Vulnérabilité énergétique des ménages

Estimation du revenu disponible des ménages à partir de modèles M-Quantiles

Eric Durieux (*)

(*) Insee, PSAR Analyse Territoriale

Résumé

Afin d'étudier la vulnérabilité énergétique des ménages, le premier travail est de dénombrer et qualifier les ménages dans de telles situations, et ce à des niveaux géographiques relativement fins. Une approche classique de la mesure de la vulnérabilité énergétique consiste en la mesure du taux d'effort énergétique des ménages, c'est-à-dire la part du revenu disponible alloué au poste énergétique. Les ménages dont le taux d'effort dépasse un certain seuil sont considérés comme vulnérables énergétiquement.

Aucune source statistique disponible ne permet de faire cette estimation : soit les enquêtes sont nationales, difficilement utilisables à un niveau fin, soit elles sont incomplètes. L'approche qui a été retenue pour développer cette méthode consiste à utiliser le recensement de la population, seule source permettant de décrire les logements, les habitants et leur déplacements.

Nous avons développé une méthode permettant de modéliser chacune des composantes du taux d'effort, c'est-à-dire les revenus et les dépenses énergétiques (logement et déplacements), à partir des informations à notre disposition dans la source du recensement .

La méthode d'estimation des revenus est innovante dans la mesure où elle utilise des méthodes qui n'ont pas été conçues pour l'estimation ponctuelle : les régressions quantiles. Ces travaux sont inspirés de ceux de R. Chambers et N. Tzavidis qui ont proposé des modèles M-Quantiles pour l'estimation sur petit domaine. Après avoir rappelé les enjeux de l'estimation des revenus, nous présenterons l'utilisation des régressions quantiles et la construction d'un estimateur de type Chambers-Tzavidis.

Abstract

In the study of fuel poverty, the first task is to put a figure on the number of households concerned, and to characterize them on a localized scale. A classic approach of the measure of fuel poverty consists in estimating an energy expenditure to income ratio for each household. Households having a ratio over a determined threshold are considered as poor as regards energy.

There is no statistical source which allows to estimate ratio : either the surveys are national, so they can not be used easily on small scales, or they are incomplete. To reach our goal, we chose tu use the population census, which is the only source that describes the housings, the inhabitants and their journeys to work.

We developped a methodology that enables to estimate each component of that ratio: the disposable incomes and the energy expenses (housing and journeys to work). These estimations are based on the information available in the census.

The method of income estimation is innovative inasmuch as it calls on models that have not been developped for point estimation : quantile regressions. This study has been inspired by R. Chambers & N. Tzavidis's work : they proposed M-Quantiles models to do small areas estimations.

First, we will remind the questions raised by the income estimation, and then, we will explain how we used the quantiles regressions to build a Chambers-Tzavidis estimator.

Mots-clés

Économétrie des revenus, régression quantile, modèles différenciés

Introduction

Parce que l'objectif est de proposer une méthode qui soit utilisable quel que soit le territoire d'analyse, le recensement de la population s'impose comme la source d'entrée principale. En effet, pour chaque ménage, on dispose des caractéristiques de son logement, des habitants du logements et des déplacements domicile-travail pour les actifs occupés.

Cependant, estimer des revenus dans le recensement annuel de la population n'est pas une démarche nouvelle. La différence entre notre approche et celles n'ayant pu aboutir provient de notre objectif qui n'est pas d'obtenir des revenus à proprement parler mais de connaître la position des ménages autour d'une grandeur s'appuyant entre-autres sur les revenus, le taux d'effort énergétique. En effet, seul nous intéresse de savoir si la part des revenus alloués aux dépenses énergétiques dépasse ou non un certain seuil (8 % pour le chauffage du logement, 4,5 % pour le carburant des déplacements), que cela le dépasse fortement ou non importe peu. L'exercice est donc sensiblement différent de l'estimation de ressources stricto-sensu.

De plus, la thématique nous amène à supposer fortement que notre population d'étude sera à revenus faibles. Ainsi, nous pouvons nous orienter vers des méthodes nous assurant une meilleure estimation pour les ménages à bas revenus, nous préoccupant peu de la précision pour les hauts revenus.

1. L'estimation des revenus, un exercice périlleux ?

1.1 - L'absence de source de synthèse

Aucune source locale ne permettant de lier dépenses énergétiques et revenus au niveau individuel, l'étude de la précarité énergétique des ménages nécessite une estimation de ces paramètres, en particulier ici, **l'estimation de revenus disponibles par ménage**.

Une approche géographique s'appuyant sur les données carroyées pourrait-elle nous permettre d'imputer des revenus au niveau ménage dans le RP ?

Notre mesure ne s'effectuerait dans ce cas non plus au niveau ménage mais au niveau de carreaux (e.g. 200m). Notre objectif serait donc d'obtenir dans un premier temps des estimations de revenus fiscaux, via le dispositif des Revenus Fiscaux Localisés (RFL), au niveau des carreaux. L'estimation en sortie pour un carreau serait le revenu médian par exemple, ou une somme de revenus.

Cette méthode présente certains avantages :

- La source RFL propose des revenus fiscaux déclarés au carreau réels et non une estimation. Si on arrive à estimer un revenu disponible au carreau, on peut espérer avoir une donnée relativement fiable à cette échelle.

- Lien avec la future source localisée des revenus - Filosofi qui permettrait de travailler au carreau cette fois sur un revenu disponible plus fiable, grâce au travail effectué sur les revenus complémentaires localisés.

Néanmoins certaines limites liées à cette approche se dessinent rapidement :

- Elle suppose une forme de « ségrégation spatiale » : il faut s'assurer de la faible dispersion des revenus au sein de mêmes carreaux (tests de variance au niveau des carreaux, intervalle inter quantile) pour ensuite imputer aux ménages-recensement du carreau le revenu médian par exemple observé sur l'ensemble du carreau. On perd en finesse par rapport à l'approche individuelle car les revenus se retrouvent lissés sur l'ensemble d'un carreau. Concrètement, soit on suppose que les revenus des ménages sont semblables au sein d'un carreau (en imputant un revenu médian à tous les ménages du carreau), soit on est amené à les lier (par une somme de revenus sur l'ensemble du carreau, le taux d'effort sera alors identique au sein d'un carreau). Ces deux possibilités présentent des limites dans l'interprétation finale du résultat.

- La source utilisée serait les Revenus Fiscaux Localisés - RFL. Nous n'avons pas dans cette source les compléments nécessaires au calcul d'un revenu disponible (RSA, APL, minimum vieillesse, API, AAH, allocations familiales et autres revenus complémentaires), ce qui est problématique concernant les bas revenus qui nous intéressent particulièrement. Il faudrait ainsi modéliser au niveau du carreau l'ensemble des revenus sociaux.

- Le passage de la source fiscale au recensement est également problématique, le recensement n'étant pas significatif à l'échelle fine qu'est le carreau pour les communes de plus de 10.000 habitants. Le biais introduit au moment du rapprochement de ces 2 sources risquerait d'être très important.

Ces limitations nous ont ainsi incités à poursuivre l'exploration des méthodes d'imputation de revenus disponible.

1.2 Les limites de la régression classique

De nombreux travaux présentent des estimations de revenus à partir de modèles plus ou moins complexes. Parmi ces travaux, ceux réalisés à l'Insee dans le cadre d'une démarche expérimentale sur une région illustrent toute la difficulté de l'exercice.

Le revenu disponible des ménages a été estimé à partir d'un modèle construit sur une enquête nationale - ERFS[1]. Les variables retenues sont :

- la catégorie socioprofessionnelle de la personne de référence (CS niveau 3) et du conjoint (CS niveau 1),

- le nombre d'individus du logement,

---

[1] *Enquête sur les Revenus Fiscaux et Sociaux pour une année donnée celle-ci fournit une évaluation du revenu disponible des ménages pour l'année considérée, enrichie des informations sociodémographiques de l'enquête Emploi de l'Insee. L'enquête Revenus fiscaux et sociaux (ERFS) consiste en un appariement statistique du fichier de l'enquête Emploi en continu (données du 4ème trimestre de l'année N) avec les fichiers fiscaux (déclarations des revenus) de la direction générale des Finances publiques (DGFiP) de l'année N et les données sur les prestations perçues au cours de l'année N collectées auprès de la caisse nationale des allocations familiales (CNAF), de la caisse nationale de l'assurance vieillesse (CNAV) et de la caisse centrale de la mutualité sociale agricole (CCMSA). L'enquête Revenus fiscaux et sociaux permet de déterminer quels sont les types de revenus perçus par le ménage :*

- les revenus individuels perçus par chaque membre des ménages : salaires, pensions, retraites, indemnités de chômage, bénéfices agricoles, industriels, commerciaux et non commerciaux ;

- les revenus non individualisables : les prestations sociales (prestations familiales, prestations logement et minima sociaux) ainsi que les revenus du patrimoine ;

- les impôts acquittés par les ménages (impôt sur le revenu, taxe d'habitation et prime pour l'emploi). L'enquête ERFS vise à analyser les revenus suivant des critères sociodémographiques usuels (catégorie socioprofessionnelle et âge des personnes composant le ménage, taille du ménage, activité de chaque individu etc.) et à mesurer le niveau de vie et la pauvreté monétaire des personnes.

France métropolitaine, en ménages dits " ordinaires " dont le revenu déclaré au fisc est positif ou nul et dont la personne de référence n'est pas un étudiant.

- l’activité de la personne de référence (actifs occupés ou non),
- son temps de travail (temps partiel/temps complet),
- sa tranche d’âge,
- son diplôme,
- le statut d’occupation du logement et
- la localisation du ménage (dans une unité urbaine ou non).

Les données de plus de 30 000 ménages résidant en France (hors Île-de-France) ont ainsi été exploitées pour élaborer cette équation log-linéaire.

$$Log(\mathrm{Re}v) = \sum_i \alpha_i X_i$$

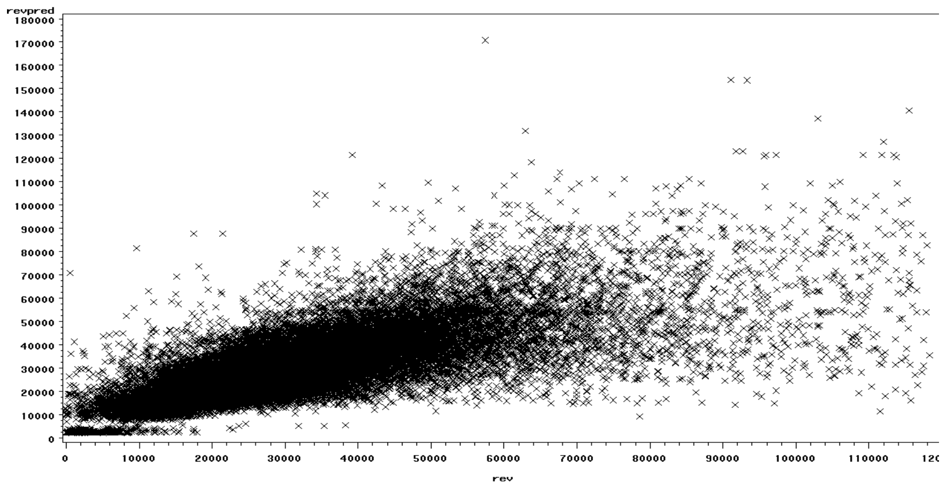
(X_i représentant l'ensemble des variables explicitées ci-dessus)

La modélisation des revenus a permis de simuler de façon satisfaisante (écart inférieur à 13 %) les déciles des revenus des ménages dans les départements tels que mesurés dans la source RDL (aucun calage effectué). Néanmoins, les étudiants ont été intégrés alors qu’une faible part d’entre eux seulement déclarent des revenus.

Le revenu disponible est le revenu à la disposition du ménage pour consommer et épargner. Comme c’est le cas dans les études sur cette source, les ménages dont le revenu déclaré est négatif ont été supprimés (on considère que disposer d’un revenu déclaré négatif ne peut correspondre au revenu réel sur la période). En revanche, les ménages dont la personne de référence est étudiante n’ont pas été supprimés alors qu’en général ils le sont dans les études (les aides versées par les parents sont mal appréhendées par les Enquêtes Revenus Fiscaux et Sociaux, alors qu’elles jouent un rôle important dans le niveau de vie de ces ménages).

Cette méthode affiche un avantage de simplicité et de "terrain connu" dans le cadre d'imputation par modélisation économétrique. En revanche une hypothèse importante des modèles linéaires est que les résidus (erreurs) sont homoscédastiques, c’est à dire, qu’ils ont tous la même variance. L’analyse des résidus, par visualisation graphique et régression des résidus sur les mêmes variables explicatives, laisse apparaître une hétéroscédasticité.

<u>Revenu prédit et revenu observé</u>
(extrait du nuage de points complet)



Ce graphique des résidus de la régression témoigne de l'hétérogénéité des résidus. Des corrections ont été apportées pour tenter de la corriger :

Nous avons testé la suppression des valeurs extrêmes et des observations influentes. Nous améliorons légèrement le modèle mais il y a toujours une surestimation encore des bas revenus et une sous-estimation des plus hauts.

Pour aller plus loin, l'enquête ERFS a été post stratifiée pour construire des modèles sur des populations plus homogènes :

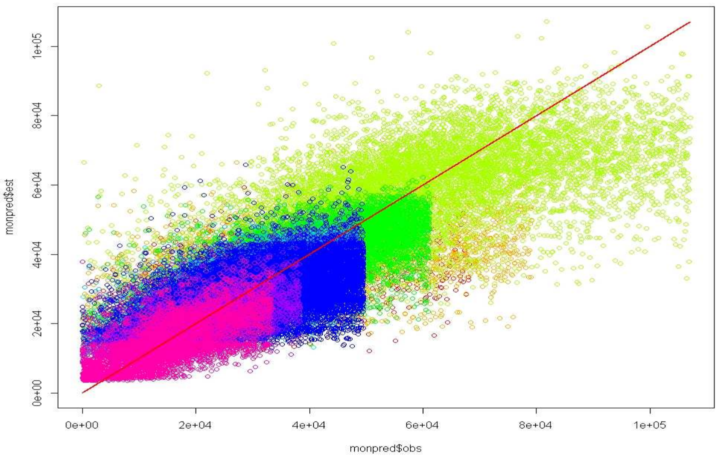
- actifs occupé par CS (6 modèles) ;
- chômeurs (1 modèle) ;
- retraités (1 modèle) ;
- autres (inactifs yc étudiants – 1 modèle)

Au total, 9 strates ont été construites. De plus, d'autres modifications ont été apportées au modèle précédent :

- ajout de variables selon les modèles (e.g. nombre d'actifs occupés du ménage, NAF 17 postes, pôle urbain/périurbain/rural, recherche d'emploi moins de 1 an ou plus), regroupement de certaines modalités
- choix d'un modèle poissonnien (de manière à conserver l'idée d'une régression de type logarithmique mais en évitant « l'explosion » des résidus et des estimations au passage à l'exponentielle)
- empilement des ERFS 2008-2009-2010
- modèle pondéré (la population enquêtée peut être différente de la population en général)
- suppression de valeurs extrêmes

L'application de modèles séparés sur des segments différents permet d'améliorer les estimations de revenus mais nous conservons une surestimation des bas revenus et sous-estimation des hauts revenus. Ci-dessous quelques résultats :

Revenus observés*revenus prédits (les couleurs correspondent aux modèles)



2. L'apport de la régression quantile

Pour répondre à 2 difficultés :

Corriger l'hétéroscédasticité

Sortir de la « dictature de la moyenne » imposée par la régression linéaire,

nous nous sommes tournés vers la régression quantile.

Les quantiles répondent en partie à ses préoccupations car ils sont beaucoup moins sensibles aux valeurs extrêmes. Une autre propriété intéressante est l'équivalence des quantiles par transformation monotone : le logarithme du 3ème décile d'une variable X est égale au 3ème décile du logarithme de la variable X :

$$Log(d_3(X)) = d_3(Log(X))$$

2.1 - Principes de la régression quantile

Pour une variable aléatoire Y de distribution F telle que F(y)=P(Y<y), le j^{ème} quantile est défini par $P(Y < q_j(Y)) = j$

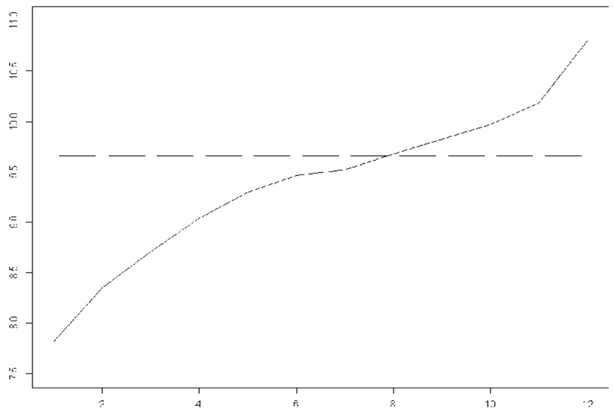
On s'intéresse plus précisément aux quantiles conditionnels à un ensemble de variable X, notés : $q_j(Y/X)$

Le principe de la régression quantile est de modéliser chaque quantile par une combinaison linéaire des X :

$q_j(Y/X) = X\beta_j$

Cette modélisation assure de tenir compte du fait que les déterminants de X ont un impact sur la distribution F, ce qui ne pourrait être capté par un modèle de régression linéaire qui vise à modéliser la moyenne : $E(Y/X) = X\beta_j$

Exemple de valeur de la constante pour chaque quantile / modèle linéaire (droite en pointillé) :



Les régressions quantiles permettent de décrire comment le décile se modifie en fonctions de certaines variables explicatives. En revanche, elles ne mesurent pas l'impact de ces variables sur les individus.

Initialement conçue pour décrire une distribution conditionnellement à un ensemble de variables explicatives, la régression quantile a été utilisée dans un domaine particulier, celui de l'estimation d'une variable dans le cadre d'un sondage (méthode M-quantile). Cette utilisation nécessite de mesurer l'effet causal de variables sur la distribution et donc d'adapter la méthode.

2.2 - Les modèles M-quantile de Chambers et Tzavidis

La méthode M-quantile proposée par Chambers et Tzavidis a pour objectif de produire des estimations sur des petits domaines basées sur des relations de type régression quantile. Ainsi, on cherche à estimer sur une population la variable y (le revenu dans notre cas). La méthode consiste à stratifier à posteriori l'échantillon des répondants pour lesquels on connaît le couple (x_i,y_i), x_i représentant un ensemble de caractéristiques connues également sur les individus non échantillonnés.

Le modèle de régression quantile s'écrit : $q_j(Y/X) = X\beta_j$. On suppose que le modèle sous-jacent est de type linéaire. Pour chaque quantile on a un jeu de coefficients, et pour chaque x_i on a un quantile q_j estimé.

On estime des droites de régressions quantiles sur l'ensemble de l'échantillon $Y = X\beta_q$, $q \in [0,1]$, Y représente le revenu. Pour chaque couple (y_i, x_i), on cherche la droite qui passe exactement par ce point.

Pour chaque strate, on fait l'hypothèse que les individus sont suffisamment homogènes au regard des caractéristiques X. Plus précisément, les individus sont situés sur le même niveau de distribution de Y conditionnellement à X.

Chambers et Tzavidis construisent ensuite un estimateur de Y en prenant une fonction des coefficients des régressions quantiles dans chacune des strates. Pour chaque strate j on a l'estimateur :

$$\hat{m}_j = \frac{1}{N_j}\left(\sum_i y_i + \sum_i x_i \hat{\beta}_q(\hat{\theta}_j)\right)$$

$\hat{\theta}_j$ représente la moyenne des estimations issues des régressions quantiles dans la strate j pour tous les individus.

Une approche alternative est également proposée, consistant à prendre la moyenne des coefficients $\hat{\beta}_q$ dans chaque strate puis d'utiliser ce nouveau coefficients $\hat{\hat{\beta}}$ à tous les individus de la même strate. Les simulations réalisées par les deux auteurs tendent à montrer de très fortes similitudes entre les deux approches. Pour estimer les revenus, nous avons testé les deux méthodes qui donnent des résultats identiques. Au final, nous avons retenu la première, la moyenne des estimations de chaque individus dans chaque strate.

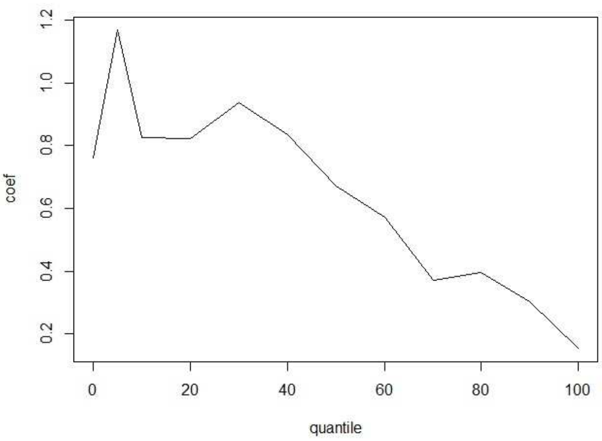
Dans leur article, Chambers et Tzavidis suggèrent de prendre la moyenne ou la médiane mais tout autre fonction est également possible. C'est sur cette fonction que nous avons particulièrement travaillé, pour répondre à notre objectif d'estimation de revenus.

2.3 - Le modèle M-quantiles appliqué à l'estimation des revenus

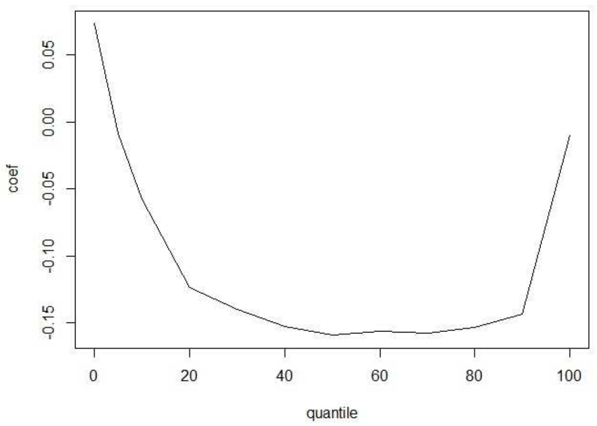
L'échantillon de l'enquête ERFS a été stratifié en 17 groupes, chaque groupe présentant une assez forte homogénéité des revenus et une structure très différente des autres.

Effets différenciés d'une variable (diplôme) selon le quantile modélisé

Agriculteurs (strate 14)



Cadres (strate 4)



Les 17 domaines d'estimation de l'enquête ERFS :

Strate	*Taille de l'échantillon*
Strate 1	41 212
Strate 2	16 281
Strate 3	13 453
Strate 4	10 892
Strate 5	10 269
Strate 6	7 670
Strate 7	5 484
Strate 8	4 330
Strate 9	4 192
Strate 10	2 789
Strate 11	2 022
Strate 12	1 998
Strate 13	1 526
Strate 14	1 391
Strate 15	1 084
Strate 16	726
Strate 17	501

Nous avons réalisé notre régression quantile sur l'enquête ERFS stratifiée. Les quantiles estimés sont les 9 déciles plus le 5ième centile.

On récupère en sortie de la régression un modèle et des coefficients :

$$\hat{Q}_j = \sum_i \alpha_i^j X_i$$

(α^j_i= coefficients estimés du modèle du quantile j et X_i = caractéristiques connues de l'individus i)

L'objectif est maintenant de passer de nos 10 modèles quantiles à un modèle unique pour estimer un revenu et non une distribution. Pour cela, on s'inspire des travaux de Chambers et Tzavidis :

pour chaque modèle j, on a des coefficients α^j_i se rapportant à des observables X_i. Nous cherchons donc une fonction F telle que $F(\alpha^1_i, \alpha^2_i \ldots)=\beta_i$, l'objectif final étant effectivement d'estimer un revenu et non une distribution de revenus. Ce procédé correspond à la première approche de Chambers et Tzavidis.

On se retrouve alors avec un nouveau modèle d'estimation de revenus, construit à partir de nos 10 modèles quantiles, modèle de la forme :

$$\hat{Revenu} = \sum_i \hat{\beta}_i X_i$$

Quel choix pour de la fonction F ? Un choix possible serait la moyenne (ou la médiane) des coefficients, comme c'est le cas dans l'exemple de Chambers. Ce choix n'est pas le plus judicieux dans notre cas car cela risquerait de trop lisser nos revenus et comme nous souhaitons conserver une dispersion, nous lui avons préféré une fonction liée à la distribution des revenus : fonction de type moyenne pondérée des coefficients.

Pondérée par quoi ? C'est l'objectif de l'utilisation de la régression multinomiale : celle-ci est réalisée sur ERFS et sert à modéliser l'appartenance d'un revenu aux intervalles inter quantiles basés sur les quantiles retenus à l'étape 1 (c'est à dire déciles + 5ème centile). Les variables (X_i) injectées dans le modèle sont identiques à celles de l'étape 1. Le positionnement des individus dans la distribution des revenus est conditionné par leurs caractéristiques individuelles. On passe donc par un modèle logit qui modélise le positionnement des individus dans les différents quantiles.

On récupère les probabilité d'appartenance aux intervalles inter-quantiles :

P_i^j=probabilité que l'individu i soit sans l'intervalle borné supérieurement par Q_j

et donc $\hat{\beta}_i = \sum_j P_i^j \alpha_i^j$ pour l'individu i

Ensuite, sur le RP, on applique le modèle multinomial pour récupérer l'ensemble des probabilités d'appartenance pour chaque individu. Grâce à ces probabilités, on calcule les coefficients β_i à partir des coefficients des modèles quantiles (cf. supra). Puis on calcule le revenu estimé de chaque individu du RP.

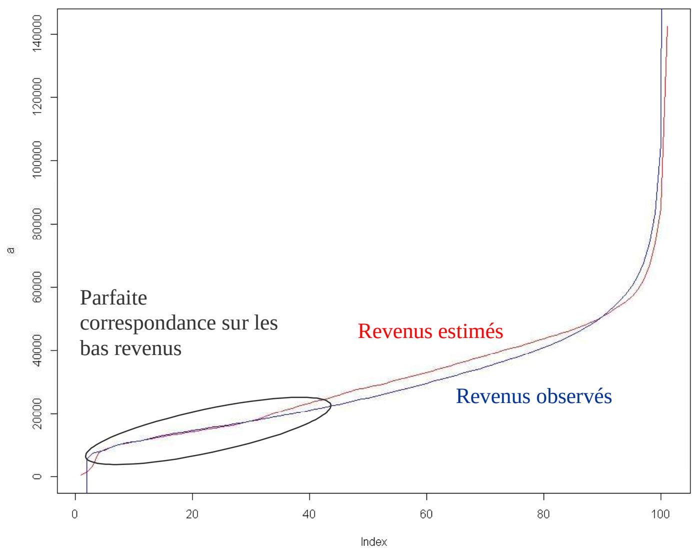
2.4- Validation de la méthode

Deux types de validation ont été menés, l'un consistant à comparer les distributions de revenus estimés avec ceux mesurés par des sources fiscales et à des niveaux géographiques fins. L'autre source de validation tient dans la robustesse du modèle.

2.4.1 Comparaison des distributions estimées puis observées

Le premier niveau de validation consiste à comparer la distribution des revenus estimées avec celle obtenue à partir d'une autre source, fiscale, RDL (revenus disponibles localisés).

Comparaison des distributions observées et estimées



Les distributions observées et estimées par le modèle M-Quantiles sont très proches, surtout pour les revenus inférieurs à la médiane. Les modèles linéaires ont de bonnes estimations uniquement au niveau de la médiane.

	Distribution observée	Distribution estimée – M-quantiles	Distribution estimée modèles linéaires simples par strates
5%	10 027	10 166	12 950
10%	12 392	11 855	14 909
20%	15 922	14 843	17 521
30%	19 128	18 391	20 067
40%	22 609	23 428	23 588
50%	26 345	28 333	26 672
60%	30 438	32 746	29 663
70%	34 953	38 004	33 744
80%	40 436	43 730	38 080
90%	47 983	52 227	45 796

2.4.2 Validation par réplication d'échantillons (bootstrap)

De manière à tester la robustesse de la méthode, nous avons implémenté une méthode type bootstrap : nous avons isolé 10% des observations ERFS pour chaque segment, calculé les modèles M-quantiles sur les 90% restants puis appliqué ces modèles aux observations isolées. Nous avons appliqué cette méthode une vingtaine de fois. Dans tous les cas, les résidus observés sur les observations isolées étaient semblables (valeurs + dispersion) à ceux obtenus par le modèle général.

3. Discussion

La régression quantile est un outil statistique puissant pour mesurer l'impact d'une variable sur une distribution. Combinée à d'autres méthodes, elle permet de faire de l'estimation ponctuelle tout en respectant la distribution d'une variable. Son apport sera d'autant plus important que la variable à estimer présente une grande dispersion au sein de la population.

Contrairement à la méthode Chambers et Tzavidis, nous avons restreint le nombre de quantiles et donc le nombre de régressions car ce n'est pas toute la distribution des revenus qui nous importait mais seuls les premiers déciles. Ceci a certainement un impact sur l'estimation que nous n'avons pas mesuré.

Par ailleurs, l'objectif de l'étude est d'estimer un ratio de type dépenses/revenus. La variable d'intérêt n'est donc pas le revenu mais son inverse. Ainsi, obtenir une estimation sans biais du revenu ($E(\hat{Y}) = Y$) ne garantit pas une estimation sans biais de son inverse en raison des propriétés de l'espérance ($E(1/\hat{Y}) \neq 1/E(\hat{Y})$).

En revanche, les régressions quantiles permettent une estimation de l'inverse du revenu car les quantiles sont conservés par les transformations monotones.

Bibliographie

[1] La régression quantile en pratique, P.Givord & X. D'Haultfoeuille, actes JMS 2012

[2] M-quantile models for small area estimation, R. Chambers & N. Tzavidis

[3] M-ESTIMATION, CONVEXITY AND QUANTILES, V.I. Koltchinskii

[4] Using quantile regression to explore the distribution of Contextual Value Added across London, N. Tzavidis & J.J. Brown

[5] M-Quantile Small Area Estimation, Poverty Mapping and Geographically Weighted Regression, Ray Chambers (Wollongong) & Nikos Tzavidis (Manchester) & Nicola Salvati & Monica Pratesi (Pisa)

[6] Vulnérabilité énergétique : Loin des pôles urbains chauffage et carburant pèsent fortement dans le budget, N. Cochez & E. Durieux & D. Levy, Insee-Première 1530, janvier 2015