Calorie Intake and Income in China: New Evidence using Semiparametric Modelling with Generalized Additive Models.

Trinh Thi Huong[1], Michel Simioni[2] & Christine Thomas-Agnan[3]

[1] *TSE, 21 allée de Brienne, 31000 Toulouse et trinhthihuong@vcu.edu.vn*
[2] *SupAgro, 1 place Pierre Viala, 34060 Montpellier et michel.simioni@supagro.inra.fr*
[3] *TSE, 21 allée de Brienne, 31000 Toulouse et christine.thomas@tse-fr.eu*

Résumé. L'existence d'une relation entre l'apport calorique des repas d'un ménage et son revenu joue un rôle important lors de l'évaluation de l'intérêt de mettre en place des politiques pour améliorer l'alimentation des plus défavorisés dans les pays en voie de développement. De nombreux travaux ont été consacrés à l'estimation de cette relation et cela pour de nombreux pays, comme résumé dans Ogundari et Abdulai (2013). La plupart de ces travaux utilisent des spécifications paramétriques et seul un petit nombre de travaux s'intéressent à la non linéarité de la relation étudiée. Dans cet article, nous contribuons à cette littérature en utilisant les apports récents quant à l'estimation de modèles additifs généralisés en utilisant la régression par splines pénalisées. Les spécifications semi paramétriques utilisées mixent des fonctions paramétriques et non paramétriques des variables et permettent ainsi de capturer des non linéarités sans contraindre la forme des fonctions étudiées. De plus, différentes distributions pour la variable de réponse sont envisagées, nous permettant d'élargir la gamme des spécifications étudiées au delà du modèle gaussien. Les données extraites de l'enquête chinoise "Chinese Health and Nutrition Survey" pour les années 2006, 2009 et 2011, sont utilisées pour estimer la relation entre l'apport calorique des repas d'un ménage et son revenu, et les résultats obtenus pour différentes spécifications paramétriques et semi paramétriques sont comparés et discutés en utilisant le test de comparaison des performances révélées de deux modèles proposé récemment par Parmeter et Racine (2015).

Mots-clés. Apport Calorique, Revenu, Modélisation semi paramétrique, Modèle Additif Généralisé, Performance Révélée , Chine.

Abstract. The knowledge of the relationship between calorie intake and income is critical for assessing the interest of implementing policies to improve the nutrition of the poor in developing countries. The estimation of this relationship has been the subject of many empirical studies involving many different countries as surveyed by Ogundari and Abdulai (2013). Most studies use parametric models and only few papers use semiparametric models to deal with the issue of non linearity. Our paper aims at contributing to this literature by using recent advances in the estimation of generalized additive models with penalized spline regression smoothing. These semiparametric models enable mixing

parametric and nonparametric functions of the explanatory variables and thus capturing nonlinearities in some variables without restricting their shapes. In addition, the distribution of response variable in the generalized additive models belongs to the exponential family which enlarges the classical family of models. Data from the "Chinese Health and Nutrition Survey" for the years 2006, 2009 and 2011, are used to investigate the relationship between calories intake and income in China and results from different specifications, either parametric or semiparametric, are compared and discussed using the recent revealed performance test of Parmeter and Racine (2015).

Keywords. Calorie Intake, Income, Semiparametric Modelling, Generalized Additive Model, Revealed Performance Test, China.

1 Résumé long

Les évolutions des modes de consommation alimentaire et la sédentarité croissante des individus sont à l'origine d'une épidémie d'obésité dans les pays développés. Les pays en voie de développement connaissent actuellement la même transition alimentaire que celle qu'ont connu les pays développés, mais en plus d'une croissance du nombre des obèses, ils font toujours face au problème de la malnutrition. Un des outils traditionnellement utilisé par les gouvernements de ces pays pour lutter contre la malnutrition consiste en des transferts soit monétaires, soit en biens de consommation courante. Parfois ces mêmes gouvernements subventionnent les prix de ces biens (par exemple, le riz aux Philippines ou encore la pain en Egypte) pour les ménages les plus pauvres. La mise en oeuvre de telles politiques repose sur l'hypothèse implicite qu'il existe une relation croissante entre la disponibilité en termes de calories pour un ménage et son revenu. De nombreux travaux en économie du développement se sont alors focalisés sur l'évaluation de l'élasticité-revenu de l'apport calorique des repas. Cette littérature génère des résultats contrastés, la relation entre la prise calorique d'un ménage et son revenu pouvant être caractérisée comme croissante et significative dans certains travaux alors qu'elle ne l'est pas dans d'autres.

Dans une étude récente, Ogundari et Abdulai (2013) recensent 40 travaux pour différents pays et proposent une méta-analyse de leurs résultats, soit une analyse des sources de la variabilité de 99 valeurs estimées de l'élasticité-revenu de la prise calorique. Il faut alors noter que sur ces 99 valeurs estimées, 86 proviennent de l'estimation d'un modèle de régression gaussien paramétrique où le logarithme de la prise calorique par individu dans un ménage est regressé sur le logarithme du revenu et diverses variables de contrôle définies au niveau du ménage. Quand une étude envisage la possibilité d'une relation non linéaire entre les deux variables d'intérêt, celle-ci est capturée par l'introduction du carré du logarithme du revenu comme variable explicative dans le modèle de régression. Il faut aussi noter que les 13 valeurs estimées de l'élasticité restantes résultent de l'estimation

d'autres spécifications ou encore de spécifications semi paramétriques additives (voir, par exemple, Gibson et Rozelle, 2002).

Notre étude s'inscrit dans la suite de ces travaux. Plus précisèment, on propose de revisiter l'estimation de la relation entre prise calorique et revenu en estimant des modèles généralisés additifs de la forme:

$$g\left(E(Y_i|X_i^*, Z_i)\right) = X_i^{*\prime}\beta + \sum_j f_j(Z_{ji}) \tag{1}$$

où

- Y_i est la prise calorique par individu dans un ménage i. Cette variable suit une distribution donnée dans la famille des distributions exponentielles,
- $g(.)$ est une fonction de lien de forme connue,
- X_i^* représente le vecteur des variables explicatives qui agissent de façon linéaire sur $g\left(E(Y_i|X_i^*, Z_i)\right)$,
- le vecteur de paramètres β mesurent l'impact des variables du vecteur X_i^*,
- Z_i représente le vecteur des variables qui agissent de façon non linéaire sur $g\left(E(Y_i|\right.$ $\left.X_i^*, Z_i)\right)$, et
- chaque fonction $f_j(.)$ est une fonction de forme inconnue qui mesure l'impact de la $j$ième composante du vecteur Z_i.

Les modélisations décrites par l'équation (1) englobent de nombreuses spécifications possibles dont le modèle gaussien usuel où $g(.)$ est l'identité, Y_i (dans les applications, le logarithme de la prise calorique par individu) est supposé être distribué selon une loi normale, et toutes les variables entrent de façon linéaire (en particulier, le logarithme du revenu par individu du ménage). Ces modélisations ont été bien étudiées dans la littérature (Wood, 2006). Leur nature semi paramétrique et additive permet de s'abstraire du problème de la malédiction de la dimension propre aux modélisations purement non paramétriques, tout en ne faisant pas d'hypothèses paramétriques quant à l'effet de certaines variables explicatives (ici, le revenu par individu du ménage). Finalement, ces modélisations peuvent être aisèment estimées en approximant les fonctions inconnues par des splines et en introduisant une pénalisation dans le critère à optimiser pour éviter un surajustement.

Plusieurs modèles généralisés additifs de type (1) sont estimés sur les données extraites de l'enquête chinoise "Chinese Health and Nutrition Survey" pour les années 2006, 2009 et 2011. Pour choisir entre ces modèles, nous implémentons le test de performance révélée récemment proposé par Parmeter et Racine (2014). Ce test consiste en

l'estimation pour chaque modèle de la "vraie erreur" moyenne comme mesure de sa performance (Efron, 1982). Considérons ainsi un échantillon sur lequel un modèle donné a été estimé, échantillon dit de calibrage. Il s'agit alors de calculer l'erreur qui est commise quant à la prédiction des valeurs pour un nouvel échantillon, ou échantillon d'évaluation, les prédictions étant calculées en utilisant le modèle estimé sur l'échantillon de calibrage et l'erreur étant calculée en utilisant une fonction de perte. La valeur de l'erreur ainsi obtenue est propre à l'échantillon d'évaluation utilisé. Or, il serait souhaitable d'avoir cette mesure pour toute réalisation possible de l'échantillon d'évaluation, cela en calculant l'espérance mathématique de la vraie erreur. Sur un jeu de données réelles de n observations, une estimation de celle-ci peut être obtenue en répliquant un grand nombre de fois une procédure qui consiste à tirer sans remise n_1 observations, à estimer le modèle sur ces observations, à prédire les valeurs pour les $n_2 = n - n_1$ observations restantes, et finalement, à calculer l'erreur commise. La moyenne des erreurs ainsi obtenues pour toutes les réplications fournit un estimateur de l'espérance mathématique de la vraie erreur. Deux modèles concurrents peuvent alors être comparés sur la base de ces "vraies erreurs" moyennes. La pertinence de ce critère de choix de modèles est évaluée sur la base de simulations dans le papier.

La mise en oeuvre de l'approche proposée sur les données chinoises permet de conclure en faveur d'un modèle généralisé additif où la fonction de lien est la fonction log, la prise calorique suit une loi binomiale négative, et l'effet du revenu par individu du ménage est capturé par une fonction de forme inconnue. Ce choix est le même quelle que soit l'année considérée. De plus, le modèle linéaire gaussien "$\log - \log$", avec carré du logarithme du revenu par individu du ménage pour capturer l'effet non linéaire de cette variable, largement utilisé dans la littérature empirique, apparaît être celui pour lequel la performance révélée est la plus faible et cela quelle que soit l'année considérée. Pour le modèle choisi, les valeurs estimées de l'élasticité-revenu de la prise calorique sont petites, variant entre 0 et 0.15, mais en cohérence avec celles trouves par Nie et Sousa-Poza (2016). Les courbes décrivant l'évolution de l'élasticité-revenu de la prise calorique en fonction du revenu pour les différentes années présentent une forme de W inversé.

Bibliographie

[1] Efron, B. (1982), *The Jacknife, the Bootstrap and Other Resampling Plans*, CBMS 38, Society for Industrial and Applied Mathematics.
[2] Gibson, J. et Rozelle, S. (2002), How Elastic is Calorie Demand? Parametric, Nonparametric, and Semiparametric Results for Urban Papua New Guinea, *Journal of Development Studies*, 38: 23 – 46.
[2] Nie, P. et Sousa-Poza, A. (2016), A Fresh Look at Calorie-Income Elasticities in China, *China Agricultural Economic Review* 8: 55-80.

[3] Ogundari, K. et Abdulai, A. (2013), Examining the Heterogeneity in Calorie-Income Elasticities: A Meta-Analysis, *Food Policy* 40, 119 – 128.
[4] Racine, J.S. et Parmeter, C.P. (2014), Data-Driven Model Evaluation: A Test for Revealed Performance, in: J.S. Racine, L., Su, L. et Ullah, A., *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, Oxford University Press, Oxford.
[5] Wood, S. (2006), *Generalized Additive Models: An introduction with R*, Chapman and Hall/CRC.