Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Housing Economics 13 (2004) 68–84
JOURNAL OF HOUSING ECONOMICS
www.elsevier.com/locate/jhe

A prediction comparison of housing sales prices by parametric versus semi-parametric regressions☆

Okmyung Bin*

Department of Economics, East Carolina University, Greenville, NC 27858-4353, USA

Received 22 July 2003

Abstract

This study estimates a hedonic price function using a semi-parametric regression and compares the price prediction performance with conventional parametric models. This study utilizes a large data set representing 2595 single-family residential home sales between July 2000 and June 2002 from Pitt County, North Carolina. Data from Geographic Information Systems (GIS) are incorporated to account for locational attributes of the houses. The results show that the semi-parametric regression outperforms the parametric counterparts in both in-sample and out-of-sample price predictions, indicating that the semi-parametric model can be useful for measurement and prediction of housing sales prices.

JEL classification: R21; C14

Keywords: Housing market; Hedonic pricing; Price prediction; Semi-parametric regression

1. Introduction

Accurate prediction of housing sales price is important in the operation of the housing market. Home sellers and buyers wish to know a fair value for their house

☆ Thanks are due to H. Pollakowski and an anonymous referee for helpful comments. I also thank Ralph Forbes of the Pitt County Management Information Systems for making floodplain and property parcel data available.

* Fax: 1-252-328-6743.
E-mail address: bino@mail.ecu.edu (O. Bin).

1051-1377/$ - see front matter
doi:10.1016/j.jhe.2004.01.001

in particular at the time of the sales transaction. A precise estimate of the sales price of a house is of real importance to investors who face choices among housing securities and other investment opportunities (Shiller, 1993). Financial institutions try to obtain an accurate estimate of the market value to manage the risk better and consequently reduce the cost related to financing homeownership. Housing price estimates have been used for mortgage-lending decisions by major financial institutions such as Fannie Mae and Freddie Mac (Goldberg and Harding, 2003). In addition, many cities and counties base property taxes on the market value of a house, which must be updated periodically. Inaccurate appraisal of house values may result in substantial property tax adjustments. However, the accurate prediction of the house price is difficult because residential housing is a composite good which is typically sold as a package of various factors, such as location, environment, structural attributes, etc. It is not obvious how to select relevant factors among others and how to account for these factors in predicting the selling price of a house.

Hedonic price models have been used as a tool to estimate the market value of a house for several decades (Mason and Quigley, 1996; Palmquist, 1980; Rosen, 1974).[1] This method assumes that the housing price reflects the value placed on a particular set of housing attributes. For instance, a house may be valued at a certain price based on quantitative characteristics such as the age of the house, the number of rooms, and garage space, and qualitative factors such as the geographical location, school districts, and environmental quality and so on. Therefore, the price of one house relative to another will differ with the amounts of various attributes inherent in one house relative to another. Regression analysis of the hedonic price models allows the researcher to construct a house price index and to predict the sales price given a set of housing attributes.

While hedonic price models have been routinely used to analyze the market price of housing, selecting an appropriate functional form has been a frequent concern in the literature (Cropper et al., 1988; Halvorsen and Pollakowski, 1981). The issue arises because there is little guidance from economic theory about the proper functional relationship between housing price and its attributes. Recognizing the potentially serious consequences of functional misspecification, earlier studies have attempted to estimate hedonic price models with flexible functional forms such as the transformation introduced by Box and Cox (1964). Despite its well-documented shortcomings (Davidson and MacKinnon, 1993; Wooldridge, 1992), the Box–Cox transformation has attracted considerable attention as it results in a better fit of the data and it can be used to test the statistical validity of alternative hypotheses about functional form (Rasmussen and Zuehlke, 1990).

More recently, a growing number of studies have applied non-parametric or semi-parametric regressions in estimating the hedonic price function. Among those are important contributions of Anglin and Gencay (1996), Gencay and Yang (1996), Pace (1998), Clapp et al. (2002). Closer inspection of this literature, however, reveals the possibility of methodological improvements that can add to both the ease of

[1] Other methods include repeat sales models that use the selling price of the same house at several points in time. For more discussions on the comparison of the two models, see Quigley (1995).

obtaining and interpreting hedonic price non-parametric functional estimates. With a few exceptions (Clapp et al., 2002; Pace, 1998), most applied non-parametric research specifies a regression class that requires the estimation of multivariate estimators.[2] This study estimates a hedonic price function using an additive semi-parametric regression based on the approach of Hastie and Tibshirani (1990). The central idea of this model is to replace the usual linear function of a covariate with an unspecified smooth function while holding the additive structure of linear regression models. This semi-parametric model is estimated by the iterative procedure known as the "backfitting algorithm," which reduces multivariate regression to successive simple bivariate regressions. The specification is free of restrictive parametric assumptions like any other non-parametric regressions, but unlike most, the effect of an individual attribute on the housing price can be easily interpreted due to its additive structure, regardless of the number of attributes. It requires only weak assumptions on the hedonic price functional form and directly estimates the association between the sales price and housing attributes. Since there is little information available on the proper functional form, this kind of generality is attractive in the hedonic price analysis.

The main objective of this study is to compare the prediction performance of the additive semi-parametric model with conventional parametric methods. Although the researchers have found that non-parametric or semi-parametric models can fit the data considerably better than the parametric counterparts, the out-of-sample prediction performance has received little attention. The out-of-sample predictions provide a better comparison between the parametric and semi-parametric models because they are non-nested. Limited studies (Clapp et al., 2002; Gencay and Yang, 1996) compared the out-of-sample predictions of a semi-parametric model with parametric alternatives and found that the semi-parametric models perform better than parametric counterparts in both in-sample and out-of-sample predictions. This study differs from the previous studies on the following methodological grounds. First, the current study specifies the hedonic price function as an additive model that avoids the problems of multivariate non-parametric regressions. Second, this study uses a local polynomial estimator that possesses a number of desirable theoretical and practical properties relative to the widely applied Nadaraya–Watson estimator. The advantages of the local polynomial estimator include its non-sensitivity to boundary data points (Fan et al., 1995; Ruppert and Wand, 1994). Third, another novelty of this study is on the estimation of the bandwidths via a plug-in method that minimizes the conditional mean average squared error. The plug-in methods are easy to compute and overcome the problem of undersmoothing that is the characteristic of the cross-validation methods (Opsomer and Ruppert, 1998).

This study utilizes two-year residential home sales data from Pitt County, North Carolina. The data were divided into in-sample and out-of-sample observations. The

[2] There are a number of practical as well as theoretical problems that emerge when estimating multivariate estimators. The well-known "curse of dimensionality" is identified by Friedman and Stuetzle (1981). From a practical perspective, multivariate estimators are difficult to compute and even with the use of sophisticated graphical analysis four or higher dimensional estimates are virtually impossible to represent or interpret.

in-sample data cover the period from July 2000 to June 2001. The out-of-sample data cover the time period from July 2001 to June 2002. In the in-sample price prediction comparisons, the root mean squared error (RMSE) of the semi-parametric model is 10.91 and 10.47% less than the semi-log model and the Box–Cox model. The mean absolute error (MAE) of the semi-parametric model is 9.97 and 9.44% less than the semi-log and the Box–Cox models, respectively. In the out-of-sample comparisons, the RMSE (MAE) of the semi-parametric model is 10.17% (11.51%) less than the semi-log model and 10.02% (11.27%) less than the Box–Cox model. With the methodological improvements described above, this study finds the superiority of the semi-parametric regression over the parametric models in house sales price predictions.

2. Study area and data

Pitt County is located in the coastal plain of eastern North Carolina. The Tar River, which goes through the middle of the County, flows into the Pamlico River and then into the Pamlico Sound. As one of the fastest growing areas in the state, the population increased by 23.3% between 1990 and 2000. According to the 2000 Census, the County has a population of 133,798 and the largest city, Greenville, has a population of 60,476. Recently, many new houses have been built due to the population growth and Hurricane Floyd that destroyed many homes with torrential rains and record flooding in September 1999. According to the Federal Emergency Management Agency (FEMA), Floyd directly affected over two million people and resulted in the largest peacetime evacuation in US history. The total number of housing units in Pitt County is 55,116, and of those housing units, a total of 50,018 are occupied.

The main data come from the Pitt County Management Information Systems database, representing a total of 2595 single-family residential homes sold between July 2000 and June 2002. The database contains extensive information on house sales transactions such as sales dates and price as well as square footage, number of bed/bath rooms, age of house, and other attributes. In addition, the data from the Pitt County Geographic Information Systems are incorporated to provide information on the important geographic locations including the Tar River, major roads and streets, business centers, and streams and creeks. This study uses the distances measured in feet from the centroid of the house to the nearest edge of these location attributes which may influence on housing sales price. All distances are measured in the Euclidean distance.

Given the recent major floods caused by Hurricane Fran in 1996 and Hurricane Floyd in 1999, whether a house is located in a floodplain or not is an important factor in the home purchase decision in eastern North Carolina. The large-scale damages caused by these hurricanes have increased public awareness of flood hazards. Government programs have also promoted both the awareness and the purchase of flood insurance. The FEMA reported that the sales of flood insurance policies increased by 24% in North Carolina after Hurricane Floyd (FEMA, 2002). Pitt

Table 1
Variables of the housing price index

Variable	Description
PRICE	House sales price in thousand dollars adjusted to a June 2002 level
GASHEAT	Dummy variable for gas heating (1 if gas heating, 0 otherwise)
FCBRICK	Dummy variable for face brick (1 if face brick, 0 otherwise)
FIREPLC	Dummy variable for fireplace (1 if fireplace, 0 otherwise)
HWFLOOR	Dummy variable for hard wood floor (1 if hard wood floor, 0 otherwise)
BEDRM1	Dummy variable for bedrooms (1 if 2 bedrooms or less, 0 otherwise)
BEDRM2	Dummy variable for bedrooms (1 if 3 bedrooms, 0 otherwise)
BEDRM3	Dummy variable for bedrooms (1 if 4 bedrooms or more, 0 otherwise)
BATHRM1	Dummy variable for bathrooms (1 if 2 bathrooms or less, 0 otherwise)
BATHRM2	Dummy variable for bathrooms (1 if 2 and a 1/2 bathrooms, 0 otherwise)
BATHRM3	Dummy variable for bathrooms (1 if 3 bathrooms or more, 0 otherwise)
QUALITY	Dummy variable for good quality (1 if good quality, 0 otherwise)
VACANT	Dummy variable for vacant house (1 if vacant house, 0 otherwise)
FLOOD	Dummy variable for house within floodplain (1 if floodplain, 0 otherwise)
TOTSQFT	Total structure square footage
AGE	Year house was built subtracted from 2002
STREAM	Distance in thousand feet to nearest creek or stream
CENTER	Distance in thousand feet to nearest business center
RIVER	Distance in thousand feet to the Tar River
TRAFFIC	Distance in thousand feet to major roads and streets

County government maintains a floodplain mapping database that contains the location and size of floodplains in the county. These floodplains are usually 100-year flood areas with a 1.0% chance of annual flooding. They often include the areas along the Tar River or streams with more significant chance of exposure to flooding.

Table 1 defines the variables used in this study and their definitions, and summary statistics are reported in Table 2. House sales prices are inflation-adjusted using a consumer price index to obtain figures in June 2002. Based on the 1397 homes sold between July 2000 and June 2001, the average selling price was $138,764 with a minimum sales price of $15,183 and a maximum of $722,018. Fig. 1 provides the nonparametric density estimation of the house sales price. Dummy variables are created for both bedrooms and bathrooms. About 6.5% of the total homes in the data are located in a floodplain. A typical home is about 19 years old and has 2350 square feet. About 46% of these homes have access to gas heating, and about 82% have a fireplace. The average distance to the nearest stream or creek is 841 feet and the average distance to the Tar River is 20,312 feet.

3. Empirical methods

Section 3 provides a brief discussion of the parametric and semi-parametric specifications for the hedonic price function and the estimation procedures. Let $\boldsymbol{X}$ represent a vector of 11 dichotomous characteristics of the house (e.g., gas heating source, hardwood floor, floodplain), and let $\boldsymbol{Z}$ represent a vector of six non-dichotomous

Table 2
Summary statistics of the variables

Variable	Mean	SD	Minimum	Maximum
PRICE	138.764	76.467	15.183	722.018
GASHEAT	0.461	0.499	0.000	1.000
FCBRICK	0.412	0.492	0.000	1.000
FIREPLC	0.820	0.385	0.000	1.000
HWFLOOR	0.233	0.423	0.000	1.000
BEDRM1	0.064	0.244	0.000	1.000
BEDRM2	0.742	0.438	0.000	1.000
BEDRM3	0.194	0.396	0.000	1.000
BATHRM1	0.650	0.477	0.000	1.000
BATHRM2	0.261	0.439	0.000	1.000
BATHRM3	0.089	0.285	0.000	1.000
QUALITY	0.039	0.195	0.000	1.000
VACANT	0.006	0.075	0.000	1.000
FLOOD	0.065	0.247	0.000	1.000
TOTSQFT	2343.310	975.789	681.000	8110.000
AGE	19.372	19.387	1.000	117.000
STREAM	0.841	0.612	0.001	4.249
CENTER	4.483	2.171	0.171	14.068
RIVER	20.312	16.730	0.202	90.751
TRAFFIC	0.150	0.120	0.012	1.115

Note. Summary statistics are based on the 1397 single-family home sales transactions occurred between July 2000 and June 2001.

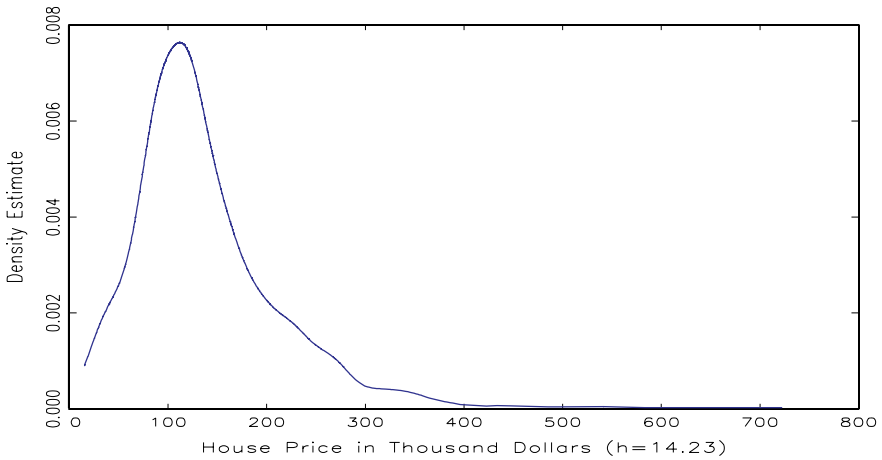

Fig. 1. Histogram of house sales price. *Notes.* In estimating the histogram, the bandwidth (h) is selected by Silverman's rule of thumb method. A Gaussian kernel function is used to assign weights for each observation.

characteristics (e.g., square footage, age, distance to river, and business center). The housing market is assumed to be in equilibrium, which requires that individuals optimize their housing choice based on the prices of alternative houses. Prices are

assumed to be market clearing, given the inventory of housing choices and their characteristics. Thus, the price of any house, P, can be described as a function of the housing characteristics:

$$P = P(\boldsymbol{X}, \boldsymbol{Z}). \tag{1}$$

Eq. (1) is referred to as the hedonic price function. With additional assumptions on individual's utility function, the estimation and partial differentiation of the hedonic price function with respect to a housing attribute reveal the marginal willingness to pay for that one attribute.[3] Furthermore, the estimation of the hedonic price function enables one to construct a house price index and to predict the sales price given a set of house characteristics.

As discussed before, selecting an appropriate functional form for Eq. (1) has been a frequent issue. Given that an incorrect choice of functional form may result in inconsistent estimates, earlier studies have attempted to estimate hedonic price models with more flexible functional forms. Most of these attempts have concentrated on parametric specifications such as the Box–Cox transformation, which includes several popular functional forms as special cases. In this study, the hedonic price function is modeled as follows:

$$E(\ln P|\boldsymbol{X}, \boldsymbol{Z}) = \alpha + \sum_{i=1}^{11} \beta_i X_i + \sum_{j=1}^{6} \beta_j Z_j^{(\lambda)}, \tag{2}$$

where $\ln P$ is the natural log of sales price, $Z_j^{(\lambda)} = ((Z_j^{\lambda} - 1)/\lambda)$ if $\lambda \neq 0$, and $Z_j^{(\lambda)} = \ln(Z_j)$ if $\lambda = 0$.

Eq. (2) is estimated using a maximum likelihood estimator. The Box-Cox transformation includes the semi-log ($\lambda = 1$) and the double-log ($\lambda = 0$) models as special cases depending on the transformation parameter λ. Only the non-dichotomous variables are subject to the transformation in order to keep the results comparable to the semi-parametric model. Likelihood ratio tests are used to compare the restricted forms with the more complex forms derived from the Box–Cox transformation.

This study also models the hedonic price function as a semi-parametric regression that is based on the approach by Hastie and Tibshirani (1990). This semi-parametric approach offers a middle ground that imposes less structure than a parametric approach but is tractable to estimate, unlike a completely general non-parametric approach. The model can be written as

$$E(\ln P|\boldsymbol{X}, \boldsymbol{Z}) = \alpha + \sum_{i=1}^{11} \beta_i X_i + \sum_{j=1}^{6} m_j(Z_j) \tag{3}$$

with $V(\ln P|\boldsymbol{X}, \boldsymbol{Z}) = \sigma^2$, an unknown parameter. Note that the usual linear function of $\boldsymbol{Z}$ is replaced with the sum of unspecified functions. The functions $m_j(Z_j)$ that appear in Eq. (3) are estimated using the iterative procedure known as the backfitting estimator, which reduces multivariate regression to successive simple regressions.[4]

[3] See Freeman (1993) for more discussions on the theoretical model.

[4] For further details of this estimation procedure, see Opsomer and Ruppert (1998).

The backfitting procedure starts with setting initial values for the unknown functions $m_j(Z_j)$ for $j = 1$–6 and then defines the partial residual of jth attribute for the vth iteration as

$$r_j^{(v)} = \ln P - \tilde{\alpha} - \sum_{i=1}^{11} \tilde{\beta}_i^{(v)} X_i - \sum_{d=1,d\neq j}^{j-1} \tilde{m}_d^{(v)}(Z_d) - \sum_{d=j+1,d\neq j}^{6} \tilde{m}_d^{(v-1)}(Z_d),$$

where $v = 1, 2, \ldots$, and $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{m}_d(Z_d)$ denote the estimated coefficients and estimated function. For the initial values, $\tilde{m}_d^{(0)}(Z_d)$ is defined as the $(n \times 1)$ vector of zeros. Each iteration completes when the six unknown functions are updated. Iterations continue until the change in the sum of squared residuals,

$$\sum_{t=1}^{n} \left(\ln P_t - \tilde{\alpha} - \sum_{i=1}^{11} \tilde{\beta}_i^{(v)} X_{ti} - \sum_{j=1}^{6} \tilde{m}_j^{(v)}(Z_{tj}) \right)^2 - \sum_{t=1}^{n} \left(\ln P_t - \tilde{\alpha} - \sum_{i=1}^{11} \tilde{\beta}_i^{(v-1)} X_{ti} - \sum_{j=1}^{6} \tilde{m}_j^{(v-1)}(Z_{tj}) \right)^2$$

is smaller than a pre-specified measure of tolerance between iterations.

During each iteration, the $\tilde{m}_j(Z_j)$ functions to be estimated are updated via the local polynomial regression that has the partial residual r_j as the dependent variable and the attribute Z_j as the independent variable for $j = 1$–6. The local polynomial estimator of p -degree for $\tilde{m}_j(Z_j)$ is defined as

$$\tilde{m}_j(Z_{tj}) = e_1'(Z_{tj}' W_{tj} Z_{tj})^{-1} Z_{tj}' W_{tj} r_j, \tag{4}$$

where e_1 is a $(p+1) \times 1$ vector having the value one in the first entry and zero elsewhere,

$$Z_{tj} = \begin{pmatrix} 1 & Z_{tj} - Z_{1j} & \cdots & (Z_{tj} - Z_{1j})^p \\ 1 & Z_{tj} - Z_{2j} & \cdots & (Z_{tj} - Z_{2j})^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{tj} - Z_{nj} & \cdots & (Z_{tj} - Z_{nj})^p \end{pmatrix},$$

W_{tj} is an n-dimensional diagonal matrix with elements given by $(1/h_j)K((Z_{tj} - Z_{sj})/h_j)$ for $s = 1, 2, \ldots, n$, K is the chosen kernel function, and h_j is a suitably chosen bandwidth.

A crucial aspect of any non-parametric estimation procedure is the selection of the bandwidths that underlie the calculation of $\tilde{m}_j(Z_j)$. The most commonly used procedure involves choosing bandwidths that minimize a jackknifed sum of squares or cross-validation function. Unfortunately, besides being extremely computationally intense, cross-validation has a tendency to undersmooth producing estimated regressions that have high variance. These undesirable characteristics have prompted the use of plug-in methods. This study uses a recent plug-in bandwidth selection method proposed by Opsomer and Ruppert (1998). The basic principle behind this plug-in method is the direct estimation of functionals that appear on the expressions

describing the optimal bandwidths. The bandwidths h_j are chosen to minimize the conditional mean average squared error (MASE):

$$\mathrm{MASE}(h_1,\ldots,h_6|Z_1,\ldots,Z_6)=\frac{1}{n}\sum_{t=1}^{n}E\left[\sum_{j=1}^{6}(\tilde{m}_j(Z_{tj})-m_j(Z_{tj}))|Z_1,\ldots,Z_6\right]^2. \tag{5}$$

Lastly, an estimated covariance matrix for each $\tilde{m}_j(Z_j)$ is obtained by $\hat{\sigma}^2R_jR_j'$ where

$$\hat{\sigma}^2=\frac{1}{n}\sum_{t=1}^{n}\left(\ln P_t-\tilde{\alpha}-\sum_{i=1}^{11}\tilde{\beta}_i^{(v)}X_{ti}-\sum_{j=1}^{6}\tilde{m}_j^{(v)}(Z_{tj})\right)^2,$$

and $\tilde{m}_j(Z_j)=R_j\ln P$. Then, the lower and upper bounds on the estimated regressions are constructed by using ± 2 times the square root of the diagonal of $\hat{\sigma}^2R_dR_d'$.

4. Estimation results

Table 3 reports the in-sample estimation results of the parametric models using the 1397 single-family residential houses sold between July 2000 and June 2001. Most slope coefficients have the same signs across the models and are statistically significant. The signs of the coefficients are consistent with the findings from previous empirical studies. The Box–Cox transformation parameter λ is estimated as 0.781. The likelihood ratio test statistics are calculated to test the semi-log and the double-log specifications. Given the critical value of 6.63, both semi-log and double-log functional forms are rejected at the 1% significance level. The data do not support the standard semi-log or double-log specification, and thus the Box–Cox transformed model is selected as a benchmark parametric model in the comparison with the semi-parametric regression fits.

The coefficient of the flood variable (FLOOD) has a negative sign and is statistically significant at the 1% level. The estimate from the Box–Cox model implies that locations within a floodplain have $9850 lower property values or a 7.1% reduction in the sales price evaluated at the sample mean. Several previous studies have found that the reduction in property values from the flood zones is equal to or greater than the capitalized value of flood insurance premiums (MacDonald et al., 1987; Shilling et al., 1985; Speyrer and Ragas, 1991). For an average-valued house ($125,000) in the study area, the estimated discount for the flood zones ($8875) is greater than the capitalized value of flood insurance premiums ($6880) when a 5% discount rate is used.[5] This finding is consistent with the notion that homebuyers are aware of flooding risks and that there may be substantial non-insurable costs including the hassle

[5] The flood insurance premium is based on the post flood insurance rate maps (FIRM) for single-family houses in flood zone A without a basement and with estimated base flood elevation of 3 feet or more. The content values of $30,000 are assumed. Deductibles for building and contents are assumed to be $500. The premium includes the federal policy fee of $80 and the increased cost of compliance (ICC) of $6.

Table 3
Estimation results of the parametric models

Variable	Semi-log ($\lambda = 1$)		Double-log ($\lambda = 0$)		Box–Cox	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Constant	3.944	0.041	−0.894	0.225	3.709	0.086
GASHEAT	0.035[b]	0.016	0.005	0.018	0.015	0.017
FCBRICK	0.078[a]	0.015	0.103[a]	0.017	0.084[a]	0.016
FIREPLC	0.274[a]	0.020	0.245[a]	0.022	0.260[a]	0.021
HWFLOOR	0.042[b]	0.019	−0.020	0.019	0.045[b]	0.019
BEDRM2	0.144[a]	0.030	0.119[a]	0.032	0.127[a]	0.030
BEDRM3	0.151[a]	0.036	0.096[b]	0.039	0.125[a]	0.037
BATHRM2	0.126[a]	0.019	0.140[a]	0.021	0.117[a]	0.019
BATHRM3	0.149[a]	0.033	0.235[a]	0.034	0.153[a]	0.033
QUALITY	0.034	0.039	0.115[a]	0.040	0.039	0.038
VACANT	−0.504[a]	0.088	−0.631[a]	0.094	−0.533[a]	0.088
FLOOD	−0.076[a]	0.028	−0.055[c]	0.029	−0.071[a]	0.027
$(TOTSQFT^{\lambda} - 1)/\lambda$	2.78e − 04[a]	1.15e − 05	0.723[a]	0.031	0.002[b]	0.001
$(AGE^{\lambda} - 1)/\lambda$	−0.010[a]	4.88e − 04	−0.134[a]	0.009	−0.021[a]	0.004
$(STREAM^{\lambda} - 1)/\lambda$	−0.038[a]	0.012	−0.022[a]	0.007	−0.035[a]	0.011
$(CENTER^{\lambda} - 1)/\lambda$	0.002	0.003	0.026[b]	0.013	0.004	0.004
$(RIVER^{\lambda} - 1)/\lambda$	−0.002[a]	4.06e − 04	−0.006	0.008	−0.004[a]	0.001
$(TRAFFIC^{\lambda} - 1)/\lambda$	−0.017	0.061	−0.004	0.016	−0.035	0.047
Sigma-sq (σ^2)	0.059[a]	0.002	0.066[a]	0.003	0.059[a]	0.002
Lambda (λ)					0.781[a]	0.062
Log-likelihood	−10.259		−87.704		−3.392	
Degrees of freedom					1378	
Likelihood ratio test statistic for semi-log functional form					13.73	
Likelihood ratio test statistic for double-log functional form					168.62	

Notes. Estimations are based on the 1397 single-family home sales transactions occurred between July 2000 and June 2001. Dependent variable is the log of sales price measured in thousand dollars. Superscripts a, b, and c denote significance at the 99, 95, and 90% levels, respectively. For proximity variables such as distance to nearest business center (CENTER) and distance to the Tar River (RIVER), a negative (positive) relationship to the dependent variable means that residents are willing to pay more (less) to live closer to the feature.

and deprivation of being displaced along with the loss of personal or family items with sentimental value.

The characteristics such as gas heating, a face brick, a fireplace, and hardwood floors have positive influences on house sales price. A four-bedroom house is sold for about $16,200 more than a two-bedroom house. Similarly, having additional bathrooms increases estimated sales price substantially. Older homes have lower property values. An additional year of age lowers the estimated sales price by $1200 evaluated at the mean value. Larger homes are more valuable. Evaluated at the average value of the houses, the results indicate that a house price increases by $40 per an additional square foot.

Table 3 also suggests that some locational variables have significant influence on house sales values. The results indicate that proximity to the nearest stream and the Tar River increases the house values. The proximity to the nearest business center

and a major road lowers house value, although the effects are not statistically significant. People like to live closer to water resources due to enhanced view quality or increased recreational opportunities. Moving 1000 feet closer to the Tar River raises estimated sales value by $220 evaluated at the sample mean. However, proximity to a business center or major roads may be undesirable due to increased traffic, congestion, and noise.

Table 4 provides the estimation results for the parametric part in the semi-parametric regression model. Note that the non-dichotomous variables are modeled into the non-parametric part and not reported here. All coefficients have the same signs with the parametric estimates in Table 3 and significant at the various levels. Magnitudes of the coefficient estimates are also quite comparable to the parametric models.

Fig. 2 displays the contribution to the housing sales price of total square footage, house age, and proximity to the locational attributes. The regression fits are presented along with their confidence intervals. The non-parametric estimates reveal that parametric functional forms might be inappropriate to approximate the complex price effects of some variables such as house age and proximity to the Tar River. The downward-slope fitted function for house age implies that the house value declines as a house gets older. The estimated regression is convex, indicating a stronger price effect for new homes. However, the negative effect dies out after the house is older than about 40 years. The estimated fit from the Box–Cox model is unable to show such relationship. The estimated function for the Tar River also clearly illustrates the advantage of using the semi-parametric regression model. While the parametric model indicates the positive effect of the proximity with a negative sign, the semi-parametric model captures a strong negative effect near the river. Tar River-adjacent houses are prone to flooding and may suffer from insect annoyances such as mosquitoes, and thus proximity to the river may decrease the house sales price for the initial distance of a mile or so. After this initial distance the proximity to the

Table 4
Estimation results of the semi-parametric model

Variable	Coefficient	SE	*t* Statistic
GASHEAT	0.038	0.014	2.720
FCBRICK	0.062	0.014	4.396
FIREPLC	0.165	0.018	8.928
HWFLOOR	0.034	0.017	2.024
BEDRM2	0.134	0.027	4.957
BEDRM3	0.125	0.033	3.791
BATHRM2	0.085	0.018	4.842
BATHRM3	0.190	0.030	6.331
QUALITY	0.117	0.035	3.342
VACANT	−0.515	0.080	−6.451
FLOOD	−0.044	0.025	−1.766
Degrees of freedom			1313

Notes. Estimation is based on the 1397 single-family home sales transactions occurred between July 2000 and June 2001. Dependent variable is the log of sales price measured in thousand dollars.

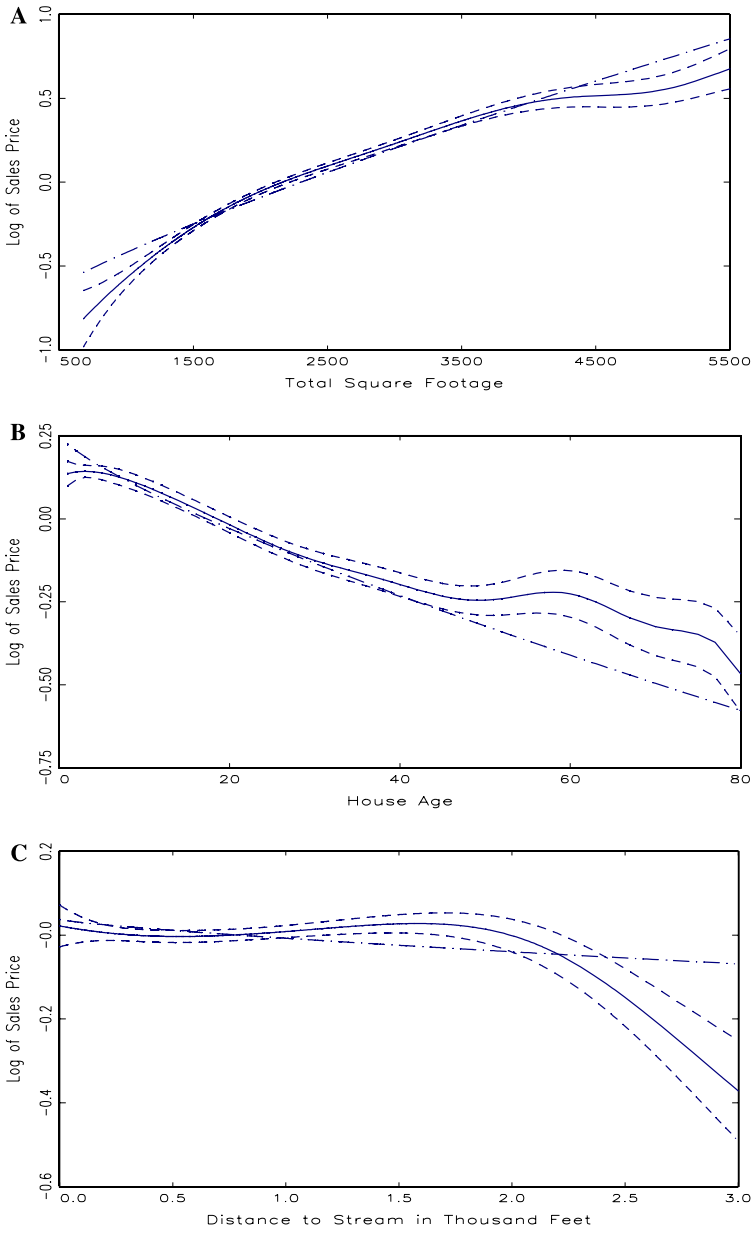

Fig. 2. Effects of housing attributes on house sales price. (A) Total square footage; (B) House age; (C) Distance to nearest stream; (D) Distance to nearest business center; (E) Distance to the Tar River; and (F) Distance to major streets. *Notes.* Figures are based on the 1397 single-family home sales transactions occurred between July 2000 and June 2001. The solid line represents the semi-parametric regression estimates. The dashed lines stand for the 95% semi-parametric confidence interval estimates. The dots and dashed line shows the parametric (Box–Cox) regression estimates.

Tar River seems to have a positive effect on the sales price because of the easy access to recreational activities along the river.

Table 5 provides the comparison of in-sample and out-of-sample price predictions for the semi-log, double-log, Box–Cox, and semi-parametric models. The semi-log

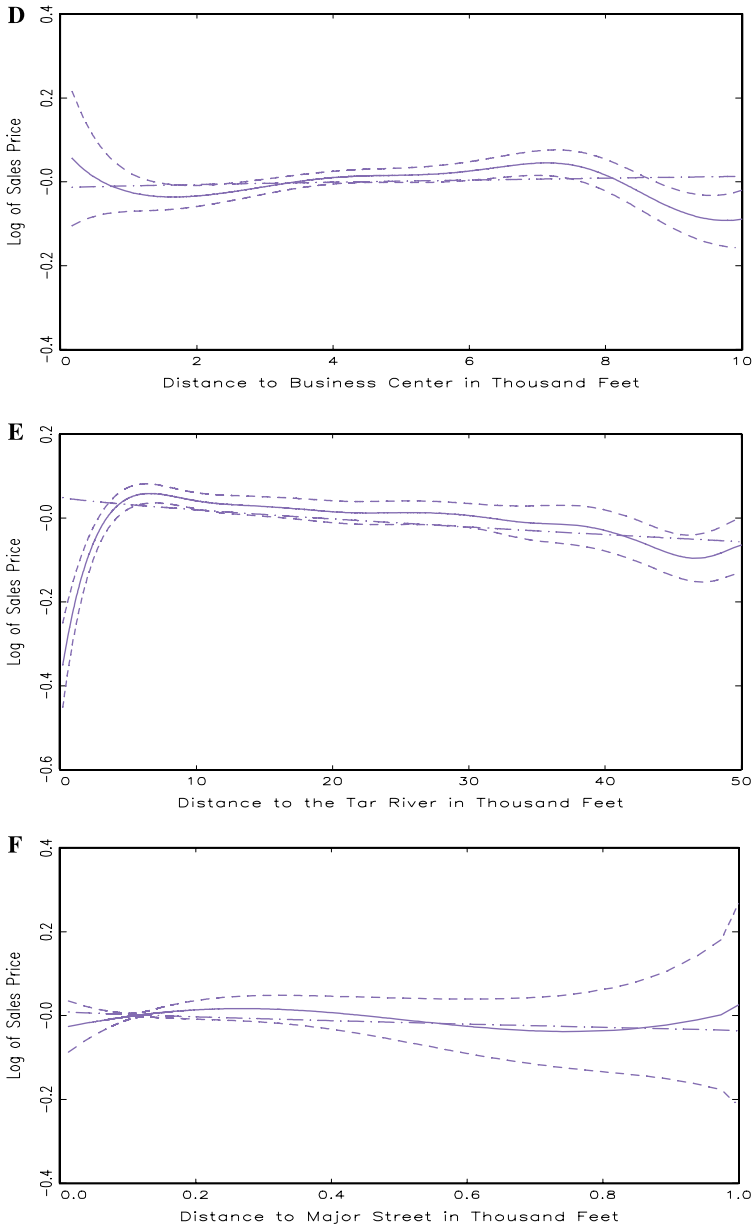

Fig. 2. (*continued*)

Table 5
In-sample and out-of-sample price prediction comparison for the semi-log (SL), double-log (DL), Box–Cox (BC), and semi-parametric (SP) regressions

	SL	DL	BC	SP (cross-validation)	SP (plug-in)
In-sample					
RMSE	0.2438	0.2576	0.2426	0.2152	0.2172
Difference (%)	10.91	15.71	10.47	−0.92	
MAE	0.1589	0.1713	0.1579	0.1420	0.1430
Difference (%)	9.97	16.52	9.44	−0.70	
Out-of-sample					
RMSE	0.2679	0.3001	0.2675	0.2397	0.2407
Difference (%)	10.17	19.79	10.02	−0.42	
MAE	0.1806	0.1994	0.1801	0.1594	0.1598
Difference (%)	11.51	19.87	11.27	−0.24	

Notes. RMSE stands for the root mean squared error and MAE stands for the mean absolute error. In-sample predictions are based on the 1397 single-family home sales transactions occurred between July 2000 and June 2001. Out-of-sample predictions are based on the 1198 single-family home sales transactions occurred between July 2001 and June 2002. Dependent variable is the log of sales price measured in thousand dollars.

and the double-log models, which can be estimated by the simple Ordinary Least Squares method, are common in practice of price predictions and thus compared to more flexible Box–Cox and semi-parametric models. This study uses two widely accepted measures of prediction accuracy of the root mean squared error (RMSE) and the mean absolute error (MAE).

The top of Table 5 shows the comparison of the in-sample prediction accuracy. The RMSE is 0.2438 for the semi-log, 0.2576 for the double-log, 0.2426 for the Box–Cox, and 0.2172 for the semi-parametric model (via the plug-in method). The semi-parametric model reduces the RMSE by 10.91% for the semi-log, 15.71% for the double-log, and 10.47% for the Box–Cox models. Similarly, the MAE of the semi-parametric model is 9.97, 16.52, and 9.44% less than the semi-log, the double-log, and the Box–Cox models, respectively. While the Box–Cox model performs better than the other parametric specifications, the semi-parametric model outperforms all the parametric models in the in-sample prediction comparison.

Table 5 also compares the prediction performance of the semi-parametric model via the plug-in method with the cross-validation estimator. Results indicate that the cross-validation estimator provides slightly smaller prediction errors than the plug-in estimator, but the differences are less than one percent. However, the plug-in estimator has shown much faster implementation, which can be an important issue with the readily available large data sets.[6] Recent hedonic studies have frequently used public records, which often include all houses in the city or county as the data source. Other than the practical advantages of the plug-in method such as computational

[6] With the 1397 observations the computation time was approximately 4 h 20 min for the plug-in method and 14 h 30 min for the cross-validation method on a 1.3 GHz Pentium IV PC. The difference between the two methods increases rapidly with the number of variables included in the non-parametric component.

efficiency, the two methods have shown quite comparable performances in price predictions for the data set used in this study.

Given the criticism that the better in-sample fit of the semi-parametric model might come at the cost of the degrees of freedom, it is useful to compare the degrees of freedom across the models. Although the degrees of freedom of the Box–Cox model (df = 1379) is larger than that of the semi-parametric model (df = 1313), the semi-parametric estimator does not seem to require an unreasonable amount of degrees of freedom. In fact, the degrees of freedom of the parametric model can be the same or even smaller than that of the semi-parametric model if the parametric model includes some interaction terms or uses a Taylor-series expansion of order two or three. The availability of large data sets would make the use of the semi-parametric regression more appealing.

The out-of-sample predictions are particularly important for comparison purposes, since the parametric and semi-parametric models are non-nested. The out-of-sample prediction evaluation is based on the 1198 single-family residential houses sold between July 2001 and June 2002. The new samples are denoted with the superscript N. For the parametric models, the in-sample parameter estimates from Table 3 are used to predict the sales price of the 1198 houses. The prediction errors are measured from these predicted sales prices ($\ln P^*$):

$$E(\ln P^* | \boldsymbol{X}^N, \boldsymbol{Z}^N) = \alpha + \sum_{i=1}^{11} \tilde{\beta}_i X_i^N + \sum_{j=1}^{6} \tilde{\beta}_j Z_j^{N(\lambda)}.$$

For the semi-parametric model, the in-sample parameter estimates from Table 4 and the non-parametric estimates with updated bandwidths are used to predict the sales price ($\ln P^*$):

$$E(\ln P^* | \boldsymbol{X}^N, \boldsymbol{Z}^N) = \alpha + \sum_{i=1}^{11} \tilde{\beta}_i X_i^N + \sum_{j=1}^{6} \tilde{m}_j(Z_j^N).$$

The bottom of Table 5 shows the comparison of out-of-sample price prediction accuracy. The RMSE is 0.2679 for the semi-log, 0.3001 for the double-log, 0.2675 for the Box–Cox, and 0.2407 for the semi-parametric models. The semi-parametric model reduces the RMSE by 10.17% for the semi-log, 19.79% for the double-log, and 10.02% for the Box–Cox models. Similarly, the MAE of the semi-parametric model is 11.51, 19.87, and 11.27% less than the semi-log, the double-log, and the Box–Cox models, respectively. In sum, the results reveal that the Box–Cox model performs better than the naive parametric models in house price predictions while the semi-parametric model outperforms the parametric alternatives.

5. Conclusions

This study estimates a hedonic price function using a semi-parametric regression and compares the price prediction performance with conventional parametric models. The results indicate that the semi-parametric model provides more accurate

housing price predictions than conventional parametric models in both in-sample and out-of-sample comparisons. The prediction errors from the semi-parametric model are smaller than those from the parametric models by roughly 10–20%. These results are consistent with the previous studies that claimed the superiority of the semi-parametric models in predicting house sales prices. The results indicate that the semi-parametric models would have great potentials in measuring and predicting residential housing prices.

References

Anglin, P., Gencay, R., 1996. Semi-parametric estimation of hedonic price functions. Journal of Applied Econometrics 11, 633–648.

Box, G., Cox, D., 1964. An analysis of transformations. Journal of the Royal Statistical Society B 26, 211–252.

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

Cropper, M., Deck, L., McConnell, K., 1988. On the choice of functional form for hedonic price functions. Review of Economics and Statistics 70, 668–675.

Davidson, R., MacKinnon, J., 1993. Estimation and Inference in Econometrics. Oxford University Press, Oxford.

Fan, J., Heckman, N., Wand, M., 1995. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Journal of the American Statistical Association 90, 141–150.

Federal Emergency Management Agency, 2002. After Floyd—North Carolina Progress.

Freeman, M., 1993. The Measurement of Environmental and Resource Values: Theory and Methods. Resources for the future, Washington, DC.

Friedman, J., Stuetzle, W., 1981. Projection pursuit regression. Journal of the American Statistical Association 76, 817–823.

Gencay, R., Yang, X., 1996. A prediction comparison of residential housing prices by parametric versus semi-parametric conditional mean estimators. Economics Letters 52, 129–135.

Goldberg, G., Harding, J., 2003. Investment characteristics of low- and moderate-income mortgage loans. Journal of Housing Economics 12, 151–180.

Halvorsen, R., Pollakowski, H., 1981. Choice of functional form for hedonic price equations. Journal of Urban Economics 10, 37–49.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, New York.

MacDonald, D., Murdoch, J., White, H., 1987. Uncertain hazards, insurance and consumer choice: evidence from housing markets. Land Economics 63, 361–371.

Mason, C., Quigley, J., 1996. Non-parametric hedonic housing prices. Housing Studies 11, 373–385.

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, K., 1998. Appraisal using generalized additive models. Journal of Real Estate Research 15, 77–99.

Palmquist, R., 1980. Alternative techniques for developing real estate price indexes. Review of Economics and Statistics 62, 442–448.

Quigley, J., 1995. A simple hybrid model for estimating real estate price indexes. Journal of Housing Economics 4, 1–12.

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

Rasmussen, D., Zuehlke, T., 1990. On the choice of functional form for hedonic price functions. Applied Economics 22, 431–438.

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

Shiller, R., 1993. Measuring asset value for cash settlement in derivative markets: hedonic repeated measures indices and perpetual futures. Journal of Finance 48, 911–931.

Shilling, J., Benjamin, J., Sirmans, C., 1985. Adjusting comparable sales for floodplain location. The Appraisal Journal, 429–436.

Speyrer, J., Ragas, W., 1991. Housing prices and flood risk: an examination using spline regression. Journal of Real Estate Finance and Economics 4, 395–407.

Wooldridge, J., 1992. Some alternatives to the Box–Cox regression model. International Economic Review 33, 935–955.