WORKING PAPERS

N° 17-842

September 2017

"A Fresh Look at the Nutrition Transition in Vietnam using Semiparametric Modeling"

Huong Thi Trinh, Michel Simioni and Christine Thomas-Agnan

Toulouse School of Economics

A Fresh Look at the Nutrition Transition in Vietnam using Semiparametric Modeling

Huong Thi TRINH
Toulouse School of Economics, INRA, University of Toulouse Capitole
21 Allée de Brienne, 31015 Toulouse Cedex 6, France
trinhthihuong@tmu.edu.vn

Michel SIMIONI*
INRA, UMR-1110 MOISA
2, place Pierre Viala - Bât. 26, 34 060 Montpellier Cedex 2, France
michel.simioni@inra.fr
Phone: + 33 (0)4 99 61 24 79

and

Christine THOMAS-AGNAN
Toulouse School of Economics, University of Toulouse Capitole
21 Allée de Brienne, 31015 Toulouse Cedex 6, France
christine.thomas@tse-fr.eu

July 2017

*Corresponding author

Abstract

Policies aimed at reducing starvation and redressing nutritional deficiencies remain among the most widely accepted policies in the world. 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. This paper revisits the issue of estimating the relationship between calorie intake and income. We present and compare estimates of this relationship for Vietnam which has undergone profound economic changes over the last 30 years. After estimating semiparametric generalized additive models are estimated, we compare their performances are compared to the performance of the classical double log model using the recently proposed revealed performance test. This methodology is implemented using successive waves of the Vietnam Household Living Standard Survey. The application delivers some new and interesting insights on nutritional transition in Vietnam between 2004 and today. The analysis focuses not only on the comparison of the general shape of the estimated curves but also on the decomposition of the evolution of average calorie intake in terms of changes in the structure of the surveyed sample (greater urbanization, for example) and changes in preferences as reflected by changes in the estimated curves over time.

Keywords: Calorie – income relationship, Nutritional transition, Vietnam, Semiparametric modeling, decomposition methods.

1 Introduction

Since the launch of economic reforms in 1986, Vietnam has recorded impressive achievements in growth performance and, at the same time, has also experienced a nutrition transition like many other middle-income countries in South East Asia. Dietary diversity from 2005 to 2015 in South-East Asia 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). On one hand, this nutrition transition to energy-dense, poor quality diets has led to obesity and diet-related chronic diseases. The percentage of overweight people in the whole population of Vietnam is 21% in 2014, the percentage of obese people being 4% (source: WHO, 2015).[1] On the other hand, a sizeable share of the population, 11%, still experiences undernutrition in Vietnam. This double burden is even more striking regarding childhood where the number of overweight individuals has more than doubled in a very short time at the same time as that of malnourished decreased (Li and Dibley, 2012).

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 are 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 a total number of 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

[1] An individual is said to be overweight (resp. obese) when her body mass index is greater than or equal to 25 (resp. 30). (Body mass index $=$ weight/height^2.)

much discussed. 86 over the 99 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), and 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 there is "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 contributes to the literature mentioned above. It aims to analyze the nutritional transition in Vietnam from 2004 to the present days by analyzing the evolution of the relationship between caloric intake and income over this period. The availability of consumption data for representative samples of the Vietnamese population over six years (the six recent waves of Vietnam Household Living Standards Survey, or VHLSS, for 2004, 2006, 2008, 2010, 2012 and 2014) makes this analysis possible. The availability of cross-sectional data every two years and not those for a panel of households followed throughout the considered years led us to propose an original methodological approach to analyze the nutritional transition in Vietnam.

In a first step, we seek the functional form that best describes the relationship between calorie intake and income for a given year. 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. The accuracy of our nonparametric estimates will be low even if we are 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 more general 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 for the GAM specifications to describe the relationship between calorie intake and income, following the approach suggested above, and we must choose among them. In a second step, we use a cross-validation procedure proposed recently by Racine and Parmeter (2014), namely "revealed performance test" or RPF, to choose among various parametric and semiparametric specifications of the relationship between calorie intake and income. 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 performances on unseen data coming from the same data generating process. The RPF 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).

The analysis of the evolution of the calorie intake - income relationship over the studied period is not easy because this relationship is estimated from 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 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) (see Fortin et al., 2011, for an extensive survey). In a third step, 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 in it the previously chosen semiparametric estimates of the relationship under investigation.

The methodology presented above is implemented using six waves of VHLSS. Among other findings, results highlight the major role played by changes in eating habits, as reflected by changes in the estimated relationship between calorie intake and income, in the consequent increase of average calorie intake in Vietnam between 2004 and 2014. Changes in sample structure play only a marginal role although their effect has increased over the period.

The paper is organized as follows. Section 2 presents the methodology used in this paper. Section 3 is devoted to the presentation of VHLSS data and to the approach chosen when converting expenditure data into quantities of calories. 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, $\left(\log(\text{INCOME})\right)^2$, 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 $\log(\text{INCOME})$ 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.

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

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{2}$$

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 (2) describes the exponential family of distributions which includes a number of well-known distributions such as the normal, Poison 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

$$\eta_i = x_i\beta + \sum_j s_j(z_{ji}) \tag{3}$$

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.

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

$$\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j \tag{4}$$

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

$$g\left(\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j)\right) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j. \tag{5}$$

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 about to the conditional distribution of calorie intake given income and various control variables.

Estimation of GAM is performed using penalized regression splines. We refer the reader to Wood and Augustin (2002), Wood (2003), and Wood (2017) for more details.

2.2 Revealed Performance test

We will estimate later on various models 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. $(y_i^*, x_i^*)_{i=n_1+1}^n$.[2]
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 $(ASPE_{jb})_{b=1}^B$.[3]

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

[2] 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.

[3] 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 $B = 10,000$.

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

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

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

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

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 (7) becomes

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

Finally we can write the difference as

$$\begin{aligned} \Delta PCCI_{t_0 \to t_1} = \quad & \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) + \\ & \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z)) \end{aligned} \tag{9}$$

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 (9) 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 (9), 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 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 (9) 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).

3 Data

3.1 Data set

This study relies 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 areas. In this study, we use the most recent six 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. 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 A2).

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 better reflect 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 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. This issue can then 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

$$\log(THCI) = \gamma_0 + \gamma_1 \, Gender + \gamma_2 \, N_a + \gamma_3 \, Family + \varepsilon. \tag{10}$$

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 (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: 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					
Vu (2009)	2348					
Nguyen and Winter (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						

Table 1 reports the average value of adult equivalent calorie intake for each year in the survey and

compares it with other studies. The average values we obtained are consistent with those obtained in other papers using the same survey data. They are just a little higher, which we could be foreseen 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 years 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: 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), $= 0$ 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 region where the household is located (Vietnam is divided into six regions). Table A3 summarizes the main characteristics of these variables.

4 Results

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

$$\log(PCCI) = \alpha_0 + \alpha_1 \log(INCOME) + \alpha_2 \left(log(INCOME)\right)^2 + \sum \beta \, Factors + \epsilon, \quad (11)$$

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

$$\mathbb{E}(PCCI|\text{INCOME}, Factors) = \alpha_0 + s(INCOME) + \sum \beta\, Factors. \tag{12}$$

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

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

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

4.1 Chosen models

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 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 -95.24. This figure indicates that the average difference between the ASPE criteria obtained for the two models, computed using the $10,000$ splits of the Vietnamese data following the procedure presented in subsection 2.2, 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, 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: -94.81 and -100.6 when comparing DLM to GAMGauLog and GAMGamLog, respectively. Thus, whatever the relative performances of the other three

Table 2: t-paired test results

Year	Model	GAMGauId	GAMGauLog	GAMGamLog	Choice
2004	DLM	-95.24***	-94.81***	-100.6***	DLM
	GAMGauId		-10.21***	-15.6***	
	GAMGauLog			-2.72**	
2006	DLM	-62.08***	-74.38***	-68.77***	DLM
	GAMGauId		-34.64***	-32.39***	
	GAMGauLog			6.2***	
2008	DLM	-13.83***	-23.46***	-17.14***	DLM
	GAMGauId		-27.41***	-11.12***	
	GAMGauLog			17.05***	
2010	DLM	1.28	10.23***	11.39***	GAMGamLog
	GAMGauId		43.89***	49.88***	
	GAMGauLog			5.29***	
2012	DLM	4.47***	-3.28**	-3.91***	GAMGauId
	GAMGauId		-31.54***	-35.84***	
	GAMGauLog			-3.56***	
2014	DLM	1.44	-37.54***	-30.65***	GAMGauId
	GAMGauId		-87.15***	-87.81***	
	GAMGauLog			15.61***	

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

specifications when compared among themselves (GAMGauId, GAMGauLog and GAMGamLog), the chosen specification for 2004 is DLM.

The same reading grid can then be applied to the other results reported in Table 2. Its last column summarizes which model is chosen after applying the revealed performance test for each year. The results clearly indicate that DLM is chosen when compared to semiparametric models for the 2004, 2006, and 2008 VHLSS waves. 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.

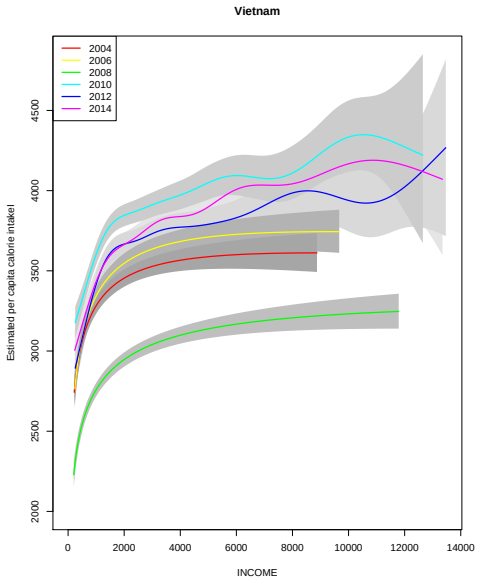
4.2 The estimated calorie intake – income relationships

Figure 1 reports per capita calorie intake as a function of income, the control variables being fixed to their mode values in 2004, for the different waves of VHLSS (shaded areas give the 95% confidence intervals around the estimated relationships.).[4] The nonlinearity of the relationship clearly appears

[4]The chosen household chosen 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. Tables A4 report the estimated values of the parameters associated to the variables entering linearly in the chosen specification for each survey wave in each country.

in view of the different curves traced in Figure 1. This result is confirmed by the various significance and linearity tests presented in Appendix 1. The relationship appears to be concave for both waves. Specifically, 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). Moreover, this general shape of the relationship is fairly constant over time.

Figure 1: Estimated calorie intake and income relationships



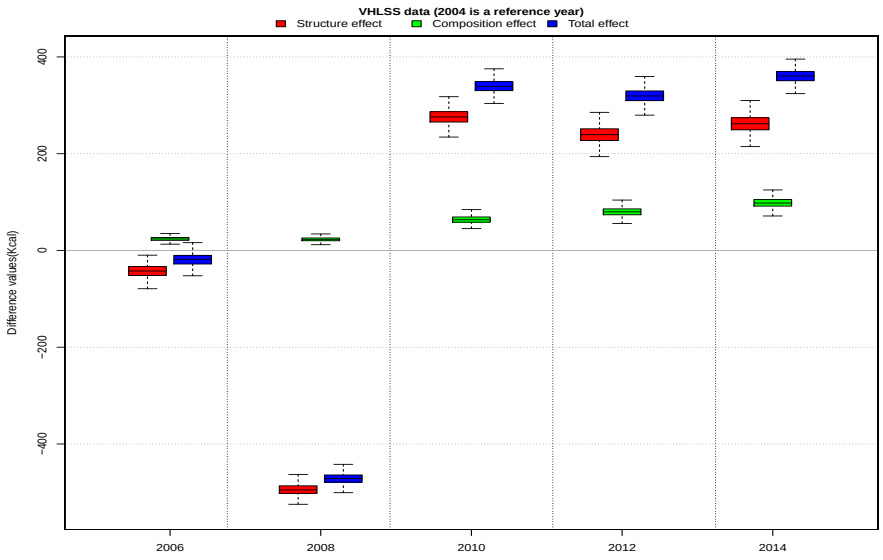
Previous results contribute to the debate on the extent to which calorie consumption responds to income changes in middle-income countries. They clearly show that income mediated policies can have an impact on nutritional goals up to a given threshold of income in Vietnam. They show the rapid improvement of nutrition in terms of calorie intake for low incomes. They do not tell us anything about improving the nutritional quality of the diet. But they also show that from a certain level of income (between US dollars $1,000$ and $1,600$ depending on the years under consideration)[5] such income mediated policies may prove to be ineffective as calorie intake seems a little responsive to an increase of income.

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

4.3 The evolution of average calorie intake over 2004 to 2014

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. For instance, 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. These two findings may explain the important drop of the 2008 curve compared to curves estimated for other years. 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 reports the results of the decomposition described in equation (9). More precisely, we report a boxplot of the distribution of the difference 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 the difference of average per capita calorie intakes



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 structure and composition effects that compensate 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 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 Vietnam. For this, we use various recent tools in semiparametric econometrics, in model choice, and in decomposition in economics. The application uses different waves VHLSS for Vietnam 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 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.

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 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.

Acknowledgements

We would like to thank the participants of the Journées de Statistiques, Montpellier, June 2016, Vietnamese Economists Annual Meeting, Da Nang, August 2016, Journées de la Recherche en Sciences Sociales, Paris, December 2016, Academy fo Policy and Development seminar, Hanoi, April 2017, and 15th EAAE Congress, Parma, Italy, August 2017, for helpful comments. We are grateful to Thibault Laurent for technical assistance in R. Supports from the TAASE project of INRA-CIRAD GloFoodS meta-program are 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.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources 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. M. and J. S. Racine (2017). The smooth colonel and the reverend find common ground. *Econometric Reviews 36*(1-3), 241–256.

Li, M. and M. Dibley (2012). Child and Adolescent obesity in Asia. In L. Baur, S. Twigg, and R. Magnusson (Eds.), *A modern Epidemic: Expert persperctives on obesity and diabetes*, pp. 171–188. Sidney: Sidney University Press.

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.

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.

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.

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.

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

Silva, J. S. 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.

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.

Appendix 1: Linearity of the calorie intake – income relationship

This appendix is devoted to the presentation of the test of the significance and linearity of the calorie intake – income relationship. Testing the linearity involves testing the nullity of the parameter α_2 in equation (11) when DLM is the chosen model. The procedure is as follows when a GAM model is chosen. The smooth function $s(x)$ in equations (12) and (13) is expressed as a linear (in parameters) basis expansion of the form

$$s(x) = \gamma_0 + \gamma_1 x + \sum_{i=1}^{n} \delta_i (x - x_i)^3 \tag{14}$$

when estimating GAM models. γ_0, γ_1, and the δ_i, $i = 1, \ldots, n$, are thus parameters to be estimated, the expansion (14) using thin plate regression splines (Wood, 2003). (14) 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 (14). 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 (14) 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.

Significance tests are reported in Table A1. 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 A1 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.

Table A1: Results of significance and linearity tests

Year:	2004	2006	2008	2010	2012	2014
Model:	DLM	DLM	DLM	GAMGamLog	GAMGauld	GAMGauld
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

Appendix 2: Tables and Figures

Table A2: Conversion table

Food item	Calories	Food item	Calories
Plain rice	344.5	Sticky rice	347
Maize	354	Cassava	146
Potato of various kinds	106	Wheat grains, bread, wheat powder	313.7
Floor noodle, instant rice noodle, porridge	349	Fresh rice noodle, dried rice noodle	143
Vermicelli	110	Pork	260
Beef	142.5	Buffalo meat	122
Chicken meat	199	Duck and other poultry meat	275
Other types of meat	-	Processed meat	-
Fresh shrimp, fish	83	Dried and processed shrimps, fish	361
Other aquatic products and seafoods	-	Eggs of chicken, ducks, Muscovy ducks, geese	103.74
Tofu	95	Peanuts, sesame	570.5
Beans of various kinds	73	Fresh peas of various kinds	59
Morning glory vegetables	25	Kohlrabi	36
Cabbage	29	Tomato	20
Other vegetables	-	Orange	37
Banana	81.5	Mango	69
Other fruits	-	Fish sauce	60
Salt	0	MSG	0
Glutamate	0	Sugars, molasses	390
Confectionery	412.2	Condensed milk, milk powder	395.7
Ice cream, yoghurt	-	Fresh milk	61
Alcohol of various kinds	47	Beer of various kinds	11
Bottled, canned, boxed beverages	47	Instant coffee	0
Coffee powder	0	Instant tea powder	0
Other dried tea	0	Cigarettes, waterpipe tobacco	0
Betel leaves, areca nuts, lime, betel pieces	0	Outdoor meals and drinks	-
Other foods and drinks	0	Lard, cooking oil	863.5

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

Table A3: VHLSS data: Some summary statistics

Variable		Description	2004	2006	2008	2010	2012	2014
$INCOME$		Household income per year (US$)	1812.2	1985.00	2059.4	2474.4	2723.3	2923.8
			(1355.9)	(1476.8)	(1644.2)	(1903.7)	(2054)	(2142.2)
$Urban$	1	Urban	76.3 %	75.39 %	74.67 %	72.25 %	71.42 %	70.7 %
	0	Rural	23.7 %	24.61 %	25.33 %	27.75 %	28.58 %	29.3 %
$Hsize$	2	$\leq$ 2 people	10.53 %	12.36 %	14.23 %	15.95 %	17.39 %	19.03 %
	3	3 people	15.36 %	16.55 %	16.98 %	19.79 %	18.96 %	20 %
	4	4 people	30.42 %	30.91 %	31.53 %	33.48 %	32.43 %	30.88 %
	5	5 people	21.58 %	20.53 %	19.41 %	16.8 %	17.46 %	16.54 %
	6	$\geq$ 6 people	22.11 %	19.65 %	17.85 %	13.98 %	13.76 %	13.54 %
$Ethnic$	1	Kinh	84.35 %	83.71 %	84.17 %	81.69 %	81.72 %	82.22 %
	0	Minorities	15.65 %	16.29 %	15.83 %	18.31 %	18.28 %	17.78 %
$Gender$	1	Male	77 %	76.43 %	76.35 %	76.21 %	76.23 %	75.7 %
	0	Female	23 %	23.57 %	23.65 %	23.79 %	23.77 %	24.3 %
Wa	1	Clean water	69.42 %	60.38 %	63.83 %	62.14 %	64.98 %	68.55 %
	0	Unclear water	30.58 %	39.62 %	36.17 %	37.86 %	35.02 %	31.45 %
$Educ$	1	Below primary	54.94 %	53.38 %	52.22 %	52.08 %	51.21 %	49.55 %
	2	Secondary, High school	40.79 %	42.22 %	43.36 %	42.13 %	43.17 %	44.09 %
	3	University	4.28 %	4.39 %	4.41 %	5.8 %	5.62 %	6.36 %
$Area$		Red River Delta	21.4 %	20.88 %	20.96 %	16.66 %	16.44 %	21.14 %
		Midlands Northern Mountains	19.73 %	19.69 %	19.17 %	13.58 %	13.67 %	18.28 %
		Northern Central Coast	19.98 %	20.26 %	20.32 %	21.91 %	21.64 %	21.45 %
		Central Highlands	6.49 %	6.3 %	6.51 %	7.03 %	7.06 %	6.76 %
		South East	11.87 %	12.38 %	12.66 %	11.36 %	11.62 %	11.68 %
		Mekong River Delta	20.52 %	20.48 %	20.38 %	29.46 %	29.57 %	20.7 %
N		Nb of observations	8439	8495	8475	8645	8641	8601

Table A4: Parameter estimates

Variable - Levels		2004	2006	2008	2010	2012	2014
Intercept		6.487 ***	6.254 ***	6.37 ***	6.497 ***	5.875 ***	6.489 ***
		(0.225)	(0.244)	(0.189)	(0.232)	(0.253)	(0.245)
$Urban$	1	0.009	0.005	-0.034 ***	-0.028 ***	0.085 ***	0.096 ***
		(0.007)	(0.007)	(0.006)	(0.007)	(0.008)	(0.008)
$Hsize$	3	0.077 ***	0.06 ***	0.074 ***	0.038 ***	0.038 ***	0.051 ***
		(0.011)	(0.011)	(0.009)	(0.01)	(0.011)	(0.01)
	4	0.064 ***	0.064 ***	0.063 ***	0.04 ***	0.013	0.018 .
		(0.01)	(0.01)	(0.009)	(0.01)	(0.01)	(0.01)
	5	0.046 ***	0.03 **	0.027 **	0.016	-0.018	-0.012
		(0.011)	(0.011)	(0.01)	(0.011)	(0.012)	(0.011)
	6	-0.017	-0.033 **	-0.04 ***	-0.042 ***	-0.083 ***	-0.07 ***
		(0.011)	(0.012)	(0.01)	(0.012)	(0.013)	(0.013)
$Ethnic$	1	-0.028 **	-0.033 ***	-0.016 .	-0.015 .	-0.028 **	-0.018 .
		(0.009)	(0.009)	(0.008)	(0.009)	(0.01)	(0.01)
$Gender$	0	0.004	-0.008	-0.008	-0.002	-0.008	-0.002
		(0.006)	(0.006)	(0.005)	(0.007)	(0.008)	(0.008)
Wa	1	0.006	-0.004	0.009	0.011 .	-0.004	-0.001
		(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
$Educ$	2	-0.041 **	-0.035 *	-0.025 .	0.017	-0.027 .	-0.023
		(0.014)	(0.015)	(0.013)	(0.014)	(0.015)	(0.014)
	3	0.005	-0.004	0.003	-0.007	0.011	-0.001
		(0.006)	(0.007)	(0.006)	(0.007)	(0.008)	(0.008)
$Area$	Midlands Northern Mountains	0.004	0.02 *	0.032 ***	0.058 ***	0.04 **	0.065 ***
		(0.009)	(0.01)	(0.009)	(0.012)	(0.013)	(0.012)
	Northern Central Coast	-0.052 ***	-0.026 **	-0.038 ***	0.035 ***	0.043 ***	0.072 ***
		(0.008)	(0.009)	(0.008)	(0.01)	(0.011)	(0.01)
	Central Highlands	-0.035 **	-0.008	-0.01	0.019	0.015	0.031 *
		(0.012)	(0.013)	(0.011)	(0.014)	(0.015)	(0.015)
	South East	-0.04 ***	-0.023 *	-0.063 ***	-0.061 ***	-0.013	-0.007
		(0.01)	(0.01)	(0.009)	(0.011)	(0.012)	(0.012)
	Mekong River Delta	-0.022 **	0.025 **	-0.031 ***	0.049 ***	0.051 ***	0.072 ***
		(0.008)	(0.009)	(0.008)	(0.009)	(0.01)	(0.01)

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