Empirical Economics (2005) 30:93–114
DOI 10.1007/s00181-004-0224-6

EMPIRICAL ECONOMICS

Estimation of hedonic price functions via additive nonparametric regression

Carlos Martins-Filho[1], **Okmyung Bin**[2]

[1]Department of Economics, Oregon State University, Ballard Hall 303, Corvallis, OR 97331-3612, USA (e-mail: carlos.martins@orst.edu)
[2]Okmyung Bin, Department of Economics, East Carolina University, A-435 Brewster, Greenville, NC 27858-4353, USA (e-mail: bino@mail.ecu.edu)

First version received: October 2002/Final version received: October 2003

Abstract. We model a hedonic price function for housing as an additive nonparametric regression. Estimation is done via a backfitting procedure in combination with a local polynomial estimator. It avoids the pitfalls of an unrestricted nonparametric estimator, such as slow convergence rates and the curse of dimensionality. Bandwidths are chosen using a novel plug in method that minimizes the asymptotic mean average squared error (AMASE) of the regression. We compare our results to alternative parametric models and find evidence of the superiority of our nonparametric model. From an empirical perspective our study is interesting in that the effects on housing prices of a series of environmental characteristics are modeled in the regression. We find these characteristics to be important in the determination of housing prices.

Key words: Additive nonparametric regression, local polynomial estimation, hedonic price models, housing markets

JEL classification: C14, R21

1. Introduction

Hedonic price models have been used extensively in applied economics since the seminal work of Rosen (1974).[1] A frequent concern in this literature is the adequacy of commonly assumed parametric specifications as hedonic price functions. This specification problem arises quite naturally from the inability

We thank B. Baltagi and two anonymous referees for their comments. The authors retain responsibility for any remaining errors.

[1] Recent examples include among many Dreyfus and Viscusi (1995), Walls (1996) and Hill et al. (1997).

of economic theory to provide guidance on how characteristics of similar products relate functionally to their market prices. Recognizing the potentially serious consequences of functional misspecification, researchers have attempted to estimate hedonic price models by specifying more flexible regression models. Most of these attempts have concentrated on parametric specifications ranging from simple data transformations, including the model introduced by Box and Cox (1964) and its variants and approximations based on second order Taylor-expansions to flexible non-linear models such as those introduced by Wooldridge (1992). A considerably smaller set of authors have proposed semi and fully (unrestricted) nonparametric specifications for hedonic price functions.

Nonparametric regression models are very flexible in that regressions are allowed to belong to a vastly broader class of functions than that in parametric models. However, their use in applied economics has not been as prevalent as one would expect, or comparable to their use in other disciplines, such as Biostatistics.[2] In fact, nonparametric estimation of hedonic price functions has appeared in just a few articles and has concentrated almost exclusively in housing markets. Among these papers are the important contributions of Hartog and Bierens (1989), Stock (1991), Pace (1993,1995,1998), Anglin and Gencay (1996), Gencay and Yang (1996), Iwata (2000) and Clapp et al. (2002). Closer inspection of this literature, however, reveals the possibility of a number of methodological improvements that can add to both the ease of obtaining, and interpreting hedonic price nonparametric functional estimates. These improvements fall into three broad categories: (a) specification of the regression class; (b) choice of the smoother underlying the estimation; and (c) choice of the bandwidths. In this paper we address each of these categories.

Most of the applied nonparametric research specify a regression class that requires the estimation of multivariate smoothers. There are a number of practical, as well as theoretical problems that emerge when estimating multivariate smoothers. First, there is the curse of dimensionality identified by Friedman and Stuetzle (1981). The problem can be especially acute in data sets used by economists in hedonic price function estimation. Although these data sets can have a large number of observations, there is normally a vast list of product attributes or characteristics which contribute to slow convergence rates of the estimators and diminished confidence on inference. Second, when defining neighborhoods in two or more dimensions for local averaging, an universal characteristic of multivariate nonparametric estimation, there is the need to assume some type of metric that is hard to justify when the variables are measured in different units or are highly correlated. Third, from a practical perspective, multivariate smoothers are extremely expensive to compute and even with the use of sophisticated graphical analysis four or higher dimensional smooths are virtually impossible to represent or interpret. Since one of the objectives of hedonic price modeling is to easily interpret and isolate the contributions of a given attribute to market price variability, holding all other product characteristics fixed, we find the use of a fully unrestricted nonparametric regression undesirable. We believe much better

[2] Yatchew (1998) provides a list of potential reasons for this relative scarcity of applied nonparametric economic modeling.

results in hedonic price modeling can be obtained by estimating an additive nonparametric regression model (ANRM), as in Hastie and Tibshirani (1986). Estimation of these models involves only univariate smoothing, but the models allow for multiple regressors, and due to their additive nature lend themselves to easy interpretation and analysis.

Regarding the choice of estimator, we depart from the standard literature in applied economics by using local polynomial estimators rather than the popular Nadaraya-Watson (NW) estimator. Recent work by Fan (1992), Fan et al. (1993) and Ruppert and Wand (1994) has shown that local polynomial estimators possess a number of desirable theoretical and practical properties relative to other smoothing methods, including the NW estimator. Lastly, we believe we make significant improvements over previous applied work by estimating bandwidths via a plug in method rather than the popular but much criticized cross validation (Park and Marron 1990; Simonoff 1996). Although plug in methods are relatively new in the statistical literature they have consistently outperformed cross validation as a data driven bandwidth selection method. They converge at faster rates, are less expensive to compute and overcome the problem of undersmoothing that is characteristic of the cross validation method (Sheather and Jones 1991; Ruppert et al. 1995).

In this paper we use data from the Portland-Oregon housing market to estimate a hedonic price function using all of the improvements described above. From an empirical perspective our model is of independent interest as a series of environmental and locational housing characteristics such as property elevation, distance to wetlands, parks and lakes are included in the hedonic price function. We also compare our nonparametric regression results to those that are obtained from two alternative parametric specifications. The rest of this paper is organized as follows: in Sect. 2 we give some generalities on the additive nonparametric model and describe the estimation strategy and its properties; Sect. 3 describes and summarizes the data that we have used in this study; Sect. 4 specifies the empirical model, gives details on the computational aspects of the estimation algorithm and introduces a viable parametric alternative; Sect. 5 presents and analyzes the results and provides an out-of-sample forecast exercise. Section 6 is the conclusion.

2. Model specification and estimation

2.1. Generalities

We model hedonic price functions as a multivariate regression model[3] given by,

$$E(Y|X_1 = x_1, \ldots, X_D = x_D) = \alpha + \sum_{d=1}^{D} m_d(x_d). \tag{1}$$

[3] We note that (1) can easily accomodate transformations of regressand and regressors. Thus, if there is a priori knowledge that allows for the specification $E(T(Y)|X_1 = x_1, \ldots, X_D = x_D) = \alpha + \sum_{d=1}^{D} m_d(\tau_d(x_d))$ for known $T(\cdot), \tau_d(\cdot)$ and $d = 1, \ldots, D$ estimation can proceed as described below using the transformed variables. See Hastie and Tibshirani (1990) and Opsomer and Ruppert (1998).

We assume that n independent observations $\{(y_t, x_{t1}, \ldots, x_{tD})'\}_{t=1}^{n}$ are taken on the random vector $(Y, X_1, \ldots, X_D)$ and that $V(Y|X_1 = x_1, \ldots, X_D = x_D) = \sigma^2$, an unknown parameter. The $m_d(\cdot)$ are real valued measurable functions with $E(m_d(\cdot)) = 0$, $E(m_d^2(\cdot)) < \infty$. Under these assumptions $E(Y) = \alpha$ and the optimal predictor for Y given $X_1, \ldots, X_D$ can be characterized by

$$m_d(x_d) = E\left(Y - \alpha - \sum_{\delta=1, \delta\neq d}^{D} m_\delta(\cdot)|X_d = x_d\right) \tag{2}$$

for $d = 1, \ldots, D$ (Buja et al. 1989). We estimate $m_d(\cdot)$ using the backfitting estimator (B-estimator) proposed by Friedman and Stuetzle (1981). Let $a_n \equiv (a, \ldots, a)'$ be the $n \times 1$ vector with the constant a as its components, I_n be the identity matrix of size n, $D_n \equiv \left(I_n - n^{-1} 1_n 1_n'\right)$, $y \equiv (y_1, \ldots, y_n)'$, $x_d \equiv (x_{1d}, \ldots, x_{nd})'$, ${}_d(x_d) \equiv (\boldsymbol{m}_d(x_{1d}), \ldots, m_d(x_{nd}))'$, S_d be the $n \times n$ smoother matrix associated with regressor d and $S_d^* \equiv D_n S_d$. The B-estimator for $\boldsymbol{m}_d(x_d)$ is the solution for the following system of normal equations,

$$\begin{pmatrix} I_n & S_1^* & \ldots & S_1^* \\ S_2^* & I_n & \ldots & S_2^* \\ \vdots & \vdots & \ddots & \vdots \\ S_D^* & S_D^* & \ldots & I_n \end{pmatrix} \begin{pmatrix} \boldsymbol{m}_1(x_1) \\ \boldsymbol{m}_2(x_2) \\ \vdots \\ \boldsymbol{m}_D(x_D) \end{pmatrix} = \begin{pmatrix} S_1^* \\ S_2^* \\ \vdots \\ S_D^* \end{pmatrix} y, \tag{3}$$

which we denote by $\left(\boldsymbol{m}_1^b(x_1), \ldots, m_D^b(x_D)\right)'$. A convenient procedure to obtain a solution for (3) involves setting initial values $\alpha^0 \equiv n^{-1} 1_n' y$ and $\boldsymbol{m}_d^b(x_d)^0 \equiv 0_n$ for all d. We then define the v^{th} iteration ($v = 1, 2, \ldots$) estimator $\boldsymbol{m}_d^b(x_d)^v$ as the smooth that results from a suitably chosen nonparametric univariate regression estimator, where the observed regressands are given by

$$y - 1_n \alpha^0 - \sum_{\delta=1}^{d-1} \boldsymbol{m}_\delta^b(x_\delta)^v - \sum_{\delta=d+1}^{D} \boldsymbol{m}_\delta^b(x_\delta)^{v-1}$$

and the regressors are x_d, for $d = 1, \ldots, D$. Iterations continue until $||y - 1_n\alpha^0 - \sum_{\delta=1}^{D} \boldsymbol{m}_\delta^b(x_\delta)^v||_2^2 - ||y - 1_n\alpha^0 - \sum_{\delta=1}^{D} \boldsymbol{m}_\delta^b(x_\delta)^{v+1}||_2^2 = 0$ or is smaller than a prespecified level of tolerance, where if $\theta \in \Re^n$, $||\theta||_2 = \left(\sum_{i=1}^{n} \theta_i^2\right)^{1/2}$. We construct S_d^* using a local polynomial estimator of order $p = 1$ *or* 3, as needed in our estimation algorithm, based on data driven bandwidths h_{dn}.[4] One of the practical conveniences of the local polynomial estimators is that provided that p is large enough and that $m_d(\cdot)$ is sufficiently smooth, the q^{th} derivative of $m_d(\cdot)$, denoted by $m_d^{(q)}(x)$ can be easily estimated. We denote such estimator by $\boldsymbol{m}_d^{b(q)}(x)$ and define $\boldsymbol{m}_d^{b(q)}(x_d) = (\boldsymbol{m}_d^{b(q)}(x_{1d}), \ldots, \boldsymbol{m}_d^{b(q)}(x_{nd}))'$.

2.2. Data driven selection of h_{dn}

One of the most important steps in estimating any nonparametric regression model is the choice of h_{dn}. In essence, after a nonparametric estimation

[4] For details see Opsomer and Ruppert (1997).

procedure is chosen the selection of the bandwidths is tantamount to the selection of the estimated regression. Here we follow the plug-in procedure suggested by Opsomer and Ruppert(1998). Specifically, we choose $h_n \equiv (h_{1n}, \ldots, h_{Dn})' \in \Re^D$ such that the conditional mean average squared error $MASE(h_n) = \frac{1}{n}\sum_{t=1}^{n} E\big(((\alpha^0 - \alpha) + \sum_{d=1}^{D}(m_d^b(x_{td}) - m_d(x_{td})))^2 | x_1, \ldots, x_D\big)$ is minimized. Given a local polynomial estimator, and under the assumption that the regressors $(X_1, \ldots, X_D)$ are independent, it can be shown that for $p = 1$,

$$MASE(h_n) = \frac{\mu_2(K)^2}{2}\sum_{d=1}^{D} h_{dn}^4 \theta_{dd}(2,2) + \sigma^2 \sum_{d=1}^{D}\frac{R(K)}{nh_{dn}} n^{-1}\sum_{t=1}^{n}\frac{1}{f_{X_d}(x_{td})} + O_p\left(\sum_{d=1}^{D}\left(\frac{1}{nh_{dn}}\right)\right) + o_p\left(\sum_{d=1}^{D}\left(\frac{1}{nh_{dn}} + h_{dn}^4\right)\right), \tag{4}$$

where $\mu_2(K) = \int x^2 K(x)dx$, $R(K) = \int K(x)^2 dx$, f_{X_d} is the marginal density of X_{td} and $\theta_{dd}(2,2) = \frac{1}{n}||\boldsymbol{m}_d^{(2)}(x_d) - E(\boldsymbol{m}_d^{(2)}(x_d))||_2^2$. Ignoring the terms O_p and o_p in (4), as in Opsomer and Ruppert (1998), we have that the vector $\hat{h}_n$ that minimizes the conditional MASE has d^{th} component given by,

$$\hat{h}_{dn} = \left(\sigma^2 \frac{R(K) n^{-1}\sum_{t=1}^{n}\frac{1}{f_{X_d}(x_{td})}}{n\mu_2(K)^2\theta_{dd}(2,2)}\right)^{\frac{1}{5}}. \tag{5}$$

The plug in strategy is to obtain $\hat{h}_{dn}$ by directly estimating σ^2, $\theta_{dd}(2,2)$ and f_{X_d}. The term $n^{-1}\sum_{t=1}^{n}\frac{1}{f_{X_d}(x_{td})}$ is estimated by $\rho_d = max(x_d) - min(x_d)$. The estimation of $\theta_{dd}(2,2)$ requires the estimation of second derivatives of m_d which in turn requires the selection of an auxiliary bandwidth vector $g_n = (g_{1n}, \ldots, g_{Dn})'$ that minimizes the conditional (asymptotic) mean squared error of the estimator $\hat{\theta}_{dd}(2,2) = \frac{1}{n}||D_n \boldsymbol{m}_d^{b(2)}(x_d)||_2^2$. When the regressors are independent, the vector $\hat{g}$ has d^{th} component given by

$$\hat{g}_{dn} = \left(C_a \frac{4! R(K_{2,3})\sigma^2\rho_d}{2n|\theta_{dd}(2,4)|\mu_4(K_{2,3})}\right)^{1/7} \tag{6}$$

for $d = 1, 2, \ldots, D$, $C_a = \begin{cases} 1 & \text{if } \theta_{dd}(2,4) < 0 \\ 2.5 & \text{if } \theta_{dd}(2,4) > 0 \end{cases}$, where $K_{2,3}$ is obtained from $K_{r,p}(u) = \left(\frac{r!det(M_{r,p})}{det(N_p)}\right)K(u)$ with N_p a $(p+1)\times(p+1)$ matrix having (i,j) entry given by $\int u^{i+j-2}K(u)du$, $M_{r,p}$ is the same as N_p except that the $r+1$ column is substituted by $(1, u, u^2, \ldots, u^p)'$, and $det(A)$ is the determinant of a square matrix A.

The other component of (5) that needs to be estimated is σ^2. We define, $\hat{\sigma}^2(\kappa_n) = n^{-1}||y - 1_n\alpha^0 - \sum_{d=1}^{D}\boldsymbol{m}_d^b(x_d)||_2^2$ as a generic estimator for σ^2 based on $\kappa_n \in \Re^D$, the bandwidth used to obtain $\boldsymbol{m}_d^b(x_d)$ for $d = 1, \ldots, D$. The κ_n that minimizes the conditional (asymptotic) mean squared error of $\hat{\sigma}^2$ has d^{th} component given by,

$$\hat{\kappa}_{dn} = \left(C_b \frac{4|R(K_{0,1}) - 2K_{0,1}(0)|\sigma^2\rho_d}{n\theta_{dd}(2,2)\mu_2(K_{0,1})^2}\right)^{1/5}, \tag{7}$$

$C_b = \begin{cases} 1 & \text{if } R(K_{0,1}) - 2K_{0,1}(0) < 0 \\ 0.25 & \text{if } R(K_{0,1}) - 2K_{0,1}(0) > 0 \end{cases}$. The expressions for the optimal bandwidth vectors in (5), (6) and (7) depend on unknown functionals, $\theta_{dd}(2,4)$, $\theta_{dd}(2,2)$ and σ^2, which are estimated as follows. First, we obtain pilot estimates for σ^2 and $\theta_{dd}(2,4)$ for $d = 1, \ldots, D$ based on a suitably defined parametric model. These initial estimates are used to compute $\hat{g}_{dn}$ according to Eq. (6). We then use $\hat{g}_{dn}$ as bandwidths to fit an additive model using a polynomial smoother of degree $p = 3$. We then use the fitted additive model to estimate $\hat{\theta}_{dd}(2,2)$. This estimate together with the initial estimate of σ^2 are then used to obtain $\hat{\kappa}_{dn}$ according to (7). $\hat{\kappa}_{dn}$ are then used to fit an additive model using a local polynomial of order $p = 1$ and to compute an updated estimate of σ^2. Finally, the updated estimate of σ^2 together with the estimate of $\hat{\theta}_{dd}(2,2)$ are used to obtain estimates $\hat{h}_{dn}$ according to (5). The $\hat{h}_{dn}$ are then used to obtain a final fit for the ANRM and σ^2.

3. The data

The housing market data that we use come from a portion of Multnomah County, Oregon-USA that lies within Portland's urban growth boundary.[5] The regressand in our empirical model is the actual recorded sales price of a dwelling. We use 1000 randomly selected recorded sales from data originally collected by MetroScan which occured between June of 1992 and May of 1994 in the study area. All sales prices were adjusted to May 1994 levels using a Multnomah County residential housing market price index. Figure 1 provides a histogram of housing prices with a bandwidth chosen by the method proposed by Sheather and Jones (1991).

Direct inspection of the histogram reveals that the vast majority of the observations falls within the (50000,250000) interval. There is however a great

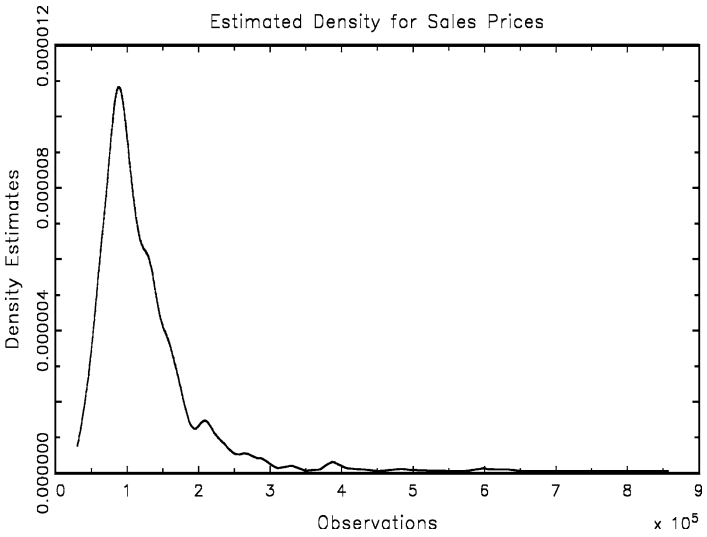


Fig. 1. Estimated Density for Sales Prices

[5] All data used is available from the first author.

Table 1. Sample statistics

Variable	Average	Standard deviation	Maximum	Minimum
Y Sale price	126783.55	83676.47	857053.90	30308.63
X_1 Number of bathrooms	1.43	0.63	6.00	1.00
X_2 Number of bedrooms	2.86	0.93	14.00	0.00
X_3 Dwelling area	134.02	56.86	584.55	40.13
X_4 Land area	727.52	593.71	10117.14	100.33
X_5 Dwelling age in 1994	44.29	27.29	106.00	0.00
X_6 Distance to nearest lake	5.38	2.09	10.43	0.00
X_7 Distance to the nearest wetland	1.09	0.75	3.46	0.01
X_8 Distance to nearest improved park	0.39	0.27	1.68	0.00
X_9 Dwelling elevation	0.08	0.04	0.27	0.00
X_{10} Distance to the nearest industrial zone	1.14	0.93	4.65	0.00
X_{11} Distance to the nearest commercial zone	0.38	0.32	2.33	0.00
X_{12} Distance to the central business district	9.54	5.34	23.71	1.28

dispersion of sales prices and some observations are quite distanced within this interval. Another clear conclusion from the histogram is the leptokurdic nature of the distribution. Table 1 provides a list of all variables used in our study as well as some sample statistics. The regressors can be grouped into two main categories. The first contains a series of dwelling attributes commonly used in hedonic price studies, such as dwelling area, land area, the total number of bedrooms and bathrooms, as well as the dwelling age in 1994. The second contains a series of locational characteristics of the dwelling. They include distances to the central business district, the nearest lake, improved park, industrial sector, commercial district and wetland. Also included is a measure dwelling elevation. All linear measurements are in kilometers, and all area measurements are in squared meters.

The data come from two major sources; MetroScan, which compiles real estate data from assessor's records for numerous U.S. cities provided most of the house's structural attributes data; Metro Regional Services, a regional government agency provided the locational characteristics. All distance calculations were made using a raster system where all data are arranged in grid cells. Each cell is a 15.6-meter square with distances measured using the Euclidean norm from the center of the dwelling land parcel to the nearest edge of the feature.

Wetland locations are based on the U.S. Fish and Wildlife Service's National Wetlands Inventory in Oregon. Wetlands vary from primarily open water to forest and grassland that is wet only part of the year. Although the area of urban wetlands has been declining in the U.S. the section of Multnomah county included in our study has more than 4500 wetlands and deep water habitats, varying in size from 1 to 358 acres (Mahan et al. 2000).

We have some a priori beliefs regarding the effects of these various characteristics on housing prices. Specifically, we expect a positive association between sales prices and the dwelling's area, number of bathrooms, land area, the dwelling's elevation and the distance to the nearest industrial zone. Conversely, we expect a negative association between sales prices and distances to lakes, parks and the central business district as well as dwelling age. We have no a priori expectations regarding the effects of wetland dis-

tance on sales prices. Although proximity to wetlands may be perceived as a desirable location characteristic due to enhanced view quality or increased pollution protection, it can also be undesirable due to development restrictions on nearby properties, bad odor and wildlife annoyances. We are also unsure about the effects of proximity to commercial districts on house prices. Although close proximity may be undesirable due to increased traffic and noise, large distances maybe undesirable due to increased transportation costs.

It is important to point out that our a priori expectations are from a descriptive point of view rather crude. One of the advantages of departing from a linear parametric model involves the possibility of unveiling much richer patterns of association between regressand and regressors. Put differently, the standard practice in applied economic work of revealing expected parameter signs is in itself evidence of how restricted linear regression parametric modeling can be.

4. Empirical modeling and computations

Since two of the variables used as regressors are categorical (number of bathrooms and bedrooms), we estimate the following semiparametric version of (1),

$$E(Y|X_1 = x_{t1}, \ldots, X_{12} = x_{tD}) = \alpha + \sum_{d=1}^{3} \alpha_d C_{td} + \sum_{d=1}^{4} \beta_d D_{td} + \sum_{d=3}^{12} m_d(x_{td}) \qquad (8)$$

for $t = 1, \ldots, n$, Y, $X_3, \ldots, X_{12}$ as indicated in Table 1,[6] $C_{td} = \begin{cases} 1, & \text{if } x_{t1} = d+1 \\ 0, & \text{otherwise} \end{cases}$ for $d = 1, 2$, $C_{t3} = \begin{cases} 1, & \text{if } x_{t2} > 3 \\ 0, & \text{otherwise,} \end{cases}$ $D_{td} = \begin{cases} 1, & \text{if } x_{t2} = d+1 \\ 0, & \text{otherwise} \end{cases}$ for $d = 1, 2, 3$ and $D_{t4} = \begin{cases} 1, & \text{if } x_{t2} > 4 \\ 0, & \text{otherwise} \end{cases}$.

The estimation procedure requires initial estimates for σ^2 and $\theta_{dd}(2,4)$. The latter requires estimates for $m_d^{(4)}(\cdot)$ and $m_d^{(2)}(\cdot)$. To obtain these initial values we first fit a linear parametric regression model that results from a Taylor's expansion of order 5 that explores the additivity of the conditional expectation of Y including all continuous regressors that appear in (8). This estimated regression provides initial estimates for the second and fourth derivatives of m_d, which are then used to construct the first estimates of $\theta_{dd}(2,4)$. These, together with an estimate for the variance σ^2 are used in Eq. (6) to obtain $\hat{g}$, which is given in Table 2. $\hat{g}$ is then used to fit (8) using the B-estimator and a local polynomial estimator of order $p = 3$. From this fit of the additive model we obtain a new estimate for the second derivatives of m_d. These new estimates of the second derivatives are used to obtain an updated estimate for $\theta_{dd}(2,2)$ which together with the

[6] The data, as well as the computer code for the implementation of the estimation procedure, which was written in the GAUSS v.4.0 (2002) programming environment, are available from the first author. A similar MATLAB code is described in Opsomer and Ruppert (1998) and is available from www.iastate.edu/~jopsomer/research.html.

Table 2. Estimated bandwidths

Variable	$\hat{g}_n$	$\hat{\kappa}_n$	$\hat{h}_n$
X_3	48.34	40.77	26.27
X_4	817.38	1205.26	776.44
X_5	9.51	10.63	6.85
X_6	0.90	0.87	0.57
X_7	0.44	0.50	0.32
X_8	0.14	0.22	0.14
X_9	0.02	0.03	0.02
X_{10}	0.51	0.54	0.34
X_{11}	0.18	0.23	0.15
X_{12}	1.58	1.29	0.83

variance estimate are used to obtain $\hat{\kappa}_{dn}$ according to (7). The $\hat{\kappa}_n$ reported in Table 2 is then used to fit (8) once again, so that a new estimate for σ^2 is obtained. This new variance estimate is used with the updated estimate of $\theta_{dd}(2,2)$ to obtain $\hat{h}_n$ according to (5) and reported in Table 2. Finally, $\hat{h}_n$ is used to obtain a final fit of (8), which is then used to obtain a final estimate for σ^2. We denote these final estimates by α^b, α_d^b, β_d^b, $\boldsymbol{m}_d^b(x_d)$ and $\hat{\sigma}^2$.

We have a series of observations to make regarding the estimation procedure. First, the fact that the estimated bandwidths depend on y produces estimators that are not projectors and are nonlinear. However, as desired, the cycles of the backfitting algorithm produced a decreasing sequence of $||y - 1_n\alpha^0 - \sum_{\delta=1}^{D} \boldsymbol{m}_\delta^b(x_\delta)^v||_2^2$. Convergence was obtained for the first fitting of the ANRM after 9 cycles, for the second fitting after 7 cycles, and for the final fitting convergence was attained after 11 cycles, all with a level of tolerance of 0.001. With $n = 1000$ the computation time for the estimates was approximately 3 hours and 16 minutes on a 1 Ghz Pentium IV PC and for each of the three ANRM that must be estimated, the computations are of order $O(vDn)$, where v is the number of iterations necessary for the convergence of the B-estimator. The estimated regressions are the solid lines that appear on Fig. 2. In the next section we provide a detailed discussion of the results.

Second, one major difficulty in fitting an ANRM via backfitting with data driven estimated bandwidths, is that we are unable to construct asymptotically valid confidence intervals for the estimated regressions. It is possible however to obtain an estimated covariance matrix for each $\boldsymbol{m}_d^b(x_d)$ for $d = 3, \ldots, 12$. This results from the fact that at convergence, $\boldsymbol{m}_d^b(x_d)$ can be written as $R_d y$ for some $n \times n$ matrix R_d. Hence, $V(\boldsymbol{m}_d^b(x_d))$ can be estimated by $\hat{\sigma}^2 R_d R_d'$. We obtain R_d by noting that at convergence it can be calculated by the following numerical procedure. We define the identity matrix of size $n = 1000$ to be I_n and denote its i^{th} column by $I_{.i}$. Using $\hat{h}_n$ we fit the ANRM in (8) with regressand given by $I_{.i}$. This produces n-dimensional vectors $\tilde{\boldsymbol{m}}_d^b(x_d)$ that correspond to the i^{th} column of R_d, for $d = 1, \ldots, n$. Hence, we run 1000 new ANRM to obtain R_d. The dashed lines that appear in Fig. 2 are pointwise lower and upper bounds on the estimated regressions constructed by multiplying by 2 the square root of the diagonals of $\hat{\sigma}^2 R_d R_d'$. In all of these 1000 auxiliary models, convergence was attained in less than 6 cycles. Although

these confidence bands are not exact, they provide what we believe are very suitable approximations.[7]

Third, our final variance estimator is obtained according to the following expression $\hat{\sigma}^2 = \frac{1}{n-tr(2R-RR')} ||y - \sum_{d=1}^{3} \alpha_d^b C_d - \sum_{d=1}^{4} \beta_d^b D_d - \sum_{d=3}^{12} \boldsymbol{m}_d^b(x_d)||_2^2$, where $R \equiv \sum_{d=1}^{12} R_d$ and $tr(\cdot)$ is the trace operator. The purpose is to account for the degrees of freedom (df) inherent in our estimation procedure. For our data we obtained $df = 889$. Although this is smaller than the degrees of freedom of a quadratic parametric approximation ($df = 927$), it is much larger than what would result if an approximation of order 3 were used. Put simply, our nonparametric estimator does not seem to require an unreasonable amount of degrees of freedom. The final variance estimate for the ANRM is $\hat{\sigma}^2 = 1.08 \times 10^9$.

It is instructive to compare our results to linear parametric specifications for this hedonic price function. We guided our choice for the parametric model by the following constraints: a) we consider only models for $E(Y|X_1, \ldots, X_D)$, ruling out specifications such as $E(T(Y)|X_1, \ldots, X_D)$, where $T(\cdot)$ is some transformation of Y. This rules out models that are well known to be ill-specified, such as the models proposed by Box and Cox (1964), as well as popular semi-log ($T(Y) = log(Y)$) models. We do so because we concur with Wooldridge(1992) in that our interest is on the conditional expectation of Y not a transformation of Y. Such transformations produce ambiguities on the analysis and interpretations of the regression results; b) we assume that the researcher has no a priori knowledge about the specific nature of $E(Y|X_1, \ldots, X_D)$ as to permit the estimation of a well defined nonlinear regression model. Within these constraints, we specified and estimated two alternative parametric regression models given by

$$E(Y|X_1 = x_{t1}, \ldots, X_{12} = x_{tD}) = \alpha + \sum_{d=1}^{3} \alpha_d C_{td} + \sum_{d=1}^{4} \beta_d D_{td} + \sum_{d=3}^{12} \theta_d x_{td} \tag{9}$$

and

$$E(Y|X_1 = x_{t1}, \ldots, X_D = x_{tD}) = \alpha + \sum_{d=1}^{3} \alpha_d C_{td} + \sum_{d=1}^{4} \beta_d D_{td} + \sum_{d=3}^{12} \theta_d x_{td} + \frac{1}{2} \sum_{d=3}^{12} \sum_{\delta=3}^{12} \theta_{d\delta} x_{td} x_{t\delta}, \tag{10}$$

where $\theta_{d\delta} = \theta_{\delta d}$. Model (9) is convenient for comparison purposes because it is a linear restriction ($m_d(x_{td}) = \theta_d x_{td}$ for $d = 3, \ldots, 12$) of (8). Model (10) is instructive because similar parametric second order approximations have been extensively used in applied econometrics. Tables 3 and 4 provide ordinary least squares estimates for the parameters in (9) and (10), as well as some other commonly reported regression statistics.

Since (9) is a restriction of (8) we follow Hastie and Tibshirani (1990) and perform a test for linearity based on the statistic,

[7] See Hastie and Tibshirani (1990). We remind the reader that even if the asymptotic distribution of the B-estimator were available, the confidence bands would still be approximations.

Table 3. Parametric models estimates and t-statistics

Model(9)					
Parameter	Estimate	t-Stat.	Parameter	Estimate	t-Stat.
α	13392.08	1.38	θ_4	15.45	6.00
α_1	3135.73	0.81	θ_5	−146.79	−1.90
α_2	10078.53	1.41	θ_6	−5311.94	−7.27
α_3	109361.11	5.29	θ_7	−3722.75	−1.76
β_1	10257.12	1.31	θ_8	−7540.62	−1.45
β_2	1979.26	0.25	θ_9	530035.51	−11.39
β_3	2277.09	0.26	θ_{10}	−10480.36	−4.87
β_4	−5726.73	−0.52	θ_{11}	27868.52	5.46
θ_3	931.19	26.31	θ_{12}	−2954.24	−8.18
$\hat{\sigma}^2 = 1.84 \times 10^9$, $df = 982$, $R^2 = 0.74$					

$$f = \frac{(RSS_{(9)} - RSS_{(8)})/\gamma_2 - \gamma_1}{RSS_{(8)}/n - \gamma_2}$$

which is approximately distributed as an $F_{\gamma_2-\gamma_1,n-\gamma_2}$, where $\gamma_1 = 17$, $\gamma_2 = tr(2R - RR')$, and $RSS_{(i)}$ is the residual sum of squares of model (i). In our case $f = 8.39$ leading to a rejection of model (9).

A comparison of (8) and (10) is more difficult since these models are nonnested due to the presence of interaction and squared terms in the parametric alternative. We observe that in parametric models such as (10), any nonlinearities on the conditional mean are captured by the terms of second order, i.e., interaction and squared terms. Hence, if one is interested on the relationship between Y and X_d, ceteris paribus, it is necessary to arbitrarily choose values for all X_δ where $\delta \neq d$ to characterize such relationship. Choosing average sample values for other regressors is common practice in the empirical literature. We observe, however, that this procedure generates simply one of (uncountably) many potential estimated regressions. The same can be said about first derivatives and elasticities that derive from these models. In contrast, the ANRM captures these nonlinearities directly through the flexible specification of m_d. For comparison with the ANRM we have estimated each regression direction based on (10) with all other regressors evaluated at their averages. Their estimated price effects are the large dash lines that appear on Fig. 2.[8] In addition, we performed a test for the presence of interaction terms proposed by Hastie and Tibshirani(1990). The test involves estimating the following artificial regression,

$$r_t = \sum_{d=3}^{12} \sum_{j>d} \psi_{dj} m_d^b(x_{td}) m_j^b(x_{tj}), \tag{11}$$

where $r_t = y_t - \alpha^b - \sum_{d=1}^{3} C_{td}\alpha_d^b - \sum_{d=1}^{4} D_{td}\beta_d^b - \sum_{d=3}^{12} m_d^b(x_{td})$ are the residuals from the estimated additive model. A conventional Student's-t statistic for $H_0 : \psi_{dj} = 0$ against $H_A : \psi_{dj} \neq 0$ is calculated. We interpret rejection of H_0

[8] Similar graphs that include plots of partial residuals $e_{t\delta} = y_t - \alpha^b - \sum_{d=1}^{3} \alpha_d^b C_{td} - \sum_{d=1}^{4} \beta_d^b D_{td} - \sum_{d=3,\delta\neq d}^{12} m_d^b(x_{td})$ are available upon request.

Table 4. Parametric models estimates and t-statistics

Model(10)					
Parameter	Estimate	t-Stat.	Parameter	Estimate	t-Stat.
α	−94151.02	−2.82	$\theta_{4,5}$	0.49	2.79
α_1	2866.82	0.88	$\theta_{4,6}$	1.25	0.74
α_2	13992.89	2.46	$\theta_{4,7}$	−2.67	−0.43
α_3	−113035.24	−4.21	$\theta_{4,8}$	−23.39	−2.01
β_1	10831.95	1.72	$\theta_{4,9}$	−80.66	−0.97
β_2	4529.74	0.70	$\theta_{4,10}$	−0.97	−0.20
β_3	8753.98	1.24	$\theta_{4,11}$	−8.33	−0.83
β_4	−5617.42	−0.62	$\theta_{4,12}$	0.69	0.95
θ_3	1476.96	8.22	$\theta_{5,6}$	−71.83	−2.37
θ_4	−26.46	−1.29	$\theta_{5,7}$	7.34	0.08
θ_5	1417.19	3.52	$\theta_{5,8}$	−13.77	−0.06
θ_6	−2480.69	−0.53	$\theta_{5,9}$	−1201.10	−0.61
θ_7	−17556.15	−1.23	$\theta_{5,10}$	−39.42	−0.45
θ_8	−49032.11	−1.56	$\theta_{5,11}$	447.51	1.85
θ_9	560523.27	1.80	$\theta_{5,12}$	−54.56	−2.94
θ_{10}	48926.78	3.44	$\theta_{6,7}$	−136.76	−0.10
θ_{11}	37124.87	1.13	$\theta_{6,8}$	4657.10	1.87
θ_{12}	3893.15	1.50	$\theta_{6,9}$	−4577.28	−0.17
$\theta_{3,3}$	0.68	1.34	$\theta_{6,10}$	3636.65	2.64
$\theta_{4,4}$	−0.00	−0.22	$\theta_{6,11}$	−9860.05	−4.01
$\theta_{5,5}$	−10.22	−2.35	$\theta_{6,12}$	138.53	0.72
$\theta_{6,6}$	1338.15	2.28	$\theta_{7,8}$	3935.11	0.53
$\theta_{7,7}$	5346.76	1.11	$\theta_{7,9}$	62305.17	0.63
$\theta_{8,8}$	35302.09	1.63	$\theta_{7,10}$	−1890.65	−0.58
$\theta_{9,9}$	13851045.00	8.50	$\theta_{7,11}$	−6117.40	−0.70
$\theta_{10,10}$	9772.72	2.27	$\theta_{7,12}$	−841.78	−1.32
$\theta_{11,11}$	69751.38	3.39	$\theta_{8,9}$	104457.00	0.68
$\theta_{12,12}$	385.55	3.02	$\theta_{8,10}$	6751.46	0.86
$\theta_{3,4}$	0.25	4.08	$\theta_{8,11}$	−1863.69	−0.11
$\theta_{3,5}$	−5.95	−4.65	$\theta_{8,12}$	1296.44	1.07
$\theta_{3,6}$	−53.23	−3.89	$\theta_{9,10}$	−654516.47	−8.63
$\theta_{3,7}$	75.07	2.06	$\theta_{9,11}$	−581740.62	−4.37
$\theta_{3,8}$	−51.75	−0.49	$\theta_{9,12}$	−56728.18	−5.35
$\theta_{3,9}$	1010.16	1.60	$\theta_{10,11}$	1265.49	0.17
$\theta_{3,10}$	−155.93	−4.48	$\theta_{10,12}$	−116.19	−0.19
$\theta_{3,11}$	228.66	2.76	$\theta_{11,12}$	1519.61	1.23
$\theta_{3,12}$	−55.13	−7.74			
$\hat{\sigma}^2 = 1.09 \times 10^9$, $df = 927$, $R^2 = 0.85$					

as evidence that the interaction of regressors is strong enough to reject additivity. Conversely, failing to reject H_0 lends support to the additivity assumption. Our test supports the additive specification in (8).[9] Models (8)

[9] Results are available upon request. Significant interaction terms can be easily incorporated into the ANRM by specifying $E(Y|X_1 = x_{t1}, \ldots, X_{12} = x_{tD}) = \alpha + \sum_{d=1}^{3} \alpha_d C_{td} + \sum_{d=1}^{4} \beta_d D_{td} + \sum_{d=3}^{12} m_d(x_{td}) + \sum_{k=1}^{K} \gamma_k z_{tk}$, where $z_{tk} \equiv x_{td} x_{t\delta}$ for some $d, \delta = 3, \ldots, 12$. We estimated this model and compared to estimates of $E(Y|X_1 = x_{t1}, \ldots, X_{12} = x_{tD}) = \alpha + \sum_{d=1}^{3} \alpha_d C_{td} + \sum_{d=1}^{4} \beta_d D_{td} + \sum_{d=3}^{12} \theta_d x_{td} + \sum_{k=1}^{K} \gamma_k z_{tk}$ using the approximate F-test described above. Once again the parametric specification is rejected in favor of the ANRM.

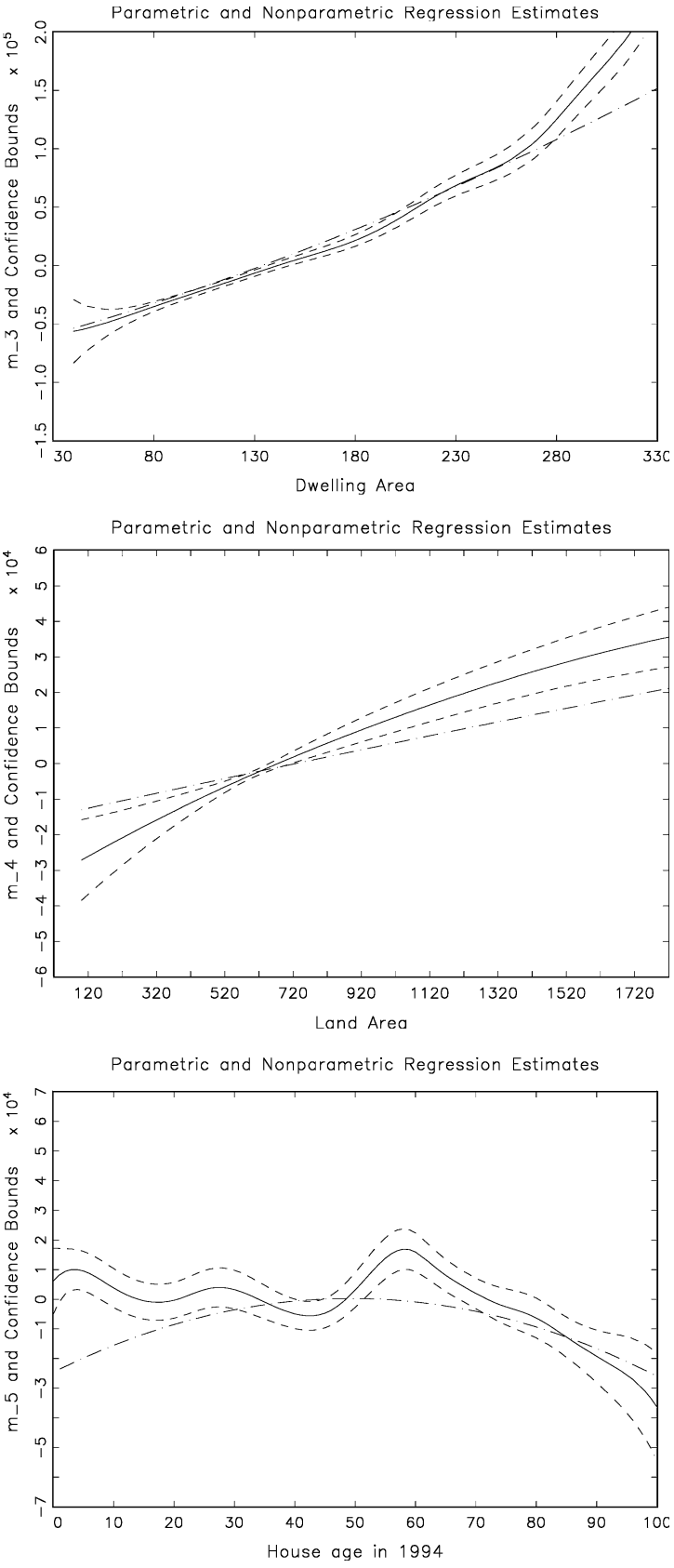
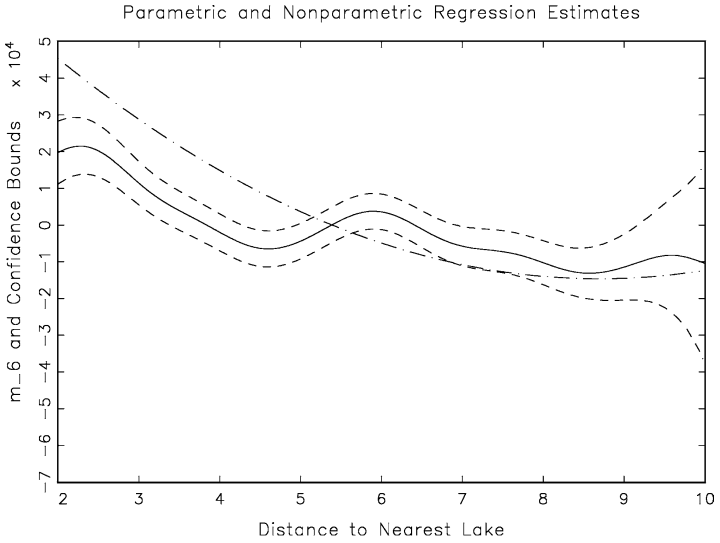
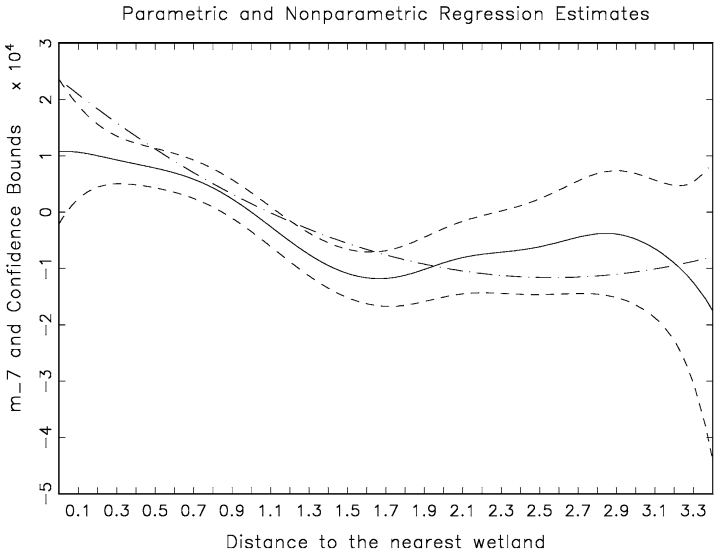


Fig. 2. Parametric and Nonparametric Regression Estimates Distance to the nearest improved park Distance to the nearest industrial zone Distance to the nearest business district


Parametric and Nonparametric Regression Estimates
m_6 and Confidence Bounds x 10^4
5 4 3 2 1 0 −1 −2 −3 −4 −5 −6 −7
2 3 4 5 6 7 8 9 10
Distance to Nearest Lake


Parametric and Nonparametric Regression Estimates
m_7 and Confidence Bounds x 10^4
3 2 1 0 −1 −2 −3 −4 −5
0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5 1.7 1.9 2.1 2.3 2.5 2.7 2.9 3.1 3.3
Distance to the nearest wetland

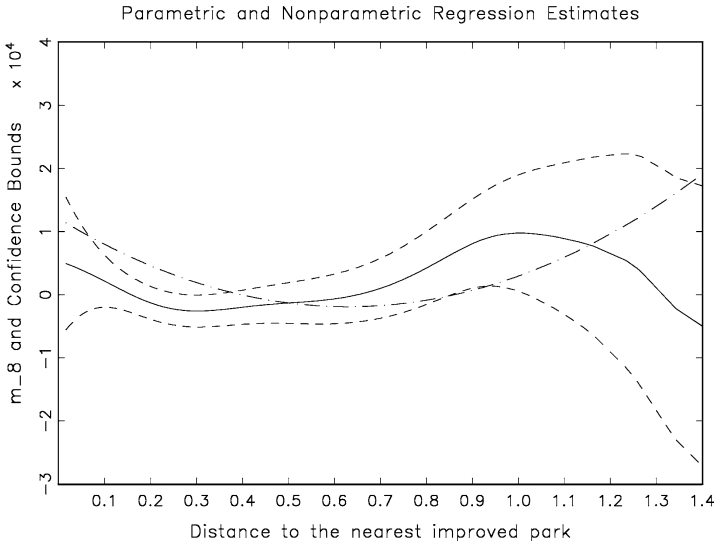
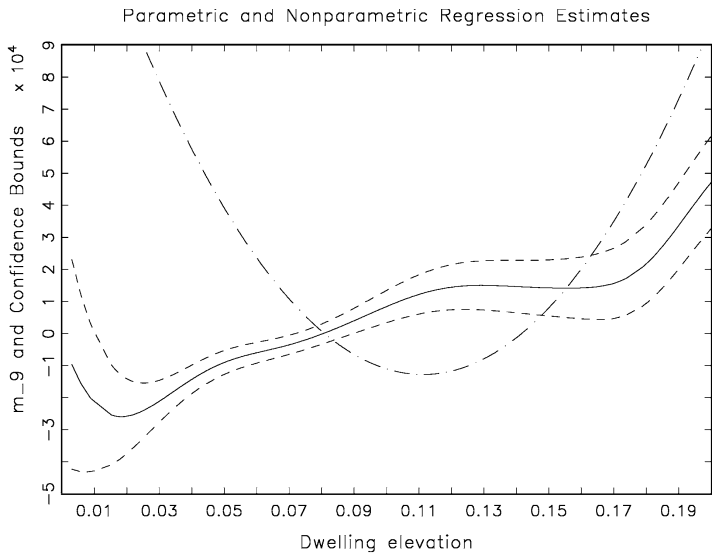
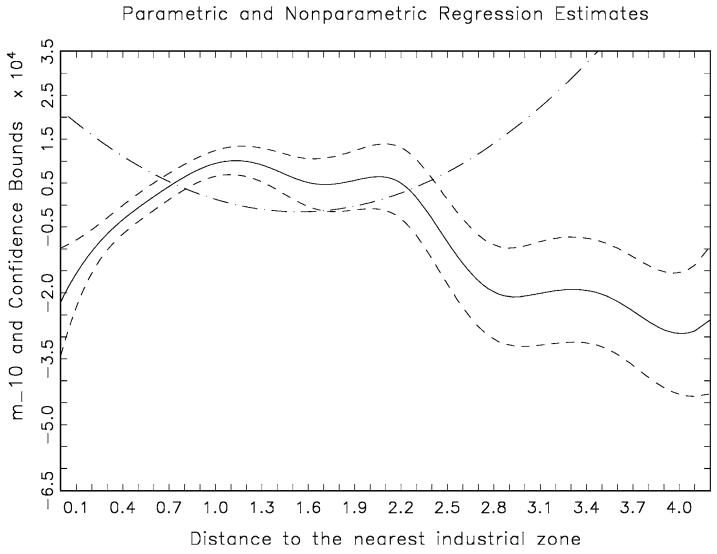

Parametric and Nonparametric Regression Estimates
m_8 and Confidence Bounds x 10^4
4 3 2 1 0 −1 −2 −3
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4
Distance to the nearest improved park

Fig. 2. (Contd.)


Parametric and Nonparametric Regression Estimates
m_9 and Confidence Bounds x 10^4
−5 −3 −1 0 1 2 3 4 5 6 7 8 9
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19
Dwelling elevation


Parametric and Nonparametric Regression Estimates
m_10 and Confidence Bounds x 10^4
−6.5 −5.0 −3.5 −2.0 −0.5 0.5 1.5 2.5 3.5
0.1 0.4 0.7 1.0 1.3 1.6 1.9 2.2 2.5 2.8 3.1 3.4 3.7 4.0
Distance to the nearest industrial zone

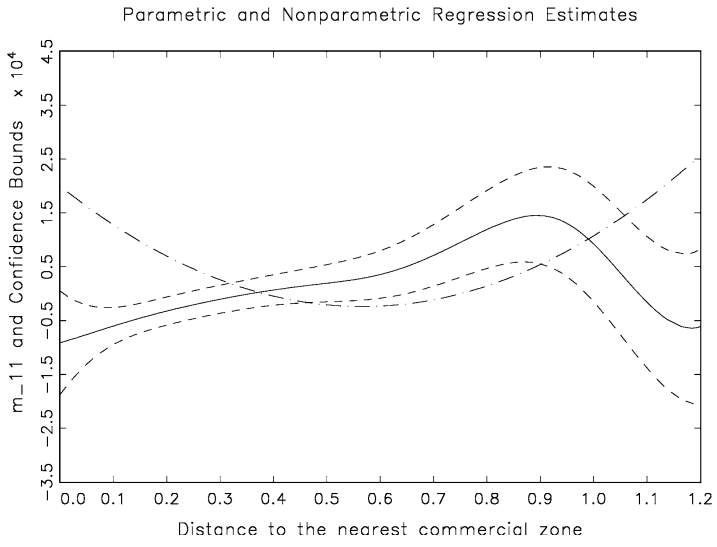

Parametric and Nonparametric Regression Estimates
m_11 and Confidence Bounds x 10^4
−3.5 −2.5 −1.5 −0.5 0.5 1.5 2.5 3.5 4.5
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
Distance to the nearest commercial zone

Fig. 2. (Contd.)

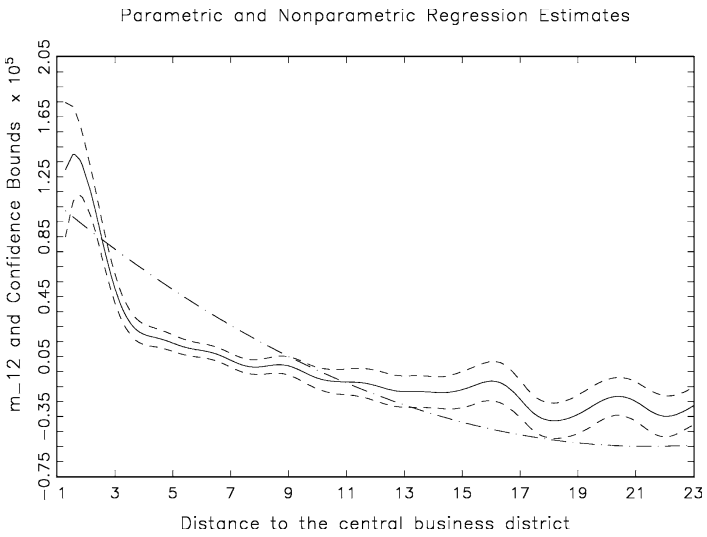


Fig. 2. (Contd.)

and (10) are also compared in subsection 5.3 by performing an out-of-sample forecast exercise.

5. Estimation results and analysis

5.1. Parametric estimates

The parametric models we estimate are simple and for the casual observer may seem rather inappropriate, especially given the availability of much richer parametric specifications. However, after informal experimentation with several alternative parametric specifications that do not expand the set of regressors, we were surprised to learn that the best overall fit resulted from model (10). Our informal search is by no means evidence that one cannot fit a better parametric model, but it is indicative that (10) does reasonably well against some obvious parametric alternatives.

The results in Tables 3 and 4 suggest that the most important attributes in determining sales prices are the dwelling area, the lot square footage and dwelling age. Among the locational attributes the most significant determinants of sales prices are the dwelling's elevation, its distance to the central business district, nearest wetland and nearest commercial district. The results are reasonable and generally in line with comparable previous studies, e.g., Iwata et al. (2000). We will make more specific comments regarding the estimated parametric model as we contrast its results with that of the ANRM. All of the variables used in the estimation of (10) are in deviation form. Hence, the parametric model can be interpreted as a second order Taylor's approximation around the sample average of the regressors.

5.2. Nonparametric estimates

We start our analysis by observing that the estimated regressions have very reasonable shapes. No clear instance of over or under smoothing is apparent

and the fact that a common bandwidth was used in each regression direction did not create problems for most of the regressors' data range. The estimated parameters associated with the dummy variables in (8) are given by $\hat{\alpha}_1 = 9700.53$, $\hat{\alpha}_2 = 8309.05$, $\hat{\alpha}_3 = 51628.19$, $\hat{\beta}_1 = 10371.50$, $\hat{\beta}_2 = 8108.13$, $\hat{\beta}_3 = 10728.55$ and $\hat{\beta}_4 = -21364.61$, which are with the exception of $\hat{\beta}_1$ different from those in the parametric models. There seem to be some similarities between the regression estimates produced by parametric model (10) and the ANRM for some regressors, e.g., dwelling area and land area, however for most regressors the two models reveal significantly different impact on sales price. These are more pronounced for the impact of distance to the nearest lake, industrial zone, central business district and dwelling elevation. Before turning to a more specific analysis of our results we point out that estimated regression behavior close to the boundaries should be interpreted with caution due to the scarcity of observations in these data ranges.

The estimated regressions for the influence of a dwelling's area and land area on price have very reasonable shape. In the range $(60, 230)$, where most of the data are concentrated, m_3^b is increasing and nearly linear, with a first derivative that oscillates between 300 and 1,000 dollars. The estimated regression is slightly convex for dwellings with area larger than 230, indicating a larger impact of size on prices for larger dwellings. It is not surprising, given the near linearity of m_3^b, that the parametric model produces results that are very similar to those suggested by m_3^b.

In general, as we expected, m_4^b suggests a positive association between sales price and land area. The impact on sales prices of land area is much less pronounced than that of dwelling area. Within ± 1 standard deviation $(134, 1,320)$ from the average land size (727) sales prices vary by less than 60 dollars for a marginal increase in land area. The marginal impact decreases with land area. The results from the parametric model are similar in that the estimated regression has positive slope, but the impact of land area on sales prices is even smaller than that suggested by the nonparametric model. We believe that in this case, even though m_4 is nearly linear, the parametric model underestimates the impact of land area on prices.

The impact of dwelling age on sales prices is difficult to intepret. The data seems to indicate that housing developments has occured in fairly well defined phases. We were able to identify four data clusters corresponding to the periods $1920 - 1930$, $1945 - 1956$, $1970 - 1981$, and $1991 - 1994$. The fact that the data is clustered has produced a bandwidth that seems to slightly undersmooth m_5^b, hence the wiggly appearence of the estimated regression. Another potential explanation for the shape of m_5^b is the existence of vintage effects. We are able, however to discern some patterns. Dwelling age impacts prices negatively in the $(1975, 1990)$, i.e., for house with age between 4 and 18 years. Ceteris paribus a house built in 1990 will cost about $10,000$ dollars more than one built in 1975. However, for dwellings that are between 20 and 50 years old the impact of age on sales prices seems to be much less pronounced. In fact, not counting the 1935-1945 period, for dwellings between 20 and 80 years old, the impact of age on sales prices is smaller than $\pm 10,000$ dollars. The sharp increase in m_5^b from $1935 - 1945$ seems to be caused by a combination of sample variability and undersmoothing. After 80 years of age sales price continuously fall with age. The results that derive from the parametric model are quite different and somewhat implausible. Age, according to model (10) has a

positive impact on price for dwellings up to approximately 40 years old. After that there is a very small negative impact.

Close proximity to a lake has a substantial positive impact on sales prices. However, this impact is restricted to dwellings that are less than 4 kilometers away from a lake. Here, the impact on prices from moving away from a lake is substantial. Ceteris paribus, moving a house from 2 to 4.5 kilometers away from a lake produces a drop in sales price of approximately $30,000$ dollars. The influence of proximity to a lake on sales price falls dramatically after a dwelling is more than 4 kilometers away. In fact, $m_5(\cdot) \equiv 0$ falls within the confidence band around m_5^b for almost every point in the data range. These results are very intutitive. After a certain distance, use of a lake for recreational purposes is limited and even prohibited by some neighborhood associations (in the case of private lakes) and transportation costs start to outweight the benefits of lake use. The parametric model produces a somewhat different interpretation of the data. The drop in sales prices as one moves away from a lake is gradual and decreasing. In contrast with the nonparametric model, the distance to a lake remains important up until approximately 7.5 kilometers. Somewhat disturbing is the fact that the parametric model seems to predict a slight increase in sales prices for dwellings that are more than 9 kilometers away from a lake.

There is a negative relationship between sales prices and distance to the nearest wetland for most of the data range. This negative impact is more pronounced within ± 1 standard deviation of the average distance (1.09 kilometers). Moving a dwelling adjacent to a wetland 2 kilometers away produces a decrease in price of about $20,000$ dollars. Although this effect is smaller than the lake effect, it is significant and suggests a market that incorporates quite well the value of environmental amenities. As in the case with distance to a lake, dwellings that are farther away from a wetland (in this case more than 2 kilometers) have prices that are impacted little by wetland distance. Note that the impact of wetland distance on sales prices dies out more rapidly than the impact of distance to a lake. The parametric model seems to do a fairly good job in estimating the impact of wetland proximity on price. There seems however to be overestimation of this impact for dwellings that are within 2 kilometers of wetlands.

Proximity to an improved park has little impact on sales prices. In fact, $m_8(\cdot) \equiv 0$ falls within the confidence band we constructed for virtually all points in the data range. Price variations where most of the data lies $(.1, .6)$ kilometers are all within $\pm 3,000$ dollars. Once again, the parametric model overestimates this regression direction in the proximity of the origin. Rather unappealing and couterintuive is the positive slope of the estimated regression for distances greater than 0.8 kilometers.

The regression estimate m_9^b predicts a positive impact of dwelling elevation on sales prices for the interval $(20, 130)$ meters. The gain in price for this data range is $40,000$ dollars. After 130 meters m_9^b levels out with a derivative that is very close to zero. For elevations above 170 meters the impact on sales prices increases sharply, however as in previous cases there are scarcely any observations on this data range. It seems reasonable to expect a positive association between elevation and prices, due to the possibility of views and quieter streets. However, higher elevations may also be associated with higher transportation costs. Hence, the fact that m_9^b levels off is expected and intuitively appealing. The parametric model produces quite different predictions.

Dwelling prices first drop (until about 110 meters) with elevation then rise continuously at an increasing rate. Clearly, there is no cost associated with elevation gain. According to the parametric model, a house at average elevation (81 meters) has a price increase of almost 80,000 dollars when it gains 100 meters in elevation.

The regression estimate m^b_{10} clearly captures the negative impact on sales prices of dwellings that are close to industrial zones. Furthermore, it indicates that this negative effect is specially intense very close to the industrial zone. For example, moving a dwelling that is adjacent to an industrial zone 400 meters away, increases its value ceteris paribus by approximately 20,000 dollars. The benefits from moving away from industrial zones diminish at an increasing rate. Interestingly, there seems to be a negative impact on sales prices after about 2 kilometers, but once again there are a small number of observations in this range of data. It is likely that this results from increased transportation costs to the work place. Once again the predictions of the parametric model are quite different. Prices are predicted to fall as we move away from industrial zones (until about 1.1 kilometers). For distances above 1.1 kilometers, prices increase continuously at an increasing rate. The fall in prices due to increased transportation costs captured by the nonparametric model is not revealed by the parametric model.

The regression estimate m^b_{11} suggests a positive relationship between sales prices and distances to the nearest commercial zone. The impact however is much smaller than that of X_{10}. That is, the problems of congestion, pollution, etc. associated with proximity to a commercial zone do not seem as intense as those related to proximity to industrial zones. As in the case of m^b_{10} we observe that after a certain distance, there is a reversal of the first derivative sign, and distance begins to have a negative impact on sales prices. Once again, we attribute this effect to a predominance of transportation costs over congestion costs. Here, the parametric model performs poorly once again reproducing the same prediction patterns observed for the case of X_{10}.

Proximity to the central business district (CBD) has a very strong positive impact on sales prices for the first 3 kilometers. Ceteris paribus moving a dwelling that is adjacent to the CBD to a location 3 kilometers away may reduce its price by as much as 100,000 dollars. However, while still negative this effect is reduced dramatically after 3 kilometers. For example, moving from 3 kilometers to 9 kilometers away from the CBD impacts price by only 20,000. This is in line with our general expectation. The effect of distance to the CBD for the parametric and nonparametric models are quite different. Once again, the very sharp initial decline on prices as a typical dwelling is moved away from the CBD is largely unaccounted by the parametric model.

Overall, we believe that the nonparametric results are more appealing than those suggested by the parametric model. It comes as no surprise that whenever the estimated regressions are close to linear the parametric model provides very adequate responses. However, even in this case estimated slopes can be over or under estimated depending on the pattern of dispersion of the data, as in the case of land area.

5.3. Out-of-sample forecast exercise

One of the characteristics of hedonic price models is that they can be used to forecast prices, given a set of product characteristics. We were able to perform

a simple out-of-sample forecast evaluation of both the parametric and nonparametric models based on a new sample of 1000 observations, which we denote by $\{(y_t^N, x_{t1}^N, \ldots, x_{t12}^N)\}_{t=1}^n$.[10] The new sample comes from the same housing market and corresponds to the same regressors and regressand. The exercise is particularly useful for comparison purposes, since as mentioned above, models (8) and (10) are nonnested. Using the new data, we obtained forecasted sales prices for the parametric models (9) and (10),

$$\hat{y}_t^p = \hat{\alpha} + \sum_{d=1}^{3} \hat{\alpha}_d C_{td}^N + \sum_{d=1}^{4} \hat{\beta}_d D_{td}^N + \sum_{d=3}^{12} \hat{\theta}_d x_{td}^N, \; \hat{y}_t^p = \hat{\alpha} + \sum_{d=1}^{3} \hat{\alpha}_d C_{td}^N + \sum_{d=1}^{4} \hat{\beta}_d D_{td}^N + \sum_{d=3}^{12} \hat{\theta}_d x_{td}^N + \frac{1}{2} \sum_{d=3}^{12} \sum_{\delta=3}^{12} \hat{\theta}_{d\delta} x_{td}^N x_{t\delta}^N,$$

where $\hat{\alpha}, \hat{\alpha}_d, \hat{\beta}_d, \hat{\theta}_d, \hat{\theta}_{d\delta}$ are the least squares estimators, and for the ANRM,

$$\hat{y}_t^b = \alpha^b + \sum_{d=1}^{3} \alpha_d^b C_{td}^N + \sum_{d=1}^{4} \beta_d^b D_{td}^N + \sum_{d=3}^{12} m_d^b(x_{td}^N).$$

We obtained new values for the bandwidth vector $\hat{h}_n$, based on the new set of observed regressors, in the estimation of the ANRM but we did not update α^b. Using $\{y_t^N\}_{t=1}^n$ we calculate the square root of the average squared forecast error for the parametric models ($FE_{(9)}$,$FE_{(10)}$) and ANRM (FE^b) estimators to be $FE_{(9)} = 42,073$, $FE_{(10)} = 40,503$ and $FE^b = 27,853$. Hence, the forecast error of the ANRM is about .69 of that corresponding to parametric model (10). This forecast difference can have a substantial effect where prices need to be forecasted. For example, assuming a property tax rate of 0.014 of dwelling market value, this represents an average 177 dollars tax adjustment per dwelling per year. Assuming 200,000 tax units, this represents about 35.4 million dollars in property tax adjustments.

6. Conclusions

In this paper we have argued that the functional form specification problem common in hedonic price models can be conveniently addressed by modeling the conditional mean of prices as an additive nonparametric regression model. The approach is in our view vastly superior to a fully nonparametric design for several reasons. First, from a practical perspective, the smooths are very easy to interpret and visualize permitting therefore an analysis of the contribution of characteristics and attributes to prices in much the same way that is obtained in classical separable parametric models. Second, from a statistical perspective, a series of well known difficulties of an unrestricted fully nonparametric design are avoided.

Our bandwidth selection method is novel in economics, entirely data driven, and has performed well in pactice vis a vis the popular cross validation selection method. To facilitate the implementation of the method among applied economists we have written a GAUSS program that permits relatively

[10] We have re-estimated the ANRM based on $n = 2000$ and the results are virtually identical to those reported here for $n = 1000$. Results are available from the first author upon request.

fast implementation of the procedure. Most importantly, after implementing the procedure using data for Multnomah County, Oregon, we verify that the nonparametric additive model is able to identify associations and patterns of dependencies among the data that a reasonably adequate parametric model fails to unveil. Besides, an out-of-sample evaluation of the forecast errors of both the parametric and nonparametric models reveals a superior performance of the ANRM. We are optimistic that the econometric modeling strategy used in this study can be successfully used in a variety of settings.

Despite our optimism, we find that the estimation procedure used in this paper needs to be understood better. We note that although the tests we perform and the confidence bands we construct have been shown to have reasonable properties in simulation studies (Hastie and Tibshirani 1990), we are concerned with our inability to construct asymptotically valid confidence intervals and test of hypotheses.

References

Anglin P, Gencay R (1996) Semiparametric estimation of hedonic price functions. Journal of Applied Econometrics 11: 633–648

Box GEP, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society B 26: 211–252

Buja A, Hastie TJ, Tibshirani R (1989) Linear smoothers and additive models. Annals of Statistics 17: 453–555

Clapp JM, Kim H-J, Gelfand A (2002) Predicting spatial patterns of house prices using LPR and Bayesian smoothing. Real Estate Economics 30: 505–532

Dreyfus MK, Viscusi WK (1995) Rates of time preference and consumer valuations of automobile safety and fuel efficiency. Journal of Law and Economics 38: 79–105

Fan J (1992) Design adaptive nonparametric regression. Journal of the American Statistical Association 87: 998–1004

Fan J, Gasser T, Gijbels I, Brockmann M, Engel J(1993) Local polynomial fitting: a standard for nonparametric regression. Department of Statistics, UNC

Friedman JH, Stuetzle W (1981) Projection pursuit regression. Journal of the American Statistical Association 76: 817-823

Gencay R, Yang X (1996) A forecast comparison of residential housing prices by parametric and semiparametric conditional mean estimators. Economic Letters 52: 129–135

Hartog J, Bierens H (1991) Estimating a hedonic earnings function with a nonparametric method. In: Ullah A (ed) Semiparametric and nonparametric econometrics: studies in empirical economics. Springer, Berlin Heidelberg New York

Hastie TJ, Tibshirani RJ (1986) Generalized additive models. Statistical Science 1: 297–318

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall, New York

Hill RC, Knight JR, Sirmans CF (1997) Estimating capital asset pricing indexes. Review of Economics and Statistics 79: 226–233

Iwata S, Murao H, Wang Q (2000) Nonparametric assessment of the effects of neighborhood land uses on the residential house values. In: Fomby T, Carter Hill R (Eds) Advances in econometrics: Applying Kernel and nonparametric estimation to economic topics 14: JAI Press, New York

Mahan B, Polasky S, Adams R (2000) Valuing urban wetlands: A property price approach. Land Economics 76: 100–113

Opsomer J, Ruppert D (1997) Fitting a bivariate additive model by local polynomial regression. Annals of Statistics 25: 186–211

Opsomer J, Ruppert D(1998) A fully automated bandwidth selection method for fitting additive models. Journal of the American Statistical Association 93: 605–619

Pace RK (1993) Nonparametric methods with applications to hedonic models. Journal of Real Estate Finance and Economics 7: 185–204

Pace RK (1995) Parametric, semiparametric, and nonparametric estimation of characteristics values within mass assessment and hedonic pricing models. Journal of Real Estate Finance and Economics 11: 195–217

Pace RK (1998) Appraisal using generalized additive models. Journal of Real Estate Research 15: 77–99

Park B, Marron JS (1990) Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85: 66–72

Rosen S (1974) Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82: 34–55

Ruppert D, Wand MP (1994) Multivariate locally weighted least squares regression. The Annals of Statistics 22: 1346–1370

Ruppert D, Sheather SJ, Wand MP (1995) An effective bandwidth selector for least squares regression. Journal of the American Statistical Association 90: 1257–1270

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society B 53: 683–690

Simonoff JS (1996) Smoothing methods in statistics. Springer, Berlin Heidelberg New York

Stock J (1991) Nonparametric policy analysis: An application to estimating hazardous waste cleanup benefits. In: Barnett W, Powell J, Tauchen G (eds) Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the 5th International Symposium in Economic Theory and Econometrics. Cambridge University Press, New York

Walls M (1996) Valuing the characteristics of natural gas vehicles: An implicit markets approach. Review of Economics and Statistics 78: 266–276

Wooldridge J (1992) Some alternatives to the box-cox regression model. International Economic Review 33: 935–955

Yatchew A (1998) Nonparametric regression techniques in economics. Journal of Economic Literature 36: 669–721