# AVERTISSEMENT



# LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n°92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992
http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *9/7/2018* **par :**

**HUONG TRINH THI**

**Adapting recent statistical techniques to the study of nutrition in Vietnam**

---

**JURY**

| | | |
|---|---|---|
| Dominique HAUGHTON | Professeur des Universités | Président du Jury |
| Germà COENDERS | Professeur des Universités | Membre du Jury |
| Phuc HO DANG | Professeur Associé | Membre du Jury |
| Tien Zung NGUYEN | Professeur des Universités | Membre du Jury |
| Christine THOMAS–AGNAN | Professeur des Universités | Membre du Jury |
| Michel SIMIONI | Directeur de Recherche | Membre du Jury |

---

**École doctorale et spécialité :**
*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**
*Toulouse School of Economics (TSE-R)*

**Directeur(s) de Thèse :**
*Christine THOMAS-AGNAN* et *Michel SIMIONI*

**Rapporteurs :**
*Dominique HAUGHTON* et *Phuc HO DANG*

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *9/7/2018* **par :**

**HUONG TRINH THI**

**Adapting recent statistical techniques to the study of nutrition in Vietnam**

---

**JURY**

| | | |
|---|---|---|
| Dominique HAUGHTON | Professeur des Universités | Président du Jury |
| Germà COENDERS | Professeur des Universités | Membre du Jury |
| Phuc HO DANG | Professeur Associé | Membre du Jury |
| Tien Zung NGUYEN | Professeur des Universités | Membre du Jury |
| Christine THOMAS–AGNAN | Professeur des Universités | Membre du Jury |
| Michel SIMIONI | Directeur de Recherche | Membre du Jury |

---

**École doctorale et spécialité :**
*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**
*Toulouse School of Economics (TSE-R)*

**Directeur(s) de Thèse :**
*Christine THOMAS-AGNAN* et *Michel SIMIONI*

**Rapporteurs :**
*Dominique HAUGHTON* et *Phuc HO DANG*

# Acknowledgments

Many thanks to my beloved husband Chinh and my son Anh Khanh for their patience and their support; great thanks to my traveling companion Khanh Linh.

I would like to thank my advisors Christine THOMAS–AGNAN and Michel SIMIONI for teaching me statistics and econometrics, from some first puzzles to complete articles; to Thibault LAURENT for teaching me the R programming language. I thank Joanna Morais for our fullfruit collaboration on CoDa. I thank Cuong LE VAN and Tien Zung NGUYEN for recommending me.

I would like to thank Dominique HAUGHTON, Germà COENDERS, Phuc HO DANG, and Tien Zung NGUYEN to have accepted to be part of my thesis jury. Your valuable discussion will help improving not only for this thesis but also my future research. I thank the School of Preventive Medicine and Public Health (Hanoi) and International Center for Tropical Agriculture (CIAT) – Asia for a great time of visiting research.

I would like to thank French colleagues who always have encouraged my passion in Food economics me over the past four years: Zhora, Olivia, Katia, Olivier. I thank Corinne and Aline for nice discussions during lunches. I also thank Linh's friends at École Matabiau for introducing her French cultures.

I would thank to all my Vietnamese friends: Ngan and Mai, my nice roommates; the vietnamese mathematics group in Toulouse: Huyen for my first year in TSE, Hang, Trang, Nga, Dat for nice photos in all Europe; Mai and Chi for helping me overcome the pressures of thesis stress and family troubles, priests and nuns in the Vietnamese catholiques community in Toulouse. VCREMEers have brought courage and ambition to my thesis. I would like thank my colleagues at Thuongmai university for their supports.

I express my gratitude to the families of Christine, Tibo, Geoffroy and Lily for considering Linh and myself as members of their family.

Cuối cùng con cảm ơn bố mẹ nội ngoại luôn động viên và chăm sóc Bin. Cảm ơn anh chị em trong nhà đã thông cảm và đồng hành trong hành trình dài này.

**Financement :** This thesis is financed by The Vietnam Ministry of Education and Training (911 Program) and TSE–R Unite and the INRA-CIRAD GloFoodS meta-program (TAASE project).

# Contents

# List of Figures

# List of Tables

# Résumé

L'objectif de cette thèse est d'adapter des méthodes récentes de statistique pour apporter une vision nouvelle de la transition nutritionnelle au Vietnam. Le Vietnam a connu un fort développement économique. Mais le Vietnam fait face aujourd'hui au double fardeau de la malnutrition caractérisé par la coexistence de la malnutrition d'un côté et du surpoids et de l'obésité de l'autre, ou des maladies non transmissibles liées à l'alimentation.

Dans le chapitre 1, nous faisons une brève introduction. Nous considérons que le Vietnam est une étude pilote sur le problème de la nutrition. Nous rappelons les fondements des principales méthodes statistiques appliquées dans cette thèse et nous mettons l'accent sur notre contribution.

Dans le chapitre 2, nous revenons sur la question de l'estimation de la relation entre la prise de calories par personne et le revenu en utilisant six vagues de l'enquête Vietnam Household Living Standard Survey sur la période 2004-2014. Quantifier la réponse au revenu de la prise de calories pour les foyers les plus pauvres est un préalable pour définir des politiques publiques visant à réduire la famine et à corriger les déficiences nutritionnelles. Pour éviter la malédiction de la dimension des méthodes purement nonparamétriques due à la présence d'un grand nombre de variables explicatives, nous adoptons plutôt la famille des modèles généralisés additifs (GAM) dans lesquels seul le revenu intervient de façon non linéaire. Nous comparons ces modèles avec une procédure récente. Les résultats mettent en relief une réponse forte de la prise de calories à un accroissement du revenu pour les foyers les plus pauvres.

Dans le chapitre 3, nous utilisons des méthodes de décomposition pour évaluer les déterminants des changements de consommation de macronutriments au Vietnam en utilisant les vagues 2004 et 2014 de l'enquête VHLSS. L'objectif commun des méthodes de décomposition est de décomposer la différence entre deux groupes pour une variable économique telle le salaire ou le revenu en deux effets: un effet de composition dû aux différences des covariables observées entre les groupes, et un effet de structure dû aux différences entre les groupes dans la relation qui lie les covariables à la variable économique d'intérêt. La méthode de décomposition récente proposée par Rothe (2015), qui peut être appliquée à une moyenne, un quantile ou tout autre paramètre caractérisant la distribution de la variable d'intérêt,

a pour but de poursuivre la décomposition plus loin en décomposant l'effet de composition en trois composantes: (1) la contribution directe de chaque covariable, (2) plusieurs effets d'interaction d'ordre deux ou supérieur et (3) un effet de la dépendance. Rothe utilise des copules pour modéliser les effets de dépendance, technique qui est bien adaptée au cas des variables continues. Nous adaptons cette approche au cas d'un mélange de variables continues et discrètes.

Dans le chapitre 4, nous nous concentrons sur la composition de la diète en modélisant les proportions de protéines, de matières grasses et de glucides dans la prise moyenne de calories par personne. Parce que ce vecteur de proportions est de nature compositionnelle, nous nous tournons naturellement vers les méthodes d'analyse de données de composition. Nous utilisons des outils descriptifs, comme les biplots compositionnels et les diagrammes ternaires, pour montrer l'évolution des trois composantes au travers du temps et modélisons ensuite la consommation de macronutriments en fonction des caractéristiques des ménages avec des modèles de régression pour données de composition. Nous établissons la formule permettant le calcul des semi-elasticités de la consommation de macronutriments par rapport à la dépense totale de nourriture. Nous comparons ensuite les interprétations de ces semi-élasticités à celle des semi-élasticités des volumes de macronutriments consommés et de la commation totale calculées à partir des modèles classiques.

Dans le chapitre 5, nous nous penchons sur la relation entre les parts de macronutriments et l'indice de masse corporelle (IMC). Nous construisons un modèle de régression compositionnelle incluant un total pour expliquer les quantiles de l'indice de masse corporelle. Cette approche nous permet de résoudre le problème de facteurs de confusion entre les parts et le volume total de macronutriments. Npus calculons ensuite les élasticités de l'IMC par rapport à chaque macronutriment. Notre travail est basé sur l'utilisation de la base de données de l'enquête "General Nutrition Survey" et nous nous restreignons aux adultes vietnamiens entre 18 et 60 ans. Les résultats révèlent d'abord des effets significatifs de facteurs socio-économiques tels que la'age, le sexe, le type d'emploi, le fait de consommer de la bière et la région géographique. Toutes les élasticités de l'IMC par rapport à tous les macronutriments sont des fonctions croissantes de l'IMC jusuq'à un seuil (MIC=20) à partir duquel elles sont stables.

# Abstract

The objective of this thesis is to adapt recent statistical techniques and to bring new insights on the nutritional transition in Vietnam. Vietnam has experienced a strong economic development that turned this poor country in the 1980s into a lower middle income country currently. But Vietnam now faces the double burden of malnutrition characterized by the coexistence of undernutrition along with overweight and obesity, or diet-related noncommunicable diseases. To fight against malnutrition, the Vietnamese government has recently defined a comprehensive strategy to improve the nutritional status of the Vietnamese population.

Chapter 1 gives a brief introduction to this thesis. We consider Vietnam is a pilot case study about nutrition. We recall the main statistical techniques applied in this thesis and we emphasize our contributions.

In chapter 2, we revisit the issue of estimating the relationship between per capita calorie intake and income using six waves of the Vietnam Household Living Standard Survey over the period 2004-2014. Characterizing the response of calorie intake to income for the poorest households is a prerequisite for considering policies aimed at reducing starvation and correcting nutritional deficiencies. The classical log-log specification does not capture the nonlinearity of this relationship. To avoid the curse of dimensionality of fully nonparametric specifications due to the presence of many control variables (age, education, region ...) we adopt rather various generalized additive models (GAM) specifications where only income is supposed to act in a nonlinear fashion and compare them with a recent procedure. The results highlight the strong response of calorie intake to an increase in income for the poorest households. A byproduct of the proposed methodology is the decomposition of the evolution of average calorie intake between the two waves into the part due to the change of population characteristics distributions and those coming from the change in calorie-income relationship, shedding new light on the nutritional transition in Vietnam.

In Chapter 3, we use decomposition methods to assess the determinants of changes in macronutrients consumption in Vietnam using the 2004 and 2014 waves of VHLSS. The common objective of decomposition methods is to decompose between-group differences in economic outcomes such as wage or income, into two components: a *composition* effect due to differences in

observable covariates across groups, and a *structure* effect due to differences in the relationship that links the covariates to the considered outcome. The recent decomposition procedure proposed by Rothe (2015), which can be applied to mean, quantiles, or other parameters characterizing the distribution of the considered outcome, aims at decomposing further the composition effect into three types of components: (1) the *direct contribution* of each covariate due to between-group differences in their respective marginal distributions, (2) several *two way* and *higher order interaction effects* due to the interplay between two or more covariates and (3) a *dependence effect* accounting for different dependence patterns among the covariates. Rothe (2015) uses a parametric copula to model the dependence effects, which is well adapted for continuous covariates. We adapt this approach to the case of a mixture of continuous and discrete covariates.

In Chapter 4, we focus on food composition in terms of diet components. We consider modeling the proportions of protein, fat and carbohydrate ($D = 3$) in the average per capita calorie intake. Because this vector of proportions is of a compositional nature, we naturally turn attention the compositional data analysis techniques. We use descriptive tools, such as compositional biplots and ternary diagrams, to show the evolution of the three components over the years and then model macronutrients composition as a function of household characteristics, using compositional regression models. We derive the expression of the semi-elasticities of macronutrients shares with respect to food expenditure. We then compare the interpretations of these shares semi-elasticities to that of volumes of macronutrients and of total calorie intake obtained using classical linear models.

In Chapter 5, we focus on the relationship between macronutrient balances and body mass index. We develop a compositional regression model including a total at various quantile orders. This approach solves the problem of confounding effects between macronutrients total volume and shares in a diet (Willett et al., 1997). We then compute the elasticities of BMI with respect to each macronutrient and to the total consumption. Our empirical research is based on the General Nutrition Survey 2009-2010 and we restrict attention to Vietnamese adults from 18 to 60 years of age. The results first reveal significant impacts of some socio–economics factors, such as the overall consumption volume, the age, the gender, the job type, the "no drinking status" and the geographical region. All elasticities of BMI with respect to each macronutrient increase as BMI increases until a threshold (BMI=20) and then remain stable.

In chapter 6, we briefly give our perspectives of future research in both mathematics and nutrition.

# Publications and Fellowships

## Publications

1. Calorie intake and income in China: New evidence using semiparametric modelling with generalized additive models, *Vietnam Journal of Mathematical Applications*, **12** (2):a-b, 2016. (with Simioni, M., and Thomas-Agnan, C.)

2. *Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis*, forthcoming in the Statistical Methods in Medical Research journal, 2018. (with Morais J., Simioni, M., and Thomas-Agnan, C.)

3. *Assessing the nonlinearity of the calorie-income relationship: an estimation strategy - With new insights on nutritional transition in Vietnam*, revised and resubmitted to World Development, 2018. (with Simioni, M., and Thomas-Agnan C.)

4. *A New Perspective on the Relationship between Calorie Intake and Income in China and Vietnam using Semiparametric Modeling*, contributed presentation, EAAE Congress 2017, Parme, Italy. (with Simioni, M., and Thomas-Agnan, C.)

5. *Decomposition of changes in the consumption of macronutrients in Vietnam between 2004 and 2014*, TSE Working Paper, n. 18-910, April 2018. (with Simioni, M., and Thomas-Agnan, C.)

6. *Measuring the progress of the timeliness childhood immunization compliance in Vietnam between 2006-2014: A decomposition analysis*, TSE working paper, number 18-920, 2018 (with Do, T.T.T, Nguyen, V.H., Nguyen, Q.D., and Thomas-Agnan, C.)

7. *Macronutrient balances and Body mass index: A new insight using compositional data analysis with a total at various quantile regressions*, TSE working paper, 2018 (with Le D. T, Thomas–Agnan C., Beal T., Simioni, M., and Nguyen D. S.)

## Prizes

- Best researcher papers, on behalf of The Eleventh Vietnam Economist Annual Meeting (VEAM) chaired by Cuong Le Van.

## Fellowships

- July- August,2017. Institute for Preventive Medicine and Public Health, Hanoi Medical University, supported by Ecole doctorale Toulouse, FRANCE
- November 2017- June 2018. International Center for Tropical Agriculture (CIAT)–Asia, Viet Nam, supported by Toulouse School of Economics (TSE-R) and International Centre for Mathematics and Computer Science Toulouse (EMDMIT), FRANCE

# Chapter 1

# Introduction

The growth and rapid development of a lower middle-income country like Vietnam raise many questions for researchers in several disciplines including the social sciences. Applied mathematics offer many tools that can be used to provide answers to these questions. But there is often a lack of expertise in these tools, as for example in Vietnam. The thesis presented below aims to fill this gap by adapting recent methods in statistics to shed light on economic issues linked to the nutritional transition in Vietnam. This thesis has been supported by the Vietnamese Government in the context of decision 211 whose goal is "training lecturers of Doctor's Degree for universities and colleges for the 2010-2020 period" (Decision No. 911/QD–TTg dated 17/6/2010). The Vietnamese government made this decision in order to improve the knowledge and skills of Vietnamese researchers. In addition, this thesis is multidisciplinary in an attempt to build successful bridges from applied mathematics to empirical questions in social sciences. Its supervision was provided by a professor of statistics from the Decision Mathematics group of TSE-R and an economist from the Institut National de la Recherche Agronomique (INRA). Its achievement has benefited from the advice of nutritionists and epidemiologists in Vietnam, France and USA.

## 1.1 A pilot case study: Nutrition in Vietnam

Nutrition in Vietnam is a good pilot case to adapt recent mathematics statistical techniques. Vietnam is a good example of a middle-income country that has recorded impressive achievements in economy and population welfare after the launch of economic reforms in 1986. At the same time, this country has also experienced a nutrition transition like many other middle-income countries. The main stages of the nutrition transition have been described in Popkin (2006), see Figure 1.1. It is only very recently that the nutrition literature has begun to highlight the development of non communicable diseases in Vietnam (pattern 4 in Figure 1.1) (Nguyen and Hoang,

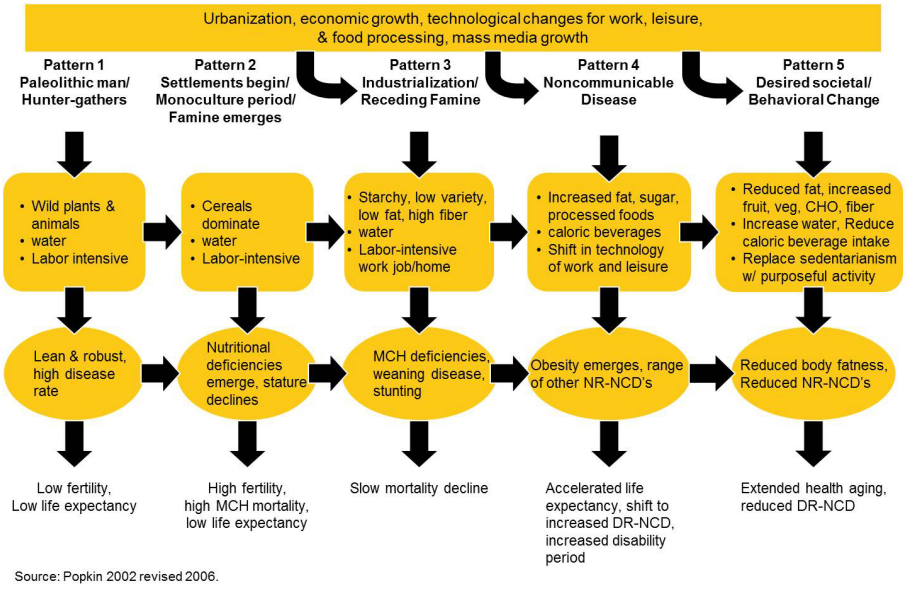
2018). However, Vietnam is facing the double burden of malnutrition, i.e. there is coexistence of undernutrition along with overweight and obesity, or diet-related noncommunicable diseases, within households and populations, and across the life course (Nguyen and Hoang, 2018). Vietnam seems to be at the crossroads of patterns 3 and 4 of the nutrition transition as summarized by Popkin (2006). The Vietnamese government has recently defined a comprehensive strategy to improve the nutritional situation of the Vietnamese population (Ministry of Health, 2012). In academic fields, there are several empirical researches focusing on the nutritional situation in Vietnam. There still is a need for immediate deep analysis about the important drivers of the nutrition transition to help Vietnamese policy-makers in implementing related policies. The complexity of the questions need multidisciplinary research.

To go into details, the structure of the diet during the 1990s in Vietnam contained less and less starchy staples and more and more proteins and lipids coming from meat, fish, and other protein-rich and higher fat food items (Nguyen and Popkin, 2004). In the 1992–1993 period, the main consumed food items by the Vietnamese people were cereals, potatoes, rice, and other starches, contributing up to 85.9% of total energy intake. In the 1997–1998 period, even though the total amount of calories consumed per capita remained at about the same level as 5 years earlier, there was a remarkable increase in daily proteins and lipids consumption while the consumption of rice and other starches reduced significantly. Recently, the National Institute of Nutrition (NIN) in Vietnam has defined the "ideal" diet balance for Vietnamese households: 14% of protein, 18% of fat and 68% of carbohydrate. NIN's goal is that 50% (resp. 75%) of Vietnamese households achieve this diet balance in 2015 (resp. 2020) (Ministry of Health, 2012). The diet balance belongs to the Vietnamese government comprehensive strategy to improve the nutritional situation of the Vietnamese population. These changes on food consumption in Vietnam correspond to pattern 3 of Nutrition transition as described by Popkin (2006).

## 1.2 Structure of the thesis

In this thesis, we apply and adapt recent statistical techniques to analyze the nutritional status of the Vietnamese population and its evolution over the 2004-2014 period. The overall structure of the thesis is summarized in Figure 1.2. For instance, Vietnam is still suffering from the undernutrition burden. In a global context, policies aimed at reducing starvation and redressing nutritional deficiencies remain among the most widely accepted policies in the world as emphasized by Banerjee (2016). Then, many different policies, such as subsidized prices of basic foodstuffs to cash transfers, attempt to cope with these nutritional deficiencies. Among them, household

Figure 1.1: Stages of the Nutrition Transition




income plays an important role and numerous papers in development and health economics deal with the issue of estimating the relationship between food demand measured in calories and household income. However, the calorie–income relationship shows controversial results (Ogundari and Abdulai, 2013). Chapter 2 revisits this issue of estimating the calorie-income relationship by proposing an estimation strategy based on recent developments on generalized additive models, model selection, and decomposition methods in economics.

To analyze in more detail the evolution of the nutritional status of individuals it is possible to focus on their diet, i.e. on the macronutrient composition of their consumption. For instance, the nutrition transition in Vietnam is characterized by an increase in per capita total calorie intake resulting from an increase in the consumption of fat and protein while the carbohydrate consumption decreases. Several factors can explain this evolution, such as the evolution of consumer preferences or of the characteristics of consumers (the population is more urbanized in 2004 than in 2014 ...). Chapter 3 proposes to highlight the socio-demographic drivers of this transition over the period 2004-2014, using a decomposition technique to evaluate the contribution of different factors in the observed evolutions, whether they are on average or for certain quantiles. The analysis focuses on the consumption of each macronutrient.

The previous chapter does not take into account that the three macro-

Figure 1.2: Stages of the Nutrition Transition in Vietnam in the context of this thesis

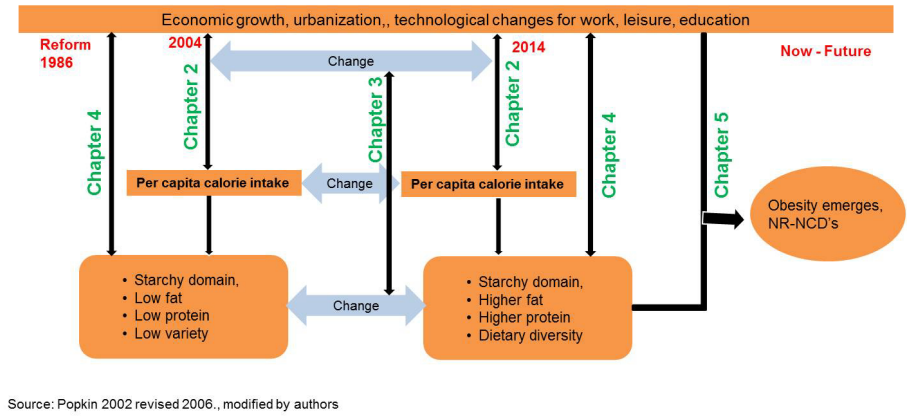
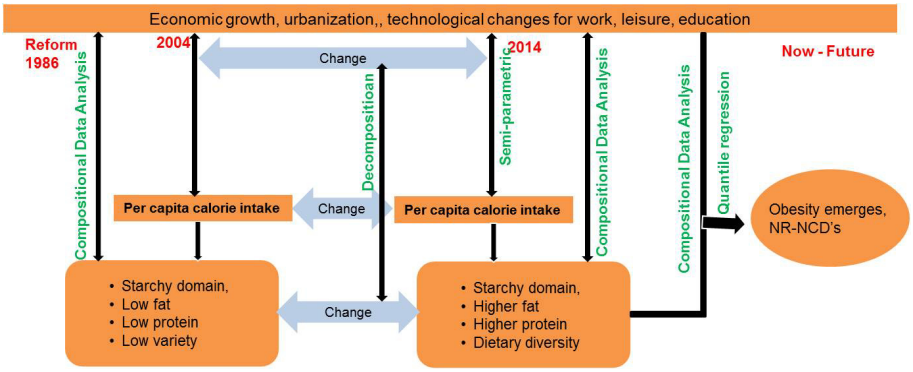



Figure 1.3: Adapting recent statistical techniques to the study of nutrition in Vietnam




nutrients constitute the whole diet of an individual so that the volumes of consumed macronutrients are not independent. For example, in the 1997–1998 period, even though the total amount of calories consumed per capita remained at about the same level as 5 years earlier, there was a remarkable increase in daily proteins and lipids consumption (4.7 points) while the consumption of rice and other starches reduced significantly (5.6 points). Moreover, the computation of consumed macronutrient volume can be criticized when using household survey data, as done in this thesis, due to the impossibility to take into account losses and wastes in food preservation, preparation and consumption. Household survey data have also limitations due to recall bias and self-reported measures (Deaton, 1997). Assuming that

these two problems affect the computation of the quantities of all macronutrients in the same way, we can expect the shares of the macronutrients not to be affected by the consecutive biases, contrary to volumes. In chapter 4, we use compositional data analysis (CODA), adapted to deal with the relative information contained in shares, to describe the evolution of diet patterns over time, and to model the impact of household characteristics on the macronutrient shares vector.

As pointed out above, Vietnam is entering in pattern 4 of the nutrition transition as defined by Popkin (2006). The Vietnamese diet patterns changed with a higher proportion of animal source, fat and protein intake (Nguyen and Popkin, 2004; Trinh et al., 2018). Among Vietnamese 18-65 years old the prevalence of overweight and obesity increased from 2.0% in 1992 to 5.2% in 2002 using a national survey (Tuan et al., 2008). Similarly, Nguyen and Hoang (2018) show that the prevalence of overweight and obesity increased from 2.3% in 1993 to 15% in 2015 in the same age group. These figures in urban sites are much higher than in rural sites. Prevalence of obesity of children under 5 has increased much faster than for adults. In the 2000-2010 period, the prevalence of overweight and obesity increased from 0.6% (resp. 0.9%, 0.5%) to 5.6% in the whole country (resp. in urban areas, in rural ones). In 2011, 14% of the children under 5 (resp. 8.6%, 4.4% ) in Vietnam were still stunted (resp. underweight, thin). In addition, both ratios for children under 5 are higher in big cities (Huynh et al., 2007). In chapter 5, we use compositional regression models with a total at various quantile orders to analyze the impact of macronutrient balances on body mass index. Here, the total variable is defined as the geometric mean of the consumption volumes.

## 1.3 Data

This thesis is firstly based on several cross-sectional data sets in Vietnam and China. First, we use the six most recent waves of the Vietnam Household Living Standard Survey, or VHLSS: 2004, 2006, 2008, 2010, 2012, and 2014. 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 of 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. The main objective of VHLSS is to collect data on Vietnamese household living standards and household members occupation, health and education status. This survey is not, by definition, constructed to assess the nutritional status of Vietnamese households. Household living

standard survey (or a similar survey, household consumption and expenditure surveys, or HCES) are often used in nutrition studies (Zezza et al., 2017). 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 macronutrient based on the calorie conversion table constructed by Vietnam National Institute of Nutrition in 2007.

Second, we use the General Nutrition Survey, or GNS, which is conducted by the Vietnam National Institute of Nutrition every 10 years. More specifically, we use the 2009-2010 wave. The objectives of this wave were to evaluate the effectiveness of the policy defined by the national strategy for the period 2001-2010, and the construction of the goals of the same strategy for the next ten-year period. As VHLSS, GNS is a nationally representative survey of all provinces in Vietnam. GNS gives individual anthropometric measurements, individual dietary intakes... GNS also allows to measure risk factors related to nutrition issues in communities such as malnourished children, and to evaluate hygiene and safety food hygiene. The 24-hour recall survey is used to collect food intakes measured in quantities at household level. Then these food intakes are transformed into macronutrients using the calorie conversion table constructed by Vietnam National Institute of Nutrition in 2007.

In related work (Trinh et al., 2018b), we use the 2006 and 2014 waves of the Multiple Indicator Cluster Survey, or MICS. MICS is conducted by UNICEF Vietnam with official support from the Vietnam General Statistics Office to capture Vietnamese achievement of Millennium Development Goals (MDGs). The purpose of this survey is to provide worthy information on Vietnamese children and Vietnamese women. This information includes child mortality, nutrition (especially breastfeeding), immunization, water and sanitation... We mainly focus on immunization to measure the progress of the timeliness childhood immunization compliance in Vietnam between 2006-2014.

Similarly in (Trinh et al., 2016), we use the China Health and Nutrition Survey, or CNHS, which is conducted by the Carolina Population Center at the University of Chapel Hill and the National Institute of Nutrition and Food Safety at the Chinese Center for Disease Control and Prevention since 1989. We use the most recent four waves of CNHS conducted in 2004, 2006, 2009, and 2011. We use the Nutrition Survey, an integral part of CHNS, which contains rich and precise information about diet, both at the household level and at individual level, and on the socio-demographic characteristics of households and individuals.

## 1.4 Statistical techniques

### 1.4.1 Generalized Additive Models (GAM)

The Generalized Additive Models (GAM) specifications were proposed by Hastie and Tibshirani (1987). We have decided to choose this technique to capture nonlinearities in the relationship between calorie intake and income. This semi–parametric regression method does not suffer from the curse of dimensionality of fully nonparametric specifications. The choice of generalized additive models (GAM) presents both advantages of allowing the presence of many control variables (age, education, region ...) as well as the non-linearity effect of income. In our case, the GAM model (1.1) is defined by two conditions as follows. The first one specifies the relationship between the expected outcome and the explanatory variables:

$$g(\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j)) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j. \tag{1.1}$$

Here the income variable is acting non–linearly through the link function $g(.)$. The second condition is about the conditional distribution of per capita calorie intake given income and various control variables. Several different assumptions can be made such as the Normal, Poisson and Gamma distributions. In addition, in all specifications, we estimate the unknown smooth function $s(.)$ using thin plate regression splines which do not require knot selection, are computationally efficient and allow for testing of the linearity assumption (Wood, 2003). Model (1.1) encompasses various regression models found in the literature review on calorie-income relationship.

We then address the problem of choice among competing specifications of the calorie-income relationship. Thus, we implement the data-driven test recently proposed by Racine and Parmeter (2014) called the "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 in out-of-sample prediction with the evaluation data. In addition, we apply a test of exogeneity (Blundell and Horowitz, 2007; Blundell et al., 2012) to solve the problem of a possible inverse impact of calorie intake on income. This test avoids using nonparametric instrumental variables for estimating the function of interest and is also likely to have better power properties.

This semi-parametric modeling approach is applied in chapter 2 of this thesis. This approach is also applied to the study of the calorie-income relationship in China (Trinh et al., 2016). In addition, we also use GAM modeling to compare this relationship between Vietnam and China (Trinh et al., 2018).

### 1.4.2 Decomposition methods

Decomposition methods were introduced by Oaxaca (1973) and Blinder (1973) in the study of inequality in economics. Recently, these techniques have been summarized in Fortin et al. (2011). Our objective of interest is a distribution feature, denoted by $\nu(F)$, where $\nu(.)$ is a function from the space of all one-dimensional distribution functions to the real line. Distribution features include mean, variance, quantiles... The objective is to compare this feature in two different situations (two dates or more generally two groups). Suppose, for ease of presentation, that we have observed two covariates for each individual in the sample of a given group: $X^0 = (X_1^0, X_2^0)$ for group 0 and $X^1 = (X_1^1, X_2^1)$ for group 1. We will denote their joint CDFs respectively by $F_X^0$ and $F_X^1$. The decomposition method aims at understanding how the observed difference between the distribution feature $\Delta_Y^\nu = \nu(F_Y^0) - \nu(F_Y^1)$ is related to differences between the distributions $F_X^0$ and $F_X^1$. By defining the counterfactual outcome distribution $F_Y^{0|1}$ that combines the conditional distribution in group 0 with the distribution of covariates in group 1, we can decompose the observed between-group difference in two parts

- The *structure effect* $\Delta_S^\nu$, solely due to differences between the two groups in the conditional distribution of the outcome given the values of the covariates,
- the *composition effect* $\Delta_X^\nu$, solely due to differences in the distribution of the covariates between the two groups.

We use three different approaches of decomposition methods in this thesis

1. First, we adapt the traditional decomposition method (Blinder, 1973; Machado and Mata, 2005; Fortin et al., 2011) to the case of GAM regression to decompose the evolution of the distribution of calorie intake between two survey waves in two parts: the "composition effect" and the "structure effect". Inspired by Machado and Mata (2005), we apply this decomposition at many different quantile levels.
2. Second, the above decomposition method cannot go further in decomposing the impact of each driver of the composition effect. Therefore we follow a more complex approach of Rothe (2015) using copula theory. This approach avoids the problem of curse of dimensionality (DiNardo et al., 1996; Leibbrandt et al., 2010), using a nonparametric approach for estimating the conditional distribution of the outcome given a possibly large number of covariates. The most important contribution of Rothe (2015) is an explicit decomposition of the composition effect in terms of the respective marginal covariate distributions typically containing "interaction terms" resulting from the interplay

of two or more covariates, and also "dependence terms" resulting from between-group difference in the dependence pattern among the covariates. Rothe (2015)'s practical implementation includes three important estimation steps:

- Univariate CDFs $\widehat{F}^t_{X_j}(x_j)$, conditional CDF of $Y^t|X^t$.
- Gaussian copula for the dependence part

$$C_\Sigma(u) = \Phi^d_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)).$$

- The estimated joint c.d.f. of the explanatory variables to construct any counterfactual experiment.

This decomposition method is applied to the evolution of per capita calorie intake of carbohydrate, fat and proteins between the two years 2004 and 2014. We decompose the change of macronutrients consumption for two measures of location: mean and median, and for the two quantiles at 10% and 90% allowing at the same time to construct a measure of dispersion.

3. Third, in Trinh et al. (2018a), we apply an extension of the Blinder-Oaxaca decomposition technique to the logistic regression framework, proposed by Fairlie (2005). We focus on the progress of the timeliness childhood immunization compliance among children aged 0-5 years in Vietnam from 2006 to 2014 and analyze the socio-economic factors that account for the changes of the compliance rate during this period. This study is based on the Multiple Indicator Cluster Survey in 2006 and 2014. Detail of this application is the sixth publication of this thesis: Trinh et al. (2018a).

### 1.4.3 Compositional data analysis (CoDa)

Compositional data analysis (CoDa) has been proposed by Aitchison (1986), Pawlowsky-Glahn and Buccianti (2011) and Pawlowsky-Glahn et al. (2015). A composition is a vector of $D$ components for which the relative information is relevant (for example a vector of $D$ shares). In order to take into account the relative information between components and to ensure the constant sum of the fitted components (equal to 1 in our case), classical regression models cannot be used directly. Thus, shares are transformed, using an isometric log-ratio (ILR) transformation into $D-1$ orthonormal coordinates which can be represented in the classical Euclidean space so that linear regression models, estimated by ordinary least squares, can be used separately on the $D-1$ coordinates (Egozcue and Pawlowsky-Glahn, 2003). An example of composition vector in our study is the vector of proportions of protein, fat and carbohydrate (in Kcal) in Vietnamese diets, i.e. $D = 3$. In addition,

we also consider macronutrient components in term of grams which include protein, fat, carbohydrate and fiber in Vietnamese diets. We first use descriptive tools of CODA, such as compositional biplots, coda-Dendrogram and ternary diagrams, to show the evolution of the components over the years.

In this thesis, we also use two kinds of compositional regression models. Due to the complexity in the interpretation, we adapt the elasticities computation, proposed by Morais et al. (2018), to these two compositional models to have direct and easy interpretations of parameters in ILR regression. The two models we consider are the following

1. First, we study the impact of socio-economics factors on the evolution of macronutrient diet in Vietnam. We analyze the shift in protein, fat and carbohydrate shares, i.e we have a composition as outcome variable. More precisely, we consider model (1.2)

$$\begin{aligned}\mathbf{S}_i &= \mathbf{a} \bigoplus_{k=1}^{K} X_{ki} \odot \mathbf{b_k} \oplus \boldsymbol{\epsilon}_i \\ &= \mathbf{a} \oplus \log(Exp)_i \odot \mathbf{b_1} \oplus Urban_i \odot \mathbf{b_2} \oplus HSize_i \\ &\odot \mathbf{b_3} \oplus Educ_i \odot \mathbf{b_4} \oplus Ethnic_i \odot \mathbf{b_5} \\ &\oplus Gender_i \odot \mathbf{b_6} \oplus Area_i \odot \mathbf{b_7} \oplus \boldsymbol{\epsilon}_i,\end{aligned} \tag{1.2}$$

   where $S_i$ is the vector of shares for household $i$, $Exp_i$ is total food expenditure, $Urban_i$ is equal to 1 when the household lives in an urban area, 0 if not, $HSize_i$ is household size, $Educ_i$ is the head of household number of years of education, $Ethnic_i$ is equal to 1 when the household belongs to the main ethnic group in Vietnam, i.e. Kinh, 0 if not, $Gender_i$ is head of household gender, $Area_i$ is the province where household lives, and $\epsilon_i$ are i.i.d error variables.

2. Second, we study the impact of diet on noncommunicable diseases (NCDs), particularly obesity and underweight. We thus apply a compositional model with shares as explanatory variables, and body mass index, or BMI, as response variable. In this study, the total food consumption in terms of calories is also important, thus we also include a total as a geometric mean of the shares among the explanatory variables as in Coenders et al. (2017), which leads us to the following model

$$\mathbb{E}(Y_i) = \alpha + \beta Cilr1 + \gamma Cilr2 + T_i.\delta + a.Z_i \qquad (EF)$$

   where $\mathbb{E}(Y_i)$ denotes the expectation of the conditional distribution of $Y_i$ given the covariates.

   In addition, because we want to study obesity and underweight together, we adapt these CODA regression models to quantile regression, in which case the model becomes

$$\mathbb{Q}_\tau(Y_i) = \alpha + \beta Cilr1 + \gamma Cilr2 + T_i.\delta + a.Z_i \qquad (QF)$$

where $\mathbb{Q}_\tau(Y_i)$ denotes the quantile of order $\tau$ of the conditional distribution of $Y_i$ given the covariates.

## 1.5 Contribution

The purpose of this thesis is to adapt recent methods in statistics to shed light on economic issues linked to the nutritional transition in Vietnam. To obtain our goals, we are going to develop mathematical tools, modify recent advanced statistical tools and find critical empirical questions in the Vietnamese nutrition context. We contribute to the literature in both mathematics and social sciences.

- We draw a strategy to choose among various semi-parametric and parametric regression models.
- We develop formulas for computing the elasticities (or semi-elasticities depending on the situation) for various compositional regression models using the same techniques as Morais et al. (2018). These authors consider the case of a CODA regression model where both the dependent and the explanatory variables are of compositional nature with the same dimension. We first adapt the formulas to the case of a compositional explanatory variable and an ordinary dependent variable. Then we deal with the case of a compositional dependent variable and ordinary explanatory variables. We also show that these techniques can be extended to quantile regression.
- We propose a compositional regression model with a total at various quantile orders. Following, we adapt semi-elasticities for a compositional quantile regression.
- By working with various datasets and recent mathematical tools, we have drawn a full picture of the nutrition transition in Vietnam with two important questions: is there any evidence of a nutrition transition in Vietnam now ? What are the main drivers of the ongoing nutrition transition in Viet Nam ?
- We analyze the relationship between calorie intake and income in Vietnam and China. Then, we also do a comparison between the two countries.
- We are the first researchers who apply various decomposition approaches for studying medical questions such as food consumption, nutrition and immunization.
- In terms of macronutrients, we are the first people to use compositional data analysis in macronutrients consumption including both relative

information of macronutrient as volume of each component and total volume.

- We use many different kinds of figures to visualize the nutrition transition and the impact of nutritional policies.

# Chapter 2

# Assessing the nonlinearity of the calorie-income relationship: an estimation strategy

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.

This chapter has been written for the 2nd revision to the *World Development* and resubmitted. Below, we reshape the content of the paper, mainly methodology and appendices.

## 2.1 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 Shabnam (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 2.2 gives a picture of the nutritional situation of the Vietnamese population. Section 2.3 presents the methodology used in this paper. Section 2.4 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 2.5. Special attention is devoted to the potential endogeneity of the measure of income we have chosen. Section 2.6 concludes.

## 2.2 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 (Huynh 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 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).

## 2.3 Methodology

### 2.3.1 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, or DLM, i.e.

$$\log(\text{PCCI}) = \alpha_0 + \alpha_1 \log(\text{INCOME}) + \alpha_2 (\log(\text{INCOME}))^2 + \sum_{j=1}^{J} \beta_j x_j + \varepsilon \qquad (2.1)$$

where PCCI denotes per capita calorie intake, INCOME is total household income (sometimes replaced by total expenditure), and the $x_j$s 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, $\alpha_0$, $\alpha_1$, $\alpha_2$, and the $\beta_j$, 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.

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 $\beta$ 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 $\mu_i$ and variance $\sigma^2$, 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

$$f(y;\theta,\phi) = \exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)] \tag{2.2}$$

where $b(.)$, $a(.)$ and $c(.)$ are arbitrary functions, and, for practical modelling, $a(\phi)$ is usually set to $\phi$. $\theta$, called the "canonical parameter" of the distribution, depends on the linear predictor. and $\phi$ is the dispersion parameter. Equation (2.2) 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

$$\eta_i = x_i\beta + \sum_j s_j(z_{ji}) \tag{2.3}$$

where $\beta$ 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

$$\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j \tag{2.4}$$

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

$$g(\mathbb{E}(\text{PCCI}|\text{INCOME}, x_j)) = \alpha_0 + s(\text{INCOME}) + \sum_j \beta_j x_j. \tag{2.5}$$

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.

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

$$g(\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J)) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \tag{2.6}$$

where (i) $g(.)$ is a link function, (ii) the variables entering with a linear effect, the $x_j$s, 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. 2.1, 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

$$\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \tag{2.7}$$

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

$$\log(\mathbb{E}(PCCI|\text{INCOME}, x_1, \ldots, x_J)) = \alpha_0 + s(INCOME) + \sum_{j=1}^{J} \beta_j x_j, \tag{2.8}$$

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 A.

### 2.3.2 Revealed Performance test

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.

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. $(y_i^*, x_i^*)_{i=n_1+1}^n$.[1]

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

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

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

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

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

### 2.3.3 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 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

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(PCCI) - \mathbb{E}_{t_0}(PCCI) \tag{2.9}$$

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

$$\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}\big(\mathbb{E}\big(PCCI|INCOME, Z\big)\big) - \mathbb{E}_{t_0}(\mathbb{E}\big(PCCI|INCOME, Z\big)\big) \tag{2.10}$$

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 (2.10) becomes

$$\begin{gathered}\Delta PCCI_{t_0 \to t_1} = \\ \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z))\end{gathered} \tag{2.11}$$

Finally we can write the difference as

$$\begin{aligned}&\Delta PCCI_{t_0 \to t_1} = \mathbb{E}_{t_1}(m_{t_1}(INCOME, Z)) - \mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) + \\ &\mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z))\end{aligned} \tag{2.12}$$

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 $t_1$.

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

Decomposition (2.12) 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 (2.12), 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 $\mathbb{E}_{t_1}(m_{t_0}(INCOME, Z)) - \mathbb{E}_{t_0}(m_{t_0}(INCOME, Z))$, measures the "composition" effect and refers to the effect of the change in the distribution of the characteristics of households.

The different terms of the decomposition (2.12) can be estimated by taking empirical counterparts of the expectations, i.e. average values of the predicted values of $PCCI$ from the different models using either the contemporaneous or the counterfactual observations. Confidence intervals can then be calculated by adapting the bootstrap procedure proposed by Machado and Mata (2005).

## 2.4 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.[3] 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[4]. 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

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

[4] 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).

conversion table constructed by Vietnam National Institute of Nutrition in 2007 (see Table 6.2). Per capita calorie intakes are then computed as adult equivalent calorie intakes following recent papers of Aguiar and Hurst (2013) and Santaeulàlia-Llopis and Zheng (2017). Details on these computations are given in the Appendix C.

Following many papers in the literature on calorie-income relationship, we measure household resources by total expenditure rather than by income.[5] 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 2.1 summarizes the main characteristics of all the variables.

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

Table 2.1: 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 | $\leq$ 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 | $\geq$ 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 2.2: 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: $^{\star}$, $^{\star\star}$, and $^{\star\star\star}$ mean significant at 10%, 5%, and 1%, respectively

## 2.5 Results

### 2.5.1 Preferred models

Table 2.2 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 the revealed performance test, 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: $-10.20$ 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.

The same reading grid can then be applied to the other results reported in Table 2.2 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.

### 2.5.2 The estimated calorie-income relationships

Figure 2.1 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.).[6] The nonlinearity of the relationship clearly appears in view of the different curves traced in Figure 2.1. This result is confirmed by the various significance and linearity tests presented in Appendix A. 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).

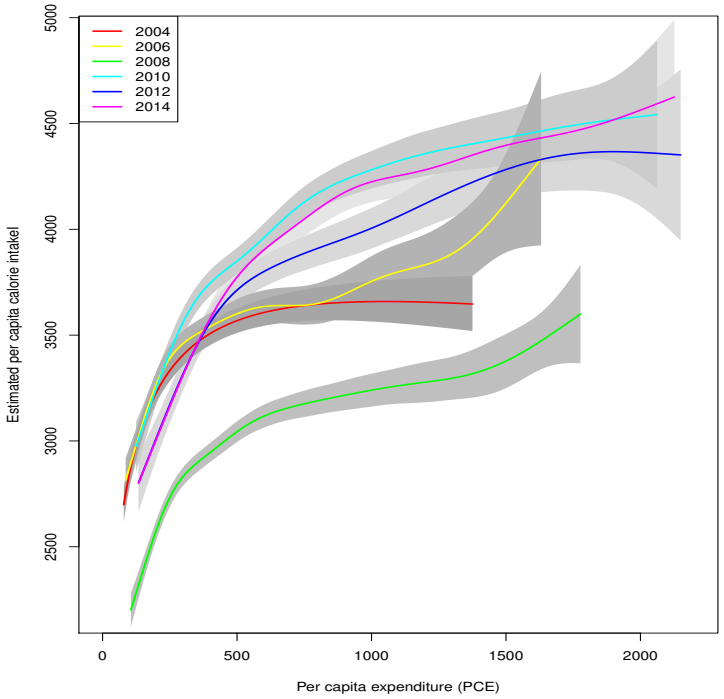
These results contribute to the debate on the extent to which calorie consumption responds to income changes in middle-income countries. They clearly show that income mediated policies can have an impact on nutritional goals up to a given threshold of income, or per capita expenditure, in Vietnam. They 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)[7] such income mediated policies may prove to be ineffective as 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 alar-

[6]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.

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

Figure 2.1: Estimated calorie-income relationships for Vietnam



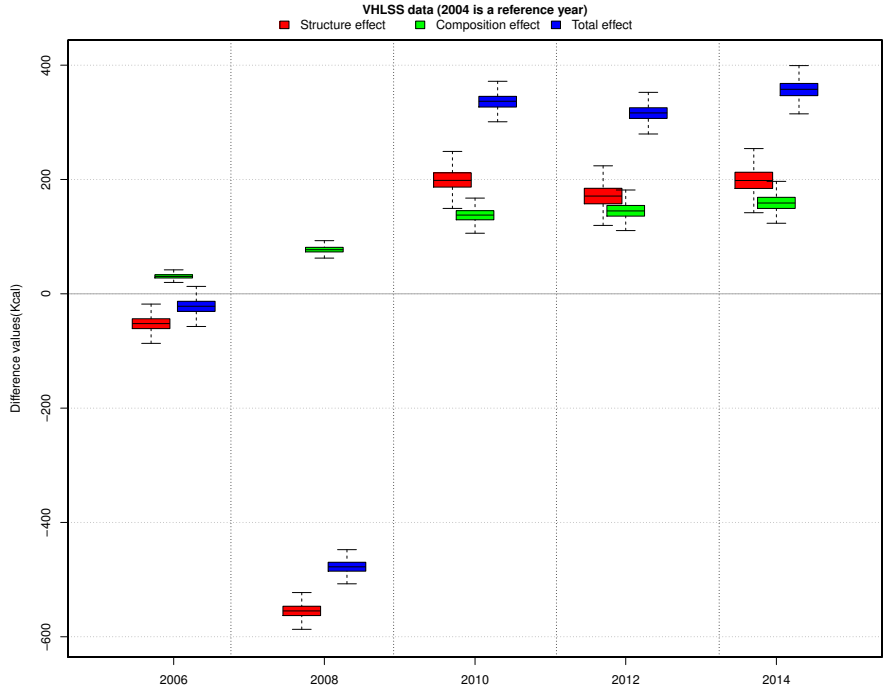

ming rates in 2008: the yearly increase of consumer price index reached 28% in September 2008 and even 65% for staple food products (rice and grains). This deterioration in macroeconomic conditions has had an impact on many Vietnamese households who have reduced their food expenditure. For instance, 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.

### 2.5.3 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 varia-

bles and what results from the change in calorie-income relationship. Thus, Figure 2.2 reports the results of the decomposition described in Eq. (2.12). 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 2.2: 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.

### 2.5.4 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 power 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 D.

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 2.3: 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 2.3. 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).

## 2.6 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.

# Chapter 3

# Decomposition of changes in the consumption of macronutrients in Vietnam between 2004 and 2014

Vietnam is undergoing a nutritional transition like many middle-income countries. This paper proposes to highlight the socio-demographic drivers of this transition over the period 2004-2014. We implement a method of decomposition of between-year differences in economic outcomes recently proposed in the literature. This method allows decomposing the composition effect on the distribution of the outcome under study, which is due to the differences in covariates across years, into direct contributions of each covariate and effects of their interactions. This method is applied to VHLSS data. The results show the importance of between-year changes in the distributions of covariates on between-year changes in the distributions of total calorie intake and calorie intakes from proteins and fat. This effect is more contrasted in case of calorie intake from carbohydrates. Food expenditure and household size appear to be the main drivers of the observed evolutions in macronutrients consumption. On the contrary, the urbanization of the population has a negative effect on these evolutions, except on fat consumption. The effect of urbanization is, nevertheless, less important than the positive effects of the previous two variables.

This chapter was recently submitted to *Economics and Human Biology*.

## 3.1 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. Using two nationally representative surveys, Ha et al. (2011) show that the nationwide prevalence of overweight (body mass index $\geq$ 25kg/m$^2$) and obesity (body mass index $\geq$ 30kg/m$^2$) was 6.6% and 0.4% respectively in 2005, almost twice the rates of 2000 (3.5% and 0.2%). Using the Asian body mass index cut-off of 23kg/m$^2$ the overweight prevalence was 16.3% in 2005 and 11.7% in 2000. According to the World Health Organization, the percentage of overweight people in the total population of Vietnam is 21% in 2014, the percentage of obese people being 4%. On the other hand, Ha et al. (2011) point out that the underweight prevalence (body mass index $<$ 18.5kg/m$^2$) of 20.9% in 2005 is lower than the rate of 25.0% in 2000. This rate has decreased by half in ten years and is currently 11%. Ha et al. (2011) also analyze the possible sources of this evolution and note that women were more likely to be both underweight and overweight compared to men in both 2000 and 2005. Urban residents were more likely to be overweight and less likely to be underweight compared to rural residents in both years. The shifts from underweight to overweight were clearer among the higher levels of food expenditure.

Many studies have been devoted to the evolution of food consumption in both developed and developing countries. Some of them aim to document how the evolution of the socioeconomic status of country's inhabitants has influenced their diets ( Nguyen and Popkin (2004), Burggraf et al. (2015)). Recently, Mayen et al. (2014) reviewed 33 studies on this issue. These studies show that (1) high socioeconomic status or living in urban areas is associated with higher intakes of calories, protein, total fat, cholesterol, polyunsaturated, saturated, and mono-unsaturated fatty acids, iron, and vitamins A and C and with lower intakes of carbohydrates and fiber, and (2) high socioeconomic status is also associated with higher fruit and/or vegetable consumption, diet quality, and diversity. The improvement of the socio-economic status of populations thus leads to a better feeding of human beings. But the other side of the coin is the link between improved diets and noncommunicable disease as emphasized by Popkin (2006) and Riera-Crichton and Tefft (2014). Thus, both policy makers and citizens are concerned by these concomitant evolutions and the fight against their

consequences in terms of malnutrition or over-food consumption. All this requires first of all knowledge of the drivers of these evolutions.

In this paper we document shifts in consumption of macronutrients in Vietnam over the period 2004 to 2014. Thanks to data from Vietnamese households living standard survey, we can calculate total calorie intakes of Vietnamese households, convert them into an adult equivalent, or per capita, calorie intakes (thus allowing comparison between households), and their decomposition into the three macronutrients : proteins, fat and carbohydrates (Thi et al., 2018). This survey also contains detailed information on the socio-demographic characteristics of Vietnamese households. Each wave of this survey is, moreover, representative of the Vietnamese population. This survey can therefore be used for a comparison of the nutritional status of the Vietnamese population between two waves.

We propose the use of decomposition methods to assess the determinants of change in macronutrients consumption in Vietnam using the 2004 and 2014 waves of VHLSS. Decomposition methods were first introduced in order to quantify the contributions of labor, capital, and unexplained factors (productivity) to economic growth (Solow, 1957). They have been extensively used in labor economics, following the seminal papers of Oaxaca (1973) and Blinder (1973). Fortin et al. (2011) provide a comprehensive overview of decomposition methods that have been developed since then. This method is recently wide used in the health sector, among them: Nie et al. (2018), (Anderson, 2018). The common objective of decomposition methods is to decompose between-group differences in economic outcomes such as wage or income, into two components: a *composition* effect due to differences in observable covariates across groups, and a *structure* effect due to differences in the relationship that links the covariates to the considered outcome. Applications to Vietnamese economy include Nguyen et al. (2007) on urban-rural income inequality, Sakellariou and Fang (2014) on wage inequality and the role of the minimum wage, and, very recently, Benjamin et al. (2017) on income inequality. To our knowledge, there is no work using decomposition methods to study the evolution of the nutritional diet and its socio-demographic determinants for Vietnam.

The Oaxaca-Blinder decomposition method has been refined in a large number of methodological papers and extended to the cases of distributional parameters besides the mean over the last four decades. Among all these methodological developments, we use the decomposition procedure recently proposed by Rothe (2015) which can be applied to mean, quantiles, or other parameters characterizing the distribution of the considered outcome (in our application, per capita calorie intake or calorie intakes coming from the three macronutrients). This decomposition method expands classical methods by adding to the usual decomposition of the composition effect into the *direct contribution* of each covariate due to between-group differences in their respective marginal distributions, and several *two way* and *higher*

*order interaction effects* due to the interplay between two or more covariates, a third effect, or *dependence effect*, accounting for the between-group difference in the dependence pattern among the covariates. To get a better understanding of the goals of the decomposition method we use, we will illustrate it by a simple example. Here, we analyze the difference in calorie intake distributions for two years, 2004 and 2014. Our outcome is measured by per capita calorie intake. We are interested in two potential drivers of the difference in per capita calorie intake distributions in 2004 and 2014: (1) evolution of Vietnamese households' food expenditures, and (2) urbanization. For instance, Vietnamese households increased their food spending between 2004 and 2014 and Vietnamese population is less urban in 2004 than in 2014. Moreover urban citizens tend to spend more on food (dependence between these two explanatory) hence leading to an extra increase in overall food expenditures. We are interested by decomposing the difference between per capital calorie intake averages in 2014 and 2004. The *structure* effect is the part of this difference that can be explained by the between-year difference in the conditional distributions of per capita calorie intake given food expenditures and location in an urban area. The *composition* effect is the part of the difference that can be explained by the between-year differences in observable characteristics (food expenditures and living in an urban area). The first *direct* contribution is the part of the composition effect that can be attributed to the fact that Vietnamese households have higher food expenditures in 2014 compared to 2004. The second *direct* effect captures the part in the composition effect due to the fact that Vietnamese population is more urban in 2014 than in 2004. The (only) *interaction* effect measures the additional contribution of the fact that Vietnamese population at the same time spends more for food and is more urban in 2014. Finally, the *dependence* effect accounts for between-year difference in association patterns among the two covariates, food expenditures and location in an urban area. In other words, the *dependence* effect captures the fact that the relative food expenditure of urban and rural households differs in the two years.

The remainder of the paper is structured as follows. Section 3.2 describes the decomposition method based on copulas and its practical implementation. Section 3.3 gives a description of the data we use in this study. Results are presented and commented in section 3.4. Section 3.5 concludes.

## 3.2 Decomposition method

### 3.2.1 Decomposing the decomposition effect

This section introduces through an example the methodology subsequently used, and draws heavily on Rothe (2015).

In the remainder of this article, we will focus on the evolution of certain characteristics of the distribution of the quantities of macronutrients consu-

med in Vietnam: average values and quantiles, between 2004 and 2014. Let us concentrate, below, on the number of calories obtained from the consumption of carbohydrates per day and per individual. The same reasoning will apply to the number of calories obtained from the consumption of protein or fat. For any household $i$ in year 2004 and any household $h$ in year 2014, we observe an outcome variable: the per capita and per day amount of calories obtained from the consumption of carbohydrates, denoted by $Y_i^{2004}$ and $Y_h^{2014}$, respectively. These observations are the realizations of two random variables, denoted by $Y^{2004}$ and $Y^{2014}$, whose marginal cumulative distribution functions, or CDFs, are $F_Y^{2004}$ and $F_Y^{2014}$, respectively. Our object of interest is a distribution feature, denoted by $\nu(F)$, where $\nu(.)$ is a function from the space of all one-dimensional distribution functions to the real line. The main features we are interested in include the mean, i.e. $\nu : F \to \int y dF(y)$, and the $\alpha$–quantiles, i.e. $\nu : F \to F^{-1}(\alpha) = \inf\{t : F(t) \geq \alpha\}$ for a given value of $\alpha \in [0, 1]$.

Suppose, for ease of presentation, that we have observed two covariates for each individual in the sample of a given year: for example, food expenditures and location in either urban or rural areas. Of course, the presentation given below can be easily generalized to more than two covariates. We denote the vectors of the two covariates by $X^{2004} = (X_1^{2004}, X_2^{2004})$ and $X^{2014} = (X_1^{2014}, X_2^{2014})$, and their joint CDFs by $F_X^{2004}$ and $F_X^{2014}$, respectively. The decomposition method aims at understanding how the observed difference between the distribution feature $\nu(F_Y^{2014})$ and $\nu(F_Y^{2004})$, i.e.

$$\Delta_Y^\nu = \nu(F_Y^{2014}) - \nu(F_Y^{2004}) \tag{3.1}$$

is related to differences between the distributions $F_X^{2004}$ and $F_X^{2014}$. For this, we can define the counterfactual outcome distribution $F_Y^{2004|2014}$ that combines the conditional distribution in year 2004 with the distribution of covariates in year 2014, as

$$F_Y^{2004|2014}(y) = \int F_{Y|X}^{2004}(y, x)\, dF_X^{2014}(x) \tag{3.2}$$

where $F_{Y|X}^{2004}(y, x)$ denotes the conditional distribution of outcome given values of the covariates in year 2004. In our example, we can interpret $F_Y^{2004|2014}(y)$ as the distribution of per day and per capita carbohydrates consumption after a counterfactual experiment in which the joint distribution of the two covariates is changed from year 2004 to year 2014, but the conditional distribution of per day and per capita carbohydrates consumption given these characteristics remains that of 2004. One can then decompose the observed between-year difference $\Delta_Y^\nu$ into

$$\begin{aligned} \Delta_Y^\nu &= \left(\nu\left(F_Y^{2014}\right) - \nu\left(F_Y^{2004|2014}\right)\right) + \left(\nu\left(F_Y^{2004|2014}\right) - \nu\left(F_Y^{2004}\right)\right) \\ &= \Delta_S^\nu + \Delta_X^\nu \end{aligned} \tag{3.3}$$

where $\Delta_S^\nu$ is a *structure effect*, solely due to differences in the conditional distribution of the outcome given values of covariates between the two years, and $\Delta_X^\nu$ is a *composition effect*, solely due to differences in the distribution of the covariates between the two years.

The different elements of the decomposition (3.3) can be easily estimated using nonparametric estimates of CDFs. One such strategy, focusing on densities instead of CDFs, is applied in DiNardo et al. (1996) or Leibbrandt et al. (2010). But the application of such a strategy soon encounters the problem of the curse of dimensionality. For a fixed sample size, the precision of the nonparametric estimators deteriorates very rapidly when the number of covariates increases, even if these estimators are free from any specification error (Silverman, 1986). In addition, it is also interesting to break down the composition effect for the different covariates. This can be easily done using the Oaxaca (1973) and Blinder (1973) approach when focusing on the between-year difference of average outcomes. But the possibility of disentangling the impact of each of the covariates in the composition effect rests on the very restrictive assumption that the data are generated from a linear model. As pointed out by Rothe (2015), in the general case, it is difficult to express the composition effect as a sum of terms which depend on the marginal distribution of a single covariate only. Instead, an explicit decomposition of the composition effect in terms of the respective marginal covariate distributions typically contains "interaction terms" resulting from the interplay of two or more covariates, and also "dependence terms" resulting from between-year difference in the dependence pattern among the covariates.

Rothe (2015) proposes to use results from copula theory in order to disentangle the covariates' marginal distributions from the dependence structure among them. Indeed, the CDF of $X^t$ can always be written as

$$F_X^t(x) = C^t(F_{X_1}^t(x_1), F_{X_2}^t(x_2)) \quad \text{for} \quad t \in \{2004, 2014\} \tag{3.4}$$

following Sklar's Theorem (Sklar, 1959). $C^t(.)$ is a copula function, i.e., a bivariate CDF with standard uniformly distributed marginals, and $F_{X_j}^t(.)$ is the marginal distribution of the $j$th component of $X^t$ (Trivedi and Zimmer, 2007). The copula describes the joint distribution of individuals' ranks in the two components of $X^t$. The copula accounts for the dependence between the covariates in a way that is separate from and independent of their marginal specifications. This result holds for continuous covariates. When some of them are discrete, some identifiability issues may arise, that can be solved by making parametric restrictions on the functional form of the copula.

In this context, the decomposition given by Eq. (3.3) can then be generalized as follows. Let $\mathbf{k}$ denote an element of the 2-dimensional product set $\{2004, 2014\}^2$, i.e. $\mathbf{k} = (k_1, k_2)$ where $k_1$ (resp. $k_2$) is equal to either 2004 or 2014. We can define the distribution of the outcome in a counterfactual

setting where the conditional distribution is as in year $t$, the covariate distribution has the copula function of year $s$, and the marginal distribution of the $l$th covariate is equal to that in group $\mathbf{k}$ by

$$F_Y^{t|s,\mathbf{k}} = \int F_{Y|X}^t(y,x) dF_X^{s,\mathbf{k}}(x) \tag{3.5}$$

with

$$F_X^{s,\mathbf{k}}(x) = C^s(F_{X_1}^{k_1}(x_1), F_{X_2}^{k_2}(x_2)). \tag{3.6}$$

For instance, the counterfactual distribution $F_Y^{2004|2014}$ in Eq. (3.3) can be written as $F_Y^{2004|2014,\mathbf{1}}$ where $\mathbf{1} = (2014, 2014)$. In other words, the computation of the counterfactual distribution $F_Y^{2004|2014}$ uses the conditional distribution of the outcome given the covariates in year 2004, the dependence structure of year 2014 , and the marginal distributions of the covariates in year 2014. Similarly, we can get $F_Y^{2004} = F_Y^{2004|2004,\mathbf{0}}$ where $\mathbf{0} = (2004, 2004)$.

Now we can write the composition effect $\Delta_X^\nu$ as

$$\begin{aligned}
\Delta_X^\nu &= \nu\left(F_Y^{2004|2014}\right) - \nu\left(F_Y^{2004}\right) \\
&= \nu\left(F_Y^{2004|2014,\mathbf{1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{0}}\right) \\
&= \left(\nu\left(F_Y^{2004|2014,\mathbf{1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{1}}\right)\right) + \left(\nu\left(F_Y^{2004|2004,\mathbf{1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{0}}\right)\right) \\
&= \Delta_D^\nu + \beta^\nu(\mathbf{1})
\end{aligned} \tag{3.7}$$

The first term of the decomposition in Eq. (3.7), or

$$\Delta_D^\nu = \nu\left(F_Y^{2004|2014,\mathbf{1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{1}}\right),$$

captures the contribution of the between-year difference of the covariates' copula functions. $\Delta_D^\nu$ is thus a *dependence effect*. The second term, or

$$\beta^\nu(\mathbf{1}) = \nu\left(F_Y^{2004|2004,\mathbf{1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{0}}\right)$$

measures the joint contribution of between-year differences in the marginal covariate distributions.

Let now $\mathbf{e^1} = (2014, 2004)$ and $\mathbf{e^2} = (2004, 2014)$. $\beta^\nu(\mathbf{1})$ can in turn be decomposed as

$$\beta^\nu(\mathbf{1}) = \left(\beta^\nu(\mathbf{1}) - \beta^\nu(\mathbf{e^1}) - \beta^\nu(\mathbf{e^2})\right) + \beta^\nu(\mathbf{e^1}) + \beta^\nu(\mathbf{e^2}) \tag{3.8}$$

with

$$\begin{aligned}
\beta^\nu(\mathbf{e^1}) &= \nu\left(F_Y^{2004|2004,\mathbf{e^1}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{0}}\right) \\
&\text{and} \\
\beta^\nu(\mathbf{e^2}) &= \nu\left(F_Y^{2004|2004,\mathbf{e^2}}\right) - \nu\left(F_Y^{2004|2004,\mathbf{0}}\right)
\end{aligned}$$

In other words, $\beta^\nu(\mathbf{e^1})$ and $\beta^\nu(\mathbf{e^2})$ measure the respective direct contributions of the first and second covariate. Let $\Delta^\nu_M(\mathbf{1}) \equiv \beta^\nu(\mathbf{1}) - \beta^\nu(\mathbf{e^1}) - \beta^\nu(\mathbf{e^2})$. $\Delta^\nu_M(\mathbf{1})$ can then interpreted as a "pure" *interaction effect*.

To sum up, the composition effect can be written as

$$\Delta^\nu_X = \beta^\nu(\mathbf{e^1}) + \beta^\nu(\mathbf{e^2}) + \Delta^\nu_M(\mathbf{1}) + \Delta^\nu_D, \tag{3.9}$$

i.e, as the sum of the respective contributions of each covariate, a term measuring the pure effect of their interaction, and a term measuring the contribution due to the between-year variation of the dependence between covariates. This decomposition can easily be generalized in the case of more than two covariates and focus either on individual effect of each of them and the pure effect of their interaction as shown above, or on the effect of groups of variables and of the interaction among these groups.

### 3.2.2 Practical implementation

Consider now the general case where the vector of the covariates has $d$ elements, and suppose we have two iid samples $\{(Y^t_i, X^t_i)\}_{i=1}^{n_t}$ of size $n_t$ from the distribution of $(Y^t, X^t)$ for $t = 2004, 2014$. The practical implementation of the decomposition procedure presented above requires the estimation of various functions or parameters.

**Univariate CDFs.** Univariate CDFs are estimated nonparametrically using the classical empirical CDF, i.e.

$$\widehat{F}^t_{X_j}(x_j) = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbb{I}(X^t_{ji} \leq x_j) \tag{3.10}$$

**Conditional CDF of $Y^t|X^t$.** The conditional CDF of $Y^t|X^t$ is a multivariate function whose dimension depends on the number of covariates. A nonparametric estimate of this function can be quite imprecise when the number of covariates is large, due to the so-called curse of dimensionality. Flexible parametric specifications can be used to overcome this drawback of nonparametric estimators (see Fortin et al. (2011)). As in Rothe (2015), conditional CDFs $F^t_{Y|X}$ are estimated using the distributional regression approach of Foresi and Peracchi (1995). The distributional regression model assumes that

$$F^t_{Y|X}(y, x) \equiv \Phi(x'\delta^t(y)), \tag{3.11}$$

where $\Phi(.)$ is the standard normal CDF. The finite-dimensional parameter $\delta^t(y)$ is estimated by the maximum likelihood estimate $\widehat{\delta}^t(y)$ in a Probit model that relates the indicator variable $\mathbb{I}(Y^t \leq y)$ to the covariates $X^t$.

**Copula choice.** The last function necessary for the implementation of the decomposition procedure of Rothe (2015) is the copula function. Let us take a copula contained in a parametric class indexed by a $k$-dimensional parameter $\theta$. A strategy for estimating the parameters characterizing the copula then consists in choosing the minimum distance estimator defined as (Weiß, 2011)

$$\widehat{\theta}^t = \arg\min_{\theta} \sum_{i=1}^{n_t} \left( \widehat{F}^t_X(X^t_{1i}, \ldots, X^t_{di}) - C_\theta(\widehat{F}^t_{X_1}(X^t_{1i}), \ldots, \widehat{F}^t_{X_d}(X^t_{di})) \right) \quad (3.12)$$

Different parametric copula functions can be used (Trivedi and Zimmer, 2007). But, here too, we must keep in mind when choosing this function to select a function that is sufficiently flexible for generating all possible types of dependence. Moreover, we are confronted here with the fact that our variables are a mixture of continuous and discrete variables. To address these issues, we choose the Gaussian copula model

$$C_\Sigma(u) = \Phi^d_\Sigma \left( \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d) \right) \quad (3.13)$$

where $\Phi^d_\Sigma(.)$ denotes the CDF of a $d$-variate standard normal distribution with correlation matrix $\Sigma$, and $\Phi^{-1}(.)$ is the inverse function of the standard normal distribution function $\Phi(.)$. The parameters $\theta \equiv \Sigma$ determine the dependence pattern among the covariates.

The flexibility and the analytical tractability of Gaussian copulas make them a handy tool in applications as emphasized by Jiryaie et al. (2016). First, This specification has a computational advantage, namely, that only the $(a, b)$ element of $\Sigma$ affects the pairwise dependence between the covariates $X^t_a$ and $X^b_b$. So minimum distance estimation (3.12) can be performed for each pair of covariates, not by taking all the covariates together simultaneously.

Second, as noted above, the copula function describes the joint distribution of individuals' ranks in the various components of $X^t$, and, here, the dependence between two components can be measured using a correlation coefficient as we are working with Gaussian copula. Indeed, in the bivariate case, we get

$$C_{\Sigma_{a,b}}(F_{X_a}(X_{ai}), F_{X_b}(X_{bi})) = \Phi^2_{\Sigma_{a,b}} \left( \Phi^{-1}(F_{X_a}(X_{ai})), \Phi^{-1}(F_{X_b}(X_{bi})) \right) \quad (3.14)$$

where $\Phi^2_{\Sigma_{a,b}}(.)$ denotes the CDF of the bivariate normal distribution with covariance matrix $\Sigma_{a,b}$, and $\Phi^{-1}(F_{X_a}(X_{ai}))$ (resp.$\Phi^{-1}(F_{X_b}(X_{bi}))$ can be interpreted as the quantile of the univariate marginal distribution associated to the observation $X_{ai}$ (resp. $X_{bi}$).

Third, Gaussian copulas make it possible to have both continuous and discrete variables in the vector of covariates. We only have to assume that

each discrete covariate $X_j^t$ can be represented as $X_j^t = t_j(\widetilde{X}_j^t)$ for some continuously distributed latent variable $\widetilde{X}_j^t$ and a function $t_j(.)$ that is weakly increasing in its argument. For instance, if $X_j^t$ is a binary, we could have $X_j^t = \mathbb{I}(\widetilde{X}_j^t > c_j)$ for some constant $c_j$. Details on the computation of the joint distribution of a vector of continuous and discrete variables using Gaussian copula can be found in Jiryaie et al. (2016).

**Counterfactual distributions.** After estimating the copula and the marginal distributions for each time period, we can construct the joint c.d.f. of the explanatory variables given by (3.6) in any counterfactual experiment where the copula is as in time $s$ and the marginals as in time $k_1$ and $k_2$. Given this joint c.d.f, using equation (3.5) and the conditional c.d.f $F_{Y|X}^t(y, x)$ at time $t$ estimated by equation (3.11), we can construct an estimation of any counterfactual distribution of the outcome.

## 3.3 Data

This study relies on the survey "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. 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 of 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 two waves of VHLSS conducted in 2004 and 2014.

### 3.3.1 Macronutrient intakes

Average annual or monthly food expenditures and quantities about 56 food items are collected for each household surveyed in each VHLSS wave.[1] The observed kilograms can then be converted into kilocalories using the conversion coefficients given in the Vietnamese Food Composition Table constructed by the Vietnam National Institute of Nutrition in 2007. Table 6.2 shows the coefficients that have been applied to perform these conversions into calorie intakes and amounts of proteins and fats, expressed as calorie intakes. Calorie intakes from carbohydrates are then obtained by difference. These annual calorie intakes, which are computed at the household level, are then converted into daily intakes and adjusted in the form of per capita calorie intakes to be comparable between households. This adjustment

[1] Only average annual food consumption was recorded in 2004 while monthly average food consumption was surveyed in 2014.

makes use of the household equivalence scale calculation procedure recently proposed by Aguiar and Hurst (2013).[2]

Figure 3.1 reports the kernel weighted estimates of the densities of per capita calorie intake for the two years. There is a shift to the right for the density from 2004 to 2014, indicating an increase in per capita calorie intake over the period, not only on average but also for all quantiles such as those reported in Table 3.2.

Figure 3.1: Density of per capita calorie intake

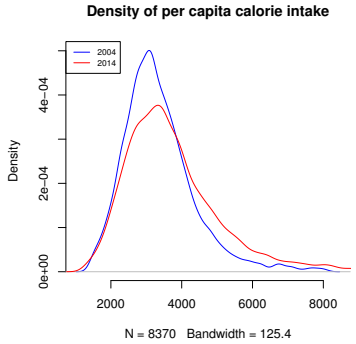



Figure 3.2: Density of per capita calorie intake by macronutrient

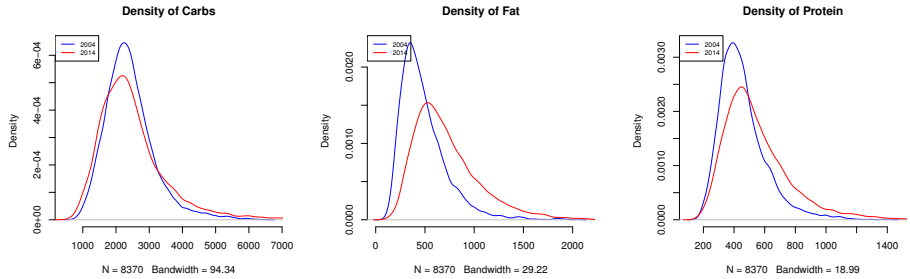



Figure 3.2 reports the kernel weighted estimates of the densities of per capita calorie intakes of carbohydrates, fat, and proteins, for the two years. Significant changes appear when comparing the estimated densities for fat and proteins, while the estimated densities for carbohydrates appear to be very close. There is a significant shift to the right for the estimated densities for fat and protein in 2014. Meanwhile, the estimated density for carbohydrates in 2014 has the same mode as in 2004, but becomes flatter. This visual observation is confirmed by the evolution of average values, standard deviations, and quantiles at 10, 50 and 90% as reported in Table 3.2. All these values increase significantly for fat and proteins. Average and median

[2]More details are given in Thi et al. (2018).

Table 3.1: Description of sociodemographic variables

| Variable | Values | Description |
|---|---|---|
| $lExp$ | | Food expenditures per year in US$ (in logarithms) |
| $Hsize$ | | Number of household members |
| $Urban$ | | Location of the household: |
| | $= 1$ | if household is located in urban area |
| | $= 0$ | if household is located in rural area |
| $Ethnic$ | | Ethnicity of head of household |
| | $= 1$ | if Kinh Ethnicity |
| | $= 0$ | if minority |
| $Yeducc$ | | Highest educational level of the head of households (year): |
| | $= 0$ | No schooling |
| | $= 5$ | Primary school level |
| | $= 9$ | Secondary school level |
| | $= 12$ | High school level |
| | $= 16$ | College degree |
| | $= 18$ | Master degree |
| | $= 21$ | Ph.D level |
| $South$ | | Region: |
| | $= 1$ | if Household is located in the South of Vietnam |
| | $= 0$ | otherwise |

values stay quite stable between 2004 and 2014 for carbohydrates while standard deviation increases, 10% quantile decreases, and 90% quantile increases. In other words, per capita calorie intakes from fat and proteins in Vietnamese households have increased over the considered period. Per capita calorie intake from carbohydrates remained stable on average, while this stability hides a contrasted picture with an increase for some households and a decrease for others.

### 3.3.2 Sociodemographic variables

Table 3.1 summarizes the sociodemographic variables we use in this paper, and detailed descriptive statistics on these variables are given in Table 3.2. These statistics show several interesting developments. First, total food expenditures of Vietnamese households increased over the considered period. Second, the population of these same households is more urbanized in 2014 than in 2004. Third, the average household size has decreased slightly, with about 65% of these households having four or fewer members in 2014 compared to about 55% ten years earlier. Fourth, heads of households are, on average, more educated in 2014 than in 2004. Furthermore, the proportion of heads with more than 12 schooling years (high school level) increased significantly from 2004 to 2014. Finally, the proportions of households with heads belonging to the Kinh ethnicity or living in South Vietnam remained stable.

Table 3.2: Descriptive statistics in VHLSS 2004 and 2014

| | Mean | SD | Q10 | Q50 | Q90 |
|---|---|---|---|---|---|
| **2004** | | | | | |
| $PCCI$ | 3359.746 | 1015.451 | 2259.852 | 3195.859 | 4609.399 |
| $V_C$ | 2415.078 | 756.170 | 1565.208 | 2318.522 | 3343.795 |
| $V_P$ | 457.920 | 156.403 | 294.643 | 428.629 | 653.904 |
| $V_F$ | 486.748 | 239.576 | 247.206 | 433.159 | 792.876 |
| $lExp$ | 6.135 | 0.547 | 5.461 | 6.125 | 6.844 |
| $Urban$ | 0.235 | | – | – | – |
| $Hsize$ | 4.355 | 1.636 | 2 | 4 | 6 |
| $Ethnic$ | 0.893 | | – | – | – |
| $Yeducc$ | 6.222 | 4.712 | 0 | 5 | 12 |
| $South$ | 0.345 | | – | – | – |
| **2014** | | | | | |
| $PCCI$ | 3764.194 | 1421.362 | 2313.206 | 3488.157 | 5528.041 |
| $V_C$ | 2493.419 | 1032.906 | 1445.146 | 2297.777 | 3764.969 |
| $V_P$ | 548.367 | 219.059 | 320.181 | 501.073 | 830.010 |
| $V_F$ | 722.409 | 343.119 | 367.404 | 647.299 | 1174.950 |
| $lExp$ | 6.638 | 0.611 | 5.843 | 6.667 | 7.399 |
| $Urban$ | 0.311 | | – | – | – |
| $Hsize$ | 3.808 | 1.526 | 2 | 4 | 6 |
| $Ethnic$ | 0.869 | | – | – | – |
| $Yeducc$ | 7.097 | 5.047 | 0 | 9 | 12 |
| $South$ | 0.339 | | – | – | – |

## 3.4 Results

To estimate the various elements of the decomposition of the composition effect, we proceed as described in section 3.2. Copulas are thereby modeled by a Gaussian copula and the joint CDF of each pair of covariates estimated using marginal empirical CDF estimators and copula estimators. Table 3.3 reports the estimated values of the copula parameters from the 2004 and 2014 VHLSS waves. Estimated copula parameters show positive and significant association between food expenditures and location in an urban area as well as food expenditures and household size. The first association decreased between 2004 and 2014 while the second remained fairly stable. The association between location in an urban area and ethnicity is negative and significant whatever the considered waves, as expected, and increases over the period. The association between location in an urban area and years of education is positive but becomes significant only in 2014. A stable positive and significant association is also shown for location in an urban area and living in South Vietnam. We also notice a negative association between household size and ethnicity in 2004, which disappears completely in 2014. As recently pointed out by Benjamin et al. (2017), the share of minorities in the rural population has risen over time, from below 15% in 2002 to over 18% in 2014. This is a consequence of a higher fertility among minorities, combined with rising urbanization among the Kinh. Finally, the association between the number of years of education and living in South Vietnam is negative and significant but decreasing between 2004 and 2014.

Table 3.3: Estimated copula parameters

| | Urban | | Hsize | | Ethnic | | Yeduc | | South | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2004 | 2014 | 2004 | 2014 | 2004 | 2014 | 2004 | 2014 | 2004 | 2014 |
| lExp | 0.50 | 0.41 | 0.54 | 0.58 | -0.17 | 0.21 | 0.16 | 0.24 | 0.36 | 0.19 |
| | (0.12) | (0.15) | (0.11) | (0.14) | (0.16) | (0.12) | (0.25) | (0.35) | (0.27) | (0.29) |
| Urban | | | -0.03 | 0.01 | -0.29 | -0.62 | 0.11 | 0.28 | 0.38 | 0.35 |
| | | | (0.27) | (0.20) | (0.08) | (0.09) | (0.09) | (0.10) | (0.07) | (0.08) |
| Hsize | | | | | -0.54 | -0.03 | -0.01 | -0.01 | 0.01 | -0.08 |
| | | | | | (0.25) | (0.20) | (0.12) | (0.13) | (0.13) | (0.108) |
| Ethnic | | | | | | | -0.19 | -0.13 | -0.25 | -0.38 |
| | | | | | | | (0.077) | (0.09) | (0.29) | (0.36) |
| Yeduc | | | | | | | | | -0.53 | -0.32 |
| | | | | | | | | | (0.13) | (0.12) |

Note: Bootstrapped standard errors, based on 300 replications, are in parenthesis.

Conditional CDFs $F^t_{Y|X}$ are modeled by a distributional regression model with a Gaussian link function. We do not report the results as they are not very helpful in the discussions that follow. Nevertheless, they are available from the authors.

Tables 3.4, 3.5, 3.6, and 3.7 then present the results of our decomposition of per capita calorie intake and calorie intake coming from the three macronutrients, for two measures of location: mean and median, and for the two quantiles at 10% and 90% allowing to construct a measure of dispersion. Row by row, we report estimates of total change, i.e. $\Delta^\nu_Y$, usual structure and composition effects, i.e. $\Delta^\nu_S$ and $\Delta^\nu_X$. Then the composition effect is in turn decomposed into the dependence effect, i.e. $\Delta^\nu_D$, and marginal distribution effect, i.e. $\beta^\nu(\mathbf{1})$. Finally, this last effect is decomposed into the direct contribution for each of the six covariates, i.e. the $\beta^\nu(\mathbf{e^l})$, and the "two-way" interaction effects, i.e. the $\Delta^\nu_M$. Figures 3.3 and 3.4 summarize these same results in the form of barplots.

Each estimated value in a decomposition is accompanied by the estimated value of its standard error. Rothe (2015) shows the asymptotic convergence of the estimator of each element in a decomposition to a mean zero normal distribution. But, as the asymptotic variance of these estimators takes a fairly complicated form, a practical way to estimate this variance is the use of a standard nonparametric bootstrap in which the estimates are recomputed a large number of times on bootstrapped samples $\{\widetilde{Y}^t_i, \widetilde{X}^t_i\}^{n_t}_{i=1}$ drawn with replacement from the original data $\{Y^t_i, X^t_i\}^{n_t}_{i=1}$. The bootstrap variance estimator then coincides with the empirical variance of the bootstrapped estimates. Here, estimated standard errors are calculated using nonparametric bootstrap with 300 replications.

Knowledge of the estimated values of total difference and the associated standard errors first allow to have an indication as to whether the chosen modeling of decomposition using parametric restrictions on copulas and conditional distributions, provides a reasonable fit. Indeed, these estimated values of total difference can be compared with the differences that can be directly calculated from the descriptive statistics given in Table 3.2. It should be noted that, in all cases, the difference computed from the descrip-

Table 3.4: Estimated decomposition of per capita calorie intake

| | Mean | | Q10 | | Median | | Q90 | |
|---|---|---|---|---|---|---|---|---|
| Total difference | 362.16 | (28.90) | 18.62 | (26.38) | 279.48 | (20.81) | 830.71 | (78.76) |
| Structure effect | -291.21 | (52.53) | -283.63 | (49.12) | -361.79 | (39.69) | -328.38 | (213.09) |
| Composition effect | 653.37 | (44.47) | 302.25 | (46.16) | 641.27 | (36.71) | 1159.09 | (202.78) |
| *Composition effect*: | | | | | | | | |
| Dependence effect | 0.91 | (23.08) | -30.6 | (22.94) | 0.25 | (23.26) | -7.97 | (135.35) |
| Marginal effect | 652.46 | (39.97) | 332.85 | (42.83) | 641.02 | (33.65) | 1167.06 | (206.92) |
| *"Direct" contributions to composition effect*: | | | | | | | | |
| lexp | 532.86 | (36.16) | 250.05 | (35.54) | 538.66 | (33.05) | 900.04 | (137.13) |
| Urban | -11.06 | (2.90) | -11.55 | (3.28) | -9.56 | (3.41) | -9.49 | (8.96) |
| Hsize | 131.12 | (8.94) | 53.62 | (8.46) | 112.77 | (12.26) | 246.08 | (27.01) |
| Ethnic | 0.69 | (1.53) | 1.90 | (1.33) | 0.34 | (1.69) | -1.61 | (3.95) |
| Yeduc | -18.16 | (7.08) | -3.09 | (6.21) | -14.00 | (5.72) | -26.18 | (12.53) |
| South | 0.99 | (1.06) | 0.88 | (0.96) | 0.83 | (1.30) | 1.11 | (1.30) |
| *"Two-way" interaction effects:* | | | | | | | | |
| lexp:Urban | -1.73 | (5.08) | 7.41 | (6.83) | -9.83 | (9.02) | 3.38 | (23.43) |
| lexp:Hsize | 23.58 | (10.38) | 50.01 | (24.08) | 30.04 | (20.97) | 34.87 | (129.97) |
| lexp:Ethnic | 0.61 | (2.70) | 3.89 | (3.58) | 2.36 | (4.12) | 2.36 | (14.9) |
| lexp:Yeduc | -6.14 | (6.01) | -7.56 | (10.87) | -11.06 | (11.26) | -14.48 | (32.18) |
| lexp:South | 0.44 | (0.70) | 0.03 | (0.80) | 0.21 | (1.28) | 0.29 | (3.96) |
| Urban:Hsize | 0.14 | (1.19) | 2.62 | (2.73) | -6.45 | (4.10) | 1.47 | (6.32) |
| Urban:Ethnic | -0.45 | (0.29) | -0.54 | (0.47) | -0.26 | (0.54) | -0.39 | (1.48) |
| Urban:Yeduc | 0.41 | (0.81) | 0.17 | (1.22) | -2.37 | (2.39) | -1.43 | (3.73) |
| Urban:South | -0.20 | (0.22) | -0.01 | (0.19) | -0.05 | (0.31) | -0.64 | (0.70) |
| Hsize:Ethnic | 0.84 | (0.48) | 0.90 | (1.09) | 1.25 | (1.37) | 0.73 | (2.68) |
| Hsize:Yeduc | -2.38 | (2.05) | -1.76 | (3.85) | -14.84 | (6.15) | -5.78 | (10.89) |
| Hsize:South | -0.06 | (0.15) | -0.43 | (0.46) | 0.63 | (0.72) | -0.22 | (0.74) |
| Ethnic:Yeduc | -0.32 | (0.40) | -0.61 | (0.50) | 0.29 | (0.62) | -0.17 | (1.63) |
| Ethnic:South | 0.03 | (0.05) | 0.05 | (0.09) | -0.01 | (0.14) | -0.10 | (0.14) |
| Yeduc:South | 0.04 | (0.07) | -0.01 | (0.14) | 0.04 | (0.38) | -0.20 | (0.44) |

Note: Bootstrapped standard errors, based on 300 replications, are in parenthesis.

tive statistics belongs to the 95% confidence interval that can be constructed from the estimated value of total difference and its estimated standard error. Moreover, the estimated values of total difference for quantiles capture well the observed shifts in empirical quantiles of calorie intake distributions. The chosen model thus provides a reasonable fit to the data.

Let us now look more closely at each of the tables. Table 3.4 presents the estimated values of the various elements in the decomposition of differences in means, median and quantiles at 10% and 90% between the two years for per capita calorie intake. The decomposition of total difference in structure effect and composition effect reveals two effects that play in opposite directions. A strong positive composition effect then appears while the structure effect is negative and quite stable among quantiles. The composition effect is only counterbalanced by the structural effect in the case of the quantile at 10%. Moreover, the composition effect increases with the quantile.

In other words, the change in the conditional distributions of per capita calorie intake given the sociodemographic characteristics, i.e. in the relationship between per capita calorie intake and these covariates, between the two years caused a significant decrease in per capita calorie intake on average as well as on the three considered quantiles. Meanwhile, the change in the composition of the sample of households between the two years led

Table 3.5: Estimated decomposition of calorie intake from fat

| | Mean | | Q10 | | Median | | Q90 | |
|---|---|---|---|---|---|---|---|---|
| Total difference | 221.51 | (8.68) | 119.61 | (6.13) | 200.73 | (7.06) | 364.92 | (28.06) |
| Structure effect | -17.63 | (13.93) | -1.92 | (6.8) | -15.85 | (10.51) | -5.33 | (55.00) |
| Composition effect | 239.14 | (12.39) | 121.53 | (8.13) | 216.57 | (9.58) | 370.25 | (55.94) |
| *Composition effect*: | | | | | | | | |
| Dependence effect | -0.77 | (8.01) | 2.34 | (5.94) | -0.67 | (5.01) | 6.33 | (46.76) |
| Marginal effect | 239.91 | (11.11) | 119.19 | (5.95) | 217.24 | (8.49) | 363.92 | (54.53) |
| *"Direct" contributions to composition effect*: | | | | | | | | |
| lexp | 178.97 | (9.34) | 80.78 | (6.57) | 173.74 | (7.12) | 296.02 | (36.63) |
| Urban | 2.51 | (0.68) | 0.77 | (0.3) | 2.39 | (0.75) | 2.91 | (2.11) |
| Hsize | 47.64 | (2.68) | 26.16 | (2.55) | 44.33 | (2.74) | 71.02 | (7.29) |
| Ethnic | 0.24 | (0.44) | -0.18 | (0.21) | 0.14 | (0.33) | -0.65 | (1.17) |
| Yeduc | -0.98 | (0.92) | -0.04 | (1.19) | 0.75 | (1.27) | -6.75 | (2.79) |
| South | 0.28 | (0.34) | 0.10 | (0.15) | 0.30 | (0.37) | 0.67 | (0.93) |
| *"Two-way" interaction effects:* | | | | | | | | |
| lexp:Urban | 0.31 | (1.37) | 2.62 | (1.28) | -1.40 | (1.24) | -3.23 | (7.07) |
| lexp:Hsize | 10.13 | (3.30) | 9.36 | (5.19) | -1.56 | (5.72) | 25.03 | (32.16) |
| lexp:Ethnic | 0.54 | (0.91) | 0.48 | (0.47) | -0.22 | (0.75) | 1.85 | (3.51) |
| lexp:Yeduc | -0.56 | (1.26) | 1.83 | (2.46) | -2.07 | (2.02) | 2.70 | (7.42) |
| lexp:South | 0.19 | (0.23) | 0.00 | (0.16) | 0.13 | (0.38) | -0.14 | (0.69) |
| Urban:Hsize | 0.54 | (0.33) | 0.62 | (0.32) | -0.74 | (0.67) | 1.63 | (1.90) |
| Urban:Ethnic | -0.02 | (0.13) | -0.01 | (0.03) | 0.04 | (0.11) | 0.07 | (0.31) |
| Urban:Yeduc | -0.05 | (0.13) | -0.01 | (0.10) | -0.32 | (0.25) | 0.20 | (0.86) |
| Urban:South | -0.04 | (0.05) | 0.00 | (0.02) | -0.03 | (0.06) | 0.07 | (0.21) |
| Hsize:Ethnic | 0.04 | (0.16) | 0.14 | (0.16) | 0.35 | (0.27) | 0.53 | (0.76) |
| Hsize:Yeduc | -0.33 | (0.32) | 0.77 | (1.03) | -0.83 | (0.90) | -0.88 | (2.71) |
| Hsize:South | 0.04 | (0.05) | 0.05 | (0.06) | 0.01 | (0.14) | 0.01 | (0.46) |
| Ethnic:Yeduc | -0.08 | (0.10) | -0.04 | (0.06) | -0.06 | (0.11) | -0.34 | (0.47) |
| Ethnic:South | 0.00 | (0.02) | 0.00 | (0.01) | 0.00 | (0.02) | 0.03 | (0.10) |
| Yeduc:South | 0.02 | (0.02) | 0.00 | (0.03) | 0.01 | (0.06) | 0.01 | (0.29) |

Note: Bootstrapped standard errors, based on 300 replications, are in parenthesis.

to a significant increase in per capita calorie intake. This increase was larger than the decrease due to changes in the relationship between per capita calorie intake and sociodemographic variables, except for the 10% quantile where the two compensate.

The dependence effect that captures the contribution of between-year differences in the covariates' copula functions plays no role in the decomposition of composition effect. The dependence effect is never significantly different from zero. The composition effect is almost always equal to the total marginal distribution effect resulting from differences in the marginal covariate distributions across the two years.

Consider now the decomposition of the total marginal distribution effect into direct effects of each covariate and "two-way" interactions effects. This decomposition shows the importance of the contribution of food expenditures and household size to total marginal distribution effect, i.e., here, the composition effect. These contributions are indeed positive, large, and significantly different from zero. It should be noted that these contributions increase according to the considered quantile order. Food expenditures and household size play a more and more important role in the increase of per capita calorie intake when moving from the 10% quantile to the 90% quantile. The effects of these two variables are barely offset by the nega-

Table 3.6: Estimated decomposition of calorie intake from protein

| | Mean | | Q10 | | Median | | Q90 | |
|---|---|---|---|---|---|---|---|---|
| Total difference | 85.74 | (4.54) | 23.76 | (4.37) | 70.93 | (3.97) | 163.34 | (11.17) |
| Structure effect | -52.32 | (9.08) | -43.97 | (7.05) | -61.23 | (7.35) | -41.46 | (31.46) |
| Composition effect | 138.06 | (8.09) | 67.73 | (7.06) | 132.16 | (7.18) | 204.8 | (30.44) |
| *Composition effect*: | | | | | | | | |
| Dependence effect | 2.94 | (6.06) | 2.80 | (4.98) | 2.66 | (4.74) | -5.93 | (25.79) |
| Marginal effect | 135.12 | (7.13) | 64.93 | (6.78) | 129.5 | (6.73) | 210.73 | (26.85) |
| *"Direct" contributions to composition effect*: | | | | | | | | |
| lexp | 108.11 | (6.08) | 49.37 | (5.10) | 108.17 | (5.54) | 169.77 | (19.08) |
| Urban | -0.57 | (0.40) | -0.82 | (0.32) | -0.91 | (0.45) | 0.14 | (1.06) |
| Hsize | 26.89 | (1.43) | 15.33 | (1.63) | 23.34 | (1.46) | 43.44 | (6.05) |
| Ethnic | -0.21 | (0.19) | -0.04 | (0.15) | -0.27 | (0.26) | -0.37 | (0.54) |
| Yeduc | -1.84 | (0.82) | -0.29 | (0.94) | -2.44 | (0.86) | -4.16 | (1.51) |
| South | -0.06 | (0.09) | 0.00 | (0.03) | -0.01 | (0.04) | -0.08 | (0.17) |
| *"Two-way" interaction effects:* | | | | | | | | |
| lexp:Urban | -0.23 | (0.98) | 1.28 | (0.99) | -2.89 | (1.42) | 4.36 | (3.93) |
| lexp:Hsize | 2.74 | (2.10) | 0.80 | (4.47) | 6.84 | (4.70) | -7.37 | (17.44) |
| lexp:Ethnic | -0.56 | (0.44) | -0.11 | (0.53) | -0.29 | (0.60) | -1.60 | (2.56) |
| lexp:Yeduc | 0.14 | (0.86) | -1.41 | (1.97) | -1.76 | (1.77) | -1.35 | (3.93) |
| lexp:South | 0.14 | (0.21) | 0.06 | (0.10) | 0.21 | (0.28) | 0.08 | (0.32) |
| Urban:Hsize | 0.28 | (0.19) | -0.10 | (0.28) | 0.25 | (0.37) | 1.51 | (1.49) |
| Urban:Ethnic | -0.04 | (0.04) | 0.00 | (0.04) | -0.06 | (0.08) | 0.03 | (0.12) |
| Urban:Yeduc | -0.03 | (0.09) | -0.10 | (0.16) | -0.01 | (0.30) | 0.12 | (0.31) |
| Urban:South | -0.03 | (0.04) | 0.00 | (0.01) | -0.01 | (0.02) | -0.08 | (0.10) |
| Hsize:Ethnic | 0.04 | (0.06) | 0.19 | (0.12) | 0.19 | (0.15) | -0.19 | (0.66) |
| Hsize:Yeduc | -0.15 | (0.28) | -0.07 | (0.78) | -0.18 | (0.74) | -1.24 | (1.85) |
| Hsize:South | -0.02 | (0.02) | -0.01 | (0.03) | 0.01 | (0.03) | -0.19 | (0.23) |
| Ethnic:Yeduc | -0.02 | (0.04) | 0.03 | (0.05) | -0.02 | (0.12) | -0.09 | (0.20) |
| Ethnic:South | 0.00 | (0.05) | 0.00 | (0.00) | 0.00 | (0.00) | 0.01 | (0.04) |
| Yeduc:South | 0.00 | (0.01) | 0.00 | (0.01) | 0.01 | (0.02) | -0.01 | (0.04) |

Note: Bootstrapped standard errors, based on 300 replications, are in parenthesis.

tive and significantly different from zero effects of urbanization and years of education of the head of the household. Moreover, almost all "two-way" interaction effects are negligible.

Similar comments can be made regarding decompositions for consumption in terms of calories from fat and protein (Tables 3.5 and 3.6). Thus, the estimated values of the total difference for the different quantiles closely trace the observed uniform changes in these distributions towards higher consumption of the two macronutrients. Again, the main source of change comes from the composition effect that the structural effect only partially compensates for. It should be noted that the structural effect is never significantly different from zero in the case of fat. The dependence effect is negligible, and the main contributors to the composition effect are still food expenditures and household size. The estimated values of the impacts of these two covariates on changes in consumed calories from fat and protein increase when moving from the 10% quantile to the 90% quantile. The number of years of education of the head of household still impacts negatively on changes, the effects being sometimes not significantly different from zero. The effect of urbanization is negligible in the case of proteins, whereas it becomes positive in the case of fat. Nevertheless, although significantly different from zero for most of considered statistics, the effect of urbanization

Table 3.7: Estimated decomposition of calorie intake from carbohydrates

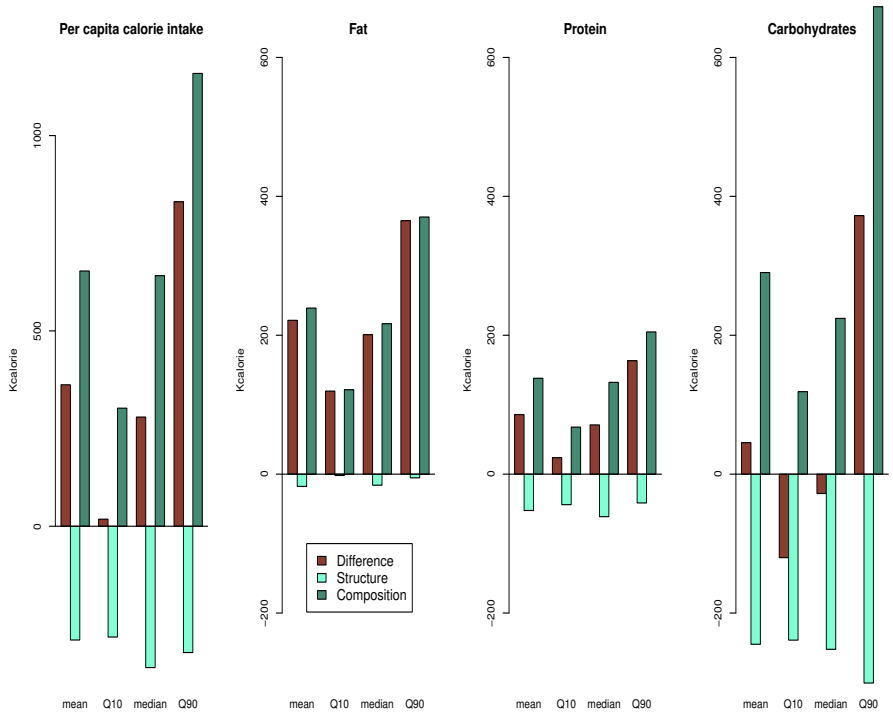
| | Mean | | Q10 | | Median | | Q90 | |
|---|---|---|---|---|---|---|---|---|
| Total difference | 45.36 | (22.04) | -120.24 | (21.14) | -27.88 | (18.25) | 372.23 | (50.91) |
| Structure effect | -244.93 | (36.73) | -238.99 | (30.56) | -252.19 | (24.66) | -300.78 | (134.48) |
| Composition effect | 290.29 | (32.66) | 118.75 | (29.32) | 224.31 | (22.52) | 673.01 | (128.18) |
| *Composition effect*: | | | | | | | | |
| Dependence effect | 0.68 | (17.57) | -6.03 | (15.75) | 5.50 | (10.70) | 18.43 | (74.42) |
| Marginal effect | 289.61 | (29.3) | 124.78 | (24.67) | 218.81 | (19.74) | 654.58 | (110.40) |
| *"Direct" contributions to composition effect*: | | | | | | | | |
| lexp | 253.01 | (25.7) | 135.13 | (20.79) | 207.47 | (19.2) | 528.18 | (88.29) |
| Urban | -12.63 | (2.34) | -16.17 | (3.88) | -13.01 | (2.78) | -11.54 | (6.03) |
| Hsize | 59.84 | (5.07) | 19.53 | (6.60) | 43.53 | (5.54) | 140.86 | (22.94) |
| Ethnic | 0.60 | (1.35) | 1.29 | (1.95) | 0.93 | (1.45) | -0.50 | (2.55) |
| Yeduc | -15.53 | (5.07) | -6.82 | (6.62) | -19.57 | (4.18) | -18.59 | (7.90) |
| South | 0.75 | (0.78) | 0.93 | (1.00) | 0.83 | (0.94) | 0.47 | (0.69) |
| *"Two-way" interaction effects:* | | | | | | | | |
| lexp:Urban | -2.20 | (4.04) | 8.73 | (5.57) | -0.93 | (5.67) | -14.17 | (17.3) |
| lexp:Hsize | 13.24 | (6.80) | -7.25 | (8.55) | -1.78 | (10.89) | 14.15 | (59.55) |
| lexp:Ethnic | 0.83 | (1.92) | 4.82 | (2.70) | -0.38 | (1.78) | 3.43 | (8.75) |
| lexp:Yeduc | -6.16 | (4.28) | -9.55 | (8.78) | -2.97 | (7.21) | 0.29 | (16.59) |
| lexp:South | 0.08 | (0.48) | -0.52 | (0.76) | -0.25 | (0.56) | 1.08 | (2.45) |
| Urban:Hsize | -0.73 | (0.79) | 0.20 | (3.37) | 0.08 | (3.28) | 3.76 | (5.80) |
| Urban:Ethnic | -0.41 | (0.25) | -0.52 | (0.61) | -0.67 | (0.54) | -0.39 | (0.78) |
| Urban:Yeduc | 0.52 | (0.52) | 1.80 | (2.12) | 3.10 | (2.39) | -0.73 | (1.91) |
| Urban:South | -0.14 | (0.16) | -0.29 | (0.26) | -0.09 | (0.27) | -0.32 | (0.35) |
| Hsize:Ethnic | 0.80 | (0.34) | 1.69 | (1.02) | 0.68 | (0.91) | 1.22 | (2.63) |
| Hsize:Yeduc | -1.93 | (1.30) | -4.10 | (2.70) | -1.57 | (4.60) | -12.03 | (7.96) |
| Hsize:South | -0.09 | (0.14) | 0.00 | (0.37) | -0.04 | (0.33) | -0.07 | (0.52) |
| Ethnic:Yeduc | -0.28 | (0.29) | -0.68 | (0.74) | -0.28 | (0.67) | -0.27 | (0.67) |
| Ethnic:South | 0.03 | (0.04) | 0.07 | (0.13) | 0.03 | (0.09) | 0.03 | (0.06) |
| Yeduc:South | 0.03 | (0.05) | -0.08 | (0.22) | -0.07 | (0.32) | 0.00 | (0.15) |

Note: Bootstrapped standard errors, based on 300 replications, are in parenthesis.

is negligible when compared to those of food expenditure or household size.

The results obtained in the case of carbohydrates are more contrasted than the previous ones (see Table 3.7). Here again, the estimated values of the total difference trace well what is observed for the empirical distributions of calories consumed from carbohydrates, whether in terms of location or spread statistics. Thus, total differences for mean and median are not significantly different from zero at the 10% and 5% threshold respectively, while total differences for 10% and 90% quantiles are significantly different from zero, the first being negative while the second is positive. The results capture well the flattening of the distribution between 2004 and 2014. But now, the structure effect compensates the composition effect in the cases of the mean and median, or even exceeds it for 10% quantile when decomposing total difference. As for the decomposition of the composition effect, it gives rise to similar comments to those made above for per capita calorie intake: negligible dependence effect, and strong positive contributions of food expenditures and household size compensated in part by negative contributions of urbanization and level of education of head of household.

Figure 3.3: Total differences, composition and structure effects

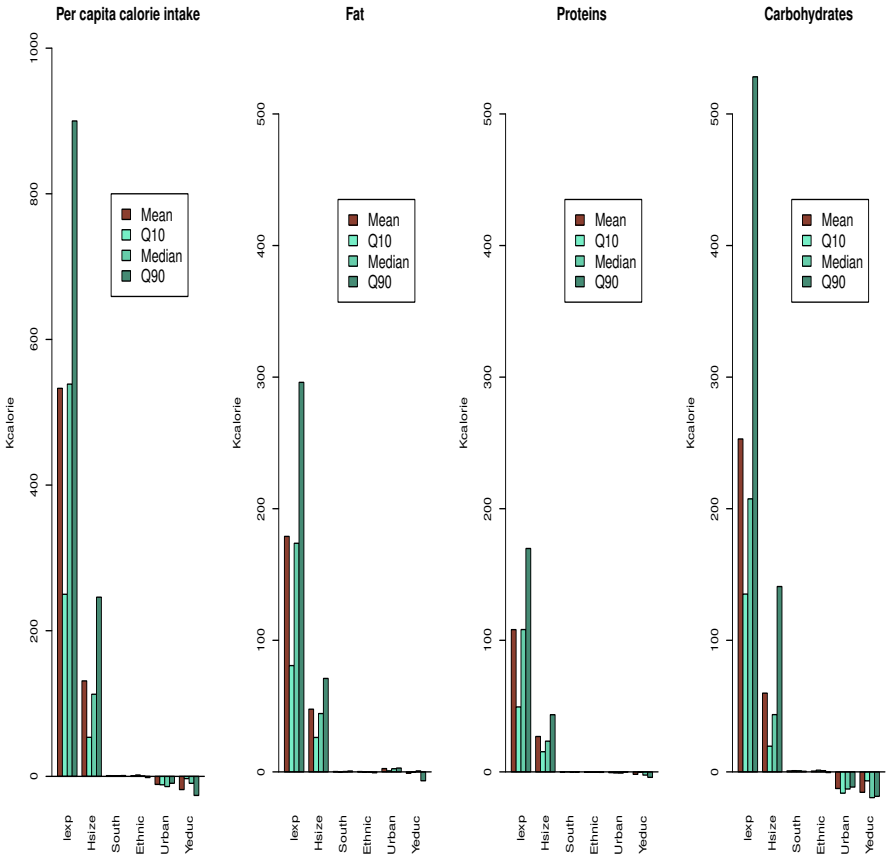



## 3.5 Conclusion

The aim of this paper is to document the evolution of Vietnamese household consumption in terms of total calorie intake and consumption of macronutrients over the period 2004-2014. The availability of VHLSS surveys makes it possible to have detailed data on these consumptions. The descriptive analysis of the data reveals an increase in per capital calorie intake over the period not only on average but also at all the quantiles of the corresponding distribution. The same evolution is observed for the consumption of proteins and that of fat. The distribution of carbohydrate consumption, on the other hand, flattens, showing an increase in low and high consumption between the two years while staying stable on average.

The characterization of the drivers of these evolutions is based on the use of a decomposition method recently proposed by Rothe (2015). In addition to the classical decomposition of between-year changes in terms of structure and decomposition effects, this method allows us to compute the direct contributions of various socio-demographic variables and the effects of their interactions in these between-year changes. We implement this method on VHLSS data to characterize the different effects on between-year mean, median, and 10% and 90% quantiles changes in per capita calorie intake

Figure 3.4: Direct contributions to the composition effects




and macronutrient consumptions in Vietnam.

The main results we have obtained can be summarized as follows (see Figures 3.3 and 3.4). First, decompositions using parametric restrictions on copulas and conditional distributions provide a reasonable fit. The estimated values of the between-year total differences clearly reflect the observed differences, either on average or for the considered quantiles. Second, the structure and composition effects play in an opposite direction, whatever the considered decomposition. Structure effects, which come from between-year differences in the relationship that links the covariates to the considered outcome, are always negative, while composition effects, which are due to differences in the distributions of observable covariates across years, are always positive. Third, the composition effect often outweighs the structure effect when considering the between-year changes in distributions of per capita calorie intake or calorie intake coming from protein or fat. The effects of changes in the composition of the Vietnamese population thus overcome

the effects of changes in preferences of the same population. This finding is particularly striking in the case of calorie intake from fat where structure effects are never distinguishable from zero. In the case of carbohydrates, this finding is reversed, with the exception of the 90% quantile. Fourth, food expenditure and household size appear to be the main contributors to the composition effect, regardless of the considered decomposition. The positive effects of these two variables explain well most of the between-year shifts observed in the calorie intake distributions. Urbanization and level of education contribute negatively to the compositional effect, with the noticeable exception of fat where the effect of urbanization is positive. In all cases, the effects of the latter two variables are negligible compared to those of food expenditure and household size. Finally, dependence effects and two-way interaction effects appear to be negligible or insignificant.

The decomposition method we use in this paper focuses on the decomposition of the composition effect into its main drivers: the direct effects of covariates or the effects of their interactions. It therefore allows a detailed analysis of one of the two sides of the decomposition, the composition effect, but it says nothing about the structure effect. Our application shows that the latter effect can play an important role. The related issue of deriving a decomposition of the structure effect, that is, dividing between-year differences in the structural functions that link the covariates and the outcome variable, into components that can be attributed to individual covariates, still is an open issue.

# Chapter 4

# Relations between socio-economic factors and nutritional diet in Vietnam: new insights using compositional data analysis

This paper contributes to the analysis of the impact of socioeconomic factors, like food expenditure level and urbanization, on diet patterns in Vietnam, from 2004 to 2014. Contrary to the existing literature, we focus on the diet balance in terms of macronutrients consumption (protein, fat and carbohydrate) and we take into account the fact that the volumes of macronutrients are not independent. In other words, we are interested in the shares of each macronutrient in the total calorie intake. We use compositional data analysis (CODA), adapted to deal with the relative information contained in shares, to describe the evolution of diet patterns over time, and to model the impact of household characteristics on the macronutrient shares vector. We compute food expenditure elasticities of macronutrient shares, and we compare them to classical elasticities for macronutrient volumes and total calorie intake. The compositional model highlights the important role of many factors in the determination of diet choices and we will focus mainly on the role of food expenditure. Our results are consistent with the rest of the literature, but they have the advantage to highlight the substitution effects between macronutrients in the context of nutrition transition.

This chapter has been accepted for publication in *Statistical Methods in Medical Research* journal, 2018, online first, 21 p. `http://journals.sagepub.com/doi/10.1177/0962280218770223`

## 4.1 Introduction

Food security and nutrient affordability have become a main concern of governmental and non-profit organizations due to their effects on health and economic development. Many empirical researches focus on the relationship between socioeconomic characteristics of households and their food consumption behavior. Food consumption is measured initially by calorie, i.e food categories in quantity are converted into calorie intake. A recent meta–analysis by Ogundari and Abdulai (2013) shows that the relationship between calorie intake and income (or expenditure) have been well studied for many countries in order to implement policies which reduce starvation and nutritional deficiencies. Then, economic development and urbanization in developing countries have affected global diet, leading to many empirical studies focusing on food sources, such as vegetable, staple cereals, meat, etc. The 2017 Global Food Policy Report shows that widespread trends include an increase of animal-source foods, sugar, oils, processed food and staple cereal refining, as results of higher incomes and urbanization, IFPRI (2017). Another concern about food consumption is its composition in terms of macro and micronutrient (such as protein, fat, carbohydrate, vitamin A, zinc). Recently, a review of a total of 26 empirical studies about income elasticities of calories macronutrients and micronutrients by Santeramo and Shabnam (2015) indicates that calories intake and proteins intake are more income–inelastic than fat intake and micronutrients intake. In addition, there are only 5 over 26 empirical studies which focus on all macronutrients, i.e protein, fat and carbohydrate.

In order to assess the relationship between nutrients consumption and socioeconomic characteristics, several regressions (one by nutrient) are usually performed in parallel with the same explanatory variables and the different nutrients as dependent variables. For example, an empirical study in Greece by Liaskos and Lazaridis (2003) performs 13 multiple linear regressions which have the same household characteristics as explanatory variables and 13 different nutrients as dependent variables. Similarly, You et al. (2016) fit three specifications of health production functions with the same explanatory variables, the response variables of the models being the macronutrients consumptions in protein, fat and carbohydrate in China. These specifications do not take into account the fact that the three macronutrients constitute the whole diet of each household (or individual) so the volumes of consumed macronutrients are not independent. Moreover, the computation of consumed macronutrient volume can be criticized when using household survey data due to the impossibility to take into account losses and wastes in food preservation, preparation and consumption. The percentage of losses and wastes varies from 5% to 12% across countries, Porkka et al. (2013). Household survey data have also limitations due to recalled bias and self-reported measures (Deaton (1997)). Assuming that these two pro-

blems affect the computation of the quantities of all macronutrients in the same way, we can expect the shares of the macronutrients not to be affected by the consecutive biases, contrary to volumes.

Vietnam is a good example of a middle-income country that has recorded impressive achievements in economy and population welfare after the launch of economic reforms in 1986. However, this country has also experienced a nutrition transition like many other middle-income countries. Nutrition transition has motivated many empirical works in Vietnam, Mishra and Ray (2009); Nguyen and Popkin (2004). The structure of the diet during the 1990s in Vietnam contained less and less starchy staples and more and more proteins and lipids coming from meat, fish, and other protein-rich and higher fat food items (Nguyen and Popkin (2004)). In the 1992–1993 period, the main consumed food items by the Vietnamese people were cereals, potatoes, rice, and other starches, contributing up to 85.9% of total energy intake, while calories coming from other food items were low: only 6.8% of total calories were obtained from meat, fish, tofu, and other protein-rich food items, and 2.4% from fats and oils. In the 1997–1998 period, even though the total amount of calories consumed per capita remained at about the same level as 5 years earlier, there was a remarkable increase in daily proteins and lipids consumption (4.7 points) while the consumption of rice and other starches reduced significantly (5.6 points). Recently, the National Institute of Nutrition (NIN) in Vietnam has defined the “ideal" diet balance for Vietnamese households: 14% of protein, 18% of fat and 68% of carbohydrate. NIN's goal is that 50% (resp. 75%) of Vietnamese households achieve this diet balance in 2015 (resp. 2020), Ministry of Health (2012).

The aim of this study is to contribute to this literature by analyzing the evolution of diet patterns in Vietnam, focusing on macronutrient shares in the diet, instead of macronutrient volumes. This approach allows us to take into account the dependence among macronutrients and to avoid the problem of overestimation of total calorie intake when using household survey data. We use compositional data analysis (CODA) in order to analyze and to model the relative information contained in those volumes and shares. CODA is a well–established field of statistics with diverse fields of application, such as geology or economics ( Pawlowsky-Glahn and Buccianti (2011); Pawlowsky-Glahn et al. (2015)).This method has been recently applied in medical and nutritional epidemiology studies( Dumuid et al. (2017); Leite (2016); Mert et al. (2016)). A composition is a vector of $D$ components for which the relative information is relevant (for example a vector of $D$ shares). It can be represented in the simplex space $\mathcal{S}^D$, where the simplicial geometry holds ( Pawlowsky-Glahn and Buccianti (2011)). In our study, diet components are the proportions of protein, fat and carbohydrate ($D = 3$) in the average per capita calorie intake. CODA allows analyzing the shift in protein, fat, and carbohydrate shares in diets. As far as we know, our study is the first to use CODA tools to analyze the evolution of diet patterns.

We first use descriptive tools of CODA, such as compositional biplots and ternary diagrams, to show the evolution of the three components over the years. Then, we model macronutrients composition as a function of household characteristics, using compositional regression models. We first check the quality of our estimates using various model diagnostics, and then we focus on the impact of food expenditure on the share of each macronutrient in the consumption, measuring elasticities of macronutrient shares relative to food expenditure. We also compare these shares elasticities to elasticities of the volumes of macronutrients, and to the elasticity of the total calorie intake using classical linear models. This study uses six waves of the Vietnam Household Living Standard Survey (VHLSS), from 2004 to 2014.

## 4.2 The diet pattern of Vietnamese households during a ten-year period

### 4.2.1 Data

This study uses data from the Vietnam Household Living Standard Survey, carried out in 2004, 2006, 2008, 2010, 2012 and 2014 by the General Statistics Office of Vietnam in collaboration with the World Bank. Each wave sample comprises nearly 9000 households and is nationwide representative for all the 63 Vietnamese provinces. Our analysis makes use of expenditures on food and drink items provided by VHLSS questionnaires[1]. Quantities for 56 food items, including purchased foods and self–subsidies, as well as expenditures for purchased food are recorded[2].

Conversion factors of grams into calories coming from the food composition table constructed by the Vietnam National Institute of Nutrition in 2007 are used to compute macronutrient consumption amounts (see Table 6.2 in the appendix). For each household, we compute the total calorie intake (in Kcal), and the protein and fat intakes (in gram) per day. Then, we convert for each household the quantity in grams of protein (resp. fat) into Kcal[3] by multiplying by 4 (resp. 9). Finally, using a recent methodology by Aguiar and Hurst (2013), we calculate a per capita calorie intake (namely $PCCI$), a per capita volume of calories obtained from protein (namely $V_P$), and a per capita volume of calories obtained from fat (namely $V_F$), by dividing by an equivalence scale computed for each household (these scales are household specific) as in Thi et al. (2018). As the total per capita calorie intake $PCCI$ comes from three types of macronutrients (protein, fat

[1] In 2004, 2006, 2008, household food consumption was surveyed using 12–month recall. In 2010, 2012, 2014, household food consumption was surveyed using 30–day recall.

[2] Self–subsidy, gift, donation, and present foods are estimated values.

[3] Protein contains 4 calories per gram and fat contains 9 calories per gram. The conversion of grams into Kcal is an example of perturbation $\oplus$ and this operator does not affect the variability from a compositional point of view.

and carbohydrate), the per capita calorie intake obtained from carbohydrate (namely $V_C$) is calculated as:

$$V_C = PCCI - V_P - V_F.$$

The macronutrient shares $S_P, S_F$ and $S_C$ are defined as the proportion of calories coming from protein, fat and carbohydrate:

$$S_P = \frac{V_P}{PCCI}, \quad S_F = \frac{V_F}{PCCI}, \quad S_C = 1 - S_P - S_F.$$

We also concentrate on many household socioeconomic characteristics such as food expenditure[4] ($Exp$), household location ($Urban$, $Area$), household size ($HSize$), the characteristics of the head of the household, including education ($Educ$), gender ($Gender$) and ethnicity ($Ethnic$). These explanatory variables can have a potential impact on macronutrient consumption (Nguyen and Popkin (2004); Mishra and Ray (2009)). Table 4.1 provides a description of our data.

The food expenditure has changed dramatically from 2004 to 2014. The average food expenditure in 2014 is twice its value in 2004 (see Table 4.1 and boxplots in Figure 4.1 where figures in red are the medians). We also calculate the arithmetic average of the Engel coefficient for each year which is the ratio of food expenditure over total expenditure[5]. The average Engel coefficients are quite stable from 2004 to 2014 (around 46%). The mean Engel coefficient has increased by 13% from 2008 to 2010. The difference is first caused by the 2009 year in the wake of the world crisis (Cling et al. (2010)). In addition, it may come from the fact that the survey is redesigned between 2008 and 2010 using different population and household census (Benjamin et al. (2017)).

[4]Expenditures are expressed in 2006 dollars, with 1 dollar being equal to 15,994.25 VNdong in 2006.

[5]Expenditure are regular consumptions which include education expenditures, health care expenditures, food and drink consumption on festive occasions, regular food and drink consumption, daily consumption of non-food items, annual consumption of non-food items, expenditures on durables over the past 12 months, recurrent expenditures on housing, electricity, water, and daily-life waste. We do not add the costs of production and business.

Table 4.1: VHLSS description variables.

| Variable | | Description | 2004 | 2006 | 2008 | 2010 | 2012 | 2014 |
|---|---|---|---|---|---|---|---|---|
| $N$ | | Nb of observations | 8244 | 8290 | 8333 | 8548 | 8670 | 8712 |
| $V_P$ | | Nb of calories from protein | 453.5 | 461.2 | 390.1 | 543.5 | 537.9 | 544.3 |
| | | | (150.0) | (159.5) | (116.5) | (194.4) | (216.7) | (218.6) |
| $V_F$ | | Nb of calories from fat | 476.4 | 510.5 | 443.8 | 658.5 | 664.1 | 709.1 |
| | | | (227.5) | (238.6) | (198.7) | (313.5) | (332.8) | (340.8) |
| $V_C$ | | Nb of calories from carbohydrate | 2416.5 | 2383.4 | 2047.3 | 2554.1 | 2516.7 | 2511.0 |
| | | | (744.7) | (757.1) | (578.7) | (893.7) | (1005.3) | (1031.2) |
| $S_P$ | | Share of calories from protein | 13.6% | 13.7% | 13.6% | 14.5% | 14.5% | 14.5% |
| | | | (1.9%) | (1.9%) | (2.0%) | (2.0%) | (2.0%) | (1.9%) |
| $S_F$ | | Share of calories from fat | 14.3% | 15.2% | 15.5% | 17.6% | 18.0% | 19.1% |
| | | | (5.2%) | (4.7%) | (5.5%) | (5.8%) | (6.0%) | (6.5%) |
| $S_C$ | | Share of calories from carbohydrate | 72.1% | 70.9% | 67.9% | 67.5% | 67.5% | 66.4% |
| | | | (6.2%) | (5.8%) | (6.6%) | (7.0%) | (6.9%) | (7.4%) |
| $Exp$ | | Food expenditure per year (US$) | 598.5 | 622.8 | 706.4 | 966.4 | 1032.4 | 1010.2 |
| | | | (330.8) | (348.1) | (383.8) | (554.1) | (612.4) | (597.9) |
| $ExpTot$ | | Total Expenditure per year (US$) | 1426.5 | 1541.2 | 1763.3 | 2173.1 | 2262.4 | 2303.4 |
| | | | (947.0) | (1008.5) | (1141.8) | (1398.7) | (1435.5) | (1424.3) |
| $Engel$ | | Engel coefficient | 46.0% | 44.2% | 44.0% | 49.8% | 48.1% | 46.0% |
| | | | (12.5%) | (12.2%) | (12.4%) | (11.3%) | (11.3%) | (10.9%) |
| $Urban$ | 1 | Urban | 23.34 % | 25.28 % | 25.86 % | 27.56 % | 28.54 % | 29.61 % |
| | 0 | Rural | 76.66 % | 74.72 % | 74.14 % | 72.44 % | 71.46 % | 70.39 % |
| $HSize$ | 2 | $\leq$ 2 people | 11.07 % | 12.98 % | 14.32 % | 16.34 % | 18.06 % | 19.72 % |
| | 3 | 3 people | 15.74 % | 17.13 % | 17.58 % | 20.12 % | 18.92 % | 20.02 % |
| | 4 | 4 people | 30.65 % | 31.54 % | 32.03 % | 33.29 % | 32.2 % | 30.84 % |
| | 5 | 5 people | 21.51 % | 20.21 % | 19.36 % | 16.66 % | 17.53 % | 16.41 % |
| | 6 | $\geq$ 6 people | 21.02 % | 18.14 % | 16.72 % | 13.58 % | 13.29 % | 13.01 % |
| $Ethnic$ | 1 | Kinh | 86.31 % | 86.14 % | 86.39 % | 83.26 % | 83.13 % | 83.67 % |
| | 0 | Minorities | 13.69 % | 13.86 % | 13.61 % | 16.74 % | 16.87 % | 16.33 % |
| $Gender$ | 1 | Male | 76.63 % | 75.78 % | 75.83 % | 75.98 % | 75.97 % | 75.2 % |
| | 0 | Female | 23.37 % | 24.22 % | 24.17 % | 24.02 % | 24.03 % | 24.8 % |
| $Educ$ | 1 | Below primary | 54.25 % | 52.06 % | 50.76 % | 51.1 % | 50.68 % | 49.15 % |
| | 2 | Secondary, High school | 41.47 % | 43.53 % | 44.77 % | 42.96 % | 43.62 % | 44.42 % |
| | 3 | University | 4.28 % | 4.4 % | 4.46 % | 5.94 % | 5.7 % | 6.43 % |
| $Area$ | 1 | Red River Delta | 21.57 % | 21.79 % | 22.13 % | 17.57 % | 17.26 % | 21.54 % |
| | 2 | Midlands Northern Mountains | 18.63 % | 18.23 % | 18.13 % | 13.35 % | 13.01 % | 17.3 % |
| | 3 | Northern Central Coast | 20.44 % | 20.53 % | 20.05 % | 22.18 % | 22.16 % | 22.08 % |
| | 4 | Central Highlands | 6.22 % | 6.15 % | 6.22 % | 7.07 % | 6.85 % | 6.65 % |
| | 5 | South East | 12.34 % | 12.75 % | 12.76 % | 11.39 % | 11.44 % | 11.96 % |
| | 6 | Mekong River Delta | 20.89 % | 20.49 % | 20.9 % | 28.35 % | 29.23 % | 20.51 % |

Averages correspond to arithmetic means for volume variables ($V_P, V_F, V_C, Exp, ExpTot, Engel$)
Averages correspond to closed geometric means for share variables ($S_P$, $S_F$, $S_C$).

### 4.2.2 Diet pattern of Vietnamese households during 2004-2014

The diet pattern of Vietnamese households has changed dramatically from 2004 to 2014. The volumes of macronutrient consumption along time are presented in Figure 4.2. The median volume of per capita calorie intake (in red color) has increased from 2004 to 2014, except that there is a strong fall of PCCI in 2008 due to a difficult climatic year and a very significant increase in food prices (double-digit inflation). With respect to the volume of macronutrient consumption, calories obtained from carbohydrate are quite stable across the six years (except a decrease in 2008) while calories obtained from protein and fat have increased gradually.

Figure 4.1: Food expenditure in US$. Each boxplot shows the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. The red numbers are the medians.

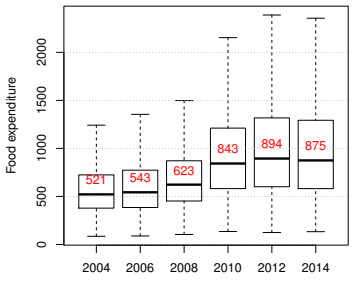
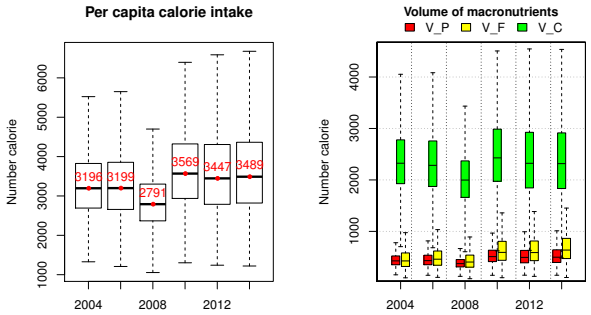



Figure 4.2: Per capita calorie intake and volume of macronutrient consumption.




Broadly speaking, during this ten-year period, the average protein share and the average fat share are between 10% and 20%, and the average carbo-

hydrate share is between 60% and 80% (see Table 4.1). Figure 4.3 represents the ternary diagrams of the share of macronutrients for the rural and urban sites. The arrows indicate the evolution over the years. Particularly, households in both type of sites tend to decrease their proportion of carbohydrate and increase their proportion of fat. The evolution of macronutrient consumption in rural and urban sites are going in the same direction. However, the starting points (in 2004) in terms of diet balance are different between rural and urban sites (see Table 4.2). Moving from ($S_P$ =13.3%, $S_F$ =12.8%, $S_C$ =73.9%) in 2004 to (14.2%, 17.6%, 68.2%) in 2014, Vietnamese rural households have increased the part of calories obtained from fat by 37.5% at the expense of calories obtained from carbohydrate while the calories obtained from protein are quite stable. In contrast, starting from (14.5%, 16.5%, 69.0%) in 2004 to (15.4%, 20.3%, 64.3%) in 2014, urban households have increased the part of calories obtained from fat by 23% at the expense of calories obtained from carbohydrate, while there is a small change in the proportion of protein (6.2%).

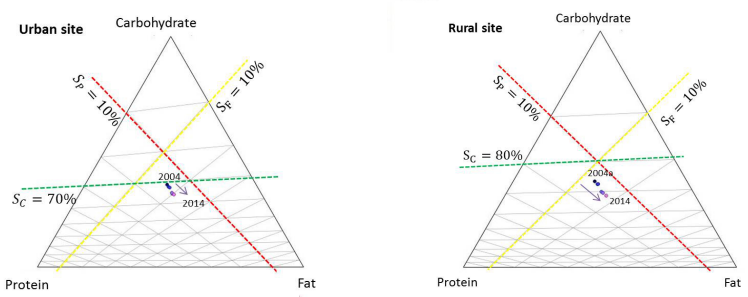
Regions in Vietnam are different in terms of socio-economic characteristics, and in terms of diet patterns. The map in Figure 4.4 shows the geometric average of macronutrient shares ($S_P, S_F, S_C$) and the arithmetic average of food expenditure ($Exp$), by region ($Area$) in 2014. Red River Delta and South East areas have the highest averages in food expenditure. They also have the largest shares of fat and protein. On the contrary, Midlands Northern Mountains and Mekong River Delta areas have the smallest values for average food expenditure. In the same line, Midlands Northern Mountains has the smallest protein share (13.4%) and Mekong River Delta has the lowest fat share (15.6%). These average macronutrient shares are similar to the results in the General Nutrition Survey 2009-2010, National Institute of Nutrition (2010). Red River Delta and South East are the two regions who have the highest food consumption of animal-based foods, eggs and milk (in kilograms of food). The General Nutrition Survey also reveals a high proportion of vegetables, such as leafy vegetables and edible flowers and tuberous vegetables for Mekong River Delta and Midlands Northern Mountains. Both our results and the General Nutrition Survey show a similar average proportion of macronutrient intake and food group consumption for the other regions.

Table 4.2: Closed geometric mean of macronutrient shares in urban and rural sites.

| | Urban site | | | Rural site | | |
|---|---|---|---|---|---|---|
| Year | $S_P$ | $S_F$ | $S_C$ | $S_P$ | $S_F$ | $S_C$ |
| 2004 | 14.5% | 16.5% | 69.0% | 13.3% | 12.8% | 73.9% |
| 2014 | 15.4% | 20.3% | 64.3% | 14.2% | 17.6% | 68.2% |

Beyond analyzing the center of the data, it is also interesting to look

Figure 4.3: Centered ternary diagrams of average macronutrient shares in urban and rural sites.

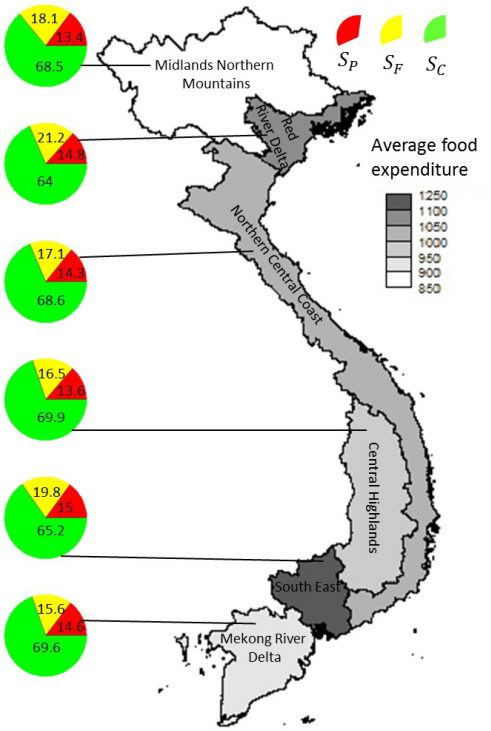



at its dispersion around this center. Figure 4.5 (left) represents in a ternary diagram the data in 2004, the data centers in 2004 and 2014, along with ellipses delimiting half of the population around these points in the simplex, Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). The "ideal" balanced diet according to the National Institute of Nutrition in Vietnam ($S_P$=14%, $S_F$=18%, $S_C$=68%) is represented by a triangle. This ternary diagram shows that half of the population in 2014 have a diet balance very close to the ideal one, closer than in 2004. In Figure 4.5 (right), the same information is represented but summarizing the three shares in two coordinates $S_1^* = \frac{1}{\sqrt{2}} \log \frac{S_F}{S_P}$ and $S_2^* = \frac{2}{\sqrt{6}} \log \frac{S_C}{\sqrt{S_F S_P}}$, which are called ILR coordinates (see next section). Due to the log–transformation, the figure in ILR coordinates reveals a larger dispersion than the figure in shares. In addition, we can see that the centers of the "very poor" and "very rich"[6] are very far from each other. In 2004, the center of the "very poor" ($S_P$ =13.0%, $S_F$ =12.1%, $S_C$ =74.9%) is far from the ideal diet point while the center of the "very rich" (15.4%, 17.8%, 66.8%) is close to the ideal diet balance. In 2014, the centers of the "very poor" and "very rich" are (13.0%, 16.8%, 69.2%) and (15.9%, 22.1%, 61.9%). Thus, the "very poor" households in 2014 still do not consume enough protein and fat, while the "very rich" households consume relatively too much fat.

Note that the information carried by a vector of $D$ shares can be summarized in $D-1$ ratios of shares, thanks to the summing up to one constraint. For example, the three macronutrient shares can be summarized in two log–ratios, $R_{CP} = \log(\frac{S_C}{S_P})$ and $R_{CF} = \log(\frac{S_C}{S_F})$. Log–ratio are preferred because their range is the whole real line. Figure 4.6 represents the dispersion of pairwise log–ratios over the years for the three log–ratios: $R_{CP}$, $R_{CF}$ and $R_{FP} = \log(\frac{S_F}{S_P})$. Looking first at the boxplots, we see that the medians of

[6]Households who have food expenditure less than 5% (217.7$) and higher than 95% quantile 1247.1$ in 2004 (resp. 304.8$ and 2165.6$ in 2014)

Figure 4.4: Macronutrient shares and food expenditure averages by area in 2014.




the log–ratios of shares $R_{CP}$ and $R_{CF}$ are larger than 1 (i.e the proportion of carbohydrate is more than twice the proportions of protein and fat). Moreover, in 2004, the median values for both $R_{CP}$ and $R_{CF}$ are quite similar, but in 2014 the median value of $R_{CF}$ is much smaller than that of $R_{CP}$. The log–ratio $R_{FP}$ has increased over the years and is larger than 0, i.e the proportion of fat is higher than the proportion of protein. The evolution shows an increase of the consumption of fat and protein at the expense of carbohydrate, and this increase is more pronounced for fat than for protein. The evolution of Vietnamese diet patterns is consistent with the global change in diets consisting of an increase in consumption of animal-source foods, fats and oils at the expense of grains and cereals, IFPRI (2017). Moreover we have added a reference line showing the value corresponding to the ideal diet for each log–ratio of share and we can see that the evolution over the years reveals a convergence to the ideal diet reference.

To give a comprehensive compositional exploratory analysis of macronutrient shares, we present a covariance biplot, often used in compositional data analysis, which represents both points and clr–variables for each year,

Figure 4.5: Plot centers in 2004 and 2014 compared to the "ideal" diet balance ($S_P$=14%,$S_F$=18%,$S_C$=68%) in ternary diagram in the simplex and in ILR coordinates.

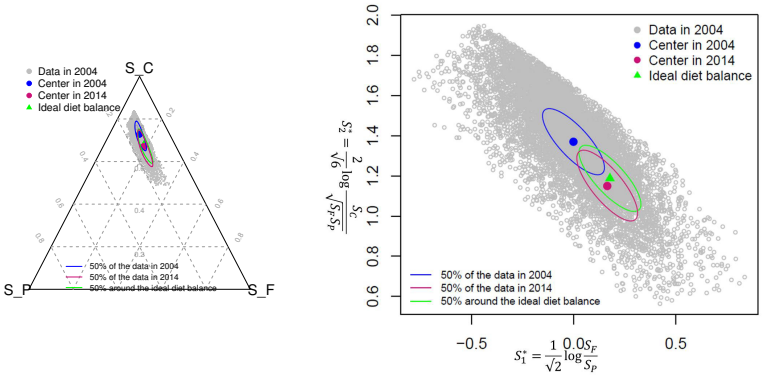
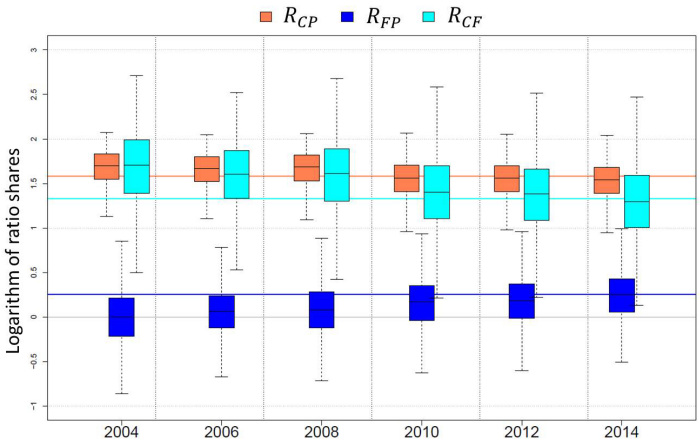
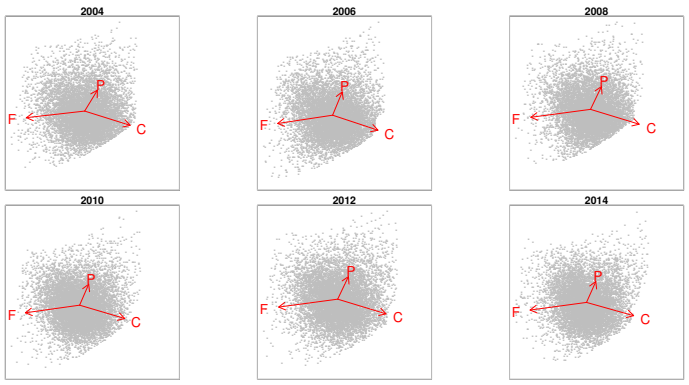



Figure 4.6: Boxplots of macronutrients log–ratio of shares by year. The line shows the value corresponding to the ideal diet for each log–ratio of share.




in Figure 4.7. Because we have here a 3–part composition, the biplot explains 100% of the variance. Interestingly, the three components point towards different directions and display very long links; moreover these trends are the same for the 6 years. The log-ratio corresponding to the longest link is that of Fat versus Carbohydrate. The Protein–Carbohydrate and Fat–Carbohyrate links appear to be orthogonal, thus revealing two possibly uncorrelated log ratios, i.e $\log(\frac{S_P}{S_C})$ and $\log(\frac{S_F}{S_C})$.

Figure 4.7: Covariance biplot of a principal component analysis of the macronutrient shares in each year. P, F, C correspond to Protein, Fat and Carbohydrate.




## 4.3 Compositional data analysis approach to describe and explain macronutrient consumption

### 4.3.1 Introduction to CODA

In the literature, different types of models are available for doing regression with shares, Morais et al. (2017). In the case where the dependent variable is a vector of shares (e.g. the composition of macronutrients) and explanatory variables are classical variables which depend only on the observations (e.g. household characteristics), a model has been proposed in the so-called CODA (compositional data analysis) literature, Aitchison (1986); Pawlowsky-Glahn and Buccianti (2011); Pawlowsky-Glahn et al. (2015). This model is very simple to implement and is based on a log-ratio transformation of shares. A composition $\mathbf{S}$ of $D$ shares can be represented in the simplex space $\mathcal{S}^D$:

$$\mathcal{S}^D = \{\mathbf{S} = (S_1, S_2, \ldots, S_D)' : S_j > 0,\ j = 1, \ldots, D;\ \sum_{j=1}^{D} S_j = 1\}.$$

In order to take into account the relative information between components and to ensure the constant sum of the fitted components (equal to 1 here), classical regression models cannot be used directly. Thus, shares are transformed, using an isometric log-ratio (ILR) transformation, Egozcue and Pawlowsky-Glahn (2003), (for example) in $D-1$ coordinates which can be represented in the classical Euclidean space so that linear regression models can be used separately on the $D-1$ coordinates. The ILR coordinates are defined as:

$$\text{ilr}(\mathbf{S}) = \mathbf{W}' \log(\mathbf{S}) = \mathbf{S}^* = (S_1^*, \ldots, S_{D-1}^*)',$$

where the $D \times (D-1)$ contrast matrix $\mathbf{W}$ allows the projection of shares onto an orthonormal basis of $\mathcal{S}^D$. For example, for $D = 3$, the following contrast matrix can be used (this is the default matrix used by the function "ilr" in the R package "compositions"):

$$\mathbf{W} = \begin{bmatrix} -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \sqrt{\frac{2}{3}} & 0 \end{bmatrix},$$

leading to the following two ILR coordinates of $\mathbf{S} = (S_1, S_2, S_3)$:

$$S_1^* = \frac{2}{\sqrt{6}} \log \frac{S_3}{\sqrt{S_2 S_1}}, \quad S_2^* = \frac{1}{\sqrt{2}} \log \frac{S_2}{S_1}.$$

In such a configuration, the first ILR coordinate $S_1^*$ contains all the relative information of $S_3$ compared to the geometric mean of the remaining shares $S_1^*$ and $S_2^*$, Muller et al. (2016)

Finally, the inverse transformation of results allows to go back to the simplex in order to interpret the model on shares. The inverse transformation is given by: $\mathbf{S} = \text{ilr}^{-1}(\mathbf{S}^*) = \mathcal{C}(\exp(\mathbf{W}\mathbf{S}^*))'$, where $\mathcal{C}(.)$ is the closure operation allowing to go from a vector of volumes $\mathbf{V}$ to a vector of shares $\mathbf{S}$: $\mathcal{C}(V_1, \ldots, V_D)' = (\frac{V_1}{\sum_{j=1}^D V_j}, \ldots, \frac{V_D}{\sum_{j=1}^D V_j})' = (S_1, \ldots, S_D)'$.
Let us introduce the following operators used in the simplex (Pawlowsky-Glahn and Buccianti (2011)): the operators $\oplus$ and $\odot$ are called perturbation operation and power transformation, and play in $\mathcal{S}^D$ a role similar to that of the operators $+$ and $\times$ in the classical Euclidean space. They are defined as follows:

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \mathcal{C}(x_1 y_1, \ldots, x_D y_D)' \qquad \text{with } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D. \\ \lambda \odot \mathbf{x} &= \mathcal{C}(x_1^\lambda, \ldots, x_D^\lambda)' \qquad \text{with } \lambda \in \mathbb{R}, \mathbf{x} \in \mathcal{S}^D. \end{aligned}$$

### 4.3.2 Compositional model for macronutrient shares

We are interested in the impact of Vietnamese household characteristics on its macronutrient composition, and the evolution of this impact across time, from 2004 to 2014. An adapted compositional regression model is the following (one model by period):

$$\begin{aligned} \mathbf{S}_i &= \mathbf{a} \bigoplus_{k=1}^{K} X_{ki} \odot \mathbf{b_k} \oplus \boldsymbol{\epsilon}_i \\ &= \mathbf{a} \oplus \log(Exp)_i \odot \mathbf{b_1} \oplus Urban_i \odot \mathbf{b_2} \oplus HSize_i \\ &\odot \mathbf{b_3} \oplus Educ_i \odot \mathbf{b_4} \oplus Ethnic_i \odot \mathbf{b_5} \\ &\oplus Gender_i \odot \mathbf{b_6} \oplus Area_i \odot \mathbf{b_7} \oplus \boldsymbol{\epsilon}_i, \end{aligned} \tag{4.1}$$

where $\mathbf{S} = (S_P, S_F, S_C)'$, and the index $i$ denotes the $i^{th}$ household. $\mathbf{S}, \mathbf{a}, \mathbf{b_k}, \boldsymbol{\epsilon} \in \mathcal{S}^D$ are compositions and $X_k$ are classical explanatory variables ($Exp$ is a positive continuous variable, used in logarithm, and others are categorical variables).

As proved in Morais et al. (2018), model (4.1) can be written in a fashion similar to the classical attraction models used in the marketing literature (Cooper and Nakanishi (1989)):

$$S_{j,i} = \frac{a_j \prod_{k=1}^{K} b_{j,k}^{X_{ki}} \epsilon_{j,i}}{\sum_{m=1}^{D} a_m \prod_{k=1}^{K} b_{m,k}^{X_{ki}} \epsilon_{m,i}}. \tag{4.2}$$

As in Dumuid et al. (2017) and Muller et al. (2016), in order to fit and interpret model (4.1), we need to run $D - 1 = 2$ ordinary linear regression models, one for each ILR coordinate of $\mathbf{S}$: $S_1^* = \frac{2}{\sqrt{6}} \log \frac{S_C}{\sqrt{S_F S_P}}$ and $S_2^* = \frac{1}{\sqrt{2}} \log \frac{S_F}{S_P}$, for each period, for $j = 1, 2$ (Egozcue et al. (2012)):

$$\begin{aligned} S_{j,i}^* &= a_j^* + \sum_{k=1}^{K} b_{j,k}^* X_{ki} + \epsilon_{j,i}^* \\ &= a_j^* + b_{j,1}^* \log(Exp)_i + b_{j,2}^* Urban_i + b_{j,3}^* HSize_i + b_{j,4}^* Educ_i \\ &+ b_{j,5}^* Ethnic_i + b_{j,6}^* Gender_i + b_{j,7}^* Area_i + \epsilon_{j,i}^*, \end{aligned} \tag{4.3}$$

where $a_j^*, b_{j,k}^*, \epsilon_j^*$ are the $j^{th}$ ILR coordinates of $\mathbf{a}, \mathbf{b_k}, \boldsymbol{\epsilon}$.

Since our VHLSS dataset includes six cross–sectional waves, we perform the two transformed models (4.3) separately for the 6 years, using OLS and the assumption that $\boldsymbol{\epsilon}^*$ follows a Gaussian distribution, that is, $\boldsymbol{\epsilon}$ follows a Gaussian distribution in the simplex.
As explained before, the estimation of the coefficients of the model in the simplex (4.1) can be obtained by inverse transformation from the estimated coefficients of the transformed model (4.3). For example, $\widehat{\mathbf{b_1}} = \mathcal{C}(\exp(\mathbf{W}\widehat{\mathbf{b}_\mathbf{1}^*}))'$, where $\widehat{\mathbf{b}_\mathbf{1}^*} = (\widehat{b}_{1,1}^*, \widehat{b}_{2,1}^*)'$.

### 4.3.3 Diagnostic model-checking

In order to determine if the above presented compositional model is reliable to explain macronutrient shares, we have to check several items.

**Significance of explanatory variables** According to the analysis of the variance of our compositional models, all household characteristics used in the model are very significant (at 1%), at all observation periods[7].

[7]Full results available upon request.

**Quality measure** The quality of compositional models can be assessed by a measure adapted to share data, called "$R^2$ based on the total variance", Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013), denoted $R^2_T$. Table 4.3 shows that our models explain around 30% of the total variability of the compositional data, but the quality of models tends to decrease over time: it could be that recently factors different from those considered explain the household diet balance.

Table 4.3: Adjusted $R^2_T$ for macronutrient shares modeling.

| | 2004 | 2006 | 2008 | 2010 | 2012 | 2014 |
|---|---|---|---|---|---|---|
| $R^2_T$ | 0.31 | 0.33 | 0.28 | 0.29 | 0.23 | 0.22 |

**Inspection of residuals** Figure 4.10 represents boxplots of share residuals by component. This figure shows that the fitted error for the share of protein is very low. Errors happen mainly in the fitting of fat and carbohydrate shares, and these two shares are more and more difficult to estimate across time. Our compositional model is based on the assumption that error terms $\boldsymbol{\epsilon}$ in (4.1) follow a "Gaussian distribution in the simplex", which is equivalent to say that error terms $\epsilon^*_j$ in (4.3) or log–ratios of error terms in $\boldsymbol{\epsilon}$ follow a Gaussian distribution. Then, we check the normality in the simplex of residuals, using QQ-plots (one by log–ratio of residuals). They show that the residuals in (4.3) are close to follow a Gaussian distribution although there is a heavy tailed distribution (see Figure 4.11 for the year 2010). Moreover, the residuals are symmetric according to the residuals log–ratios boxplots (see Figure 4.12 for the year 2010).
We thus conclude that our compositional model is relevant and reliable to explain the diet balance between calories intakes from protein, fat and carbohydrates.

### 4.3.4 Regression results

As we will see, interpretations of the results involves looking at rates of changes. For two observations $X_1$ and $X_2$ of a variable $X$, we will call "rate of change" the proportion $\frac{X_2}{X_1} - 1$: a rate of change of 1% between $X_1$ and $X_2$ meaning that $X_2 = 1.01X_1$. Therefore a positive rate of change corresponds to $X_2 > X_1$ and reversely for a negative rate of change. The first ILR component $S^*_1 = \frac{2}{\sqrt{6}} \log \frac{S_C}{\sqrt{S_F S_P}}$ corresponds to Carbohydrate versus the geometric mean of other shares and the second component $S^*_2 = \frac{1}{\sqrt{2}} \log \frac{S_F}{S_P}$ corresponds to Fat versus Protein. Table 4.4 summarizes the coefficients of the compositional model in ILR coordinates over the years whereas Table 4.7 gives the corresponding coefficients in the simplex. In general, the sign of the ILR coefficients associated to $\log(Exp)$, $Urban$, $Hsize$, $Ethnic$, $Gender$

and *Educ* are opposite for $S_1^*$ and $S_2^*$ for all years.

The interpretation of regression parameters is complex for practical purposes, Dumuid et al. (2017); Muller et al. (2016). We start by doing an interpretation in the same spirit as Muller et al. (2016), but keeping the natural logarithm. Let us imagine an increase in food expenditure of 1% for a given household. This corresponds to an additive increase of $\delta$ of the logarithm of expenditure, where $\exp(\delta) = 1.01$, yielding $\delta = \log(1.01)$. Keeping all else fixed, this would result in an increase of $\beta\delta$ in the first ILR coordinate $\frac{2}{\sqrt{6}} \log(\frac{S_C}{\sqrt{S_P S_F}})$, where $\beta = -0.265$ is the coefficient of log of food expenditure in the regression of this coordinate. Therefore this would result in the relative dominance of the share of carbohydrates with respect to the geometric average of other parts being multiplied by $\exp(\frac{\sqrt{6}}{2}\beta\delta) \simeq 0.997$, which is a decrease of 0.3%. This is consistent with the fact that larger households live in rural sites[8] and rural households have a large share of calories obtained from carbohydrate while the calories obtained from fat and protein are low. As explained in Muller et al. (2016), if we were to interpret instead the impact on the relative dominance of the share of fat with respect to the geometric average of other parts, theoretically, we would have to make a permutation of shares before running again the regression models. However, in practice, there is a matrix formulation to do that, so that you do not need to run the regression models again.

[8]It was especially true at the beginning of the period: in 2004, 80% of the household made of 5 people and more were living in rural sites, whereas in 2014 it was 73% (77% on average on the period).

Table 4.4: Coefficients of the compositional regression model in ILR coordinates.

| Estimator | Description | 2004 | 2006 | 2008 | 2010 | 2012 | 2014 |
|---|---|---|---|---|---|---|---|
| $S_1^* = \frac{2}{\sqrt{6}} \log \frac{S_C}{\sqrt{S_F S_P}}$ **(Carbohydrate against other shares) is outcome variable** | | | | | | | |
| $(Intercept)$ | | 2.722 *** | 2.561 *** | 2.505 *** | 2.369 *** | 2.269 *** | 2.136 *** |
| $log(Exp)$ | Log of food expenditure per year (US$) | -0.265 *** | -0.241 *** | -0.232 *** | -0.214 *** | -0.194 *** | -0.182 *** |
| $Urban$ | Rural | 0.064 *** | 0.069 *** | 0.093 *** | 0.072 *** | 0.045 *** | 0.037 *** |
| $HSize$ | 3 people | 0.178 *** | 0.142 *** | 0.152 *** | 0.135 *** | 0.119 *** | 0.115 *** |
| | 4 people | 0.25 *** | 0.212 *** | 0.232 *** | 0.2 *** | 0.187 *** | 0.16 *** |
| | 5 people | 0.32 *** | 0.281 *** | 0.301 *** | 0.271 *** | 0.24 *** | 0.212 *** |
| | $\geq$ 6 people | 0.423 *** | 0.36 *** | 0.384 *** | 0.345 *** | 0.305 *** | 0.27 *** |
| $Ethnic$ | Minorities | 0.067 *** | 0.05 *** | 0.061 *** | 0.049 *** | 0.053 *** | 0.069 *** |
| $Gender$ | Female | -0.02 *** | -0.027 *** | -0.025 *** | -0.028 *** | -0.035 *** | -0.031 *** |
| $Educ$ | Secondary, High school | -0.024 *** | -0.019 *** | -0.018 *** | -0.033 *** | -0.024 *** | -0.014 * |
| | University | -0.071 *** | -0.047 *** | -0.063 *** | -0.061 *** | -0.063 *** | -0.035 ** |
| $Area$ | Midlands Northern Mountains | 0.001 | 0.011 . | 0.03 *** | 0.038 *** | 0.042 *** | 0.055 *** |
| | Northern Central Coast | 0.02 ** | 0.048 *** | 0.033 *** | 0.076 *** | 0.098 *** | 0.129 *** |
| | Central Highlands | 0.011 | 0.05 *** | 0.042 *** | 0.096 *** | 0.095 *** | 0.128 *** |
| | South East | -0.02 * | 0.009 | -0.007 | 0.025 ** | 0.036 *** | 0.048 *** |
| | Mekong River Delta | 0.014 * | 0.044 *** | 0.057 *** | 0.064 *** | 0.061 *** | 0.142 *** |
| $S_2^* = \frac{1}{\sqrt{2}} \log \frac{S_F}{S_P}$ **(Fat against Protein) is outcome variable** | | | | | | | |
| $(Intercept)$ | | -0.719 *** | -0.524 *** | -0.276 *** | -0.455 *** | -0.38 *** | -0.139 *** |
| $log(Exp)$ | Log of food expenditure per year (US$) | 0.147 *** | 0.117 *** | 0.079 *** | 0.105 *** | 0.091 *** | 0.061 *** |
| $Urban$ | Rural | -0.04 *** | -0.034 *** | -0.057 *** | -0.04 *** | -0.015 ** | -0.011 * |
| $HSize$ | 3 people | -0.1 *** | -0.07 *** | -0.061 *** | -0.066 *** | -0.047 *** | -0.033 *** |
| | 4 people | -0.137 *** | -0.102 *** | -0.105 *** | -0.089 *** | -0.074 *** | -0.038 *** |
| | 5 people | -0.174 *** | -0.144 *** | -0.145 *** | -0.129 *** | -0.097 *** | -0.058 *** |
| | $\geq$ 6 people | -0.244 *** | -0.184 *** | -0.195 *** | -0.175 *** | -0.136 *** | -0.086 *** |
| $Ethnic$ | Minorities | -0.039 *** | -0.026 *** | -0.024 ** | 0.017 ** | 0.014 * | -0.032 *** |
| $Gender$ | Female | 0.015 ** | 0.023 *** | 0.017 ** | 0.021 *** | 0.026 *** | 0.023 *** |
| $Educ$ | Secondary, High school | 0.035 *** | 0.028 *** | 0.026 *** | 0.042 *** | 0.045 *** | 0.023 *** |
| | University | 0.058 *** | 0.035 *** | 0.042 *** | 0.045 *** | 0.067 *** | 0.028 ** |
| $Area$ | Midlands Northern Mountains | 0.015 . | 0.009 | 0.009 | -0.017 * | -0.015 . | 0.002 |
| | Northern Central Coast | -0.055 *** | -0.077 *** | -0.051 *** | -0.079 *** | -0.104 *** | -0.11 *** |
| | Central Highlands | -0.005 | -0.042 *** | -0.007 | -0.069 *** | -0.077 *** | -0.088 *** |
| | South East | -0.053 *** | -0.072 *** | -0.029 *** | -0.032 *** | -0.047 *** | -0.056 *** |
| | Mekong River Delta | -0.125 *** | -0.145 *** | -0.134 *** | -0.103 *** | -0.104 *** | -0.173 *** |

## 4.4 Food expenditure elasticity of macronutrient consumption shares and volumes

### 4.4.1 Elasticities computation in compositional models

In order to interpret share models, elasticity is often a more adapted tool to overcome complex interpretations of parameters in ILR regressions. The elasticity of a dependent variable $Y$ with respect to an explanatory variable $X$ measures the rate of change between two values of the dependent variable $Y$ corresponding to an infinitesimal rate of change in $X$. This corresponds to the following formula:

$$Elast(Y, X) = \frac{\frac{\partial Y}{Y}}{\frac{\partial X}{X}} = \frac{\partial \log Y}{\partial \log X}. \tag{4.4}$$

From equation (4.2) and Morais et al. (2018), we derive the elasticity of the consumption share $S_j$ of household $i$ with respect to $\log(Exp)$, and then with the chain rule the following elasticity of the consumption share $S_{j,i}$ with respect to $Exp$ as follows for household $i$:

$$Elast(S_{j,i}, Exp_i) = \log b_{j,1} - \sum_{m=1}^{D} S_{m,i} \log b_{m,1}, \tag{4.5}$$

where $b_{j,1}$ are the coefficients associated to $\log(Exp)$ for each macronutrient $S_j$ in the simplex, and not in the coordinate space.

### 4.4.2 Elasticity of macronutrient shares

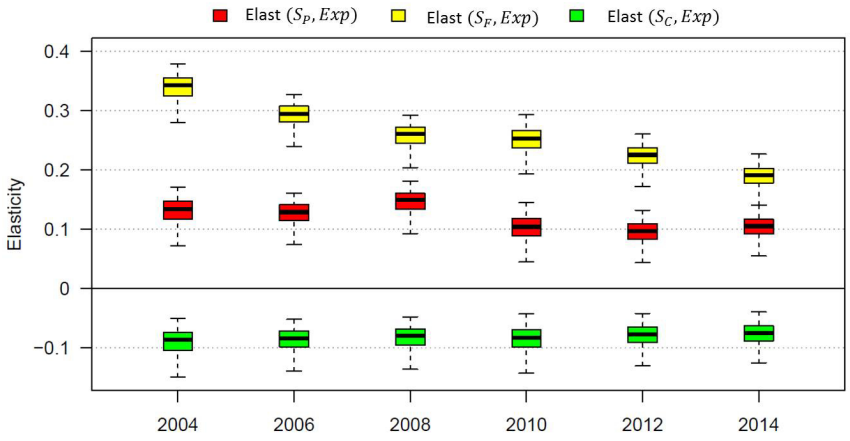
For applications to medicine, it is interesting to recall the following relationship between elasticities and odds ratios, due to the fact that odds ratios are ratios of share, Morais et al. (2018). For a small rate of change $\delta$ between two values of an explanatory variable, the odds ratio $OR$ between two components $S_j$ and $S_k$ of the share vector $S$ is related to the corresponding elasticity by $Elast(S_j/S_k, X) \simeq \frac{OR-1}{\delta}$.

Elasticities of macronutrient shares relative to the household food expenditure are presented in the boxplots in Figure 4.8, and are summarized in Table 4.5, for all observation periods. We can see that the fat share is the most elastic macronutrient with respect to food expenditure: in 2004, the food expenditure was quite low compared to the rest of the periods, and at that time, a positive rate of change of 1% of the food expenditure between households corresponds on average to a positive rate of change of 0.34% in the shares of fat in the total caloric intake, of 0.13% in the shares of protein whereas it corresponds to a negative rate of change of 0.09% in the share of carbohydrate. Let us notice that carbohydrate elasticities are negative at

all periods: it could correspond to the fact that households increasing their food expenditure tend to substitute fat and protein to carbohydrates.

To give an example of interpretation of elasticity, let us consider for example a household in 2014 having an average diet balance, i.e (14.5%, 19.1%, 66.4%) for protein, fat and carbohydrate, and a food budget of US$1000. The corresponding elasticities are (0.1031, 0.1890, -0.0769) thus if we imagine a rate of change of US$50 (an increase of 5%) for this household (all else being equal), it would correspond to a new diet balance of (14.6%, 19.3%, 66.1%). We see that this interpretation allows to directly measure the impact of a change in an explanatory variable on the whole vector of shares rather than on some complex ratios measuring the dominance of one share with respect to the other ones. Note that the elasticity of the share of fat decreases across time, whereas we know that the food expenditure tends to progress (on average from US$599 in 2004 to US$1010 in 2014). This means that for low food budget households, an increase in food expenditure tends to benefit much more to fat consumption than for high food budget households.

Figure 4.8: Boxplot of food expenditure elasticities of macronutrient consumption shares. Boxplot in red (resp. green, yellow) represents the food expenditure elasticities of protein shares (resp. carbohydrate, fat).




### 4.4.3 Elasticity of macronutrient volumes

In order to compare these results with the existing literature, we also perform the usual double-log regression models explaining the consumption volume of each macronutrient and of the total calorie intake ($PCCI$) by the same household characteristics than in model (4.1) (one model by macronutrient

and one for the total, estimated separately by OLS):

$$\begin{aligned} \log(V_{j,i}) &= \alpha_j + \beta_{j,1}\log(Exp_i) + \sum_{k=2}^{K}\beta_{j,k}X_{ki} + \varepsilon_{j,i} \qquad \text{for } j = 1,2,3 \\ \log(PCCI_i) &= \alpha + \beta_1\log(Exp_i) + \sum_{k=2}^{K}\beta_k X_{ki} + \varepsilon_i. \end{aligned} \tag{4.6}$$

Then, the elasticities of macronutrient volumes relative to food expenditure are equal to:

$$Elast(V_{j,i}, Exp_i) = \frac{\frac{\partial V_{j,i}}{V_{j,i}}}{\frac{\partial Exp_i}{Exp_i}} = \frac{\partial \log V_{j,i}}{\partial \log Exp_i} = \beta_{j,1},$$

and the elasticity of the total calorie intake relative to food expenditure is equal to $\beta_1$. Note that for double-log regression models, the elasticity is a constant term which does not depend on the considered household $i$, whereas the elasticity of the macronutrient share $S_j$ for household $i$ depends on all $S_{m,i}, m = 1,\ldots,D$ (on the full composition of macronutrient shares), that is on the diet balance of household $i$.

In this application, estimated coefficients $\hat{\beta}_{j,1}$ and $\hat{\beta}_1$ are all significantly different from zero at 0.1%, at all periods, meaning that the food budget has a real impact on the consumption of macronutrients and on the total calorie intake. Figure 4.9 represents the volume elasticities relative to the food expenditure across time. Table 4.5 compares elasticities obtained from the share model (4.1) and the volume model (4.6). All elasticities are positive for macronutrient volumes, meaning that a positive rate of change of food budget results in a positive rate of change in all types of caloric intakes. This is consistent with the fact that the food expenditure elasticities of $PCCI$ are positive and significant too. However, as for the study of macronutrient shares, we conclude that fat is the more elastic macronutrient and carbohydrate is the less elastic macronutrient to the food budget. If the food expenditure of two households differ in percentage by 1%, the calories coming from fat differ in percentage by 0.62% in 2004 and by 0.53% in 2014 on average. Our results are consistent with those of previous studies (Liaskos and Lazaridis (2003)).

Note that the log of food expenditure is very significant (P-value $< 2e-16$) for all macronutrients and all periods. The quality measures ($R^2$) of models relative to the volumes of macronutrient consumption in Table 4.6 indicate that the volume of carbohydrate is the most complicated to estimate using household characteristics. In contrast, fat and protein consumptions are well determined by the household characteristics we are using.

Figure 4.9: Food expenditure elasticities of macronutrient volumes and PCCI.


Table 4.5: Food expenditure elasticities of macronutrients shares and volumes.

| | Protein | | Fat | | Carbohydrates | | PCCI |
|---|---|---|---|---|---|---|---|
| Year | Share | Volume | Share | Volume | Share | Volume | Volume |
| 2004 | 0.1296 | 0.4071 | 0.3377 | 0.6152 | -0.0911 | 0.1863 | 0.2795 |
| 2006 | 0.1261 | 0.4063 | 0.2921 | 0.5723 | -0.0866 | 0.1936 | 0.2813 |
| 2008 | 0.1450 | 0.5123 | 0.2564 | 0.6237 | -0.0836 | 0.2837 | 0.3703 |
| 2010 | 0.1011 | 0.4023 | 0.2494 | 0.5507 | -0.0862 | 0.2150 | 0.3003 |
| 2012 | 0.0946 | 0.3807 | 0.2227 | 0.5088 | -0.0795 | 0.2067 | 0.2848 |
| 2014 | 0.1031 | 0.4437 | 0.1890 | 0.5296 | -0.0769 | 0.2637 | 0.3400 |

* Average in the case of shares

Table 4.6: Adjusted $R^2$ for macronutrient volume models.

| | 2004 | 2006 | 2008 | 2010 | 2012 | 2014 |
|---|---|---|---|---|---|---|
| Protein | 0.36 | 0.32 | 0.52 | 0.31 | 0.30 | 0.39 |
| Fat | 0.46 | 0.41 | 0.48 | 0.39 | 0.38 | 0.42 |
| Carbohydrate | 0.10 | 0.09 | 0.20 | 0.11 | 0.09 | 0.14 |
| PCCI | 0.19 | 0.17 | 0.33 | 0.19 | 0.18 | 0.25 |

## 4.5 Conclusion and discussion

This paper analyzes the evolution of diet patterns in terms of macronutrients (protein, fat and carbohydrate) and the impact of socioeconomic factors on diet balance in Vietnam, using six waves of the VHLSS data, from 2004 to 2014.

In the existing literature, food consumption is usually analyzed in terms of nutrient volumes, leading to biases due to the over-declaration of households in survey data, to the failure to account for waste, and to ignoring the dependence between the different macronutrients consumption. In order to avoid these problems, we propose to focus on the diet balance in terms of macronutrient shares in the total consumption. We use the compositional data analysis (CODA) tools and regression models to highlight the nutrition transition and to explain it according to household characteristics.

The compositional analysis reveals that the share of fat, which was almost equal to the share of protein at the beginning of the period (around 14%), increases a lot at the expense of the carbohydrate share. Even though the focus of this paper is more on the effect of food expenditure in the determination of diet choices, the compositional model highlights the important role of many household socioeconomic characteristics such as food expenditure ($Exp$), household location ($Urban$, $Area$), household size ($HSize$), the characteristics of the head of the household, including education ($Educ$), gender ($Gender$) and ethnicity ($Ethnic$).

For example, the larger the household is, the lower the fat share tends to be. Concerning the role of food expenditure, elasticities of macronutrient shares have been computed and compared to classical elasticities for macronutrient volumes and total calorie intake. Our results are consistent with the existing literature: the fat is the most elastic macronutrient (in a positive way) to the food expenditure, but this elasticity tends to slowly decrease over time (from 0.34 to 0.19 on average from 2004 to 2014). The carbohydrate share is negatively elastic to food expenditure (between -0.09 and -0.08). This reflects the substitution effects in a context of nutrition transition. Moreover, the positive elasticities of the three macronutrient volumes capture the positive impact of food expenditure on the total calorie intake of households.

This research contributes to important findings in the literature about the evolution of diets at the country level. As nutrition transition is well–known to be correlated with the rise of non-communicable diseases like obesity and heart disease national policies are needed to encourage Vietnamese people to improve their diet balance in terms of macronutrients (Bloom et al. (2012)). Indeed, policies should be targeted toward different groups. For example, they should tend to encourage "very poor" households to consume a higher share of fat and protein, and "very rich" households to stabilize their fat share in order to limit the risk of obesity. A limitation of our study comes from the fact that our data does not allow to distinguish between different types of fat. With adequate data, the same methodology could be applied taking into account the different types of fat.

In further research, similar empirical studies about macronutrients shares in the diet can be done for other countries in order to design a whole picture about food consumption composition. Moreover, it could be interes-

ting to focus on the relationship between macronutrients shares and non-communicable diseases as obesity at the country level.

Table 4.7: Coefficients of the compositional regression model in the simplex.

| Estimator | Description | 2004 | | | 2006 | | | 2008 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $S_P$ | $S_F$ | $S_C$ | $S_P$ | $S_F$ | $S_C$ | $S_P$ | $S_F$ | $S_C$ |
| (Intercept) | | 0.06 | 0.02 | 0.92 | 0.06 | 0.03 | 0.91 | 0.05 | 0.04 | 0.91 |
| $log(Exp)$ | Log of food expend. | 0.33 | 0.41 | 0.26 | 0.33 | 0.39 | 0.27 | 0.34 | 0.38 | 0.27 |
| $Urban$ | Rural | 0.33 | 0.31 | 0.35 | 0.33 | 0.32 | 0.35 | 0.33 | 0.31 | 0.36 |
| $HSize$ | 3 people | 0.33 | 0.29 | 0.38 | 0.33 | 0.30 | 0.37 | 0.33 | 0.30 | 0.38 |
| | 4 people | 0.33 | 0.27 | 0.40 | 0.33 | 0.28 | 0.39 | 0.32 | 0.28 | 0.40 |
| | 5 people | 0.32 | 0.25 | 0.42 | 0.32 | 0.26 | 0.41 | 0.32 | 0.26 | 0.42 |
| | $\geq$ 6 people | 0.32 | 0.23 | 0.45 | 0.32 | 0.25 | 0.44 | 0.32 | 0.24 | 0.44 |
| $Ethnic$ | Minorities | 0.33 | 0.32 | 0.35 | 0.33 | 0.32 | 0.35 | 0.33 | 0.32 | 0.35 |
| $Gender$ | Female | 0.34 | 0.33 | 0.33 | 0.34 | 0.33 | 0.33 | 0.34 | 0.33 | 0.33 |
| $Educ$ | Second-high school | 0.33 | 0.34 | 0.33 | 0.33 | 0.34 | 0.33 | 0.33 | 0.34 | 0.33 |
| | University | 0.33 | 0.36 | 0.31 | 0.33 | 0.35 | 0.32 | 0.33 | 0.35 | 0.32 |
| $Area$ | Mid-North Mountains | 0.33 | 0.34 | 0.33 | 0.33 | 0.33 | 0.34 | 0.33 | 0.33 | 0.34 |
| | North-Central Coast | 0.34 | 0.32 | 0.34 | 0.34 | 0.31 | 0.35 | 0.34 | 0.32 | 0.34 |
| | Central Highlands | 0.33 | 0.33 | 0.34 | 0.34 | 0.32 | 0.35 | 0.33 | 0.33 | 0.34 |
| | South East | 0.35 | 0.32 | 0.33 | 0.35 | 0.32 | 0.34 | 0.34 | 0.33 | 0.33 |
| | Mekong River Delta | 0.36 | 0.30 | 0.34 | 0.36 | 0.29 | 0.34 | 0.36 | 0.29 | 0.35 |
| | | 2010 | | | 2012 | | | 2014 | | |
| | | $S_P$ | $S_F$ | $S_C$ | $S_P$ | $S_F$ | $S_C$ | $S_P$ | $S_F$ | $S_C$ |
| (Intercept) | | 0.06 | 0.03 | 0.91 | 0.07 | 0.04 | 0.88 | 0.07 | 0.06 | 0.87 |
| $log(Exp)$ | Log of food expend. | 0.34 | 0.39 | 0.27 | 0.34 | 0.38 | 0.28 | 0.34 | 0.37 | 0.29 |
| $Urban$ | Rural | 0.33 | 0.31 | 0.35 | 0.33 | 0.32 | 0.35 | 0.33 | 0.33 | 0.34 |
| $HSize$ | 3 people | 0.33 | 0.30 | 0.37 | 0.33 | 0.31 | 0.37 | 0.32 | 0.31 | 0.37 |
| | 4 people | 0.33 | 0.28 | 0.39 | 0.32 | 0.29 | 0.39 | 0.32 | 0.30 | 0.38 |
| | 5 people | 0.32 | 0.27 | 0.41 | 0.32 | 0.28 | 0.40 | 0.32 | 0.29 | 0.39 |
| | $\geq$ 6 people | 0.32 | 0.25 | 0.43 | 0.32 | 0.26 | 0.42 | 0.31 | 0.28 | 0.41 |
| $Ethnic$ | Minorities | 0.32 | 0.33 | 0.35 | 0.32 | 0.33 | 0.35 | 0.33 | 0.32 | 0.35 |
| $Gender$ | Female | 0.33 | 0.34 | 0.33 | 0.33 | 0.34 | 0.32 | 0.33 | 0.34 | 0.33 |
| $Educ$ | Second-high school | 0.33 | 0.35 | 0.32 | 0.33 | 0.35 | 0.33 | 0.33 | 0.34 | 0.33 |
| | University | 0.33 | 0.35 | 0.32 | 0.33 | 0.36 | 0.32 | 0.33 | 0.34 | 0.32 |
| $Area$ | Mid-North Mountains | 0.33 | 0.33 | 0.34 | 0.33 | 0.32 | 0.34 | 0.33 | 0.33 | 0.35 |
| | North-Central Coast | 0.34 | 0.30 | 0.35 | 0.34 | 0.30 | 0.36 | 0.34 | 0.29 | 0.37 |
| | Central Highlands | 0.34 | 0.31 | 0.36 | 0.34 | 0.30 | 0.36 | 0.34 | 0.30 | 0.37 |
| | South East | 0.34 | 0.32 | 0.34 | 0.34 | 0.32 | 0.34 | 0.34 | 0.31 | 0.35 |
| | Mekong River Delta | 0.35 | 0.30 | 0.35 | 0.35 | 0.30 | 0.35 | 0.35 | 0.28 | 0.37 |

Figure 4.10: Boxplots of values of residuals by component and year.

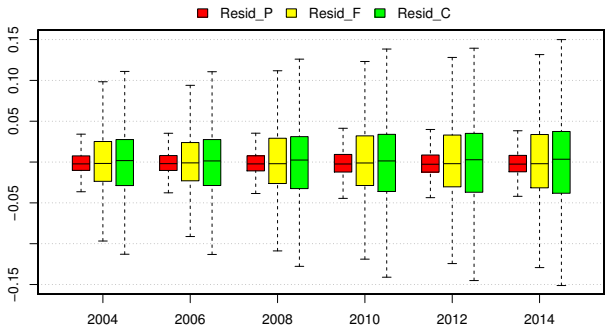
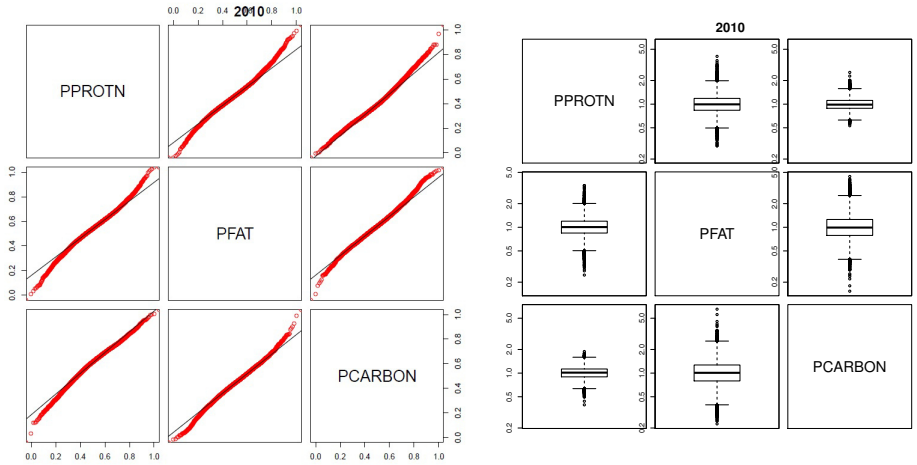



Figure 4.11: QQ-plot of residuals log ratios in 2010.

Figure 4.12: Boxplots of residuals log ratios in 2010.



# Chapter 5

# Macronutrient balances and body mass index: A new insight using compositional data analysis with a total at various quantile orders

The impact of food consumption on diseases is complex due to the confounding effects between macronutrients on a diet. We are interested in the impact of both the volume and the proportions of macronutrients on body mass index. We develop a compositional regression model with a total at various quantile orders. Then we compute the elasticities of BMI with respect to each macronutrient. Our methodology is applied to Vietnamese adults from 18 to 60 years of age. The results first reveal significant impacts of some socio–economics factors, such as the total as geometric mean, age, gender, job type, no drinking status and geographical region. All elasticities of BMI with respect to each macronutrient increase as BMI increases until a threshold (BMI=20) and then remain stable.

This chapter will be submitted to *TSE Working Paper*, May 2018

## 5.1 Introduction

The Nutrition transition has occurred in both developing countries and developed countries (Popkin (2006)). There is an increase of the double burden of malnutrition characterized by the coexistence of undernutrition along with overweight and obesity, called diet-related noncommunicable diseases (Organization et al., 2016). The World Health Organization (WHO, 2018) declares that "the fundamental cause of obesity is an energy imbalance be-

tween calories consumed and calories expended". Researchers from several disciplines have focused on the relationship between diet composition and disease (see some review in Hooper et al., 2001; Riera-Crichton and Tefft, 2014; Hall et al., 2011; Albar et al., 2014). These findings remain controversial due to the complex associations between total energy intake, physical activity, body size and the prevalence of disease and due to limitations of the datasets.

Total energy intake consists of macronutrient and micronutrient and each specific nutrient is correlated with the total energy intake: i.e each nutrient provides directly a part of the energy intake. A person who has a larger total energy intake also consumes larger volumes of all specific nutrients, on average. In addition, the contribution of each macronutrient in a total energy intake (measured by kcal) may have a different effect. Several empirical studies show the impact of a diet with the same amount of caloric content but different compositions of macronutrients on health (for example, Camacho and Ruppel, 2017). In the US, "Dietary Guidelines for Americans", issued by the US Department of Agriculture and the US Department of Health, Education and Welfare (now the Department of Health and Human Services) in 1980, recommended a reduction in the consumption of the share of total macronutrients attributable to fat and saturated fat, and a reduction in the absolute consumption of cholesterol. To compensate, the guidelines recommended increasing the share (in grams) of carbohydrates in the total consumption of calories because carbohydrates contain less than half the number of calories per ounce than fats (Cohen et al., 2015).

From a mathematical perspective, to control confounding in epidemiologic analysis, Wacholder et al. (1994), Willett et al. (1997), Trichopoulou et al. (2002), and Randi et al. (2007) have discussed various methods of adjustment for total energy intake, such as: nutrient density model, standard multivariate model, nutrient residual (energy-adjusted) model, partition regression. However, "the specific effects of individual macronutrients and the generic effect of energy cannot be disentangled by multivariate analysis" (Wacholder et al., 1994). All of the above regression models still fail to solve the comprehensive effect of total energy from that of each component of energy, i.e protein, fat, and carbohydrate. We recall the compositional nature of the dietary intake in Kcal, i.e total energy = energy from protein + energy from fat + energy from carbohydrate. Thus the four variables: total energy, energy from protein, energy from fat and energy from carbohydrate are perfectly linearly related. Recently, Leite (2016), Dumuid et al. (2017) and Trinh et al. (2018) propose to use a compositional data approach (CoDa) to analyze dietary data and show its advantages over the usual methods. Leite and Prinelli (2017) has applied this approach to analyze the associations between macronutrient balances and diseases. This study, conducted in 1992–1993 from the database of the Italian Bollate Eye study, focuses on adults of between 40 and 70 years of age. The authors discuss a

diet which consists of three macronutrients and then go into further detail by widening the composition, including now: saturated versus unsaturated fats.

Our empirical study focuses on Vietnam. This country has experienced a strong economic development after Doi Moi reforms in the 1980s. Now, Vietnam is a lower middle-income country. Due to an increase in income and changes in other socioeconomic characteristics, there is an increase in per capita calorie intake (Thi et al., 2018). In addition, the Vietnamese diet patterns have also changed with a larger proportion of animal source, fat and protein intake (Nguyen and Popkin, 2004; Trinh et al., 2018). However, Vietnam still faces the double burden of malnutrition as many developed countries. According the the United Nations, Vietnam ranks always among the thirty-six countries with the highest stunting rates in the world. Among Vietnamese 18-65 years old, the prevalence of overweight and obesity increased from 2.0% in 1992 to 5.2% in 2002 using a national survey (Tuan et al., 2008). Similarly, Nguyen and Hoang (2018) show that the prevalence of overweight and obesity increased from 2.3% in 1993 to 15% in 2015 in the same age group. The figures in urban sites are much higher than in rural sites. Cuong et al. (2007) show that 26.2% (resp. 6.4%) of adults living in the urban area of Ho Chi Minh City[1] were already considered as overweight (resp. obese) in 2004. Prevalence of obese among children under 5 has increased much faster than among adults. In the 2000-2010 period, the prevalence of overweight and obesity increased from 0.6% (resp. 0.9%, 0.5%) to 5.6% in the whole country (resp. in urban areas, in rural ones). In 2011, 14% of children (resp. 8.6%, 4.4% ) in Vietnam under 5 were still stunted (resp. underweight, thin). In addition, both figures for children under 5 are higher in big cities (Huynh et al., 2007).

This paper contributes to the literature by focusing on the impact of the macronutrient diet and socio-economics characteristics, such as age, gender, job, and living location, on the body mass index (BMI) of 18–65 years old adults by using the 2009 - 2010 wave of the General Nutrition Survey in Vietnam. We contribute to the literature in various ways:

- We apply CoDa regression with a total variable to take into account both the relative importance of each macronutrient in the whole diet and total energy.

- We perform regression both for the average BMI to obtain a general relationship and for the 15% and 90% conditional quantiles of BMI in order to be more precise for vulnerable groups. These limits correspond to underweight and overweight thresholds in the marginal distribution of BMI.

[1]This is the biggest city in Vietnam

- We adapt semi-elasticity computations in the above two regression models to obtain a direct interpretation of a change of the volume of a given macronutrient on BMI.

## 5.2 Descriptive analysis of the nutrition issue of adults aged 18–60 years old in Vietnam using compositional data analysis

We use the General Nutrition Survey 2009 - 2010 in Vietnam which was conducted by the Vietnam National Institute of Nutrition (NIN) (National Institute of Nutrition, 2010). This cross-sectional survey is representative of the Vietnamese population and has been conducted every ten years since 1981. Household dietary intake is based on a 24-hour dietary recall. Food categories in quantities are converted into calorie intake and grams using the Food composition table for Vietnam in 2007. We use the average daily intake of households. In this survey, we only focus on adults between 18 to 60 years of age. Diet intake can be divided according to macronutrient sources. From a macronutrient component perspective in terms of Kcal, we divide the diet intake into three macronutrients: protein ($P$), fat ($F$) and carbohydrates ($C$). From a macronutrient component perspective in term of grams, we divide the diet intake into four macronutrients: protein ($P$), fat ($F$), carbohydrate ($C$) and fiber[2] ($Fi$). Table 5.1 displays some summary descriptive statistics of the Vietnamese diets and their macronutrient intakes. In terms of kcal, the average per capita calorie intake (PCCI) is 1923.9 Kcal: note that this number follows the recommendation of NIN[3]. In terms of grams, per capita per day food intake is around 440 grams. In addition, the volumes of fiber are quite small compared to other macronutrients (6 (g) per person per day and it only accounts for 1.4% of total diet intake). The average total number of fiber grams is lower than in the recommendation but this number is reasonable in Vietnam due to the fact that there is only a small quantity of fiber in ordinary polished rice – the most common rice in Vietnamese meals[4].

Figure 5.1 shows the prevalence of obesity and underweight in 2010 in Vietnam, based on the cut-off of BMI classification of World Health Organization[5]. 16% of Vietnamese adults are underweight and about 7% are

[2]Fiber do not provide any calories.

[3]A household with energy intake below 1800 Kcal will be considered as a low energy intake.

[4]Ordinary polished rice has 0.4g Fiber per 100g.

[5]Body Mass Index (BMI)is defined as the weight in kilograms divided by the square of the height in meters (kg/m$^2$). According to WHO (2004), people with a BMI less than 18.49 are underweight. The normal range of BMI is 18.50 - 24.99. People with a BMI larger than 25 are overweight. In addition, people are obese if BMI is larger than 30

Table 5.1: Descriptive statistics of Vietnamese diets and their macronutrients composition

| Variable | Description | Value |
|---|---|---|
| N | Number of observations | 15035 |
| $PCCI$ | Per capita calorie intake (Kcal) | 1923.9 ( 501.8 ) |
| $PCCIg$ | Per capita per day food intake (gram) | 440 ( 114.6 ) |
| $V_P$ | Volume of calorie obtained from protein (Kcal) | 318.5 ( 98.7 ) |
| $V_F$ | Volume of calorie obtained from fat (Kcal) | 338.4 ( 177.1 ) |
| $V_C$ | Volume of calorie obtained from carbohydrate (Kcal) | 1267 ( 369.7 ) |
| $S_P$ | Share of calorie obtained from Protein (%) | 16.7 ( 3.5 ) |
| $S_F$ | Share of calorie obtained from Fat (%) | 17.4 ( 7.5 ) |
| $S_C$ | Share of calorie obtained from Carbohydrate (%) | 65.9 ( 8.8 ) |
| $Vg_P$ | Volume of intake per day from protein (gram) | 79.6 ( 24.7 ) |
| $Vg_F$ | Volume of intake per day from fat (gram) | 37.6 ( 19.7 ) |
| $Vg_C$ | Volume of intake per day from carbohydrate (gram) | 316.8 ( 92.4 ) |
| $Vg_{Fi}$ | Volume of intake per day from fiber (gram) | 6 ( 3.1 ) |
| $Sg_P$ | Share of intake per day from protein (gram) | 18.3 ( 4.1 ) |
| $Sg_F$ | Share of intake per day from fat (gram) | 8.6 ( 4.2 ) |
| $Sg_C$ | Share of intake per day from carbohydrate (gram) | 71.7 ( 7 ) |
| $Sg_{Fi}$ | Share of intake per day from fiber (gram) | 1.4 ( 0.6 ) |

Standard errors are in parenthesis

overweight. These figures are less than in developed countries but they are increasing every year.

Figure 5.1: Prevalence of obesity and underweight in Vietnam - 2010.

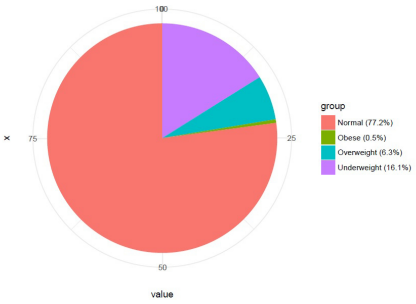
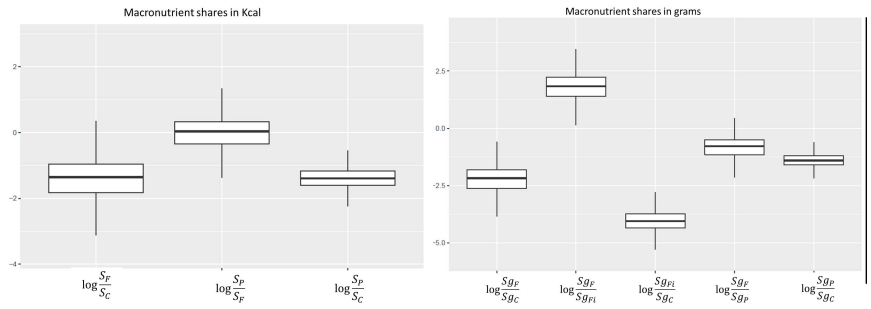



Figure 5.2 reports the ratios of macronutrient intakes expressed in logarithm, or log ratios for both Kcal and grams measurements. The first figure shows log ratios when macronutrient are measured in Kcal. The median of the two boxplots of log ratio $log(\frac{S_P}{S_C})$ and log ratio $log(\frac{S_F}{S_C})$ are negative. Carbohydrates represent the largest source of calories in the Vietnamese diet. Although having similar median values, the log ratio $log(\frac{S_P}{S_C})$ exhibits more variation than the log ratio $log(\frac{S_F}{S_C})$. The boxplot in the middle shows that the median value of log ratio $log(\frac{S_F}{S_P})$ is close to zero and that its distribution seems symmetric around zero. The right figure shows the log ratios of the four macronutrients when they are measured in grams. In the left figure, the median log ratios between protein, fat versus carbohydrate, i.e $log(\frac{Sg_P}{Sg_C})$ and $log(\frac{Sg_F}{Sg_C})$ are always negative but their absolute values are larger than when we measure macronutrient intakes in Kcal. The shares of fiber are very small compared to other macronutrients shares.

Recently, Ministry of Health (2012) has issued recommendations on the

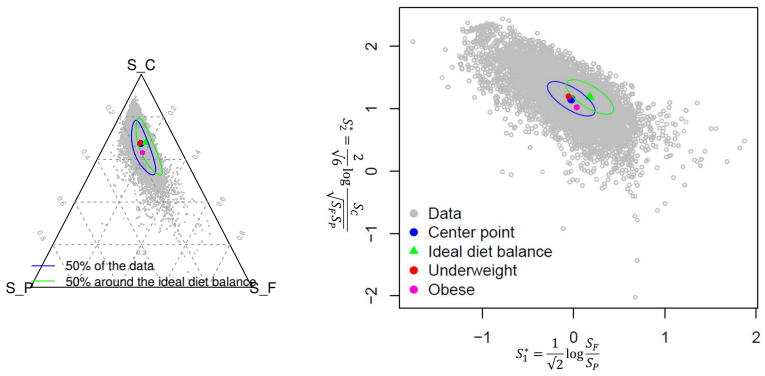
Figure 5.2: Boxplots of macronutrients log shares ratios.




ideal balanced diet for the Vietnamese population (in Kcal), namely (Protein:Fat:Carbohydrate = 14% : 18% : 68%). A ternary diagram can be used to plot this ideal diet and compare it with the observed center point of the sample. A ternary diagram is the adequate representation of shares data, incorporating information that these shares sum to one. The left panel of Figure 5.3 shows the scatterplot of the observed vectors of shares and the three center points for the whole population and for the two vulnerable groups: obese and underweight, respectively. Ellipses are added to show where half of the population is located around these center points in the simplex (Mahalanobis distance level curves). The same is done for the ideal balanced diet. The right panel of Figure 5.3 is simply a transformation of the previous one using ilr coordinates. Its lecture is easier as data are projected onto a plane (Van den Boogaart, K. G. and Tolosana-Delgado, R., 2013), but the interpretation stays the same whatever representation of the data we use... In our data, the center point is not far from the ideal point. The line passing through the two center points for underweight and obese groups is parallel to the edge $S_C$–$S_F$ of the triangle which means that underweight and obese groups have a similar proportion of protein. However, the diets of the obese group has a larger fat share (similarly, smaller carbohydrate share) than the underweight group.

Covariance biplots in Figure 5.4 show a comprehensive compositional exploratory analysis of the three macronutrient shares (in Kcal) and of the four macronutrient shares (in grams). The left biplot has a 3–part composition, the biplot explains 100% of the variance. The three components protein, fat, and carbohydrate are very long and they point towards different directions (making angles of approximately $90^0$ to $120^0$). The log-ratio corresponding to the longest link is that of Fat versus Carbohydrate. The right biplot has a 4–part components, i.e adding share of fiber. Three group links (P, F, Fi) points towards different directions as in the left biplot. In the above descriptive statistics of fiber, we see the small amount of fiber in

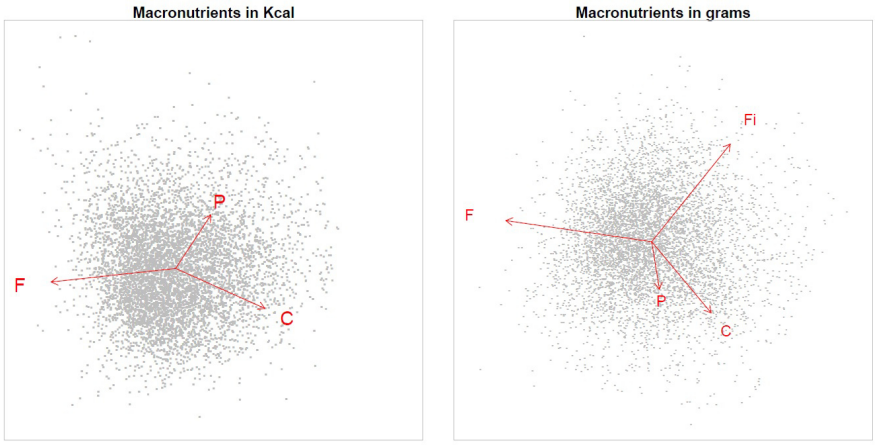
Figure 5.3: Plotof centers diets of the whole population, of he overweight people and of the obese people compared to the "ideal" diet balance ($S_P$=14%,$S_F$=18%,$S_C$=68%) in a ternary diagram in the simplex and in ILR coordinates.




the diet. But the three group links (P, F, Fi) indicate that the fiber share, although small, is very important in the diet.

The Other protein (OP) and Carbohydrate (C)links appear to be close to each other, thus revealing possibly a collinearity between (OP) and (C). The sets of rays: protein–fat and carbohydrate–fiber appear to be orthogonal, thus revealing two possibly uncorrelated log ratios, i.e $\log(\frac{Sg_P}{Sg_F})$ and $\log(\frac{Sg_{Fi}}{Sg_C})$.

Figure 5.4: Covariance biplot of a principal component analysis of the macronutrient shares for each year.



## 5.3 A compositional data perspective on studying the associations between macronutrient balances and BMI

### 5.3.1 A total as geometric mean as an determinant of obesity

As suggested byPawlowsky-Glahn et al. (2015), Coenders et al. (2017), Ferrer-Rosell and Coenders (2017), we use a total variable defined as the geometric mean of macronutrients volumes. This total corresponds to an average value in the space of the logarithm of absolute volumes values. The choice of logarithm in this total has some advantages: (1) it naturally converts an absolute positive value to a value belonging to $\mathbb{R}$, (2) it allows interpreting regression coefficients using the link between coefficients and elasticities (in economics studies) or odd ratios (epidemiologic studies).

We use the following two total variables as geometric means denoted by $T$ (resp. $Tg$) when macronutrients are measured in Kcal (resp. in grams).

$$\ln T = \frac{1}{3}\big[\ln(V_P) + \ln(V_C) + \ln(V_F)\big]$$

$$\ln Tg = \frac{1}{4}\big[\ln(Vg_P) + \ln(Vg_C) + \ln(Vg_F) + \ln(Vg_{Fi})\big]$$

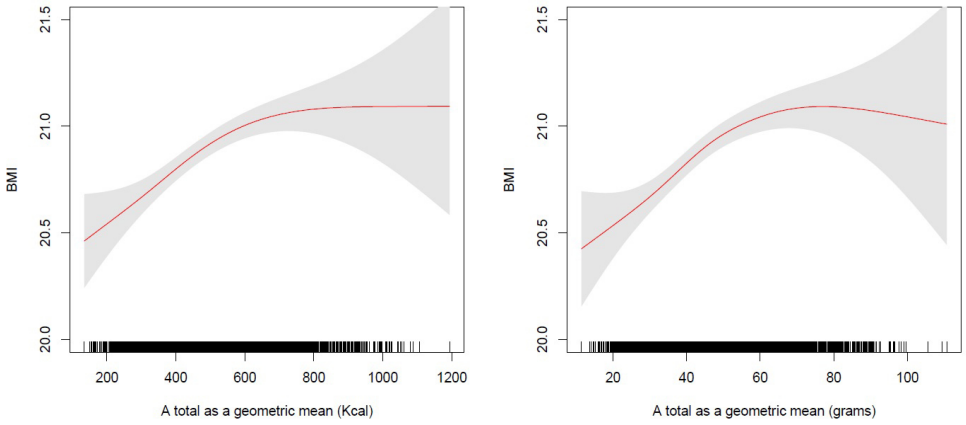
Table 5.2 shows the descriptive statistics of these two totals. In terms of Kca, the average of this geometric mean is equal to 497.5 Kcal. This number is smaller than one third of PCCI, i.e $\frac{1923.9}{3} = 641.3$ Kcal. The difference between $T$ and $\frac{PCCI}{3}$ is due to the logarithm. Similarly, an average of $Tg$ of macronutrients in grams is 46.8 (g). This number is smaller than one fourth of $PCCIg$, i.e $\frac{440}{4} = 110$ (g).

Table 5.2: Descriptive statistics for the total variables

| Variable | Description | Average value |
|---|---|---|
| T | Total in kcal | 497.5 ( 149.2 ) |
| Tg | Total in gram | 46.8 ( 14.1 ) |

Figure 5.5 shows a scatterplot of BMI and the total variable, together with a semi-parametric regression curve (Wood, 2017). There figures show a potential non-linear relationship between BMI and totals. In both figures, at the beginning of the range of totals, BMI indicators increase as totals increase. Then, when totals exceeeds a threshold, say 600Kcal and 55 grams, BMI tends to remain constant.

Figure 5.5: BMI indicator as a function of total




### 5.3.2 Various regression models with compositional predictor and a total

Compositional data describe parts of a whole and, consequently, convey only relative information. A model has been proposed in the so-called CODA (compositional data analysis) literature, which is the standard method of statistic to deal with a positive vector which carries only relative information (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015). In our approach, we are interested in the BMI indicator, denoted by $Y_i$, $Y_i \in \mathbb{R}, Y_i > 0$ and several explanatory variables. Among the explanatory variables, we will include the macronutrient shares of a diet. Due to the constant sum of the fitted components (equal to 1 here), classical regression models cannot be used directly. For example, the three macronutrient shares (in Kcal) $(S_P, S_F, S_C)$ have the following constraint $S_P + S_C + S_C = 1$. Each vector of shares $(S_P, S_F, S_C)$ belongs to the simplex $S^3$. To overcome this difficulty, shares are transformed, using an isometric log-ratio (ILR) transformation (Egozcue and Pawlowsky-Glahn, 2003). We will illustrate our strategy in the case of three macronutrient shares, a similar strategy will be applied in the case of four macronutrient shares (in grams). For three components in the simplex, the Ilr transformation transforms them into two isometric log ratios (Ilr) coordinates $Ilr1$ and $Ilr2$ that vary in $\mathbb{R}$.[6] Importantly, coefficients of compositional regression in simplex are invariant to the choice of sequential binary partition.

[6]The Ilr coordinates we are using here are based on a sequential binary partition: Carbohydrate vs protein and fat, fat vs protein. We can apply alternative sequential binary partitions, such as protein vs fat and carbohydrate, fat vs carbohydrate

$$Ilr1 = \sqrt{\frac{2}{3}} log \frac{S_C}{\sqrt{S_P S_F}}, \quad Ilr2 = \sqrt{\frac{1}{2}} log \frac{S_P}{S_F}$$

Using the ilr coordinates, a linear compositional model can be formulates to estimate the impact of some explanatory variables $Z_i$ and $(S_P, S_F, S_C)$ on the average of the outcome variable $Y_i$

$$\mathbb{E}(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + a.Z_i \qquad (EC)$$

where $\mathbb{E}(Y_i)$ denotes the expectation of the conditional distribution of $Y_i$ given the covariates. Here, $Z$ includes several explanatory variables described in Table 5.7. They are total expenditure per week (ExpWeek), age, gender, ethnicity, education levels, job (farmer or non-farmer), region[7], drinking beer status, smooking status.

The coefficients of model $(EC)$ are estimated using ordinary least squares. Examples of applications of these models in social sciences can be found in Muller et al. (2016), Leite (2016), Leite and Prinelli (2017).

The above compositional model ignores the information about total abundance of all components while focusing only on relative information between shares. In this epidemiologic study, the totals, i.e per capita calorie intake or per capita per day food intake, are also important due to their impact on BMI. Then, we adapt a compositional model including these totals, initially proposed by Pawlowsky-Glahn et al. (2015) and Coenders et al. (2017), called the T–space model. In this $T$–space model, the total is defined as in the previous subsection such that its logarithm equals to the geometric mean of the volumes.

We can then formulate a compositional model with the two ilr coordinates together with a total by

$$\mathbb{E}(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + T_i.\delta + a.Z_i \qquad (EF)$$

In addition, the classical linear model explaining $Y$ with the total only is nested in model $(EF)$ and can be used to estimate the impact of the total on the outcome variable $Y$. Thus, our "total only" regression model will be

$$\mathbb{E}(Y_i) = \alpha + T_i.\delta + a.Z_i \qquad (ET)$$

Finally, model $(EC)$, $(EF)$ and $(ET)$ can be extended to the quantile regression framework, as in Koenker and Hallock (2001). Here, we are interested in the estimation of the impact of explanatory variables $Z_i$ and $(S_P, S_F, S_C)$ on the $\tau^{th}$ conditional quantile of the outcome variable $Y_i$, so that we write

$$\mathbb{Q}_\tau(Y_i) = \alpha_\tau + \beta_\tau CIlr1 + \gamma_\tau Cilr2 + a.Z_i \qquad (QC)$$

[7]the variable region corresponds to the division of Vietnam into 5 areas : Delta, Midlands-mountainous, Low mountains, High mountains, and Coastline

$$\mathbb{Q}_\tau(Y_i) = \alpha_\tau + \beta_\tau CIlr1 + \gamma_\tau Cilr2 + T_i.\delta + a.Z_i \qquad (QF)$$

$$\mathbb{Q}_\tau(Y_i) = \alpha_\tau + T_i.\delta + a.Z_i \qquad (QT)$$

where $\mathbb{Q}_\tau$ denotes a $\tau-$ quantile level of $Y_i$ given the explanatory variables. The interpretation of the coefficients in the three quantile models are similar to that in the classical regression models $(EC)$, $(EF)$ and $(ET)$.

To obtain a comprehensive and complex impact of diet pattern on BMI, the above models are also applied to the case of four shares $Sg_F, Sg_P, Sg_C, Sg_{Fi}$.

Figure 5.8 shows the density of $log(BMI)$ which has a shape similar to a normal density. This figure supports our choice of using $log(BMI)$ as an outcome variable and shows that its distribution is approximately gaussian. To decide which quantile order to focus on, we use the cut-off for underweight (BMI is less than 18.5) and for overweight (BMI is larger than 25) which are based on BMI Asian populations (WHO, 2004). We then fit quantile regression at 15% quantile and 90% quantile levels of the marginal distribution of BMI. Table 5.3 shows all potential regression models at various quantile orders. To choose among these various models, we use an analysis of variance table (resp. an analysis of deviance table) comparing conditional mean linear models (resp. quantile regression (Koenker and Bassett, 1982). Table 5.4 shows the corresponding $F-$value and significance levels of the tests. For all these various models defined at mean or or for quantiles, results show that the full model is always preferred. This strategy is similar to Coenders et al. (2017) when choosing among alternative compositional models.

Table 5.3: Strategy to study the associations between macronutrient balances and BMI

| Shares | Macronutrient measured in Kcal<br>$S_P, S_F, S_C$ | Macronutrient measured in gram<br>$Sg_P, Sg_F, Sg_C, Sg_{Fi}$ |
|---|---|---|
| A total | $\ln T = \frac{1}{3}\left[\ln(V_P) + \ln(V_C) + \ln(V_F)\right]$ | $\ln Tg = \frac{1}{5}\left[\ln(Vg_P) + \ln(Vg_C) + \ln(Vg_F) + \ln(Vg_{Fi})\right]$ |
| Ilr coordinates | $Ilr1, Ilr2$ | $Ilrg1, Ilrg2, Ilrg3$ |
| Models<br>Conditional mean | $\mathbb{E}(Y_i) = \alpha + T_i.\delta + a.Z_i \quad (ET)$<br>$\mathbb{E}(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + a.Z_i \quad (EC)$<br>$\mathbb{E}(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + T_i.\delta + a.Z_i \quad (EF)$ | $\mathbb{E}(Y_i) = \alpha + Tg_i.\delta + a.Z_i \quad (ET)$<br>$\mathbb{E}(Y_i) = \alpha + \beta Ilrg1 + \gamma Ilrg2 + \nu Ilrg3 + a.Z_i \quad (EC)$<br>$\mathbb{E}(Y_i) = \alpha + \beta Ilrg1 + \gamma Ilrg2 + \nu Ilrg3 + Tg_i.\delta + a.Z_i \quad (EF)$ |
| $\tau$ quantile | $\mathbb{Q}_\tau(Y_i) = \alpha_\tau + T_i.\delta + a.Z_i \quad (QT)$<br>$\mathbb{Q}_\tau(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + a.Z_i \quad (QC)$<br>$\mathbb{Q}_\tau(Y_i) = \alpha + \beta Ilr1 + \gamma Ilr2 + T_i.\delta + a.Z_i \quad (QF)$ | $\mathbb{Q}_\tau(Y_i) = \alpha_\tau + Tg_i.\delta + a.Z_i \quad (QT)$<br>$\mathbb{Q}_\tau(Y_i) = \alpha + \beta Ilrg1 + \gamma Ilrg2 + \nu Ilrg3 + a.Z_i \quad (QC)$<br>$\mathbb{Q}_\tau(Y_i) = \alpha + \beta Ilrg1 + \gamma Ilrg2 + \nu Ilrg3 + Tg_i.\delta + a.Z_i \quad (QF)$ |

Table 5.4: Analysis of Variance table for alternative models

| Models | Macronutrient in Kcal | | Macronutrient in Kcal | |
|---|---|---|---|---|
| | Full vs Total | Full vs Composition | Full vs Total | Full vs Composition |
| Conditional mean | 12.76*** | 25.31*** | 9.35*** | 25.84*** |
| $\tau = 0.15$ quantile | 2.91. | 20.16*** | 2.49. | 25.78*** |
| $\tau = 0.9$ quantile | 5.03** | 9.61** | 4.23** | 8.49** |

Note: ., $^\star$, $^{\star\star}$, and $^{\star\star\star}$ mean significant at 10%, 5%, 1% and 0.1%, respectively

Muller et al. (2016), Leite (2016), Leite and Prinelli (2017) gives interpretation of the coefficients of the Ilr coordinates. We will rather adopt the same kind of interpretation as Morais et al. (2018) and adapting it to our case, i.e. using semi-elasticities in the next subsection to have direct interpretation of the impact of each macronutrient. Table 5.5 shows the coefficients of (traditional, non-compositional) explanatory variables[8] for both two kinds of food intake measures.

Table 5.5: Multiple linear regression analysis of the relationship between the first ilr coordinate and the total as geometric mean and BMI.

| Regressors | Macronutrient in Kcal | | | Macronutrient in grams | | |
|---|---|---|---|---|---|---|
| | * $\tau = 0.15$ | Mean | $\tau = 0.9$ | $\tau = 0.15$ | Mean | $\tau = 0.9$ |
| (Intercept) | 1.19*** | 1.502*** | 2.000*** | 1.154*** | 1.512*** | 2.057*** |
| A total as geometric mean | | | | | | |
| T | 0.01*** | 0.1*** | 0.01** | 0.1*** | 0.1*** | 0.1** |
| Age | | | | | | |
| log(Age) | 0.935*** | 0.797*** | 0.58*** | 0.954*** | 0.797*** | 0.551*** |
| $\log^2$(Age) | -0.125*** | -0.102*** | -0.068*** | -0.128*** | -0.102*** | -0.064*** |
| Expenditure per week | | | | | | |
| log(EXP) | $10^{-5}$ | 0.001* | 0.002 | 0.0002 | 0.002* | 0.002. |
| Gender | | | | | | |
| Female | -0.019*** | -0.017*** | -0.011* | -0.019*** | -0.018*** | -0.011* |
| Ethnicity | | | | | | |
| Kinh | -0.009* | -0.016*** | -0.023*** | -0.008. | -0.015*** | -0.024*** |
| Education levels | | | | | | |
| Secondary school | 0.003 | -0.005* | -0.01* | 0.003 | -0.005* | -0.011* |
| High school | 0.006 | 0 | -0.005 | 0.006 | 0 | -0.006 |
| Univeristy | 0.006 | -0.002 | -0.009 | 0.006 | -0.003 | -0.01 |
| Job | | | | | | |
| Non–Farmer | 0.015*** | 0.019*** | 0.023*** | 0.015*** | 0.019*** | 0.023*** |
| Smoking status | | | | | | |
| Non smoker | 0.015*** | 0.018*** | 0.017** | 0.015*** | 0.018*** | 0.016*** |
| Drinking beer status | | | | | | |
| 1–4 times per months | -0.005 | 0.002 | 0.004 | -0.004 | 0.002 | 0.003 |
| (no drinking) | -0.019*** | -0.009** | -0.003 | -0.019*** | -0.009** | -0.003 |
| Geographical region | | | | | | |
| Coastline | 0.005 | 0.005 | 0 | 0.006 | 0.006. | 0.001 |
| Midlands-mountainous | 0.005 | -0.015** | -0.034* | 0.006 | -0.015** | -0.031* |
| Low mountains | 0.015*** | -0.002 | -0.029*** | 0.015*** | -0.002 | -0.029*** |
| High mountains | 0.023*** | -0.002 | -0.033*** | 0.023*** | -0.003 | -0.034*** |

The coefficients of T are multiplied by 1000.
Note: ., *, **, and *** mean significant at 10%, 5%, 1% and 0.1%, respectively

The interpretation of Table 5.5 is as follows

- For both measures (in Kcal and grams), the coefficients of totals as geometric mean are significant positive in all regression models.
- The logarithm of age is significant and positive and the square of the logarithm of age is significant and negative, i.e BMI increases as age increases, then after a given threshold, BMI tends to decrease.[9] For example, in terms of macronutrients in Kcal, the thresholds at 15% quantile, mean and 90% quantile are 42.1, 49.7 and 71.1 years old. It is quite interesting that the threshold of the obese group is the highest number.

[8] These coefficients are dependent on the choice of Ilr coordinates.

[9] When we interpret the coefficients of Age, we assume that all other variables remain constant. Then, the peak of Age is equal to $\exp(\frac{a1}{-2*a2})$ where $a1$ and $a2$ are coefficients of $log(Age)$ and $log^2(Age)$.

- The coefficients of the logarithm of expenditure are all positive but significant only for the conditional mean regression.
- When gender takes the female level, its coefficient is significant and negative. It means that women tend to have a lower BMI than men conditionally on other characteristics.
- In comparison to minority level of ethnicity, the coefficient of the Kinh ethnicity level is significant and negative. It means that the Kinh people have a lower BMI indicator than minority people on average conditional on other characteristics.
- The coefficients of secondary school levels are significant and negative at the mean and at the 90% quantile level. All coefficients of other education levels are insignificant.
- The coefficients of the non-farmers job level are significant and positive. Then, on average, non–farmers tend to have higher BMI than farmers at all regression levels conditional on other characteristics. These results are reasonable since in this study, job type, i.e farmers or non–farmers, plays the role of activities levels. People who have more intensive activities will consume more energy.
- About the drinking beer status, it is interesting that the coefficient of the non-drinking beer people is significant and negative. Then, on average, non–drinking people have smaller BMI than drinking people conditional on other characteristics.
- The coefficients of geographical regions have various signs (positive and negative, significant or insignificant). These mixed effects are due to the fact that there is a confounding effect between the impact of the regions and that of the other characteristics.

### 5.3.3 Elasticities computation in these compositional models

In order to interpret the share regression models, Morais et al. (2018) suggest to use elasticities to overcome complex interpretations of the parameters in ILR coordinates regressions. These elaticities are similar to odds ratios which are popular in medical research. The elasticity quantifies the relative variation of an outcome variable due to the relative variation of an explanatory variable, measured in percentage. We adapt the elasticity calculation of Morais et al. (2018) to the case of our preferred model, i.e a the compositional model with a total. In our case, since the dependent variable is not a composition, the adapted tool is a semi-elasticity but since our ourcome is the log of BMI, the semi elasticity of the outcome corresponds to the elasticity of BMI. The mathematical computation of the semi-elasticities

are given in the Appendix E. We also prove that these elasticity formulas are invariant to the choice of ilr coordinates. In the case of three macronutrient shares (in Kcal), the elastiscity of BMI with respect to the volumes of macronutrients are given by:[10]

$$\frac{\partial Y}{\partial lnV_C} = \beta\sqrt{\frac{2}{3}} + \frac{\delta T}{3}, \qquad \frac{\partial Y}{\partial lnV_P} = \frac{-\beta}{\sqrt{6}} + \frac{\gamma}{\sqrt{6}} + \frac{\delta T}{3}, \qquad \frac{\partial Y}{\partial lnV_F} = \frac{-\beta}{\sqrt{6}} - \frac{\gamma}{\sqrt{6}} + \frac{\delta T}{3}$$

and

$$\frac{\partial Y}{\partial \ln T} = \delta$$

Table 5.6 displays the average elasticities of BMI with respect to macronutrients at various quantile orders and for both units: Kcal and grams. Results indicate that

- The elasticities of BMI with respect to carbohydrate are always negative.
- Positive semi-elasticities are associated to fat, protein and fiber.
- Generally, BMI is more elastic to protein.

Table 5.6: Average elasticities of BMI with respect to macronutrients at various quantile orders

| Macronutrient | Macronutrient in Kcal | | | Macronutrient in grams | | |
|---|---|---|---|---|---|---|
| | $\tau = 0.15$ | Mean | $\tau = 0.9$ | $\tau = 0.15$ | Mean | $\tau = 0.9$ |
| Carbohydrate | -0.003 | -0.011 | -0.012 | -0.006 | -0.014 | -0.016 |
| Fat | 0.005 | 0.005 | 0.003 | 0.004 | 0.004 | 0.004 |
| Protein | 0.020 | 0.024 | 0.030 | 0.020 | 0.021 | 0.025 |
| Fiber | | | | 0.006 | 0.008 | 0.006 |

Figure 5.6 and 5.7 show the average elasticities as a function of BMI for all macronutrients in the two units obtained by smoothing the scatterplot of elasticity as a function of BMI. Average elasticities of BMI with respect to all macronutrients in both units increase rapidly with BMI and then become stable after a threshold around BMI = 20, meaning that individuals with a low BMI are more affected by the composition of their nutrition.

## 5.4 Conclusion

This article focuses on the relationship between food consumption and the BMI indicator. This is an important issue since there is a close link between eating habits and the occurrence of chronic diseases. These topics are currently analyzed by many multi-disciplinary researchers but the findings

[10]These formulas are based on sequential binary partitions: Carbohydrate vs protein and fat, protein vs fat.

Figure 5.6: Elasticity (in Kcal) as function of BMI.

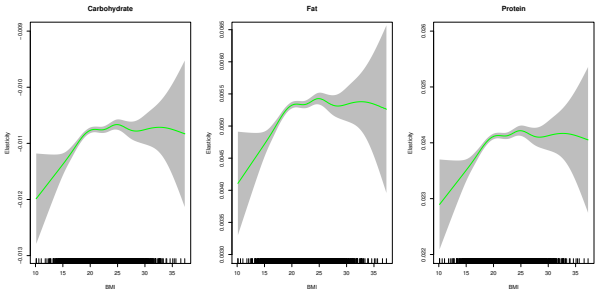
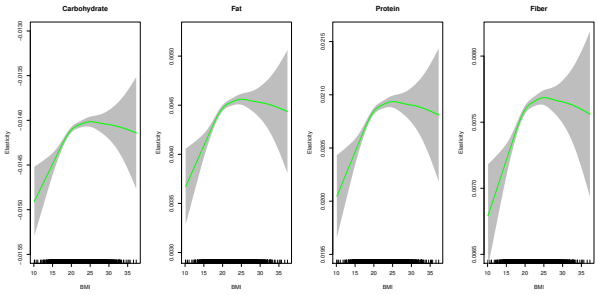



Figure 5.7: Elasticity (in grams) as function of BMI.




remain controversial due to the complex associations between total energy intake, physical activity and the metabolism of each person.

There are many different approaches to find the relationship between the BMI indicator and the diet intake in epidemiologic analysis. However, the current models in the literature fail to disentangle the comprehensive effect of total energy from that of each component of energy. Our proposal is based on the compositional data approach (CoDa). There are only few empirical epidemiologic studies using the CoDa approach, such as Leite (2016), Dumuid et al. (2017) and Trinh et al. (2018). This advanced methodology has much to bring to epidemiology.

We propose to estimate various regression models: compositional models, total only models and compositional models with a total. We use geometric mean of macronutrient shares as total, as a determinant variable of the BMI indicator. These models are estimated at the conditional mean and two quantile orders: 15% quantile regression (corresponding to the underweight cut–off), and 90% quantile regression (corresponding to the overweight cut–off). Macronutrients are measured in two different units: Kcal and grams.

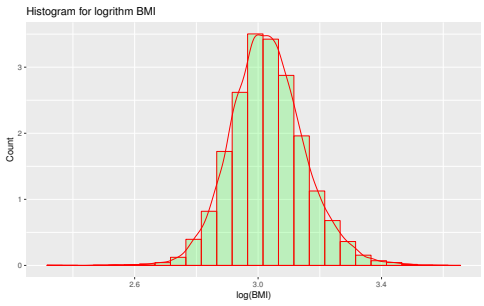
From an analysis of variance comparing alternative models, we conclude that the full model, i.e compositional model with a total, is preferred whatever the conditional mean or quantile regressions.

Average elasticity values increase as BMI increases until a threshold ($BMI = 20$). Some of these results could be due to confounding effects. Protein could be acting as a proxy for unhealthy behaviors: individuals who consume higher amounts of protein may be wealthier, less active, smoke more, and consume more processed foods, sugar-sweetened beverages, and total energy than individuals who consume lower amounts of protein. It is impossible to account for all of the potential confounders (i.e., unhealthy behaviors) using the available dataset.

Table 5.7: General Nutrition Survey for 2009-2010 description variables

| Variable | Description | Value |
|---|---|---|
| N | Number of observation | 15035 |
| BMI | Body mass index (BMI) | 20.9 ( 2.6 ) |
| ExpWeek | Total expenditure per week (thousand vnd) | 3416.4 ( 3211 ) |
| Age | Year olds | 36.4 ( 11.4 ) |
| Gender | Male | 51.25 % |
| | Female | 48.75 % |
| Ethnic | Minority | 15.32 % |
| | Kinh | 84.68 % |
| Educ | Below primary | 35.1 % |
| | Secondary school | 34.04 % |
| | High school | 19.92 % |
| | Univeristy | 10.95 % |
| Job | Farmer | 43.25 % |
| | Non-Farmer | 56.75 % |
| Region | Delta | 53.47 % |
| | Coastline | 8.92 % |
| | Midlands-mountainous | 3.15 % |
| | Low mountains | 21.31 % |
| | High mountains | 13.15 % |
| Drinking beer status | Bear1 (more than 1 time per week) | 16.72 % |
| | Bear2 (1-4 times per months) | 15.46 % |
| | Bear0 (no drinking) | 67.82 % |
| Smooking status | Smook | 25.3 % |
| | Nonsmook | 74.7 % |

Figure 5.8: Density of log(BMI).



# Chapter 6

# Further research

## 6.1 In terms of mathematical perspective

### 6.1.1 Decomposition method using copulas with discrete variables

We have applied a decomposition method recently proposed by Rothe (2015) in Chapter 3. This decomposition method is based on the estimation of the cumulative distribution function of the covariates. One solution would be to use nonparametric techniques to estimate this function but then we would face the curse of the dimensionality problem if the number of covariates is too large. This motivates the use of copulas because, as proved by Sklar (1959), the CDF of $X = (X_1, X_2, \ldots, X_k)$ can always be written as

$$F_X(x) = C(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_k}(x_k))$$

The estimation of the joint CDF then consists of estimating marginal CDFs and copula dependency parameters.

But, the above theorem only holds for continuous covariates. When some of them are discrete, some identifiability issues may arise. Rothe (2015) first proposes to represent each discrete covariate $X_j$ as $X_j = t_j(\widetilde{X}_j)$ for some continuously distributed latent variable $\widetilde{X}_j$ and a function $t_j(.)$ that is weakly increasing in its argument. These argument does not show clearly how to build a latent variable, especially in the case when a variable includes several factors. Second, the author uses Gaussian copulas without any comparison with alternatives. Third, the author uses probit regression in the calculation of the counterfactual distributions without detailed explanation.

In the literature, several articles propose to build specific copulas for discrete variables or mixed-variables. For example, Panagiotelis et al. (2011) has extended the principles of vine Pair Copula Constructions (PCCs) to discrete margins. However these vine PCCs, though appealing on the theoretical point of view, are quite complex to apply in practice. Kolesárová et al. (2006) introduce some interesting properties of discrete copulas in the

case when the marginals coincide and correspond to the uniform probability distribution on a grid. However, they only consider the case of uniform discrete probability distribution. Thus, their approach is narrow, specific and it is difficult to extend to general discrete variables.

There will be two main points in our extension of Rothe (2015). **First**, our proposal for constructing counterfactual copulas is based on the empirical distribution of the discrete variables. We plan to go further on defining these copulas estimates in two important cases:

- Copulas in the case of two discrete variables,
- Copulas in the case of two discrete variables and one continuous variable,

then to adapt the procedure to more general situations. **Second**, we propose to choose among some alternative families of copulas and types of regression in the calculation of the counterfactuals. We plan to apply the inference on counterfactual distributions of Chernozhukov et al. (2013). The implementation in R will be adapted from the R package counterfactual (Chen et al., 2016).

### 6.1.2 Decomposition method and compositional models

Shorrocks (1982) has introduced a decomposition by factor components to analyze inequality. The author considers the case of the different components of total income. Benjamin et al. (2017) has applied Shorrock's decomposition to study inequality of income in Vietnam with six components: Crop income, Sideline income, Family business, Wages, Remittances, Other income. The six components here constitute a compositional vector in the simplex with $D = 6$, i.e we will have six shares: income sources from the six components.

Compositional data analysis (CoDa) has been applied in many different socio-economics contexts. Then, a natural question is whether we can build a decomposition method adapted to the case of an explanatory variable which is a compositional vector, possibly including also a total.

## 6.2 In terms of nutrition perspective, several empirical articles are in progress

Nutrition should be seen in a comprehensive picture where we consider together the relationship between: nutrition - food - agriculture and environment. Many researchers have recently focused on these relations due to the Sustainable Development Goals (SDGs) of the United Nations. Among 17 SDGs indicators, there are several indicators related to nutrition - food -

agriculture and environment. In Vietnam, the Vietnamese government, institutes and researchers are focusing on these indicators which are related to nutrition - food - agriculture and environment. These empirical researches are multidisciplinary and require collaboration with several Vietnamese ministries such as: Ministry of Health, General Statistical Office, Ministry of Agriculture and Minister of Natural Resources and Environment...

To go further in applying recent mathematics methods to various topics related to the nutrition - food - agriculture and environment relationships, there are several projects listed below that I am working on

- Determinants of stunting of children in Vietnam, using the Nutrition Surveillance Profiles 2013 (National Institute of Nutrition, 2013). Potential mathematic tools are: multilevel regression, non-linear decomposition.
- Extreme climate change and food security using Vietnam Living Standard Survey and rainfall data.
- Climate change and health status using Multiple Indicator Cluster Surveys of Unicef, Nutrition Surveillance Profiles 2013 (National Institute of Nutrition, 2013) and climate change data.
- Impact of an increasing availability of imported or processed consumer goods on the food diet composition in Vietnam. We will apply compositional data analysis (CoDa) in this empirical research. Our primary results have been presented at the Workshop TAASE, June 12th, 2016, Thang Long University, Hanoi Vietnam.

# Bibliography


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.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall London.

Albar, S. A., N. A. Alwan, C. E. Evans, and J. E. Cade (2014). Is there an association between food portion size and bmi among british adolescents? *British Journal of Nutrition 112*(5), 841–851.

Anderson, L. R. (2018). Adolescent mental health and behavioural problems, and intergenerational social mobility: A decomposition of health selection effects. *Social Science & Medicine 197*, 153–160.

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

Benjamin, D., L. Brandt, and B. McCaig (2017). Growth with equity: income inequality in Vietnam, 2002-14. *Journal of Income Inequality 8*, 436–455.

Bhalotra, S. and C. Attfield (1998). Intrahousehold resource allocation ibn 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.

Bloom, D. E., E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, A. Hamandi, M. Mowafi, et al. (2012). The global economic burden of noncommunicable diseases. Technical report, Program on the Global Demography of Aging.

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.

Burggraf, C., R. Teuber, S. Brosig, and T. Glauben (2015). Economic growth and the demand for dietary quality: Evidence from russia during transition. *Economics & Human Biology 19*, 184–203.

Camacho, S. and A. Ruppel (2017). Is the calorie concept a real solution to the obesity epidemic? *Global health action 10*(1), 1289650.

Chen, M., V. Chernozhukov, I. Fernández-Val, and B. Melly (2016). Counterfactual: An r package for counterfactual analysis. *arXiv preprint arXiv:1610.07894*.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Cling, J. P., M. Razafindrakoto, and F. Roubaud (2010). Assessing the potential impact of the global crisis on the labour market and the informal sector in vietnam. *Journal of Economics and Development 38*, 16–25.

Coenders, G., J. A. Martín-Fernández, and B. Ferrer-Rosell (2017). When relative and absolute information matter: Compositional predictor with a total in generalized linear models. *Statistical Modelling 17*(6), 494–512.

Cohen, E., M. Cragg, A. Hite, M. Rosenberg, B. Zhou, et al. (2015). Statistical review of us macronutrient consumption data, 1965–2011: Americans have been following dietary guidelines, coincident with the rise in obesity. *Nutrition 31*(5), 727–732.

Cooper, L. G. and M. Nakanishi (1989). *Market-share analysis: Evaluating competitive marketing effectiveness*, Volume 1. Springer Science & Business Media.

Cuong, T. Q., M. J. Dibley, S. Bowe, T. T. M. Hanh, and T. T. H. 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.

DiNardo, J., N. Fortin, and T. Lemieux (1996). Labor market institutions and the distributions of wage, 1973-1992: a semiparametric approach. *Econometrica 81*, 1001–1044.

Dumuid, D., T. E. Stanford, J. Martin-Fernández, Ž. Pedišić, C. A. Maher, L. K. Lewis, K. Hron, P. T. Katzmarzyk, J. P. Chaput, M. Fogelholm, et al. (2017). Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical methods in medical research*, 0962280217710835.

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

Egozcue, J., J. Daunis-I-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). *Simplicial regression. The normal model*. na.

Egozcue, J. and V. Pawlowsky-Glahn (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Fairlie, R. W. (2005). An extension of the blinder-oaxaca decomposition technique to logit and probit models. *Journal of economic and social measurement 30*(4), 305–316.

Ferrer-Rosell, B. and G. Coenders (2017). Airline type and tourist expenditure: Are full service and low cost carriers converging or diverging? *Journal of Air Transport Management 63*, 119–125.

Foresi, S. and F. Peracchi (1995). The conditional distribution of excess returns: an empirical analysis. *Journal of the American Statistical Association 90*, 451–466.

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.

Ha, D. T. P., E. J. M. Feskens, P. Deurenberg, B. M. Le, C. K. Nguyen, and F. J. Kok (2011). Nationwide shifts in the double burden of overweight and underweight in Vietnamese adults in 2000 and 2005: two national nutrition surveys. *BMC Public Health 62*(11).

Hall, K. D., G. Sacks, D. Chandramohan, C. C. Chow, Y. C. Wang, S. L. Gortmaker, and B. A. Swinburn (2011). Quantification of the effect of energy imbalance on bodyweight. *The Lancet 378*(9793), 826–837.

Hastie, T. and R. Tibshirani (1987). Generalized additive models: some applications. *Journal of the American Statistical Association 82*(398), 371–386.

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

Hooper, L., C. D. Summerbell, J. P. Higgins, R. L. Thompson, N. E. Capps, G. D. Smith, R. A. Riemersma, and S. Ebrahim (2001). Dietary fat intake and prevention of cardiovascular disease: systematic review. *Bmj 322*(7289), 757–763.

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

Huynh, T. T. D., M. J. Dibley, D. Sibbritt, and T. M. H. Tran (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.

IFPRI (2017). 2017 global food policy report. Technical report, International Food Policy Research Institute Pub, Washington, DC.

Jiryaie, F., N. Withanage, B. Wu, and de Leon A. R. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation 86*(9), 1643–1659.

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

Koenker, R. and G. Bassett (1982). Tests of linear hypotheses and l" 1 estimation. *Econometrica: Journal of the Econometric Society*, 1577–1583.

Koenker, R. and K. F. Hallock (2001). Quantile regression. *Journal of economic perspectives 15*(4), 143–156.

Kolesárová, A., R. Mesiar, J. Mordelová, and C. Sempi (2006). Discrete copulas. *IEEE Transactions on Fuzzy Systems 14*(5), 698–705.

Le, N. B., T. H. Le, D. V. Nguyen, T. N. Tran, H. C. Nguyen, T. T. 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.

Leibbrandt, M., J. A. Levinsohn, and J. McCrary (2010). Incomes in South Africa after the fall of Apartheid. *Journal of Globalization and Development 1*(1), Article 2.

Leite, M. and F. Prinelli (2017). A compositional data perspective on studying the associations between macronutrient balances and diseases. *European journal of clinical nutrition 71*(12), 1365.

Leite, M. L. (2016). Applying compositional data methodology to nutritional epidemiology. *Statistical methods in medical research 25*(6), 3057–65.

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

Liaskos, G. and P. Lazaridis (2003). The demand for selected food nutrients in greece: The role of socioeconomic factors. *Agricultural Economics Review 4*(2).

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.

Mayen, A. L., P. Marques-Vidal, F. Paccaud, P. Bovet, and S. Stringhini (2014). Socioeconomic determinants of dietary patterns in low- and middle-income countries: a systematic review. *American Journal of Clinical Nutrition 100*, 1520–1531.

Mert, M. C., P. Filzmoser, G. Endel, and I. Wilbacher (2016). Compositional data analysis in epidemiology. *Statistical Methods in Medical Research 6*, 0962280216671536.

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.

Morais, J., C. Thomas-Agnan, and M. Simioni (2017). Using compositional and dirichlet models for market share regression. *Journal of Applied Statistics 24*, 1–20.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018). Interpreting the impact of explanatory variables in compositional models. *17*.

Muller, I., K. Hron, and E. Fiserova (2016). Interpretation of compositional regression with application to time budget analysis. *arXiv* (1609.07887).

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.

National Institute of Nutrition (2010). *General Nutrition Survey 2009 - 2010*. Medical Publishing House.

National Institute of Nutrition (2013). *Nutrition Surveillance Profiles 2013*. UNICEF, Alive & Thrive.

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, M. T. and B. M. Popkin (2004). Patterns of food consumption in vietnam: effects on socioeconomic groups during an era of economic growth. *European journal of clinical nutrition 58*(1), 145.

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., A. A. Leon, M. E. D. Sánchez, and A. Sousa-Poza (2018). The rise in obesity in cuba from 2001 to 2010: An analysis of national survey on risk factors and chronic diseases data. *Economics & Human Biology 28*, 1–13.

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.

Organization, W. H. et al. (2016). The double burden of malnutrition: policy brief.

Panagiotelis, A., C. Czado, and H. Joe (2011). Pair copula constructions for discrete data.

Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Pawlowsky-Glahn, V., J. J. Egozcue, and D. Lovell (2015). Tools for compositional data with a total. *Statistical Modelling 15*(2), 175–190.

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.

Porkka, M., M. Kummu, S. Siebert, and O. Varis (2013). From food insufficiency towards trade dependency: a historical analysis of global food availability. *PloS one 8*(12), e82714.

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

Randi, G., C. Pelucchi, S. Gallus, M. Parpinel, L. Dal Maso, R. Talamini, L. S. Augustin, A. Giacosa, M. Montella, S. Franceschi, et al. (2007). Lipid, protein and carbohydrate intake in relation to body mass index: an italian study. *Public health nutrition 10*(3), 306–310.

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

Riera-Crichton, D. and N. Tefft (2014). Macronutrients and obesity: revisiting the calories in, calories out framework. *Economics & Human Biology 14*, 33–49.

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

Sakellariou, C. and Z. Fang (2014). The Vietnam reforms, change in wage inequality and the role of the minimum wage. *Economics of Transition 22*(2), 313–340.

Santaeulàlia-Llopis, R. and Y. Zheng (2017). Why is food consumption inequality underestimated? a story of vices and children.

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

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica: Journal of the Econometric Society*, 193–211.

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Sklar, A. (1959). Fonction de répartition dont les marges sont données. *Inst. Stat. Univ. Paris 8*, 229–231.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics 39*, 312–320.

Stiglitz, J. E. (1976). The efficiency wage hypothesis, surplus labour, and the distribution of income in ldcs. *Oxford economic papers 28*(2), 185–207.

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.

Thi, H. T., M. Simioni, and C. Thomas-Agnan (2018). Assessing the nonlinearity of the calorie-income relationship: An estimation strategy–with new insights on nutritional transition in vietnam. *World Development 110*, 192–204.

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.

Trichopoulou, A., C. Gnardellis, V. Benetou, P. Lagiou, C. Bamia, and D. Trichopoulos (2002). Lipid, protein and carbohydrate intake in relation to body mass index. *European journal of clinical nutrition 56*(1), 37.

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*, 0962280218770223.

Trinh, T., T. Do, V. Nguyen, Q. D. Nguyen, and C. Thomas-Agnan (2018a). Measuring the progress of timeliness childhood immunization compliance in vietnam between 2006-2014: A decomposition analysis. *TSE working papers Xxx*, xx–xx.

Trinh, T. H., T. Do, V. Nguyen, Q. Nguyen, and C. Thomas-Agnan (2018b). Measuring the progress of the timeliness childhood immunization compliance in vietnam between 2006-2014: A decomposition analysis. *TSE working paper* (18-920).

Trinh, T. H., M. Simioni, and C. Thomas-Agnan (2016). Calorie intake and income in china: New evidence using semiparametric modelling with generalized additive models. *Vietnam Journal of Mathematical Applications 14*(1), 11–26.

Trinh, T. H., M. Simioni, and C. Thomas-Agnan (2018). A new perspective on the relationship between calorie intake and income in china and vietnam using semiparametric modeling. Technical report.

Trivedi, P. K. and D. M. Zimmer (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics 1*(1), 1–111.

Tuan, N., P. Tuong, and B. Popkin (2008). Body mass index (bmi) dynamics in vietnam. *European journal of clinical nutrition 62*(1), 78.

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data with R.* Springer.

Wacholder, S., A. Schatzkin, L. S. Freedman, V. Kipnis, A. Hartman, and C. C. Brown (1994). Can energy adjustment separate the effects of energy from those of specific macronutrients? *American journal of epidemiology 140*(9), 848–855.

Weiß, G. (2011). Copula parameter estimation by maximum likelihood and minimum-distance estimators: a simulation study. *Computational Statistics 26*, 31–54.

WHO, E. C. (2004). Appropriate body-mass index for asian populations and its implications for policy and intervention strategies. *Lancet (London, England) 363*(9403), 157.

Willett, W. C., G. R. Howe, and L. H. Kushi (1997). Adjustment for total energy intake in epidemiologic studies. *The American journal of clinical nutrition 65*(4), 1220S–1228S.

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.

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

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

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

# Appendix

## A 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 $\alpha_2$ in equation (2.1) when DLM is the chosen model. The procedure is as follows when a GAM model is chosen. The smooth function $s(x)$ in equations (2.7) and (2.8) is expressed as a linear (in parameters) basis expansion of the form

$$s(x) = \gamma_0 + \gamma_1 x + \sum_{i=1}^{n} \delta_i (x - x_i)^3 \tag{6.1}$$

when estimating GAM models. $\gamma_0$, $\gamma_1$, and the $\delta_i$, $i = 1, \ldots, n$, are thus parameters to be estimated, the expansion (6.1) using thin plate regression splines (Wood, 2003). (6.1) 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 (6.1). 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 (6.1) 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 6.1. The tests clearly reject null hypothesis $\mathrm{H}_0 : \alpha_1 = 0$ and $\alpha_2 = 0$ when the chosen model is DLM, or $\mathrm{H}_0 : s(.) = 0$ when it is GAM. Table 6.1 reports also the results from linearity tests. The parameter $\alpha_2$ 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 6.1: 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 $\alpha_1$ and $\alpha_2$ in DLM models.
(3) $\widehat{\gamma}_1$ is estimated value of parameter $\gamma_1$ in GAM models.
(4) $^{\star}$, $^{\star\star}$, and $^{\star\star\star}$ mean significant at 10%, 5%, and 1%, respectively

## B 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 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.

## C 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 6.2. 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).[1]

Table 6.2: Conversion table Calories for Vietnam.

| Food | Energy Kcal | protein gr | fat gr |
|---|---|---|---|
| Plain rice | 344.5 | 8.5 | 1.55 |
| Sticky rice | 347 | 8.3 | 1.6 |
| Maize | 354 | 8.3 | 4 |
| Cassava | 146 | 0.8 | 0.2 |
| Potato of various kinds | 106 | 1.4 | 0.15 |
| Wheat grains, bread, wheat powder | 313.7 | 10.2 | 1.1 |
| Floor noodle, instant rice noodle, porridge | 349 | 11 | 0.9 |
| Fresh rice noodle, dried rice noodle | 143 | 3.2 | 0.2 |
| Vermicelli | 110 | 1.7 | 0 |
| Pork | 26016.5 | 21.5 | |
| Beef | 142.5 | 20.3 | 7.15 |
| Buffalo meat | 122 | 22.8 | 3.3 |
| Chicken meat | 199 | 20.3 | 13.1 |
| Duck and other poultry meat | 275 | 18.5 | 22.4 |
| Other types of meat | - | - | - |
| Processed meat | - | - | - |
| Fresh shrimp, fish | 83 | 17.75 | 1.2 |
| Dried and processed shrimps, fish | 361 | 49.16 | 14.6 |
| Other aquatic products and seafoods | - | - | - |
| Lard, cooking oil | 863.5 | 0 | 99.8 |
| Eggs of chicken, ducks, Muscovy ducks, geese | 103.74 | 8.34 | 7.74 |
| Tofu | 95 | 10.9 | 5.4 |
| Peanuts, sesame | 570.5 | 23.8 | 45.5 |
| Beans of various kinds | 73 | 5 | 0 |
| Fresh peas of various kinds | 596 | 0.4 | |
| Morning glory vegetables | 25 | 3 | 0 |
| Kohlrabi | 36 | 2.8 | 0 |
| Cabbage | 29 | 1.8 | 0.1 |
| Tomato | 20 | 0.6 | 0.2 |
| Other vegetables | - | - | - |
| Orange | 37 | 0.9 | 0 |
| Banana | 81.5 | 1.2 | 0.2 |
| Mango | 69 | 0.6 | 0.3 |
| Other fruits | - | - | - |
| Fish sauce | 60 | 12.55 | 0 |
| Salt | 0 | 0 | 0 |
| MSG | 0 | 0 | 0 |
| Glutamate | 0 | 0 | 0 |
| Sugars, molasses | 390 | 0.55 | 0 |
| Confectionery | 412.2 | 8.9 | 10.7 |
| Condensed milk, milk powder | 395.7 | 23.4 | 11.9 |
| Ice cream, yoghurt | - | - | - |
| Fresh milk | 61 | 3.9 | 4.4 |
| Alcohol of various kinds | 47 | 4 | 0 |
| Beer of various kinds | 11 | 0.5 | 0 |
| Bottled, canned, boxed beverages | 47 | 0.5 | 0 |
| Instant coffee | 0 | 0 | 0 |
| Coffee powder | 353 | 12 | 0.5 |
| Instant tea powder | 0 | 0 | 0 |
| Other dried tea | 0 | 0 | 0 |
| Cigarettes, waterpipe tobacco | 0 | 0 | 0 |
| Betel leaves, areca nuts, lime, betel pieces | 0 | 0 | 0 |
| Outdoor meals and drinks | - | - | - |
| Other foods and drinks | - | - | - |

Notes:
(1) Amount per 100gr food ; protein contains 4 calories per gram and fat contains 9 calories per gram
(2) Source: National Institute of Nutrition (2007).

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

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. 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

$$\log(THCI_i) = \gamma_0 + \gamma_1 \, Gender_i + \gamma_2 \, N_{a,i} + \gamma_3 \, Family_i + \varepsilon. \tag{6.2}$$

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 6.1 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.

Figure 6.1: Comparison of equivalence scales using 2012 VHLSS data

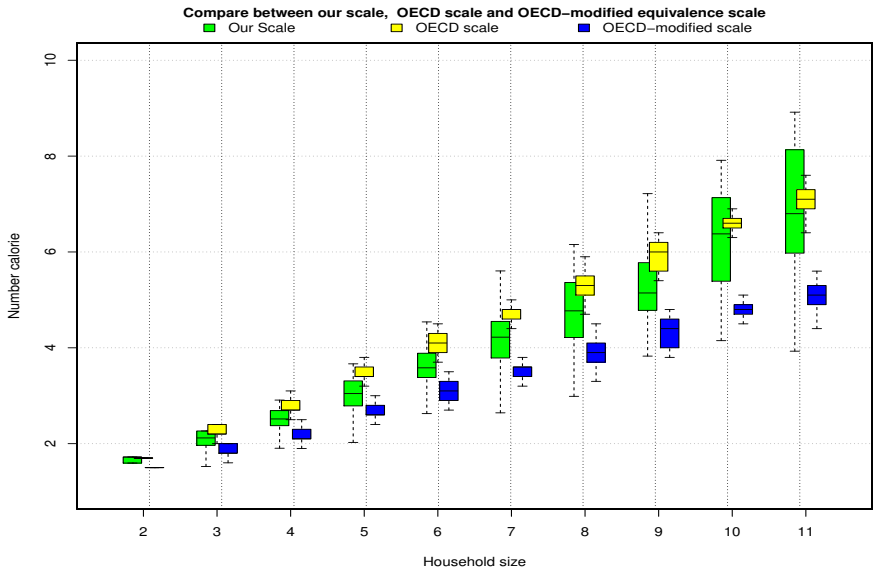



Table 6.3 reports the average value of adult equivalent calorie intake for each VHLSS wave and compares it with other available studies on Vietnam. The average values we obtained are consistent with those obtained in other papers using the same survey data. They are just a little higher, which we could be foreseen as the other studies use total calorie intake divided by household size.

Table 6.3: 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 6.3. 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. 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.

## D 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:

$$\mathbb{E}[Y - g(X)|W] = 0 \tag{6.3}$$

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. (6.3), 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

$$\tau_n = \int S_n^2(x)dx \tag{6.4}$$

where $S_n(x)$ is the sample analogue of $S(x) = \mathbb{E}\{[Y - G(X)]f_{XW}(x, W)\}$ 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 $\tau_n$ 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 $\tau_n$ 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.

## E Marginal effect and elasticity calculus on ILR

We are going to demonstrate how to compute the semi–elasticities of the dependent variable $Y$ relative to an explanatory variable $X_j$, using compositional models. The demonstration is made for a CODA model with a compositional explanatory variable and a real valued dependent variable. These semi-elasticities calculations are valid for both linear regression and quantile regression.

Consider for $D = 3$, the ILR transformation defined by the transformation matrix:

$$\mathbf{W} = \begin{bmatrix} \sqrt{\frac{2}{3}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{6.5}$$

Let us remind that $X^* = ilr(X) = V' \ln(X)$, i.e

$$X_1^* = \sqrt{\frac{2}{3}} \ln X_1 - \frac{1}{\sqrt{6}} \ln X_2 - \frac{1}{\sqrt{6}} \ln X_3$$

$$X_2^* = \frac{1}{\sqrt{2}} \ln X_2 - \frac{1}{\sqrt{2}} \ln X_3$$

We define the total as

$$T = exp(\frac{1}{3}\big[\ln(V_1) + \ln(V_2) + \ln(V_3)\big]) \tag{6.6}$$

i.e

$$\ln T = \frac{1}{3}\big[\ln(V_1) + \ln(V_2) + \ln(V_3)\big]$$

where $V_1, V_2, V_3$ are the three volumes of macronutrients. Then,

$$\frac{\partial T}{\partial V_1} = \frac{\partial T}{\partial \ln T}.\frac{\partial \ln T}{\partial V_1} = \frac{T}{3V_1}, \qquad \frac{\partial T}{\partial V_2} = \frac{T}{3V_2}, \qquad \frac{\partial T}{\partial V_3} = \frac{T}{3V_3}$$

We define the following transformations

$$F_T : (V_1, V_2, V_3)' \to (ilr_1, ilr_2, T)'$$
$$M : (ilr_1, ilr_2, T)' \to Y = M(ilr_1, ilr_2, T) = \alpha + \beta ilr_1 + \gamma ilr_2 + \delta T,$$

whether $M = \mathbb{E}(ilr_1, ilr_2, T)$ is a mean level or $M = \mathbb{Q}_\tau(ilr_1, ilr_2, T)$, $\tau$ is a quantile level.

We are going to use the following property of Jacobian matrices: $J = J_M J_{F_T}$.

$$J_{F_T} = \begin{bmatrix} \sqrt{\frac{2}{3}}\frac{1}{V_1} & -\frac{1}{\sqrt{6}}\frac{1}{V_2} & -\frac{1}{\sqrt{6}}\frac{1}{V_3} \\ 0 & \frac{1}{\sqrt{2}}\frac{1}{V_2} & -\frac{1}{\sqrt{2}}\frac{1}{V_3} \\ \frac{T}{3V_1} & \frac{T}{3V_2} & \frac{T}{3V_3} \end{bmatrix} \tag{6.7}$$

and

$$J_M = \begin{bmatrix} \frac{\partial Y}{\partial V_1^*} & \frac{\partial Y}{\partial V_2^*} & \frac{\partial Y}{T} \end{bmatrix} = \begin{bmatrix} \beta & \gamma & \delta \end{bmatrix}. \tag{6.8}$$

Then

$$J = J_M J_{F_T} = \begin{bmatrix} \frac{\partial Y}{\partial V_1} \\ \frac{\partial Y}{\partial V_2} \\ \frac{\partial Y}{\partial V_3} \end{bmatrix} = \begin{bmatrix} \beta & \gamma & \delta \end{bmatrix} \begin{bmatrix} \sqrt{\frac{2}{3}}\frac{1}{V_1} & -\frac{1}{\sqrt{6}}\frac{1}{V_2} & -\frac{1}{\sqrt{6}}\frac{1}{V_3} \\ 0 & \frac{1}{\sqrt{2}}\frac{1}{V_2} & -\frac{1}{\sqrt{2}}\frac{1}{V_3} \\ \frac{T}{3V_1} & \frac{T}{3V_2} & \frac{T}{3V_3} \end{bmatrix}$$
$$= \begin{bmatrix} \beta\sqrt{\frac{2}{3}}\frac{1}{V_1} + \frac{\delta T}{3V_1} \\ \frac{-\beta}{\sqrt{6}}\frac{1}{V_2} + \frac{\gamma}{\sqrt{2}}\frac{1}{V_2} + \frac{\delta T}{3V_2} \\ \frac{-\beta}{\sqrt{6}}\frac{1}{V_3} - \frac{\gamma}{\sqrt{2}}\frac{1}{V_3} + \frac{\delta T}{3V_3} \end{bmatrix}$$

Then,

$$\frac{\partial Y}{\partial lnV_1} = \beta\sqrt{\frac{2}{3}} + \frac{\delta T}{3}, \quad \frac{\partial Y}{\partial lnV_2} = \frac{-\beta}{\sqrt{6}} + \frac{\gamma}{\sqrt{2}} + \frac{\delta T}{3}, \quad \frac{\partial Y}{\partial lnV_3} = \frac{-\beta}{\sqrt{6}} - \frac{\gamma}{\sqrt{2}} + \frac{\delta T}{3}$$

We are now going to demonstrate that the semi–elasticities are invariant to the choices of the transformation matrix. In addition, the demonstration is made in a general case, i.e with D components. Assume $V$ has D components, i.e $V = (V_1, V_2, ..., V_D)$. Assume there are two tranformation matrices $\mathbb{V}^{\mathbb{A}}$ and $\mathbb{V}^{\mathbb{B}}$, the corresponding Ilr coordinates are

$$Ilr^A = (Ilr_1^A, ..., Ilr_{D-1}^A) = \left[\mathbb{V}^{\mathbb{A}}\right]_{(D-1)\times D} [\ln V]_{D\times 1} = \left[\mathbb{V}^{\mathbb{A}}\right]_{(D-1)\times D} \begin{bmatrix} \ln V_1 \\ \ln V_2 \\ ... \\ \ln V_D \end{bmatrix}$$

and

$$Ilr^B = (Ilr_1^B, ..., Ilr_{D-1}^B) = \left[\mathbb{V}^{\mathbb{B}}\right]_{(D-1)\times D}[\ln V]_{D\times 1} = \left[\mathbb{V}^{\mathbb{B}}\right]_{(D-1)\times D}\begin{bmatrix}\ln V_1\\ \ln V_2\\ ...\\ \ln V_D\end{bmatrix}$$

A total as geometric mean of $V$ is

$$\ln T = \frac{1}{D}(\ln V_1 + \ln V_2 + ... + \ln V_D).$$

Then, given that there are two ways to construct Ilr coordinates, there are two regression models:

$$Y = \alpha^A + \sum_{j=1}^{D-1}\beta_j^A Ilr^A + \delta^A T + \epsilon^A = \alpha^A + [\beta_1^A ... \beta_{D-1}^A]\left[\mathbb{V}^{\mathbb{A}}\right]_{(D-1)\times D}[\ln V]_{D\times 1} + \delta^A T + \epsilon^A \quad (6.9)$$

$$Y = \alpha^B + \sum_{j=1}^{D-1}\beta_j^B Ilr^B + \delta^B T + \epsilon^B = \alpha^B + [\beta_1^B ... \beta_{D-1}^B]\left[\mathbb{V}^{\mathbb{B}}\right]_{(D-1)\times D}[\ln V]_{D\times 1} + \delta^B T + \epsilon^B \quad (6.10)$$

Models (6.9) and (6.10) are estimated by the ordinary least squares method. They have the same dependent variable, i.e $Y$ and the same explanatory variables, i.e $\ln V_1, \ln V_2, ..., \ln V_D$ and $T$. Then, the corresponding coefficients estimated from the two models must be equal. Thus we have

$$\delta^A = \delta^B = \delta \quad \text{and} \quad [\beta_1^A \quad ... \quad \beta_{D-1}^A]\left[\mathbb{V}^{\mathbb{A}}\right]_{(D-1)\times D} = [\beta_1^B \quad ... \quad \beta_{D-1}^B]\left[\mathbb{V}^{\mathbb{B}}\right]_{(D-1)\times D} \quad (6.11)$$

In addition, we define the following transformations

$$F_T^A : (V_1, ..., V_D)' \to (ilr_1^A, ..., ilr_{D-1}^A, T)'$$

$$M^A : (ilr_1^A, ..., ilr_{D-1}^A, T)' \to Y = M^A(ilr_1^A, ..., ilr_{D-1}^A, T)' = \alpha^A + \sum_{j=1}^{D-1}\beta_j^A Ilr^A + \delta^A T,$$

and

$$F_T^B : (V_1, ..., V_D)' \to (ilr_1^B, ..., ilr_{D-1}^B, T)'$$

$$M^B : (ilr_1^B, ..., ilr_{D-1}^B, T)' \to Y = M^B(ilr_1^B, ..., ilr_{D-1}^B, T)' = \alpha^B + \sum_{j=1}^{D-1}\beta_j^B Ilr^B + \delta^B T,$$

then, we have

$$J^A = J_{M^A}J_{F_T^A} \qquad J^B = J_{M^B}J_{F_T^B}$$

In detail

$$J_{M^A} = \begin{bmatrix}\beta_1^A & ... & \beta_{D-1}^A & \delta\end{bmatrix} \qquad J_{M^B} = \begin{bmatrix}\beta_1^B & ... & \beta_{D-1}^B & \delta\end{bmatrix}$$

$$J_{F_T^A} = \begin{bmatrix} \mathbb{V}^{\mathbb{A}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix} \left[\frac{1}{V}\right]_{D\times 1} \qquad J_{F_T^B} = \begin{bmatrix} \mathbb{V}^{\mathbb{B}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix} \left[\frac{1}{V}\right]_{D\times 1}$$

Then,

$$J^A = \begin{bmatrix} \beta_1^A & ... & \beta_{D-1}^A & \delta \end{bmatrix} \begin{bmatrix} \mathbb{V}^{\mathbb{A}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix} \left[\frac{1}{V}\right]_{D\times 1}$$

and

$$J^B = \begin{bmatrix} \beta_1^B & ... & \beta_{D-1}^B & \delta \end{bmatrix} \begin{bmatrix} \mathbb{V}^{\mathbb{B}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix} \left[\frac{1}{V}\right]_{D\times 1}$$

The semi-elasticity computed from the two different sets of Ilr coordinates are

$$\left[\frac{\partial Y}{\partial lnV}\right]_{D\times 1} = \begin{bmatrix} \beta_1^A & ... & \beta_{D-1}^A & \delta \end{bmatrix} \begin{bmatrix} \mathbb{V}^{\mathbb{A}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix}$$

$$\left[\frac{\partial Y}{\partial lnV}\right]_{D\times 1} = \begin{bmatrix} \beta_1^B & ... & \beta_{D-1}^B & \delta \end{bmatrix} \begin{bmatrix} \mathbb{V}^{\mathbb{B}}{}_{(D-1)\times D} \\ \left[\frac{T}{D}\right]_{1\times D} \end{bmatrix}$$

Applying the results of equation (6.11), we infer that the calculation of the semi-elasticity is invariant to the choices of transformation matrix $\mathbb{V}^{\mathbb{A}}{}_{(D-1)\times D}$ and $\mathbb{V}^{\mathbb{B}}{}_{(D-1)\times D}$.