REVUE DE STATISTIQUE APPLIQUÉE

WERNER HILDENBRAND
ALOIS KNEIP
KLAUS J. UTIKAL

Une analyse non paramétrique des distributions du revenu et des caractéristiques des ménages

Revue de statistique appliquée, tome 47, n° 3 (1999), p. 39-56
<http://www.numdam.org/item?id=RSA_1999__47_3_39_0>

Rev. Statistique Appliquée, 1999, **XLVII** (3), 39-56

UNE ANALYSE NON PARAMÉTRIQUE DES DISTRIBUTIONS DU REVENU ET DES CARACTÉRISTIQUES DES MÉNAGES

Werner Hildenbrand[1], Alois Kneip[2], Klaus J. Utikal[1]

[1]*Département d'Economie, Université de Bonn, Bonn, Allemagne*

[2]*Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgique*

RÉSUMÉ

Nous étudions des familles de distributions bivariées du revenu et de l'âge, mais aussi du statut professionnel par l'intermédiaire de leurs densités marginales et conditionnelles sur plusieurs années. Un concept d'invariance est introduit afin de décrire l'évolution dans le temps de ces distributions. Une estimation non paramétrique est réalisée en utilisant des estimateurs fonctionnels à noyau obtenus en sélectionnant des fenêtres optimales. Les données se composent d'échantillons en coupe transversale provenant de la population des ménages britanniques au cours des années 1968 à 1995.

Mots-clés : *estimation non paramétrique de la densité, fonctions noyau, sélection de fenêtre, invariance, données en coupe transversale, distributions âge-revenu, enquête Family Expenditure Survey.*

ABSTRACT

We study families of bivariate distributions of income with age and occupation via their marginal and conditional densities over several years. A concept of invariance is introduced to describe the time changing nature of these distributions. Estimation is carried out nonparametrically using kernel smoothers with optimally selected bandwidths. The data consists of cross sectional samples from the population of Brithish households drawn over the years of 1968–1995.

Keywords : *nonparametric density estimation, kernel functions, bandwidth selection, invariance, cross sectional data, income-age distributions, Family Expenditure Survey.*

1. Introduction

En économie, comme dans d'autres champs d'application, il est d'une importance majeure d'analyser la distribution du revenu et de variables socio-économiques apparentées. Le terme «variables apparentées» désigne des caractéristiques de la population, telles que l'âge, la taille de la famille, le statut professionnel, etc..., qui sont fortement corrélées au revenu et qui jouent un rôle important dans de nombreux modèles économiques.

Dans cet article, notre étude porte principalement sur l'étude du revenu et des caractéristiques des ménages (âge et catégorie sociale) en Grande Bretagne. Les données proviennent de l'enquête Family Expenditure Survey (FES) conduite au cours des années 1968 à 1995. Des enquêtes similaires ont été menées dans d'autres pays comme, par exemple, l'Enquête Budget Famille en France (voir section 2); elles pourraient être analysées de façon analogue.

Une littérature abondante traite des distributions du revenu. Beaucoup de travaux théoriques et appliqués partent du point de vue de l'analyse du bien-être, voir, entre autres, Gottschalk et Smeeding (1997). Ces travaux ont pour objet de mesurer l' «inégalité» de la distribution du revenu parmi des populations données. Une approche usuelle consiste à construire des paramètres spécifiques qui quantifient le degré d'inégalité, comme les quantiles de la distribution du revenu, le coefficient de Gini, la courbe de Lorentz, voir Atkinson (1983). Ces paramètres sont comparés d'un pays à un autre et leur évolution dans le temps est étudiée.[1]

Des variables telles que le revenu, l'âge, la taille de la famille jouent aussi un rôle important en analyse de la demande. Elles influencent de façon déterminante le budget domestique des biens de consommation (tels que les produits alimentaires, les carburants, les transports, les services, etc...). Ce qu'on apelle «systèmes de demande», voir entre autres Deaton et Muellbauer (1980) ou Blundell et al. (1993), modélisent la consommation d'un foyer particulier comme une fonction de ces variables et des prix. Un problème important en macroéconomie consiste à modéliser l'évolution temporelle de la consommation moyenne de la population. Pour modéliser la consommation moyenne, on utilise de façon usuelle les modèles autorégressifs basés sur le revenu *moyen*, voir Deaton (1992). D'autre part, il est clair que les modifications de la consommation moyenne dépendent généralement des modifications de la distribution toute entière du revenu. Supposer que cette distribution ne rentre dans l'analyse que par l'intermédiaire de sa moyenne est une démarche simplificatrice. En fait, dans la littérature économique, on note une controverse au sujet de l'influence additionnelle que pourraient avoir des effets de distribution comme l'«inégalité croissante» du revenu. Hildenbrand et Kneip (1999) proposent une approche générale pour modéliser l'influence des distributions du revenu et des caractéristiques de la population sur la consommation lorsque ces distributions évoluent dans le temps. L'idée fondamentale consiste à chercher des invariances dans le temps de telles distributions. L'exemple suivant en donne une illustration. Soit $f_1, f_2, \ldots$ les densités du revenu en différentes années $t = 1, 2, \ldots$. En général, lorsque les distributions évoluent dans le temps et qu'elles ne rentrent dans l'analyse que par l'intermédiaire de leur moyenne, on n'obtient des résultats corrects que si l'on a l'invariance des distributions du revenu relatif :

$$\bar{x}_t \, f_t(\bar{x}_t x) = \bar{x}_s \, f_s(\bar{x}_s x) \qquad \text{pour les années } t, s$$

où $\bar{x}_t$ désigne le revenu moyen de l'année t. Alors, les évolutions des distributions du revenu sont complètement paramétrisées par leur moyenne $\bar{x}_t$ puisque la densité du revenu relatif $x_t/\bar{x}_t$ ne change pas dans le temps.

[1] Dans ces travaux, on s'intéresse d'habitude à l'analyse du revenu «individuel». En règle générale, un tel revenu individuel est établi à partir du revenu des ménages en utilisant ce qu'on appelle des échelles d'équivalence. Dans cet article, nous n'adoptons pas cette approche.

La notion d'invariance introduite ci-dessus nous amène à poser une condition spécifique sur l'évolution des densités du revenu dans le temps. Nous montrerons plus loin que cette transformation simple peut déjà apporter une première approximation raisonnable, et que l'on peut améliorer l'invariance en utilisant d'autres transformations plus générales du revenu aussi bien que d'autres variables concommitantes. En général, la recherche de l'invariance des densités transformées de façon appropriée peut se résumer par l'énoncé suivant : *étant donnés des échantillons annuels d'observations, trouver une famille de transformations simples qui conduisent à une famille de densités évoluant très peu dans le temps*. Une transformation simple satisfait le principe de parcimonie si elle dépend seulement de peu de paramètres qui peuvent évoluer dans le temps. L'idéal serait que ces paramètres puissent s'interpréter facilement (comme la moyenne ou la variance de la distribution du revenu). De plus, leur interprétation pourrait se révéler pertinente dans l'analyse des prédictions de transformations futures (donc des densités futures du revenu) à partir du passé.

Avant de pouvoir chercher des transformations des densités du revenu, le problème qui se pose à nous est celui de l'estimation de la densité. Il s'avère que l'on ne peut pas considérer ces densités comme une forme paramétrique simple sauf pour des sous-groupes particuliers de la population. Elles doivent alors être estimées de façon non paramétrique. La littérature dans le domaine de l'estimation non paramétrique de la densité est abondante, et plusieurs méthodes différentes ont été développées, voir Silverman (1986). La méthode la plus largement admise et la plus simple à concevoir est la méthode de l'estimation à noyau de la densité. De plus, d'après notre expérience (et on peut aussi le démontrer mathématiquement), lorsque les échantillons sont de taille élevée, les estimations sont proches les unes des autres bien que les estimateurs soient dérivés de méthodes non paramétriques différentes. Nous disposons de tels échantillons que nous analysons plus loin aux sections 4 et 5. Une description des données se trouve à la section 2.

Dans le cadre de l'estimation des distributions du revenu, des aspects techniques importants de l'utilisation des estimateurs à noyau ont déjà été discutés dans Wand *et al.* (1991) ainsi que dans Schmitz et Marron (1992). Notre approche méthodologique décrite à la section 3 se base sur ces travaux, mais il se distingue par une meilleur procédure pour choisir le paramètre de lissage.

Les transformations examinées aux sections 4 et 5 sont basées sur des méthodes simples, comme la standardisation ou encore la transformation logarithmique. La performance des transformations est généralement améliorée en partitionnant la population en sous-groupes. Dans plusieurs cas, on peut argumenter, que, de cette manière, l'invariance est déjà atteinte de façon satisfaisante, alors que pour d'autres variables telles que le revenu, l'utilisation de transformations reste indispensable.

Dans cet article, il reste quelques points à éclairer par les recherches à venir, tels que le problème de la prédiction des densités futures et l'élaboration de méthodes plus générales pour trouver une transformation appropriée. Nous nous référons à un article de Kneip et Utikal (1999) pour une solution mathématique de ce problème, basée sur l'analyse semi paramétrique d'une famille générale de transformations d'échelle. Les implications économiques de ces résultats restent jusqu'à ce jour inexplorées.

2. Les Données

Dans les sections suivantes, les analyses sont conduites sur la base des données provenant de l'enquête britannique Family Expenditure Survey (FES). Cette enquête est effectuée chaque année depuis 1957, à l'initiative du gouvernement britannique. Chaque année, environ 7000 ménages, ce qui revient à 0.5 % des ménages britanniques, notent leurs dépenses concernant une grande variété de biens de consommation, comme le pain, différentes sortes de viande, par exemple. Un «ménage» est défini globalement comme un groupe de personnes vivant sous le même toit et qui partagent au moins un repas commun. Les informations sont recueillies au moyen d'entretiens avec les membres du foyer, mais aussi à partir d'«agendas» dans lesquels les personnes interrogées ont noté toutes leurs dépenses sur une période de quatorze jours. Le taux de réponses se situe autour de 68 % de tous les ménages sélectionnés et varie chaque année. Dans cette enquête, se trouvent également des informations sur les différentes formes du revenu ainsi que sur d'autres caractéristiques des ménages. Afin de définir précisément les variables, les unités expérimentales, les modèles d'échantillonage, le travail d'enquête, la confidentialité et la fiabilité de l'information, entre autres, nous nous référons au manuel de l'enquête FES publié annuellement comme à l'ouvrage de Kemsley et al., Family Survey Handbook (1980).

Dans cette étude, nous utilisons des informations sur le revenu des ménages ainsi que sur l'âge et le statut professionnel du chef de ménage. On désigne par «chef de ménage» l'époux d'un couple marié; dans tous les autres cas, il s'agit principalement de la personne qui possède ou qui loue le lieu de résidence. Nous disposons de la variable de revenu net, c'est-à-dire du revenu disponible hebdomadaire des ménages, ce qui revient essentiellement au revenu brut minoré des taxes et des charges de Sécurité Sociale. Notons cependant, que les cotisations retraite n'ont pas été déduites du revenu brut.

Des enquêtes similaires ont été menées dans d'autres pays, comme le CEX (Etats-Unis), EPF (Espagne), EBF (France). Toutes ces enquêtes sont dites en coupe transversale (c'est-à-dire qu'elles portent sur différents ménages à différentes années). Elles sont considérablement différentes des enquêtes de panel, dans lesquelles une cohorte de ménages est étudiée sur une période donnée, comme cela a déjà été fait dans le PSID (Etats-Unis) et le GSEP (Allemagne).

Le lecteur néophyte doit être averti qu'aucune conclusion ne peut être tirée à partir des données sans examiner attentivement la signification précise des variables observées. Il est particulièrement important de garder en mémoire que les définitions changent un peu avec le temps, comme c'est le cas dans notre étude. Ce problème sera illustré dans la section suivante.

3. Les méthodes statistiques

La méthode de l'histogramme (c'est-à-dire le diagramme des fréquences) constitue un outil utile pour estimer des densités. Elle est aussi facile à calculer que simple à comprendre et à expliquer. Cependant, elle présente aussi des inconvénients certains. Entre autres, il est clair qu'une densité continue ne peut raisonnablement

être estimée par une fonction discontinue. Par ailleurs, il serait peu pertinent d'étudier des densités variables dans le temps en représentant plusieurs histogrammes sur le même graphique. Plusieurs méthodes ont été élaborées pour surmonter les défauts des histogrammes et pour produire des estimateurs plus précis, voir Silverman (1986). Nous utiliserons uniquement les estimateurs à noyau de la densité. Des résultats théoriques ont été pleinement développés et des logiciels de calcul sont facilement accessibles. On peut montrer aussi que les histogrammes sont un cas particulier de ces estimateurs à noyau.

Etant données n observations $X_1, \ldots, X_n$ d'une variable aléatoire X, un estimateur non paramétrique bien connu de sa densité $f(x)$ est l'estimateur à noyau

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.1}$$

où K est une fonction noyau déterminée, qui satisfait aux conditions suivantes : K est une fonction non négative et dont l'intégrale est égale à 1. Par exemple, nous utilisons dans nos analyses le noyau gaussien

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3.2.}$$

Nous ferons par la suite des commentaires plus approfondis sur le choix du noyau et du *paramètre de lissage* h (encore appelé fenêtre). Notons que pour $x = x_1, x_2, \ldots$ où $x_{i+1} - x_i = h$ et pour un noyau $K(x) = 1/h$ pour $-h/2 < x \leqslant h/2$ et zéro ailleurs, nous retrouvons l'histogramme si l'on calcule $\widehat{f}(x_i)$ au moyen de (3.1) et définissons $\widehat{f}(x) = \widehat{f}(x_i)$ pour tout autre x tel que $x_i - h/2 < x \leqslant x_i + h/2$. Il est bien connu que, pour de grands échantillons, cet estimateur est asymptotiquement non biaisé et normalement distribué, de variance

$$\operatorname{var}\{\widehat{f}(x)\} = \frac{c_k}{nh} f(x)$$

où

$$c_k = \int K^2(u)\, du\,.$$

L'erreur quadratique moyenne (mse) approchée de $\widehat{f}(x)$ s'exprime comme la somme de la variance et du biais au carré,

$$\operatorname{mse}\{\widehat{f}(x)\} = \operatorname{var}\{\widehat{f}(x)\} + b^2(x) \tag{3.3}$$

$$b(x) = \frac{1}{2} h^2 f''(x)\, k_2 \tag{3.4}$$

où

$$k_2 = \int u^2\, K(u)\, du$$

(Notons que pour le noyau gaussien $c_k = 1/(2\sqrt{\pi})$ et $k_2 = 1$).

En règle générale, il est établi que l'estimation est peu sensible au choix du noyau, sauf pour des propriétés de continuité, alors que la fenêtre joue un rôle crucial. Plusieurs méthodes de sélection de fenêtre ont été mises au point, voir Simonoff (1996). Dans les analyses menées en section 4, nous utilisons le critère de minimisation de l'erreur quadratique moyenne intégrée MISE $= \int \mathrm{mse}(\widehat{f}(x))\,dx$. La fenêtre optimale h_{opt} doit alors satisfaire

$$h_{\mathrm{opt}} = k_2^{-2/5} c_k^{1/5} \left\{ \int f''(x)^2\,dx \right\}^{-1/5} n^{-1/5} \tag{3.5}$$

Pour résoudre cette équation, nous avons besoin d'estimer la dérivée de la densité inconnue f'', ce qui requiert de choisir un second paramètre de lissage. Par conséquent, la fenêtre obtenue $\widehat{h}_{\mathrm{opt}}$ est seulement une estimation de h_{opt}. Une approximation simple, décrite par Silverman (1986), consiste à remplacer dans (3.5) la fonction inconnue f'' par la dérivée seconde d'une approximation de la densité inconnue f. Cette méthode nous paraît fort utile pour donner une première idée de la largeur de la fenêtre lorsque des densités symétriques en forme de cloche doivent être estimées. Dans une approche plus générale, on se propose de spécifier la dépendance par rapport à $\widehat{h}_{\mathrm{opt}}$ du paramètre de lissage dans l'estimation de f'' et ensuite de résoudre itérativement l'équation consécutive. Engel *et al.* (1994) ont montré que l'algorithme proposé (et que nous utilisons aussi) est convergent. De plus, la solution $\widehat{h}_{\mathrm{opt}}$ est asymptotiquement convergente vers la solution h_{opt} de (3.5) lorsque la taille de l'échantillon croît. En effet, la convergence de $|h_{\mathrm{opt}} - \widehat{h}_{\mathrm{opt}}|/h_{\mathrm{opt}}$ vers 0 est d'ordre stochastique $n^{-1/2}$, ce qui est assez rapide[2]. De l'estimateur obtenu avec cette fenêtre optimale estimée, on déduit une erreur quadratique moyenne estimée d'ordre $O_P(n^{-4/5})$[3] . On peut espérer que cet estimateur soit d'une grande précision pour des échantillons de taille modérément grande.

Notre approche pour estimer des densités du revenu consiste cependant en une légère modification de cette procédure générale. Pour étudier les densités sur une période de plusieurs années, nous effectuons des transformations logarithmiques du revenu, et nous nous concentrons tout d'abord sur l'estimation des densités f_{Y_t} du log revenu $y(x) = \log(x)$. Les fenêtres optimales sont calculées séparément pour chaque année $t = 1, 2, \ldots$. Ensuite, nous estimons les différentes densités f_{Y_t} en utilisant comme fenêtre la moyenne de toutes ces fenêtres optimales. Sur chaque graphique, les fenêtres moyennes sont notées en légende. On peut les utiliser pour évaluer la variabilité des estimateurs, en s'appuyant sur la formule (3.3).

Comme dernière étape, ces estimations des f_{Y_t} sont utilisées pour construire des estimations des densités originales f_t du revenu. Ceci se fait par une simple

---

[2] Schmitz et Marron (1992) sélectionnent la fenêtre de lissage par la méthode de validation croisée qui a seulement une convergence d'ordre stochastique $n^{-1/10}$. Wand *et al.* (1991) utilise une procédure purement approximative.

[3] Soit $\{Z_n\}_{n\in\mathbb{N}}$ une séquence de variables aléatoires réelles. On dit que $Z_n = O_p(n^{-\gamma})$, $\gamma \geqslant 0$, si pour chaque $\epsilon > 0$ il existe un $M > 0$ tel que $P(n^{\gamma}|Z_n| > M) < \epsilon$ pour chaque n suffisamment grand.

transformation selon la formule suivante :

$$f_t(x) = f_{Y_t}(y(x)) \left| \frac{dy(x)}{dx} \right|$$

Il n'est pas difficile de voir que cette procédure équivaut à estimer les densités f_t du revenu (non transformé) en employant un noyau de fenêtre variable $x\,h$ au lieu d'une fenêtre fixe h. Cette méthode employée pour transformer les données a été discutée dans Wand *et al.* (1991). Si l'on ne l'emploie pas, la structure des densités, pour les tranches des très faibles revenus, risque d'être mal reproduite (voir Wand *et al.*, 1991).

Pour illustrer cette méthode, considérons le revenu relatif des ménages observé en 1984. Un histogramme classique de ces données est représenté en figure 3.1 [à gauche]. Une estimation à noyau de la densité [à droite] obtenue avec une fenêtre de 0.3 montre une courbe estimée ayant en gros la même allure.

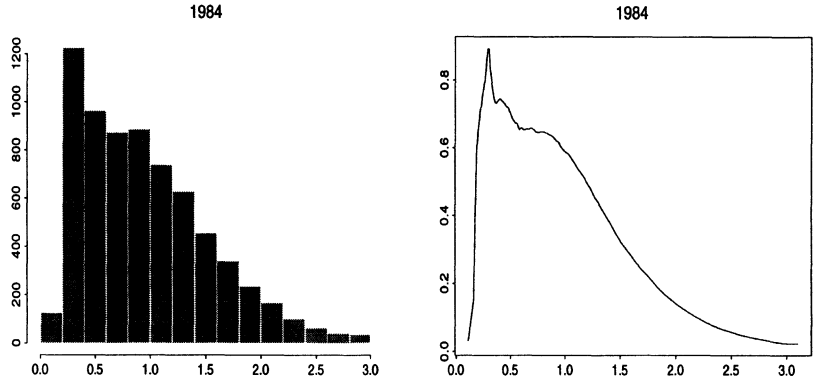


FIGURE 3.1
Histogramme (15 classes) et estimateur à noyau de la densité (fenêtre = 0.3).

Un estimateur à noyau de la densité, obtenu avec la fenêtre optimale (0.08) aboutit à une forme différente, voir figure 3.2 [à gauche]. Une singularité importante apparaît pour des niveaux de revenus très faibles. En reprenant les mêmes données, un histogramme, réalisé avec une partition en classes très fines, révèle l'existence de plusieurs centaines de ménages ayant de très faibles revenus (voir figure 3.2 [à droite]).

En outre, nous observons que cette particularité reste présente chaque année postérieure à 1984, alors qu'elle n'apparaît pas les années précédentes, voir figure 3.3. Une explication peut en être la suivante. Dès 1984, la définition du revenu net a été modifiée : les allocations logement ne rentrent plus en compte dans le calcul du revenu net. Par conséquent, nous limiterons notre analyse aux données du revenu antérieures à 1984, ce qui nous permettra d'illustrer ces méthodes tout aussi bien. De plus, on peut espérer que certains sous-groupes de la population, tels que celui des personnes employées à temps plein, seront peu affectés par ce changement à

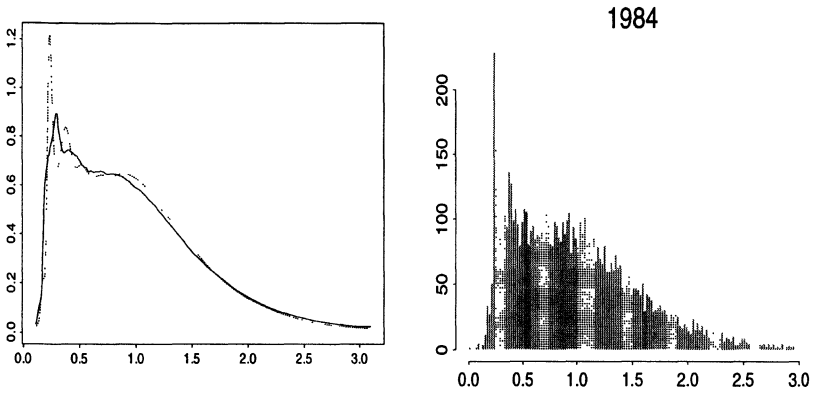


FIGURE 3.2
Estimateur à noyau de la densité [à gauche] fenêtre = 0.3 (trait plein) et fenêtre optimale = 0.08 (ligne brisée). Histogramme obtenu avec une partition en fines classes [à droite].

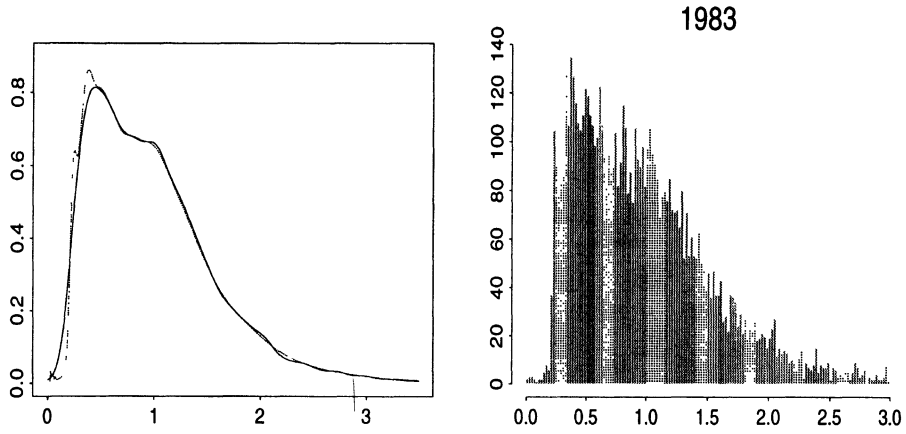


FIGURE 3.3
Estimateur à noyau de la densité [à gauche] fenêtre = 0.3 (trait plein) et fenêtre optimale = 0.09 (ligne brisée). Histogramme avec une partition en fines classes [à droite].

l'extrêmité inférieure de l'échelle des revenus. C'est la raison pour laquelle nous avons étendu l'analyse de ce groupe à l'ensemble de l'intervalle de temps.

Nous insistons à nouveau sur le fait suivant : un choix «intuitif» de la fenêtre, conduisant à des estimations «crédibles», aurait abouti à masquer cette singularité importante des données. La méthode de sélection optimale de la fenêtre a le mérite de rendre l'estimateur sensible à ce type de caractéristique.

Nous terminons cette section en mentionnant brièvement le problème de l'estimation d'une densité multivariée. Une bonne introduction à ce sujet se trouve

dans la monographie de Scott (1992). De façon similaire à (3.1), on peut définir, pour un échantillon bivarié $(X_1, Y_1), \ldots, (X_n, Y_n)$ l'estimateur à noyau

$$\widehat{f}(x,y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K_{12}\left(\frac{x - X_i}{h_1}, \frac{y - Y_i}{h_2}\right)$$

où K_{12} est une fonction bivariée non négative, d'intégrale égale à 1. Par commodité, nous considérons cette fonction comme le produit de deux noyaux gaussiens univariés, c'est-à-dire

$$K_{12}(x,y) = K(x)\,K(y)$$

où K est définie en (3.2). De façon analogue au cas univarié, l'erreur quadratique moyenne, le choix de la matrice fenêtre etc... ont été étudiés dans le cas multivarié. En outre, de nouveaux problèmes sont apparus dans ce domaine intéressant encore en plein développement; pour une liste actuelle de références, voir entre autres Simonoff (1996).

4. La distribution du revenu des ménages

Les données décrites en section 2 constituent des échantillons en coupe transversale du revenu des ménages sur une période de 28 années. Les échantillons sont de taille importante, comprenant environ 7000 ménages chaque année. Etant donnés de grands échantillons, on peut espérer que les estimations à noyau des densités du revenu soient très précises.

On peut visualiser par la figure 4.1 les densités estimées du revenu net hebdomadaire (en livres) pour les années 1968, 1973, 1978 et 1983. Puisque le revenu (nominal) augmente régulièrement dans le temps, il n'est pas surprenant de constater que les densités estimées changent beaucoup dans le temps.

Il semble naturel de considérer le revenu relatif plutôt que le revenu nominal; le revenu relatif s'obtient en divisant le revenu nominal de chaque ménage par la moyenne de la population. Les densités du revenu relatif sont alors comparables sur une échelle commune des multiples du revenu moyen, voir figure 4.2.

On s'aperçoit que seulement très peu de ménages ont un revenu extrêmement faible et l'on ne peut tirer de conclusions claires concernant le majorant des hauts revenus, pour lesquels les queues de distribution décroissent assez lentement. Notons que, pour des raisons de présentation, les densités ont été représentées pour des revenus allant seulement jusqu'à trois fois le revenu moyen.

La structure multimodale de ces densités pourrait s'expliquer comme étant le résultat d'une superposition de densités unimodales caractérisant des sous-populations influentes. C'est ce que suggèrent les figures 4.3 et 4.4 qui montrent les estimations des densités du revenu relatif des chefs de famille employés à temps plein et de ceux sans emploi. Notons que, pour chaque sous-groupe, le revenu relatif se réfère au revenu normalisé par le revenu moyen de la sous-population correspondante. Les densités obtenues sont à peu près unimodales, et leur mode se situe en des

points très différents[4] . Les modes correspondants de la densité du revenu relatif de la population totale (figure 4.2) apparaissent clairement.

Lorsque nous avons considéré l'évolution des densités du revenu dans le temps, nous avons déjà remarqué plus haut que la distribution du revenu nominal se modifiait rapidement dans le temps. La transformation du revenu nominal en un revenu relatif conduit à des distributions beaucoup plus «invariantes». Néanmoins, ce caractère d'invariance est loin d'être parfait. Un examen minutieux de la figure 4.2 met quand même en évidence une tendance dans l'évolution temporelle de ces densités du revenu relatif : le nombre de ménages à faibles revenus est en augmentation, alors que la hauteur du pic de classe moyenne, qui se situe autour du revenu moyen, décroît avec le temps. Ce sujet a davantage été approfondi dans Kneip et Utikal (1999). Une stratification en sous-populations appropriées peut déboucher sur des densités plus stables dans le temps comme on peut le voir sur les figures 4.3 et 4.4. En particulier, les densités relatives aux chefs de famille employés à temps plein sont plus invariantes dans le temps que celles relatives à la population totale. En fait, une partie de la tendance temporelle caractérisant ces dernières peut s'expliquer simplement par la croissance de la population de tels sous-groupes; il est bien connu que le pourcentage de personnes employées à temps plein décroît régulièrement au Royaume-Uni alors que celui des personnes sans emploi a sans cesse augmenté jusqu'à ces derniers temps.

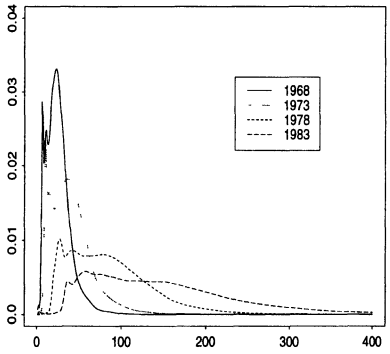


FIGURE 4.1
Population totale : densités du revenu nominal.

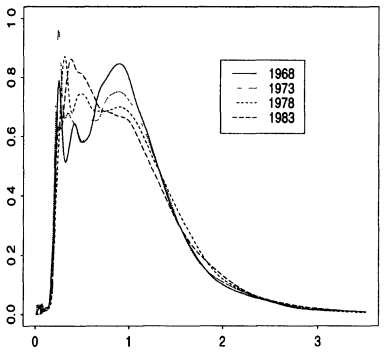


FIGURE 4.2
Population totale : densités du revenu relatif (0.079)[5].

On pourrait essayer d'éliminer les différences qui restent présentes entre les densités en figures 4.2-4.4 en appliquant des transformations plus sophistiquées.

[4] En section 2, nous avons déjà signalé une modification apparue en 1984 de la définition du revenu. Cette modification n'affecte pas la population des personnes employées à temps plein. Pour cette raison, seuls les revenus de ce dernier groupe ont été étudiés à partir de 1983.

[5] Les fenêtres moyennes optimales utilisées pour le lissage du log revenu $y(x) = \log(x)$ sont données entre parenthèses ( ).

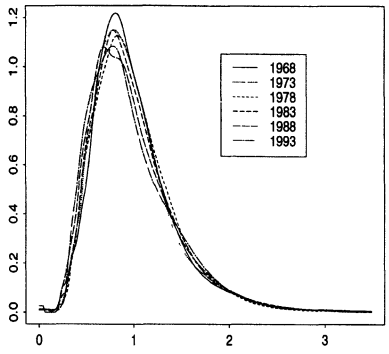


FIGURE 4.3
Employés à temps plein :
densités du revenu relatif (0.084).

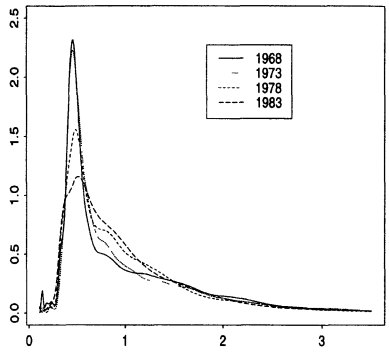


FIGURE 4.4
Sans emploi :
densités du revenu relatif (0.089).

Celles-ci incorporent des moments d'ordre plus élevé ainsi que nous le ferons par la suite. Soit X_{it} le revenu nominal d'un ménage individuel i, définissons le *log revenu standardisé* Z_{it} par

$$Z_{it} = \frac{\log(X_{it}) - \mu_t}{\sigma_t}$$

où μ_t et σ_t^2 désignent la moyenne et la variance de $\log(X_{it})$ dans la population. Si les distributions sous-jacentes étaient exactement normales alors, pour toute année t, la densité obtenue f_t^* de Z_{it} générée par cette transformation serait une normale standard. Donc, sous cette hypothèse, les densités f_t^*, $f_{t+1}^*, \ldots$ seraient complètement invariantes dans le temps. Notons toutefois qu'il n'est pas nécessaire d'avoir la log-normalité pour obtenir l'invariance temporelle. En fait, assez généralement, il semble raisonnable d'espérer que l'invariance temporelle s'améliore après avoir appliqué cette transformation : elle élimine les différences en localisation et *en dispersion* entre les distributions. Par exemple, il est bien connu que les variances du log-revenu relatif augmentent dans le temps. La standardisation élimine donc cet effet.

Les figures 4.5 et 4.6 montrent les densités estimées f_t^* du log-revenu standardisé pour la population totale et pour le sous-groupe des personnes employées à temps plein. Comme résultat, il semble que les densités des personnes employées à temps plein soient de façon très satisfaisante invariantes dans le temps. En principe, cela peut être testé, bien que cela n'ait pas encore été fait par les auteurs. Cette hypothèse est d'un intérêt certain puisque l'étude de l'évolution de la densité du revenu se réduirait alors entièrement à une étude de l'évolution des paramètres μ_t et σ_t.

Manifestement, les densités du log-revenu standardisé de la population totale ne sont pas exactement invariantes, mais il semble qu'elles gagnent en stabilité comparées aux densités du revenu relatif.

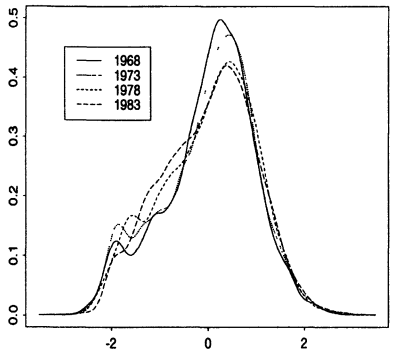


FIGURE 4.5
Population totale :
densités du log-revenu standardisé
(0.127)

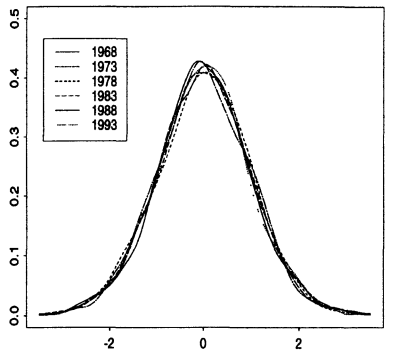


FIGURE 4.6
Employés à temps plein :
densités du log-revenu standardisé
(0.188)

5. Les distributions conjointes du revenu et des caractéristiques des ménages

Dans la section précédente, nous avons vu qu'une stratification par le statut professionnel aboutit à des densités du revenu dont la structure est plus simple. En général, d'un point de vue économique, l'analyse conjointe du revenu et d'autres caractéristiques des ménages (tels que l'âge, la taille du ménage, entre autres) est d'un intérêt considérable. Cette démarche est particulièrement importante en analyse de la demande. La consommation et l'épargne dépendent du revenu aussi bien que des caractéristiques des ménages. Dans la littérature économique, l'âge est souvent considéré comme la plus importante des covariables, voir Deaton (1992).

Par la suite, nous concentrons donc notre attention sur la distribution conjointe de l'âge et du revenu. Dans l'enquête FES, la variable «âge» se réfère à l' «âge du chef de famille». Par conséquent, la distribution de l'âge correspondante ne représente pas la distribution de l'âge de tous les individus en Grande Bretagne. De toute évidence, très peu chefs de ménage sont âgés de moins de 20 ans.

Dans la section précédente, nous avons analysé en détail la distribution marginale du revenu. On peut à nouveau utiliser des estimateurs à noyau de la densité pour étudier la distribution marginale de l'âge. La figure 5.1 montre les densités estimées de l'âge pour les années 1968-1971 et 1980-1983. On peut constater qu'il y a réellement peu de ménages dont le chef de famille est jeune ou très âgé. Les densités sont élevées dans un intervalle allant de 25 à 70 ans. Les détails structuraux sont plutôt irréguliers et peu aisés à interpréter. En fait, les fenêtres optimales estimées sont très petites. Ceci se présente comme conséquence du fait qu'une petite fenêtre est indispensable pour une modélisation adéquate de la croissance rapide des densités après 20 ans ainsi que de la décroissance rapide autour de 70 ans. Des estimations plus stables pourraient être obtenues par un choix adaptif local des fenêtres, en utilisant des fenêtres plus larges dans la région entre 30 et 70 ans.

Nous nous intéressons principalement à la comparaison des densités de l'âge dans le temps. Il n'y a pas eu d'évolution apparente *au cours de l'une ou l'autre* de ces deux périodes de quatre ans. Outre des fluctuations mineures, il s'avère que les densités de l'âge entre 1968 et 1971 sont fort semblables les unes aux autres. On peut faire la même observation pour les densités entre 1980 et 1983, ou toute autre période de quatre années consécutives. Cependant, une différence frappante distingue ces deux familles de densités sur la figure 5.1 correspondant l'une au début des années soixante-dix et l'autre au début des années quatre-vingt. Un pic caractérisant les «jeunes ménages» émerge au cours des années quatre-vingt, phénomène qui n'apparaît pas dans les années soixante-dix, ce qui indique donc un changement socio-économique à long terme.

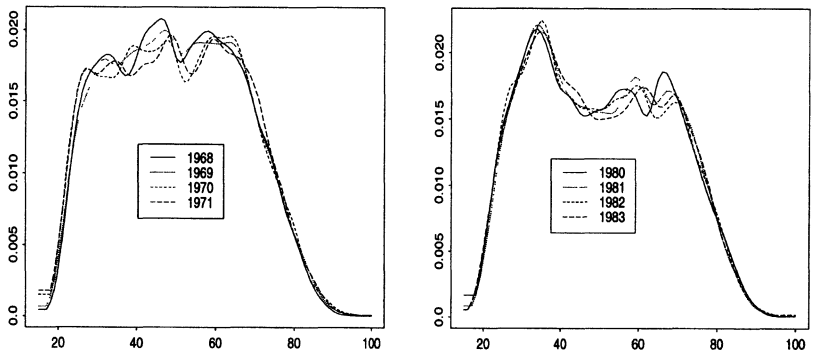
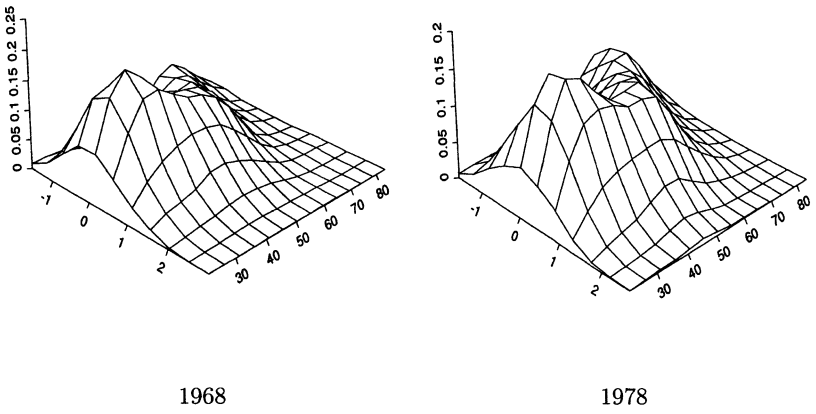


FIGURE 5.1
Population totale : densités de l'âge (1.84) [à gauche], (1.91) [à droite].

Rappelons la discussion menée en section 4 au sujet des invariances temporelles. Nous avons vu que les densités du revenu nominal se modifient rapidement d'une année à l'autre. Ce n'est qu'après avoir appliqué des transformations appropriées que nous pouvions parler d'une invariance approximative des densités du revenu transformé. La situation est clairement différente en ce qui concerne la distribution marginale de l'âge. Il n'est ni nécessaire ni réalisable de chercher de telles transformations. Les densités de l'âge évoluent très peu d'une année à l'autre, elles sont, à court terme, à peu près invariantes dans le temps. Les modifications des densités de l'âge doivent être prises en compte dans le seul cas où l'on s'intéresse principalement à une analyse de long terme.

Revenons maintenant à l'étude de la distribution conjointe de l'âge et du revenu. Considérer la distribution conjointe de l'âge et du revenu *nominal* n'a pas de sens. Il est clair, au regard de la distribution marginale du revenu nominal, que cette distribution évoluera rapidement dans le temps. Cependant, consécutivement à notre discussion sur les densités marginales, nous avons quelques espoirs de trouver des régularités dans la densité conjointe de l'âge et du *log-revenu standardisé*. La figure 5.2 montre les estimations à noyau en deux dimensions de cette densité conjointe pour deux années distinctes.

Les deux densités semblent être similaires dans une certaine mesure. Comme tendance globale, on reconnaît que, pour chacune de ces deux années, les ménages très



1968 1978

FIGURE 5.2
Population totale : Densité conjointe du log-revenu standardisé et de l'âge.

jeunes ainsi que les ménages très âgés ont des revenus plus faibles que les ménages ayant atteint la cinquantaine. La grande majorité des ménages à faibles revenus ont plus de 60 ans. Cependant, les jeunes ménages dont les revenus sont très faibles forment également un groupe important. Il apparaît que la plupart des ménages de la classe moyenne et de la classe moyenne supérieure sont âgés d'environ 50 ans. On trouve, de façon assez surprenante, peu de ménages à hauts revenus ayant plus de 70 ans.

Une connaissance plus approfondie de l'évolution temporelle de ces particularités peut être acquise en analysant les densités de l'âge stratifié par classes de revenus. Nous utilisons quatre classes de revenus que l'on détermine par leur position dans la densité du log-revenu standardisé (noté x) : $-2 \leqslant x < -1$ (faibles revenus); $-1 \leqslant x < 0$ (classe moyenne inférieure); $0 \leqslant x < 1$ (classe moyenne supérieure); $1 \leqslant x < 2$ (hauts revenus).

La figure 5.3 représente les distributions de l'âge au cours des années 1968-1972, obtenues pour quatre classes de revenus. Nous remarquons très vite que les distributions de l'âge sont très différentes pour des classes de revenus distinctes. Plus précisément, comme nous l'avons déjà vu plus haut pour la densité bivariée, la plupart des ménages à faibles revenus sont âgés d'environ 70 ans. La distribution bimodale de la classe moyenne inférieure se concentre autour de 30 et 70 ans. Le groupe des ménages à hauts revenus présente une distribution unimodale concentrée autour de 50 ans.

La figure 5.3 montre également que, à l'intérieur de *chaque groupe*, les densités stratifiées évoluent très peu de 1968 à 1972. On obtient ce même résultat pour d'autres périodes de cinq années consécutives. C'est ce que l'on peut voir, par exemple en figure 5.5, qui montre les densités de l'âge des ménages à faibles revenus au cours des années 1991-1995. D'autre part, les évolutions à long terme des densités stratifiées sont considérables, les plus prononcées apparaissant pour le groupe des ménages à faibles revenus. C'est ce que montre la figure 5.4 qui représente l'évolution obtenue au cours des années 1968 à 1995. On détecte clairement l'apparition d'un sous-groupe des jeunes ménages à faibles revenus. Nous devrions noter que ce phénomène socio-

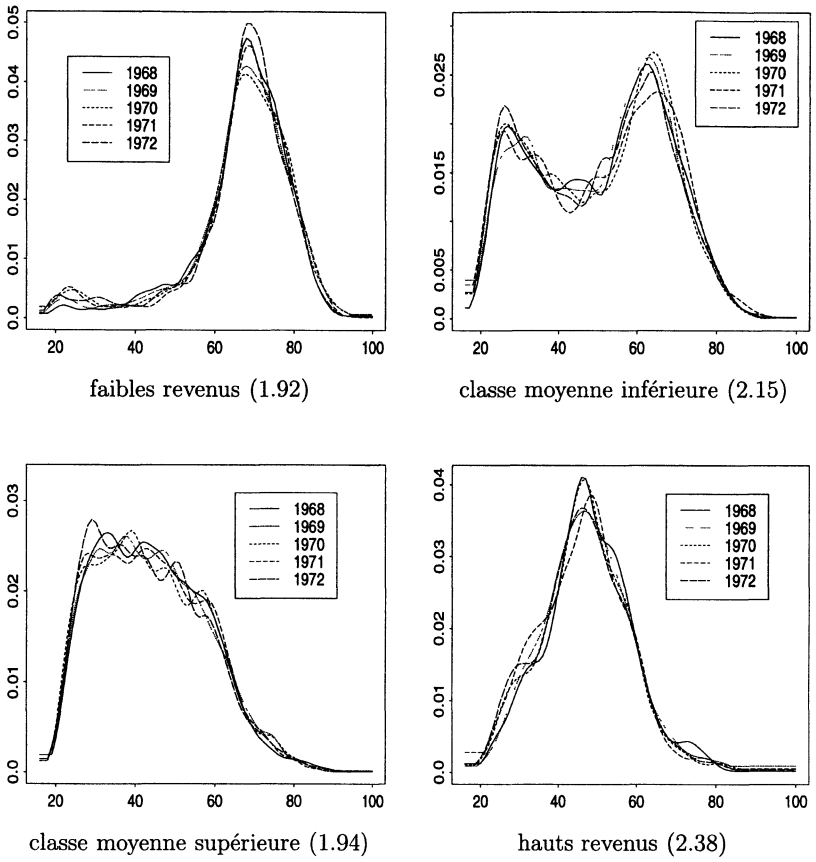


FIGURE 5.3
Population totale : densités de l'âge, pour six années consécutives, stratifiées en classes de revenus; les fenêtres optimales moyennes sont présentées entre parenthèses.

économique explique en partie l'augmentation du «pic caractérisant les ménages à faibles revenus» dans la distribution du revenu. En résumé, de façon similaire à notre discussion sur la distribution marginale de l'âge, nous pouvons tirer les conclusions qualitatives suivantes au sujet de l'évolution temporelle de la distribution conjointe de l'âge et du log-revenu standardisé :

a) Les distributions stratifiées de l'âge évoluent très lentement d'une année à l'autre, elles sont *à court terme approximativement invariantes dans le temps.*

b) A long terme, une tendance apparaît clairement pour le groupe des ménages à faibles revenus. Cette tendance doit être prise en compte dans une analyse de long terme.

c) Les distributions de l'âge présentent des différences *radicales* d'une classe de revenus à une autre. Ces différences sont beaucoup plus importantes que l'évolution temporelle à l'intérieur de chaque classe.

Les distributions conjointes du revenu et d'autres caractéristiques des ménages, telles que la taille de la famille et le statut professionnel, peuvent être étudiées de façon similaire. La nature discrète de ces variables simplifie même l'analyse. On peut montrer que les conclusions qualitatives énoncées en a) - c) demeurent valides lorsque l'on remplace l'âge par la taille de la famille ou le statut professionnel. Une analyse détaillée dépasse le cadre de cet article.

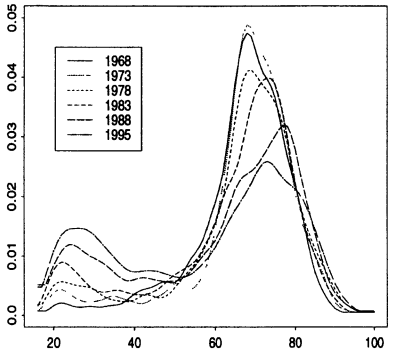


FIGURE 5.4
Ménages à faibles revenus : évolution des densités à long terme (2.31)

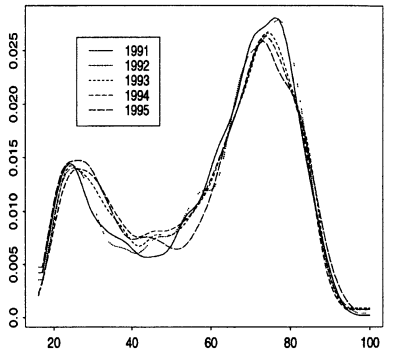


FIGURE 5.5
Ménages à faibles revenus : évolution des densités à court terme (2.56)

6. Conclusion

Pour étudier les distributions du revenu et de certaines caractéristiques des ménages, nous avons utilisé des méthodes d'estimation non paramétrique de la densité. Nous avions comme objectif principal de trouver des invariances dans l'évolution temporelle de ces densités après avoir appliqué certaines transformations adéquates. Dans ce contexte, l'utilisation de méthodes non paramétriques paraît naturelle puisque nous disposons de grands échantillons, et qu'aucune forme fonctionnelle de la distribution de la population posée *a priori* n'est acceptable. Il s'avère que l'estimation par la méthode des fonctions à noyau est bien meilleure que celle de l'histogramme. Le problème du choix de la fenêtre est crucial. On peut montrer que la méthode utilisée pour déterminer la fenêtre optimale est supérieure à tout autre choix «intuitif».

Les estimations de la densité bivariée sont visualisées graphiquement comme des surfaces dans des espaces à trois dimensions, uniquement pour donner une appréciation d'ensemble, bien qu'en général, une étude détaillée des densités multivariées soit menée par l'intermédiaire des densités marginales et conditionnelles.

La recherche de l'invariance est le thème de base de l'étude de l'évolution temporelle des densités. On cherche des transformations simples des variables de telle manière que les densités correspondantes deviennent «approximativement» identiques. Les problèmes de description, de modélisation et de prédiction des densités

qui évoluent dans le temps se réduisent de cette façon à l'étude des paramètres de ces transformations. Notons que dans la littérature économétrique, il y a, par exemple, plusieurs approches pour modéliser l'évolution temporelle du revenu moyen (voir Deaton, 1992).

Dans le cas particulier des densités du revenu, on obtient une approximation grossière de l'invariance simplement en standardisant les revenus par leur moyenne. Appliquée à certaines sous-populations, cette invariance est améliorée. Une meilleure invariance des densités du revenu est obtenue en standardisant le logarithme du revenu nominal. Cette transformation donne également des résultats pour la distribution conjointe du revenu et de l'âge.

Références

ATKINSON A.B. (1983). *The Economics of Inequality*. Clarendon Press, Oxford.

BLUNDELL R., PASHARDES P., WEBER G. (1993). What do we learn about consumer demand patterns from micro data? *The American Economic Review 83*, 570–597.

CEX. Institute for Social Research, University of Michigan, Consumer Expenditure Survey, United States Department of Labor, Bureau of Labor Statistics.

DEATON A. (1992). *Understanding Consumption*. Clarendon Press, Oxford.

DEATON A., MUELLBAUER J. (1980). An almost ideal demand system, *American Economic Review 70*, 312–326.

EBF. Enquête Budget de Famille (1979, 1984-85, 1989). Division «Condition de Vie des Ménages», Institut National de la Statistique et des études économiques, Paris.

ENGEL J., HERRMANN E., GASSER T. (1994). An iterative bandwidth selector for kernel estimation of densities and their derivatives, *Nonparametric Statistics 4*, 21–34

EPF. Instituto Nacional de Estadistica (INE), Encuesta de Presupuestos Familiares, Madrid, Spain.

ESCR Data Archive at the University of Essex, Family Expenditure Survey, Annual Tapes 1968-1986, Department of Employment, Statistics Division, Her Majesty's Stationary Office, London.

GOTTSCHALK P., SMEEDING T. (1997). Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature 35*, 633–687.

GSOEP Deutsches Institut für Wirtschaftsforschung, Berlin, German Socio-Economic Panel.

HILDENBRAND W., KNEIP A. (1999). Demand aggregation under structural stability. Journal of Mathematical Economics, 31, 81–109.

KEMSLEY W.F., REDPATH R.D., HOLMES M. (1980). Family Expenditure Survey Handbook, Her Majesty's Stationery Office, London.

KNEIP A., UTIKAL K.J. (1999). Inference for density families using functional principal component analysis, manuscript.

PSID. Institute for Social Research, University of Michigan, The Panel Study of Income Dynamics.

SCHMITZ H.P., MARRON J.S. (1992). Simultaneous estimation of several size distributions of income. *Econometric Theory 8*, 476–488

SCOTT D.W. (1992). *Multivariate Density Estimation*. John Wiley.

SILVERMAN B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SIMONHOFF J.S. (1996). *Smoothing Methods in Statistics*. Springer Verlag, New York.

UTIKAL K.J. (1996). *Invariant points of low dimensional curve families* Department of Economics, University of Bonn, SFB 303, Projektbereich A, Discussion Paper No. A–516

WAND M.P., MARRON J.S., RUPPERT D. (1991). Transformations in density estimation. *Journal of the American Statistical Association 86*, 343–361 (with discussion)