A new perspective on the relationship between calorie intake and income in China and Vietnam using semiparametric modeling

Thi Huong Trinh, Michel Simioni, Christine Thomas-Agnan

A new perspective on the relationship between calorie intake and income in China and Vietnam using semiparametric modeling

Trinh Thi Huong*, Simioni Michel**, Thomas-Agnan Christine*

* TSE, Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

** INRA, UMR 1110 MOISA, F-34398 Montpellier, France

Contribution presented at the XV EAAE Congress 'Towards Sustainable Agri-Food Systems: Balancing Between Markets and Society'

August, 29th – September 1st, 2017

Parma, Italy

UNIVERSITÀ
DI PARMA

A New Perspective On The Relationship Between Calorie Intake And Income In China And Vietnam Using Semiparametric Modeling

February 4, 2017

Abstract

This paper revisits the issue of estimating the relationship between calorie intake and income, and presents and compare estimates of this relationship for China and Vietnam. Semiparametric generalized additive models are estimated and their performances are compared to the performance of the classical double log model using the revealed performance test of Racine and Parmeter (2014). This methodology is implemented using successive waves of the Chinese Health and Nutrition Survey and of the Vietnam Household Living Standard Survey. The application delivers some new and interesting insights on nutritional transition in the two countries between 2004 and today.

Keywords: Nutrition, Income, Semiparametric Econometrics, China, Vietnam

JEL Classification: I31, J21, J22

1 Introduction

Policies aimed at reducing starvation and redressing nutritional deficiencies remain among the most widely accepted policies in the world as emphasized by Banerjee (2016). These policies can take many different forms, from subsidized prices of basic foodstuffs to cash transfers, and their effectiveness depends on the existence of a sensitivity of food demand to income variation and its magnitude. Several papers in development economics were thus interested in the estimation of the relationship between food demand measured in calories and household income. Ogundari and Abdulai (2013) recently summarized this literature by providing a meta analysis of results given in forty papers published on this issue for several countries in the world, i.e. covering 99 estimated calorie intake income-elasticities. Unlike other literatures in applied econometrics , the choice of the functional form to characterize the relationship between calorie intake and income does not seem much discussed. 86 estimated income-elasticities were thus obtained using the parametric double-log specification. Nonlinearity is sometimes introduced through the addition of the square of the logarithm of income, after having checked the concavity of the relationship between logarithm of calorie intake and logarithm of income with nonparametric regression tools (Abdulai and Aubert 2004). Only few papers use semiparametric specifications to deal with the issue of nonlinearity (see, for instance, Gibson and Rozelle 2002, Vu 2009a, Tian and Yu 2015, Nie and Sousa-Poza 2016). The last two papers investigate the relationship for China and use the same data sets but produce conflicting results. On one hand, Nie and Sousa-Poza (2016) suggests that "no clear nonlinearity, regardless of whether parametric, nonparametric, or semiparametric approaches are used," while, on the other hand, Tian and Yu (2015) claim that "nutrition improvement and dietary change will continue in China but will slow down in the future with further income growth."

Our paper aims to contribute to the aforementioned literature in several ways. First, we want to compare changes over time in the relationship between calorie intake and income for China and Vietnam. We can think that China began its nutritional transition much earlier than Vietnam, as China began economic reforms ten years before Vietnam. The relationship between calorie intake and income may have converged to a similar shape in both countries, as both countries have undergone profound changes over the last 30 years. To our knowledge, this is the first time this issue is addressed in the literature while many papers have studied the differences in growth between the two countries since they have experienced economic reforms (see, among others, Vu 2009b).

To answer the previous question, it is necessary to seek the functional form that best describes the relationship between calorie intake and income. Here is our second contribution. A natural choice would be to adopt a fully nonparametric specification of the relationship. The estimate of the relationship involves many control variables (age, education, region ...) and then we would be faced with the problem of the curse of dimensionality. We then choose to estimate various semi-parametric additive specifications in which the control variables included in the parametric part of the model. Income is then supposed to impact calories intake through a smooth function of unknown form. A similar choice has also been done by Gibson and Rozelle (2002), Tian and Yu (2015), and Nie and Sousa-Poza (2016). We consider more general specifications belonging to the family of generalized additive models, or GAM (Wood 2006). The conditional distribution of calorie intake given income and various control variables is thus chosen in a list of conventional statistical distributions. The conditional expectation of calorie intake given income and control variables is expressed as a linear function of the control variables and a smooth function of income, up to a monotone transformation or link function. The papers cited just above actually use GAM specifications where the conditional distribution is the classical normal distribution and the link function the identity function.

Several potential options are possible for the GAM specifications to describe the relationship between calorie intake and income, following the approach suggested above, and we must choose among them. Here is the third contribution of our paper. We use a cross-validation procedure proposed recently by Racine and Parmeter (2014), namely "revealed performance test." This procedure is a data-driven method for testing whether or not two competing approximate models are equivalent in terms of their expected true errors, i.e., their expected performance on unseen data from the same data generating process. The proposed test is quite flexible with regard to the types of models that can be compared (nested versus non-nested, parametric versus nonparametric) and is applicable in cross-sectional and time-series settings. The proposed test can be also applied to model selection (Kiefer and Racine 2017).

The comparison between China and Vietnam, based on the approach described above, is illustrated on the China Health and Nutrition Survey, or CHNS, and the Vietnam Household Living Standards Survey,

or VHLSS, in recent years. The methodology outlined above is carried out for different years from the early 2000s until today.

The analysis of the evolution of the calorie intake - income relationship over the studied period is not easy because this relationship is estimated from samples whose structure has evolved over time to remain representative of the population of Chinese or Vietnamese households. Nevertheless, estimates of the relationship between calorie intake and income for each survey wave can be used to decompose the difference between average calorie intakes between two waves in two effects: the effect of change in the surveyed populations and that due to changes in eating habits as reflected by the differences between the estimates of the calorie intake - income relationship. This is the objective of decomposition methods in economics initiated by Oaxaca (1973) and Blinder (1973). We modify the approach proposed by Machado and Mata (2005) by applying it to the case of a difference between mean values and by incorporating in it the semi-parametric estimates of the relationship under investigation.

The paper is organized as follows. Section 2 presents the methodology used in this paper. Section 3 is devoted to the presentation of the CHNS and VHLSS data. Results are presented and discussed in Section 4. Section 5 concludes.

2 Methodology

2.1 Generalized Additive Models

Following Abdulai and Aubert (2004), most empirical works about estimating the relationship between calorie intake and income, use the classical double-log specification, i.e.

$$\log(\text{PCCI}) = \alpha_0 + \alpha_1 \log(\text{INCOME}) + \alpha_2 \left(\log(\text{INCOME})\right)^2 + \sum_j \beta_j x_j + \varepsilon \tag{1}$$

where PCCI denotes per capita calorie intake, INCOME is total household income (sometimes replaced by total food expenditure), and x_j are other covariates (usually discrete covariates describing the structure of the household) . The squared term, $(\log(\text{INCOME}))^2$, is introduced to capture the nonlinearity of the income elasticity of calorie intake as a function of income. The unknown coefficients, or α_0, α_1, α_2, and the β_j, can be easily estimated by using the classical estimation techniques for linear models.

Although apparently flexible, the double-log specification constrains the form of the response of calorie intake to a change in income. Of course, it is easy to give a direct interpretation to the estimated values of coefficients associated with log(INCOME) and its squared value in terms of income-elasticity, which explains the frequent choice of this specification in empirical studies. However, taking the conditional expectation of the logarithm of the calorie intake as the object to be estimated rather than directly the conditional expectation of calorie intake can lead to misleading conclusions about the relationship studied as shown by Silva and Tenreyro (2006). More general, or less restrictive, specifications belonging to the family of generalized additive models, or GAM (Wood 2006), can be chosen to provide clearer statistical foundations to the estimation of the relationship between calorie intake and income and to capture nonlinearities in this relationship.

GAMs can be viewed as extensions of Generalized Linear Models, or GLM, which have the following structure

$$g(\mu) = \eta = X\beta \tag{2}$$

where $g(.)$ is a smooth monotonic link function, $\mu \equiv \mathbb{E}(y|X)$, y is a vector of independent response variables $(Y_1, \ldots, Y_n)'$, n is the sample size, η is called the linear predictor, X is an $n \times k$ matrix of k covariates, and β represents the $k \times 1$ vector of unknown regression coefficients. The response variable Y follows an exponential family distribution whose probability density functions are of the form

$$\exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{3}$$

where $b(.)$, $a(.)$ and $c(.)$ are arbitrary functions, θ is known as the "canonical parameter" of the distribution, and ϕ is the dispersion parameter. For practical modelling, $a(\phi)$ is usually set to ϕ. The expected value and variance of such a distribution are $\mathbb{E}(Y) = \partial b(\theta)/\partial\theta = \mu$ and $Var(Y) = \phi\partial\mu/\partial\theta = \phi V(\mu)$, where

$V(.)$ denotes the variance function. Among the most popular distributions of the exponential family are the Gaussian, Poisson and Gamma distributions.

GAMs extend GLMs by allowing the determination of possible non-linear effects of predictors on the response variable. The linear predictor of a GAM is typically given by

$$\eta = X\beta + \sum_j s_j(z_j) \tag{4}$$

where β represents the vector of unknown regression coefficients for the covariates acting linearly in the matrix X (usually discrete covariates), and the s_j are unknown smooth functions of the covariates z_j. The smooth functions can be function of a single covariates as well as of interactions between several covariates.

In recent papers, Tian and Yu (2015) and Nie and Sousa-Poza (2016) generalize the usual double-log model by introducing an unknown smooth function to capture the impact of income on per capita calorie intake. They estimate models whose expressions can be summarized as

$$\text{PCCI} = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j + \varepsilon \tag{5}$$

This equation can be viewed as a special case of the general GAM specification presented above. We also estimate more general semiparametric specifications whose expression is

$$g\left(\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j)\right) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j. \tag{6}$$

The logarithmic transformation is chosen as the link function, i.e., $g(.) = \log(.)$, ensuring that the conditional expectation is always positive. Different assumptions are then made as to the conditional distribution of calorie intake given income and various control variables.

2.2 Revealed Performance test

Various models will be estimated later to describe the relationship between calorie intake and income, as explained above. So we will be facing the problem of choice among these models. We approach the issue of selecting among these models from the perspective that fitted statistical models are approximations. The revealed performance test proposed by Racine and Parmeter (2014) uses random sample splits of the available data to construct evaluation and training data sets, estimating the competing models with the training data sets and then engaging out-of-sample prediction with the evaluation data. This process is repeated a large number of times and then the average out-of-sample squared prediction error, or $ASPE$, is used to compare models. The model with the smallest $ASPE$ is deemed the model with the lowest average prediction error.

Assuming that the data represent independent draws, as they would in a standard cross-sectional setup, the steps involved in the procedure proposed by Racine and Parmeter (2014) are:

1. Resample without replacement pairwise from $(y_i, x_i)_{i=1}^n$ and call these resamples $(y_i^*, x_i^*)_{i=1}^n$
2. Let the first n_1 of the resampled observations represent the training sample, i.e. $(y_i^*, x_i^*)_{i=1}^{n_1}$. The remaining $n_2 = n - n_1$ observations represent the evaluation sample, i.e. $(y_i^*, x_i^*)_{i=n_1+1}^n$.[1]
3. Fit each model, we denote here by $\widehat{m}_j(.)$, $j = 1, \ldots, k$, the estimates obtained using only the training observations $(y_i^*, x_i^*)_{i=1}^{n_1}$. Then compute predicted values from the evaluation observations $(y_i^*, x_i^*)_{i=n_1+1}^n$, i.e. $\widehat{y}_{i,j} = \widehat{m}_j(x_i^*)$, $i = n_1 + 1, \ldots, n$.

[1] Racine and Parmeter (2014) do not give any theoretical guidance in selecting n_2, or equivalently n_1, as a function of the sample size. They just advise the user to investigate the stability of their results with respect to the choice of n_2.

4. Compute average out-of-sample squared prediction error, or $ASPE$, for each model j as

$$ASPE_j = \frac{1}{n_2} \sum_{i=n_1+1}^{n} (y_i - \widehat{y}_{i,j})^2$$

5. Repeat steps $1-4$ a large number B of times, yielding B draws for each model j, or $(ASPE_{jb})_{b=1}^{B}$.[2]

These draws are used to discriminate between models. Paired t-test of difference in means for the two distributions can be used to choose between these models.

2.3 Decomposition methods

The procedure presented above allows us to select a specification for the relationship between calorie intake and income for each wave of the surveys we use (see below). It is then interesting to see in the evolution of the distribution of calorie intake between two waves what comes from the change in the distribution of explanatory variables and what results from the change in the chosen models. For this we will focus on the decomposition of average calorie intake between two waves and break it down into two effects: one specific to the change in the distribution of the explanatory variables and the other related to the model change. Or, put differently, we focus on

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(PCCI) - \mathbb{E}_{t_0}(PCCI) \tag{7}$$

where the two waves are denoted by t_0 and t_1, and $\mathbb{E}_t(PCCI)$ denotes the expectation of calorie intake using the distribution of the explanatory variables for wave t. Using the law of iterated expectations, the difference $\Delta PCCI_{t_0 \to t_1}$ can be written as

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(\mathbb{E}(PCCI|INCOME, Z)) - \mathbb{E}_{t_0}(\mathbb{E}(PCCI|INCOME, Z)) \tag{8}$$

Note that $\mathbb{E}(PCCI|INCOME, Z) = m_t(INCOME, Z)$ where $m_t(.)$ denotes the model chosen for wave t by the revealed performance test. Equation (8) becomes

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z)) \tag{9}$$

Finally we can write the difference as

$$\begin{aligned} \Delta PCCI_{t_0 \to t_1} = \quad & \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) + \\ & \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z)) \end{aligned} \tag{10}$$

where $\mathbb{E}_{t_1}(m_{t_0}(INCOME, Z))$ is the counterfactual expectation of calorie intake using the model chosen for wave t_0 and the distribution of explanatory variables of wave t_1.

Decomposition (10) can be viewed as a generalization of the well-known Oaxaca-Blinder decomposition (Oaxaca 1973, Blinder 1973) to semiparametric models. The first term in the right han side of equation (10), or $\mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z))$, measures what is usually called the "structure" effect. This effect can capture the chang of impact of household behavior in their choice of consumption due to changes in their environment. For instance, such changes may make these choices more or less income sensitive. The second term, or $\mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z))$, measures the "composition" effect and refers to the effect of the change in the distribution of the characteristics of households.

The different terms of the decomposition (10) can be estimated by taking empirical counterparts of the expectations, i.e. average values of the predicted values of $PCCI$ from the different models using either the contemporaneous or the counterfactual observations. Confidence intervals can then be calculated by adapting the bootstrap procedure proposed by Machado and Mata (2005).

[2] Here too, there is no theoretical guidance as to the number B in Racine and Parmeter (2014). They just advise to take a large number such as $B = 10,000$.

3 Data and Variables

3.1 Data sets

This study is based on two surveys conducted in both countries over several years. The first survey is China Health and Nutrition Survey, or CNHS. This survey is an on-going project between the Carolina Population Center at the University of Chapel Hill and the National Institute of Nutrition and Food Safety at the Chinese Center for disease Control and Prevention. Diet, health, demographic and socioeconomic information from about 4000 households is collected in nine geographically diverse Chinese provinces since 1989. In this study, we use the most recent four waves of the CNHS conducted in 2004, 2006, 2009, and 2011. This study relies also on Vietnam Household Living Standard Surveys, or VHLSS. VHLSS is conducted by the General Statistics Office of Vietnam, or GSO, with technical assistance of the World Bank, every two years since 2002. Each VHLSS survey contains modules related to household demographics, education, health, employment, income generating activities, including household businesses, and expenditures. The survey is conducted in all the 64 Vietnamese provinces and data are collected from about 9000 households for each wave. The survey is nationally representative and covers rural and urban area. In this study, we use the most recent six waves of the VHLSS conducted in 2004, 2006, 2008, 2010, 2012, and 2014.

The surveys of the two countries we used in this study are very different in terms of how household consumption data are collected. The Nutrition Survey, an integral part of the CHNS, documents a variety of food items that each and all household members consume (in grams), both at home and away from home, over a three-day window. The result is a highly detailed account of hundreds of types of food consumed during the day, whose precision is suitable for nutrition studies as emphasized by Batis et al. (2014). These very precise data are converted to energy intake and amounts of macronutrients using food composition tables for China (see Popkin et al. 2002, for more details). Calorie intake is thus available at the individual level for each member of a surveyed household in CNHS data set.

The main objective of VHLSS is to collect data on Vietnamese household living standards, as measured by households income and expenditure, as well as household members occupation, health and education status. This survey is not, by definition, constructed to assess the nutritional status of Vietnamese households. Only data on food expenditures and consumption are collected in this survey. Information on food expenditures and consumption are obtained for both regular and holiday expenses. These data are collected for both purchased goods and self-supplied food (home production) for 56 food items. All food consumption is transformed into calories based on the calorie conversion table constructed by Vietnam National Institute of Nutrition in 2007. The Vietnamese food composition table used in this study differs from that used in Mishra and Ray (2006) who use the food composition table builded by FAO international (the first FAO food composition table was published in 1949) to obtain calorie consumption. The calorie conversion table used in this study should reflect better calorie consumption in Vietnam because it is based on Vietnamese diets while the FAO table is constructed based on the most common food items consumed around the world.

3.2 Calculating per capita calorie intake

Once estimated the number of calories consumed per household, it is common practice to convert household-level calorie intake into individual-level calorie intake using equivalence scales. Household total calorie intake, or $THCI$, can be expressed as

$$THCI = CI^h + \sum_{i \ neq h} CI^i_{g,a}$$

where CI^h is calorie intake of the head of the household, taken as the reference, and $CI^i_{g,a}$ is calorie intake of the non-head household member i of gender g and age a. Calorie intake of the adult reference member can then be computed as

$$CI^h = \frac{THCI}{1 + \sum_{i \neq h} \mathbb{1}_{i \in \{g,a\}} \theta_{g,a}}$$

where $\theta_{g,a} = CI^i_{g,a}/CI^h$ defines the equivalence scale for a non-head member of the household of gender g and age a.

It is no frequent to observe calorie intake for each member of a household, making it impossible to calculate directly the equivalence scales. Most papers in the literature do not use any equivalence scales, and calculate the adult equivalent of household calorie intake by dividing household total calorie intake by the total number of members in the household. Hence, $\theta_{g,a} = 1$, whatever the age or gender of the household members. However, this issue can be addressed using OECD equivalence scales, i.e. setting $\theta_{g,a} = 0.7$ for each adult other than the head of the household, whatever the gender, and $\theta_{g,a} = 0.5$ for the children, whatever their age or gender (OECD 2013). Here, to calculate our equivalence scales, we proceed as Aguiar and Hurst (2013) and we estimate the following regression model

$$\log(THCI) = \gamma_0 + \gamma_1\, Gender + \gamma_2\, N_a + \gamma_3\, Family + \varepsilon. \tag{11}$$

where $THCI$ is total household calorie intake, $Gender$ is the gender of the head of the household (male is taken as the reference), N_a is the number of adults in the household other than the head, and $Family$ counts the numbers of children by gender and age categories ($0-2$, $3-5$, $6-13$, and $14-17$). This regression is estimated separately by area of residence, i.e. rural or urban, and by year as in Santaeulàlia-Llopis and Zheng (2016). Then we use the exponentiated predicted value of $THCI$, normalized by the value for singleton households, i.e. $\exp(\widehat{\gamma_0})$ if the individual is a male, or $\exp(\widehat{\gamma_0} + \widehat{\gamma_1})$, otherwise, as the equivalence scale. An equivalence scale is thus defined for each household. Per capita calorie intake, or adult equivalent calorie intake, is then computed as the ratio of household total calorie intake and household equivalence scale.

[Table 1 about here.]

Table 1 reports the average value of adult equivalent calorie intake for each year in the survey and each country and compares it with other studies. The average values we obtained for the two countries are consistent with those obtained in the other papers using the same survey data. They are just a little higher, which we could foresee as the other studies do not use any equivalence scale.

3.3 Income and control variables

The following variables are used as explanatory variables when estimating the relationship between calorie intake and income: $INCOME$: total household income per year; $INCOME$ was converted to 2006 dollars to make comparisons between the two countries easier; $URBAN$: dummy variable = 1 if the household is located in an urban area, = 0 if not; $HSIZE$: household size (this variable is discretized in several classes: five for China and six for Vietnam, the last class for households with 5 (resp. 6) or more members in China (resp. Vietnam)); HAN for China or $KINH$ for Vietnam: ethnicity of the head of household, = 1 if the head of the household belongs to the major ethnic group of the country (Han for China or Kinh for Vietnam), = 0 if not; $EDUCH$: the highest education level of the head of the household (this ordered variable takes three levels: = 1 for primary school, = 2 for secondary school, and = 3 for university); $GENDER$: gender of the head of the household, = 1 if male, = 0 if not; WA: this variable indicates if the household is located in a house having access to clean water or not; $AREA$: the province where the household is located. In China, we use the nine provinces which are surveyed. We divide Vietnam in six biggest regions.

4 Main results

4.1 Chosen models

We estimate four different models. The first one is the classical double-log model, or DLM,

$$\log(PCCI) = \alpha_0 + \alpha_1 \log(INCOME) + \alpha_2 \Big(log(INCOME)\Big)^2 + \sum \beta\, Factors + \epsilon, \tag{12}$$

where $Factors$ include $URBAN$, $HSIZE$, $ETHNIC$, WA, $EDUCH$, $GENDER$, $AREA$. The second model is the semiparametric model where the distribution of PCCI belongs to the Gaussian family and

$$\mathbb{E}(PCCI|\text{INCOME}, Factors) = \alpha_0 + s(INCOME) + \sum \beta\, Factors. \tag{13}$$

This model is denoted by GAMGauSId. This model corresponds to models used by Gibson and Rozelle (2002), Tian and Yu (2015), and Nie and Sousa-Poza (2016). The third and fourth models are semiparametric models such that

$$log(\mathbb{E}(PCCI|\text{INCOME}, Factors)) = \alpha_0 + s(INCOME) + \sum \beta\, Factors, \tag{14}$$

where the distribution of PCCI belongs to the Gaussian family, model denoted by GAMGauSLog, or to the Gamma family, model denoted by GAMGamLog.

Table 2 reports the results of the t-paired tests used to compare the out-of-sample predictive performances of the four models for each year. This table should be read like that. Consider, for example, the value obtained of the test statistics shown for China in 2004 at the intersection of the line for DLM and the column for GAMGauId, namely 78.80. This figure indicates that on average the difference between the ASPE criteria obtained for the two models for each of the 10,000 splits of the Chinese data using the procedure presented in subsection 2.2 is positive. On average, the value of ASPE for the DLM is therefore higher than that obtained for GAMGauId, and this difference is significatively different from zero, indicating that DLM underperforms when compared to GAMGauId. A negative and significantly different from zero value of the test statistics would have indicated the opposite.

[Table 2 about here.]

The last column of Table 2 indicates which model is chosen after the revealed performance test for each country and each survey year. The results clearly indicate that DLM is never chosen when compared to semiparametric models for China. GAMGauId is chosen three times over the four years studied when compared to other semiparametric models. GAMGAULog is only chosen in 2006. The results obtained for Vietnam show the best predictive capability for DLM for the first three years. This model is rejected for the three following years where GAMGauId is chosen two times, in 2012 and 2014, and GAMGauLog only one time in 2010.

4.2 Comparison between China and Vietnam

Figure 1 reports per capita calorie intake as a function of income, the control variables being fixed to their model values in 2004,[3] for the different waves for China and Vietnam. The nonlinearity of the relationship clearly appears in the view of the different curves traced in Figure 1. This confirms the tests performed above. The relationship appears to be concave for both countries. For Vietnam, the relationship is strongly increasing for low income levels up to a point at which it continues to grow but at a much slower rate (or even zero rate). Also this general shape of the relationship is fairly constant over time. The results are more mixed for China, but the general shape of the curve still is concave.

[Figure 1 about here.]

The comparison of the evolution of the curves over time has little meaning since these curves are estimated from samples whose structure varies over time and for years that may have been subject to external shocks on households' consumption in a given country. The sample structure of the CHNS survey changed between 2009 and 2011 due to the addition of the three autonomous cities, Beijing, Shanghai, and Chongqing, in 2011. In addition, the year 2008 was a year of strong decrease in food consumption in Vietnam due to difficult climatic year and a very significant increase in food prices due to double-digit inflation. Nevertheless, it is interesting to disentangle, in the evolution of the distribution of calorie intake between two waves, what comes from the change in the distribution of explanatory variables and what results from the change in the relationship between calorie intake and explanatory variables. Figure 2

[3] The chosen household for China is living in a rural area in Heilongjiang region, with three members from HAN ethnicity, with a male head having secondary school degree, and having access to clean water. As for the household chosen for Vietnam, it comes from a rural area in Mekong province. Its head is a man with primary education level. It comprises four members from Kinh ethnicity and has access to clean water.

reports the results of the decomposition described in equation (10). More precisely, we report a boxplot of the bootstrapped distribution of the difference of average $PCCI$ between a given survey wave and 2004, for each country, and the boxplots of the corresponding bootstrapped distributions coming from its decomposition into a structure and a composition effects.

[Figure 2 about here.]

Figure 2 shows a contrasting trend over the period considered for China. The difference between the average calorie intake between 2006 and 2004 is not significantly different from zero and the elements of its decomposition also. The difference between the average calorie intake between 2008 and 2004 is positive and larger than that between 2006 and 2004, but it is still not significantly different from zero. This is now due to the compensation between the two effects, the structure effect being predominantly negative, while the composition effect is always positive. The latter effect increases when comparing 2011 to 2004. This positive composition effect is completely counterbalanced by the significant negative structure effect. This results in a significant decrease in the average calorie intake between the two waves. The introduction of the three autonomous cities in 2011 has certainly changed the composition of the CHNS sample by increasing the proportion of well-educated, small (less than two members), and urban households in it. Nevertheless, this also had the consequence of modifying significantly the estimated relation between calorie intake and income as seen in Figure 1.

Results are much clearer for Vietnam. They are more stable, with the noticeable exception of the 2008 wave, an atypical year already mentioned above. The difference in average calorie intake between 2006 and 2004 is not significantly different from zero and this is due to the two effects that compensate over the period. Wave comparisons of 2010, 2012 and 2014 to 2004 show a stagnation of the difference between average calorie intakes which is positive and significantly different from zero. Now the two effects are also positive and significantly different from zero, the structure effect being always larger than the composition effect. It should be noted that the samples for the 2010, 2012 and 2014 waves are composed of more urban and small (less than two members) households than the 2004 wave. The difference between the average calorie intakes is certainly due to an effect coming from the difference in the composition of the samples but it is also the result of a significant change in the relationship between calorie intake and income, as reflected in the structural effect.

5 Conclusion

This paper revisits the issue of estimating the relationship between calorie intake and income, and presents and compare estimates of this relationship for China and Vietnam. For this, we use various recent tools in semiparametric econometrics, in model choice, and in decomposition in economics. The application uses different waves of CHNS surveys for China and VHLSS for Vietnam over a similar period, from 2004 to the present day.

The different models chosen at the end of the model selection procedure include both the classical double-log model and more general semiparametric specifications. Most of them highlight a relationship between caloric intake and income that is strongly increasing for low income levels and that becomes increasing with a much lower slope or even constant from a certain income threshold. The analysis of the evolution of these curves is not easy because they are estimated from samples whose structure has evolved over time to remain representative of the population of Chinese or Vietnamese households. Nevertheless, estimates of the relationship between calorie intake and income for each survey wave can be used to decompose the difference between average calorie intakes between two waves in two effects: the effect of change in the surveyed populations and that due to changes in eating habits as reflected by the differences between the estimates of the calorie intake - income relationship. The analysis reveals a significant negative difference between the average calorie intakes in China between 2011 and 2004, a finding already made by You et al. (2016). This difference is mainly due to the negative and significant effect of the change in consumption patterns which is barely compensated by the positive and also significant effect due to the change in the structure of the studied population between the two years. Unlike China, the two effects play in the same direction over the period 2004 - 2014 for Vietnam. They are positive and significantly different from zero. Their addition explains the increasing evolution of the average calorie intake observed over this period.

References

Abdulai, A. and D. Aubert (2004). Nonparametric and parametric analysis of calorie consumption in Tanzania. *Food Policy 29*(2), 113–129.

Aguiar, M. and E. Hurst (2013). Deconstructing life cycle expenditure. *Journal of Political Economy 121*(3), 437–492.

Banerjee, A. (2016). Policies for a better-fed world. *Review of World Economics 152*(1), 3–17.

Batis, C., D. Sotres-Alvarez, P. Gordon-Larsen, M. A. Mendez, L. Adair, and B. Popkin (2014). Longitudinal analysis of dietary patterns in Chinese adults from 1991 to 2009. *British Journal of Nutrition 111*(08), 1441–1451.

Blinder, A. (1973). Journal of Human resources. *British Journal of Nutrition 111*(08), 436–455.

Gibson, J. and S. Rozelle (2002). How elastic is calorie demand? Parametric, nonparametric, and semiparametric results for urban Papua New Guinea. *Journal of Development Studies 38*(6), 23–46.

Kiefer, N. and J. Racine (2017). The smooth colonel and the reverend find common ground. *Econometric Reviews 36*(1-3), 241–256.

Machado, J. and J. Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics 20*(4), 445–465.

Mishra, V. and R. Ray (2006). Dietary pattern, calorie intake and undernourishment: the vietnamese experience. Technical report, Discussion Paper 2006-02, School of Economics and Finance, University of Tasmania.

Nie, P. and A. Sousa-Poza (2016). A fresh look at calorie-income elasticities in China. *China Agricultural Economic Review 8*(1), 55–80.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693–709.

OECD (2013). OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth. *OECD Publishing, Paris*.

Ogundari, K. and A. Abdulai (2013). Examining the heterogeneity in calorie–income elasticities: A meta-analysis. *Food Policy 40*, 119–128.

Popkin, B. M., B. Lu, and F. Zhai (2002). Understanding the nutrition transition: measuring rapid dietary changes in transitional countries. *Public health nutrition 5*(6a), 947–953.

Racine, J. and C. Parmeter (2014). Data-Driven Model Evaluation: A Test for Revealed Performance. *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, 308–345.

Santaeulàlia-Llopis, R. and Y. Zheng (2016). Missing consumption inequality: direct evidence from individual food data.

Silva, J. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

Tian, X. and X. Yu (2015). Using semiparametric models to study nutrition improvement and dietary change with different indices: The case of China. *Food Policy 53*, 67–81.

Vu, L. (2009a). Analysis of calorie and micronutrient consumption in Vietnam. *Development and Policies Research Center Working Paper Series* (2009/14).

Vu, M. (2009b). Economic Reform and Growth Performance: China and Vietnam in Comparison. *China: An International Journal 7*(2), 189–226.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

You, J., K. Imai, and R. Gaiha (2016). Declining nutrient intake in a growing China: Does household heterogeneity matter? *World Development 77*, 171–191.

6 References

Appendix: Figures and Tables

Figure 1: Estimated calorie intake and income relationships for China and Vietnam

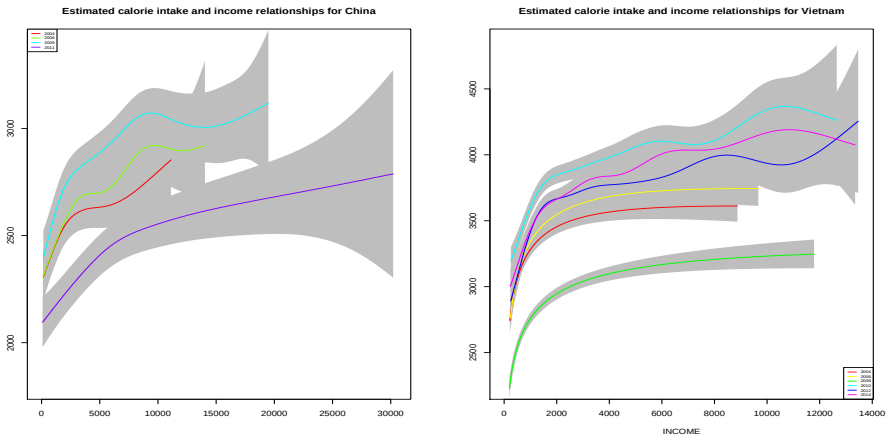

Figure 2: Decomposition of the difference of average PCCI for China and Vietnam

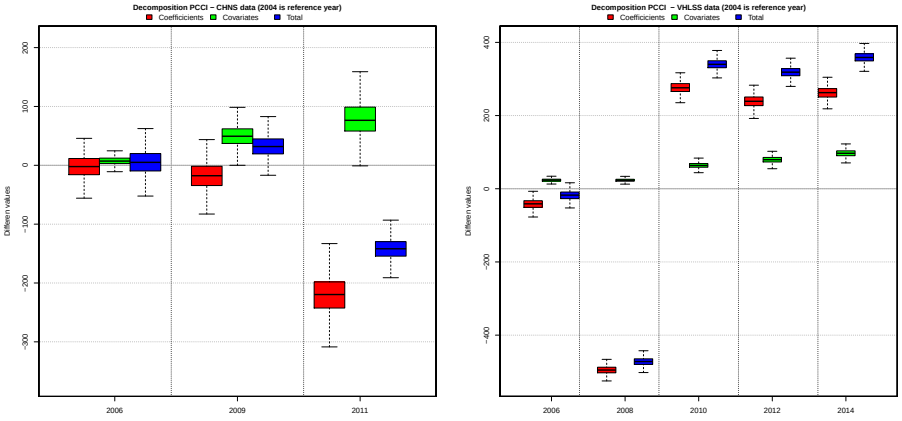

Table 1: Average Per Capita Calorie Intake: Comparison with other papers

	China					
	2004	2006	2009	2011		
Tian and Yu (2015)	2173	2103	2086			
Nie and Sousa-Poza (2016)	2348	2259	2253			
Our study	2398	2402	2428	2254		
	Vietnam					
	2004	2006	2008	2010	2012	2014
Mishra and Ray (2006): Rural	3206					
Mishra and Ray (2006): Urban	2824					
Vu (2009a)	2348					
Nguyen and Winter (2011)	3144	3074				
Our study	3291	3272	2818	3632	3611	3651
Note: unit = KCal						

Table 2: t-paired test results

China					
Year	Model	GAMGauId	GAMGauLog	GAMGamLog	Choice
2004	DLM	78.80***	39.08***	40.47***	GAMGauId
	GAMGauId		-109.91***	-107.67***	
	GAMGauLog			-14.37***	
2006	DLM	15.91***	39.92***	32.85***	GAMGauLog
	GAMGauId		57.03***	18.61***	
	GAMGauLog			-26.04***	
2009	DLM	76.26***	16.40***	17.01***	GAMGauId
	GAMGauId		-143.71***	-122.56***	
	GAMGauLog			-7.09***	
2011	DLM	33.06***	8.30***	6.98***	GAMGauId
	GAMGauId		-61.15***	-122.56***	
	GAMGauLog			-3.51***	
Vietnam					
Year	Model	GAMGauId	GAMGauLog	GAMGamLog	Choice
2004	DLM	-95.80***	-100.79***	-95.35***	DLM
	GAMGauId		-16.41***	-9.32***	
	GAMGauLog			4.26***	
2006	DLM	-62.99***	-69.92***	-75.03***	DLM
	GAMGauId		-34.73***	-34.36***	
	GAMGauLog			-4.14***	
2008	DLM	-7.73***	-11.28***	-17.66***	DLM
	GAMGauId		-11.70***	-29.21***	
	GAMGauLog			-18.23***	
2010	DLM	-1.6	8.50***	7.64***	GAMGauLog
	GAMGauId		50.56***	44.55***	
	GAMGauLog			-4.056***	
2012	DLM	6.71***	-1.66	-0.86	GAMGauId
	GAMGauId		-35.74***	-31.21***	
	GAMGauLog			4.23***	
2014	DLM	3.10***	-30.33***	-36.47***	GAMGauId
	GAMGauId		-90.41***	-89.48***	
	GAMGauLog			-14.36***	
Note: *, **, and *** significant at 10%, 5%, and 1%, respectively					