}^{n} w_i(\mathbf{x})\, \mathbf{x}_i p_i \,, \tag{9}$$

which is a well-defined, closed-form mathematical expression. Given the distributional assumptions underlying the data-generating process in (1), prediction intervals can easily be constructed:

$$CI(\mathbf{x}_j) = \left(\overline{p}_j - c\, s \left\| \boldsymbol{l}(\mathbf{x}_j) \right\|,\ \overline{p}_j + c\, s \left\| \boldsymbol{l}(\mathbf{x}_j) \right\| \right), \tag{10}$$

where $\boldsymbol{l}(\mathbf{x}_j) = \left\{ l_i(\mathbf{x}_j) \right\}_{i=1}^{n}$ is a weight diagram vector such that $\overline{p}_j = \sum_{i=1}^{n} l_i(\mathbf{x}_j) p_i$, i.e. *predicted response values are simply quality-weighted averages of all observed in-sample land prices*. c denotes an appropriate quantile of the standard normal distribution, e.g. $c = 1.96$ for 95% confidence. s is the standard error of the regression (see below). Now:

$$l_i(\mathbf{x}_j) = \mathbf{x}'_j \left(\sum_{i=1}^{n} w_i(\mathbf{x})\, \mathbf{x}_i \mathbf{x}'_i \right)^{-1} w_i(\mathbf{x})\, \mathbf{x}_i \,. \tag{11}$$

The hat matrix $\mathbf{H}$ is the $n \times n$ matrix that takes the vector of observations of the response to the vector of fitted values. Its $(j, i)^{th}$ element is $l_i(\mathbf{x}_j)$ so that:

$$\overline{\mathbf{p}} = \mathbf{H}\mathbf{p} \ \text{ and } \ \mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{p} \,, \tag{12}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{W}$. Normal-based inferences about the unknown f utilize two quantities:

$$\delta_k = tr((\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H}))^k \tag{13}$$

for $k = 1$ and $k = 2$. The *standard error of regression* can now be estimated:

$$s = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{\delta_1}} \ . \tag{14}$$

Furthermore δ_k for $k = 1$ and $k = 2$ can be used to construct an approximate F-test statistic, which is a useful statistical indicator when errors are normally distributed (e.g. Cleveland and Devlin, 1988; McMillen, 1996).

The *exact hedonic prices* can be calculated, if needed, separately for each evaluation point using (Loader, 1999):

$$\hat{f}'(\mathbf{x}) = \mathbf{e}_2' \left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{p} \ + \ \mathbf{e}_1' \left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1} \mathbf{X}' \frac{\partial \mathbf{W}}{\partial \mathbf{x}} \left(\mathbf{I} - \mathbf{H} \right) \mathbf{p} \, , \tag{15}$$

In other words, the exact derivatives are sums of the local slope estimates and weighted residuals. $\mathbf{e}_2$ is otherwise as defined in (3), but having the value 1 in the second entry and zero elsewhere. Often $\mathbf{e}_2' \left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{p}$, the local slope estimate, is used as an estimate of the exact derivatives, which is a reasonable approximation when observed residuals $\left(\mathbf{I} - \mathbf{H} \right) \mathbf{p}$ are negligible.

9 Empirical Analysis of Local Hedonic Models

This section considers the empirical applicability of local modelling tools to the *prediction* of observed prices of undeveloped land in two different Finnish submarkets (Espoo and Nurmijärvi). The estimation and testing of hedonic prices and, in particular, the problematic issues of estimating average hedonic prices are not tackled in this paper.

Parametric localisation is the fundamental characteristic underlying much of the following empirical research in this section, which enables hedonic analysis to proceed with less restrictive and unrealistic a priori assumptions that is expected to reduce any strong theory dependency of results. One of the main attractions of the local modelling scheme is that it can effectively utilise knowledge from underlying economic rationales of global modelling, which makes local regression *at least a low-level theory consistent*, highly flexible tool for hedonic modelling, not a "black box" in any sense. In this paper parametric localisation is achieved by applying locally weighted least squares to some useful and well-known global hedonic models of land prices (McMillen, 1996; Hannonen, 2005): (1) double-log model, (2) translog model, (3) cubic expansion model and (4) different error correction models. Subsequently, the predictive ability of these localised hedonic models is analysed in depth and compared to predictions of their global analogues.

10 Sample Data

Empirical studies of land prices tend to exhibit significant sensitivity to changes in data, i.e. different submarket, time period and land use can lead to widely differing results, even in the context of unified methodology. The majority of that variation is explained by spatio-temporal movements: functional forms and parameters tend to vary with location and are not homogeneous throughout the data set, whereas temporally changing market conditions cause data-generating processes to evolve over time. To reduce the sample dependency, i.e. to improve the invariance of empirical study, the paper examines two data sets that are located in different submarkets and associated with partially non-overlapping time frames. The land type (land use) is, in contrast, fixed in order to reduce unnecessary heterogeneity of land prices. It represents undeveloped land that has not yet reached its highest and best use: vacant sites without a local detailed plan that are reserved for residential housing purposes.

The first sample data involve observations on land prices and the associated characteristics in the municipality of Espoo, a highly polycentric city, which lies inside the Helsinki metropolitan area with circa 225,000 habitants; its population is the second largest of the cities in Finland, which has experienced a rapid growth in its late history. The study period is from January, 1990 to December, 2001 with a total number of observations of 400. The observations from the last year (total of 39) are held back for post-sample predictive testing. In terms of quality, this data set is preferable over the second sample, i.e. it has been pre-checked for any errors and analysed before (Hannonen, 2005). In table 1 are documented some standard sample statistics for the study variables in the case of Espoo.

The second sample contains observations of land prices and the associated attributes on 793 unbuilt land parcels without a local detailed plan sold in the period spanning from January, 1985 to March, 2004 in the municipality of Nurmijärvi, which lies just outside the Helsinki metropolitan area with approximately 36,000 habitants and three distinctive population centres (the parish village, Klaukkala and Rajamäki). Nurmijärvi has recently also witnessed several years of rapid expansion. The last 50 observations were left aside for evaluating the predictive accuracy of estimated local models outside the estimation sample, which corresponds approximately to a one-year period from the late February 2003 to the early March 2004. In terms of quantity, this data set offers more opportunities to flexible modelling than the Espoo case, albeit it is more pronounced to any errors (e.g. recording errors), since it has not been pre-checked for hedonic modelling purposes. In table 2 are documented some standard sample statistics for the study variables in the Nurmijärvi case.

***Table 1**. Some common sample statistics for the municipality of Espoo.*

Variable (unit)	**Arithmetic mean**	**Minimum**	**Maximum**	**Std. Deviation**
Total price (€)	59,126.40	3,027.00	756,846.00	61,976.88
Square price (€/m^2) (unit price)	22.99	0.24	127.55	19.09
Parcel size (m^2)	4,207.49	1,000.00	28,400.00	4,613.75
Distance to CBD of Helsinki, L_2-metric (km)	17.22	7.61	27.29	4.34
Northing	667,922.61	666,540.00	669,650.00	695.24
Easting	253,796.25	252,815.00	254,638.00	406.98
Quarterly price index of single-family houses	154.06	116.80	187.30	22.35
Parcel type (=0 if whole site; 1 otherwise)	–	0	1	–

***Table 2**. Some common sample statistics for the municipality of Nurmijärvi.*

Variable (unit)	**Arithmetic mean**	**Minimum**	**Maximum**	**Std. Deviation**
Total price (€)	22,019.03	673.00	479,336.00	21,262.25
Square price (€/m^2) (unit price)	3.80	0.34	22.83	3.27
Parcel size (m^2)	7,387.36	1,000.00	30,000.00	4,651.66
Distance to CBD of Helsinki (km)	33.11	22.28	44.91	5.22
Distance to parish village of Nurmijärvi (km)	8.55	0.32	16.06	3.37
Distance to population center of Klaukkala (km)	9.33	0.50	21.44	5.00
Distance to population center of Rajamäki (km)	11.86	0.43	20.92	4.58
Northing	67,0354.00	669,240.00	671,736.00	600,600.00
Easting	25,4091.50	253,031.00	255,695.00	250,546.70
Quarterly price index of single-family houses	154.30	100.00	226.00	32.00
Parcel type (=0 if whole site; 1 otherwise)	–	0	1	–

11 Specification of Local Hedonic Models

This subsection considers some of the most prominent global hedonic models of land prices (Hannonen, 2005) that are subsequently localised, which enables estimated local counterparts to be *simultaneously linear, convex or concave in different regions of attribute space*, if needed. Localisation is implemented by

using locally weighted least squares with adaptive bandwidth selection to each model candidate separately.

The most obvious global model to consider is the conventional *multiplicative form of the double-log model*:

$$p_i = e^{\beta_0} \prod_{j=1}^{k} x_{ij}^{\beta_j} \prod_{j=1}^{s} e^{\gamma_j d_{ij}} e^{\varepsilon_i} = e^{\beta_0} x_{i1}^{\beta_1} x_{i2}^{\beta_2} \cdot \ldots \cdot x_{ik}^{\beta_k} e^{\sum_{j=1}^{s} \gamma_j d_{ij}} e^{\varepsilon_i} \;\; \forall \;\; i \in \mathrm{U} \subset n, \quad (16)$$

which has k quantitative attributes x_{ij} and s qualitative characteristics d_{ij}; p_i is the total sales or unit price of a land parcel. This classical double-log form has a constant elasticity property with an advantage that the price elasticities are independent of the unit of measurement. It is intrinsically linear since the logarithmic transformation for positive real values leads to a hedonic model which is linear in the parameters. In many studies of land prices, only the observed total price (or unit price) – and possibly the parcel size – is transformed to the logarithmic scale; the rest of the model's characteristics enter structure linearly. However, this practice is problematic mainly because it biases mean land prices when the back-transformation is made to the original scale. Sometimes the feature of multiplying the powers of explanatory variables together is tentatively undesirable and as reconciliation the mutual effect of regressors is expected to be additive, rather than multiplicative. This yields to an additive form of double-log model, which is intrinsically a nonlinear specification. However, this additive form seems to be empirically inferior to the multiplicative one in modelling land prices (Hannonen, 2005).

In the *translog model* the assumption of unitary elasticity of substitution between land parcels is relaxed. The classical double-log model is obtained by the restriction that all the parameters relating to second-order interactions of model variables are zero. From the perspective of land markets the underlying theoretical rationale of the translog model is convenient since land parcels are heterogeneous products that cannot be considered as perfect substitutes. The translog model can be formulated as:

$$p_i = e^{\beta_0} \prod_{j=1}^{k} x_{ij}^{\beta_j} \prod_{j=1}^{s} e^{\gamma_j d_{ij}} \prod_{j=1}^{k} \prod_{l=1}^{k} \left(e^{\beta_{jl} \ln x_{ij} \ln x_{il}} \right)^{\frac{1}{2}} \prod_{j=1}^{s} \prod_{l=1}^{s} \left(e^{\gamma_{jl} d_{ij} d_{il}} \right)^{\frac{1}{2}} e^{\varepsilon_i} \;\; \forall \;\; i \in \mathrm{U} \subset n, \quad (17)$$

which undergoes a logarithmic transformation for estimation purposes.

The *cubic expansion model* further expands the logic underlying the translog model to allow for third-order interactions, if needed. In land valuation many elements relating to physical space or distance measures are sometimes considered to be so highly nonlinear that cubic terms are needed[15].

[15] The awkward mathematical formulation is not reproduced here; It simply builds on the translog model in a similar vain as the double-log model is expanded to the translog model (see e.g. McMillen, 1996).

The conventional double-log model, the translog model and the cubic expansion model can be understood as representations of the class of static hedonic models, which can adequately be used for modelling stationary processes. However, there seems to be evidence that behaviour of land prices is nonstationary (Hannonen, 2005) so that the error correction models could be incorporated as useful model contenders. Equilibrium (or error) correction mechanisms are generalisations of standard growth-rate models, where modelling is carried out in differenced values instead of the original levels, combined with an appropriate error correction term. These error correction models can separate the impact effect, the feedback effect and the long-run response, and thus incorporate the important concept of *cointegration* directly into the model formulation. Cointegration itself represents a certain kind of nonstationary behaviour that tends to describe many economic time series data. A *standard error correction model* can be formulated as:

$$\Delta ln(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \Delta ln(x_{ij}) + \sum_{j=1}^{s} \gamma_j d_{ij} + \boldsymbol{\theta}' \mathbf{z}_{i-1} + \varepsilon_i \quad \forall \; i \in \mathrm{U} \subset n \;, \quad (18)$$

where $\boldsymbol{\theta}$ denotes an error correction parameter vector; $\mathbf{z}_{i-1}$ represents a combination of attributes that possess a long-term relation to response vector and Δ is a difference operator. However, *there is no need to restrict*, on the one hand, $\boldsymbol{\theta}$ *to be a linear combination of relevant long-run characteristics or*, on the other hand, *that short-term and long-term impacts are independent in any ways*. For example, if a local quadratic approximation is made:

$$\begin{aligned} \Delta ln(p_i) = \beta_0 &+ \sum_{j=1}^{k} \beta_j \Delta ln(x_{ij}) + \sum_{j=1}^{s} \gamma_j d_{ij} + \frac{1}{2} \sum_{j=1}^{k} \sum_{l=1}^{k} \beta_{jl} \Delta ln(x_{ij}) \Delta ln(x_{il}) \\ &+ \frac{1}{2} \sum_{j=1}^{s} \sum_{l=1}^{s} \gamma_{jl} d_{ij} d_{il} + \frac{1}{2} \sum_{j=1}^{k} \sum_{m=1}^{h} \beta_{jm} \Delta ln(x_{ij}) f(\theta_m z_{im}) \\ &+ \frac{1}{2} \sum_{j=1}^{s} \sum_{m=1}^{h} \gamma_{jm} d_{ij} f(\theta_m z_{im}) + \varepsilon_i \;, \end{aligned} \quad (19)$$

might be a useful starting point, where f is a member of a polynomial class and h denotes the dimension of cointegrating vector.

12 Forecasting Accuracy of Local Hedonic Models

12.1 Measures of Predictive Accuracy

Forecasting accuracy is a critical factor in the evaluation of empirical validity of the chosen hedonic model. There are numerous different indicators for post-sample predictive assessment of hedonic models (e.g. Case et al., 2004) and the relative ranking of the performance of various models varies according to the applied accuracy measure. However, not all measures can be considered equally

informative in the context of hedonic modelling of land prices: *of paramount importance are those criteria underlying the mean prediction error and the strength of association between the model's predictions and observed land prices*. Variance-based indicators, albeit commonly utilised in studies, are secondary for most practical land valuation purposes.

Mean prediction error is evaluated in this study by (1) arithmetic average prediction error, (2) median prediction error, (3) weighted arithmetic average and median prediction errors, (4) error centre of gravity and (5) by arithmetic average and median percentage errors. If unit land prices are analysed (as in this paper), the weights are directly obtained by using parcel size to give the total mass of errors, which obviates the need of estimating the unknown density function. The use of average prediction errors is often, in land value studies, a more plausible alternative than median-based indicators of central tendency of prediction errors, since latter measures down-weight too heavily the influence of unusual prediction errors as compared to typical prediction errors since they ignore relevant information about distances between errors. Also median-based mean error indicators tend to produce, in many cases, highly similar and indecisive results between different model candidates. The major problem of using arithmetic average prediction error or median prediction error, and their weighted analogues, is that these are absolute measures, which depend on the unit of measurement. Therefore, the relative versions of the basic indicators, i.e. arithmetic average and median percentage errors, are also documented for comparing mean prediction errors between models, where responses are measured on a different scale.

Three measures of strength of the association between predictions and observed out-of-sample land prices are reported. First, the usual correlation coefficient is calculated, which is a useful measure of statistical relation in the case of normally distributed error terms and when the focus is on the co-variation of errors. The major problem of using the classical correlation measure in land valuation studies lies in its strong dependency on the normality assumption, which is typically violated by the influence of aberrant error terms, whose effect is squared in the denominator, which, in turn, tend to lead to highly similar standard deviations between different model alternatives. Consequently, calculated correlation coefficients tend to produce not only invalid but also indecisive results. The use of the scaled conventional inner product (or some weighted version of it, but this seems to add very little new information) is preferable, which measures the strength of prediction in the direction of actual out-of-sample land prices scaled by the total uncentered variation of land prices. It generates more valid and decisive results that are not strongly dependent on any particular distributional assumptions. Also the gravity (see McMillen, 2001) is reported, which seems to be a viable measure of strength of association; it usually produces results that parallel the use of the scaled inner product.

Root mean squared error is the most commonly used measure of success of numeric prediction, which mainly controls the reliability or variability of predictions, not the actual predictive validity or predictive unbiasedness. Furthermore, this statistic is overly sensitive to outlying observations tending to

exaggerate the variance of prediction errors of model choices in which the prediction error is larger than the others (which is typical in land valuation studies). This sensitivity or lack of robustness to extreme errors can also cause the root mean square error to produce results that are indecisive. Mean absolute error is generally a more appropriate indicator of predictive variability, and is especially suitable in cases of outlying prediction errors. Another useful measure of predictive variability is simply the predictive range expressed in terms of minimum and maximum predictions; the shorter the interval, the less variability is expected in predictions. A widely used measure of predictive variability is mean absolute percentage error (see e.g. Makridakis and Hibon, 2000), which is reported here with the associated robust version of it: median absolute percentage error. They have also been criticized for the problems of asymmetry and instability, when the data is small (as here).

13 The Espoo Case

This subsection reports the main results of out-of-sample predictive testing for land prices in the municipality of Espoo. Unit land price was selected as the proper response, since this seemed to yield superior results to those models using total sales price of a land parcel as the dependent variable. Table 3 documents the forecasting results from the localised versions of double-log, translog and cubic expansion models. They indicate *highly nonlinear relationship between land prices and* the best fitting *attribute variables* (parcel size, distance measure and parcel type; see (Hannonen, 2005). The localised cubic expansion hedonic model encompasses the local double-log and translog models in most measurable aspects; fifteen (out of seventeen) different measures of predictive accuracy are superior (and all the fundamental ones) in the case of localised cubic expansion model. Beyond any reasonable doubt, the superiority of empirical validity of the cubic expansion model over the other investigated local model structures is strongly supported by the data. The results of table 3 should be contrasted with the results of table 4, which document the corresponding measures from the global alternatives.

Table 3. *Explanatory variables: parcel size, L_2-distance to CBD of Helsinki and parcel type.*

Measure of predictive accuracy	Local hedonic model		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	0.12	0.093	0.071
Median prediction error	0.36	0.31	0.26
Weighted average error	0.96	0.78	0.57
Weighted median error	3.01	2.70	2.43
Error centre of gravity	0.12	0.095	0.069
Average percentage error	4.68	3.50	2.67
Median percentage error	12.94	11.09	9.36
Predictive range[min, max]	[–2.27, 1.64]	[–2.22, 1.44]	[–2.16, 1.32]
Mean absolute error	0.64	0.63	0.57
Median absolute error	0.51	0.54	0.45
Mean absolute percentage error	63.87	64.90	60.10
Median absolute percentage error	19.14	17.29	18.32
Root mean squared error	0.80	0.79	0.75
Root median squared error	0.51	0.54	0.45
Correlation (Pearson)	0.81	0.79	0.81
Scaled inner product	86.75	89.60	90.28
Gravity	3,711.02	3,913.52	4,115.28

In terms of predictive accuracy, the most congruent global hedonic model seems to be the double-log model, which encompasses other global models by only seven (out of seventeen) different indicators, but these are the most important ones concerning predictive unbiasedness. The strength of association is very similar between different global alternatives. It should be noted that the global cubic expansion model encompasses the other global model choices by nine different criteria, but seven of these relate to the reliability aspect of forecastability, which for most practical valuation purposes is secondary to predictive validity. The global cubic expansion model also has the highest classical correlation coefficient and gravity in the post-sample predictive testing, albeit the improvement in these numbers is minimal over the other model candidates. The most precise local model, the local cubic expansion model, encompasses the optimal global alternative, the conventional double-log model, by fourteen (out of seventeen) different forecasting criteria. *Especially*, the measures relating to absolute or relative mean prediction errors, i.e. *predictive validity indicators are significantly lower in the optimal local model than in the most valid global alternative*.

In table 4 the estimated global translog model indicates that the interaction term between parcel size and distance variables is statistically the most significant factor in the determination of land prices, which is followed by first-

order main effects of parcel size and distance variables, respectively[16]. Parcel type per se and its interaction with parcel size are also statistically highly significant. Second order main effects of parcel size and distance variables are statistically insignificant as well as the interaction between distance and parcel type variables. In the global cubic expansion model, the statistically significant third-order terms appearing in the model structure are the cubic distance variable and parcel size times parcel type times distance interaction. Second-order effects consist of squared distance, distance times parcel size interaction, distance times parcel type interaction and parcel size times parcel type interaction. All the first-order (original) variables are also included. Both global translog and cubic expansions models are centered.

Table 4*. Explanatory variables: parcel size, L_2-distance to CBD of Helsinki and parcel type.*

Measure of predictive accuracy	**Global hedonic model**		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	0.18	0.20	0.25
Median prediction error	0.37	0.42	0.41
Weighted average error	1.37	1.65	1.97
Weighted median error	3.29	3.54	3.51
Error centre of gravity	0.17	0.20	0.24
Average percentage error	6.58	7.37	9.30
Median percentage error	14.53	15.05	15.17
Predictive range [min, max]	[–2.05, 1.54]	[–2.06, 1.48]	[–1.91, 1.53]
Mean absolute error	0.63	0.61	0.58
Median absolute error	0.52	0.58	0.49
Mean absolute percentage error	59.36	58.88	54.56
Median absolute percentage error	19.79	17.76	17.68
Root mean squared error	0.75	0.74	0.72
Root median squared error	0.52	0.58	0.49
Correlation (Pearson)	0.84	0.83	0.87
Scaled inner product	86.36	87.03	85.14
Gravity	3,925.58	4,042.57	4,085.73

[16] If translog or cubic expansion models are used with original variables (as in this research), there is always, more or less, the problem of multicollinearity, which heavily confounds the classical inference.

Table 5. *Explanatory variables: parcel size, northing, easting and parcel type.*

Measure of predictive accuracy	Local hedonic model		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	0.23	0.15	0.049
Median prediction error	0.36	0.34	0.23
Weighted average error	1.83	1.21	0.52
Weighted median error	3.36	2.61	1.67
Error centre of gravity	0.22	0.15	0.063
Average percentage error	8.64	5.64	1.79
Median percentage error	12.95	12.23	8.04
Predictive range [min, max]	[–2.25, 1.90]	[–1.98, 1.50]	[–2.87, 1.35]
Mean absolute error	0.69	0.59	0.59
Median absolute error	0.59	0.54	0.35
Mean absolute percentage error	63.97	57.60	57.31
Median absolute percentage error	20.91	15.48	17.91
Root mean squared error	0.87	0.71	0.83
Root median squared error	0.59	0.54	0.35
Correlation (Pearson)	0.79	0.85	0.77
Scaled inner product	82.46	88.18	92.04
Gravity	3,261.19	4,239.23	3,769.30

Table 5 documents the forecasting results from the localised versions of double-log, translog and cubic expansion models, when L_2-distance to CBD of Helsinki variable is replaced by the northing and easting variables of each land parcel in the spirit of (Clapp, 2003 & 2004). The estimated models indicate a strong nonlinear relation between land prices and the associated attribute variables, albeit not as strong as in the case of table 3 as the choice between the local translog and local cubic expansion models is a bit more difficult here. However, the local cubic approximation appears to encompass the quadratic alternative in most measurable respects (12 out of 17 times), so eventually it can be chosen as the most adequate description of the data-generating process in table 5. The preference between the estimated local cubic expansion models of the table 3 and 5 is subjective; in mean prediction error there exists only minor discrepancies and the choice over the other forecasting criteria is difficult and subjective. I would conclude with the cubic expansion model of the table 3, as it is more parsimonious and the different predictive criteria behave in a more coherent manner in that model structure. Table 6 reports the global modelling analogues to the table 5.

The estimated global translog model of table 6 improves significantly the prediction results over the conventional global double-log model, which is in slight conflict over the results of table 4. The parametric inference is now plagued with very high multicollinearity, which makes it difficult to choose the relevant set of explanatory variables in a classical sense. The final estimated translog model, which is centered, consisted of first-order main effects of northing, easting and parcel type variables without the first-order influence of

parcel size variable. The estimated second-order terms included squared main effects of northing and easting in conjunction with the interactions of northing times parcel size, easting times parcel size and northing times parcel type. All these explanatory variables were statistically significant and most of them also statistically highly significant.

***Table 6**. Explanatory variables: parcel size, northing, easting and parcel type.*

Measure of predictive accuracy	**Global hedonic model**		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	0.33	0.22	0.29
Median prediction error	0.47	0.42	0.43
Weighted average error	2.63	1.87	2.33
Weighted median error	3.97	3.48	3.58
Error centre of gravity	0.32	0.23	0.28
Average percentage error	12.41	8.38	10.85
Median percentage error	16.71	15.20	15.37
Predictive range [min, max]	[–1.96, 1.94]	[–1.60, 1.68]	[–1.45, 1.56]
Mean absolute error	0.70	0.63	0.58
Median absolute error	0.60	0.55	0.55
Mean absolute percentage error	59.02	51.58	46.55
Median absolute percentage error	22.33	20.47	20.21
Root mean squared error	0.86	0.74	0.70
Root median squared error	0.61	0.55	0.55
Correlation (Pearson)	0.81	0.84	0.89
Scaled inner product	80.19	86.23	84.14
Gravity	3,193.01	4,002.94	4,164.97

The estimated global cubic model also appears in many respects to be highly attractive. Although the mean prediction error could not been improved over the estimated global translog model many variance based prediction accuracy measures are better. In particular, root mean squared error and mean absolute percentage errors are the lowest of all the model candidates in the tables 3–6. Thus the choice between global translog and cubic expansion models in table 6 is very difficult; I would prefer predictive validity and choose the translog model. The final global cubic model, which is centered, consisted of first-order easting, northing and parcel type variables without the first-order influence of parcel size variable. The estimated second-order interactions included northing times parcel size, easting times parcel size and northing times parcel type with the second-order main effect of northing. The estimated third-order variables interactions consisted of parcel type times easting times northing, parcel type times easting times parcel size and parcel type times northing times parcel size with the third-order main effect of northing. All these variables were statistically highly significant. The local cubic model of table 5 is superior to its most congruent global analogue, the conventional translog model, encompassing this by twelve (out of seventeen) different criteria.

The results of tables 3–6 indicate, beyond any reasonable doubt, that the use of local models significantly enhances the predictive accuracy of hedonic models; in particular *predictive validity of hedonic models can be substantially improved by using local modelling scheme*. Furthermore, these results suggest that most of the improvement resulted from increased nonlinearity rather than variable interactions.

Table 7 represents estimation results of global error correction and two most data-congruent local error correction models. Global error correction is the conventional equilibrium model (Engle and Granger, 1987). Local error correction 1 denotes a model, in which the applied equilibrium correction term is a linear combination of parcel size, L_2-distance to CBD of Helsinki and parcel type variables. This error correction term is subsequently estimated in conjunction with the other attributes (in differenced values) using a local quadratic approximation to an unknown hedonic relation. This yielded the most plausible local approximation, when the error correction term entered the model structure as a linear combination. Local error correction 2 denotes a model, where the applied equilibrium correction term is a nonlinear combination of parcel size, L_2-distance to CBD of Helsinki and parcel type variables; in particular, the error correction term itself is constructed using local quadratic approximation. This estimated error correction term is used as an explanatory variable along with the other characteristics using locally linear approximation. This appears to generate the most adequate results, if the error correction term enters the model structure as a nonlinear combination.

Table 7. *Explanatory variables: parcel size, L_2-distance to CBD of Helsinki, parcel type and error correction.*

Measure of predictive accuracy	**Equilibrium correction model**		
	Global error correction	**Local error correction 1**	**Local error correction 2**
Average prediction error	–0.055	–0.13	0.081
Median prediction error	0.10	0.017	0.17
Weighted average error	0.027	0.12	–0.044
Weighted median error	–0.0021	0.00	0.017
Error centre of gravity	13.44	59.85	–22.18
Average percentage error	52.94	127.10	77.08
Median percentage error	73.67	13.38	132.08
Predictive range [min, max]	[–2.50, 1.25]	[–2.61, 1.10]	[–2.64, 1.92]
Mean absolute error	0.58	0.60	0.59
Median absolute error	0.45	0.45	0.45
Mean absolute percentage error	213.05	240.36	136.94
Median absolute percentage error	37.04	47.41	36.72
Root mean squared error	0.78	0.81	0.82
Root median squared error	0.45	0.45	0.45
Correlation (Pearson)	0.91	0.90	0.91
Scaled inner product	69.68	73.53	64.61
Gravity	46.23	49.93	39.99

In essence, estimating locally the conventional error correction model does not result in remarkable improvements in predictive accuracy. In fact, *most mean*

prediction error indicators are the lowest when using global equilibrium correction mechanism among different error correction processes; there is some enchainment in local error correction model 1 relating to median-based measures of central tendency of errors (and also some improvement in the general strength of association in this model structure). Furthermore, the mean absolute error and root mean squared errors are lowest in the global error correction model in table 7. Striking feature of all error correction models in table 7 is that the relative mean prediction errors and the error centre of gravity are notoriously large, albeit absolute and standard weighted mean prediction errors are small. As a result, *although the absolute mean prediction errors, and their weighted analogues, are marginal, their relative effects, when the unit of measurement is taken into account, are unacceptably large for any practical land valuation purpose*. Also all error centres of gravities are prohibitively large. Furthermore, scaled inner product, and gravity in particular, indicate that the genuine strength of association between predictions and actual outcomes is low; an important fact that is not revealed simply by looking at the usual correlation coefficient. The most disturbing aspect is, however, that the optimal error correction model, the global error correction model, passes the conventional specification tests, such as the RESET test and the CUSUM t-test, which measure functional form mis-specification and zero forecast innovation mean, respectively. Furthermore, the unit-root t-tests strongly suggest that the observed data is co-integrated, i.e. the equilibrium correction model would be a valid description for the data-generating mechanism[17]. This is in direct conflict with the unusual high relative mean prediction errors and low association between the model's predictions and observed land prices.

14 The Nurmijärvi Case

This subsection reports the main results of out-of-sample predictive testing for land prices in the municipality of Nurmijärvi. Square land price was selected as the proper response, since this seemed to yield superior results to those models using total sales price of a land parcel as the dependent variable. Unlike in the Espoo case, here the price index variable would lower the bias of model's predictions. Table 8 documents the forecasting results from the localised versions of double-log, translog and cubic expansion models.

[17] The unit roots have been tested in the global error correction model structure using the augmented Dickey-Fuller method that derives from the Dickey-Fuller test by adding lagged differences. The unit-root t-test does not in fact have a t-distribution, but the significance (here 5%) is based on the correct critical values (the critical values are based on response surfaces in MacKinnon (1991)). The unit-root test is made for the cointegrating vector of the error correction model.

Table 8. *Explanatory variables: parcel size, L_2-distance to the CBD of Helsinki, L_2-distance to the CBD of parish village of Nurmijärvi, price index and parcel type.*

Measure of predictive accuracy	Local hedonic model		
	Double-log	Translog	Cubic expansion
Average prediction error	–0.12	–0.25	–0.19
Median prediction error	–0.052	–0.09	–0.042
Weighted average error	–1.03	–2.07	–1.56
Weighted median error	–0.42	–0.77	–0.35
Error centre of gravity	–0.12	–0.24	–0.18
Average percentage error	7.72	16.32	12.69
Median percentage error	3.64	6.36	2.94
Predictive range [min, max]	[–1.78, 0.90]	[–2.09, 0.53]	[–2.48, 0.55]
Mean absolute error	0.36	0.34	0.43
Median absolute error	0.23	0.13	0.29
Mean absolute percentage error	424.95	491.35	581.17
Median absolute percentage error	13.95	8.79	19.93
Root mean squared error	0.55	0.57	0.64
Root median squared error	0.23	0.13	0.29
Correlation (Pearson)	0.71	0.75	0.70
Scaled inner product	0.96	1.07	1.05
Gravity	2,477.55	2,651.96	2,318.95

Table 8 indicates that the most plausible local description of land prices is the local double-log model, which has significantly lower mean prediction errors than higher order local polynomial approximations. It seems that nonlinear relation between observed prices and the attributes is much weaker here than in the Espoo case. Besides the different mean prediction errors, only minor discrepancies exist among the rival local models in table 8. Table 9 reports the prediction results from the corresponding global alternatives. The most data-congruent global model, I believe, is the conventional double-log model (although cubic expansion model is highly desirable in some ways). The estimated global translog model consists of first-order effects of parcel type and L_2-distance to the CBD of Helsinki variables. The second-order terms comprise the squared main effects of parcel size, L_2-distance to the CBD of Helsinki, L_2-distance to the CBD of parish village of Nurmijärvi and price index variables; the interaction terms include inner products of parcel size with parcel type and the Helsinki CBD distance variables. This estimated model is centered; the main effects and interactions are statistically highly significant. The global cubic expansion model consists of first-order effects of parcel size and L_2-distance to the CBD of Helsinki variables; second and third-order main effects of those variables; second-order main effect of price index variable; second-order interaction of parcel type and L_2-distance to the CBD of parish village of Nurmijärvi variables, and third-order interaction of parcel type times parcel size times L_2-distance to the CBD of parish village of Nurmijärvi variables. The

estimated model is centered, whose all variables are statistically highly significant.

Table 9. *Explanatory variables: parcel size, L_2-distance to the CBD of Helsinki, L_2-distance to the CBD of parish village of Nurmijärvi, price index and parcel type.*

Measure of predictive accuracy	**Global hedonic model**		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	–0.066	–0.12	–0.091
Median prediction error	0.037	–0.077	0.035
Weighted average error	–0.66	–1.12	–0.87
Weighted median error	0.32	–0.69	0.31
Error centre of gravity	–0.07	–0.13	–0.10
Average percentage error	4.62	8.12	6.39
Median percentage error	2.70	5.54	2.51
Predictive range [min, max]	[–2.31, 1.05]	[–2.43, 1.01]	[–2.26, 0.94]
Mean absolute error	0.43	0.41	0.39
Median absolute error	0.25	0.24	0.25
Mean absolute percentage error	390.86	407.62	427.34
Median absolute percentage error	18.67	16.41	16.67
Root mean squared error	0.64	0.63	0.60
Root median squared error	0.25	0.24	0.25
Correlation (Pearson)	0.67	0.72	0.72
Scaled inner product	0.86	0.89	0.90
Gravity	1,868.12	2,005.37	2,083.46

If comparison is made between the optimal local and global models in tables 8 and 9, it turns out that the global version, the conventional double-log model, clearly encompasses the corresponding local analogue in terms of significantly lower mean prediction error. However, all measures concerning the strength of association between predictions and actual outcomes as well as most variance-based measures are better in the local case. Therefore the choice between these model candidates is difficult and, ultimately, somewhat subjective; I would favour the predictive validity as before and conclude with the conventional double-log model.

The estimated local double-log model appears to yield to the most congruent description among the local model candidates in table 10, where Euclidean distance measures are replaced by northing and easting variables of each land parcel. In particular, the use of higher order polynomial approximations would significantly increase mean prediction error. This gives more evidence that nonlinear association between observed land prices and its attributes is a less significant determinant than in the Espoo case.

Table 11 documents the corresponding results from their global analogues. The choice of the most plausible global model structure is again not evident; using overall assessment with the focus on the mean prediction error indicators

(notice also the reduction in median-based indicators) and the strength of association statistics, I would eventually choose the global translog model, which consists of first-order effects of parcel type, easting and northing variables. The second-order terms comprise the squared main effects of parcel size, northing and price index variables; the interaction includes the parcel size times parcel type term. The estimated model is uncentered; the main effects and interactions are statistically highly significant. The cubic expansion model consists of first-order effects of parcel type and northing; second-order main effect of northing and interaction of parcel size and parcel type variables; and third-order main effect of northing, parcel size and price index and third-order interaction of parcel size, easting and northing variables. This is a centered model, whose all attributes are statistically highly significant.

Table 10. *Explanatory variables: parcel size, northing, easting, parcel type and price index.*

Measure of predictive accuracy	**Local hedonic model**		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	–0.11	–0.21	–0.18
Median prediction error	0.021	–0.051	–0.11
Weighted average error	–0.95	–1.78	–1.49
Weighted median error	0.19	–0.44	–1.02
Error centre of gravity	–0.11	–0.20	–0.17
Average percentage error	7.08	14.13	11.75
Median percentage error	1.47	3.58	7.71
Predictive range [min, max]	[–1.79, 0.82]	[–2.10, 0.57]	[–2.31, 0.75]
Mean absolute error	0.35	0.34	0.40
Median absolute error	0.24	0.14	0.31
Mean absolute percentage error	427.11	491.47	511.73
Median absolute percentage error	11.91	11.31	17.17
Root mean squared error	0.55	0.56	0.57
Root median squared error	0.24	0.14	0.31
Correlation (Pearson)	0.71	0.75	0.75
Scaled inner product	0.95	1.05	1.04
Gravity	2,462.52	2,658.24	2,600.35

When comparing the global translog model and the optimal local model (double-log model), the former encompasses the latter nine out of seventeen different forecasting criteria. In particular, most mean prediction error indicators are lower in the global translog model case. Otherwise, there is relatively little difference between the predictive performance of these models. In essence, *local estimation adds no value to hedonic modelling in terms of out-of-sample forecasting power as compared to global modelling alternatives*, which agrees with results in tables 8 and 9 using Euclidean distances.

Table 11. *Explanatory variables: parcel size, northing, easting, parcel type and price index.*

Measure of predictive accuracy	Global hedonic model		
	Double-log	**Translog**	**Cubic expansion**
Average prediction error	0.015	–0.052	–0.064
Median prediction error	0.13	–0.021	–0.042
Weighted average error	0.06	–0.54	–0.65
Weighted median error	1.16	–0.19	–0.33
Error centre of gravity	0.01	–0.06	–0.07
Average percentage error	0.98	3.44	4.29
Median percentage error	8.78	1.44	2.91
Predictive range [min, max]	[–1.66, 0.99]	[–1.66, 1.02]	[–1.73, 1.03]
Mean absolute error	0.39	0.37	0.35
Median absolute error	0.26	0.25	0.21
Mean absolute percentage error	386.27	398.94	413.63
Median absolute percentage error	17.42	16.06	16.21
Root mean squared error	0.56	0.52	0.51
Root median squared error	0.26	0.25	0.21
Correlation (Pearson)	0.69	0.77	0.78
Scaled inner product	0.86	0.91	0.92
Gravity	2,201.95	2,531.50	2,601.83

Equilibrium correction mechanisms are also estimated for the Nurmijärvi case. Local error correction 1 in table 12 corresponds to local quadratic approximation, where the error correction term is formulated as a linear combination of the original attribute variables (parcel size, parcel type, L_2-distance to the CBD of Helsinki, L_2-distance to the CBD of parish village of Nurmijärvi and price index variables). Local error correction 2 in table 12 denotes a model formulation, where the error correction term itself is constructed using locally linear approximation and subsequently used as explanatory variable in a local quadratic model. This model structure seemed to yield the most adequate description for local equilibrium correction mechanisms.

The results of table 12 are similar to results in table 7: All estimated error correction models possess very low absolute mean prediction errors and large relative mean prediction errors, albeit local error correction model 2 posseses significantly smaller average percentage error than any other error correction model in tables 7 and 12. Scaled inner product and gravity effectively reveal that the genuine strength of association between predictions and actual outcomes is low (especially the gravity); an aspect that is not captured by the usual correlation coefficient. The only major departure here is that the estimation of an error correction model locally can improve the prediction results over the global error correction model in most measurable respects. Otherwise the same problems are encountered here as in the table 7.

Table 12. *Explanatory variables: parcel size, L_2-distance to CBD of Helsinki, L_2-distance to CBD of parish village of Nurmijärvi, parcel type and error correction.*

Measure of predictive accuracy	Equilibrium correction model		
	Global error correction	Local error correction 1	Local error correction 2
Average prediction error	–0.11	–0.11	–0.011
Median prediction error	–0.030	–0.12	0.050
Weighted average error	0.010	–0.17	–0.010
Weighted median error	0.00	–0.02	0.00
Error centre of gravity	117.92	–2089.68	–140.76
Average percentage error	472.22	238.21	23.12
Median percentage error	61.26	∞	∞
Predictive range [min, max]	[–2.22, 1.00]	[–4.00, 2.23]	[–1.74, 2.03]
Mean absolute error	0.43	0.68	0.40
Median absolute error	0.28	0.49	0.24
Mean absolute percentage error	∞	∞	∞
Median absolute percentage error	∞	∞	∞
Root mean squared error	0.65	1.04	0.59
Root median squared error	0.28	0.49	0.24
Correlation (Pearson)	0.84	0.47	0.85
Scaled inner product	69.47	34.30	66.89
Gravity	20.83	10.80	24.35

15 Conclusions

This paper has investigated in depth the predictive accuracy of several different static and dynamic local hedonic models of land prices with adaptive bandwidth selection learning. Static models whose out-of-sample predictions were analysed, consisted of local double-log, local translog and local cubic expansion models. Dynamic local hedonic models were constructed and their ex-sample forecasting ability analysed using equilibrium correction mechanisms under two different scenarios: error correction term entered model structure as a linear combination of the original attributes, or as a flexible nonlinear transformation of those characteristics. All local models' post-sample predictions were compared to their proper corresponding global analogues with two data sets differing both in terms of quality and quantity of observations.

It turned out that predictive performance of different hedonic models varied significantly across the data sets. In the municipality of Espoo, whose data set can be regarded qualitatively superior (it is manually pre-checked) to the Nurmijärvi case, the use of static local models significantly enhanced the forecasting accuracy as compared to their global alternatives. Specifically, the predictive validity of hedonic models may substantially be improved by the local modelling scheme. The results also suggested, beyond any reasonable doubt, that most of the improvement resulted from increased nonlinearity rather than variable interactions. The most data-congruent local hedonic model, the local

cubic expansion model, encompassed the most adequate global alternative, the conventional global double-log model, in most measurable respects.

The predictive testing provided quite contradictory results in the Nurmijärvi case; this data set contained twice as many observations as the former, yet was more susceptible to any errors due to lack of specific manual pre-checking. In essence, local estimation added no value to hedonic modelling in terms of out-of-sample forecasting power. Instead, the most data-congruent specification was either global double-log or translog model, which depended on the measurement of location effect. It appears that the quality, not the quantity, of the data set is the critical factor, which determines the utility of flexible estimation techniques in forecasting the land prices, even if the results of in-sample estimation are robustified by some procedure. There also seems to be a genuine association between land prices and housing prices (quarterly price index of single-family houses) in the Nurmijärvi case: the use of housing price index significantly improved the predictive unbiasedness of any chosen hedonic model. In contrast, for the Espoo case, the use of this index variable would seriously undermine the predictive validity of all model combinations used in the research.

The innovative part of the paper, an attempt to unify the concepts of nonlinearity and nonstationarity simultaneously into hedonic model specification by using local regression and error correction mechanisms, utterly failed. This predictive failure is, however, an appearance of potentially a far more profound problem relating to the proper use of equilibrium correction models, and thus to co-integrated processes, at least if they are to be used in the determination of land prices. Over both data sets, the absolute mean prediction errors are marginal, yet their relative effects, when the unit of measurement is taken into account, are unacceptably large (ranging from 23% to infinity!) for any practical land valuation purpose. Furthermore, scaled inner product and gravity indicate that the genuine strength of association between predictions and actual outcomes is low; an important fact that is not revealed simply by looking at the usual correlation coefficient. The most disturbing aspect is, however, that the optimal error correction models clearly pass the conventional specification tests, such as the RESET test and the CUSUM t-test, which measure functional form mis-specification and zero forecast innovation mean, respectively. Furthermore, the unit-root t-tests strongly suggest that the observed data is co-integrated, i.e. equilibrium correction model would be a valid description for the data-generating mechanism. This is in direct conflict with the unusual high relative mean prediction errors and low association between model's predictions and observed land prices.

Finally, if predictive testing is performed on land markets, the choice of a relevant set of forecasting criteria tends to be critical. It is a sound practice to document several statistics that measure somewhat different aspects of post-sample predictive performance. It is vital to remember to calculate the relative mean prediction errors and, I believe, weighted average-based measures should be preferred over median-based analogues in the evaluation of predictive validity of hedonic models. The strength of association between model's predictions and actual out-of-sample prices should be considered and measures

such as scaled inner product, gravity, etc. ought to be preferred to the simple correlation coefficient. In analysing the reliability aspect, the usual root mean squared error and mean absolute prediction error indicators are easily misleading, and the median-based alternatives would be a better solution; or simply, to move to using mean absolute error indicators.

References

Allen, D.M. (1974). "The Relationship Between Variable Selection and Data Augmentation and a Method of Prediction", *Technometrics* 16, 125–127.

Anglin, Paul M. and Gencay, Ramazan (1996). "Semiparametric Estimation of a Hedonic Price Function", *Journal of Applied Econometrics*, Vol. 11, 633–648.

Breiman, Leo. (2001). "Statistical Modeling: The Two Cultures", *Statistical Science*, Vol. 16, No. 3, 199–231.

Case, Bradford, Clapp, John, Dubin, Robin and Rodriquez, Mauricio (2004). "Modelling Spatial and Temporal House Price Patterns: A Comparison of Four Models", Journal of Real Estate Finance and Economics, Vol. 29, No. 2, 167–191.

Clapp, John M. (2003). "A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation", *Journal of Real Estate Finance and Economics*, Vol. 27, No. 3, 303–320.

Clapp, John M. (2004). "A Semiparametric Method of Estimating Local House Price Indices", *Real Estate Economics*, Vol. 32, No. 1, 127–160.

Clapp, John M., Kim, Hyon-Jung and Gelfand, Alan E. (2002). "Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing", *Real Estate Economics*, Vol. 30, No. 4, 505–532.

Cleveland, William S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association* 74, 829–836.

Cleveland, William S. and Devlin, Susan J. (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting", *Journal of the American Statistical Association*, Vol. 83, No. 403, 596–610.

Cleveland, William S. and Grosse, E. (1991). "Computational Methods for Local Regression", *Statistics and Computing* 1, 47–62.

Cleveland, W.S., Loader, C.R. (1996). "Smoothing by Local Regression: Principles and Methods", *Statistical Theory and Computational Aspects of Smoothing*, (eds.) Härdle, W. and Schimek, M.G., Physica-Verlag.

Craven, P. and Wahba, G. (1979). "Smoothing Noisy Data with Spline Functions", *Numerische Mathematik* 31, 377–403.

Day, R.H. (1992). "Complex Economic Dynamics: Obvious in History, Generic in Theory, Elusive in Data", *Journal of Applied Econometrics*, Vol. 7, 9–23.

Engle, R.F. and Granger, C.W. (1987). "Co-integration and Error Correction: Representation, Estimation and Testing", *Econometrica* 55, 251–276.

Fan, Jianqing and Gijbels, Irene (1995). "Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation", *Journal of Royal Statistical Society*, B57, No. 2, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Monographs on Statistics and Applied Probability 66, Chapman & Hall.

Friendman, J.H., Bentley, J.L and, Finkel, R.A. (1977). "An Algorithm for Finding Best Matches in Logarithmic Expected Time", *ACM Transactions on Mathematical Software* 3, 209–226.

Gencay, Ramazan and Yang, Xian (1996). "A Forecast Comparison of Residential Housing Prices by Parametric versus Semiparametric Conditional Mean Estimators", *Economic Letters*, 52, 129–135.

Hannonen, Marko (2005). "On the Recursive Estimation of Hedonic Prices of Land", Nordic Journal of Surveying and Real Estate Research, 2:2, 30–55.

Hastie, Trevor and Loader, Clive (1993). "Local Regression: Automatic Kernel Carpentry", Statistical Science, Vol. 8, No. 2, 120–143.

Heinonen, Tuomo (1993). The Price and Price Development of Single-Family Housing Sites in the Finnish Cities in Years 1985–1991 (in Finnish), Helsinki.

Hendry, David F. (1997). "The Econometrics of Macroeconomic Forecasting", *The Economic Journal*, 107 (September), 1330–1357.

Hiltunen, Ari (2003). *On the Price Formation of Single-Family House Properties: An Analysis of Comparable Sales Prices for Single-Family Houses in Some Regional Areas in Finland* (in Finnish), Doctoral Thesis, Helsinki University of Technology, Espoo.

LaCour-Little, Michael, Marschoun, Michael and Maxam, Clark L. (2002). "Improving Parametric Mortgage Prepayment Models with Non-parametric Kernel Regression", *Journal of Real Estate Research*, Vol. 24, No. 3, 299–327.

Leväinen, Kari I. (1991). A Calculation Method for a Site Price Index, Doctoral Thesis, Helsinki University of Technology, Espoo.

Loader, Clive (1999). *Local Regression and Likelihood*. Springer Series in Statistics and Computing, Springer-Verlag.

Loader, Catherine (2004). "Smoothing: Local Regression Techniques", *Handbook of Computational Statistics*, (eds.) Gentle, J., Härdle, W. and Mori, Y., Springer-Verlag.

MacKinnon, J. (1991). "Critical Values for Cointegration Tests", *Long-Run Economic Relationships*, (eds.) Engle, R.F. and Granger, C. W. J., 267–276.

Makridakis, Spyros and Hibon, Michele (2000). "The M3-Competition: Results, Conclusions and Implications", *International Journal of Forecasting*, 16, 451–476.

Maxam, Clark L. and LaCour-Little, Michael (2001). "Applied Nonparametric Regression Techniques: Estimating Prepayments on Fixed-Rate Mortgage-Backed Securities", *Journal of Real Estate Finance and Economics*, Vol. 23, No. 2, 139–160.

McDonald, J.F. and Bowman, H.W. (1979). "Land Value Functions: A Re-evaluation", *Journal of Urban Economics*, 6, 25–41.

McMillen, Daniel P. (1996). "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach", *Journal of Urban Economics*, 40, 100–124.

McMillen, Daniel P. (2001). "Nonparametric Employment Subcenter Identification", *Journal of Urban Economics*, 50, 448–473.

McMillen, Daniel P. and Thorsnes, Paul (2003). "The Aroma of Tahoma: Time-Varying Average Derivates and the Effect of a Superfund Site on House Prices", *Journal of Business and Economics Statistics*, Vol. 21., No. 2, 237–246.

Meese, Richard and Wallace, Nancy (1991). "Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residental Housing Price Indices", *AREUEA Journal*, Vol. 19, No. 3, 1991.

Meese, Richard and Wallace, Nancy (1997). "The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression and Hybrid Approaches", Journal of the Real Estate Finance and Economics, 14, 51–73.

Pace, R. Kelley (1993). "Nonparametric Methods With Applications to Hedonic Models", *Journal of Real Estate Finance and Economics,* 7, 195–204.

Pace, R. Kelley (1995). "Parametric, Semiparametric, and Nonparametric Estimation of Characteristic Values Within Mass Assessment and Hedonic Pricing Models", *Journal of Real Estate Finance and Economics*, 11, 195–217.

Robinson, P.M. (1988). "Root-N-consistent Semiparametric Regression", *Econometrica*, Vol. 56, No. 4, 931–954.

Ruppert, D. and Wand, M.P (1994). "Multivariate Locally Weighted Least Squares Regression", *The Annals of Statistics*, Vol. 22, No. 3, 1346–1370.

Schulz, M.A. Rainer (2003). *Valuation of Properties and Economic Models of Real Estate Markets*. Dissertation, Berlin.

Stock, J.H. (1989). "Nonparametric Policy Analysis", *Journal of the American Statistical Association*, Vol. 84, No. 406, 567–575.

Thorsnes, Paul and McMillen, Daniel P. (1998). "Land Value and Parcel Size: A Semiparametric Analysis", *Journal of Real Estate Economics and Finance*, Vol. 17, No. 3, 233–244.

Wallace, Nancy E. (1996). "Hedonic-Based Price Indexes for Housing: Theory, Estimation and Index Construction", *Economic Review*, Vol. 3, 34–48.

Appendix

***Table 1A**. Some in-sample goodness-of-fit statistics of the hedonic models of study.*

Hedonic Model Table No.	Column No.	**Standard Error of Regression**	**Coefficient of Determination**	**Residual Sum of Squares**	**Fitted Degrees of Freedom**
3	1	0.43	0.82	64.05	4.03
3	2	0.40	0.84	55.71	9.10
3	3	0.39	0.85	52.22	16.16
4	1	0.66	0.57	156.87	4
4	2	0.63	0.60	142.98	5
4	3	0.61	0.64	129.46	10
5	1	0.44	0.81	67.01	5.03
5	2	0.42	0.83	61.62	14.12
5	3	0.37	0.88	44.24	30.02
6	1	0.65	0.59	150.90	5
6	2	0.62	0.62	135.36	8
6	3	0.59	0.69	120.70	11
7	1	0.66	0.76	152.22	5
7	2	0.47	0.88	75.92	14.13
7	3	0.43	0.90	63.54	5.07
8	1	0.37	0.76	100.22	6.02
8	2	0.38	0.76	100.50	20.06
8	3	0.33	0.83	71.40	50.03
9	1	0.57	0.42	242.92	6
9	2	0.55	0.46	227.46	8
9	3	0.54	0.48	218.15	9
10	1	0.38	0.75	103.50	6.03
10	2	0.37	0.77	98.40	20.05
10	3	0.33	0.83	71.10	50.03
11	1	0.58	0.41	249.04	6
11	2	0.56	0.45	231.02	8
11	3	0.56	0.45	230.30	8
12	1	0.58	0.65	250.32	6
12	2	0.41	0.83	118.74	20.16
12	3	0.44	0.81	137.38	6.07