ESTIMATION DE DISTRIBUTIONS DE REVENU PAR DES METHODES NON PARAMETRIQUES

Denis Fougère[1], Daniel Verger[2]

1. Introduction

Cette communication a pour objet l'analyse non paramétrique des distributions de revenus pour les années 1979, 1984 et 1990. Pour ce faire, elle met en oeuvre deux types de méthodes :

- l'estimation des fonctions de densité de revenus par la méthode du noyau ;
- l'application des méthodes de régression non paramétrique pour l'estimation d'indicateurs d'inégalité.

Les sources statistiques utilisées sont les enquêtes DGI-INSEE sur les revenus fiscaux des ménages. L'étude s'intéresse à deux concepts de revenus, les revenus fiscaux et les revenus disponibles, et porte sur le sous-échantillon des ménages déclarant un revenu (disponible ou fiscal) strictement positif, pour lesquels ni la personne de référence, ni son conjoint éventuel ne sont agriculteurs ou indépendants non agricoles (en activité).

Les tailles des sous-échantillons analysés sont indiquées dans le Tableau 1.

Tableau 1 : Tailles des échantillons

Année d'enquête	1979	1984	1990
Revenus disponibles	18 793	20 186	14 690
Revenus fiscaux	18 767	20 136	14 648

[1] CNRS et CREST
[2] CREST-INSEE

Il est à remarquer que ces échantillons ont été tirés avec probabilités inégales : en particulier, les ménages aisés y sont fortement sur-représentés. De ce fait, toutes les estimations ici reproduites tiennent compte des pondérations rendant l'échantillon représentatif de la population de l'ensemble des ménages ordinaires.

Le revenu fiscal inclut les traitements et salaires, les pensions, retraites et rentes, les revenus de la propriété (soumis à l'impôt sur le revenu), ainsi que les bénéfices agricoles, industriels et commerciaux ou non commerciaux. Le revenu disponible est calculé en retirant de ce montant les impôts directs payés (impôts sur le revenu et taxe d'habitation) et en y ajoutant une estimation des prestations sociales (familiales, aides au logement, RMI, minimum vieillesse). Tous les revenus sont exprimés en francs constants 1990 ; l'actualisation est réalisée sur la base de l'indice des prix à la consommation.

2. Estimation des distributions de revenus par la méthode du noyau

Dans la mise en oeuvre de la méthode du noyau (ici appliquée à l'estimation de fonctions de densité), le praticien fait face à deux problèmes : le choix du noyau lui-même et le choix de la largeur de la fenêtre. Rappelons tout d'abord la forme générale d'un estimateur à noyau pour une fonction de densité *f*. Celui-ci s'écrit :

$$\hat{f}_n(y) = \frac{1}{h_n} \sum_{i=1}^{n} \rho_i \quad K\left(\frac{y - y_i}{h_n}\right)$$

où ρ_i et y_i sont respectivement la pondération et le revenu observé pour le ménage *i (i=1,... n)*, *n* la taille de l'échantillon, *K* la fonction noyau et h_n la largeur de la fenêtre. La fonction *K* doit vérifier :

$$K(u) \geq 0, K(u) < \infty, K(-u) = K(u), \forall u \in R,$$
$$et \int_R K(u)du = 1$$

Trois noyaux usuels seront utilisés dans notre étude :

- le noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1-u2)1|u| \leq 1$
- le noyau biweight : $K(u) = \frac{15}{16}(1-u2)21|u| \leq$

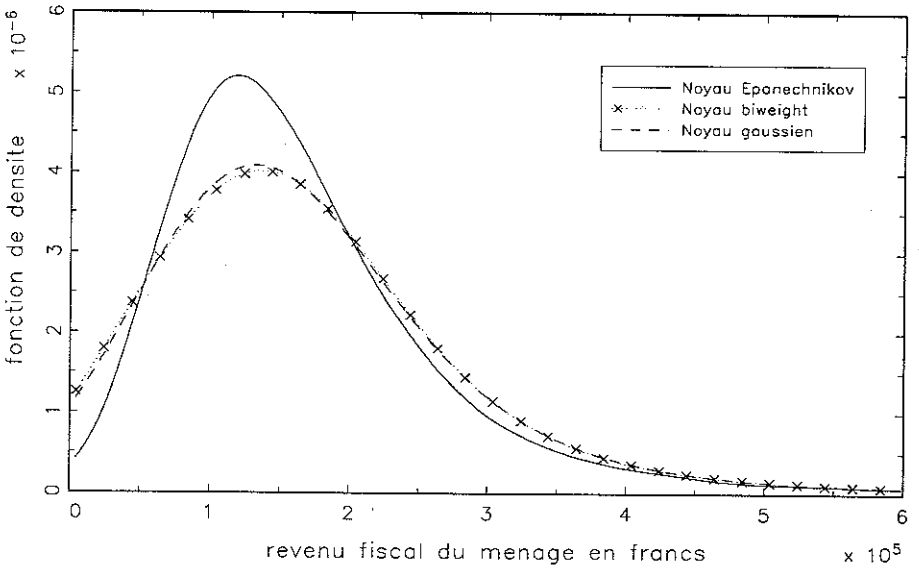
- le noyau gaussien : $K(u) = \frac{1}{\sqrt{2x}} \exp(-\frac{u^2}{2})$

Pour chacun de ces noyaux, nous avons choisi d'utiliser la largeur de fenêtre h_n déduite de la « règle du pouce » (rule-of-thumb) correspondante :

- pour le noyau d'Epanechnikov : $h_n \approx 1 \quad \hat{\sigma}_y \quad n^{-1/5}$,
- pour le noyau biweight : $h_n = 2{,}78 \quad \hat{s}_y \quad n^{-1/5}$,
- pour le noyau gaussien : $h_n = 1{,}06 \quad \hat{s}_y \quad n^{-1/5}$,

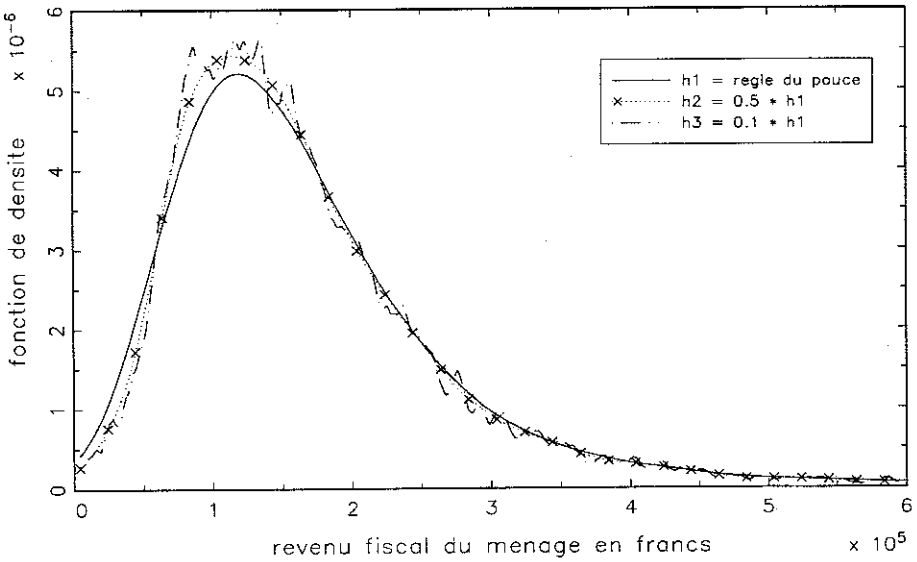
où $\hat{\sigma}_y$ est l'écart-type empirique de la variable analysée (ici, le revenu fiscal ou le revenu disponible). Le graphique 1 illustre la sensibilité de l'estimation au choix du noyau, la largeur de la fenêtre pour chacun des trois noyaux étant ici calculée à l'aide de la formule précédente. Les trois estimations obtenues pour les revenus fiscaux 1990 diffèrent assez notablement, en particulier dans le bas de la distribution. Une comparaison graphique (non reproduite ici) avec l'estimation par histogramme incite à choisir le noyau d'Epanechnikov pour le reste de l'analyse conduite dans cette section.

Graphique 1 : Estimation non-parametrique de la densite des revenus fiscaux 90 par la methode du noyau



La sensibilité de l'estimation non paramétrique à la largeur de la fenêtre est illustrée par le graphique 2, qui montre les résultats obtenus à l'aide de trois largeurs de fenêtre (fenêtre « règle du pouce », divisée par deux, divisée par dix) appliquées au noyau d'Epanechnikov. Bien évidemment, l'estimation est d'autant plus lissée que la fenêtre est large. Remarquons toutefois qu'une fenêtre plus étroite accroît la densité au mode de la distribution. Une approche plus rigoureuse consisterait ici à utiliser des méthodes de choix de fenêtre moins frustes (par exemple, méthodes de « plug-in » ou de validation croisée).

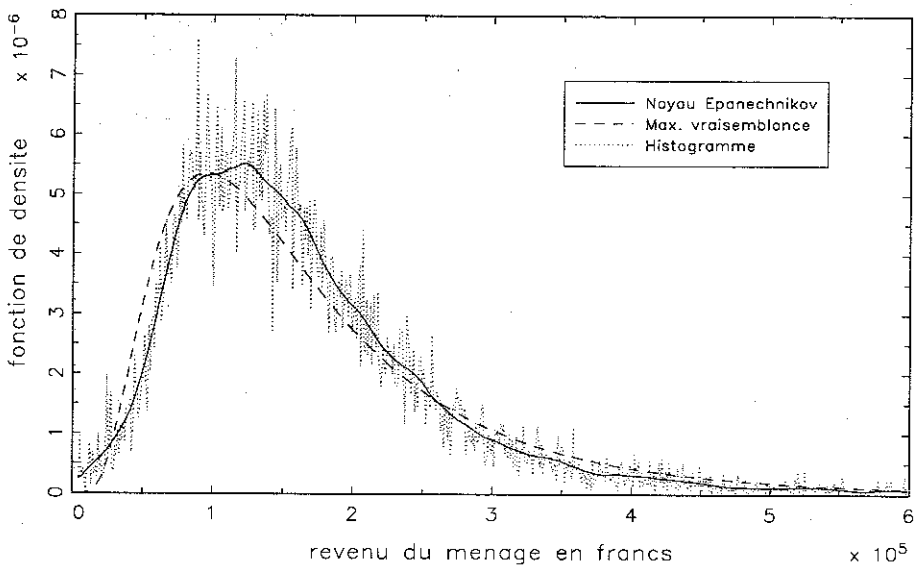
Graphique 2 : Estimation non-parametrique de la densite des revenus fiscaux 90 par la methode du noyau (Epanechnikov)



Les analyses paramétriques de la distribution des revenus font souvent l'hypothèse que les revenus suivent une loi log-normale. Le graphique 3 montre qu'une telle hypothèse, donnant lieu à une procédure d'estimation par maximisation de la vraisemblance, conduit à une sur-estimation de la densité dans la partie basse de la distribution et à une sous-estimation dans la partie intermédiaire. Cet exercice montre que l'estimation non paramétrique préalable par méthode du noyau peut guider utilement le praticien dans le choix d'une hypothèse paramétrique sur la loi de la variable étudiée.

Graphique 3 : Trois estimations de la densite

des revenus fiscaux 90

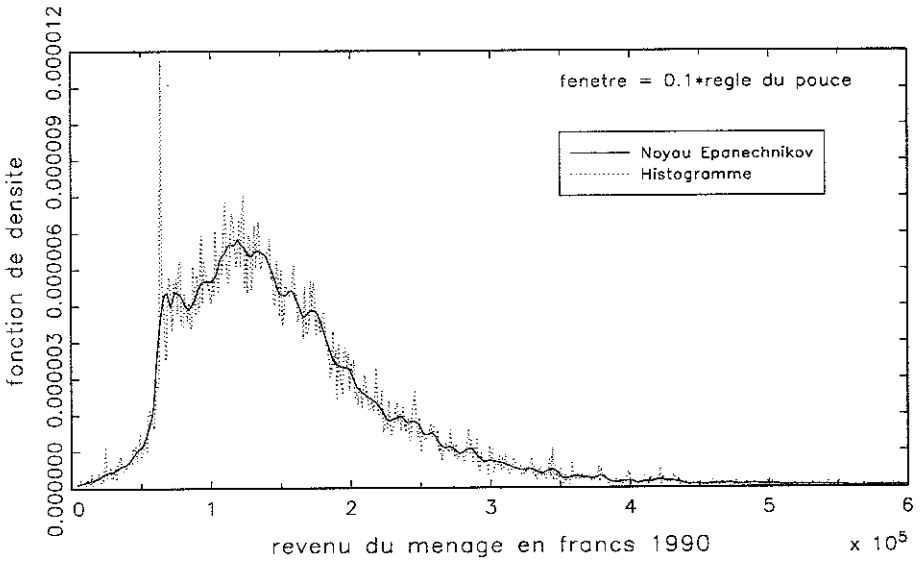


Le graphique 4 met l'accent sur la difficulté soulevée par l'existence d'une masse ponctuelle : celle-ci dans l'exemple choisi correspond à l'existence d'un minimum social, à savoir le minimum vieillesse. Non imposable, il a été imputé sur barème. En 1984, tous les foyers fiscaux ayant-droits se sont vus attribuer le même revenu disponible d'une valeur approximativement égale à 80 % du minimum vieillesse[3]. Une procédure identique a été appliquée pour les foyers fiscaux ayant droit au RMI en 1990. Ce type de minima sociaux crée une masse dans la distribution des revenus disponibles, au point correspondant au montant du seuil. Le graphique 4 montre qu'une estimation non-paramétrique réalisée avec une fenêtre pourtant étroite (égale à un dixième de la fenêtre « règle-du-pouce ») ne peut rendre compte de cette masse, à l'inverse d'une estimation par histogramme.

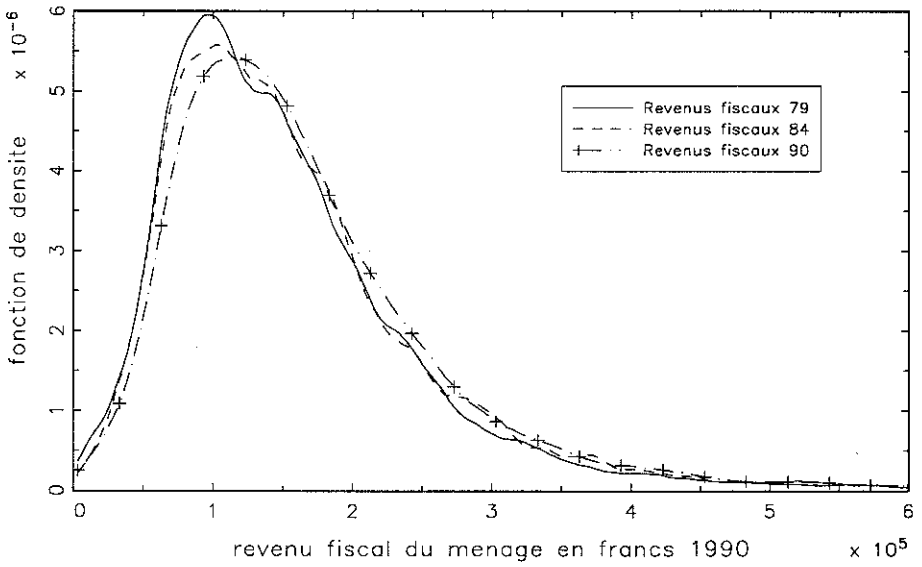
La méthode peut ensuite être appliquée pour l'analyse de l'évolution globale des revenus disponibles entre 1979 et 1990. Sur le graphique 5 sont reportées les estimations par méthode du noyau (Epanechnikov, avec largeur de fenêtre « règle-du-pouce ») des densités de revenus disponibles aux trois dates. De 1979 à 1990, on remarque un déplacement assez sensible de la densité vers la droite, déplacement qui toutefois ne concerne pas l'extrémité basse de la distribution. Par ailleurs, le mode s'est accru entre ces deux dates d'environ 20 000 Francs.

[3] valeur calculée pour retrouver les masses globales connues par ailleurs.

Graphique 4 : Estimation non-parametrique de la densite des revenus disponibles 84 par la methode du noyau

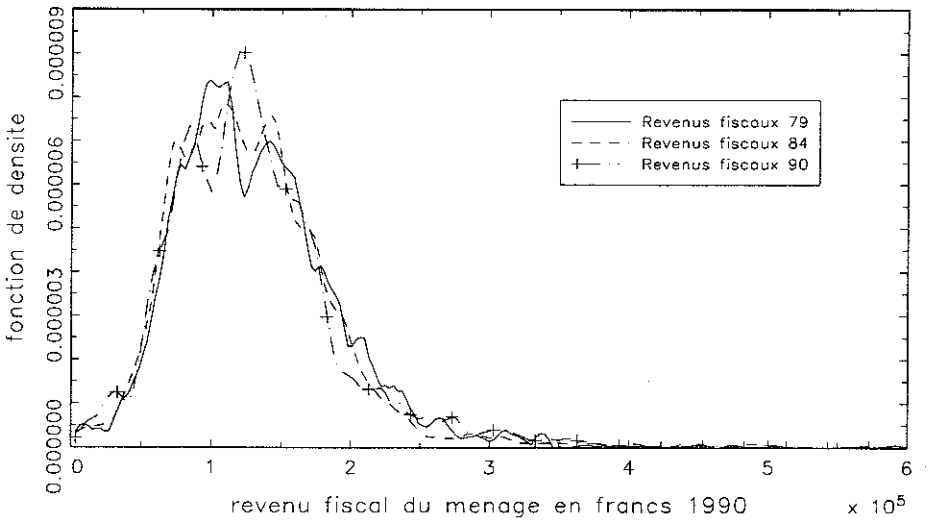


Graphique 5 : Estimation non-parametrique de la densite des revenus fiscaux par la methode du noyau (Epanechnikov)

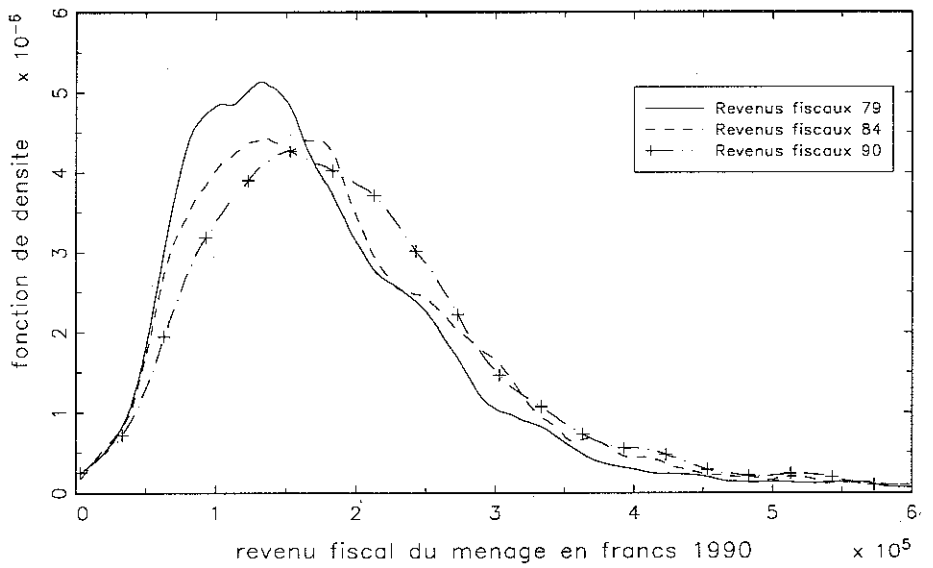


Une analyse plus fine selon l'âge de la personne de référence dans le ménage (cf. graphiques 6a, 6b, 6c) révèle que le déplacement le plus significatif entre 1979 et 1990 concerne les ménages d'âge intermédiaire (voir graphique 6b). Le déplacement est moins marqué, mais encore net, pour les ménages âgés (cf. graphique 6c), alors que les distributions des revenus disponibles des ménages jeunes sont quasiment identiques aux trois dates (cf. graphique 6a).

Graphique 6a : Estimations par noyau de la densite des revenus fiscaux (age du chef de menage compris entre 21 et 30 ans)

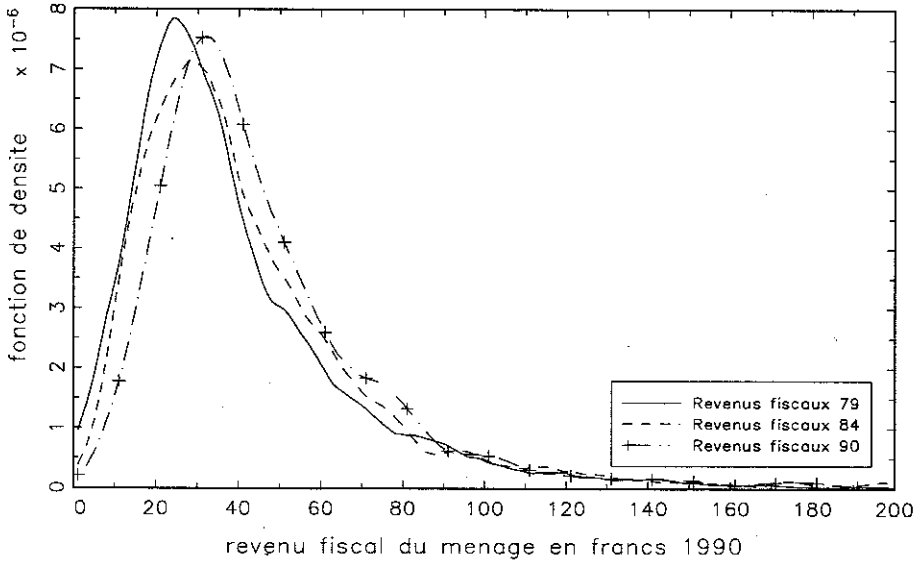


Graphique 6b : Estimations par noyau de la densite des revenus fiscaux (age du chef de menage compris entre 41 et 53 ans)



Graphique 6c : Estimations par noyau de la densite des revenus fiscaux

(age du chef de menage compris entre 66 et 85 ans)



3. Estimation non-paramétrique des indicateurs d'inégalité

Outre les distributions de revenus, les méthodes non-paramétriques peuvent également servir à estimer de manière robuste des indicateurs usuels d'inégalité, tels que l'indicateur de Theil ou celui d'Atkinson, en fonction des valeurs prises par les variables caractérisant les foyers fiscaux (par exemple, l'âge de la personne de référence au sein du ménage). Cette approche repose toutefois sur l'estimation préalable de l'espérance conditionnelle des revenus par des techniques de régression non paramétrique.

3. 1 Estimation non-paramétrique de l'espérance conditionnelle des revenus

Considérons par exemple, l'âge de la personne de référence du ménage *i* ($i = 1, \ldots n$), noté a_i, et supposons que le revenu de ce ménage soit défini par la relation de régression linéaire :

$$y_j = m(a_i) + e_i, \quad i = 1,\ldots n,$$

où les erreurs ε_i sont des variables aléatoires i.i.d. de moyenne nulle. $m(a_i)$ est l'espérance conditionnelle de y_i sachant a_i, soit :

$$m(a_i) = E(y_i|a_i), i = 1,\ldots,n.$$

Un estimateur non-paramétrique de l'espérance conditionnelle des revenus à l'âge a, notée m(a), est l'estimateur de Nadaraya-Watson défini comme :

$$\hat{m}_h(a) = \frac{\sum_{i=1}^{n} r_i K\left(\frac{a-a_i}{h}\right) y_i}{\sum_{i=1}^{n} r_i K\left(\frac{a-a_i}{h}\right)} \quad ,$$

où la largeur de fenêtre h peut être calculée, selon la « règle-du-pouce », à l'aide de la variance empirique $\hat{\sigma}_a$ des âges dans l'échantillon.

L'approche paramétrique habituelle consiste à spécifier une loi particulière, par exemple à faire l'hypothèse de log-normalité :

$$Y|a \sim N(m(a), s_a^2), \quad avec \quad Y = \log y.$$

En ce cas, on peut estimer $m(a) = E[Y|a]$ *et* $s_a^2 = \text{var}[Y|a]$ par maximisation de la vraisemblance, et en déduire une estimation de $E[y|a] = \exp[m(a) + s_a^2/2]$.

Les estimations de l'espérance conditionnelle des revenus fiscaux selon l'âge obtenues à l'aide de ces deux méthodes sont représentées sur le graphique 7a. La régression non-paramétrique permet d'obtenir un estimateur lissé, à l'inverse de l'approche paramétrique par maximum de vraisemblance. Il est à remarquer que la moyenne des revenus fiscaux croît jusqu'à 47 ans puis décroît au-delà. Son évolution n'est toutefois pas représentable par un polynôme de degré 2, comme on le fait parfois dans les approches paramétriques en posant :

$$E(y|a) = b_0 + b_1 a + b_2 a^2$$

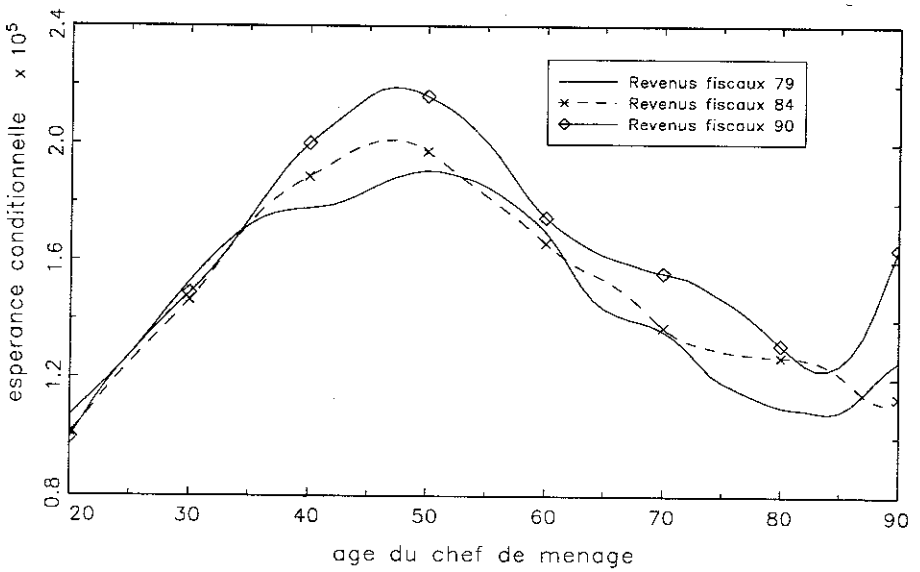
Il serait ici préférable de spécifier cette espérance conditionnelle à l'aide d'un polynôme d'ordre supérieur ou, éventuellement, d'une approximation linéaire par morceaux (par exemple, sur les intervalles [20, 47],[47,70] et [70,90] ans). Ici encore, l'approche non-paramétrique peut aider au choix d'une forme paramétrique

permettant de représenter correctement l'effet d'une variable explicative dans un modèle de régression.

GRAPHIQUE 7a : Estimations non-parametrique et parametrique de l'esperance conditionnelle des revenus fiscaux 84 selon l'age

Considérons maintenant les estimations des expériences conditionnelles des revenus fiscaux selon l'âge aux trois dates d'enquête (cf. graphique 7b). La moyenne de ces revenus a crû significativement entre 1979 et 1990 pour les ménages dont la personne de référence avait entre 40 et 55 ans, ou plus de 65 ans. Le revenu fiscal moyen n'a absolument pas varié pour les ménages les plus jeunes (moins de 33 ans). L'ensemble de ces résultats, ainsi que ceux de la sous-section suivante, ont été obtenus à l'aide d'un noyau biweight (avec une fenêtre de type « règle-du-pouce »).

Graphique 7b : Estimation non-parametrique de l'esperance conditionnelle des revenus fiscaux en fonction de l'age



3.2 Estimation non-paramétrique des indicateurs d'inégalité

L'analyse de l'inégalité des revenus (ou des salaires) est généralement fondée sur l'utilisation d'indicateurs d'inégalité (de « concentration ») tels que ceux de Theil et d'Atkinson. Si l'on s'intéresse à l'inégalité des revenus selon l'âge (de la personne de référence), ces indicateurs s'écrivent respectivement :

$$I_t = E(y|a) \times E(y \log y|a) - \log[E(y|a)],$$

et

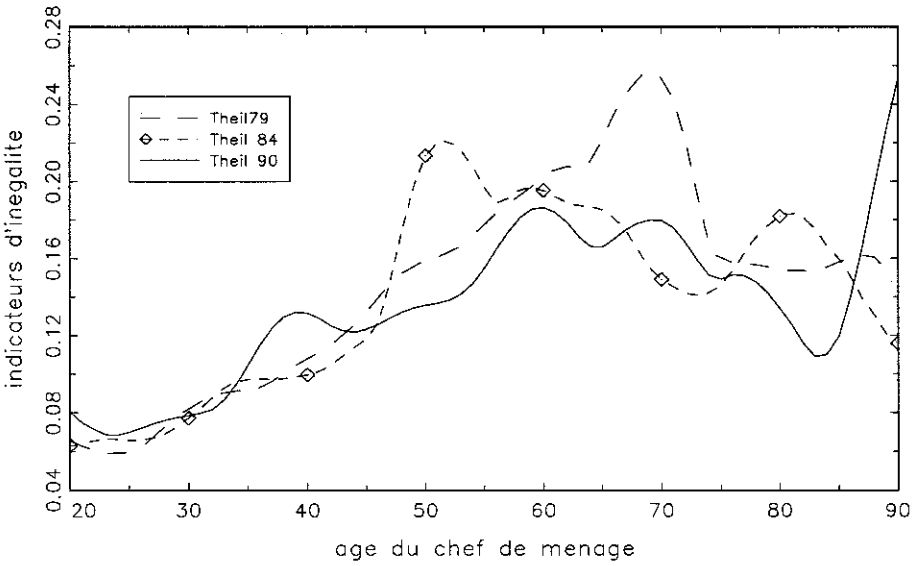
$$I_a(\epsilon) = 1 - \left\{ \left[E(y^{1-\epsilon}|a) \right]^{1/(1-\epsilon)} E(y|a)^{-1} \right\}.$$

Les expressions de ces indicateurs font intervenir les espérances conditionnelles des revenus (ou de fonctions des revenus) qui peuvent être estimées à l'aide des techniques de régression non-paramétrique présentées dans la sous-section précédente. Rappelons que l'indicateur d'Atkinson est d'autant plus sensible aux modifications dans le bas de la distribution que le paramètre ε est élevé. Dans les exemples présentés, nous avons donné deux valeurs à ce paramètre : 0,5 et 2. Il faut remarquer que $\varepsilon = 2$ est une valeur assez élevée par rapport à ce que l'on choisit habituellement : mais l'intérêt est ici précisément d'apprécier le comportement des estimateurs non-paramétriques dans le cas d'un indicateur très sensible.

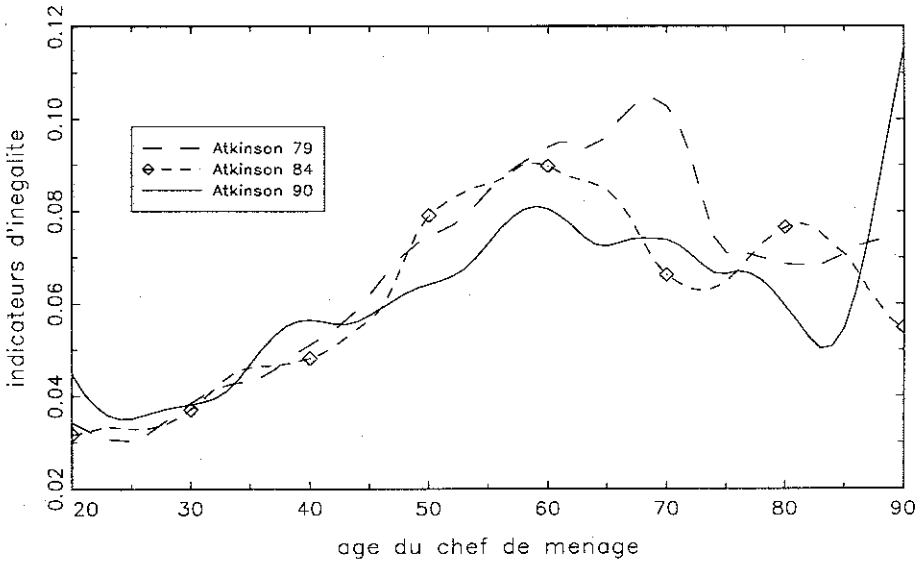
Les graphiques 8a, 8b, 8c représentent les estimations non-paramétriques des indicateurs d'inégalité de Theil et d'Atkinson pour les revenus disponibles aux trois

dates d'enquête. Le résultat le plus notable est la baisse des indicateurs d'inégalité aux âges compris entre 50 et 70 ans. Un problème apparaît toutefois pour les ménages jeunes (dont la personne de référence a moins de 33 ans) avec l'indicateur d'Atkinson $\varepsilon = 2$: à ces âges-là, les revenus seraient devenus plus inégaux en 1990. Comment interpréter ce dernier résultat ? Les évolutions du marché du travail constatées au cours de la dernière décennie (accroissement du chômage des jeunes et des moins qualifiés, augmentation du nombre d'emplois précaires, etc) le rendent certes vraisemblable. Toutefois, résiste-t-il à une analyse plus rigoureuse des données, et en particulier au traitement des valeurs « aberrantes » présentes dans le fichier ?

Graphique 8a: Estimation non-parametrique des indicateurs d'inegalite de Theil pour les revenus disponibles



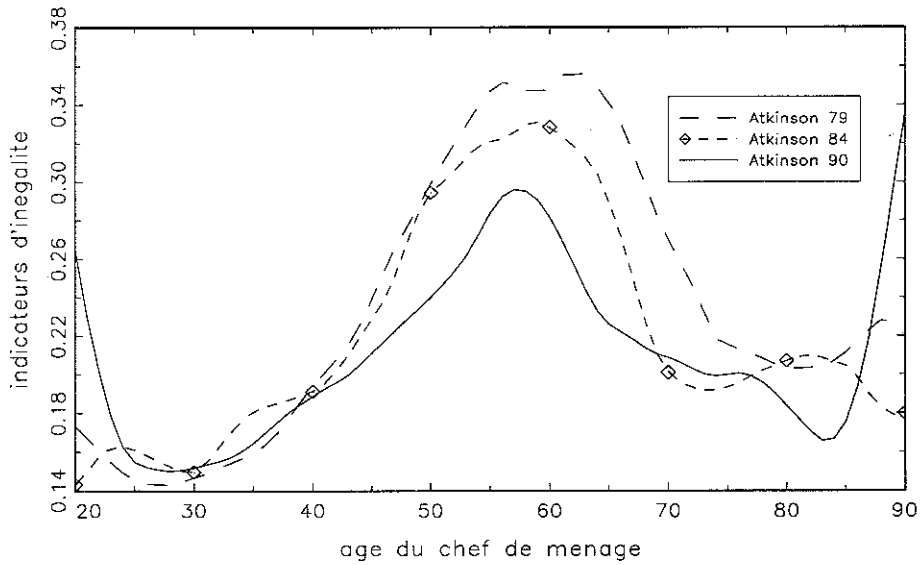
Graphique 8b : Estimation non-parametrique des indicateurs d'inegalite d'Atkinson (epsilon=0.5) pour les revenus disponibles



Graphique 8c : Estimation non-parametrique des indicateurs d'inegalite d'Atkinson (epsilon=2) pour les revenus disponibles

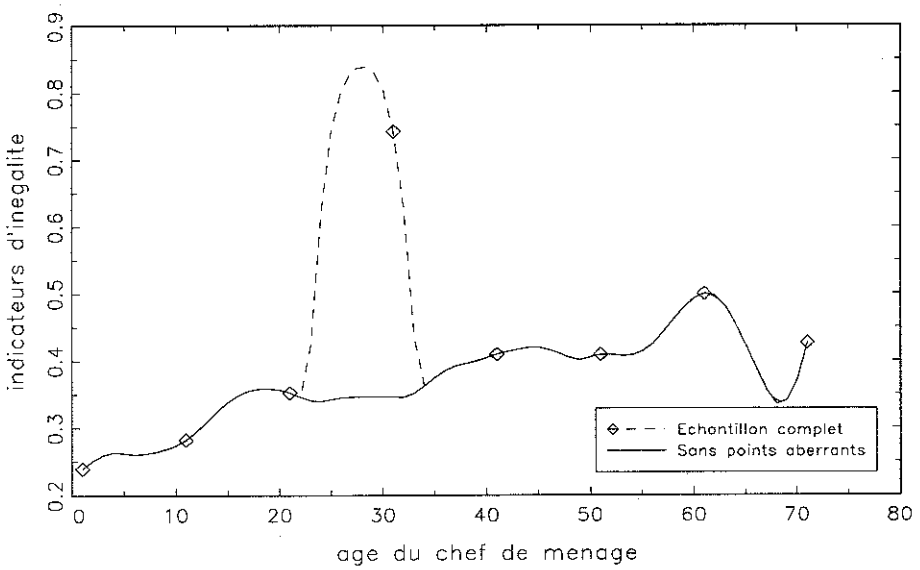
Pour répondre à cette question, nous avons ré-estimé l'indicateur d'Atkinson 2 en excluant les quelques revenus disponibles particulièrement bas. Ainsi, pour les enquêtes de 1979 et 1984, nous avons supprimé les valeurs inférieures à 1 510 Francs (au nombre de 27 et 17, respectivement). Pour l'année 1990, nous avons exclu les 5 ménages dont le revenu disponible était cette année-là inférieur à 3 500 Francs. Le résultat est net : l'omission de ces cinq ménages, dans lesquels par ailleurs l'âge de la personne de référence était en 1990 au plus égal à 27 ans, provoque une baisse significative de l'indicateur 1990 d'Atkinson 2 aux âges compris entre 25 et 35 ans, de sorte que cet indicateur devient alors quasiment identique sur cette classe d'âge à l'indicateur relatif à l'année 1979 (voir graphique 9). La conclusion de cet exercice est que, pour des indicateurs particulièrement sensibles aux valeurs extrêmes tel que l'indicateur d'Atkinson 2, le praticien doit veiller à utiliser des méthodes non-paramétriques de traitement des « points aberrants » (voir par exemple, Härdle [1990, chapitre 6]).

Graphique 9 : Estimation non-parametrique des indicateurs d'Atkinson (epsilon=2) pour les revenus disponibles (sans points aberrants)



Une conclusion du même type peut être obtenue pour l'indicateur d'Atkinson 2 relatif aux revenus fiscaux de l'année 1984. Le graphique 10 montre que l'estimation non-paramétrique de cet indicateur est particulièrement pathologique aux âges compris entre 40 et 52 ans. Cette pathologie disparaît totalement dès lors que l'on exclut de l'échantillon les huit ménages dont le revenu fiscal était inférieur à 650 Francs en 1984 : parmi ces huit ménages, deux seulement appartenaient à la classe d'âge 40-50 ans.

Graphique 10 : Estimation non-parametrique des indicateurs d'Atkinson (epsilon = 2) pour les revenus fiscaux 1984



4. Conclusions

Le but de cette communication était d'offrir des exemples d'application des méthodes non-paramétriques pour l'analyse des distributions et des inégalités de revenus. L'estimation des densités de revenus des foyers fiscaux par méthode du noyau a mis en évidence un déplacement « vers la droite » des distributions de ces revenus entre 1979 et 1990. L'effet est plus particulièrement net pour les ménages d'âge intermédiaire. La distribution des revenus des ménages jeunes n'a en revanche connu aucune modification durant la décennie étudiée. L'estimation par régression non-paramétrique des revenus moyens et des indicateurs usuels d'inégalité confirme l'analyse précédente : les revenus fiscaux moyens ont crû notablement entre 1979 et 1990 pour les ménages dont la personne de référence avait entre 40 et 50 ans à ces dates, alors qu'ils sont restés stables pour les ménages âgés de moins de 33 ans. L'analyse non-paramétrique fait en outre apparaître une baisse sensible de l'inégalité des revenus disponibles au sein des ménages adultes et plus âgés. Ces résultats sont conformes aux conclusions émises à partir d'autres types de méthodes et qui soulignent l'émergence de problèmes pécuniaires au sein de la fraction jeune de la population ; mais ici la limite d'âge en deçà de laquelle il y a stagnation relative des revenus est mesurée avec plus de précision. Elle s'avère d'ailleurs plus élevée que ce que l'on a coutume d'associer à la fin de la « jeunesse » : à 33 ans, par exemple, l'insertion professionnelle devrait être stabilisée depuis plusieurs années.

Härdle W. (1990) : Applied Nonparametric Regression. Cambridge University Press.