MÉTHODES SEMIPARAMÉTRIQUES EN ÉCONOMÉTRIE APPLIQUÉE : UNE APPLICATION AUX PRIX HEDONIQUES

Michel SIMIONI

INRA-ESR, Toulouse.

Introduction

La méthode des prix hédoniques appliquée aux prix des maisons est l'une des méthodes couramment utilisées par les économistes pour évaluer les pertes ou les gains monétaires liés à la qualité de l'environnement (voir **[6]** pour un tour d'horizon). Cette méthode repose sur l'idée que la variabilité observée du prix des maisons selon leur localisation peut être utilisée pour estimer la valeur que les consommateurs attribuent à un changement dans la qualité de leur environnement. Une maison y est ainsi considérée comme un ensemble de caractéristiques (nombre de pièces principales, âge, …) qui peut inclure des caractéristiques propres à l'environnement de la maison telles que la qualité de l'air (voir, par exemple, **[3]**) ou la qualité de l'eau (voir, par exemple, **[16]**). Le différentiel de prix entre des maisons de caractéristiques environnementales différentes peut alors constituer une information sur le prix implicite, ou prix hédonique, de cette caractéristique. En effet, il est possible d'envisager que, lorsque la qualité de l'environnement varie et lorsque les consommateurs préfèrent une meilleure qualité, le prix d'une maison sera, *ceteris paribus*, affecté par le niveau de qualité de l'environnement. L'information sur la qualité sera reflétée par le prix.

La méthode des prix hédoniques a été initiée dans les années 1920 par Waugh lors d'une étude de l'impact de la qualité des asperges sur leur prix (**[24]**). Un cadre formel à cette méthode a été fourni en termes d'équilibre partiel dans les années 1970 par Rosen (**[21]**). Jusqu'à récemment, la méthode des prix hédoniques était vivement critiquée quant aux problèmes d'identification des préférences des consommateurs et des coûts des producteurs que posait sa mise en œuvre en utilisant une approche s'appuyant sur le choix d'une forme paramétrique pour résumer la relation liant le prix d'une maison à ses caractéristiques, ou fonction de prix hédonique. Ekeland, Heckman et Nesheim (**[7]** et **[8]**) ont montré récemment qu'il s'agissait là d'un faux problème. Ils montrent que les auteurs proposaient l'utilisation d'approximations paramétriques linéaires de la fonction de prix hédonique qui n'englobaient pas toute l'information contenue dans le modèle économique sous-jacent. L'utilisation de toute cette information permet, en effet, d'établir un résultat d'identification non paramétrique des préférences et des coûts, c'est à dire, très général et ne dépendant pas d'un choix particulier quant à la forme de la fonction de prix hédonique. De plus, ce résultat est valable même lorsqu'on ne dispose de données que sur un seul marché.

Dans cet article, nous allons nous intéresser à la première étape de la méthode des prix hédoniques, à savoir : l'estimation de la fonction de prix hédonique. Trois modèles sont proposés : le modèle linéaire usuel, un modèle partiellement linéaire et un modèle à indice simple. Notre objectif est de montrer que les deux derniers modèles qui sont usuellement appelés semi paramétriques et qui ont été amplement

discutés lors de la conférence de Michel Delecroix (voir **[5]**), sont aisément implémentables à partir d'un jeu de données et que des tests de spécification récemment publiés peuvent être mis en œuvre pour juger de leur validité au regard des données.[1]

L'application qui est présentée ci-dessous, s'intéresse aux impacts des nuisances liés à l'élevage intensif et à la dégradation du bocage en Bretagne. Il s'agit là de résultats préliminaires d'une étude en cours de réalisation : voir **[2]**.

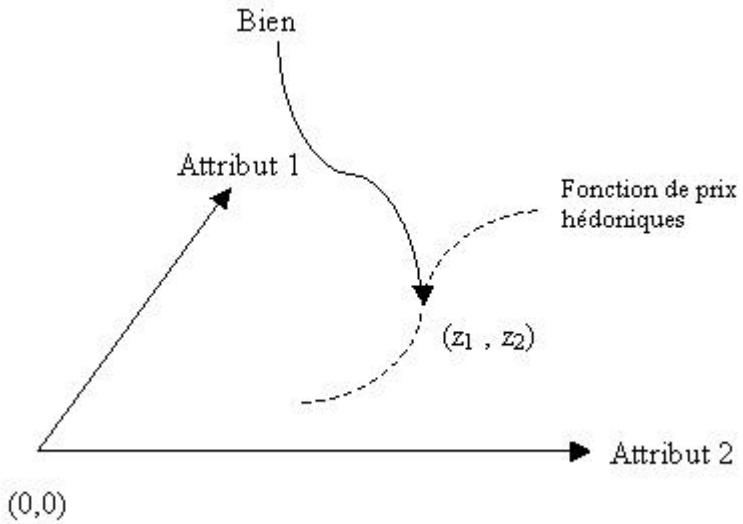
L'article est organisé comme suit. La première section présente les éléments théoriques qui sont à la base de la méthode des prix hédoniques. Les trois modélisations proposées de la fonction de prix hédonique et les méthodes d'estimation qui leur sont associées, sont détaillées dans une deuxième section. Cette section donne aussi diverses références quant à l'utilisation de telles modélisations lors de l'estimation de fonctions de prix hédoniques. La troisième section décrit les données utilisées dans l'application. Les résultats des estimations sont présentés et discutés dans une quatrième section. Divers tests de spécification y sont aussi présentés et mis en œuvre. Une dernière section conclue en donnant quelques pistes quant à la modélisation.

1. Les prix hédoniques : quelques éléments théoriques

Rosen (1974) a été le premier économiste à donner un cadre formel à la méthode des prix hédoniques. Le modèle de Rosen s'inscrit dans la littérature sur les biens différenciés. Un bien y est ainsi décrit par le vecteur de ses caractéristiques ou attributs. [Voir, par exemple, le graphique 1 où un bien est supposé être décrit par deux attributs notés (z_1, z_2).] Dans le cas d'une maison, ces attributs peuvent inclure les caractéristiques physiques de la maison (le nombre de pièces principales, par exemple), les caractéristiques de l'environnement social de celle-ci (revenu moyen dans la commune où est située la maison, par exemple), et les caractéristiques de l'environnement physique de la maison (qualité de l'eau, par exemple). Dans ce cadre, le prix d'une maison, noté P_i, est écrit comme une fonction de ces n caractéristiques, notées (z_{1i}, …, z_{ni}), soit :

$$P_i = P(z_{1i}, \ldots, z_{ni})$$

La dérivé de la fonction de prix hédoniques $P(.)$ par rapport au j-ème attribut, ou $\partial P/\partial z_j$, est alors interprétée comme le prix marginal implicite de cet attribut.



Graphique 1. Définition de la fonction de prix hédoniques

[1] Les estimations et les tests présentés dans cet article ont été réalisées à partir de programmes écrits pour le logiciel Gauss par l'auteur et utilisant des procédures mises à disposition du public des chercheurs par Joël Horowitz et Sokbae Lee dans le cadre de leur publication **[12]** (voir http:/www.faculty.econ. northwestern.edu/horowitz/papers).

Le modèle de Rosen repose sur une hypothèse de concurrence pure et parfaite sur le marché du bien étudié. La fonction de prix hédonique y est alors déterminée par l'équilibre entre les demandes des consommateurs et les offres des producteurs définies dans l'espace des attributs. Plus précisément, cette fonction est donnée par l'ensemble des points de tangences dans l'espace des attributs entre les fonctions de consentements à payer des consommateurs et les fonctions d'offre des producteurs[2]. [Ce lieu est décrit par la fonction tracée en pointillés sur le graphique 1.]

Dans le cas du marché des maisons, l'offre est supposée être fixe à court terme. Il s'agit là d'une hypothèse assez réaliste pour ce marché. La variation observée dans les prix des maisons va alors refléter les variations existantes dans les consentements à payer des consommateurs pour les différentes combinaisons d'attributs que représentent les maisons. Autrement dit, la fonction de prix hédoniques alloue les consommateurs en différentes localisations dans l'espace des attributs selon leurs préférences. En une combinaison donnée des attributs, le gradient de la fonction de prix hédoniques va donc donner la variation de prix qui va compenser le consommateur lorsqu'il est confronté à une qualité de l'eau qui se détériore. Des maisons dans une commune où la qualité de l'eau est médiocre, doivent avoir un prix plus faible pour pouvoir attirer des acheteurs potentiels. Remarquons alors qu'en identifiant les consentements à payer des consommateurs en tout point de l'espace des attributs, la méthode des prix hédoniques permet une analyse en termes de bien être de ces consommateurs lorsque la valeur d'un des attributs change.

2. Modèles et estimateurs

Cette section décrit les modèles qui vont être utilisés par la suite pour étudier la relation entre le prix d'achat d'une maison, variable que nous notons P, et les variables décrivant les caractéristiques d'une maison, représentées par le vecteur X, et les indicateurs de pollution dans un voisinage de la maison, représenté par le vecteur Z. Nous nous intéressons plus particulièrement à l'espérance conditionnelle de P sachant X et Z que nous noterons

$$m(x, z) = \mathrm{E}(P \mid X=x, Z=z) \qquad (1)$$

Celle-ci exprime le prix moyen payé pour une maison quand les caractéristiques de cette maison ont pour valeurs celles données dans le vecteur x, et quand les indicateurs de pollution prennent celles données par le vecteur z. Elle définira la fonction de prix hédonique.

2.1. Modèle linéaire

Dans le cas du modèle linéaire, l'espérance conditionnelle est exprimée comme une fonction linéaire, soit :

$$m(x,z) = x'\beta + z'\gamma \qquad (2)$$

où β et γ sont deux vecteurs de paramètres à estimer. Dans l'application, ils seront estimés par moindres carrés ordinaires. Il s'agit là du modèle le plus couramment utilisé dans la littérature empirique utilisant la méthode des prix hédoniques, que ce soit sous la forme linéaire présentée ci-dessus ou sous des formes paramétriques plus élaborées de type log-log, semi-log, etc. Plusieurs études proposent l'utilisation de transformations de type box-cox pour pallier à un choix trop restreint quant à la transformation de la variable à expliquer et de certaines des variables explicatives. Une discussion quant au choix parmi toutes ces formes fonctionnelles paramétriques est donnée dans Cropper, Deck et McConnell (1988) (**[4]**).

[2] Voir Rosen (1974) ou Ekeland, Heckman et Nesheim (2002 et 2003) pour plus de détails.

2.2. Modèle partiellement linéaire

Dans le cadre d'un modèle partiellement linéaire, le vecteur des variables explicatives est partitionné en deux vecteurs ne possédant aucune variable commune. Le premier ensemble de variables est supposé entrer de façon linéaire dans l'espérance mathématique alors que le second entre de façon non linéaire dans celle-ci. Autrement dit, le modèle est de la forme suivante

$$m(x, z) = x'\beta + H(z) \qquad (3)$$

où β est un vecteur de paramètres et $H(.)$ est une fonction inconnue. L'identification du vecteur de paramètres β requiert qu'aucune des variables composant le vecteur X ne soit parfaitement prédictible à partir des variables composants le vecteur Z. Un estimateur du vecteur β peut être obtenu en observant que (2) implique que

$$P - E(P|Z=z) = (X - E(X|Z=z))'\beta + V \qquad (4)$$

où V est une variable aléatoire non observée telle que $E(V|X=x, Z=z) = 0$. Un estimateur de β peut ainsi être obtenu par moindres carrés linéaires en estimant le modèle (4) où les quantités $E(P|Z=z)$ et $E(X|Z=z)$ ont été remplacés par leurs estimations non paramétriques. On peut alors montrer que l'estimateur de β, noté β_n, converge à la vitesse $n^{-1/2}$ et est asymptotiquement distribué selon une loi normale (**[20]**). La fonction $H(.)$ peut alors être estimée via la régression non paramétrique de $P - X'\beta_n$ sur Z.

Ce type de modélisation a été appliquée, entre autres, à l'évaluation de l'impact de la proximité de sites polluants dans onze agglomérations du Massachusetts par Stock (1991) et à celle de la présence de hauts fourneaux de l'industrie du cuivre dans la ville de Tacoma dans l'état de Washington par McMillen et Thorsnes (2000) (voir **[23]** et **[17]**). Une étude des performances de cette modélisation comparée à celles de formes fonctionnelles paramétriques usuelles figure dans Anglin et Gencay (1996) (**[1]**).

2.3. Modèle à indice simple

Dans le cadre d'un modèle à indice simple, l'espérance conditionnelle est de la forme :

$$m(x, z) = G(x'\beta + z'\gamma) \qquad (5)$$

où β et γ sont deux vecteurs de paramètres inconnus et $G(.)$ est une fonction elle aussi inconnue. L'identification des vecteurs de paramètres β et γ , ainsi que de la fonction $G(.)$ repose sur les trois restrictions suivantes :

- Les vecteurs X et Z ne contiennent aucune variable constante.
- La valeur d'un des paramètres (soit du vecteur β, soit du vecteur γ) est fixée à un.
- Le vecteur X ou le vecteur Z contiennent au moins une variable continue dont le coefficient n'est pas nul.

Sous ces conditions, l'estimation des vecteurs β et γ peut être réalisé en remarquant que, lorsque toutes les composantes des vecteurs X et Z sont continues, alors

$$\beta \propto E\left[\frac{\partial p(X,Z)}{\partial G(X'\beta + Z'\gamma)} / \partial X\right] = -2\,E\left[P \frac{\partial p(X,Z)}{\partial X}\right] \qquad (6a)$$

et

$$\gamma \propto E[\, p(X,Z)\, \partial G(X'\beta + Z'\gamma) / \partial Z] = -2\, E[P\, \partial p(X,Z)/\partial Z] \quad (6b)$$

où $p(X,Z)$ représente la densité jointe de X et Z (la deuxième égalité est obtenue en intégrant par parties la première espérance). Les vecteurs β et γ peuvent donc être estimés à un paramètre d'échelle près en estimant le dernier terme des égalités précédentes. **[19]** propose de remplacer $p(.,.)$ par un estimateur non paramétrique et de remplacer l'espérance mathématique $E(.)$ par la moyenne empirique calculée sur toutes les observations. Cet estimateur direct des paramètres β et γ a été étendu au cas où certaines des composantes des vecteurs X et Z peuvent être discrètes dans **[14]**.

A notre connaissance, peu d'études utilisent une telle modélisation de la fonction de prix hédoniques. Signalons, néanmoins, l'étude de Pace (1995) qui propose une comparaison des performances relatives de trois modélisations : modèle linéaire, modèle à indice simple et modèle non paramétrique, à partir d'un échantillon américain de ventes de maison (voir **[18]**).

3. Les données

Les données utilisées ici proviennent de différentes sources dont la base de données des notaires sur les transactions immobilières en Bretagne et la Direction Régionale de l'Agriculture et de la Forêt de Bretagne. Une description complète de ces données est donnée dans **[15]**. De cette base, nous avons extrait les caractéristiques suivantes des maisons :

- Le prix payé par l'acquéreur d'une maison (exprimé en francs) se décompose de la manière suivante :

  Prix = prix hors taxe + TVA + droits d'enregistrement (fixes ou proportionnels)
  + taxe de publicité foncière + frais de notaire + frais d'agence

- Le nombre de pièces principales
- La surface du terrain mesurée en m^2.
- L'âge mesuré en années.

Deux variables ont été choisies comme indicateurs de la qualité de l'environnement, ou, plus précisément, du niveau de pollution, dans la commune où a eu lieu l'achat d'une maison :

- L'azote hors sol par hectare de Surface Agricole Utile : Il s'agit là de la quantité d'azote produite par l'élevage hors sol (c'est-à-dire, élevages de bovins, de volailles ou de porcins) rapportée à la surface agricole utile de la commune.
- La part des prairies temporaires : la surface des prairies temporaires est égale à la différence entre celle des prairies totales et celle des prairies naturelles. Cette surface est ramenée à la surface agricole utile de la commune.

La première de ces deux variables est aisément interprétable en termes de pollution, les élevages hors sol (surtout de porcins) étant source de nuisances vis à vis du voisinage (odeurs), de nuisances paysagères et de pollution des eaux (algues vertes, …) et de l'air. La deuxième variable correspond, quant à elle, à un phénomène d'intensification fourragère observé en Bretagne. La conversion des prairies naturelles en prairies temporaires ou prairies cultivées s'y est faite parallèlement à la suppression des haies (opérations de remembrement). La proportion de prairies temporaires est ainsi un indicateur de la densité du bocage. Plus elle est élevée, moins le bocage est dense.

Le tableau 1 donne quelques caractéristiques descriptives de ces variables sur l'échantillon étudié.

Tableau 1. Description des données

Variables :	Moyenne	Ecart-type	Maximum	Minimum
Prix (francs)	501768	222593	1066477	101015
Nombre pièces principales	4.43	1.35	7	1
Superficie du terrain (m^2)	1793.59	2550.62	21880	102
Age	47.84	42.02	298	0
Azote hors sol par ha de SAU	45.17	51.12	339.48	0
Pourcentage de prairies temporaires	29.42	9.97	70.10	0
Effectifs = 2092 observations				

4. Résultats des estimations

4.1. Commentaires

4.1.1. Quelques détails techniques

Comme nous l'avons indiqué ci-dessus, le modèle linéaire a été estimé en utilisant la technique des moindres carrés ordinaires. Dans le cadre du modèle partiellement linéaire, les espérances conditionnelles des différentes variables entrant dans la partie linéaire du modèle, étant données la quantité d'azote hors sol et le pourcentage de prairies temporaires ont été estimées par la méthode du noyau de convolution. Le noyau choisi est un noyau normal bivarié dont la matrice de covariances est égale à la matrice des covariances empiriques des deux indicateurs de pollution, et la fenêtre est calculée en employant la règle de Silverman (1986), c'est à dire qu'elle est égale au produit d'une constante et de la taille de l'échantillon élevée à la puissance –1/6 (**[22]**). La fonction *H(.)* est estimée en utilisant aussi la méthode du noyau.

L'identification des paramètres et de la fonction *G(.)* dans le cadre du modèle à indice simple, nécessite la fixation de la valeur d'un des paramètres à un. Nous avons donc fixé la valeur du paramètre correspondant au nombre de pièces principales à un. Les estimations des paramètres des vecteurs b et g sont obtenues par la technique exposée ci-dessus. Les variables des vecteurs X et Z sont d'abord standardisées. Leur densité jointe est estimé par la méthode du noyau de convolution en utilisant un noyau quadratique et une fenêtre égale à cinq fois la taille de l'échantillon élevée à la puissance –1/6 (voir **[14]** pour plus de justifications). La fonction *G(.)* est elle aussi estimée par la méthode du noyau, la fenêtre étant choisie en multipliant par deux la fenêtre obtenue en employant la règle de Silverman.

4.1.2. Commentaires

Les valeurs estimées des paramètres figurant dans les trois modèles présentés ci-dessus sont données dans le tableau 2. Les estimations des fonctions *H(.)* et *G(.)* sont tracées respectivement dans les graphiques 2 et 3. La lecture de ces différentes estimations suscite les commentaires suivant :

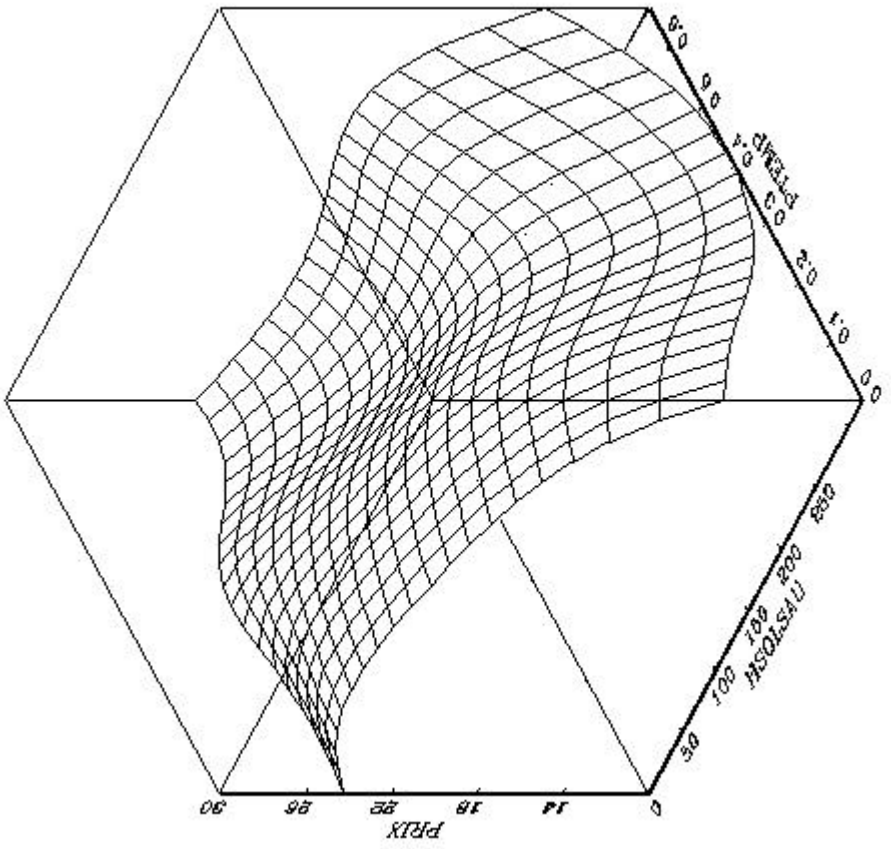
- Remarquons premièrement que les estimations des paramètres du modèle paramétrique sont toutes significatives et possèdent toutes leurs signes attendus. Ainsi, les indicateurs de pollution ont un effet négatif quant au prix d'une maison : la production d'une unité

supplémentaire d'azote par hectare par les élevages hors sol engendre une diminution du prix d'un maison de 449 francs et la présence de 1% de prairies temporaires dans une commune fait décroître ce prix de 2079 francs.

- Les valeurs estimées des paramètres du modèle partiellement linéaire sont de même grandeur et de même signe que celles comparables du modèle paramétrique. Cela semble confirmer l'hypothèse nécessaire à la mise en œuvre de la méthode d'estimation de Robinson (**[]**), à savoir, l'orthogonalité entre les variables de la partie linéaire du modèle et celles de la partie non linéaire. L'allure générale de l'estimation de la fonction *H(.)* fait, elle aussi, apparaître un effet négatif de chacun des indicateurs de pollution[3].

- Considérons maintenant les résultats pour le modèle à indice simple. Remarquons qu'aux effets de bord près, l'estimation de la fonction *G(.)* est pratiquement linéaire et croissante. Les signes des valeurs estimées des paramètres peuvent être ainsi être interprétés directement. Toutes ces valeurs sont significatives et ont leurs signes attendus, à l'exception notable de celle du paramètre associé à la quantité d'azote produite par les élevages hors sol qui n'est pas significativement différente de zéro.

Graphique 2. Modèle partiellement linéaire : résultat de l'estimation de la relation entre prix d'une maison et indicateurs de pollution



[3] Les retournements apparaissant sur les bords dans le graphique 2 sont fréquemment observés lorsqu'on a recours à l'estimateur du noyau pour estimer une fonction. Ces effets de bord peuvent être corrigés en utilisant d'autres types d'estimateurs non paramétriques tel que les polynômes locaux (voir **[9]**)

Graphique 3. Modèle à indice simple : Estimation de la fonction *G(.)*.

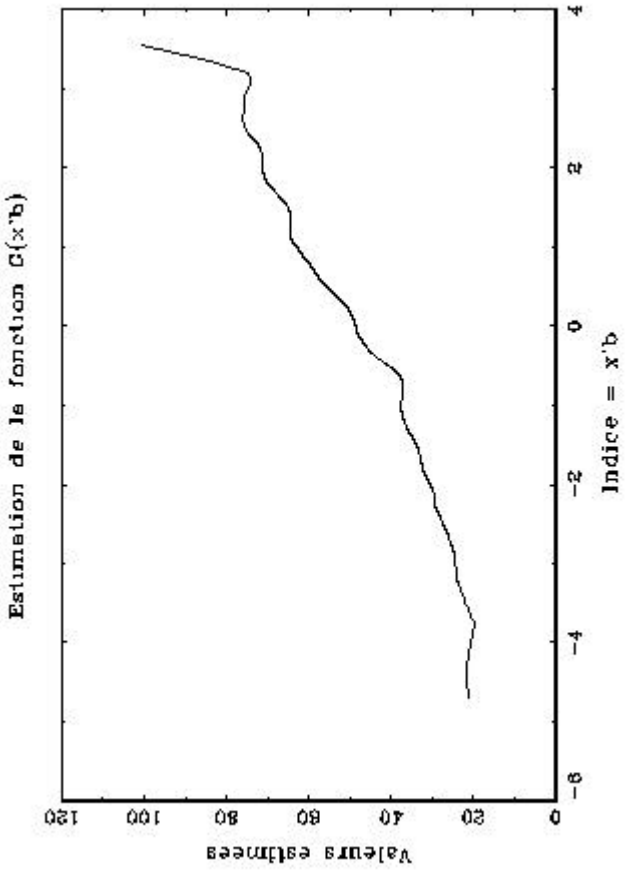


Tableau 2. Valeurs estimées des paramètres.

Variables explicatives	Modèle paramétrique	Modèle partiellement linéaire	Modèle à indice simple
Constante	32.6797 (2.0262)	*****	*****
Nombre de pièces principales	7.3929 (0.2894)	7.4313 (0.2851)	1.0000
Superficie du terrain	0.8433 (0.1439)	0.8322 (0.1414)	0.2334 (0.1129)
Age	-0.1801 (0.0094)	-0.1789 (0.0093)	-0.7420 (0.1716)
Azote hors sol par ha de SAU	-0.0449 (0.0074)	*****	-0.0143 (0.1377)
Pourcentage de prairies temporaires	-0.2079 (0.0378)	*****	-0.3068 (0.1748)

Note : les valeurs entre parenthèses sont les écarts-types estimés.

En résumé, les deux premières modélisations font apparaître un effet négatif significatif de chacun des deux indicateurs de pollution alors que la troisième modélisation ne met en évidence que l'effet négatif du pourcentage de prairies temporaires. Dans une optique d'économétrie appliquée, les résultats présentés ci-dessus nécessiteraient maintenant d'être complétés par la mesure de ces effets pour les deux modèles non linéaires via le calcul des dérivés des fonctions *G(.)* et *H(.)*. Les performances relatives des différents modèles pourraient être étudiées sur cette base. Au lieu de procéder de la sorte, nous allons maintenant effectuer différents test de spécification pour juger de la validité de ces modélisations au regard des données.

4.2. Tests de spécification

4.2.1. Test du modèle linéaire

Pour tester de la validité de représenter l'espérance conditionnelle de P étant donnée une réalisation des variables X et Z par un modèle linéaire comme dans (2), il est possible d'avoir recours au test proposé par Horowitz et Spokoiny dans **[13]**. Ce test repose sur la valeur de la «distance» entre l'estimateur non paramétrique de l'espérance conditionnelle (1) et l'estimateur paramétrique lissée, ou :

$$S_h(b_n, t_n) = \sum_{i=1,\ldots,N} [\, m_h(X_i, Z_i) - F_h(X_i, Z_i)]^2 \qquad (7)$$

où b_n et t_n représentent les estimateurs des moindres carrés ordinaires des vecteurs b et g, $m_h(X_i, Z_i)$ représente l'estimateur non paramétrique de l'espérance conditionnelle (1) et $F_h(X_i, Z_i)$, l'estimateur paramétrique lissé. Ces deux estimateurs sont définis comme suit :

$$m_h(X_i, Z_i) = \sum_{j=1,\ldots,N} w_h[\,(X_i, Z_i), (X_j, Z_j)\,]\, P_j \qquad (8a)$$

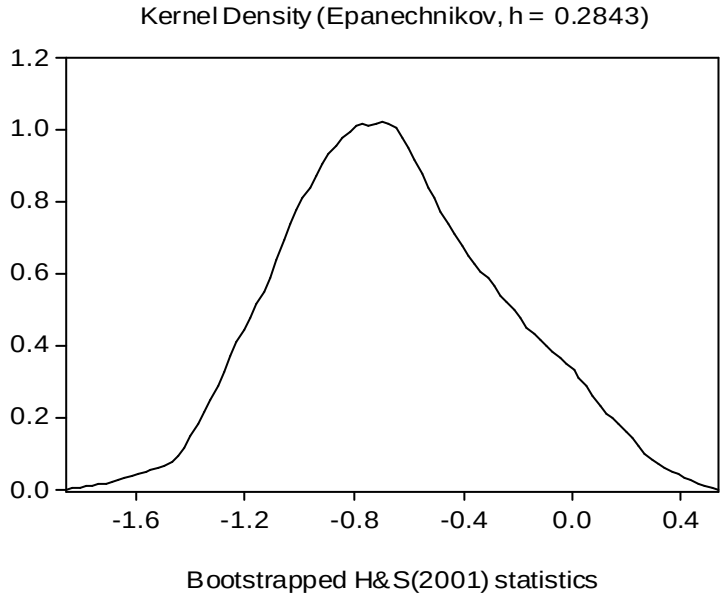
et

$$F_h(X_i, Z_i) = \sum_{j=1,\ldots,N} w_h[\,(X_i, Z_i), (X_j, Z_j)\,]\, [X_j' b_n + Z_j' t_n] \qquad (8b)$$

Les termes $w_h[\,(X_i, Z_i), (X_j, Z_j)\,]$ représentent les poids usuels dans les estimateurs non paramétriques de type noyau et sont fonction d'une fenêtre notée h.
La procédure de test repose alors sur la valeur maximale de la version centrée et réduite de la statistique S_n calculée pour différentes fenêtres. Sa loi limite n'étant pas connue, la distribution empirique de la statistique de test sous l'hypothèse nulle selon laquelle la représentation de l'espérance conditionnelle par un modèle linéaire est exacte, est simulée par bootstrap (pour plus de détails, voir **[13]**).

Dans le cadre de notre application, les estimations paramétriques utilisées sont celles présentées dans la deuxième colonne du tableau 1. Les variables sont toutes réduites pour éviter des problèmes dus aux différentes unités les définissant. La valeur maximale est choisie pour les valeurs suivantes de la fenêtre h : 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, et la procédure de bootstrap consiste en 120 réplications de la procédure de test.

Graphique 4. Densité des valeurs échantillonnées de la statistique de Horowitz et Spokoiny (2001)



Pour information, une estimation non paramétrique de la densité de la statistique bootstrappée est donnée dans la figure 4. La statistique de test obtenue, soit 3.6719, est largement plus grande que le quantile à 95% de la distribution des valeurs bootstrappées qui est égal à 0.2191. Le test de Horowitz et Spokoiny conclue donc en la rejection de l'hypothèse selon laquelle la représentation de l'espérance conditionnelle par un modèle linéaire est correcte.

4.2.2. Tests du modèle partiellement linéaire et du modèle à indice simple

Pour tester de la validité de représenter l'espérance conditionnelle de P étant donnée une réalisation des variables X et Z soit par un modèle partiellement linéaire comme dans (3), soit par un modèle à indice simple comme dans (4), il est possible d'avoir recours au test proposé par Fan et Li dans **[10]**. Ce test repose sur l'idée suivante. Considérons la variable aléatoire U_i définie comme suit :

$$U_i = Y_i - m(X_i, Z_i) \qquad (9)$$

Sous l'hypothèse nulle selon laquelle l'espérance conditionnelle (1) peut être représenté par un modèle partiellement linéaire du type (3) (ou un modèle à indice simple du type (4)), alors

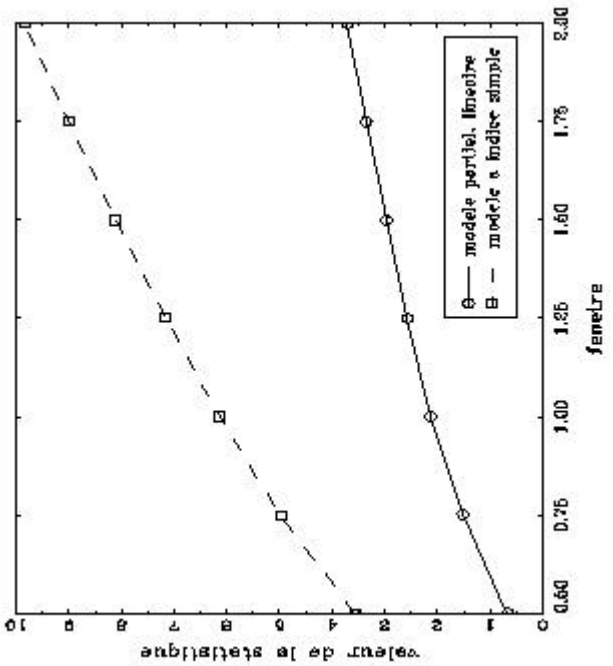
$$E(U_i \mid X_i, Z_i) = 0 \qquad (10)$$

Pour tester cette hypothèse nulle contre une alternative totalement non paramétrique, Fan et Li proposent d'utiliser la statistique suivante :

$$E[\,U_i\, E(U_i \mid X_i, Z_i)] = E[\,E(U_i \mid X_i, Z_i)^2] \qquad (11)$$

dont la valeur est positive ou nulle. En effet, cette statistique est nulle si et seulement si l'hypothèse nulle est vérifiée. Si U_i et $E(U_i \mid X_i, Z_i)$ étaient disponibles, la statistique de test pourrait être calculée à partir de la moyenne empirique : $N^{-1} \sum_{j=1,\ldots,N} U_i\, E(U_i \mid X_i, Z_i)$. L'équivalent empirique de (11) est alors construit à partir des résidus estimés et d'un estimateur non paramétrique de $E(U_i \mid X_i, Z_i)$. Pour éviter les problèmes posés par la présence de l'estimateur non paramétrique de la densité au dénominateur du précédent estimateur, la statistique de test n'est plus la moyenne arithmétique élémentaire mais une moyenne pondérée par la densité des variables explicatives (pour plus de détails, voir **[10]**). Fan et Li ont montré que, sous l'hypothèse nulle, cet équivalent empirique divisé par son écart-type est distribué selon la loi normale centrée et réduite.

Graphique 5. Résultats du test de Fan et Li 1996).



D'un point de vue pratique, le test de Fan et Li nécessite l'utilisation d'un noyau très lisse du type noyau normal. De plus deux fenêtres doivent être choisies : la première correspond à l'hypothèse nulle (modèle partiellement linéaire ou modèle à indice simple), la seconde à l'hypothèse alternative. Comme proposé par Horowitz et Lee dans **[12]**, après avoir centré et réduit les variables explicatives, nous avons choisi comme fenêtre $c\, n^{-1/5}$ sous l'hypothèse nulle et $c\, n^{-1/4}$ sous l'hypothèse alternative, c étant une constante. La théorie ne donnant aucune indication quant au choix de cette constante, nous avons effectué le test de Fan et Li pour plusieurs valeurs possibles de celle-ci, soit : 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2.
Les valeurs obtenues de la statistique de Fan et Li pour les sept choix de fenêtre et les deux hypothèses nulles sont représentées dans la figure 5. Ces valeurs indiquent clairement un rejet de l'hypothèse selon laquelle la représentation de l'espérance conditionnelle (1) par un modèle à indice simple est correcte. Les valeurs obtenues dans le cas de l'hypothèse nulle selon laquelle la représentation de l'espérance conditionnelle (1) par un modèle partiellement linéaire est correcte, ne permettent pas de conclure en un rejet clair et net de cette hypothèse.

Conclusion

Dans cet article, nous avons montré que des estimateurs semiparamétriques proposés dans la littérature pouvaient être aisément mis en œuvre dans le cadre de travaux en économétrie appliquée. Leur utilisation dans le cadre d'une application de la méthode des prix hédoniques a généré des résultats contrastés quant à l'impact d'indicateurs de pollution sur les prix des maisons en Bretagne. Bien entendu, l'analyse empirique devrait être approfondie via l'amélioration des spécifications empiriques utilisées : transformation de certaines des variables, introduction d'autres variables explicatives, etc. D'autres spécifications récemment présentées dans la littérature théorique telles que les modèles additifs généralisés (voir Horowitz, 2001) pourraient elles aussi être envisagées. Le problème qui se pose alors, est celui du choix de la spécification qui semble être la plus valide au regard des données. La dernière partie de l'article a aussi montré que des tests de spécification récemment construits dans la littérature théorique pouvaient être mis en œuvre pour répondre à cette question.

Références.

[1] Anglin, P.M., et R. Gencay (1996), "Semiparametric Estimation of a Hedonic Price Function", *Journal of Applied Econometrics*, 11, 633-648.

[2] Bontemps, C., Simioni, M., et Y. Surry (2002), "Semiparametric Assessment of the Effects of Neighborhood Land Uses on Residential House Values", en progrès.

[3] Chay, K.Y., et M. Greenstone (2000), "Does Air Quality Matter? Evidence from the Housing Market", document de travail, Department of Economics, University of California, Berkeley.

[4] Cropper, M.L., Deck, L.B., et K.E. McConnell (1988), "On the Choice of Functional Form for Hedonic Price Functions", *Review of Economics and Statistics*, 70, 668-675.

[5] Delecroix, M. (2003), "L'usage des "modèles linéaires généralisés" en analyse de la régression est-il encore indispensable aux statisticiens ?", *INSEE Méthodes*, ce numéro.

[6] Desaigues, B., et P. Point (1993), *Economie du patrimoine naturel : la valorisation des bénéfices de protection de l'environnement*, Economica, Paris.

[7] Ekeland, I., Heckman, J.J., et L. Nesheim (2002), "Identifying Hedonic Models", *American Economic Review*, 92: 304-309.

[8] Ekeland, I., Heckman, J.J., et L. Nesheim (2003), "Identification and Estimation of Hedonic Models", *Journal of Political Economy*, à paraître.

[9] Fan, J., et I. Gijbels (1996), *Local Polynomial Modelling and Its Application*, Chapman & Hall, Londres.

[10] Fan, Y., et Q. Li (1996), "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms", *Econometrica*, 64: 865-890.

[11] Horowitz, J.L. (2001), "Nonparametric Estimation of a Generalized Additive Models with an Unknown Link Function", *Econometrica*, 69, 499-513.

[12] Horowitz, J.L., et S. Lee (2002), "Semiparametric Methods in Applied Econometrics: Does the Models Fit the Data?", *Statistical Modelling*, 2, 3-22.

[13] Horowitz, J.L., et V.G. Spokoiny (2001), "An Adaptative, Rate-Optimal Test of a Parametric Mean-Regression Model against a Nonparametric Alternative", *Econometrica*, 69: 599-631.

[14] Horowitz, J.L., et W. Härdle (1996), "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariables", *Journal of the American Statistical Association*, 91, 1632-1640.

[15] Koïdou, C.D. (1999), *Impacts de l'agriculture sur les prix immobiliers*, mémoire du DEA ERNEA, Université Toulouse 1, Toulouse.

[16] Legett, C.G., et N.E. Bockstael (2000), "Evidence of the Effects of Water Quality on Residential Land Prices", *Journal of Environmental Economics and Management*, 39, 121-144.

[17] McMillen, D.P., et P. Thorsnes (2000), "The Reaction of Housing Prices to Information on Superfund Sites: A Semiparametric Analysis of the Tacoma, Washington, Market ", dans : T.B. Fomby et R. Carter Hill, *Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics*, JAI Press Inc., Stamford, Connecticut.

[18] Pace, R.K. (1995), "Parametric, Semiparametric, and Nonparametric Estimation of Characteristic Values Within Mass Assessment and Hedonic Price Models", *Journal of Real Estate Finance and Economics*, 11, 195-217.

[19] Powell, J.L., Stock, J.H., et T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients", *Econometrica*, 51, 1403-1430.

[20] Robinson, P.M. (1988), "Root-*N*-Consistent Semiparametric Regression", *Econometrica*, 56: 931-954.

[21] Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy*, 82: 34-55.

[22] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, Londres.

[23] Stock, J.H. (1991), "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits", dans : W.A. Barnett, J. Powell et G. Tauchen, éditeurs, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press, Cambridge, 77-98.

[24] Waugh, F.V. (1928), "Quality Factors Influencing Vegetable Prices", *Journal of Farm Economics*, 10, 185-196.