Hedonic Housing Prices and Agricultural Pollution: An Empirical Investigation on Semiparametric Models

Christophe Bontems,
University of Toulouse, INRA-ESR, France
bontems@toulouse.inra.fr

Michel Simioni,
University of Toulouse, INRA-ESR and IDEI, France

and

Yves Surry,
Swedish University of Agricultural Sciences, Uppsala, Sweden.

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Providence, Rhode Island, July 24-27, 2005

1We thank Anna Alberini, Pascal Lavergne, Céline Nauges, and the participants of the (EAERE) thirteenth annual conference (Budapest) for their remarks. Copyright 2005 by [Bontems, Simioni and Surry]. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on such copies.
Hedonic Housing Prices and Agricultural Pollution: An Empirical Investigation on Semiparametric Models

Abstract:
Many empirical studies challenged the parametric approach to estimate nonlinear house price models and to use instead nonparametric or semiparametric model specifications. In the area of environmental analysis using house price models, Stock (1991) and McMillen and Thorsen (2000) applied semi-parametric models to estimate the impact of hazardous wastes on house prices. Furthermore, estimating nonparametrically house price models is gaining now additional momentum because Ekeland et al. (2002) provided a theoretical justification to use such an empirical approach to estimate hedonic price models.

The objective of this paper is to assess the impact of agricultural pollution in a livestock intensive region of France using alternative non-parametric and semi-parametric hedonic price models. We propose a two stage estimation procedure were the first step is based on the partially linear model introduced by Robinson (1988). In the second step, once the parametric part of the model has been estimated on the house characteristics, we propose four specifications for the remaining functional form on the environmental variables. A parametric, a fully nonparametric, an additive, and a single index specifications are proposed and estimated using appropriate econometric procedures similar to those developed by Horowitz and Lee (2001), Iwata et al. (2000) and Robinson (1988). The performances of these models are examined, compared with respect to the benchmark nonparametric model and specification tests are proposed.
1 Introduction

Intensive agriculture activities generate negative externalities that are increasingly significant over time and space. Thus, growing concerns on the impacts of intensive livestock operations in rural areas with denser and more urbanized population are representative of such a phenomenon. Agricultural economists have attempted to quantify such effects, using hedonic models of house prices. In such a framework, negative agricultural externalities (pollution) are measured by relevant indicators that are supposed to be inversely related to the prices of houses. Then, estimating the first order derivatives of the hedonic price function with respect to the pollution indicators yields estimates of the prices of these environmental attributes and, indirectly, an estimate of consumers’ willingness to pay for these “disamenities”. Although well grounded from a theoretical point of view, the hedonic price model, when implemented empirically, raised several problems associated with the identification of the underlying structural model parameters. The true nature of the relationships between house prices and its various attributes, which thus might be complex and nonlinear, would be rather be represented by nonparametric models than the traditional parametric specifications (Ekeland et al., 2004).

Up to now, the study of the relationships between agricultural pollution and house prices has been conducted using parametric model specifications (see, for instance, Herriges et al., 2003, Huang et al., 2003; and Palmquist et al., 1997), and to the authors’ knowledge, nonparametric and/or semi-parametric models have never been employed to address this problem of agricultural pollution in housing price models\(^2\). The objective of this paper is to fill this gap by using a partially linear semi parametric hedonic price function in order to assess the impact of agricultural pollution on the prices of residential houses in a livestock-intensive region of France.

Three alternative specifications - a fully nonparametric one, a nonparametric additive one, and a single index one - are used to capture different possible forms of nonlinearity associated with these pollution variables in a hedonic price function. We also consider the parametric counterpart of the three aforementioned non- and semi-parametric model specifications. Indeed,

the performances of non- and semi-parametric models are usually investigated by comparing their goodness of fit to the parametric model "benchmark". This work differs from previous studies in two respects. First the "true" model benchmark is the fully nonparametric model, and second specification tests are performed in order to compare these specifications (fully parametric, nonparametric additive and single index) with respect to this former model benchmark.

The empirical application deals with a set of transaction prices of residential houses sold in 1996 and 1997 in Brittany, France, the leading French region for a number of livestock products and vegetable products. Agriculture in this region has two major impacts on the environment. First, the existence of intensive livestock units induce harmful effects on the environment in various forms such as the emergence of unpleasant odors and the emissions of nitrate which pollute soil, affect water quality and seep into the underground water table. The second effect of agriculture on the environment concerns the degradation of the country’s landscape resulting from intensive agricultural practices and operations. In our study, these two effects are captured by two aggregate indicators: the livestock nitrogen emissions per hectare of arable land in the rural township where the residential house is located, and the proportion of permanent pasture land converted into tilled land.

In the hedonic price model, we specify the prices of residential houses as a function of its physical characteristics, of the two environmental indicators but also of variables representing the economic structure of rural townships where the residential houses are located. Since many housing characteristics are discrete and since our main interest in this work hinges on measuring the impact of environmental variables on house prices, all the explanatory variables but the two former environmental indicators enter the hedonic price function in a linear fashion. This makes up the linear part of the hedonic price function. The two pollution indicators enter the hedonic price function in a nonparametric or semi parametric way and this constitutes the nonlinear part of the model.

The empirical strategy to estimate our housing price model follows a general-to-specific specification search involving three stages. In the first stage, the parameters involved in the linear part of the hedonic price models are estimated using Robinson's approach of partially linear model. In the second stage, the four - parametric, nonparametric, nonparametric additive and single index - specifications of the nonlinear part of the hedonic price func-
tion are estimated using the estimated residuals of the first stage estimation procedure. The econometric procedures being used involve local polynomial regression (Fan and Gijbels, 1996), average derivative estimation (Horowitz and Hardle, 1996) and marginal integration (Linton and Nielsen, 1995). Finally, specification tests similar to those of Fan and Li (1996), Horowitz and Spokoiny (2001), and Gozalo and Linton (2001), are performed. The three specifications (parametric, nonparametric additive and single index) are thus compared to the more general one, i.e. the fully nonparametric one.

The specification tests only select the nonparametric additive specification. Willingness to pay for pollution reduction is then computed for this selected model specification using the estimation procedure of derivatives for additive separable models, proposed by Severance-Lossin and Sperlich (1999). The main result is that the pollution resulting from livestock operations in rural townships is a more crucial environmental issue than the pollution due to intensive crop practices, although both affect significantly, and in a nonlinear way, house prices.

The rest of the paper is organized as follows: the semiparametric house price models are defined in Section 2. In Section 3, our general-to-specific specification search procedure is described. Section 4 is devoted to the presentation of data, while the estimation results of the proposed models are examined and discussed in Section 5. The paper ends by a concluding section highlighting the main findings of this empirical exercise.

2 Semiparametric house price models

In this section is presented a discussion of the different specifications for the hedonic price function. This function can be defined as follows. Suppose that each house can be viewed by economic agents as a bundle of different amounts of a vector of characteristics. All these characteristics are observed by the economic agents when making their choices. Hereafter, we assume that the econometrician only observe some of these characteristics we denote by $X$ when considering characteristics of the house and its surrounding (e.g. number of rooms, state of repair, age of the house, population of the city, stock of existing houses, ...) and by $Z$ when considering environmental characteristics defining the impact of agricultural pollution. The hedonic price function specifies how the price of a house varies as the characteristics
vary, i.e.

\[ Y = m(X, Z, \xi) \]  

(1)

where \( \xi \) denote the vector of unobserved (by the econometrician) characteristics of the house. For ease of simplicity, we assume that this vector is unidimensional.

Rosen (1974) provided a theoretical equilibrium framework in which the interactions of consumers and suppliers determine the hedonic price function and where differentiating the hedonic price function with respect to a given characteristic enables to derive the marginal willingness to pay for that characteristic. Rosen also proposed a two-step parametric procedure for estimating the demand for each characteristic in the case where the hedonic price function has the following additive structure:

\[ Y = m(X, Z) + \xi \]  

(2)

and where a parametric functional form of \( m(\ldots) \) is chosen. In a first step, the hedonic price function is estimated by regressing observed house prices, \( Y \), on all their characteristics, \( X \) and \( Z \), using the best fitting parametric functional form. Next a set of implicit marginal prices for a given characteristic, or marginal willingness to pay, can be computed by taking the partial derivative of the estimated hedonic price function with respect to this characteristic evaluated at the amounts of characteristics actually bought for each house sale. Finally the demand equations for all the characteristics can be recovered using these estimated implicit marginal prices as endogenous variables in a second-stage simultaneous estimation of a system of parametric functional forms for the demands.

Since Brown and Rosen (1982), the lack of identification of the marginal willingness to pay or bid functions has been a discussed issue in the hedonic model literature. Indeed, the main shortcoming of the Rosen’s second stage demand estimation is that the estimated implicit marginal price may not contain any information beyond what the first stage provides. The only new information is the functional form restriction placed on the demand equation (Horowitz, 1987, and Chattopadhyay, 1999). As shown by Brown
and Rosen (1982), if there is no new information, the estimated demand equation simply reproduces the results of the hedonic regression from which it was initially produced, i.e., the demand equation cannot be identified from the hedonic price function. Only recently, Ekeland, Heckman and Nesheim (2004) showed that the main source of the identification problems underlined in this literature lies on the commonly used linearization strategies made to simplify estimation problems. They emphasized that the hedonic price model is generically nonlinear. In the same way, Bajari and Benkard (2004) considered the identification of hedonic price models when some product characteristics are not observed by the econometrician. Using the results of Matzkin (2003), they showed that, given data on a single market, the hedonic price function and the distribution of the unobserved product characteristic are identified if the unobserved product characteristic is independent of the observed ones. Thus, the hedonic price function may have a general non-additive structure.

Armed with these observations, it would seem appropriate to consider non-parametric regression estimators as natural candidates to estimate the hedonic price function (1). But, unfortunately in this process, we would face two problems. First, the characteristic $\xi$ cannot be observed. As in Bajari and Kahn (2005), we take the view that in $Y = m(X, Z) + \xi$, the additive error term to the hedonic price function is interpreted as a vertical product characteristic observed by the consumer but not by the econometrician$^3$. Then, a second problem crops up: the curse of dimensionality given that the vectors $X$ and $Z$ may involve a large number of characteristics. Since many housing characteristics are discrete and since our main interest in this work hinges on measuring the impact of environmental variables on house prices, we assume a partially linear specification given by

$$Y = \beta'X + m(Z) + \xi$$

Expression (3) is the conceptual model that is estimated empirically in this paper using parametric and nonparametric estimation procedures. For this purpose, the following four alternative empirical model specifications are proposed and estimated.

$^3$We could use Matzkin’s (2003) nonparametric estimation procedure that directly deals with a non-additive structure. But this procedure supposes that all the observed characteristics are continuous and many housing characteristics in our dataset are discrete.
Specication of $m(z)$ & Resulting model specication

<table>
<thead>
<tr>
<th>Parametric specification</th>
<th>Linear and parametric model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(Z) = \gamma'Z$</td>
<td>$Y = \beta'X + \gamma'Z + \xi$ (M1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonparametric specification</th>
<th>Partially linear and nonparametric model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(Z) = m(Z_1, \ldots, Z_L)$</td>
<td>$Y = \beta'X + m(Z_1, \ldots, Z_L) + \xi$ (M2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Additive specification</th>
<th>Partially linear and additive model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(Z) = \sum_{t=1}^L g_t(Z_t)$</td>
<td>$Y = \beta'X + \sum_{t=1}^L g_t(Z_t) + \xi$ (M3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single index specification</th>
<th>Partially linear and single index model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(Z) = G(\gamma'Z)$</td>
<td>$Y = \beta'X + G(\gamma'Z) + \xi$ (M4)</td>
</tr>
</tbody>
</table>

These four empirical models differ from each other according to the way that the function $m(z)$ is defined. Although model (M1) is similar to a typical linear regression, the remaining model specifications, (M2), (M3) and (M4), include non-parametric and semi-parametric elements with their own specific characteristics.

3 Specification search procedure

To estimate these four empirical models requires to devise a method able to estimate the coefficients $\beta$ and the function $m(z)$. To do so, we estimate this former unknown function, using $Y - \beta X$ as the dependent variable. Even though $\beta$ has to be estimated, our estimation approach follows a two-stage procedure. In the first stage, we estimate the less restrictive available model specification, that is the specification proposed by Robinson (1988), where the function $m(.)$ is left unspecified. In the second stage, we investigate the four empirical specifications of the function, $m(.)$ defined earlier. Specification tests aimed at selecting the proper specification of the function $m(.)$ are then developed and presented.
3.1 Estimating the linear part of the hedonic price function

The first stage leading to the estimation of $\beta$ is based on the procedure proposed by Robinson (1988). It is motivated by observing that, if we subtract on both sides of (3) the conditional expectation relative to $z$, we obtain:

$$Y - E(Y|Z = z) = \beta' \left( X - E(X|Z = z) \right) + \xi \quad (4)$$

The estimation procedure can be described as follows:

1. Regress both $y_i$ and $x_i$ on $z_i$ nonparametrically to obtain residuals $\hat{Y}_i \equiv y_i - E(Y|Z = z_i)$ and $\hat{X}_i \equiv x_i - E(X|Z = z_i)$.

2. Then perform OLS on these residuals to get an estimate of $\beta$ in (4).

Robinson (1988) showed that, under regularity conditions, this procedure yields to a $\sqrt{n}$-consistent and asymptotically normal estimator $\hat{\beta}$ for $\beta$, and that there exists a consistent estimator of its limiting covariance matrix.

As a nonparametric estimator of $E(Y|Z = z_i)$ (or $E(X|Z = z_i)$), we use the local polynomial estimator $\hat{\theta}_0$ which is the solution of the following minimization problem$^4$:

$$\min_{(0,...,\theta_L)} \sum_{k=1}^n \left[ Y_k - \theta_0 - \sum_{j=1}^L \theta_j (Z_{kj} - z_{ij}) \right]^2 K_h(Z_k - z_i)$$

where $K_h(.)$ is a multidimensional kernel depending on a vector of bandwidths denoted by $h$. The solution of this weighted least squares regression problem is given by

$$\hat{\theta} = (Z^T Z)^{-1} Z^T Y$$

where $Y = (Y_1, \ldots, Y_n)'$, and

$^4$A local polynomial estimator possesses a number of desirable theoretical and practical properties relative to other smoothing methods including the widely applied Nadaraya-Watson kernel estimator. See Fan and Gijbels (1996).
\[ Z = \begin{bmatrix}
1 & Z_{11} - z_{i1} & \ldots & Z_{1L} - z_{iL} \\
1 & Z_{21} - z_{i1} & \ldots & Z_{2L} - z_{iL} \\
\vdots & \vdots & \ddots & \vdots \\
1 & Z_{n1} - z_{i1} & \ldots & Z_{nL} - z_{iL}
\end{bmatrix}, \]

and \( \Upsilon = \text{diag}(K_h(Z_k - z_i)) \). The estimate of the function \( E(Y|Z = z_i) \) is thus the first component of the vector \( \hat{\theta} \), i.e. \( E(Y|Z = z_i) \equiv \hat{\theta}_0 \). Similar formulas can be written for the estimator of \( E(X|Z = z_i) \).

As for choosing the bandwidth \( h \), in the estimations of \( E(Y|Z = z_i) \) and \( E(X|Z = z_i) \), automatic bandwidth selection is possible, such as plug-in methods or cross-validation methods. Hereafter, we apply the following cross-validation criterion that has been the most popular for partially linear house price model estimation (see Stock, 1991, Anglin and Gençay, 1996, and, more recently, Kondo and Lee, 2003). It minimizes

\[
CV(h) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{\beta} \tilde{X}_i)^2
\]

with respect to the bandwidth \( h \) (recall that \( \tilde{Y}_i, \tilde{X}_i, \) and \( \tilde{\beta} \) are functions of \( h \)).

### 3.2 Estimating the nonlinear part of the hedonic price function

Once \( \beta \) is estimated, we use \( W = Y - \tilde{\beta} X \) for dependent variable in all of the models proposed, and only specify the function \( m(z) \) according to models M1-M4. We turn now to the description of the three non and semiparametric models M2-M4 and to the way we can get estimates of these models.

#### 3.2.1 Fully nonparametric model (M2)

This model is used as benchmark for our investigation on the ability of semiparametric model to estimate hedonic functions for real datasets. The func-
tion $m(z)$ is estimated using the second stage of Robinson’s procedure based on a nonparametric regression of $Y - \tilde{\beta} X$ on $Z$ leading to an estimate of $m(\cdot)$. As a nonparametric estimator of $m(z)$, we use a local polynomial estimator.

3.2.2 Additive model (M3)

The additive house price model is of the form:

$$m(Z) = \sum_{i=1}^{L} g_i(Z_i)$$  \hspace{1cm} (6)

where $\left(g_i(\cdot)_{i=1}^{L}, \text{resp.} \right)$ is a set of $L$ unknown functions satisfying the identifiability condition

$$E(g_i(Z_i)) = 0$$ \hspace{1cm} (7)

for every $l = 1, \ldots, L$.

Additive models are usually estimated using the Hastie and Tibshirani’s (1990) backfitting algorithm\(^5\). This algorithm proceeds as follows:

1. **Initialization**: Select initial non parametric estimates of the functions $\left(g_i(\cdot)_{i=1}^{L}\right)$, denoted by $\left(g_i^0(\cdot)_{i=1}^{L}\right)$.

2. **Iteration**: Obtain an estimate of each function $g_i^l(\cdot)$ by nonparametric regression of $w - \sum_{l=1, l \neq k}^{L} g_{i}^{l-1}(Z_i)$ on $Z_k$.

3. **Convergence**: Continue iteration until there is little change in the individual function estimates.

While this algorithm converges to a unique solution independent of the starting values, the statistical properties of the resulting estimates are not well understood. An alternative procedure for the estimation of additive models based on marginal integration has been proposed by Linton and Nielsen (1995) in the case of two independent variables and extended by Tjøstheim

\(^5\)see, for instance, Pace, 1998, Iwata et al., 2000, and Martins-Filho and Bin, 2005
and Auestad (1995) and Chen et al. (1996) to larger dimensions. The idea behind this estimator is quite straightforward: if the function $m(Z)$ is the sum of univariate functions, i.e. $m(Z) = E(Y \mid Z = z) = \sum_{l=1}^{L} g_l(z_l)$, then each function can be recovered as follows

$$g_l(z_l) = \int \ldots \int m(z) \varphi_{-l}(z_{-l}) \, dz_{-l}$$

(8)

where $z_{-l}$ is the vector of all components in $z$ but the $l$-th, and $\varphi_{-l}(z_{-l})$ denotes the joint density function of this $(L-1)$-dimensional vector. Hence, each component function $g_l$ is identified from $m(z)$ through a simple integration procedure. Given an i.i.d. sample $(W_i, Z_i)$, a sample version of (8) is

$$\hat{g}_l(z_l) = \frac{1}{n} \sum_{i=1}^{n} m(Z_{i1}, \ldots, Z_{i,l-1}, z_l, Z_{i,l+1}, \ldots, Z_{iL})$$

(9)

An estimate of the function $\hat{g}_l(z_l)$ can be obtained by replacing the unknown multivariate function $m(Z_{i1}, \ldots, Z_{i,l-1}, z_l, Z_{i,l+1}, \ldots, Z_{iL})$ by a nonparametric estimate. As a nonparametric estimator, we can use a local polynomial estimator.

3.2.3 Single-index model (M4)

A single-index house price model rests on the assumption that all the information conveyed by the independent variables can be summarized into a single index $\gamma'Z$ where $\gamma$ is a vector of unknown coefficients, linked to the endogenous variable through an unknown link function $G(.)$ as:

$$m(z) = G(\gamma'Z)$$

(10)

The main idea underlying these models is to avoid the curse of dimensionality, by reducing the dimension of the regressor space to one, through the index. There is a cost being paid in terms of identification since for any arbitrary $\delta$ and $\nu$, equation (10) is equivalent to $m(z) = G^*(\nu + \delta(\gamma'Z))$, and thus size and scale normalization are needed.
If \( G \) were known, the estimation of \( \gamma \) would be achieved by solving the following nonlinear least square estimation problem:

\[
\hat{\gamma} = \arg \min_{\gamma} \sum_{i=1}^{n} \left( W_i - G(\gamma'Z_i) \right)^2
\]

where \((W_i)_{i=1,\ldots,n}\) denotes the first step residuals \((Y_i - \tilde{\beta}X_i)_{i=1,\ldots,n}\). Since \( G(\cdot) \) is unknown, the estimation procedures require an iterative and difficult numerical resolution procedure. Fortunately, if \( Z \) only have continuous components, the Density Average Derivative methods (DWADE) may be used to estimate \( \gamma \) without solving an optimization problem (see Powell et al., 1989). The main idea for estimating \( \gamma \) rely on the fact that, up to a multiplicative parameter,

\[
\gamma \propto E \left[ \frac{\partial G(\gamma'Z)}{\partial Z} \right].
\]

This leads to the estimator:

\[
\gamma_{DWADE} = -\frac{2}{n} \sum_{i=1}^{n} y_i \cdot \hat{\varphi}(Z_i)
\]

where \( \hat{\varphi}(\cdot) \) is a nonparametric estimate of the unknown density \( \varphi(\cdot) \) of \( Z \).

### 3.3 Third stage: Specification tests

As stressed at the beginning of this paper, one of the contribution of this work is to adopt specification tests whereby the various empirical specifications of the function \( m(z) \) are tested and compared to the “true” model benchmark model that is the full non-parametric specification. In our case, the “true” non-parametric model specification is represented by model (M2) while the three remaining ones, (M1), (M3) and (M4) represent the alternative model specifications.

#### 3.3.1 Parametric vs nonparametric

Several papers have proposed tests of a parametric specification of a regression model against a nonparametric alternative (see among others, Lavergne
and Vuong, 1996 or Härdle and Mammen, 1993). Recently, Horowitz and Spokoiny (2001) have developed a new test used here. We test the null hypothesis, \( H_0 \), that \( m(Z) \) belongs to some parametric family i.e. there exists some \( \gamma \in \Gamma \) such that \( m(Z) = M(Z, \gamma) \) against the alternative, \( H_1 \) that there is no such \( \gamma \). The test is based on the distance \( S_h(\gamma) \) between the kernel estimation of \( m(Z) \) and the kernel-smoothed estimation of the parametric regression \( M(Z, \gamma) \).

\[
S_h(\gamma) = \sum_{i=1}^{n} \left( \hat{m}_h(Z_i) - \hat{M}_h(Z_i, \gamma) \right)^2
\]  

(12)

were \( \hat{M}_h(Z_i, \gamma) = \sum_{j=1}^{n} W_h(Z_i, Z_j) M(Z_j, \gamma) \) is the kernel-smoothed parametric estimator with kernel weight \( W_h(\cdot, \cdot) \). The test statistic, \( T^* \) is then computed with a rate-optimal and adaptative bandwidth, based on set of bandwidth values \( h \in H_n \), and is centered and studentized:

\[
T^* = \max_{h \in H_n} \frac{S_h(\gamma) - \hat{N}_h}{\hat{V}_h}
\]  

(13)

We use \( T^* \) for the empirical tests presented in section 4.

### 3.3.2 Additive vs nonparametric

Gozalo and Linton (2001) propose several nonparametric tests of a very general additive structure in nonparametric regression. The null hypothesis \( H_0 \) is set in a more general framework than the additive model we are testing here and allows for discrete covariates and unknown link function \( L(\cdot) \) belonging to a parametric family of transformations. Hereafter, we will use this framework with no unknown link function (\( L=\text{Identity} \)). Under \( H_0 \) it is assumed that for some parameters \( \gamma \in \Gamma \), \( m(Z) = M(Z; \gamma) = \gamma_0 + \sum_{d=1}^{D} g_d(Z_d) \). Four tests statistics are proposed based on Hausman-like statistics, \( \tau_0 \), on measures of the distance between the nonparametric estimates of the function, (resp. its residuals) under the null and under the general alternative that \( H_0 \) is false, \( \tau_1 \) (resp. \( \tau_3 \)), and on Lavergne and Vuong-like statistics, \( \tau_2 \):

\[
\tau_0 = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{m}(Z_i) - \hat{M}(Z_i; \gamma) \right)^2 \pi(Z_i)
\]
\[ \tau_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i \left( \hat{m}(Z_i) - \hat{M}(Z_i; \gamma) \right) \pi(Z_i) \]
\[ \tau_2 = \frac{1}{n^2 h_d} \sum_{i \neq j} K_{ij} \hat{u}_i \hat{u}_i \pi(Z_i) \pi(Z_j) \]
\[ \tau_3 = \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i^2 - \hat{u}_i)^2 \pi(Z_i) \]

where \( \hat{u}_i = Y_i - \hat{m}(Z_i) \) are the unrestricted residuals, \( \tilde{u}_i = Y_i - \hat{M}(Z_i; \gamma) \) are the additive (restricted) residuals, \( \pi(\cdot) \) is a weighting function and \( K_{ij} = K((Z_i - Z_j)/h) \mathbf{1}(Z_i = Z_j) \).

### 3.3.3 Single index vs nonparametric

To test the single index specification of the regression function \( m(z) \), we use the procedure proposed by Fan and Li (1996). It is based on the null \( H_0 \) that \( m(z) = G(\gamma' z) \), for some \( \gamma \in \mathbb{R}^d \) and some unknown function \( G(\cdot) : \mathbb{R} \to \mathbb{R} \) against the general alternative that \( H_0 \) is not true. The test is based on the quantity \( \nu = Y - G(\gamma Z) \). Under \( H_0 \), \( \mathbb{E}[\nu|Z] = 0 \). Then the statistic
\[ \mathbb{E}[\nu \mathbb{E}[\nu|Z]] = \mathbb{E}[\mathbb{E}[\nu^2|Z]] \geq 0 \]
and the equality holds if and only if \( H_0 \) is true. More precisely, the test statistic \( T^c \) is based on \( \text{U-statistics constructed as moments of } \nu : \)
\[ T^c = \frac{n h^{d/2} I_n^c}{\sqrt{(2 \sigma_c)}} \]
with \[ I_n^c = \frac{1}{n(n-1)h^d} \sum_{i=1}^{n} \left( \hat{u}_i \hat{f}(\hat{\gamma}' Z_i) \right) \left( \sum_{j \neq i} (\hat{\varphi}_j \hat{f}(\hat{\gamma}' Z_j)) K_{ij} \right) \]

where \( \hat{u}_i = Y_i - \hat{G}(\hat{\gamma}' Z_i) \) are the single index (restricted) residuals, \( \hat{f}(\hat{\gamma}' Z_i) \) is a kernel estimator of the density \( f(\gamma' Z) \) of \( \gamma' Z \), and \( K_{ij} = K((\hat{\gamma}' Z_i - \hat{\gamma}' Z_j)/h) \) with \( K(\cdot) \) a kernel function. Under \( H_0 \), the asymptotic distribution of the statistic \( T^c \) is \( N(0,1) \).
4 Data

The variables used in the hedonic regression analysis fall into three broad categories: (i) the price and the physical attributes of the home and the lot, (ii) the characteristics of the surrounding township, and (iii) the environmental "disamenities". Data on house sales were obtained from the MIN (Marché Immobilier des Notaires) database. This database provides a very detailed description of every house sale including the sale price and the physical attributes of the home and the size of the associated lot. Taxes and various fees linked to the sale were incorporated into the computation of the price really paid by the buyer of the house. This variable is denoted by PRICE. Four physical characteristics of the home and lot are used in the empirical analysis: The age of the house, AGE, its state of repair, REPAIR, the number of rooms, ROOMS, and the lot size, LOT. A description of these five variables is given in table 1.

The second category of explanatory variables are characteristics of the surrounding township. We associated each home with the township where the house is located, and used the INSEE\(^6\) database to obtain the total population of the township (POP) and the average family income (AVINC). The MIN database enables us to compute for each township the percentage of vacant houses. This variable, denoted by VACANT, measures the state of the property market in each township. The last variable in this category (denoted by COUNTY), is a dummy variable indicating the location of the township in the Ille et Vilaine county or not. The Brittany region is composed of four counties the Ille et Vilaine county being the less rural one. This variable also measures the state of the property market but now in the Brittany area. A description of the four variables: POP, AVINC, VACANT, and COUNTY, is given in table 1.

To measure consumer’s willingness to pay for environmental "disamini-
ties" generated by agriculture would require to dispose of, not only personal and confidential information on consumers’ views on such issues, but also of detailed information on the location of livestock operations relative to the consumers’ residential houses. Collecting such quantitative information is so

---

\(^6\)INSEE stands for Institut National de la Statistique et des Études Économiques, the French government agency collecting and producing information on the French economy and society.
sensitive that it is impossible to undertake relevant surveys to generate the relevant data. Given this situation, agricultural pollution is measured by two aggregate indicators which were provided by the French Direction Régionale de l’Agriculture et de la Forêt de Bretagne:

- The first one (NITRO) is the per hectare of arable land amount of nitrogen emissions of livestock operations in the rural township where the residential house is located.

- The second indicator considered (TMEAD) is the proportion of permanent pasture land converted into tilled land. A high value associated with this variable would indicate a degradation of country’s landscape.

Deriving the first order derivatives of house prices with respect to these variables will give an estimate of the prices of these two environmental attributes and indirectly an estimate of consumers’ willingness to pay for these two agricultural "disamenities".

A description of the two pollution indicators is given in table 1. We also present in figure 1 the results of the nonparametric estimation of the joint density of these two indicators. This joint density appears to be single peaked density and the majority of the observations corresponds to values of TMEAD and NITRO belonging to the [20%, 45%] and [0, 50 kg/ha] ranges, respectively.

5 Empirical results

In the hedonic price models, we specify the prices of residential houses as a function of its physical characteristics, of the two environmental indicators but also of variables representing the economic structure of rural townships where the residential houses are located. All the explanatory variables but the two former environmental indicators enter the hedonic price models in a linear fashion. This makes up the linear part of the hedonic price function. The two pollution indicators enter the hedonic price function in a nonparametric or semiparametric way and this constitutes the nonlinear part of the

---

7A bivariate normal kernel was used and the bandwidth was chosen using the Silverman’s rule. See Silverman (1986).
Table 1: Variables and Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Units</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRICE</td>
<td>Market Price</td>
<td>Euro</td>
<td>15400</td>
<td>162583</td>
<td>167494</td>
<td>34593</td>
</tr>
<tr>
<td>AGE</td>
<td>Age</td>
<td>Year</td>
<td>0</td>
<td>298</td>
<td>47.835</td>
<td>42.018</td>
</tr>
<tr>
<td>REPAIR</td>
<td>State of repair</td>
<td></td>
<td>= 1 if good</td>
<td>0</td>
<td>1</td>
<td>0.687</td>
</tr>
<tr>
<td>ROOMS</td>
<td>Number of rooms</td>
<td>#</td>
<td>1</td>
<td>7</td>
<td>4.429</td>
<td>1.353</td>
</tr>
<tr>
<td>LOT</td>
<td>Lot size</td>
<td>#</td>
<td>102</td>
<td>21880</td>
<td>1793</td>
<td>2551</td>
</tr>
<tr>
<td>COUNTY</td>
<td>County location</td>
<td>= 1 if “Île et Vilaine”</td>
<td>0</td>
<td>1</td>
<td>0.478</td>
<td>0.499</td>
</tr>
<tr>
<td>VACANT</td>
<td>Vacant Housing</td>
<td>Percent</td>
<td>0.000</td>
<td>20.000</td>
<td>6.275</td>
<td>3.157</td>
</tr>
<tr>
<td>POP</td>
<td>County population</td>
<td># (x1000)</td>
<td>0.104</td>
<td>4.972</td>
<td>2.047</td>
<td>1.215</td>
</tr>
<tr>
<td>AVINC</td>
<td>Average income</td>
<td>Euro</td>
<td>571</td>
<td>2854</td>
<td>1082</td>
<td>250</td>
</tr>
<tr>
<td>TMEAD</td>
<td>Temporary meadows</td>
<td>Percent</td>
<td>0.010</td>
<td>70.143</td>
<td>29.420</td>
<td>9.972</td>
</tr>
<tr>
<td>NITRO</td>
<td>Nitrogen concentration</td>
<td>kg/ha</td>
<td>0.000</td>
<td>339.48</td>
<td>45.469</td>
<td>51.118</td>
</tr>
</tbody>
</table>

N=2092 observations 1 Euro = 6.55937 FF

model. The four partially linear house price models we estimate are defined as followed:

- Parametric specification (M1):

\[
\ln \text{Price}_i = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{REPAIR}_i + \beta_3 \text{ROOMS}_i + \beta_4 \text{LOT}_i + \beta_5 \text{COUNTY}_i + \beta_6 \text{VACANT}_i + \beta_7 \text{POP}_i + \beta_8 \text{AVINC}_i + \gamma_1 \text{TMEAD}_i + \gamma_2 \text{NITRO}_i + \xi_i
\]

- Nonparametric specification (M2):

\[
\ln \text{Price}_i = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{REPAIR}_i + \beta_3 \text{ROOMS}_i + \beta_4 \text{LOT}_i + \beta_5 \text{COUNTY}_i + \beta_6 \text{VACANT}_i + \beta_7 \text{POP}_i + \beta_8 \text{AVINC}_i + m(\text{TMEAD}_i, \text{NITRO}_i) + \xi_i
\]

- Nonparametric additive specification (M3):

\[
\ln \text{Price}_i = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{REPAIR}_i + \beta_3 \text{ROOMS}_i + \beta_4 \text{LOT}_i + \beta_5 \text{COUNTY}_i + \beta_6 \text{VACANT}_i + \beta_7 \text{POP}_i + \beta_8 \text{AVINC}_i + g_1(\text{TMEAD}_i) + g_2(\text{NITRO}_i) + \xi_i
\]
Figure 1: Joint iso-density curves for the two environmental variables \((z_1, z_2)\)

- Single index specification (M4):

\[
\ln Price_i = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{REPAIR}_i + \beta_3 \text{ROOMS}_i + \beta_4 \text{LOT}_i + \\
\beta_5 \text{COUNTY}_i + \beta_6 \text{VACANT}_i + \beta_7 \text{POP}_i + \beta_8 \text{AVINC}_i + \\
G(\gamma_1 TMEAD_i + \gamma_2 \text{NITRO}_i) + \xi_i
\]

In the following, we report the results of the specification search procedure presented in the previous section. We first pay attention to the estimated values of the parameters involved in the linear part of the hedonic price function (first stage). We then present a 2D-graphical comparison of the four specifications of \(m(z)\) (second stage), followed by the specification test results (third stage). Marginal willingness-to-pay for pollution reduction is then computed for the selected model specification at the end of this section.

5.1 Linear part

The estimates of the parameters \(\beta\) involved in the linear part of the first step estimation are reported in table 2. The bandwidth used in the estimation
Table 2: Estimates of the parameters the first step and of the linear and Single index models

<table>
<thead>
<tr>
<th>First step estimates</th>
<th>Partial linear estimates</th>
<th>standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
<td><strong>estimates</strong></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.002</td>
<td>0.0002</td>
</tr>
<tr>
<td>Repair</td>
<td>0.359</td>
<td>0.0174</td>
</tr>
<tr>
<td>Rooms</td>
<td>0.140</td>
<td>0.0057</td>
</tr>
<tr>
<td>Lot</td>
<td>0.029</td>
<td>0.0028</td>
</tr>
<tr>
<td>County</td>
<td>0.051</td>
<td>0.0169</td>
</tr>
<tr>
<td>Vacant</td>
<td>-0.017</td>
<td>0.0032</td>
</tr>
<tr>
<td>Pop</td>
<td>0.016</td>
<td>0.0074</td>
</tr>
<tr>
<td>Avinc</td>
<td>0.050</td>
<td>0.0061</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second step estimates</th>
<th>Fully parametric model estimates (standard error)</th>
<th>Single Index model estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
<td><strong>estimates</strong></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.889</td>
<td>(0.0253)</td>
</tr>
<tr>
<td>Tmead</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td>Nitro</td>
<td>-0.0006</td>
<td>(0.0001)</td>
</tr>
</tbody>
</table>

For comparison purpose, the coefficient in the Single index model \( T_{\text{mead}} \) has been normalized to its corresponding value in the linear model.

of the conditional expectations in equation 4 is selected using Kondo and Lee’s (2003) cross-validation criterion. More details about the bandwidth choice and the grid search procedure are given in appendix. In the lower part of this table are also reported estimates of the parameters involved in the fully parametric and single index specifications of \( m(z) \). We have followed the guidance given by Härdle and Tsybakov (1993) and used by Horowitz and Härdle (1996) in order to choose the bandwidth in the single index estimation. They show that the asymptotically optimal bandwidth \( h_{DWADE} \) for the DWAD estimator has the form:

\[
h_{DWADE} = h_0 n^{-\frac{2}{3p+4k+2}}
\]

where \( k \) is the dimension of vector \( Z \), and it is assumed that all the partial derivatives of the density of \( Z \) up to order \( p + 1 \) exist. This leads here with \( p = 1 \) and \( k = 2 \) to \( h_{DWADE} = h_0 n^{-1/3} \). We then have used an empirical rule of selection with a grid search on \( h_0 \) to select the bandwidth in our ap-
All the estimated parameters belonging to the linear part of the housing price models\(^9\) are significant and have the expected signs and magnitudes. Examining first the influence of the physical characteristics of the houses on their prices, we note that older houses are characterized by prices declining at a rate of 0.2% per year. Undertaking major renovations to a residential house in Brittany leads to a 35% appreciation in its price, \textit{ceteris paribus.} A larger number of rooms or a bigger lot size are factors contributing to increase the values of the houses. Variables defining the township where the houses are located have signs that are conformed to our expectations. Hence, the prices of any houses located in the rural townships of the most urbanised county of Brittany (Ile-et-Vilaine) experience an average increase by 9.1% of their prices. By contrast, residential houses located in rural townships with higher housing vacancy rates would experience lower prices while opposite effects would take place with rural municipalities that either are more populated or have households with higher average incomes.

Looking at the bottom part of Table 1, we note that the environmental indicators have the expected influence of the prices of the houses. Their estimated coefficients are statistically significant and negative. Thus, one percent point increase in the share of permanent pasture converted into tillable land results in a 0.3% decline in the price of houses. A similar interpretation could be made for the influence of the (livestock) nitrogen emissions on the property values. Although the magnitudes of these two coefficients look quite small at first glance, their true economic significance will come to the fore when corresponding WTP are estimated and discussed further in this paper. For comparative purposes we report at the bottom part of Table 2 the parameter estimates of the index model. As expected, these two coefficients are negative. No straightforward interpretation can be provided due to the existence of the nonparametric function \(G(.).\)

\(^8\)Another approach is based on a resampling procedure. See Horowitz (1998) for details.
\(^9\)As the price of houses are expressed in a logarithmic form, the estimated coefficients could be interpreted as the percentage variation in the price of the house resulting from one unit change in the explanatory variables.
5.2 Nonlinear part of the hedonic price function

Like in any empirical study of nonparametric models, this one reports the role and importance of nonlinearity in the form of a graphical analysis. Hence, we develop the estimated response surfaces linking housing prices to the two pollution indicators for the four specifications of the function \( m(z) \) in figure 3. We have restricted the representation of these curves to an area where the density of the joint distribution of the environmental variables \( z_1 \) and \( z_2 \) was high, i.e. where the values of TMEAD belong to the [20\%, 45\%] interval and those of NITRO are located in [0, 50kg/ha] range (see figure 1). The bandwidth choice for the single index specification has been discussed in section 5.1 above. For the nonparametric specification, we built a grid search and use a cross-validation criterion for the bandwidth choice, solving the following program:

\[
\min_h CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left( W_i - \hat{m}_h^{-i}(z_i) \right)^2
\]  

(15)

where:

- \((W_i)_{i=1,...,n}\) are the first step residuals \((Y_i - \tilde{\beta}X_i)_{i=1,...,n}\).

- \(\hat{m}_h^{-i}(z_i)\) denotes the leave-one out local polynomial estimator used for the cross validation.

- Each bandwidth is computed with respect to the distribution of \( z_i \) (through its standard deviation \( \sigma(z_i) \)) and the theoretical rate of convergence, so that \( h_i = h_0 \sigma(z_i) n^{-\frac{1}{5}} \) for \( i = 1, 2 \)

For the additive specification, the derivation of the nonparametric estimator for the functions \( (g_l(\cdot))_{l=1}^{L} \) involves the choice of a vector of bandwidths. As \( L = 2 \) in our application, we follow the Linton and Nielsen’s (1995) rule of thumb, controlled by the theoretical guidance given by their theorem, giving the bandwidth \( h_l \) for \( \hat{g}_l(z_i) \), for \( l = 1, 2 \) as the following one:

\[
\]
Table 3: Estimation of $m(z_1, z_2)$ for the four specifications
\[ h_l = h_{0l} \left( \frac{\hat{\sigma}^2 \nu(k)(\max(z_l) - \min(z_l))}{\mu(k)^2(\hat{\theta}_1 + \hat{\theta}_2)^2} \right)^{\frac{1}{2}} n^{-\frac{1}{6}} \]

where:

- \( h_{0l} \) is a multiplicative factor used to tune the bandwidth according to this rule of thumb.

- \( \hat{\theta}_l \) are respective coefficients of \( z_l^2 / 2 \), \( l = 1, 2 \), obtained in an auxiliary regression of \( W_l \) on a constant, \( z_1, z_2, z_1z_2, z_1^2 / 2 \) and \( z_2^2 / 2 \) while \( \hat{\sigma}^2 \) is obtained from the residual of this regression.

- The terms \( \nu(k) \) and \( \mu(k) \) are constant that only depend on the kernel used.

A visual inspection of the four estimated surfaces suggests that the responses of the house prices to the two environmental indicators look very similar in terms of shape and steepness for the specifications (M2) and (M3). On the other hand, the fully parametric specification of the hedonic price function seems to be unable to capture all the features of our data sample. The same pattern would also apply to the partially linear specification involving a single index in the nonlinear part of the hedonic price function.

### 5.3 Specification tests

Each specification test presented in section 3 was performed to compare the more general nonparametric model with the restricted specifications. All these tests require the choice of bandwidths, which is fully explained in a technical appendix at the end of the paper. In the absence of theoretical guidance for some of these tests we also report some sensitivity analysis of the test statistics to the bandwidths in the same appendix.

Table 4 summarizes the main results of the three specification tests. The results show that the nonparametric additive model is clearly not rejected, while the two others are. This result is consistent with the informal graphical finding that the parametric and single index specifications fail to capture important nonlinear features of the data and that the nonparametric additive specification fits the data satisfactorily.
5.4 Marginal prices

Based on the former specification test diagnosis, we now report the willingness to pay estimates for pollution reduction for the nonparametric additive model (M3). These computed values are derived using the estimator developed by Severance-Losin and Sperlich’s (1999):

\[
\hat{g}_i^l(z_i) = \frac{1}{n} \sum_{i=1}^{n} m_i^l(Z_{i,1}, \ldots, Z_{i,l-1}, z_l, Z_{i,l+1}, \ldots, Z_{i,L})
\]  

(16)

An estimate of the function \(\hat{g}_i^l(z_i)\) can be obtained by replacing the unknown multivariate function \(m_i^l(Z_{i,1}, \ldots, Z_{i,l-1}, z_l, Z_{i,l+1}, \ldots, Z_{i,L})\) by a nonparametric estimate. As a nonparametric estimator, we use a local polynomial estimator.

In the figures 5a and 5b\(^{10}\), we report the estimated values of the willingness to pay for each housing transaction, as well as a nonparametric estimation of the mean willingness to pay function and its 95% confidence bounds. In figure 6a and 6b are reported the same mean willingness to pay function expressed as a percentage of the corresponding housing prices.

\(^{10}\)It is important to note that WTP estimates presented in both figures 5 and 6 are computed, assuming the following units of measurements of the environmental indicators: TMEAD: 10% and NITRO: 100kgs/ha. Thus, parameter estimates presented in Table 2 associated with TMEAD and NITRO must be interpreted with these new units of measurement in mind.
Figure 5.a : WTP for landscape degradation (Tmead)

Figure 5.b : WTP for livestock nitrogen emission (Nitro)

Note: In both figures, dots represents WTP estimated for each observation of the data sample, while lines represent nonparametric estimation of the WTP function with 95% confidence bounds.
An inspection of all these figures 5 and 6 clearly indicates that the relationships linking the willingness to pay to the pollution indicators are highly nonlinear for specific ranges of the values taken by $z_1$ and $z_2$. Up to a certain threshold that is significantly different from zero, the derivative of the hedonic price function with respect to the "landscape degradation" indicator "TMEAD" is rather small relative to average observed house prices, but exhibiting a marked and sharp decline when the share of permanent pasture converted into tillable land increases from 20 to 40%. Then the willingness to pay tends to flatten when this indicator takes values greater than 40%. In addition, the examination of figures 5.a and 6.a clearly shows that the relationship landscape degradation exhibit variations in the degree of curvature, which change with the values taken by the indicator TMEAD.

On the other hand, a different pattern seems to take place for WTP estimates associated with livestock nitrogen emissions (NITRO). An examination of figures 5.b and 6.b reveals that the relationship between the mean WTP function and this variable is steep and convex for small values of nitrogen emissions until it reaches 80kgs per hectare of arable land. Then, the mean WTP function for nitrogen emissions tends towards an asymptotic value that is equal to 7% of the residential house prices.

It is interesting to compare these WTO estimates obtained with the nonparametric model specification with similar ones obtained with a parametric specification. If we perform this exercise with model specifications (M1) and (M3) estimated in this work, we note that the WTP estimates obtained with the two model specifications are comparable and very similar for large values of the two environmental indicators. Let us take for instance the case of the landscape degradation indicator: WTP estimate in model (M1) is constant and equal to 3% for the house prices (assuming a 10 percent point change in the share of pasture land converted into tillable land), while the estimate obtained with model (M3) is equal to 3.5% when TMEAD is greater than 50%. Similar conclusions can be reached for WTP estimates for livestock nitrogen emissions.
Figure 6.a: WTP for landscape degradation (expressed as a percentage of house prices)

Figure 6.b: WTP for livestock nitrogen emission (expressed as a percentage of house prices)

Note: In both figures, dots represents WTP estimated for each observation of the data sample
6 Concluding remarks

The main objective of this paper has been to show the relevance of non- and semi-parametric models to study the relationship between agricultural pollution and property values. For this purpose, four model specifications - one parametric and three non- and semi-parametric - were applied to the prices of residential houses in a livestock intensive region of France in order to study the influences of landscape degradation and livestock nitrogen emissions on the house prices. Using appropriate specification tests, we conclude that a nonparametric additive model structure is the most appropriate model to explain the relationships between property values and agricultural pollution. Results on the willingness to pay estimates for agricultural pollution seem reasonable, conformed with a priori expectations, in line with estimates obtained using a parametric (semi-log) model specification but enhancing substantial nonlinearities.

Although the application of these various nonparametric models to an agricultural-related hedonic pricing case looks promising to capture complex nonlinearities, it is still too early to give a definitive appreciation on its merits. Further works and applications to other agricultural related situations are needed. In this vein, it would for instance be fruitful to analyse and quantify the role of positive and negative agricultural amenities in a common (nonparametric) model framework (Ready and Abdalla, 2005) and to compare it to a more conventional parametric modeling approach.
References.


Appendices

A  Bandwidth choices for the first step

![Graph](image)

Figure 2: Kondo and Lee (2003) cross-validation criterion.

B  Specification test results

B.1  Parametric versus Nonparametric (Horowitz and Spokoiny (2001))

The test is based on the distance $S_h(\gamma)$ between the kernel estimation of $m(Z)$ and the kernel-smoothed estimation of the parametric regression $M(Z, \gamma)$.

\[
S_h(\gamma) = \sum_{i=1}^{n} \left( \tilde{m}_h(Z_i) - \tilde{M}_h(Z_i, \gamma) \right)^2
\]  

(17)

were $\tilde{M}_h(Z_i, \gamma) = \sum_{j=1}^{n} W_h(Z_i, Z_j)M(Z_j, \gamma)$ is the kernel-smoothed parametric estimator with kernel weight $W_h(\cdot, \cdot)$. The test statistic, $T^*$ is then
computed with a rate-optimal and adaptive bandwidth, based on set of bandwidth values $h \in H_n$, and is centered and studentized:

$$T^* = \max_{h \in H_n} T_h$$

$$= \max_{h \in H_n} \frac{S_h(\gamma) - \widehat{N}_h}{\widehat{V}_h}$$

As for the bandwidth choice set, we select the bandwidth $h = c \times n^{-\frac{1}{6}}$ where $c$ belongs to the set $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}$ and with normalized variables. Table 5 reports the sensitivity analysis made with this set.

Table 5: Horowitz and Spokony tests statistics for various choices of the bandwidth

<table>
<thead>
<tr>
<th>$c$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.1</th>
<th>1.3</th>
<th>1.5</th>
<th>1.7</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_h$</td>
<td>6.107</td>
<td>6.025</td>
<td>4.083</td>
<td>2.721</td>
<td>2.021</td>
<td>1.648</td>
<td>1.166</td>
<td>0.774</td>
<td>0.513</td>
<td>0.335</td>
</tr>
</tbody>
</table>

B.2 Additive versus Nonparametric (Gozalo and Linton (2001))

Under $H_0$, it is assumed that for some parameters $\gamma \in \Gamma$, $G(m(Z)) = G(M(Z; \gamma)) = \gamma_0 + \sum_{d=1}^{D} m_d(Z_d)$, where $G$ belongs to a family of transformations. The four tests statistics proposed are:

$$\tau_0 = \frac{1}{n} \sum_{i=1}^{n} \left( H(\widehat{m}(Z_i)) - H(\widehat{M}(Z_i; \gamma)) \right)^2 \pi(Z_i)$$

$$\tau_1 = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i \left( \widehat{m}(Z_i) - \widehat{M}(Z_i; \gamma) \right) \pi(Z_i)$$

$$\tau_2 = \frac{1}{n^2 h_d^2} \sum_{i \neq j} K_{ij} \widehat{u}_i \widehat{u}_j \pi(Z_i) \pi(Z_j)$$

$$\tau_3 = \frac{1}{n} \sum_{i=1}^{n} (\widehat{u}_i^2 - \widehat{u}_i)^2 \pi(Z_i)$$

35
where the function $H$ can be the link function $G$ or simply $Id$; $\hat{u}_i = Y_i - \hat{m}(Z_i)$ are the unrestricted residuals, $\tilde{u}_i = Y_i - \tilde{M}(Z_i, \gamma)$ are the additive (restricted) residuals, $\pi(\cdot)$ is a weighting function and $K_{ij} = K((X_i - X_j)/h_n)1(Z_i = Z_j)$.

All the estimated functions $\hat{m}(Z_i)$, $\tilde{M}(Z_i; \gamma)$, and thus the residuals $\hat{u}_i$, and $\tilde{u}_i$ depend on smoothing parameters. Following Gozalo and Linton (2001), we use the same bandwidth for the restricted and unrestricted estimators. This bandwidth is defined as $h_n = c \times n^{-\frac{1}{4}}$ with $c \in \{0.5, 1.0, 1.5, 2.0\}$ and with normalized variables. We report in table 6 the results for the test statistics $\tau_0$, $\tau_1$ and $\tau_3$.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$\tau_0$</th>
<th>$\tau_1$</th>
<th>$\tau_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.5731766</td>
<td>1.5734135</td>
<td>1.5736503</td>
</tr>
<tr>
<td>1.0</td>
<td>0.3353748</td>
<td>0.33475074</td>
<td>0.33412665</td>
</tr>
<tr>
<td>1.5</td>
<td>0.14924605</td>
<td>0.14882588</td>
<td>0.14840571</td>
</tr>
<tr>
<td>2.0</td>
<td>0.037371477</td>
<td>0.037191986</td>
<td>0.037012494</td>
</tr>
<tr>
<td>2.5</td>
<td>0.022026428</td>
<td>0.021377212</td>
<td>0.020727995</td>
</tr>
</tbody>
</table>

### B.3 Single index versus Nonparametric (Fan and Li (1996))

Following the procedure proposed by Fan and Li (1996), the test $T^c$ is based on U-statistics constructed as moments of the residuals $\hat{u}_i = Y_i - \hat{G}(\gamma'Z_i)$ of the Single Index regression:

\[
T^c = \frac{nh^{d/2} I^c_n}{\sqrt{(2)\sigma_c}}
\]

with

\[
I^c_n = \frac{1}{n(n-1)h^d} \sum_{i=1}^{n} \left( \hat{u}_i \hat{f}(\gamma'Z_i) \right) \left( \sum_{j \neq i} (\hat{u}_j \hat{f}(\gamma'Z_j)) K_{ij} \right)
\]

We should note here a smoothing parameter $h_\gamma$ is used for the density estimator $\hat{f}(\gamma'Z_i)$ and another in the remaining kernel elements $K_{ij} = K((\gamma'Z_i - \gamma'Z_j)/h)$. For the former we use the usual bandwidth for nonparametric unidimensionnal density estimator $h_\gamma = c \times n^{-\frac{1}{4}}$, while we choose
Table 7: Grid search on Fan and Li (1996) tests statistics for bandwidth $h_{\gamma}$ and $h$

<table>
<thead>
<tr>
<th>$c$</th>
<th>$h_{\gamma}$</th>
<th>$h$</th>
<th>$T^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.021671435</td>
<td>0.014786302</td>
<td>5.8505905</td>
</tr>
<tr>
<td>0.6</td>
<td>0.13002861</td>
<td>0.088717814</td>
<td>3.3423753</td>
</tr>
<tr>
<td>1.1</td>
<td>0.23838578</td>
<td>0.16264932</td>
<td>3.4405514</td>
</tr>
<tr>
<td>1.6</td>
<td>0.34674296</td>
<td>0.23658084</td>
<td>3.3586483</td>
</tr>
<tr>
<td>2.1</td>
<td>0.45510013</td>
<td>0.31051235</td>
<td>3.2614862</td>
</tr>
<tr>
<td>2.6</td>
<td>0.56345730</td>
<td>0.38444386</td>
<td>3.3344222</td>
</tr>
<tr>
<td>3.1</td>
<td>0.67181448</td>
<td>0.45837537</td>
<td>3.5194407</td>
</tr>
<tr>
<td>3.6</td>
<td>0.78017165</td>
<td>0.53230688</td>
<td>3.6740766</td>
</tr>
<tr>
<td>4.1</td>
<td>0.88852882</td>
<td>0.60623839</td>
<td>3.7213201</td>
</tr>
<tr>
<td>4.6</td>
<td>0.99688600</td>
<td>0.68016990</td>
<td>3.6568200</td>
</tr>
<tr>
<td>5.1</td>
<td>1.1052432</td>
<td>0.75410142</td>
<td>3.5103628</td>
</tr>
<tr>
<td>5.6</td>
<td>1.2136003</td>
<td>0.82803293</td>
<td>3.3173418</td>
</tr>
<tr>
<td>6.1</td>
<td>1.3219575</td>
<td>0.90196444</td>
<td>3.1064701</td>
</tr>
<tr>
<td>6.6</td>
<td>1.4303147</td>
<td>0.97589595</td>
<td>2.8970483</td>
</tr>
<tr>
<td>7.1</td>
<td>1.5386719</td>
<td>1.0498275</td>
<td>2.7001645</td>
</tr>
<tr>
<td>7.6</td>
<td>1.6470290</td>
<td>1.1237590</td>
<td>2.5209337</td>
</tr>
<tr>
<td>8.1</td>
<td>1.7553862</td>
<td>1.1976905</td>
<td>2.3606013</td>
</tr>
<tr>
<td>8.6</td>
<td>1.8637434</td>
<td>1.2716220</td>
<td>2.2181650</td>
</tr>
<tr>
<td>9.1</td>
<td>1.9721006</td>
<td>1.3455535</td>
<td>2.0915016</td>
</tr>
</tbody>
</table>

$h = c \times n^{-\frac{1}{4}}$ for the latter according to theoretical guidance given by Fan and Li (1996). A grid search was performed on $c$. We report in table 7 the corresponding values of the Fan and Li's test statistic.