Supplementary Material

Supermarkets and household food acquisition patterns in Vietnam in relation to population demographics and socioeconomic strata: insights from public data

Thi Huong TRINH [1,2*], Dharani DHAR BURRA [1,*], Michel SIMIONI [3], Stef DE HAAN [4], Tuyen Thi Thanh HUYNH [1], Tung Van HUYNH [5] and Andrew D JONES [6]

[1]International Center for Tropical Agriculture (CIAT)-Asia Office, Hanoi, Vietnam

[2]Thuongmai University, Hanoi, Vietnam

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

[4]International Potato Center (CIP), Peru

[5]Can Tho City Institute for Socio-Economic Development, Vietnam

[6]School of Public Health, University of Michigan, USA

*** Correspondence:**
Corresponding Author
T.Huong@cgiar.org, D.Burra@cgiar.org

Poisson regression analysis

Poisson regression analysis (Cameron and Trivedi 2013) was performed to estimate the relationship between socio-economic demographic characteristics and year on acquired food diversity score (aFDS), as response variable. For each supermarket cluster, a full model with all explanatory variables and a null model with only the intercept as explanatory variable were constructed. A step wise variable selection and regression approach was employed. Residuals from both forward and backward step wise selection approach were used to assess the residual deviance, and a goodness of fit test was performed for each model, obtained from forward and backward selection. The model that was significantly different from the null model, and not different from a full model, and the one that obtained a relatively lower AIC (Akaike Information Criterion) score, in comparison to the full model, was selected and further described. *Sandwich* and *lmtest* packages in R were used to calculate the standard errors of the final model using a robust variance estimator to correct for over-dispersion in the residuals. The *msm* package was used to calculate incidence rate ratios. Each incidence rate ratio measures by which factor expected aFDS will increase (when the ratio is greater than one) or decrease (when the ratio is smaller than one) when the corresponding explanatory variable increases by one point, while holding all other variables in the model constant.

Compositional data analysis (CODA)

To estimate the relationship between socioeconomic, demographic and time variable impact on diet quality indices, CODA approach was performed separately for each supermarket cluster. Compositional data are those which contain only relative information and are parts of some whole. In most cases they are recorded as closed data, i.e. data summing to a constant, such as 100%-macronutrient shares in diet being a good example. Compositional data are always positive and range only from 0 to 100, or any other constant, when given in closed form. If one component increases, others must, perforce, decrease, whether or not there is a genetic link between these components. This means that the results of standard statistical analysis of the relationships between raw components or parts in a compositional data set are clouded by spurious effects. Several recent papers (Trinh et al. 2018; Solans et al. 2018) have applied compositional analysis to food consumption patterns.

A vector of macronutrient shares, or (S_P, S_F, S_P) is a composition of three positive components for which only the relative information is of interest as $S_C + S_F + S_P = 1$.

As three components are involved in the composition, these composition lies in an equilateral triangle surface in R^3, and it is common to represent the data in a ternary diagram such as Figure A-4, which is an equivalent representation. Compositional models are “transformation models" in the sense that they assume a distribution (generally Gaussian) for a log-ratio transformation of shares. Transformation models have several advantages compared to other share models: they are easy to estimate (usually by OLS on coordinates) and flexible in terms of explanatory variables (they can be compositional or non-compositional variables, with or without component-specific parameters) (Morais, Thomas-Agnan, and Simioni 2018). For instance, in our application, the three macronutrient shares are summarized in two Isometric Log-Ratios, or Ilr, coordinates.

It has been shown that ILR transformation of shares into real coordinates is very flexible as it preserves all the metric properties that characterize shares in the simplex (Egozcue et al. 2003).

The interpretation of regression parameters of above compositional regression models are complex. In our analysis, we follow the interpretation proposed in (Müller et al. 2018). Alternative interpretation is elasticity of shares, as in (Trinh et al. 2018). The entire regression analysis and processing of the compositional data was conducted in R statistical computing environment, using the *compositions* and *robCompositions* packages.

Table A-1. Incidence rate ratios obtained from Poisson regressions

	Supermarket clusters		
Predictor variables:	High SM	MedSM	LowSM
Per capita expenditure (US)	1.08***	1.09***	1.12***
Proportion of food consumed from purchases	1.21*	1.10***	1.13***
Urban residence (Reference level: Rural residence)		1.01*	1.01
Minority ethnic (Reference level: Kinh Ethnic)		0.98***	0.93***
(Reference level: Household with 2 members)	3.63*	1.05***	1.05***
Household with 4 members (Reference level: Household with 2 members)	3.75**	1.09***	1.10***
Household with 5 members (Reference level: Household with 2 members)	3.82***	1.12***	1.13***

Household with 6 members (Reference level: Household with 2 members)	3.91***	1.15***	1.14***
Midlands Northern Mountains (Reference level: Red River Delta)		1.00	0.94***
Northern Central Coast (Reference level: Red River Delta)		0.97**	0.96***
Central Highlands (Reference level: Red River Delta)		0.95*	0.96**
South East (Reference level: Red River Delta)	0.92***	0.97*	0.95**
Mekong River Delta (Reference level: Red River Delta)		0.96*	0.90***
Year 2012 (Reference level: Year 2010)		0.97*	0.98*
Year 2014 (Reference level: Year 2010)		0.99	0.98*

Results from Poisson regressions with aFDS as the response variable and household level social, economic and demographic variables obtained from VHLSS as predictor variables.
Poisson regressions are performed for each supermarket cluster. Reported figures are incidence rate ratios, wherein a value above 1 (resp. below 1) indicates positive (resp. negative) relation- ship between aFDS and the corresponding predictor variable. Asterisks correspond to significance at p-value < 0.05 for *, p-value < 0.01 for **, and p-value < 0.001 for ***.

Table A-2. Results for the compositional data regression model estimation

	Supermarket clusters					
Predictor variables	HighSM			MedSM		LowSM
Per capita expenditure (US)	-1e-04 ***	-1e-04 ***	-2e-04 ***	1e-04 ***	-3e-04 ***	1e-04 ***
Proportion of food consumed from purchases	-0.4206 ***	0.1764 ***	-0.1758 ***	0.00777	-0.2549 ***	0.1273 ***
Urban residence (Reference level: Rural residence)			-0.0247 ***	-0.0049	-0.0285 ***	0.0052
Minority ethnic (Reference level: Kinh Ethnic)			0.0409 ***	-0.0015	-1.4886 ***	0.0353 ***
Household with 3 members (Reference level: Household with 2 members)	1.5298 ***	0.0587 *	1.5217 ***	-0.0007	0.0371 ***	-0.0009
Household with 4 members (Reference level: Household with 2 members)	1.5029 ***	0.0615 *	1.5403 ***	-0.006	0.0475 ***	-0.0046
Household with 5 members (Reference level: Household with 2 members)	1.5171 ***	0.0719 *	1.571 ***	-0.0216 *	0.0721 ***	-0.0085
Household with 6 members (Reference level: Household with 2 members)	1.5169 ***	0.0523 .	1.5908 ***	-0.026	0.0922 ***	-0.0279 ***
Midlands Northern Mountains (Reference level:					0.0211 *	0.0231 ***

Red River Delta)						
Northern Central Coast (Reference level: Red River Delta)					0.1129 ***	-0.171 ***
Central Highlands (Reference level: Red River Delta)					0.0897 ***	-0.0928 ***
South East (Reference level: Red River Delta)	1.61E-02	-0.0878 ***			0.1025 ***	-0.0945 ***
Mekong River Delta (Reference level: Red River Delta)					0.101 ***	-0.2044 ***
Year 2012 (Reference level: Year 2010)			-0.0365 ***	0.0725 ***	-0.0345 ***	0.0692 ***
Year 2014 (Reference level: Year 2010)			-0.0209 **	0.0485 ***	-0.0366 ***	0.0457 ***
R^2 −adjust	0.657	0.894	0.324	0.958	0.294	0.953

Results from composition regression analysis with Ilr_1 and Ilr_2 as the outcomes and household level social, economic and demographic variables obtained from VHLSS as predictor variables. Each regression corresponds to a supermarket cluster. Numbers indicate regression coefficients, wherein values positive values signify positive and negative values signify negative relationship between the outcome variables and the corresponding predictor variable from VHLSS. Asterisks correspond to significance at p-value < 0.05 for *, p-value < 0.01 for ** and p-value < 0.001 for ***.

Reference

Cameron, A C, and P K Trivedi. 2013. *Regression Analysis of Ccount Data*. Cambridge: Cambridge University Press.

Egozcue, J J, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barceló-Vidal. 2003. “Isometric Logratio Transformations for Compositional Data Analysis.” *Mathematical Geology* 35 (3): 279–300.

Morais, Joanna, Christine Thomas-Agnan, and Michel Simioni. 2018. “Interpretation of Explanatory Variables Impacts in Compositional Regression Models.” *Austrian Journal of Statistics* 47 (5): 1–25.

Müller, Iva, Karel Hron, Eva Fiserova, Jan Smahaj, Panajotis Cakirpaloglu, and Jana Vancakova. 2018. “Interpretation of Compositional Regression with Application to Time Budget Analysis.” *Austrian Journal of Statistics* 47 (2): 3–19.

Solans, M, G Coenders, R Marcos-Gragera, A Castelló, E Gràcia-Lavedan, Y Benavente, V Moreno, et al. 2018. “Compositional Analysis of Dietary Patterns.” *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280218790110.

Trinh, Thi Huong, Joanna Morais, Christine Thomas-Agnan, and Michel 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.