Nonparametric Assessment of the Effects of Neighborhood Land Uses on Residential House Values*

Abbreviated Title: Nonparametric Assessment of the Effects of Land Uses on House Values

Shigeru IWATA**, Department of Economics, University of Kansas, Lawrence, KS 66045
Phone: (785) 864-2867; Fax: (785) 864-5270; e-mail: shigeru@ukans.edu.
Hiroshi Murao, Aomori Public College, Aomori, Japan. Phone: (0177) 28-1111; Fax: (0177) 28-1110;
e-mail: murao@nebuta.ac.jp
Qiang Wang, Providian Bancorp, Inc., 201 Mission St., San Francisco. Phone: (415) 222-8819;
e-mail: john_wang@providian.com

Abstract

The effects of land uses on residential property values are crucial when evaluating costs and benefits of land projects for the purpose of public policy prescription or business decision making. It is widely recognized that a nuclear plant or a prison, for example, may often have an adverse effect on the property values of the nearby houses, while a park, a museum or a university usually has a beneficial effect. The effect of a land use defined as a function of distance between the locations of the land use factor and a particular house is, however, inherently nonlinear (in an unknown form) and the use of a simple linear regression method could lead to a misleading conclusion.

The purpose of this paper is to estimate the land use effect function by using the recently developed techniques of nonparametric regression method. There are three important features of our statistical model. First, it is a semiparametric model, which keeps a conventional linear form with respect to the dwelling attributes of the house just like in the popular hedonic model, but treats its location characteristics in a nonparametric fashion using the kernel method (Robinson 1988). Second, unlike the usual nonparametric regression, it keeps additive structure in the nonparametric component (Hastie and Tibshirani 1986, 1990), so that it retains much of the interpretative features of the linear models. Third, it uses the local linear smoother developed by Fan (1992, 1996), which is superior to other smoothers in terms of avoiding the boundary effect and other features.

We estimate the effects of three land use factors: (1) golf courses, (2) a university, and (3) a nitrogen plant on the neighborhood home values in Lawrence, Kansas. The data on the sales price and other attributes of the house with 6,400 observations over the period from 1986 to 1995 are obtained from the Douglas County Appraisal Office and the data on distance to the three sites above are constructed using the Geographic Information System (GIS)

---

* The authors would like to thank three anonymous referees for their helpful comments. The work of the first author was supported by University of Kansas, General Research allocation 3442-20-0038-1001. The earlier version of the paper was presented at the American Real Estate and Urban Economics Association annual meeting in New York in January 1999.

** Corresponding author.

Key Words: Land use, property value, semi-parametric regression, additive model, local linear estimator.
JEL Classification: R21, R31, C14.

1 Introduction

The e®ect of nearby land uses on residential property values has long been a popular topic among a variety of agents such as city designers, property tax collectors, housing developers, and possible house buyers as well as sellers. It is widely recognized that a nuclear plant or a prison, for example, may have an adverse e®ect on the property value of the nearby houses, while a park, a museum or a university usually has a bene¯cial e®ect. Assessment of such e®ects is essential for designing public projects, evaluating property tax of nearby houses, planning housing development and setting bid and ask prices of the houses in the market.

The impact of land uses on house prices often cannot be appropriately described by a simple linear function of distance. For instance, the houses located close enough to overlook a golf course entertain direct bene¯cial impact (wide open view, clean air, etc.) on their property values. This direct impact is expected to decline rapidly and becomes zero at a certain point, as distance gets large. This, however, is not the end of the story. The above group of houses generates a `good neighborhood,' which in turn has a bene¯cial e®ect on the houses located further away from the golf course. This secondary impact is expected to decline much more slowly as distance grows.

Estimating such a nonlinear e®ect of a land use factor is complicated by the need to control for many other factors including a variety of dwelling characteristics of the house such as the size of the house and the number of the bedrooms, etc., as well as other land use factors.

This paper introduces a partly linear and partly nonparametric regression procedure

that treats its nonparametric part in additive manner and therefore provides a convenient framework for analysis of this problem. A traditional approach to the house value assessment in economics is based on the hedonic price model (see e.g. Rosen, 1974), in which the value of a house is viewed as a sum of the values of its dwelling attributes. What lacks in this approach is the appropriate evaluation of the location characteristics of the house. The approach adopted by this paper retains a linear structure of the hedonic price model with respect to the dwelling characteristics of the house, while it models the location characteristics in nonparametric but additive fashion. In this way the model preserves an important interpretation feature of the linear model that would be lost with the usual nonparametric regression models. In particular, the nature of the e®ect of a variable on the response surface does not depends on the values of the other variables. Therefore, we can plot the function for each coordinate separately to examine the roles of the variables in predicting the response.

There are several studies which attempted to quantify the e®ect of land uses empirically. Nelson et al. (1992) estimate the e®ect of one Minnesota land¯ll on the values of 708 nearby homes located within 2 miles of the land¯ll. Do and Grudnitski (1995) estimate the e®ect on the values of 717 houses in San Diego when they are directly adjacent to golf courses. They both found evidence of the presence of the land use e®ects (see also Waddell et al. 1993). Their investigations, however, are restricted to the analysis based on conventional linear regression with dummy variables, which allows them to capture only qualitative aspects of the land use e®ect. Stock (1989, 1991), on the other hand, uses a semiparametric regression to estimate the e®ect of removing hazardous waste on house prices. McMillen and Thorsnes (1999) construct a house price index using a semiparametric regression (see also McMillen

1996). Other applications of nonparametric or semiparametric estimation techniques to the hedonic price model include Meese and Wallace (1991), Pace (1993, 1998), Goetzmann and Spiegel (199?), and Anglin and Gencay (1996).

This paper identi¯es the e®ect of land uses on the value of a particular house as an unrestricted function of distance to the land use factor, estimates this function by using recently developed techniques of nonparametric regression, and assesses the e®ects in detail. Speci¯cally, the goal of this paper is to make a nonparametric assessment of the e®ects of three land use factors: (1) golf courses, (2) a university (a major employment and education center of the city) and (3) a nitrogen plant (the main polluter), on the nearby home values in Lawrence, Kansas.

The organization of the rest of the paper is as follows: In Section 2, the semiparametric additive model is introduced and described for this application. Section 3 discusses the data, Section 4 explains the estimation procedure and Section 5 gives the results. A brief conclusion is given in Section 6.

2 The Semiparametric Additive Model

2.1 Semiparametric Model

To describe the semiparametric procedure, suppose that the i-th observation is given by a $(k + p + 1) \times 1$ vector $(y_i; x_i^{\prime}; z_i^{\prime}); i = 1; :::; n$, which is generated by the model

$$y_i = f(z_i) + g(x_i; \bar{}) + u_i \tag{1}$$

where $f(z)$ is an arbitrary function of z, while $g(x;\beta)$ is a known parametric function of x and a vector of unknown parameters β. The disturbance term u_i is assumed to satisfy

$$E(u_i|x_i,z_i) = 0 \tag{2}$$

and

$$E(u_iu_j|x_i,x_j,z_i,z_j) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The most popular functional form of $g(\cdot;\cdot)$ is linear, i.e.

$$g(x;\beta) = x'\beta. \tag{4}$$

In our application y_i stands for the natural log of the sale price of the house, x_i is a vector of the dwelling characteristics of the house, and z_i is a vector of the location characteristics including distance to the land use factors. Rather than considering an arbitrarily chosen parametric form, such as polynomials, for $f(\cdot)$, we estimate it by a nonparametric function.

There are by now a variety of techniques available for applied researchers to estimate a nonparametric regression model [see e.g. Härdel (1990), Eubank (1988)]. These techniques have much in common and may be referred to as `smoothers.' They are characterized in essence by local averaging, that is, averaging the y-values of observations having predictor values z close to a target value. Smoothers di®er mainly in their method of averaging. We restrict our attention to linear smoothers, i.e. smoothers that are linear in y. Examples of the linear smoothers include the Kernel, spline, and orthogonal series regression estimators [see Eubank (1988) and Härdel (1990) for a review of these smoothers].

Now let P denote the projection matrix $X(X'X)^{-1}X'$ and S be a linear smoothing operator, where X is an $n \times k$ matrix with its i-th row equal to x_i'. For example, for the Nadaraya-

Watson kernel regression estimator, the $(i; j)$ element of S is given by $S_{ij} = \frac{K(z_i \, ¡ \, z_j)}{\sum_j K(z_i \, ¡ \, z_j)}$ for $i; j = 1; :::n$; where $K(¢)$ is a kernel function, which generates the weights with a maximum at zero and satis¯es certain moment conditions. Write equation (1) in matrix form

$$y = f + g + u \tag{5}$$

where $y; f; g;$ and u are $n £ 1$ vectors with the i-th element equal to $y_i; f(z_i); g(x_i; ¯)$ and u_i. Assuming (2), (3), and (4), the estimators of f and g are given by

$$\begin{aligned} \hat{f} &= S(y ¡ \hat{g}) = S(y ¡ X\hat{¯}) \\ \hat{g} &= P(y ¡ \hat{f}) = X(X'X)^{-1}X'(y ¡ \hat{f}) = X\hat{¯} \end{aligned}$$

Combining the above two equations, we obtain (Green, Jennison, and Seheult, 1985)

$$\hat{¯} = [X'(I ¡ S)X]^{-1}X'(I ¡ S)y \tag{6}$$

$$\hat{f} = S(y ¡ X\hat{¯}): \tag{7}$$

Under some regularity conditions on the bandwidth parameter, $\hat{¯}$ and $\hat{f}$ are consistent estimators of $¯$ and $f = E(y ¡ X¯|Z)$, respectively. In particular, it is well known that the convergence rate of $\hat{¯}$ as $n \to 1$ is the same as in the parametric case [see e.g. Robinson(1988) for this property and its limit distribution; see Bierens (1987) for the limit distribution of the kernel estimator).

2.2 Additive Model

We now make a further assumption that nonparametric part f in model (1) or (5) takes an additive form, i.e.

$$f(z_i) = ® + \sum_{j=1}^{X} f_j(z_{ji}) \tag{8}$$

where z_{ji} is the j-th components of vector z_i. An underlying assumption here is that di®erent land use factors do not have interactive e®ects on the house values. A simple test for the validity of this assumption will be discussed in the later section.

There are three distinctive merits for the above speci¯cation [see Hastie and Tibshirani (1989, 1990) for more detail]. First, it can avoid what is called the `curse of dimensionality' problem that plagues standard non- or semi-parametric methods, i.e. far large sample size is required to obtain a reasonable estimate of f with high dimensions of z. The additive model speci¯cation converts a high dimension problem into that of a single dimension, thereby getting around this problem.

Second, it provides a simple interpretation. Economists tend to ask ceteris paribus questions: What is the impact of z_j (the j-th variable of z) on the left hand variable if all other variables are kept unchanged? The additive model gives an immediate answer to such questions. Because the impact of z_j on y does not depend on all other z_k's ($k \neq j$), we can plot the p-coordinate functions separately to examine the roles of the variables in predicting y.

Third, it makes computation easy. With the help of the `back¯tting algorithm' developed by Friedman and Stuetzle (1981), estimation of the additive model becomes especially attractive.

We refer to the model consisting of (5) and (8) as the \Semi-parametric Additive Model", which is a version of the \Generalized Additive Model" introduced originally by Hastie and Tibshirani (1986). Under this speci¯cation the property value of a house is explained by a linear combination of conventional dwelling characteristics plus a sum of the unrestricted functions of distance to each factor.

2.3 Smoother

Among linear smoothers most popular are probably kernel smoothers and spline smoothers. We adopt a special type of kernel smoother in our application. The reason for choosing it is simply the availability of relatively more complete theoretical results than other smoothers. In fact, many linear smoothers can be expressed as kernel smoothers.

It is sometimes useful to view kernel smoothers as local polynomial ¯ts. Consider a small neighborhood $N(z)$ of z and let $z_i \in N(z)$ for $i = 1; :::; n$. Then, assuming the existence of the q-th derivative of $f(z)$, we have a Taylor series approximation

$$f(z_i) \approx b_0 + b_1(z_i - z) + \cdots + b_q(z_i - z)^q;$$

where $b_j = \partial^j f(z)/\partial z^j$ for $j = 1; \cdots; q$. The problem of estimating $f(z)$ is therefore equivalent to that of estimating b_0 in a local polynomial regression that minimizes

$$\sum_{i=1}^{n} [y_i - b_0 - b_1(z_i - z) - \cdots - b_q(z_i - z)^q]^2 K\left(\frac{z_i - z}{h}\right);$$

where $K(\cdot)$ stands for a kernel function and h is the bandwidth parameter. When $q = 0$; the solution $\hat{f}(z)$ is known as the Nadaraya-Watson (NW) estimator:

$$\hat{f}_{NW}(z) = \sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right) y_i \Big/ \sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right);$$

which is probably the most popular kernel regression estimator. It, however, often su®ers a large bias. When $q = 1$, the estimator $\hat{f}(z)$ is referred to as the local linear (LL) estimator [Fan (1992), Fan and Gijbels (1996)] and is given by

$$\hat{f}_{LL}(z) = \sum_{i=1}^{n} [\hat{b}_2(z) - \hat{b}_1(z)(z_i - z)] K\left(\frac{z_i - z}{h}\right) y_i \Big/ \sum_{i=1}^{n} [\hat{b}_2(z) - \hat{b}_1(z)(z_i - z)] K\left(\frac{z_i - z}{h}\right)$$

where $s_k(z) = \sum_{i=1}^{n}(z_i - z)^k K\left(\frac{z_i - z}{h}\right)$ for $k = 1; 2$: The LL estimator simply provides a higher order approximation than the NW estimator. There are two advantages of using this estimator. First, it has high asymptotic e±ciency among all possible linear smoothers [Fan(1992)]. Second, it does not require any modi¯cations at boundary points, where the usual kernel estimator su®ers a large bias. Its disadvantage is its computational cost (n^2 iterations are required, compared to n iterations in NW case).

3 Data

Our data are on the residential houses in Lawrence, Kansas (see Figure 1 for a map and Appendix A for a brief description).

FIGURE 1 ABOUT HERE

Table 1 summarizes our data on dwelling attributes (`x-variables') and location characteristics (`z-variables'), which are two fundamental determinants of house price (`y-variable').

TABLE 1 ABOUT HERE

Dwelling attributes are a set of variables that describe the characteristics of a house. Age of a house, total square footage, and number of rooms are typical examples of dwelling attributes. Location characteristics measure the characteristics of a real estate site, such as proximity to various land uses or facilities. Four location characteristics: golf course, university, nitrogen plant, and site elevation were evaluated. All three golf courses in Lawrence are located in western part of the city. Houses around the golf courses are usually priced higher than the

city average. The university, located at the center of the city, is the primary employment center; students and university employees accounted for approximately one third of the city population. The nitrogen plant, at the east edge of the city, is the major industrial establishment and is perceived as a pollution source. Land elevation is an indicator of °ood potential. Part of the residents can be victimized by rainstorms in a normal year, let alone the °ood of 1993. This last location variable was apparently not explicitly evaluated in previous research.

House sale data were obtained from Douglas County Appraisal O±ce and site investigations. The data include the dwelling attributes and sale values of each residential transaction in Douglas County from January 1986 to May 1995. Douglas County TIGER (composed in 1990) ¯le was used as the basic coverage. Lawrence was clipped out as the study area. Address matching was accomplished in PC Arc/Info to identify the house locations. Boundaries of golf courses, university and nitrogen plant were digitized. GIS 'overlay' and 'near' were then used to determine the distance of each house to a land use boundary [see Arono® (1989)]. Eventually, 6,415 residential sales were available for our analysis with a set of variables necessary for modeling. The dependent variable is the logarithm of the sales price of each property at the date of sale. Independent variables, summarized in Table 1 in the Appendix, include multiple measures of dwelling attributes and four location characteristics: golf course, university, nitrogen plant, and elevation. Dwelling characteristics include the log of the living area, log of the lot area, log of the ¯nished living area, the age of the house in order to estimate an age depreciation e®ect, and dummy variables for the number of full and half baths, the number of bedrooms, the story height, the presence and type of ¯replace, heating and cooling. The main city is on the south side of Kansas River; a dummy variable

for whether a house falls in this side of the river is included to measure an expected positive transportation convenience.

Three important variables, crime rate, school quality, and jurisdictional tax rates, are not included. The omissions are mainly due to the consideration that Lawrence is relatively homogeneous in these factors. We expect that the absence of these Tiebout variables would not diminish our results.

One limitation of our data is that the distances to the land use factors were recorded only for the houses located within the preselected maximum distance from the factors: 1,000 meters for the golf courses, 4,000 meters for the university, and 8,000 meters for the nitrogen plant, respectively (those are 43.4%, 99.8%, and 79.1% of the total observations). Those numbers were selected because, on the basis of the preliminary examination of a small subset of the data, the price e®ect of a land use factor appears to be negligible for those located beyond that distance. For each of such observations, a random distance larger than the maximum was assigned; namely, a randomly generated number between 1,000 and 6,000 for the golf courses, 4,000 and 5,000 for the university, and 8,000 and 9,000 for the nitrogen plant, respectively. [In fact, the impacts of the factors were empirically found dying out within the distance of 50{60% of the above maximum to the factors: about 600 meters, 2 kilometers, and 5 kilometers for the golf courses, the university, and the nitrogen plant, respectively (see Figure 3 A{C).]

4 Estimation and Computation

The model we shall estimate is given by

$$y_{it} = \alpha_t + x_{it}^{0}\beta + \sum_{j=1}^{p} f_j(z_{jit}) + u_{it} \tag{9}$$

for $i = 1; : : : ; n$; and $t = 1; : : : ; T$, where y_{it} is the value of the house, x_{it} is a vector of the dwelling characteristics of the house (such as the size of the house, the number of bedrooms and etc.), z_{jit} is the j-th location characteristics including distance to the land use factors, and u_{it} is an unknown disturbance term.

The parametric part $x_{it}^{0}\beta$ corresponds to the conventional hedonic price model, while $\sum f_j(z_{jit})$ represents the nonparametric part. The individual function $f_j(\cdot)$ is not restricted to any functional form except that it is smooth, but the whole nonparametric part is restricted to be additive. A most popular computational approach (among economists) to such a model appears to be a two-step estimation procedure (e.g. Robinson 1988). In this procedure, the ¯rst step consists of the usual nonparametric regressions of y on z, and x on z, respectively, and the second step is the ordinary least squares (OLS) regression of the residuals of the former regression on those of the latter to obtain the `semiparametric estimate' of $\hat{\beta}$ of β. To obtain the estimate of f, we run the nonparametric regression of $y - x\hat{\beta}$ on z.

The actual computation goes as follows: After estimating $E(y_i|z_i)$ and $E(x_i|z_i)$ with any usual nonparametric regression techniques, we calculate the `residuals' $e_i^y = y_i - \hat{E}(y_i|z_i)$ and $e_i^x = x_i - \hat{E}(x_i|z_i)$. The estimator $\hat{\beta}$ in (6) may be obtained by regressing e_i^y on e_i^x with the ordinary least squares. Our goal is then to estimate $\hat{p} = \hat{E}(e_i|z_i)$ with constraint (8), where $e_i = y_i - x_i^{0}\hat{\beta}$. To this end, we follow the steps below:

1. We ¯rst approximate each f_j by a step function and use it as a starting value for the iteration.

   (a) Let c_{jh} denote the appropriately chosen equally spaced points in the support of z_j for $h = 1; ::::; H_j$ and $j = 1; ::::; p$ with $c_{j0} = 0$. De¯ne the location dummy $d_{jh} = 1(z_j \in (c_{j;h-1}; c_{jh}])$, which takes the value unity when z_j falls into the h-th interval.

   (b) Regress e_i linearly on d_{jhi}'s with constant and obtain the ¯tted value of e_i given by $\hat{e}_i = \hat{\mu}_0 + \sum_{j=1}^{p} \sum_{h=1}^{H_j} \hat{\mu}_{jh} d_{jhi}$. Write the ¯tted regression in the ¯rst step as $y_i = \hat{\theta} + x_i'\hat{\gamma} + \sum_j d_{ji}'\hat{\mu}_j + \hat{u}_i \equiv \hat{\theta} + x_i'\hat{\gamma} + \hat{e}_i^0$, where d_{ji} is a $H_j \times 1$ vector of j-th location dummies and $\hat{\theta}$, $\hat{\gamma}$ and $\hat{\mu}_j$ are OLS estimates.

2. We now apply the iterative procedure known as the back¯tting algorithm, which is explained brie°y as follows

   (a) Set the initial values $s_j^0(z_{ji}) = \sum_{h=1}^{H_j} \hat{\mu}_{jh} d_{jh}$ for $j = 1; ::::; p$. This is the step for $m = 0$.

   (b) The m-th iteration is described as follows: Starting with $j = 1$, set

$$r_{ji}^m = y_i - \hat{\theta} - x_i'\hat{b}_{j-1}^m - \sum_{l=1}^{j-1} s_l^m(z_{li}) - \sum_{l=j+1}^{p} s_l^{m-1}(z_{li})$$

   where $\hat{b}_j^0 = \hat{\gamma}$ for $j = 1; : : : ; p$, and estimate

$$s_j^m(z_{ji}) = \hat{E}(r_{ji} | z_{ji})$$

   with some nonparametric regression techniques.

   (c) Compute $\hat{b}_j^m = \left[\sum_{i=1} x_i x_i'\right]^{-1} \sum_{i=1} x_i \left[y_i - \sum_{l=1}^{j} s_l^m(z_{li}) - \sum_{l=j+1}^{p} s_l^{m-1}(z_{li})\right]$.

   (d) Repeat this step from $j = 1$ to p, and repeat the entire p-steps until $ESS \equiv \sum_{i=1}^{n} [y_i - s_0 - \sum_{j=1}^{p} s_j^m(z_{ji})]^2$ fails to decrease.

Clearly the error sum of squares (ESS) does not increase at any step of the algorithm and therefore converges. Breiman and Friedman (1985) show that $\sum_{j=1}^{p} s_j^1(z_{ji})$ is unique and provides the best additive approximation to the nonparametric part or $f = E(y_i - x_i^{\prime}\beta | z_i)$ in our model. This does not mean, however, that the individual functions $f_j(\cdot)$'s are uniquely estimated. Buja et al. (1989) show that if the smoothers S_j are symmetric with eigenvalues in $[0; 1)$ such as the cubic spline smoother, the normal equations corresponding to the algorithm are consistent for every y.

The essential idea of this algorithm is to reduce computation of multiple regression to that of successive simple regressions. We use a local linear estimator described in subsection 2.3 with the Epanechnikov kernel[1] given by $K(u) = 0.75(1 - u^2)I(|u| \leq 1)$. The cross-validation functions (see the Appendix B) are computed to guide the selection of the bandwidth. The SAS code for all procedures used in this paper is available at http://falcon.cc.ukans.edu/~econ/workpaper/wp99.htm.

5 Empirical Results

We ¯rst show the result of the parametric linear model and then report nonparametric estimation results.

5.1 Parametric Estimates

The left columns of Table 2 reports the estimates based on the linear regression with dummy variables for equally spaced intervals of z_j's.

TABLE 2 ABOUT HERE

[1] The Epanechnikov kernel is frequently used because of certain optimality properties, such as minimizing mean integrated squared error.

The results suggests that the most important three dwelling attributes of the house (in terms of statistical signi¯cance) are the size of the total living area (LVG_AREA), the size of the land (LAND), and the age of the house (AGE). The age has a nonlinear e®ect on the house value, which declines at a decreasing rate as the house gets older. Other important characteristics are the number of full bathrooms (FL_BATH), whether the basement is at least half-¯nished (BASEMT), whether it has central heating system (HEAT) and or air conditioner (HEAT_AIR), and the number of openings of brick ¯replaces (BRICK). The number of bedrooms (BED_RMS) is totally insigni¯cant, and the number of family rooms (FAM_RMS) has a negative e®ect. These results might sound odd but it is not unreasonable that when LVG_AREA variable is included in the regressors, BED_RMS variable does not have any additional explanatory power. Also, given LVG_AREA, an increase in FAM_RMS could lower the house value, because many family rooms often imply that the house is for rent to multifamily tenants, in particular to a group of students, and usually such a house is likely to be of low quality. The coe±cients of the year dummies measure the time e®ect of each year from 1986 to 1994 with 1995 as the base year. The estimates show the upward movement from 1986 to 1988, followed by the downward movement from 1989 to 1991, which is followed again by the upward movement from 1992 to 1995. This trend roughly coincides with the local business cycle of the real estate market in the Midwest during the period. In any case, most of the results above are quite standard in empirical studies of the hedonic price of a house.

5.2 Semiparametric Estimates

5.2.1 Bandwidth Selection

We now estimate equation (9) according to the procedure outlined in section 4. The ¯rst important step is the choice of the bandwidth, ¸, which determines how much to smooth. There is a well-known trade-o® in kernel estimation between the bias and variance of the estimate: The use of a small ¸ reduces the bias but generates a large variance, while choosing a large value of ¸ will reduce the variance at the expense of introducing bias into the estimation. Although there are many approaches to this problem [see, for example, Eubank (1988), Härdel (1990) and Silverman (1986)], we use the cross validation procedure, which is described in the Appendix B. The idea of this procedure is to compute the squared prediction errors for each selected value of ¸ by sample reuse techniques. We compute the cross-validation function separately for four location characteristics: (i) distance to the nearest golf course (z_1), (ii) distance to the university (z_2), (iii) distance to the nitrogen plant (z_3), and (iv) land elevation (z_4), for a range of values of each $¸_j$ and plot them in Figure 2(A)-(D).

FIGURE 2 ABOUT HERE

These ¯gures provide data driven criteria for bandwidth selection. In addition to this automatic criteria, we also have our own subjective criteria. Although we do not have prior information about the exact shape of each regression curve, our natural expectation is summarized as follows: (i) All curves are reasonably smooth, (ii) $f_1(z_1)$ and $f_2(z_2)$ are monotone decreasing functions of $z_1$and z_2, respectively, (iii) $f_3(z_3)$ is a monotone increasing function of z_3, (iv) $f_4(z_4)$ is expected to be strictly increasing at least up to some point and then

possibly may go down.

Each bandwidth is then selected by combining the automatic criteria with our subjective criteria. More speci¯cally, starting from the bottom point of each cross-validation curve in Figure 2, we search for the value of $¸_j$ (j = 1;¢¢¢4) closest to the bottom value that satis¯es four subjective criteria described above. Each of the selected bandwidth for $f_j(z_j)$ is marked along the respective cross-validation curves in Figures 2(A)-(D). As we can see, the bandwidths selected for all but the golf-course e®ect exceed that suggested by the mechanical cross-validation procedure. These sizes of smoothing are needed for removing spurious noise such as the e®ect of arterial streets.

The nonparametrically estimated e®ects of three land use factors and elevation on the property value of the house are displayed in Figures 3(A) through (D).

FIGURE 3 ABOUT HERE

For comparison purpose, a polynomial function based on the cubic curve ¯t and a step function estimated using dummy variables are superimposed in each ¯gure. The step function, although not smooth, can capture rough shapes of the true curve. As we can observe in Figures 3(A){(D), the kernel estimates smooth out the discontinuity of the step function without losing important local features. The polynomial estimates, on the other hand, tend to mask the local details of the curve in favor of the overall ¯t.

The vertical axis of each ¯gure measures the proportional change in house values. For example, according to the kernel estimates, the house located 100 meters away from the nearest golf course is expected to value 6 percentage point higher than the comparable house

located 200 meters away ($\hat{f}_1(100) - \hat{f}_1(200) = 0.06$). The 95% pseudo con¯dence bands of each kernel regression curve is shown in Figures 4(A)-(D).

FIGURE 4 ABOUT HERE

Strictly speaking, these bands are not exactly the con¯dence bands because the function estimates $\hat{f}(z_j)$ are asymptotically biased.

5.3 Nonparametric Assessment

In the following we summarize the assessment of the e®ects of neighborhood land uses on the residential house values on the basis of nonparametric regression curves.

5.3.1 Golf course e®ect and University e®ect

All four curves look very reasonable. The positive e®ects of the golf courses and the university are quite sizable; the house directly adjacent to one of the golf courses values more than 20% higher than the comparable house 600 meters or more away from it, while the house directly adjacent to the university values more than 40% higher than the comparable house located 2,000 meters or more away from it. As the distance to the golf course or the university gets large, such e®ects initially declines rapidly and then in a more moderate pace. The parametric and nonparametric regression curves are quite close in both cases.

In the golf course case, about 22% price premium at 0 meter declines to 12% at 100 meters and to just 6% at 200 meters. This sharp decline stops around 200 meters, which is followed by a more gradual decline. In particular, between 300 and 450 meters, the curve has a plateau of 2.5% price premium. In the university case, the initial 42% price premium

declines to 28% at 200 meters and to 15% at 400 meters. In this range the curve is close to linear with a slope equal to minus 6.5% per 100 meters. Then the decline becomes much gradual; 8% price premium at 500 meters and 3% at 1,000 meters | the slope in this range is minus 1% per 100 meters.

The main source of the price e®ect of the golf course appears to be the direct and physical bene¯ts that a house can entertain from the golf course, namely those such as big open space, attractive view, and fresh air. The upper limit of the distance that allows such bene¯ts seems to be about 200 meters. The small but positive price premium in the range of 300 to 450 meters, therefore, should have a di®erent source. It is commonly observed that the houses located close to the golf courses tend to form a good neighborhood. The latter, in turn, has a positive e®ect on the price of the houses located near but not so close to the golf courses. The reason for this kind of `good neighborhood e®ect' to disappear around 500-600 meters is most likely to be a wide street which cuts the otherwise continuous residential area and prevents the e®ect to continue to spread.

The price premium for the houses near the university appears to have a di®erent source. It is true that the direct physical bene¯ts similar to those described in the golf course case also apply to the houses near the university. However, this explains only a fraction of total price premium for the university. The major source of the price e®ect appears to be simply the length of time required to travel to the university. The university is the center of activities for 27,000 students as well as an employment center of many workers. A reason for the university e®ect to disappear around 1,800 meters seems to be that the walking distance of twenty minutes is likely to be the maximum for the usual person to choose to commute on

foot.

5.3.2 Nitrogen-plant e®ect

The negative e®ect of the nitrogen plant, a pollution source of the city, is sizable as well as wide spread. The house located 1,500 meters from the plant values 17% lower than the comparable house located 6,000 meters or more from it. The negative price premium decreases almost constantly by 2% per 500 meters from minus 11% at 2,000 meters to minus 4% at 4,000 meters and then the pace slows down. Although the thick smoke produced by the plant is sometimes visible from the center of the city, its negative externality is not likely to be entirely based on the measured pollution level.

5.3.3 Elevation e®ect

As seen in Figure 3(D), the discrepancy between parametric and semiparametric estimates is largest for the elevation e®ect. The parametric estimates show a large variation of the e®ect; the house at the lowest elevation values 16% lower than the house at the highest elevation, whereas the semiparametric estimates suggest much smaller variation, just 6% di®erence. A reason for this discrepancy appears to be that our subjective criteria make us to choose a rather wide bandwidth, which °attens the curve.

5.4 Stability of the E®ects

So far we have implicitly assumed that the price e®ects of the four location characteristics are time invariant. To check their stability, the model was estimated separately for 1986, 1990, and 1994 (the number of observations are 868, 595, and 655 for 1986, 1990, and 1994, respectively). The results are displayed in Figures 5(A)-(D).

FIGURE 5 ABOUT HERE

The most stable among the four characteristics is clearly the university e®ect, for which there appears to be no structural change over time between 1986 and 1994. Interesting cases are the e®ects of the golf course and nitrogen plant. The golf course e®ect looks relatively stable over time in the range of distance between 200 and 600 meters, whereas there seems to be a large negative shift in the e®ect on the houses within 200 meters from the golf course. The house adjacent to a golf course values 24% higher in 1986 but only 16% higher in 1994. This seems to re°ect a speci¯c situation on the supply side in that a new area next to one of the golf courses had been rapidly developed for housing during 1990 and 1995. The e®ect of the nitrogen plant exhibits a little di®erent pattern of change over time. There is an upward shift in the estimated regression curve from 1984 to 1990, while the curve for 1994 is not much di®erent from that for 1990. The negative e®ect of the nitrogen plant appears to have decreased during this period. A possible explanation is that the plant became less pollutant.

5.5 Speci¯cation Tests and Prediction Performance

Two speci¯cation tests for checking the validity of the underlying assumptions of our Semiparametric Additive Model given by (9) are conducted.[2] The ¯rst test examines the additivity assumption (8), whereas the second one tests the semiparametric model against the linear model.

A simple diagnostic test to check the additivity assumption, proposed by Hastie and Tibshirani (1990), is to regress the residuals from the semiparametric regression (9) on the

[2] Another potentially important misspeci¯cation is the omission of any major centers (land uses), which would cause spatial autocorrelation of the errors. To test this possibility, a standard test (such as Moran I) for spatial autocorrelation can be used. Such a test, however, is not conducted in this paper.

interaction terms of estimated f_j's and examine the signi¯cance of each. More speci¯cally, we run the regression

$$\hat{\varepsilon}_i = \sum_{\substack{j,k=1 \\ j<k}}^{J} \gamma_{jk} \hat{f}_{ji} \hat{f}_{ki} + e_i$$

where $\hat{\varepsilon}_i = y_i - \hat{\alpha} - x_i'\hat{\beta} - \sum \hat{f}_{ji}$ is the residual of semiparametric regression (9), $\hat{f}_{ji} = \hat{f}(z_{ji})$, and γ_{jk} is the unknown parameter. If the coe±cient γ_{jk} were found to be signi¯cantly di®erent from zero from conventional standard, we would suspect that location characteristics z_j and z_k have enough interaction to prevent the additive speci¯cation (8). A preferred model in such a case would be

$$f(z_i) = \alpha + \sum_{h \neq j,k} f_h(z_{hi}) + f_{jk}(z_{ji}, z_{ki})$$

rather than (8), where $f_{jk}(\cdot)$ is an unknown function of z_j and z_k. The result of the regression of the residuals on six interaction terms using 1993 data is reported in Table 3.

TABLE 3 ABOUT HERE

None of the terms is found signi¯cant at the 1 % level, providing support for the additivity assumption (8).

Next, to test the linear speci¯cation (H_0) against the semiparametric speci¯cation, we conduct a simple Wu-Hausman test, which is described in Robinson (1988). The test is based on the contrast between the OLS estimate $\bar{\beta}$ and the semiparametric estimate $\hat{\beta}$. If H_0 is true, that is, the linear speci¯cation is correct, then $\bar{\beta}$ is consistent and more e±cient than $\hat{\beta}$, while $\hat{\beta}$ is consistent whether or not H_0 is true. Although our primary interest is

in estimation of f rather than ¯, this test provides a simple and convenient way to examine the validity of the semiparametric speci¯cation. Since $\hat{\beta}$ is $\sqrt{n}$ consistent, the test statistic has the usual limit distribution and the test may be conducted in the manner same as in the parametric case. The test statistic is given by

$$\chi = (\hat{\beta} - \tilde{\beta})' \left[\mathrm{var}(\hat{\beta}) - \mathrm{var}(\tilde{\beta}) \right]^{-1} (\hat{\beta} - \tilde{\beta})$$

which, under H_0, has Chi-Squared distribution with degrees of freedom equal to the dimension of ¯. The right hand columns of Table 2 presents the semiparametric estimates of the coe±cients of the dwelling attributes (x-variables). The Wu- Hausman statistic computed from the 1993 data is 240.47. Since the 99 % quantile of the Chi-Squared distribution with 27 degrees of freedom is 46.96, the consistency of the OLS regression estimates is rejected, providing support for our semiparametric model.

We may also compare the semiparametric regression with the OLS regression in terms of prediction performance. For this purpose, we use 1993 observations $(y_j; x_j; z_j)$ to predict 1994 prices of houses, y_i^*, given their dwelling attributes, x_i^*, as well as location characteristics, z_i^*. The detailed prediction procedure is described in Appendix C. After obtaining $\hat{y}_i^*$'s, the measures of prediction accuracy are computed for a subset of 1994 houses that are located near either one of golf courses, the university, or the nitrogen plant (more speci¯cally, $z_1 < 1{,}000$, $z_2 < 4{,}000$ and $z_3 < 8{,}000$). The total number of such houses is 158. The results are reported in Table 4.

TABLE 4 ABOUT HERE

As shown in the table, the prediction performance of the semiparametric regression is slightly better than the OLS regression.

6 Conclusion

In this paper we estimated the e®ects of three land use factors and elevation on the residential house values without assuming any parametric restrictions on the functional forms of the distances. Our use of semi-parametric additive model with a local linear smoother enabled us to reveal salient features of the price e®ect curves of the golf courses, the university, the nitrogen plant, and the elevation, which are consistent with our natural expectations. Our procedure can be applied to a broad range of similar studies.

Appendix A

City of Lawrence

Lawrence is located in the northeast corner of the State of Kansas, about 40 miles west of Kansas City and about 30 miles east of Topeka, the state capital. Lawrence is a university town with the University of Kansas being the center of the city. The city has a population of about 75,000 and the university has more than 27,000 students. Its geological size is about 5 to 6 miles in each direction (see also Figure 1).

Appendix B

Cross-Validation

The problem of deciding how much to smooth is of great importance in nonparametric regression. The choice of a bandwidth or smoothing parameter has frequently a more important impact on the shape of the regression curve than the choice of a kernel has. To guide our bandwidth selection, we compute the cross-validation function

$$CV(\lambda_1;\cdots;\lambda_p) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - x_i'\hat{\beta}_{(i)} - \sum_{j=1}^{p}\hat{f}^{(i)}_{\lambda_j;j}(z_{ji})\right]^2$$

for an appropriate range of values of λ_j, $j = 1;\cdots;p$, where the estimate of the j-th function is given by

$$\hat{f}^{(i)}_{\lambda_j;j}(z_{ji}) = \sum_{k\neq i;k=1}^{n} K\left(\frac{z_{ji} - z_{jk}}{\lambda_j}\right) r_{ji} = \sum_{k\neq i;k=1}^{n} K\left(\frac{z_{ji} - z_{jk}}{\lambda_j}\right);$$

$$\hat{\beta}_{(i)} = \left[X_{(i)}'X_{(i)}\right]^{-1} X_{(i)}'y_{(i)} - \sum_{j=1}^{p}\hat{f}^{(i)}_{j}(z_j);$$

$X_{(i)}$ and $y_{(i)}$ stand for X and y with i-th rows are removed.

In practice, 1,164 observations were randomly chosen out of 6,415 total observations and this subset was used for the cross-validation. The bandwidth selection depends on the sample size and it must be rescaled for the whole data set.

Appendix C

Prediction Procedure

Denote 1993 observations by $(y_i; x_i^{\prime}; z_i^{\prime})$ for $i = 1; : : : ; n$, and 1994 observations by $(y_i^{\alpha}; x_i^{\alpha\prime}; z_i^{\alpha\prime})$ for $i = 1; : : : ; n^{\alpha}$. Let $r_i = y_i - \hat{\alpha} - x_i^{\prime}\hat{\beta}$ where $\hat{\alpha}$ and $\hat{\beta}$ are the semiparametric regression estimates based on 1993 data. Now compute successively

$$\hat{f}_k(z_{ji}^{\alpha}) = \frac{\sum_{h=1}^{n} \left[\hat{\beta}_{j2}(z_{ji}^{\alpha}) - \hat{\beta}_{j1}(z_{ji}^{\alpha})\, z_{jh} - z_{ji}^{\alpha}\right] K\left[\frac{(z_{jh} - z_{ji}^{\alpha})}{\delta_j}\right] r_h^j}{\sum_{h=1}^{n} \left[\hat{\beta}_{j2}(z_{ji}^{\alpha}) - \hat{\beta}_{j1}(z_{ji}^{\alpha})\, z_{jh} - z_{ji}^{\alpha}\right] K\left[\frac{(z_{jh} - z_{ji}^{\alpha})}{\delta_j}\right]}$$

$$r_h^j = y_h - \hat{\alpha} - x_h^{\prime}\hat{\beta} - \sum_{k=0}^{j-1} \hat{f}_k(z_{kh})$$

from $j = 1$ to p, where $\hat{\beta}_{jk}(z) = \sum_{i=1}^{n} (z_{ji} - z)^k K\left[\frac{z_{ji} - z}{\delta_j}\right]$ for $k = 1; 2$, and $\hat{f}_0(z_{kh}) \equiv 0$.

The predicted value of y_i^{α} based on 1993 data is given by

$$\hat{y}_i^{\alpha} = \hat{\alpha} + x_i^{\alpha\prime}\hat{\beta} + \sum_{j=1}^{p} \hat{f}_j(z_{ji}^{\alpha})$$

and the root mean squared error is given by

$$RMSE = \left[(1/n^{\alpha}) \sum_{i=1}^{n^{\alpha}} (\hat{y}_i^{\alpha} - y_i^{\alpha})^2\right]^{1/2}.$$

References

[1] Arono®, S. (1989) Geographic Information System: A Management Perspective. Ottawa, Canada: WDL Publications.

[2] Anglin, P.M., and R. Gencay (1996) \Semiparametric Estimation of a Hedonic Price Function," Journal of Applied Econometrics 11, 633{648.

[3] Bierens, H.J. (1987) \Kernel Estimators of Regression Functions," in T.F. Bewley(ed.), Advances in Econometrics: Fifth World Congress, Cambridge: Cambridge University Press, 99{144.

[4] Breiman, L., and J. H. Friedman (1985) \Estimating Optimal Transformations for Multiple Regression and Correlation," Journal of the American Statistical Association 80, 580{619.

[5] Buja, A., T. J. Hastie, and R. Tibshirani (1989) \Linear Smoothers and Additive Models," (with discussion), Annals of Statistics 17, 453{555.

[6] Do, A. Q., and G. Grudnitski (1995) \Golf Courses and Residential House Prices: An Empirical Examination," Journal of Real Estate Finance and Economics 10, 261{270.

[7] Engle, R.F., Granger, C.W.J., Rice, J., and A. Weiss (1986) \Semiparametric Estimates of the Relation Between Weather and Electricity Sales," Journal of the American Statistical Association 81, 310{320.

[8] Eubank, R. L. (1988), Spline Smoothing and Nonparametric Regression, New York: Marcel Dekker.

[9] Fan, J. (1992) \Design-Adaptive Nonparametric Regression," Journal of the American Statistical Association 87, 998{1004.

[10] Fan, J., and I. Gijbels (1996) Local Polynomial Modeling and Its Applications, London: Chapman & Hall.

[11] Friedman, J.H., and W. Stuetzle (1981) \Projection Pursuit Regression," Journal of the American Statistical Association 76, 817{823.

[12] Green, P., C. Jennison, and A. Seheult (1985) \Analysis of Field Experiments by Least Squares Smoothing," Journal of Royal Statistical Society B 47, 299{315.

[13] Härdel,W. (1990), Applied Nonparametric Regression, Cambridge: Cambridge University Press.

[14] Hastie, T., and R. Tibshirani (1986) \Generalized Additive Models" (with discussion), Statistical Science 1, 297{318

[15] Hastie, T., and R. Tibshirani (1990) Generalized Additive Models. Chapman-Hall

[16] McMillen, D. P. (1996) \One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach," Journal of Urban Economics 40, 100-124.

[17] McMillen, D. P., and P. Thorsnes (1999) \The Reaction of Housing Prices to Information on Superfund Sites: A Semiparametric Analysis of the Tacoma, Washington Market," this volume.

[18] Nelson, A., J. Genereux, and M. Genereux (1992) \Price E®ects of Land¯lls on House Values," Land Economics 68, 359{365.

[19] Robinson, P. M. (1988) \Root-N-Consistent Semiparametric Regression," Econometrica 56, 931{954.

[20] Rosen, S. (1974) \Hedonic Prices and Implicit Markets: Product Di®erentiation in Pure Competition," Journal of Political Economy 82, 34{55.

[21] Stock, J.H. (1989) \Nonparametric Policy Analysis," Journal of the American Statistical Association 84, 1461{1481.

[22] Stock, J. H. (1991) "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Bene¯ts," in W. Barnett, J. Powell and G. Tauchen (eds), Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, Cambridge University Press, New York.

[23] Waddell, P., Berry, B.J.L., and I. Hoch (1993) \Residential Property Values in a Multinodal Urban Area: New Evidence on the Implicit Price of Location," Journal of Real Estate Finance and Economics 8, 117{141.

Table 1: List of Variables

Type of Variable	Variable Name	Description	MEAN	STD	MIN	MAX
y-variable	HOUSE_VALUE	The sale price of the house in 1983 dollars (log)	10.88	0.46	8.2 1	12.59
x-variables	AGE	The number of years after the house was built	24.87	25.87	0	135
	AGE_SQ	Square of AGE	1287.68	2437.39	0	18225
	REMODL	A dummy variable for remodeling (1 if the house is remodeled)	0.08	0.27	0	1
	AGERM	The number of years after the house was remodeled	0.75	3.98	0	61
	AGERM_SQ	Square of AGERM	16.38	136.71	0	3721
	LVG_AREA	The size of the total living area (log of square feet)	7.33	0.39	5.95	8.70
	LAND	The size of the primary site (log of square feet)	9.10	0.47	6.74	12.21
	BED_RMS	The number of bedrooms	3.08	0.74	1	7
	FAM_RMS	The number of family rooms	0.36	0.49	0	2
	FL_BATH	The number of full bathrooms (any three fixtures)	1.67	0.66	0	4
	HLF_BATH	The number of half bathrooms (any two fixtures)	0.39	0.52	0	3
	HEIGHT	A dummy variable for story height (1 if the story height is 20 feet or higher)	0.15	0.35	0	1
	BASEMT	A dummy variable for basement (1 if the basement is either half- or full- finished)	0.62	0.49	0	1
	HEAT	A dummy variable for central heating (1 if the house is equipped with central heating)	0.13	0.34	0	1
	HEAT_AIR	A dummy variable for central heating and air conditioning (1 if the house is equipped with both)	0.84	0.37	0	1
	BRICK	The number of openings of brick fireplaces	0.42	0.63	0	4
	METAL	The number of openings of metal fireplaces	0.25	0.45	0	3
	NORTH	A dummy variable for north (1 if the house is in North Lawrence)	0.04	0.20	0	1
	YEAR86-94	Dummy variables for each year 1986 to 1994				
z-variables	GOLF	Distance from the house to the nearest golf course (meters)	2176.69	1862.29	0.03	5997
	UNIV	Distance from the house to the university (meters)	2290.68	934.57	42	4991
	NITRO	Distance from the house to the nitrogen plant (meters)	5954.43	2125.33	1504	9000
	ELEV	Elevation of the house site (meters)	277.01	19.78	247	322

Table 2: Parameter Estimates of the Linear Model with Dummy Variables vs. the Semiparametric Model

Variable	Parametric Regression			Semiparametric Regression		
	Estimate	Standard Error	t-value	Estimate	Standard Error	t-value
Constant	6.23735	0.09033	69.045			
AGE	-0.00905	0.00041	-22.275	-0.00724	0.00039	-18.686
AGE_SQ	0.00005	0.00000	15.090	0.00004	0.00000	12.026
REMODL	0.07403	0.01292	5.729	0.07706	0.01275	6.043
AGERM	-0.00462	0.00183	-2.532	-0.00491	0.00181	-2.713
AGERM_SQ	0.00007	0.00004	1.581	0.00008	0.00004	1.833
LVG_AREA	0.44103	0.01206	36.573	0.42861	0.01183	36.233
LAND	0.13607	0.00579	23.516	0.13502	0.00558	24.186
BED_RMS	-0.00263	0.00427	-0.062	-0.00083	0.00417	-0.199
FAM_RMS	-0.03812	0.00585	-6.513	-0.02079	0.00580	-3.581
FL_BATH	0.07138	0.00590	12.104	0.06713	0.00580	11.576
HLF_BATH	0.03666	0.00572	6.414	0.03337	0.00559	5.972
HEIGHT	0.06824	0.00774	8.815	0.05844	0.00761	7.683
BASEMT	0.09533	0.00591	16.127	0.08273	0.00571	14.498
HEAT	0.21144	0.01513	13.978	0.22396	0.01492	15.015
HEAT_AIR	0.28632	0.01524	18.784	0.31337	0.01504	20.830
BRICK	0.06884	0.00533	12.915	0.07103	0.00520	13.668
METAL	0.02519	0.00683	3.686	0.02006	0.00664	3.021
NORTH	-0.17328	0.01958	-8.849	-0.15891	0.01740	-9.134
YEAR86	-0.11726	0.01503	-7.802	-0.09985	0.01490	-6.702
YEAR87	-0.07820	0.01511	-5.173	-0.06265	0.01498	-4.182
YEAR88	-0.05507	0.01512	-3.636	-0.04291	0.01498	-2.864
YEAR89	-0.05519	0.01554	-3.552	-0.04808	0.01540	-3.123
YEAR90	-0.07758	0.01552	-4.999	-0.07645	0.01536	-4.978
YEAR91	-0.10881	0.01557	-6.986	-0.10392	0.01545	-6.726
YEAR92	-0.11395	0.01529	-7.487	-0.10821	0.01508	-7.173
YEAR93	-0.08632	0.01519	-5.681	-0.08050	0.01505	-5.347
YEAR94	-0.04366	0.01526	-2.860	-0.03890	0.01512	-2.574

Table 3: Regression of Residuals on Interaction Terms

Variable	Estimate	Standard Error	t-value
FGFK	-0.676729	1.98228160	-0.341
FGFN	1.719703	3.95649979	0.435
FGFE	-11.854903	10.04750721	-1.180
FKFN	1.870161	4.24298318	0.441
FKKE	4.361601	10.84225800	0.402
FNFE	-4.346213	10.89307381	-0.399
R^2	0.0033		

Notes:

1. The regression does not include an intercept term, and R^2 is computed in uncorrected form.
2. FGFK stands for the interaction term between the golf-course effect and the university effect, i.e., $FGFK = \hat{f}_1 \cdot \hat{f}_2$. Similarly, $FGFN = \hat{f}_1 \cdot \hat{f}_3$, $FGFE = \hat{f}_1 \cdot \hat{f}_4$, $FKFN = \hat{f}_2 \cdot \hat{f}_3$, $FKFE = \hat{f}_2 \cdot \hat{f}_4$, $FNFE = \hat{f}_3 \cdot \hat{f}_4$.
3. To compute the above regression, 1993 year data are used. The sample size is 693

Table 4: Prediction Performance of Parametric vs. Semiparametric Regressions

Measures of Prediction Accuracy	Parametric regression	Semiparametric regression
Root Mean Square Error	0.1959051	0.1904084
Root Mean Square Percentage Error	0.0174042	0.0168488
Theil's U Statistic	0.0177319	0.0172344

Notes:

1. Both linear and semiparametric regressions were first run using 1993 data with 693 observations and then those estimates were used for predicting 1994 house sales prices. To compute the measures of prediction accuracy, only 158 observation points of 1994 data that are located near the land use factors were utilized.
2. Measures of prediction accuracy are defined as follows:

   (A) Root mean square error (RMSE)

$$RMSE = \sqrt{(1/n)\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$$

   (B) Root mean square percentage error (RMSPE)

$$RMSPE = \sqrt{(1/n)\sum_{t=1}^{n}\left[(\hat{y}_t - y_t)/y_t\right]^2}$$

   (C) Theil's U statistic

$$U_1 = \sqrt{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 / \sum_{t=1}^{n} y_t^2}$$

Figure 1: City Map

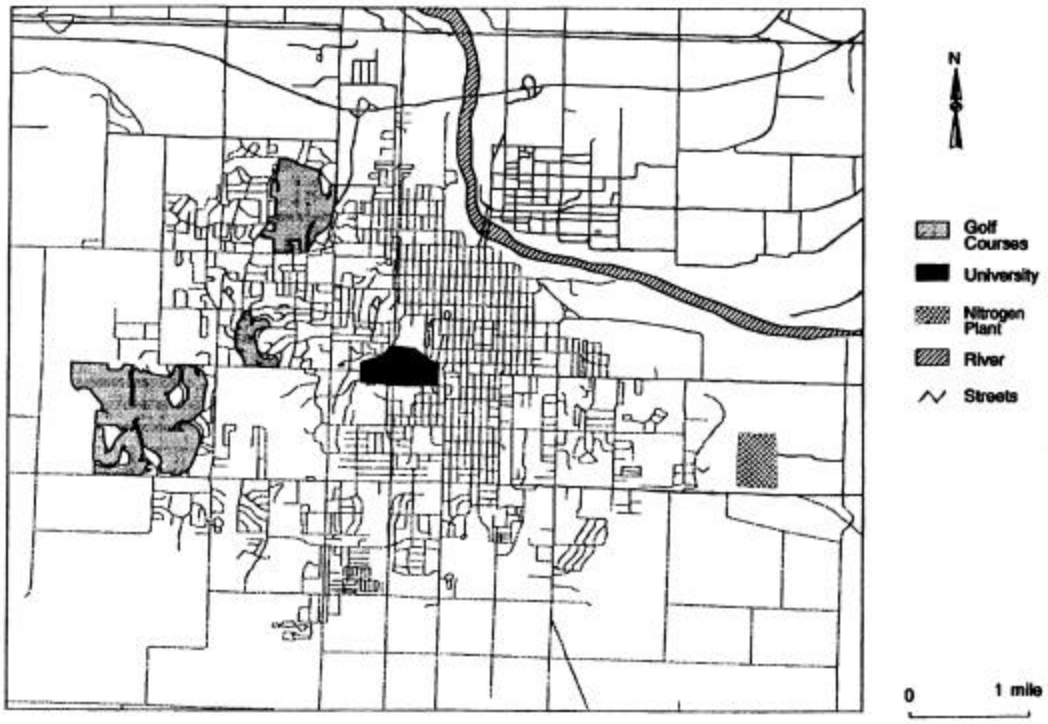
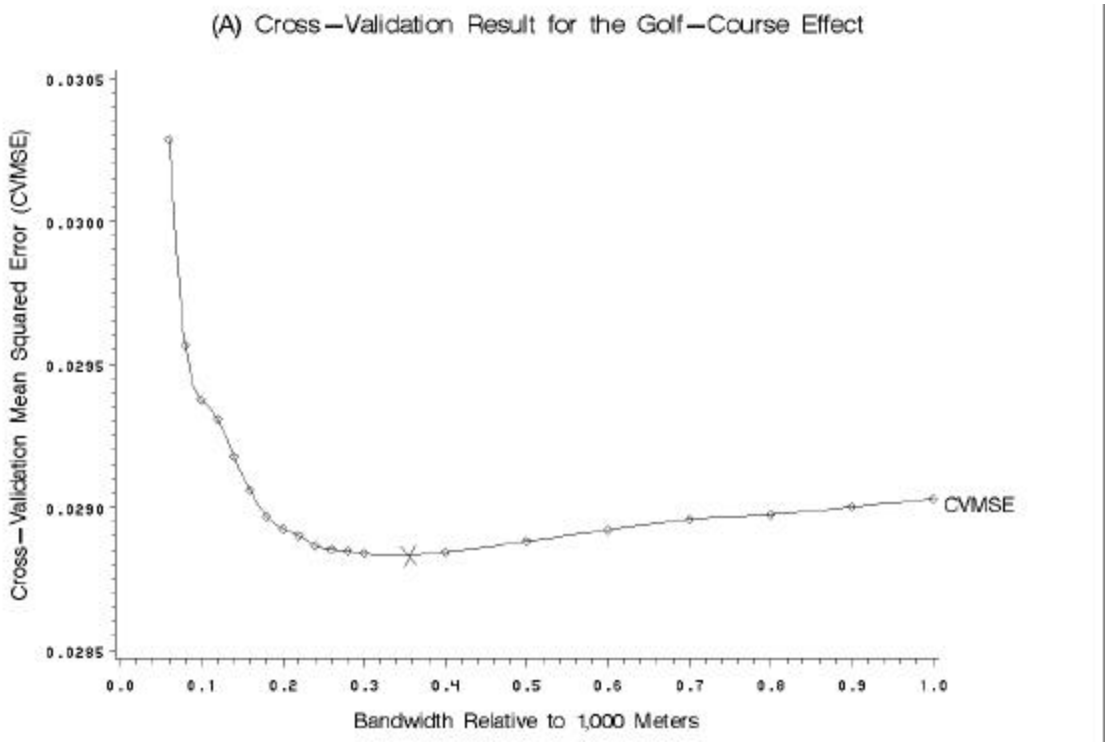
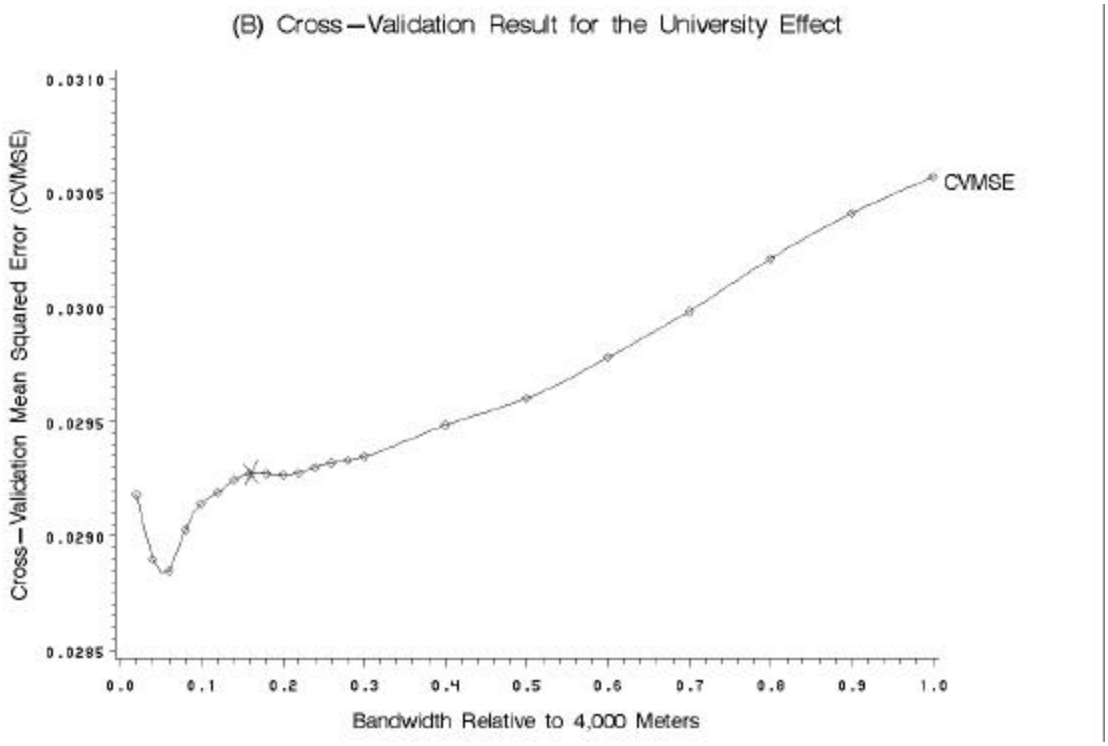
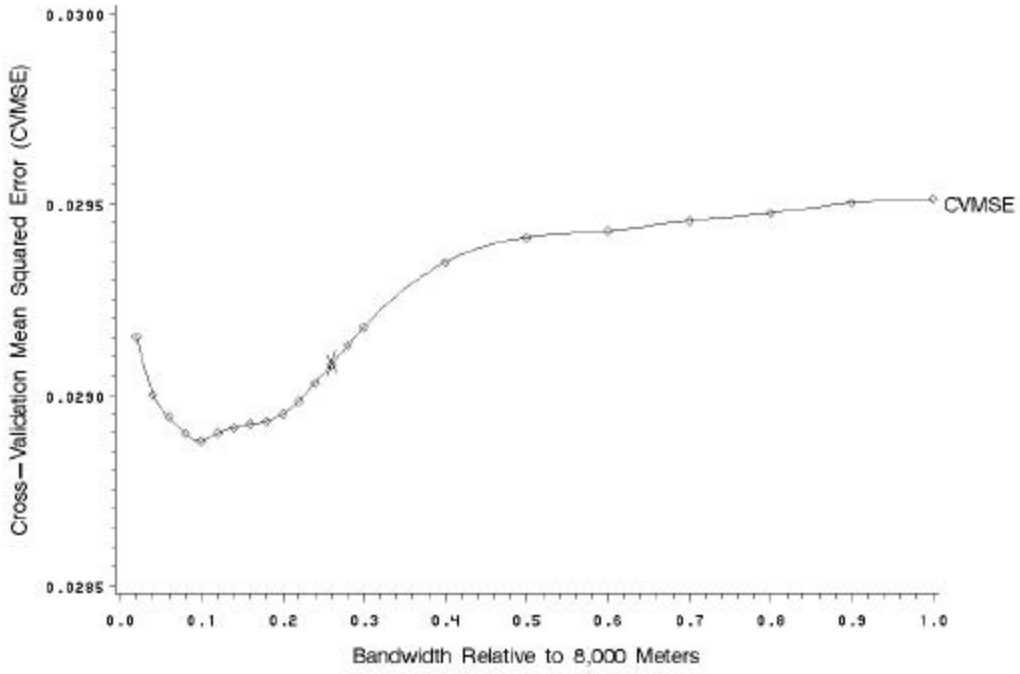


Figure 2: Cross-Validation Curves





(C) Cross–Validation Result for the Nitrogen–Plant Effect



(D) Cross–Validation Result for the Elevation Effect

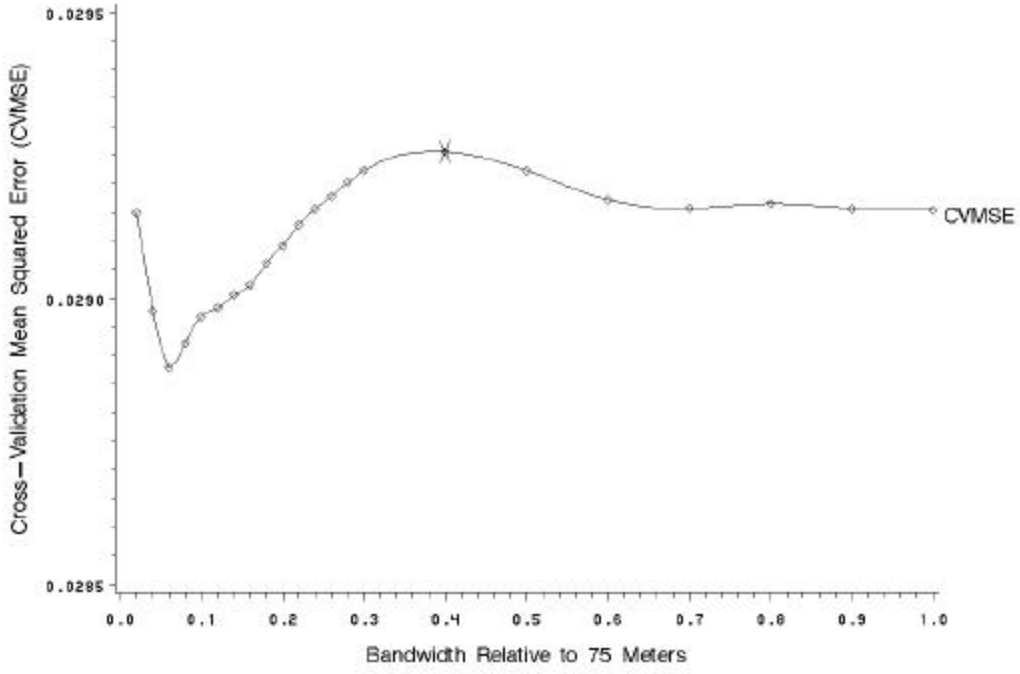
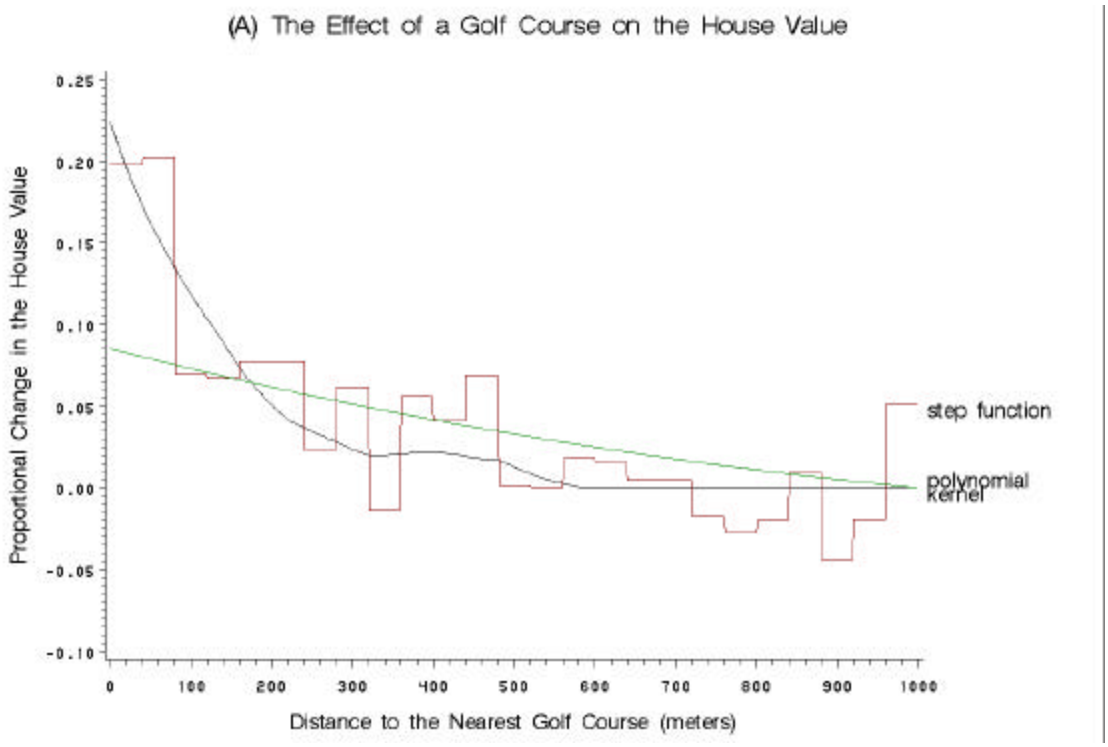
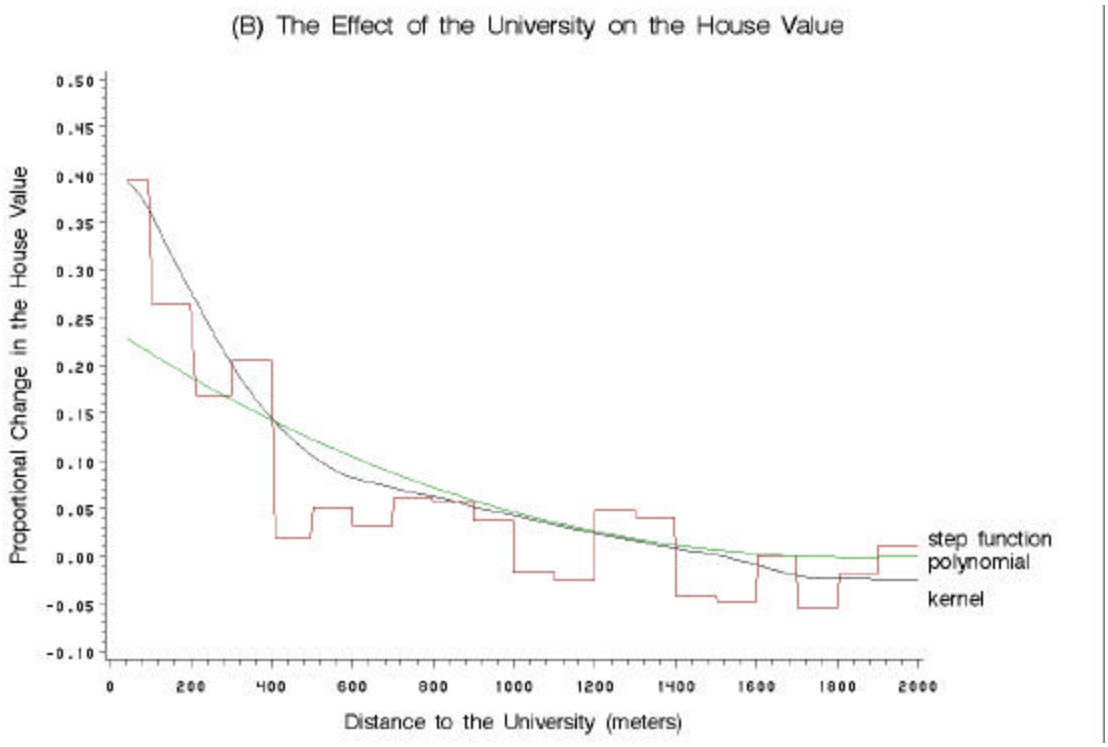
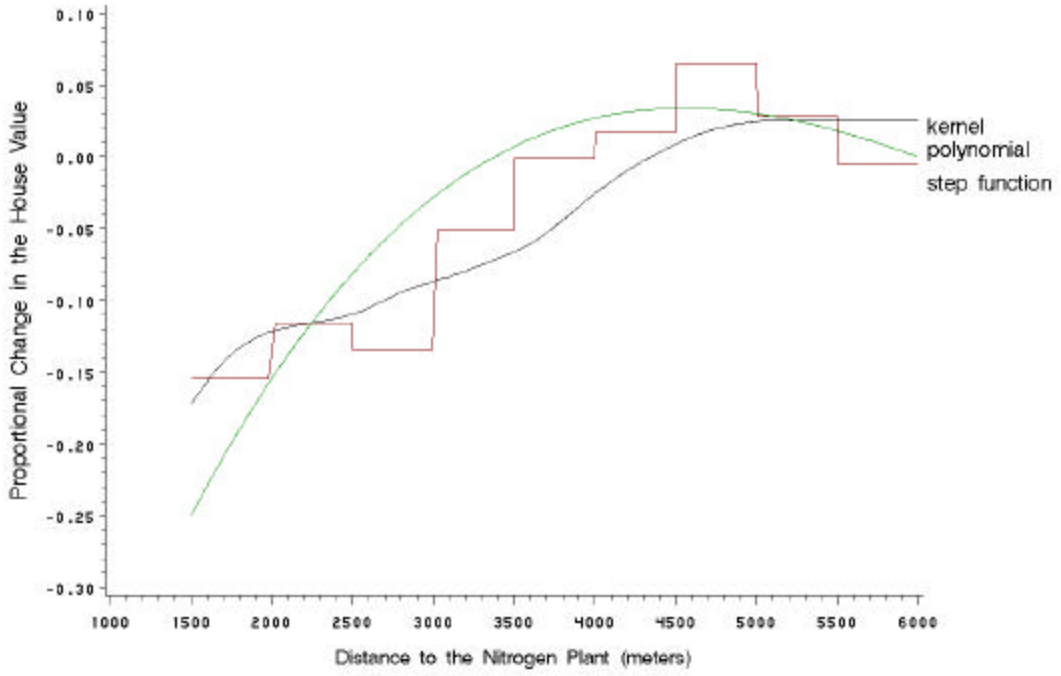


Figure 3: Nonparametric Estimates of the Effects of Location Characteristics





(C) The Effect of the Nitrogen Plant on the House Value



(D) The Effect of the Elevation on the House Value

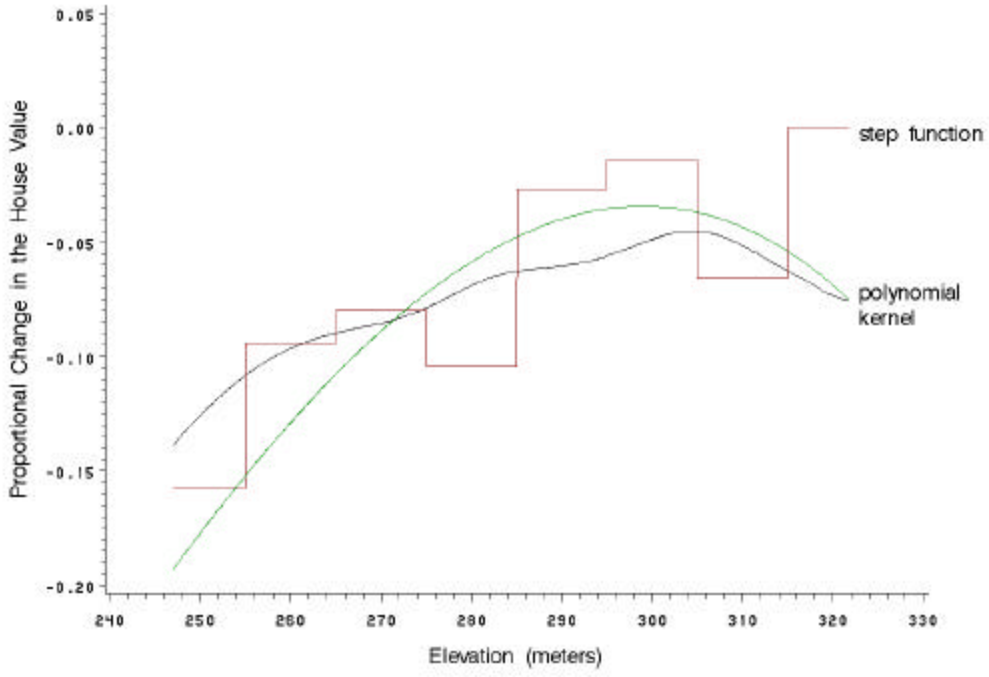
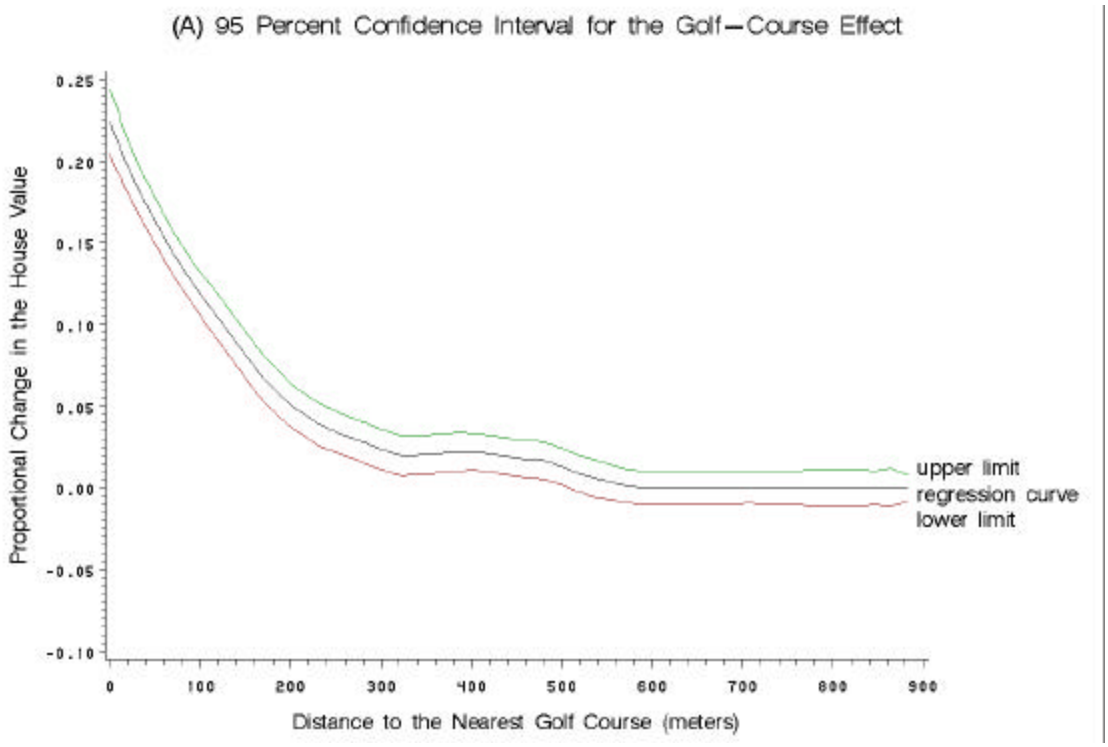
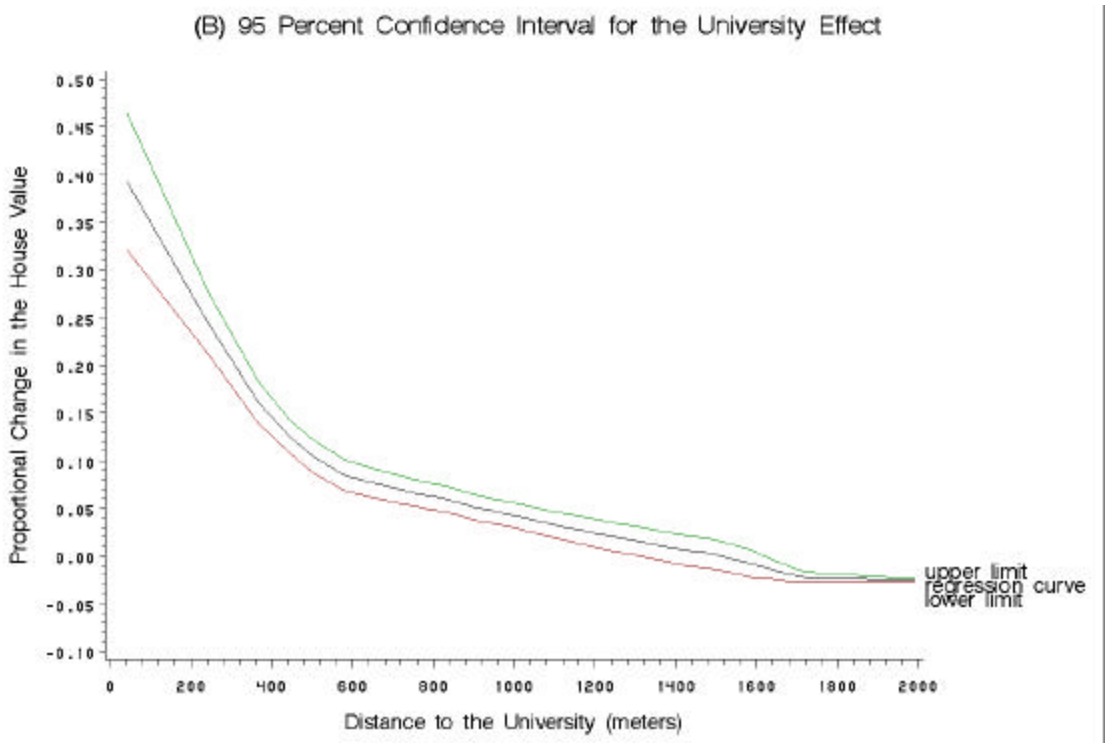
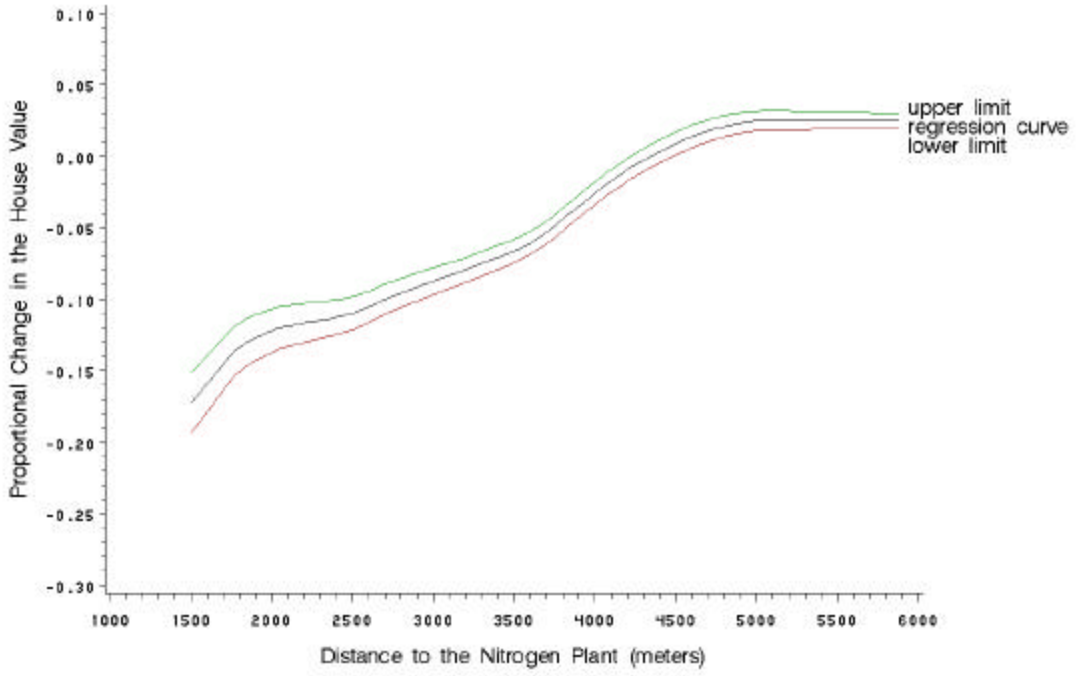


Figure 4: Confidence Bands of Nonparametric Estimates





(C) 95 Percent Confidence Interval for the Nitrogen—Plant Effect



(D) 95 Percent Confidence Interval for the Elevation Effect

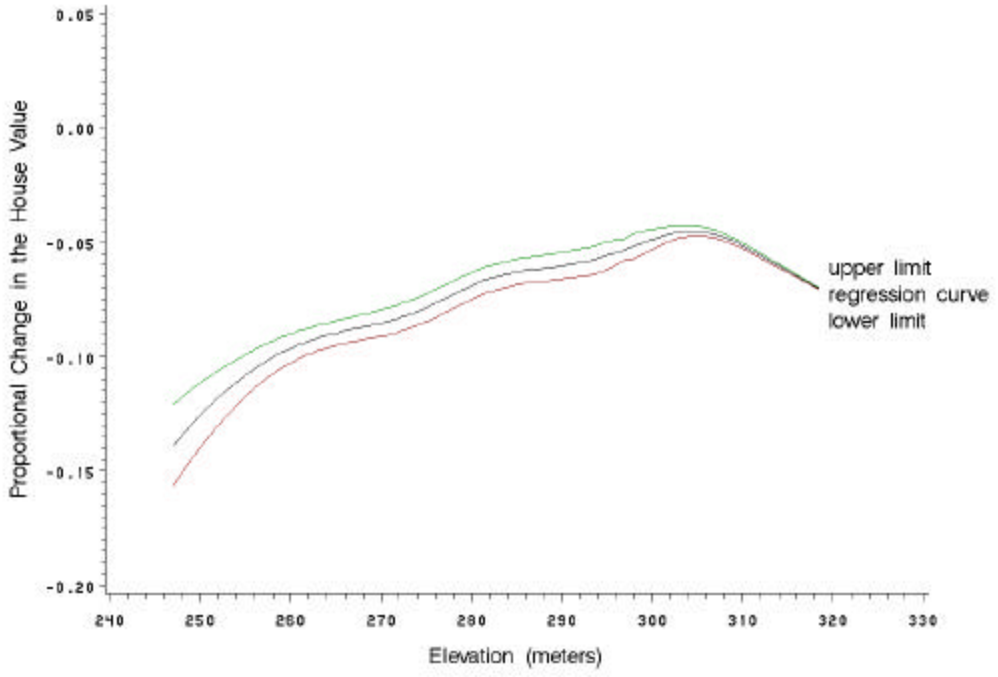
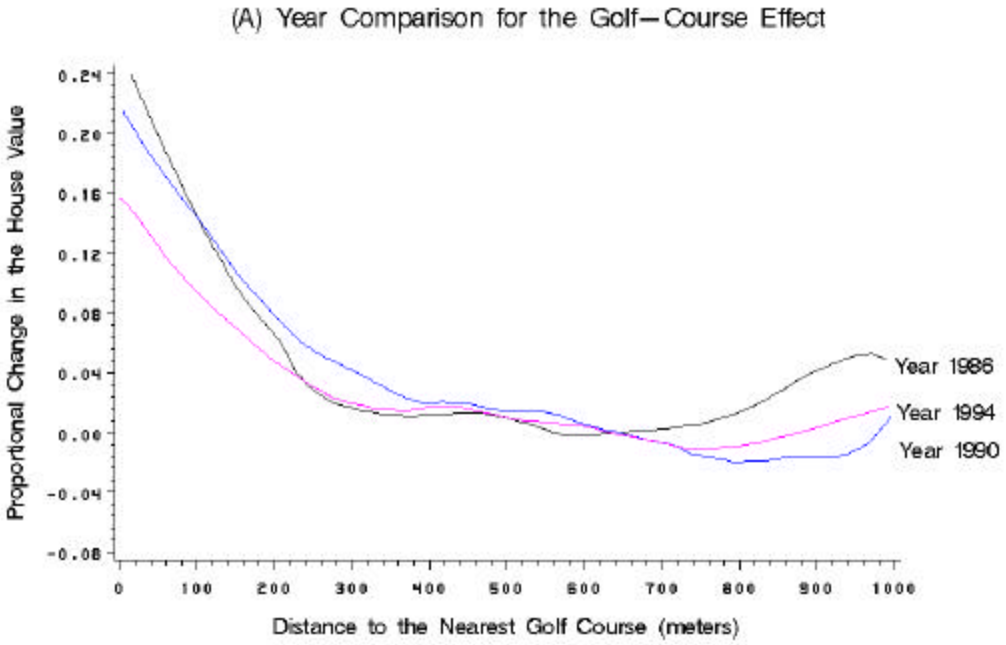
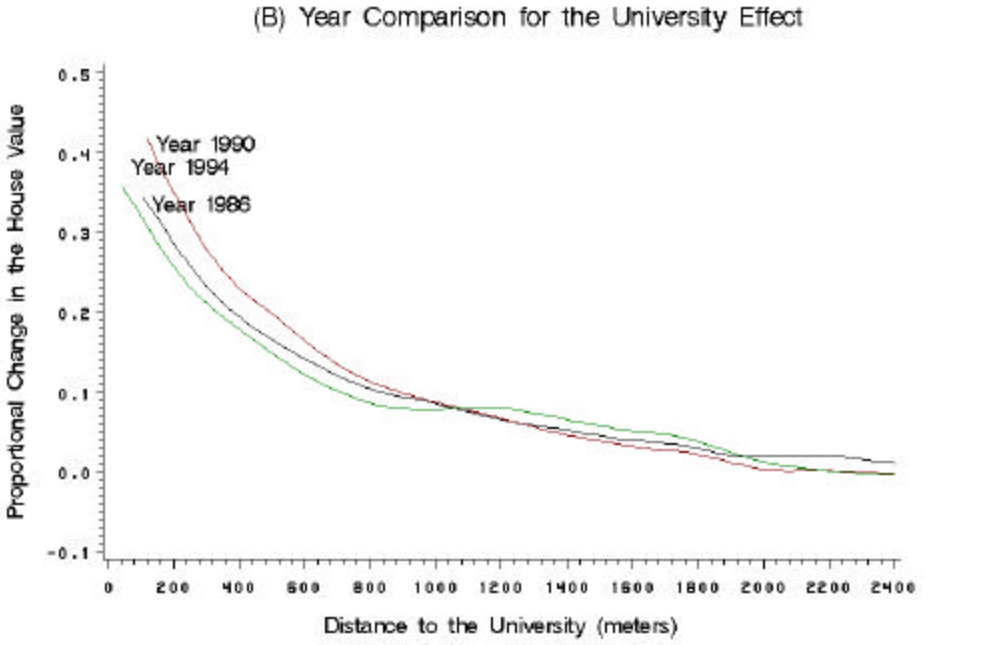
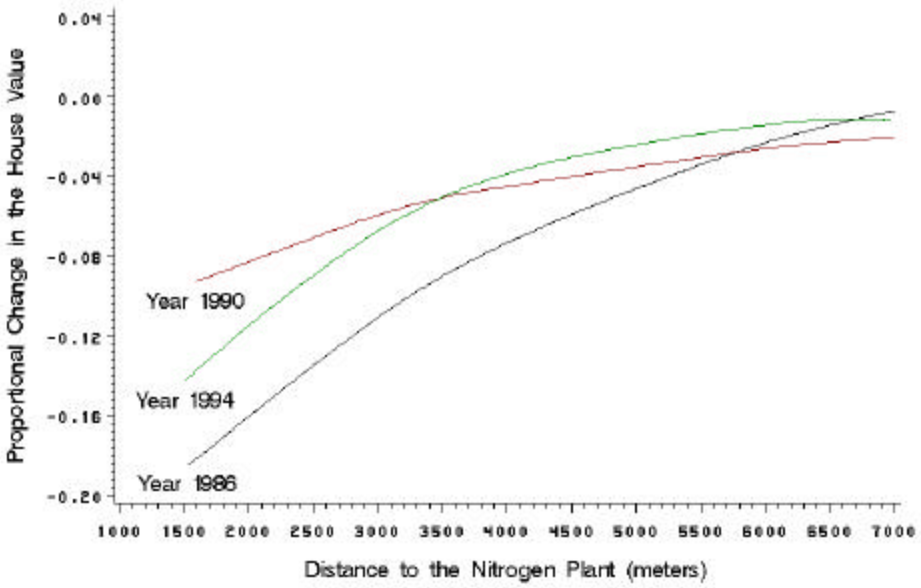


Figure 5: Year Comparisons





(C) Year Comparison for the Nitrogen–Plant Effect



(D) Year Comparison for the Elevation Effect