Assessing the nonlinearity of the calorie-income relationship: an estimation strategy - With new insights on nutritional transition in Vietnam

Huong Trinh Thi[1,2], Michel Simioni[3,4] *and Christine Thomas–Agnan[5]

[1]Toulouse School of Economics, INRA, University of Toulouse Capitole, France

[2]Department of Mathematics and Statistics, Thuongmai University, Hanoi, Vietnam

[3]MOISA, INRA, University of Montpellier, Montpellier, France

[4]IREEDS-VCREME, Hanoi, Vietnam

[5]Toulouse School of Economics, University of Toulouse Capitole, France

May 2018

Abstract: Assessing the nonlinearity of the calorie-income relationship is a crucial issue when evaluating policies aimed at fighting against malnutrition. A natural choice would be to adopt a fully nonparametric specification of the relationship in order to let the data reveal its nonlinearity. But, we would be faced with the problem of the curse of dimensionality due to the presence of many control variables in addition to income. Here, we first propose to estimate generalized additive models where only income is supposed to enter nonlinearly in the specification. Second, we use a recent cross-validation procedure in order to choose among various competing specifications including the parametric double-log specification widely used in the literature in addition to GAM specifications. This methodology is implemented for each of the six waves of the Vietnam Household Living Standard Survey from 2004 to 2014. The calorie-income relationship is nonlinear whatever the wave. A strong response of calorie intake to an increase in income for poorest households is highlighted, showing that there is still room for income-based policies to fight against malnutrition. A byproduct of this methodology is the decomposition of the evolution of average calorie intake between the two waves in the part due to population change and that coming from the change in calorie-income relationship, shedding new light on the nutritional transition in Vietnam.

Keywords: Calorie-income relationship, Generalized additive model, decomposition methods, nutritional transition, Vietnam.

*Corresponding author: michel.simioni@inra.fr

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. Numerous papers in development and health economics deal with the issue of estimating the relationship between food demand measured in calories and household income, and lead to controversial results. Recently, Ogundari and Abdulai (2013), Santeramo and Shabnamb (2015), and Zhou and Yu (2015) provide surveys of this literature, and summarize the main issues that have been encountered. Thus, following Ravallion (1990), the literature generally agrees that the calorie-income relationship is nonlinear. Its general shape is popularly assumed to change with income dynamics. Calorie intake increases rapidly as income increases for consumers with low income. These consumers spend most of their additional income on food, and calorie intake therefore grows rapidly with income. Calorie intake increases then with income growth up to a threshold, called subsistence level. Beyond this threshold, calorie intake increases only slowly or even decreases, the marginal utility of additional calories going down significantly and finally staying relatively low. Many empirical studies tackle this issue by estimating the classical double-log specification where the log-income parameter possesses a direct interpretation as calorie-income elasticity and nonlinearity is captured by adding the square of log-income. 86 of the 99 elasticities recorded by Ogundari and Abdulai (2013) were thus obtained by estimating this parametric specification. Following Gibson and Rozelle (2002), only few papers use semiparametric specifications to deal with the nonlinearity of the calorie-income relationship (Tian and Yu, 2015; Nie and Sousa-Poza, 2016).

This paper aims at contributing to the literature on estimating the calorie-income relationship. It proposes to mobilize recent developments in semiparametric estimation (Wood, 2017) and model selection (Racine and Parmeter, 2014) to revisit the nonlinearity problem mentioned above. The objective is to find a functional form that best describes the relationship between calorie intake and income from cross-sectional data. A natural choice would be to adopt a fully nonparametric specification of the relationship. Since the estimate of the relationship involves many control variables (age, education, region ...) in addition to income, we would be faced with the problem of the curse of dimensionality (Stone, 1980). The accuracy of our nonparametric estimates would be low even if we were lucky enough to have large samples. Semiparametric specifications then make it possible to seek a balance between the problem of the curse of dimensionality and the choice of totally nonparametric specifications to measure the impact of certain variables such as income in our case. We choose to estimate various semiparametric additive

specifications in which the control variables are included in the parametric part of the model, and income is supposed to impact calorie 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). Here, we consider general semiparametric specifications belonging to the family of generalized additive models, or GAM (Wood, 2017). The conditional distribution of calorie intake given income and various control variables is thus chosen in a list of conventional statistical distributions, and the conditional expectation of calorie intake given income and various control variables is expressed as the sum of linear functions of the control variables and a smooth function of income, up to a monotone transformation or link function. For instance, 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 to describe the relationship between calorie intake and income: not only semiparametric GAM specifications as suggested above, but also the classical parametric double-log specification, and we must choose among them. We use a cross-validation procedure recently proposed by Racine and Parmeter (2014), namely "revealed performance test" or RPT, to choose among these various competing parametric and semiparametric specifications. This procedure is a data-driven method for testing whether or not two competing specifications are equivalent in terms of their expected true errors, i.e., their expected performances on unseen data coming from the same data generating process. The RPT procedure 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. This procedure can thus be applied to model selection as shown in Kiefer and Racine (2017).

Empirical analysis focuses on Vietnam. Indeed, although Vietnam has experienced a strong economic development that turned this poor country in the 1980s into a lower middle income country currently, Vietnam faces the double burden of malnutrition. This double burden of malnutrition is characterized by the coexistence of undernutrition along with overweight and obesity, or diet-related noncommunicable diseases, within individuals, households and populations, and across the life course (Nguyen and Hoang, 2018). Policies to fight against malnutrition are already relevant in Vietnam. The Vietnamese government has recently defined a comprehensive strategy to improve the nutritional situation of the Vietnamese population (Ministry of Health, 2012). The characterization of the shape of the calorie-income relationship is therefore relevant in order to assess the appropriateness of public policies affecting incomes of poor Vietnamese households.

The empirical analysis is based on six waves of the Vietnam Household Living Standard Survey, or VHLSS: 2004, 2006, 2008, 2010, 2012, and 2014. Expenditure data of each survey are transformed into nutritional data using

energy conversion factors of food kilograms into kilocalories that are specific to Vietnam (National Institute of Nutrition, 2007). These data are used to characterize the shape of the calorie-income relationship for each wave of VHLSS, using the methodology presented above. The shapes of the chosen estimated calorie-income relationships are consistent with what was expected. Calorie intake increases as income increases.This growth is strong up to an income threshold from which it noticeably reduces. This result shows that there is still room for income-based policies to fight against malnutrition in Vietnam.

A by-product of the previous work is the analysis of the evolution of the calorie-income relationship over the studied period. The aim is to provide new insights into the nutrition transition in Vietnam. It then needs to be stressed that this analysis is not easy because the calorie-income relationship is estimated from different cross-sectional samples whose structure has evolved over time to remain representative of the population of Vietnamese households. Nevertheless, estimates of the relationship between calorie intake and income for each VHLSS 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 between the two waves, and that due to changes in eating habits as reflected by the differences between the estimates of the calorie-income relationship for these two waves. This is the usual objective of decomposition methods in economics initiated by Oaxaca (1973) and Blinder (1973) and surveyed by Fortin et al. (2011). We modify the approach proposed by Machado and Mata (2005) and Nguyen et al. (2007) by applying it to the case of a difference between mean values and by incorporating the previously chosen parametric or semiparametric estimates of the relationship under investigation.

The results of the decomposition show that both effects contributed positively to the increase in average calorie intake over the studied period. Nevertheless, the effect of changes in eating habits, as reflected by changes in the estimated relationship between calorie intake and income, is a little higher than the effect due to changes in the structure of the population (mainly increasing urbanization and decreasing household size), the first effect remaining fairly stable while the latter is slowly increasing over the period.

The paper is organized as follows. Section 2 gives a picture of the nutritional situation of the Vietnamese population. Section 3 presents the methodology used in this paper. Section 4 is devoted to the presentation of the VHLSS data and to the approach chosen when converting expenditure data into quantities of calories. Results are presented and discussed in Section 5. Special attention is devoted to the potential endogeneity of the measure of income we have chosen, i.e. total expenditure. Section 6 concludes.

2 Nutritional issues in Vietnam

Vietnam's development record over the past 30 years is remarkable. Economic and political reforms under Doi Moi, launched in 1986, have spurred rapid economic growth and development and transformed Vietnam from one of the world's poorest nations to a lower middle-income country. According to World Bank, per capita Gross National Income rose from 435 to 1691 constant 2010 US dollars between 1989 and 2016. Moreover, the poverty rate decreased gradually from 58% in 1993 to 28.9% in 2002, 14.5% in 2008 and 12% in 2011.

At the same time, Vietnam has also experienced a nutrition transition like many other middle-income countries in South-East Asia (Popkin, 2006). Dietary diversity from 2005 to 2015 in this region and China has considerably increased: the share of cereal demand (in terms of quantity) has decreased by 12% while the share of meat and fish demand and those of dairy and eggs have increased by 8% and 30% respectively, the share of fruits and vegetables staying steady (IFPRI, 2017). Moreover, in terms of macronutrients, from 2004 to 2014, the share obtained from fat in total calorie intake has increased by 37.5% (resp. 23%) for Vietnamese rural households (resp. urban households), at the expense of calories obtained from carbohydrates, calories obtained from proteins staying quite stable (Trinh et al., 2018).

This nutrition transition to energy-dense, poor quality diets has led to obesity and non-communicable diseases. Among Vietnamese 18-65 years old, the prevalence of overweight and obesity increased from 2.3% in 1993 to 15% in 2015 (Nguyen and Hoang, 2018). Figures in big cities are higher. For instance, ten years ago, Cuong et al. (2007) were reporting that 26.2% (resp. 6.4%) of adults living in Ho Chi Minh City urban areas were already considered as overweight (resp. obese). Nevertheless, despite these changes, a sizeable share of the population, 11%, still experiences undernutrition in Vietnam. This double burden of undernutrition and overnutrition concerns more and more early childhood. In children under 5, the prevalence of overweight and obesity increased from 0.6% to 5.6% (overall), 0.9% to 6.5% in urban area, and 0.5% to 4.2% in rural ones, in the 2000-2010 period. As for adults, figures in big Vietnamese cities are larger then the averages for the whole country. Overweight and obesity among preschool children in Ho Chi Minh City urban areas already reached 20.5% and 16.3%, respectively, in 2005 (Dieu et al., 2007). But, approximately 14% of children in Vietnam under 5 were still stunted, 8.6% underweight and 4.4% thin in 2011 (Le Nguyen et al., 2013). According the the United Nations, despite a huge decrease in stunting and underweight rates, Vietnam remained among the thirty-six countries with the highest stunting rates in the world.

Improving the nutritional status of the Vietnamese population is now considered as a major concern by the Vietnamese government. The "National Nutrition Strategy for 2011-2020, with a vision toward 2030," defines the

main objectives and instruments of the nutrition policy in Vietnam (Ministry of Health, 2012). One of the objectives of this strategy, amongst others, is to simultaneously reduce the proportion of households with low caloric intake (below 1800 Kcal) to 5% and reach a proportion of households with a balanced diet (Protein: 14%; Lipid: 18%; Carbohydrate: 68%) equal to 75% by 2020. Emphasis is also placed on improving the nutritional status of mothers and children. It is then proposed to develop specific food and nutrition interventions to improve the nutritional status of target groups, and therefore, to give priority to the poor, disadvantaged and ethnic minority areas, as well as those at risk. Food and nutrition policy instruments, such as subsidized prices of basic foodstuffs or cash transfers, are not clearly envisaged in the strategy defined by the Vietnamese government. Nevertheless, it is interesting to see if there is still room for such instruments to improve the nutritional situation of Vietnamese households. This assessment requires knowledge of the responsiveness of calorie intake as income increases for different levels of income, and so requires the characterization of the calorie-income relationship form as emphasized by Zhou and Yu (2015).

3 Methodology

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

$$\log(\text{PCCI}) = \alpha_0 + \alpha_1 \log(\text{INCOME}) + \alpha_2 \left(\log(\text{INCOME})\right)^2 + \sum_{j=1}^{J} \beta_j x_j + \varepsilon \quad (1)$$

where PCCI denotes per capita calorie intake, INCOME is total household income (sometimes replaced by total expenditure), and the x_js are J 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, 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, 2017), 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. Appendix A gives more details on GAM.

In our application, the GAM specifications we estimate are of the form

$$g(\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J)) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \quad (2)$$

where (i) $g(.)$ is a link function, (ii) the variables entering with a linear effect, the x_js, are dummies or ordered variables such as gender of head of household or household size, and (iii) the variable entering with a non linear effect captured by an unknown smooth function $s()$, is the continuous variable $INCOME$.

To sum up, in addition to the classical double-log model described in Eq. 1, we estimate three competing specifications belonging to the GAM family. The first specification is a semiparametric one where the distribution of PCCI belongs to the Gaussian family and

$$\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \quad (3)$$

specifying the link function as the identity function. This specification has been used recently by Tian and Yu (2015) and Nie and Sousa-Poza (2016) in line with the pioneering paper of Gibson and Rozelle (2002). We denote this specification by GAMGauId. The second, third and fourth specifications are also semiparametric ones with

$$\log(\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J)) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \quad (4)$$

with $\log(.)$ as the link function and where the distribution of PCCI belongs either to the Gaussian family, specification denoted by GAMGauLog, or to the Gamma family, specification denoted by GAMGamLog.

Estimation of GAM is usually performed using penalized regression with splines (Wood, 2017). In all GAM specifications, we use thin plate regression splines, which do not require knots selection and are computationally efficient (Wood, 2003). Moreover, the choice of this type of splines allows for testing the linearity of the response $s(.)$ as explained in the Appendix B.

We then face the problem of choice among these models. We approach the issue of selecting among these models from the perspective that fitted statistical models can be viewed as approximations and they must be evaluated on the basis of their predictive performance when new samples are available (Efron, 1982). Thus, we implement the data-driven test recently

proposed by Racine and Parmeter (2014) and called "revealed performance test". This test 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 computed and used to compare models. The model with the smallest $ASPE$ is deemed the model with the lowest average prediction error and is therefore chosen. Details on the implementation of the revealed performance test are given in the Appendix C.

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 joint 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 the 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

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(PCCI) - \mathbb{E}_{t_0}(PCCI) \tag{5}$$

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 joint distribution of the outcome variable $PCCI$ and 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

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(\mathbb{E}(PCCI|INCOME, Z)) - \mathbb{E}_{t_0}(\mathbb{E}(PCCI|INCOME, Z)) \tag{6}$$

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 (6) becomes

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z)) \tag{7}$$

Finally we can write the difference as

$$\begin{aligned} \Delta PCCI_{t_0 \to t_1} = \quad & \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) + \\ & \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z)) \end{aligned} \tag{8}$$

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 (8) 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 hand side of equation (8), or

$\mathbb{E}_{t_1}\left(m_{t_1}\left(INCOME, Z\right)\right) - \mathbb{E}_{t_1}\left(m_{t_0}\left(INCOME, Z\right)\right)$, measures what is usually called the "structure" effect. This effect can capture the change 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 $\mathbb{E}_{t_1}\left(m_{t_0}\left(INCOME, Z\right)\right) - \mathbb{E}_{t_0}\left(m_{t_0}\left(INCOME, Z\right)\right)$, 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 (8) 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).

4 Data

This study relies on the Vietnam Household Living Standard Survey, or VHLSS. This survey is conducted by the General Statistics Office of Vietnam, or GSO, with technical assistance of the World Bank, every two years since 2002.[1] 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 areas. In this study, we use the six most recent waves of the VHLSS conducted in 2004, 2006, 2008, 2010, 2012, and 2014.

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.[2] Only data on food expenditures and quantities are collected in this survey. Information on food expenditures and quantities 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. Food consumption is transformed into calories based on the calorie conversion table constructed by Vietnam National Institute of Nutrition in 2007 (see Table 3). Per capita calorie intakes are then computed as adult equivalent calorie intakes following recent papers of Aguiar and Hurst (2013)

[1] A detailed description of the design of the survey and the way data are collected is given in Appendix D.

[2] We refer the reader to Bouis (1994) for an insightful discussion of the comparative advantages of household expenditure surveys and 24-recall surveys of nutritionists. See also Zezza et al. (2017).

and Santaeulàlia-Llopis and Zheng (2017). Details on these computations are given in the Appendix E.

Following many papers in the literature on calorie-income relationship, we measure household resources by total expenditure rather than by income.[3] As emphasized by Deaton (1997), households generally underestimate their income making total expenditure a more reliable proxy for household income. Other papers argue that current incomes are more volatile than current expenditure, making them a more noisy measure of permanent income (Bhalotra and Attfield, 1998). Total expenditures are thus converted to 2006 dollars to make comparisons between VHLSS waves easier. Household per capita expenditure is computed as household total expenditure divided by the number of members in the household.

Control variables include: $URBAN$: 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: six, the last class being for households with 6 or more members); $KINH$: ethnicity of the head of household, = 1 if the head of the household belongs to the major ethnic group of the country (Kinh for Vietnam), = 0 otherwise; $EDUCH$: 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 region where the household is located (Vietnam is divided into six ecological regions). Table 5 summarizes the main characteristics of all the variables.

Insert Table 5.

5 Results

5.1 Preferred models

Table 6 reports the results of the t-paired tests used to compare the average out-of-sample squared prediction error (ASPE) performances of the four models for each year. This table should be read as follows. Consider, for example, the value of the test statistic shown at the intersection of the line for DLM and the column for GAMGauId for 2004, namely -11.64. This figure indicates that the average difference between the ASPE criteria obtained for the two models, computed using the $10,000$ splits of the VHLSS data following the procedure described in Appendix C, is negative. On average, the value of ASPE for the DLM is therefore smaller than that obtained for GAMGauId. Moreover, this difference is significatively different from zero,

[3] In Ogundari and Abdulai (2013), 64 over the 99 calorie-income elasticities reported in the literature were computed with expenditure as proxy for income.

indicating that DLM outperforms clearly GAMGauId. A positive and significantly different from zero value of the test statistics would have indicated the opposite. The values given on the same line also indicate that the DLM model has better predictive performances than the other two models: -10.20 and -14.70 when comparing DLM to GAMGauLog and GAMGamLog, respectively. Thus, whatever the relative performances of the other three specifications when compared among themselves (GAMGauId, GAMGauLog and GAMGamLog), the chosen specification for 2004 is DLM.

Insert Table 6.

The same reading grid can then be applied to the other results reported in Table 6 for each VHLSS waves. Its last column summarizes which model is preferred after applying the revealed performance test for each wave. The results clearly indicate that DLM is chosen when compared to semiparametric models for 2004 wave, and that GAMGauId is always chosen when compared to the other parametric or semiparametric models for the other waves.

5.2 The estimated calorie-income relationships

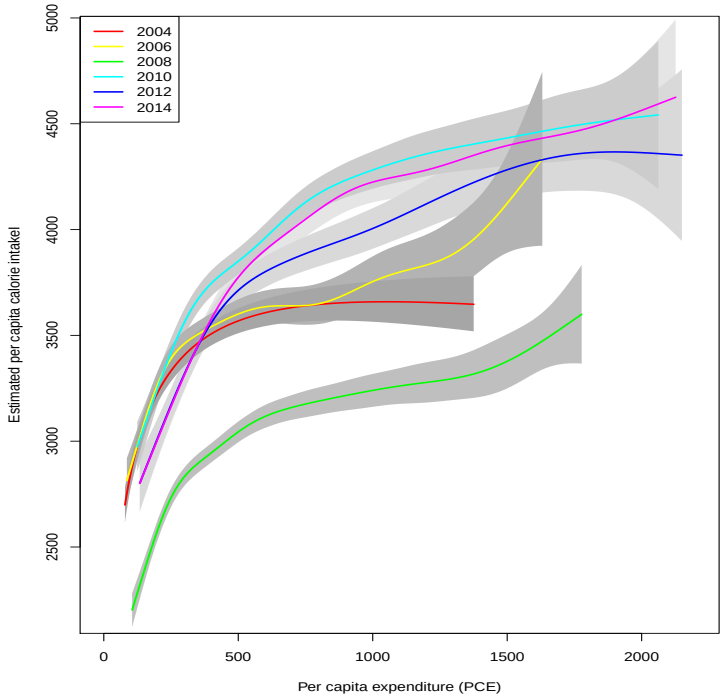
Figure 1 reports per capita calorie intake as a function of per capita expenditure and the control variables being fixed to their mode values in 2004, for the different VHLSS waves (shaded areas give the 95% confidence intervals around the estimated curves.).[4] The nonlinearity of the relationship clearly appears in view of the different curves traced in Figure 1. This result is confirmed by the various significance and linearity tests presented in Appendix B. The relationship appears to be concave for most waves. Generally, the relationship is strongly increasing for low per capita expenditure levels up to a point at which it continues to grow but at a much slower rate (or even zero rate).

These results contribute to the debate on the extent to which calorie consumption responds to income changes in middle-income countries. They clearly show that income mediated policies can have an impact on nutritional goals up to a given threshold of income, or per capita expenditure, in Vietnam. They show the rapid improvement of nutrition in terms of calorie intake for low per capita expenditure. They do not tell us anything about improving the nutritional quality of the diet. But they also show that from a certain level of per capita expenditure (between 250 and 750 dollars depending on year)[5] such income mediated policies may prove to be ineffective as

[4] The chosen household chosen comes from a rural area in the Mekong province. Its head is a man with primary education level. It comprises four members from Kinh ethnicity and has access to clean water.

[5] For comparison, the Gross Domestic Product per capita in Vietnam was recorded at US dollars 1162 US dollars in 2006.

Figure 1: Estimated calorie-income relationships for Vietnam



calorie intake seems little responsive to an increase of per capita expenditure.

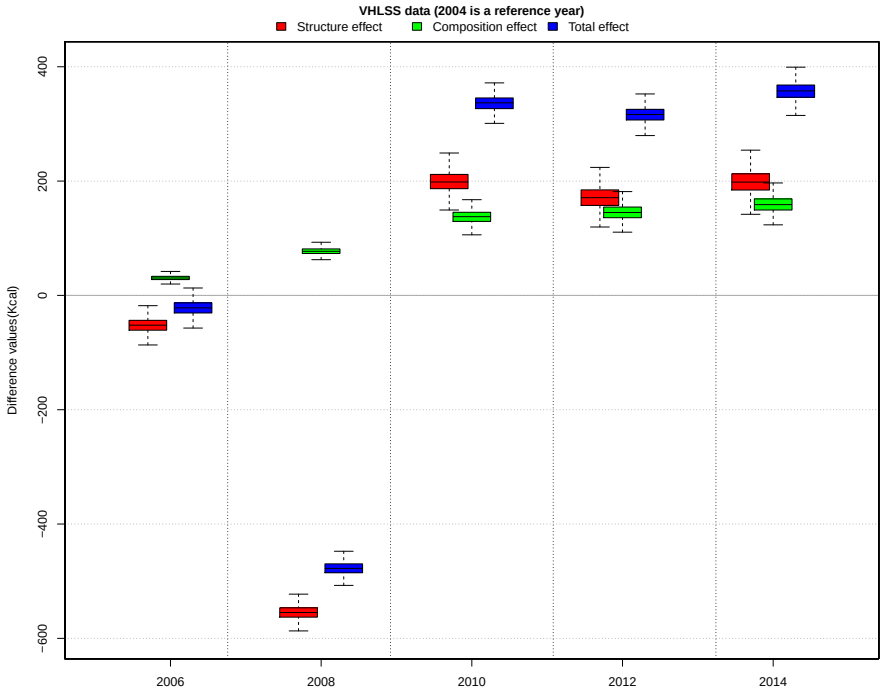
The comparison made above only makes sense because it concerns the evolution of the shape of the calorie-income relation over the period 2004-2014. Conversely, the comparison of the evolution of per capita calorie intake for a given value of the per capita expenditure is meaningless. The comparison only makes sense for the chosen values of the control variables that were set to their modes in 2004. Nevertheless, the significant drop in the estimated relationship in 2008 compared to other years deserves special comment. Due to the world economic crisis, the yearly growth rate of Vietnam GDP slowed down from 8.5% in 2007 to 6.3% in 2008, then 5.3% in 2009, before recovering to 6.5% in 2010. Moreover, inflation reached alarming rates in 2008: the yearly increase of consumer price index reached 28% in September 2008 and even 65% for staple food products (rice and grains). This deterioration in macroeconomic conditions has had an impact on many Vietnamese households who have reduced their food expenditure. For instance, the economic crisis has led to a significant increase in informal sector employment compared to the formal sector in the country's two largest cities, Hanoi and Ho Chi Minh City (World Bank, 2010). Almost one half (46%) of households involved in informal sector who were surveyed in HCMC as part of

2009 round of Household Business and Informal Sector survey declared that they have suffered from a decrease in income between 2008 and 2009. They reacted mainly by drawing on their savings (48.1% of these households) and cutting food expenditures (37% of these households) Situation was less acute for the same category of households in Hanoi.

5.3 The evolution of average calorie intake over 2004 to 2014

As shown above, the estimated calorie-income relationships can be used 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 calorie-income relationship. Thus, Figure 2 reports the results of the decomposition described in Eq. (8). More precisely, we report a boxplot of the distribution of the differences of average $PCCI$ between a given survey wave and 2004, based on 1000 bootstrap replications, and the boxplots of the corresponding distributions coming from its decomposition into a structure and a composition effects.

Figure 2: Decomposition of average per capita calorie intake difference



Decomposition results show a clear pattern in the evolution of average calorie intake between the successive waves of VHLSS and that of 2004, with the noticeable exception of the 2008 VHLSS wave, an atypical year already

mentioned above. The difference in average calorie intakes, i.e. total effect, between 2006 and 2004 is not significantly different from zero and this is due to the compensation between the structure and composition effects over the period. The total effect is always positive and significantly different from zero when comparing 2010, 2012 or 2014 to 2004. But the value of this effect remains stable for the three considered years. The structure and composition effects are also positive and significantly different from zero, the structure effect being always larger than the composition effect. It should be noted that samples for the $2010, 2012$ and 2014 waves are composed of more urban and small (less than three members) households and higher level of education of the head of a household 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.

5.4 Testing for exogeneity of income

An important concern in the estimation of calorie-income relationship is the potential endogeneity of income, or per capita expenditure as in our application to Vietnamese data. Following many empirical studies on calorie-income relationship estimation, we have so far assumed nutrition to be conditioned by income or food expenditure. But, if one follows the efficiency wage hypothesis (Stiglitz, 1976), it is conceivable that productivity of workers depends on their wages through the nutrition that their earnings enable them to purchase. This reverse causality can be a source of endogeneity of income or even of food expenditure when estimating the calorie-income relationship, thus leading to biased estimates.

The problem of endogeneity has recently received attention in nonparametric estimation. Nonparametric instrumental variables methods have been proposed by Darolles et al. (2011) and Horowitz (2011), among others. Testing the exogeneity assumption of an explanatory variable can be based on comparing a nonparametric estimate of the function of interest under exogeneity with an estimate obtained by using nonparametric instrumental variables methods. However, the moment condition that identifies the function of interest in the presence of endogeneity is a nonlinear integral equation of the first kind, which leads to an ill-posed inverse problem. Because of this problem, the rate of convergence of a nonparametric instrumental variables estimator is typically very slow. Therefore, a test based on a direct comparison of nonparametric estimates obtained with and without assuming exogeneity will have low power.

Blundell and Horowitz (2007) has developed a different approach to testing for endogeneity that avoids nonparametric instrumental variables estimation of the function of interest and then is likely to have better po-

wer properties. This test of exogeneity of explanatory variables directly exploits the conditional mean restriction that can be used to identify a nonparametric instrumental variables model. Its implementation requires only finite-dimensional matrix manipulations, kernel nonparametric regression, and kernel nonparametric density estimation as explained in Appendix F.

Below, we question the assumption of exogeneity of food expenditure that has been maintained throughout the study of the calorie-income relationship using different VHLSS waves. To address this concern, we follow Blundell and Horowitz (2007) and, to simplify computations, we use the univariate version of the test by focusing on the nonparametric estimation of the relationship between per capita calorie intake and per capita total expenditure. Following Subramanian and Deaton (1996), we use per capita nonfood expenditure as an instrumental variable for per capita total expenditure.

Table 1: Exogeneity test results (p-values)

Year	Base case	Bandwidth sensitivity		
		0.80	1.25	1.50
	(1)	(2)	(3)	(4)
2004	0.1070	0.0867	0.1419	0.1902
2006	0.3273	0.3067	0.3701	0.4118
2008	0.0053	0.0045	0.0061	0.0084
2010	0.1911	0.1742	0.2320	0.3019
2012	0.3897	0.3505	0.4244	0.4749
2014	0.3417	0.2589	0.4803	0.6615

Results of the test of exogeneity for the different VHLSS waves are reported in Table 1. Column (1) presents our baseline estimates while columns (2) to (4) show a sensitivity analysis with respect to the bandwidth choice required for the kernel nonparametric estimations involved in the test statistics computation. The bandwidths chosen in the baseline case are multiplied by 0.8, 1.25, and 1.5 in this sensitivity analysis. The p-values obtained for the 2006, 2010, 2012 and 2014 VHLSS waves are above 0.1 throughout, and thus there is no evidence of a violation of exogeneity of per capita total expenditure for these waves. A borderline p-value of 0.0867 is obtained for 2004 wave when baseline bandwidths are multiplied by 0.8. But, overall, the other p-values are larger than 0.1, and we interpret this evidence as suggesting exogeneity of per capita expenditure for the 2004 VHLSS wave too.

The results for the 2008 VHLSS wave are quite different from those for the other waves. p-values clearly indicate rejection of the null hypothesis of exogeneity of per capita total expenditure. For waves other than 2008, there is reason to doubt that calorie intake has had an impact on household spending. These are years characterized by sufficient economic growth to absorb new entrants into the labor market and strong productivity gains. However, 2008 is characterized by a sharp deterioration of macroeconomic conditions in Vietnam due to the global economic crisis. We can then conjecture that

this economic situation has led to a deterioration of the living conditions of many Vietnamese households: for example the decrease in food expenditure and thus in calorie intake they experienced may have had a feedback effect on their productivity and therefore their total expenditure (see the results of World Bank (2010) mentioned above).

6 Conclusion

This paper revisits the issue of estimating the relationship between calorie intake and income, and presents and compare estimates of this relationship for Vietnam. For this, we use various recent tools in semiparametric econometrics, in model choice, in decomposition methods in economics, and in testing exogeneity. The application uses six different waves of VHLSS for Vietnam from 2004 to 2014.

Different parametric and semiparametric models are estimated and compared for each VHLSS wave. 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 calorie 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 Vietnamese households. Moreover, the preferences of Vietnamese consumers have evolved over this ten years period. Estimates of the relationship between calorie intake and income for each survey wave can then 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 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 increase of average calorie intake observed in Vietnam over this period. Finally, we check whether the exogeneity assumption of income we have done throughout our analysis can be supported. The test we use does not reject the hypothesis of exogeneity except for the 2008 VHLSS wave, the year in which Vietnam experienced the maximum impact of the global economic crisis.

The methodology proposed in this paper stops at the decomposition of the evolution of average per capita calorie intake into a structure and a composition effects. This paper does not go further, i.e. does not propose a decomposition of the structure and composition effects, i.e. dividing differences between years into components which can be attributed to the characteristics of the households. To our knowledge, such decompositions have

never been proposed in the literature for semiparametric models. Moreover, as pointed out by Rothe (2015), such decompositions seem impossible for very general nonlinear models with interactions between the covariates.

Acknowledgements

We would like to thank the Editor of the review and two referees for their thoughtful comments and suggestions on the earlier draft of this paper. This paper has also benefited from valuable discussions with participants at Journées de Statistiques, Montpellier, June 2016, Vietnamese Economists Annual Meeting, DaNang, August 2016, Journées de la Recherche en Sciences Sociales, Paris, December 2016, Academy for Policy and Development seminar, Hanoi, April 2017, University of Economics and Law seminar, Ho Chi Minh City, August 2017, and 15th EAAE Congress, Parma, August 2017. We are grateful to Dao The Anh for providing us VHLSS data, and to Thibault Laurent for technical assistance in R. Financial Support from INRA-CIRAD GloFoodS meta-program (TAASE project) is fully acknowledged.

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

Bhalotra, S. and C. Attfield (1998). Intrahousehold resource allocation in rural Pakistan: A semiparametric analysis. *Journal of Applied Econometrics 13*(4), 463–480.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human resources 111*(08), 436–455.

Blundell, R. and J. L. Horowitz (2007). A non-parametric test of exogeneity. *Review of Economic Studies 74*, 1035–1058.

Blundell, R., J. L. Horowitz, and M. Paray (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics 3*, 29–51.

Bouis, H. E. (1994). The effect of income on demand for food in poor countries: Are our food consumption databases giving us reliable estimates? *Journal of Development Economics 44*, 199–226.

Bouis, H. E. and L. J. Haddad (1992). Are estimates of calorie-income elasticities too high? A recalibration of the plausible range. *Journal of Development Economics 39*, 333–364.

Cuong, T., M. Dibley, S. Bowe, T. Hanh, and T. Loan (2007). Obesity in adults: an emerging problem in urban areas of Ho Chi Minh City, Vietnam. *European journal of clinical nutrition 61*(5), 673–681.

Darolles, S., Y. Fan, J. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

Deaton, A. (1997). *The analysis of household surveys:amicro-econometric approach to development policy*. The John Hopkins University Press,Baltimore and London.

Dieu, H. T. T., M. J. Dibley, D. Sibbritt, and T. T. M. Hanh (2007). Prevalence of overweight and obesity in preschool children and associated socio-demographic factors in Ho Chi Minh City, Vietnam. *Pediatric Obesity 2*(1), 40–50.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, Pennsylvania 19103: Society for Industrial and Applied Mathematics.

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition methods in economics. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, pp. 1–102. Amsterdam: North Holland.

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.

Hoang, L. V. (2009). Analysis of calorie and micronutrient consumption in vietnam. Technical report, DEPOCEN working paper 2009/14.

Horowitz, J. (2011). Applied nonparametric instrumental variables estimation. *Econometrica 79*(2), 347–394.

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

Le Nguyen, B., H. Le Thi, V. Nguyen Do, N. Tran Thuy, C. Nguyen Huu, T. Thanh Do, P. Deurenberg, and I. Khouw (2013). Double burden of undernutrition and overnutrition in Vietnam in 2011: Results of the SEANUTS study in 0.5-11-year-old children. *British Journal of Nutrition 110*(S3), S45–S56.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: Theory and practice*. Princeton and Oxford: Princeton University Press.

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

Ministry of Health (2012). *National Nutrition Strategy for 2011-2020, with a Vision Towards 2030*. Hanoi: Medical Publishing House.

Mishra, V. and R. Ray (2009). Dietary diversity, food security and undernourishment: The Vietnamese evidence. *Asian Economic Journal 23*(2), 225 – 247.

Muth, M. K., S. A. Karns, S. J. Nielsen, J. C. Buzby, and H. F. Wells (2011). Consumer-level food loss estimates and their use in the ERS loss-adjusted food availablity data. *USDA-ESR, Technical Bulletin Number 1927*.

National Institute of Nutrition (2007). *Vietnamese Food Composition Table*. Ministry of Health, Hanoi, Vietnam.

Nguyen, B. T., J. W. Albrecht, S. B. Vroman, and M. D. Westbrook (2007). A quantile regression decomposition of urban–rural inequality in Vietnam. *Journal of Development Economics 83*(2), 466–490.

Nguyen, M. C. and P. Winters (2011). The impact of migration on food consumption patterns: The case of Vietnam. *Food Policy 36*, 71–87.

Nguyen, T. T. and M. V. Hoang (2018). Non-communicable diseases, food and nutrition in Vietnam from 1975 to 2015: The burden and national response. *Asia Pacific Journal of Clinical Nutrition 27*(1), 19–28.

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. (2006). Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases (NCDs). *American Journal of Clinical Nutrition 84*, 289–298.

Racine, J. and C. Parmeter (2014). Data-Driven Model Evaluation: A Test for Revealed Performance. In J. Racine, L. Su, and A. Ullah (Eds.), *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pp. 308–345. Oxford: Oxford University Press.

Ravallion, M. (1990). Income effects on undernutrition. *Economic Development and Cultural Change 38*(3), 323–337.

Rothe, C. (2015). Decomposing the composition effect: the role of covariates in determining between-group differences in economic outcomes. *Journal of Business and Economic Statistics 33*(3), 323–337.

Santaeulàlia-Llopis, R. and Y. Zheng (2017). Why is food consumption inequality underestimated? A story of vices and children. *Barcelona GSE Working Paper 969*.

Santeramo, F. G. and N. Shabnamb (2015). The income-elasticity of calories, macro- and micro-nutrients: What is the literature telling us? *Food Research International 76*, 932–937.

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

Stiglitz, J. (1976). The efficiency wage hypothesis, surplus labor, and the distribution of income in LDCs. *Oxford Economic Papers 28*, 185?07.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics 8*, 1348–1360.

Subramanian, S. and A. Deaton (1996). The demand for food and calories. *Journal of Political Economy 104*(1), 133–162.

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.

Trinh, H. T., J. Morais, C. Thomas-Agnan, and M. Simioni (2018). Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical Methods in Medical Research*, DOI:10.1177/0962280218770223.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wood, S. N. (2017). *Generalized additive models: An introduction with R.* 2nd Edition, Chapman and Hall/CRC.

Wood, S. N. and N. H. Augustin (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling 157*(2), 157–177.

World Bank (2010). *How deep was the impact of the economic crisis in Vietnam? A focus on the informal sector in Hanoi and Ho Chi Minh City.* Washington, DC: World Bank.

Zezza, A., C. Carletto, J. L. Fiedler, P. Gennari, and D. Jolliffe (eds) (2017). *Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES).* Food Policy, Special Issue, Volume 72.

Zhou, J. and X. Yu (2015). Calorie elasticities with income dynamics: Evidence from the literature. *Applied Economic Perspectives and Policy 37*(4), 575–601.

Appendices

A Generalized additive models

GAMs can be viewed as extensions of Generalized Linear Models, or GLM. Classical linear regression model for a conditionally normally distributed response y

assumes that (i) the linear predictor through which $\mu_i \equiv \mathbb{E}(y_i|x_i)$ depends on the vector of the observations of the covariates for individual i, or x_i, can be written as $\eta_i = x_i'\beta$ where β represents a vector of unknown regression coefficients (ii) the conditional distribution of the response variable y_i given the covariates x_i is normally distributed with mean μ_i and variance σ^2, and (iii) the conditional expected response is equal to the linear predictor, or $\mu_i = \eta_i$. GLMs extend (ii) and (iii) to more general families of distributions for y and to more general relations between the expected response and the linear predictor than the identity. Specifically, y_i given x_i may now follow a probability density functions of the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{A-1}$$

where $b(.)$, $a(.)$ and $c(.)$ are arbitrary functions, and, for practical modelling, $a(\phi)$ is usually set to ϕ. θ, called the "canonical parameter" of the distribution, depends on the linear predictor. and ϕ is the dispersion parameter. Equation (A-1) describes the exponential family of distributions which includes a number of well-known distributions such as the normal, Poisson and Gamma. Finally, the linear predictor and the expected response are now related by a monotonic transformation $g(.)$, called the link function, i.e. $g(\mu_i) = \eta_i$

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

$$\eta_i = x_i\beta + \sum_j s_j(z_{ji}) \tag{A-2}$$

where β represents the vector of unknown regression coefficients for the covariates acting linearly (usually discrete covariates), and the $s_j(.)$ are unknown smooth functions of the covariates z_{ji}. The smooth functions can be function of a single covariate as well as of interactions between several covariates.

Recent papers in the literature on the estimation of calorie-income relationship, Tian and Yu (2015) and Nie and Sousa-Poza (2016), generalize the traditional 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

$$\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j \tag{A-3}$$

with the assumption that $PCCI$ is normally distributed. This equation can be viewed as a special case of the general GAM specification presented above. More general semiparametric specifications such as

$$g\left(\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j)\right) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j. \tag{A-4}$$

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

Estimation of GAM is usually performed using penalized regression splines and

can be implemented using package mgcv in R. We refer the reader to Wood and Augustin (2002), Wood (2003), and Wood (2017) for more details.

B Testing linearity of the calorie-income relationship

This appendix is devoted to the presentation of the test of the significance and linearity of the calorie-income relationship. Testing the linearity involves testing the nullity of the parameter α_2 in equation (1) when DLM is the chosen model. The procedure is as follows when a GAM model is chosen. The smooth function $s(x)$ in equations (3) and (4) is expressed as a linear (in parameters) basis expansion of the form

$$s(x) = \gamma_0 + \gamma_1 x + \sum_{i=1}^{n} \delta_i (x - x_i)^3 \tag{A-5}$$

when estimating GAM models. γ_0, γ_1, and the δ_i, $i = 1, \ldots, n$, are thus parameters to be estimated, the expansion (A-5) using thin plate regression splines (Wood, 2003). (A-5) which includes a linear function in x, is very useful when testing the linearity of the smooth function. This amounts to test the nullity of the nonlinear part in expansion (A-5). This test can be implemented by

1. estimating the chosen GAM specification
   - including now $INCOME$ in the regressors entering linearly, and
   - setting $\gamma_0 = \gamma_1 = 0$ in the expansion (A-5) of the smooth function with $x = INCOME$,
2. testing the nullity of the nonlinear remaining term of the expansion, we denoted by $s_{NL}(.)$, i.e. $s_{NL}(x) \equiv \sum_{i=1}^{n} \delta_i (x - x_i)^3$

This amounts to perform a F-type test.

Table 2: Results of significance and linearity tests

Year:	2004	2006	2008	2010	2012	2014
Model:	DLM	DLM	DLM	GAMGamLog	GAMGauId	GAMGauId
Significance test when DLM chosen:						
$H_0: \alpha_1 = 0$ and $\alpha_2 = 0$	128.81***	135.21***	238.92***	—	—	—
Linearity test when DLM chosen:						
$\widehat{\alpha}_1$	0.365***	0.414***	0.333***	—	—	—
$\widehat{\alpha}_2$	-0.02***	-0.023***	-0.016***	—	—	—
Significance test when GAM chosen:						
$H_0: s(.) = 0$	—	—	—	32.543***	26.831***	29.115***
Linearity test when GAM chosen:						
$\hat{\gamma}_1$	—	—	—	3.544***	5.168***	3.144**
$H_0: s_{NL}(.) = 0$	—	—	—	16.459***	16.693***	14.8***

Note:
(1) Reported values for testing either $H_0: \alpha_1 = 0$ and $\alpha_2 = 0$, $H_0: s(.) = 0$, or $H_0: s_{NL}(.) = 0$ are F-statistics.
(2) $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are estimated values of parameters α_1 and α_2 in DLM models.
(3) $\widehat{\gamma}_1$ is estimated value of parameter γ_1 in GAM models.
(4) *, **, and *** mean significant at 10%, 5%, and 1%, respectively

Significance tests are reported in Table 2. The tests clearly reject null hypothesis $H_0: \alpha_1 = 0$ and $\alpha_2 = 0$ when the chosen model is DLM, or $H_0: s(.) = 0$ when it is GAM. Table 2 reports also the results from linearity tests. The parameter α_2 is significantly different from zero when the chosen model is DLM. Moreover the nullity of $s_{NL}(.)$ is clearly rejected when the chosen model is GAM. Linearity is thus rejected whatever the chosen model.

C Revealed performance test

Assuming that the data represent independent draws, as they would in a standard cross-sectional setup like a wave of VHLSS, the implementation of the revealed performance test proposed by Racine and Parmeter (2014) involves the following steps:

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. $(y_i^*, x_i^*)_{i=n_1+1}^n$.[6]
3. Fit each model using only the training observations $(y_i^*, x_i^*)_{i=1}^{n_1}$. Denote here by $\widehat{m}_j(.)$, $j = 1, \ldots, k$, these estimates. Then compute predicted values for the evaluation observations $(y_i^*, x_i^*)_{i=n_1+1}^n$, i.e. $\widehat{y}_{i,j} = \widehat{m}_j(x_i^*)$, $i = n_1 + 1, \ldots, n$.
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 $(ASPE_{jb})_{b=1}^B$.[7]

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.

D VHLSS

This study relies on Vietnam Household Living Standard Surveys, or VHLSS. VHLSS is conducted by the General Statistics Office of Vietnam, with technical assistance of the World Bank, every two years since 2002. Its main objective is to collect information to be used as foundation for rating living standards, poverty and rich-poor gap, which helps Vietnamese policy-makers to define programs to improve household living standards across the country, regions and provinces. Each VHLSS wave consists of two surveys: for household and for commune. Household

[6] 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.

[7] 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 $B = 10,000$.

survey includes information reflecting living standards, including income and expenditure, assets, housing and key household facilities, and some key information affecting living standards such as levels of education, employment and involvement in poverty reduction programs. Commune survey reports socio-economic features affecting household living standards in the commune such as key socio-economic infrastructure structures, agricultural production, off-farm job opportunities, and some key information on social order and safety, and environmental protection.

The target population of VHLSS comprises the civilian, non-institutionalized population of Vietnam. The sampling unit is the household. The VHLSS defines household membership on the basis of physical presence: Individuals must eat and live with other members for at least six out of the past twelve months, and contribute to collective income and expenses. Among other things, this means that family members who have moved away to work or school (e.g., migrants) are not considered household members.

Sample design used in VHLSS is a two-stage area sample design where communes are selected in first stage, and three enumeration areas, or EAs, per commune are selected in second stage. EAs are defined by Population Census (1999 and 2009) and are of comparable sizes (around 105 or 99 households in urban or rural areas, respectively). This sample design solves the problem due to the large size of some communes because only one of the selected EAs is surveyed in each waves. Moreover, the design allows for rotation of EAs rather than households in each EA, which is operationally simpler. Communes are stratified on province and urban/rural and the sample is allocated over strata proportionally to the square root of the total number of households in each strata. Both communes and EAs are then selected with probability proportionate to the number of households according to Population Census. Surveyed households in each selected EA are selected based on the most recent list of households in the selected EAs (three months before the field work of surveyors).

VHLSS can be viewed as a rotating panel. Sample design for each waves of VHLSS implies 50% rotation of households and a household can only be tracked for three years. In this study, we consider each wave independently to keep enough waves in our analysis.

E Calculating per capita calorie intake

VHLSS is not, by definition, constructed to assess the nutritional status of Vietnamese households. Thus, the most difficult task in cleaning data is the computation of total household calorie intake and, then, per capita calorie intake. The survey collect data on both purchased goods and self-supplied food (home production) for a wide range of food items. Food expenditures are transformed into kilocalories using a conversion table built by the Vietnamese National Institute of Nutrition in 2007. Conversion factors are summarized in Table 3. In this table, when a given food item such as other types of meat does not appear in the conversion table, we associate a caloric content calculated following Hoang (2009). First, we compute the price of one calorie of all the food items which we have both quantity (and thus the corresponding calorie intake) and expenditure. Second, for each food item with only expenditure information, we approximate calorie intake by dividing the expenditure by the average calorie price taken from a list a corresponding food items (for instance, pork, beef, buffalo meat, chicken meat, duck and other poultry

meat for other types of meat).[8]

Table 3: Conversion table

Food item	Calories	Food item	Calories
Plain rice	344.5	Cabbage	29
Sticky rice	347	Tomato	20
Maize	354	Other vegetables	–
Cassava	146	Banana	37
Potato of various kinds	106	Orange	81.5
Wheat grains, bread, wheat powder	313.7	Mango	69
Flour noodle, instant rice noodle	349	Other fruits	–
Fresh rice noodle, dried rice noodle	143.0	Fish sauce	60
Vermicelli	110	Salt	0
Pork	260	MSG	0
Beef	142.5	Glutamate	0
Buffalo meat	122	Sugar, molasses	390
Chicken meat	199	Confectionery	412.2
Duck and other poultry meat	275	Condensed milk, milk powder	395.7
Other types of meat	–	Ice cream, yoghurt	–
Processed meat	–	Fresh milk	61
Lard, cooking oil	827	Alcohol of various kinds	47
Fresh shrimp, fish	83	Beer of various kinds	11
Dried and processed shrimps, fish	361	Bottled, canned, boxed beverages	47
Other aquatic products and seafood	–	Instant coffee	0
Eggs of chickens, ducks, Muscovy	103.7	Coffee powder	353
Tofu	95	Instant tea powder	0
Peanuts, sesame	570.5	Other dried tea	0
Beans of various kinds	73	Cigarettes, waterpipe tobacco	0
Fresh peas of various kinds	59	Betel leaves, Areca nuts, lime, betel pieces	0
Morning glory vegetables	25	Outdoors meals and drinks	–
Kohlrabi	36	Other food and drinks	–

Notes:
(1) Unit = KCal per 100gr.
(2) Source: National Institute of Nutrition (2007).

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

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

[8] Details on the chosen approximation method are available upon request to the authors.

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 not 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 scale, and calculate the adult equivalent of household calorie intake by dividing household total calorie intake by the total number of members in the household, leading to $\theta_{g,a} = 1$, whatever the age or gender of the household members. Some papers address this issue using either the "old" 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 each child, whatever their age or gender, or the modified OECD equivalence scale, i.e., setting $\theta_{g,a} = 0.5$ for each adult other than the head of the household, whatever the gender, and $\theta_{g,a} = 0.3$ for each child, whatever their age or gender (OECD, 2013). Here, to calculate equivalence scales, we proceed as Aguiar and Hurst (2013). First, we estimate the following regression model It is not 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 scale, and calculate the adult equivalent of household calorie intake by dividing household total calorie intake by the total number of members in the household, leading to $\theta_{g,a} = 1$, whatever the age or gender of the household members. Some papers address this issue using either the "old" 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 each child, whatever their age or gender, or the modified OECD equivalence scale, i.e., setting $\theta_{g,a} = 0.5$ for each adult other than the head of the household, whatever the gender, and $\theta_{g,a} = 0.3$ for each child, whatever their age or gender (OECD, 2013). Here, to calculate equivalence scales, we proceed as Aguiar and Hurst (2013). First, we estimate the following regression model

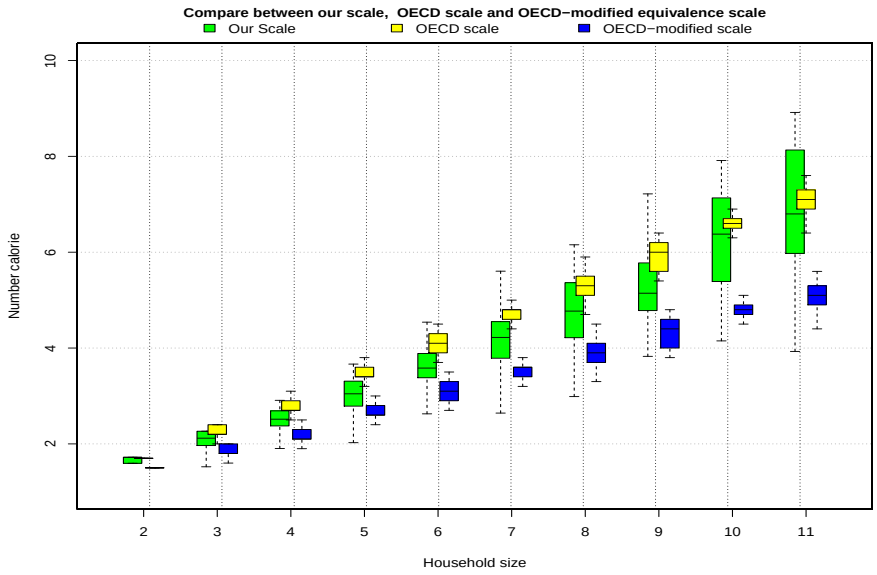
$$\log(THCI_i) = \gamma_0 + \gamma_1\, Gender_i + \gamma_2\, N_{a,i} + \gamma_3\, Family_i + \varepsilon. \tag{A-6}$$

where $THCI_i$ is total household i calorie intake, $Gender_i$ is the gender of the head of the household (male is taken as the reference), $N_{a,i}$ is the number of adults in the household other than the head, and $Family_i$ 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 VHLSS wave as in Santaeulàlia-Llopis and Zheng (2017). Then we use the exponentiated predicted value of $THCI_i$, 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.

Figure 3 gives the computed values of equivalence scales using either OECD or Aguiar and Hurst (2013) methodologies for 2012 VHLSS wave. As expected, equivalence scales are increasing with respect to household size. Equivalences scales computed using Aguiar and Hurst (2013) are between the equivalence scales calculated according to OECD for most household size, and exhibit more variability than the two other scales.

Table 4 reports the average value of adult equivalent calorie intake for each VHLSS wave and compares it with other available studies on Vietnam. The average values we obtained are consistent with those obtained in other papers using the same

Figure 3: Comparison of equivalence scales using 2012 VHLSS data



survey data. They are just a little higher, which we could be foreseen as the other studies use total calorie intake divided by household size.

Table 4: Average per capita calorie intake: Comparison with other papers

	2004	2006	2008	2010	2012	2014
Mishra and Ray (2009): Rural	3206					
Mishra and Ray (2009): Urban	2824					
Hoang (2009)	2348					
Nguyen and Winters (2011)	3144	3074				
Our study	3291	3272	2818	3632	3611	3651
FAO, IFAD and WFP (2015)	2478	2483	2615	2678	2713	na
Note: unit = KCal						

The average values of $PCCI$ can compared with similar values provided by public agencies working on food security in the world. The survey data seem to lead to overestimation of average individual calorie intakes when compared to figures from FAO, IFAD and WFP (2015), as shown in Table 4. It should then be emphasized that different data collection procedures as well as different procedures for computing per capita calorie intake can explain these differences. For their part, figures given by FAO are obtained from food balance sheets at the country level. The per capita supply of each food item is then obtained by dividing the quantity of the food item available for human consumption in the country by its total number of inhabitants. Data on per capita food supplies are expressed as quantities. Then applying appropriate food composition factors for all primary and processed products produces data in terms of dietary energy value, protein and fat content.

VHLSS data, however, are not collected for the purpose of providing information

on nutrition. It is well known that data such as those of VHLSS surveys always overestimate calorie intakes (Bouis and Haddad, 1992). They give a measure of calorie availability at the household level rather than calorie intake of members of that same household. Indeed, they do not include losses and waste from food preservation and preparation. These losses were evaluated for each food item in the US (Muth et al., 2011). They range from 4% for low-fat cottage cheese to 69% for fresh pumpkin, with a remarkable 33% for rice. Such reliable data on food losses and waste are not still available for Vietnam, and differences in consumption habits between the two countries prevent us from applying the estimated loss coefficients for the US to Vietnamese data. The correction as proposed in Muth et al. (2011) is based on the assumption that there is a systematic bias to overestimation when transforming consumption data into nutrition data. This bias is assumed to be the same regardless of the considered household. Due to lack of data allowing a thorough treatment of this assumption, we maintain it in this paper.

Another source of overestimation of calorie intake is the possible substitutability within each of the food groups. As emphasized by Bouis and Haddad (1992), household expenditure for a food aggregate may increase in response to higher income, without a proportionate increase in calorie intake because of within-group substitution toward more expensive calorie sources. The availability of the total quantity purchased only for each food aggregate does not make it possible to evaluate this substitution effect towards better calorie sources when income increases. Further analysis of the impact of these potential substitutions would require more detailed data on household food purchases such as, for example, the brands purchased and the nutritional composition of these brands, data that are not available in a survey such as VHLSS. Nevertheless, the availability of a fairly large number of very detailed food groups may help mitigating this substitution effect.

F Test of exogeneity

The test of exogeneity proposed by Blundell and Horowitz (2007) exploits directly the conditional mean restriction that can be used to identify a nonparametric instrumental variable model. This condition can be written as follows. Let Y be a scalar variable, X, an endogenous explanatory variable, and W, an instrumental variable. The function g is a nonparametric function that is identified by the the conditional mean restriction:

$$\mathbb{E}\left[Y - g(X)|W\right] = 0 \tag{A-7}$$

Now, define the conditional mean function $G(x) = \mathbb{E}(Y|X = x)$. X is said to be exogenous if $g(x) = G(x)$. Otherwise, X is said to be endogenous. From Eq. (A-7), testing the null hypothesis, H_0, that X is exogenous, against the alternative hypothesis, H_1, that X is endogenous, is equivalent to testing the hypothesis $\mathbb{E}(Y - G(X)|W) = 0$.

The test statistics proposed by Blundell and Horowitz (2007) is

$$\tau_n = \int S_n^2(x)dx \tag{A-8}$$

where $S_n(x)$ is the sample analogue of $S(x) = \mathbb{E}\left\{\left[Y - G(X)\right] f_{XW}(x, W)\right\}$ which is obtained by replacing the unknown regression model G and joint density f_{XW} by

leave-one-observation-out kernel estimators. H_0, the null hypothesis of exogeneity, is rejected if τ_n is large.

Blundell and Horowitz (2007) show that, under H_0, the test statistics can be written as an infinite weighted sum of independent chi-square random variables. Notice that, under H_0, $G = g$, so knowledge of or estimation of g is not needed to obtain the asymptotic distribution of τ_n under H_0. Weights are eigenvalues of a matrix whose sample analogue can be easily computed using nonparametric kernel estimate of f_{XW} and estimated errors $\widehat{U}_i = Y_i - \widehat{G}(X_i)$. The test statistics can then be approximated by a finite sum of independent chi-square distributed random variables where the weights are now the non vanishing eigenvalues of this sample analogue. An application of the test is given in Blundell et al. (2012).

Here, the bandwidths we use to estimate f_{XW} and G are selected by cross-validation, and the kernel is the Epanechnikov kernel (Li and Racine, 2007). The selected number of eigenvalues used for calculating the simulated values of the test statistic under H_0 is 25. $100,000$ values are simulated and the p-value corresponding to the computed test statistics is obtained as the share of simulated values larger than it.

Table 5: VHLSS data: Some summary statistics

Variable		Description	2004	2006	2008	2010	2012	2014
PCE		Per capita expenditure (US$)	335.3	374.6	435.8	570.5	597.3	622.2
			(211.8)	(239.4)	(272.3)	(337.2)	(342.8)	(343.6)
$Urban$	1	Urban	23.32 %	24.42 %	25.11 %	27.49 %	28.4 %	29.04 %
	0	Rural	76.68 %	75.58 %	74.89 %	72.51 %	71.6 %	70.96 %
$Hsize$	2	≤ 2 people	10.39 %	12.11 %	14.04 %	15.91 %	17.36 %	18.96 %
	3	3 people	15.36 %	16.56 %	17.07 %	19.87 %	18.95 %	19.97 %
	4	4 people	30.61 %	31.27 %	31.78 %	33.81 %	32.52 %	31.09 %
	5	5 people	21.74 %	20.72 %	19.45 %	16.81 %	17.55 %	16.6 %
	6	≥ 6 people	21.9 %	19.34 %	17.66 %	13.6 %	13.63 %	13.37 %
$Ethnic$	1	Kinh	84.88 %	84.25 %	84.74 %	82.26 %	82.2 %	82.76 %
	0	Minorities	15.12 %	15.75 %	15.26 %	17.74 %	17.8 %	17.24 %
$Gender$	1	Male	77.1 %	76.36 %	76.36 %	76.14 %	76.28 %	75.63 %
	0	Female	22.9 %	23.64 %	23.64 %	23.86 %	23.72 %	24.37 %
Wa	1	Clean water	69.17 %	60.5 %	63.97 %	62.38 %	65.22 %	68.9 %
	0	Unclear water	30.83 %	39.5 %	36.03 %	37.62 %	34.78 %	31.1 %
$Educ$	1	Below primary	54.92 %	53.27 %	52.04 %	52.08 %	51.23 %	49.64 %
	2	Secondary, High school	41.07 %	42.52 %	43.82 %	42.48 %	43.39 %	44.35 %
	3	University	4.01 %	4.2 %	4.14 %	5.43 %	5.38 %	6.02 %
$Area$		Red River Delta	21.44 %	21 %	21 %	21.03 %	20.99 %	21.23 %
		Midlands Northern Mountains	19.58 %	19.54 %	18.92 %	17.94 %	18.14 %	18.1 %
		Northern Central Coast	20.01 %	20.29 %	20.46 %	22.03 %	21.65 %	21.53 %
		Central Highlands	6.41 %	6.22 %	6.42 %	6.88 %	6.79 %	6.49 %
		South East	11.79 %	12.2 %	12.53 %	11.35 %	11.59 %	11.72 %
		Mekong River Delta	20.76 %	20.74 %	20.67 %	20.77 %	20.84 %	20.93 %
N		Nb of observations	8269	8325	8305	8469	8439	8427

Table 6: t-paired test results

Year	Model	GAMGauId	GAMGauLog	GAMGamLog	Choice
2004	DLM	-11.64***	-10.20***	-14.70***	DLM
	GAMGauId		4.67***	-7.78***	
	GAMGauLog			-11.89***	
2006	DLM	17.14***	12.79***	9.3***	GAMGauId
	GAMGauId		-19.6***	-29.49***	
	GAMGauLog			-11.9***	
2008	DLM	62.38***	21.77***	13.67***	GAMGauId
	GAMGauId		-87.88***	-95.8***	
	GAMGauLog			-19.98***	
2010	DLM	19.26***	-10.02***	-16.74***	GAMGauId
	GAMGauId		-73.06***	-79.93***	
	GAMGauLog			-15.04***	
2012	DLM	58.25***	2.41*	-5.34***	GAMGauId
	GAMGauId		-164.72***	-149.59***	
	GAMGauLog			-16.28***	
2014	DLM	70.01***	-23.97***	-49.93***	GAMGauId
	GAMGauId		-174.34***	-163.31***	
	GAMGauLog			-31.15***	

Note: *, **, and *** mean significant at 10%, 5%, and 1%, respectively