Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis

Huong Thi Trinh, Joanna Morais, Michel Simioni, Christine Thomas-Agnan

Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis

Huong Trinh Thi [1,2], Joanna Morais [3], Christine Thomas-Agnan [4], Michel Simioni [5]

[1] Toulouse School of Economics, INRA, University of Toulouse Capitole, France
[2] Department of Mathematics and Statistics, Thuongmai University, Hanoi, Vietnam
[3] BVA, 52 rue Marcel Dassault, Boulogne-Billancourt, France
[4] Toulouse School of Economics, University of Toulouse Capitole, France
[5] MOISA, INRA, University of Montpellier, Montpellier, France

April 25, 2018

Post-print version of the article published in: Statistical Methods in Medical Research, 2018, *online first, 21 p.*
http://journals.sagepub.com/doi/10.1177/0962280218770223

Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis

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

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

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

[3]BVA, 52 rue Marcel Dassault, Boulogne-Billancourt, France

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

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

April 25, 2018

Abstract: 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.

Keywords: Macronutrient shares, diet pattern, compositional regression models, expenditure elasticity, Vietnam

*Corresponding author:trinhthihuong@tmu.edu.vn

†michel.simioni@inra.fr

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 problems affect the

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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 descrip-

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

tive 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.

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

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 7 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 Trinh et al. (2017). 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.

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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 5 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 5 and boxplots in Figure 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)).

2.2 Diet pattern of Vietnamese households during 2004-2014

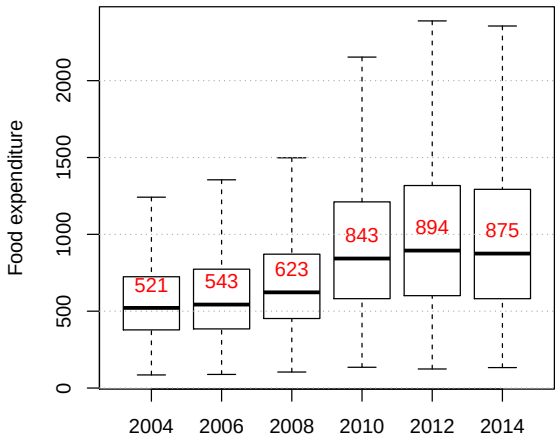
The diet pattern of Vietnamese households has changed dramatically from 2004 to 2014. The volumes of macronutrient consumption along time are presented in Figure 2. The median volume of per capita calorie intake (in red color) has increased from 2004 to 2014, except that there is a strong fall of PCCI 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

[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.

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 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.



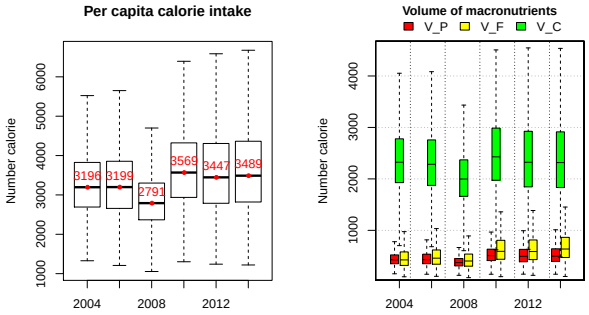
stable across the six years (except a decrease in 2008) while calories obtained from protein and fat have increased gradually.

Broadly speaking, during this ten-year period, the average protein share and the average fat share are between 10% and 20%, and the average carbohydrate share is between 60% and 80% (see Table 5). Figure 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 1). 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 characteris-

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 2: Per capita calorie intake and volume of macronutrient consumption.



tics, and in terms of diet patterns. The map in Figure 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.

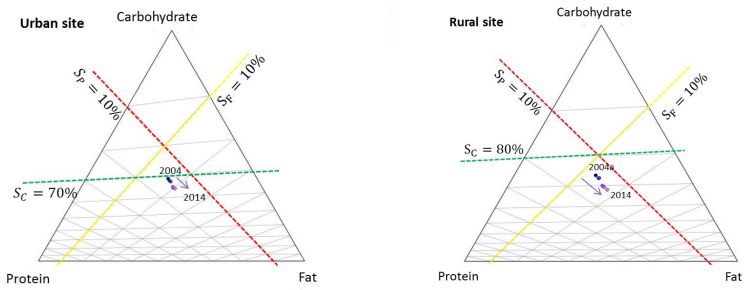
Table 1: Closed geometric mean of macronutrient shares in urban and rural sites.

	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 at its dispersion around this center. Figure 5 (left) represents in a ternary diagram the data in 2004, the data centers in 2004 and 2014, along with ellipses

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 3: Centered ternary diagrams of average macronutrient shares in urban and rural sites.



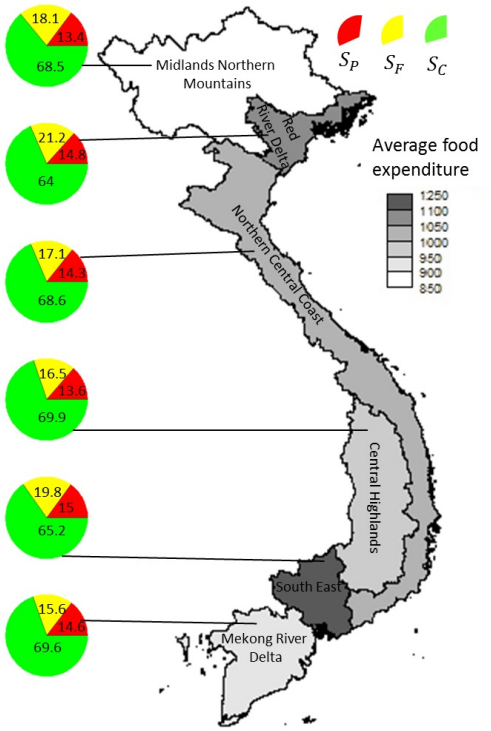
delimiting half of the population around these points in the simplex, Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). The "ideal" balanced diet according to the National Institute of Nutrition in Vietnam (S_P=14%, S_F=18%, S_C=68%) is represented by a triangle. This ternary diagram shows that half of the population in 2014 have a diet balance very close to the ideal one, closer than in 2004. In Figure 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 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 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). Mo-

[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)

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 4: Macronutrient shares and food expenditure averages by area in 2014.



reover, 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, in Figure 7. Because we have here a 3–part composition, the biplot explains 100% of the variance. Interestingly, the three components point towards dif-

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 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.

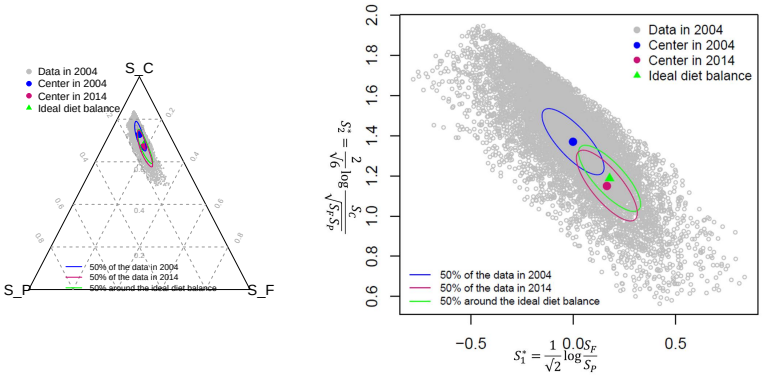
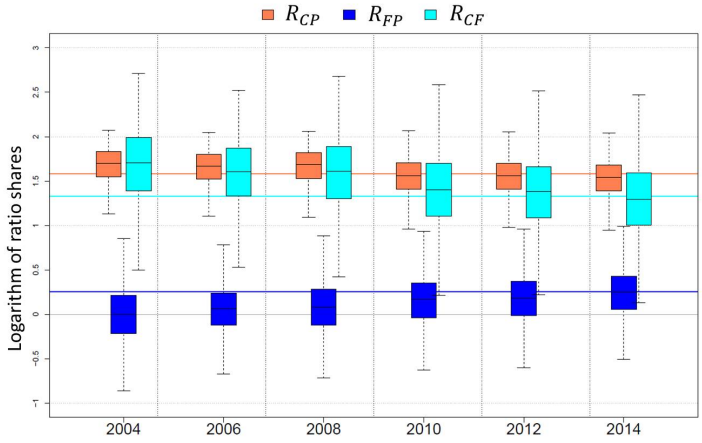


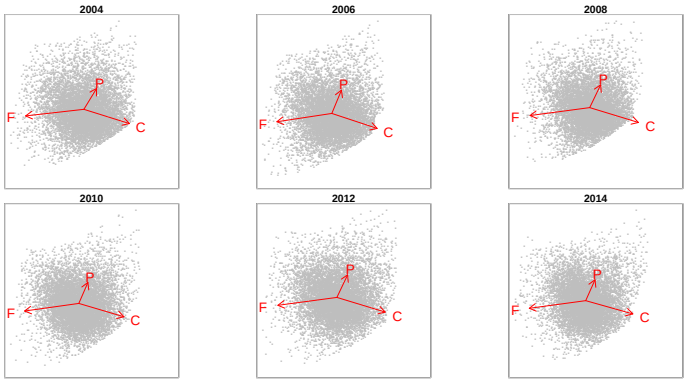
Figure 6: Boxplots of macronutrients log–ratio of shares by year. The line shows the value corresponding to the ideal diet for each log–ratio of share.



ferent 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 $\log(\frac{S_F}{S_C})$.

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 7: Covariance biplot of a principal component analysis of the macronutrient shares in each year. P, F, C correspond to Protein, Fat and Carbohydrate.



3 Compositional data analysis approach to describe and explain macronutrient consumption

3.1 Introduction to CODA

In the literature, different types of models are available for doing regression with shares, Morais et al. (2017b). 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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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:

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

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):

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

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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

positive continuous variable, used in logarithm, and others are categorical variables).

As proved in Morais et al. (2017a), model (1) can be written in a fashion similar to the classical attraction models used in the marketing literature (Cooper and Nakanishi (1989)):

$$S_{j,i} = \frac{a_j \prod_{k=1}^{K} b_{j,k}^{X_{ki}} \epsilon_{j,i}}{\sum_{m=1}^{D} a_m \prod_{k=1}^{K} b_{m,k}^{X_{ki}} \epsilon_{m,i}}. \tag{2}$$

As in Dumuid et al. (2017) and Muller et al. (2016), in order to fit and interpret model (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)):

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

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 (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 (1) can be obtained by inverse transformation from the estimated coefficients of the transformed model (3). For example, $\widehat{\mathbf{b}}_\mathbf{1} = \mathcal{C}(\exp(\mathbf{W}\widehat{\mathbf{b}}_\mathbf{1}^*))'$, where $\widehat{\mathbf{b}}_\mathbf{1}^* = (\widehat{b}_{1,1}^*, \widehat{b}_{2,1}^*)'$.

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].

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_T^2.

[7] Full results available upon request.

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Table 2 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 2: 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 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 (1) follow a "Gaussian distribution in the simplex", which is equivalent to say that error terms ϵ^*_j in (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 (3) are close to follow a Gaussian distribution although there is a heavy tailed distribution (see Figure 11 for the year 2010). Moreover, the residuals are symmetric according to the residuals log–ratios boxplots (see Figure 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.

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 6 summarizes the coefficients of the compositional model in ILR coordinates over the years whereas Table 8 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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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 δ 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.

4 Food expenditure elasticity of macronutrient consumption shares and volumes

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:

$$Elast(Y, X) = \frac{\frac{\partial Y}{Y}}{\frac{\partial X}{X}} = \frac{\partial \log Y}{\partial \log X}. \tag{4}$$

From equation (2) and Morais et al. (2017a), 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:

$$Elast(S_{j,i}, Exp_i) = \log b_{j,1} - \sum_{m=1}^{D} S_{m,i} \log b_{m,1}, \tag{5}$$

[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).

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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.

4.2 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. (2017a). For a small rate of change δ 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 8, and are summarized in Table 3, 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.

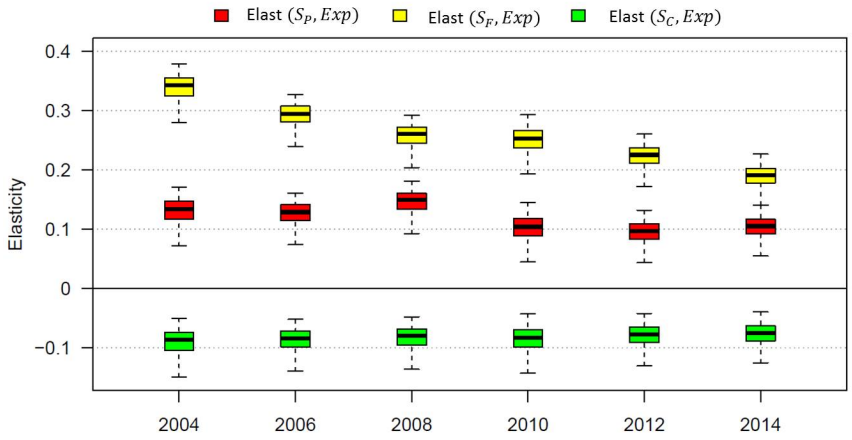
To give an example of interpretation of elasticity, let us consider for example a household in 2014 having an average diet balance, i.e (14.5%, 19.1%, 66.4%) for protein, fat and carbohydrate, and a food budget of US$1000. The corresponding elasticities are (0.1031, 0.1890, -0.0769) thus if we imagine 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.

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 (1) (one model by macronutrient

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Figure 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).



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

$$\begin{aligned} \log(V_{j,i}) &= \alpha_j + \beta_{j,1}\log(Exp_i) + \sum_{k=2}^{K}\beta_{j,k}X_{ki} + \varepsilon_{j,i} \qquad \text{for } j = 1,2,3 \\ \log(PCCI_i) &= \alpha + \beta_1\log(Exp_i) + \sum_{k=2}^{K}\beta_k X_{ki} + \varepsilon_i. \end{aligned} \tag{6}$$

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 β_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 9 represents the volume elasticities relative to the food expenditure across time. Table 3 compares elasticities obtained from the share

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

model (1) and the volume model (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)).

Figure 9: Food expenditure elasticities of macronutrient volumes and PCCI.

Table 3: 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

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 indicate that the volume of carbohydrate is the most complicated to estimate

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

using household characteristics. In contrast, fat and protein consumptions are well determined by the household characteristics we are using.

Table 4: 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

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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Table 5: VHLSS description variables. Averages correspond to arithmetic means for volume variables ($V_P, V_F, V_C, Exp, ExpTot, Engel$) and to closed geometric means for share variables (S_P, S_F, S_C).

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	$\leqslant$ 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	$\geqslant$ 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 %

Table 6: 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 ***

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Acknowledgements

We would like to thank the Editor of the review and two referees for their thoughtful comments and suggestions on the earlier draft of this paper. We would like to thank John GALLUP and NGUYEN Ngoc Hieu who re-calculated expenditure data used in this article. We are grateful to Thibault LAURENT for technical assistance in R. We thank the market research company BVA and the TAASE project of INRA-CIRAD GloFoodS meta-program for their support.

References

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.

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

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.

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

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

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.

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.

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

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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

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

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 (2017a). Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper 17*(805).

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

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

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

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.

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

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.

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.

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.

Trinh, T. H., M. Simioni, and C. Thomas-Agnan (2017). A fresh look at the nutrition transitionin vietnam using semiparametric modeling. *TSE working paper Sept*(17-842).

Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

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

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.

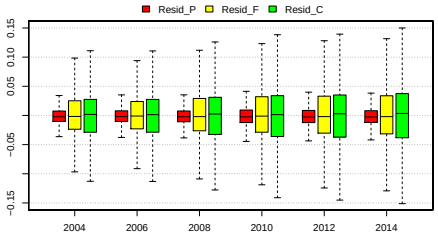
Appendix

Table 7: Conversion table Calories for Vietnam.

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

Amount per 100gr food ; protein contains 4 calories per gram and fat contains 9 calories per gram

Figure 10: Boxplots of absolute values of residuals by component and year.



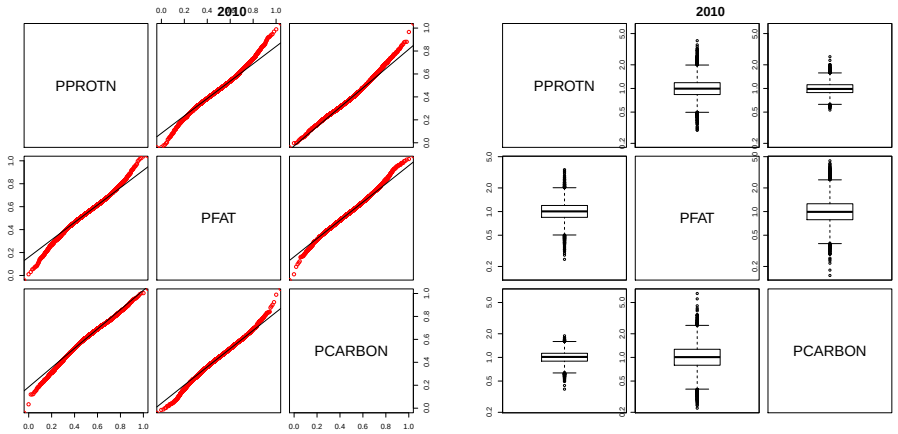
Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223

Table 8: 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 11: QQ-plot of residuals log ratios in 2010.

Figure 12: Boxplots of residuals log ratios in 2010.



Post-print version of the article published in : Statistical Methods in Medical Research, 2018, online first, 21p. http://journals.sagepub.com/doi/10.1177/0962280218770223