JNS
JOURNAL OF NUTRITIONAL SCIENCE

RESEARCH ARTICLE

Log-ratio transformations for dietary compositions: numerical and conceptual questions

Maria Léa Corrêa Leite*

National Research Council/Institute of Biomedical Technologies, Milan, Italy

(Received 9 June 2021 – Final revision received 12 October 2021 – Accepted 15 October 2021)

Journal of Nutritional Science (2021), vol. 10, e97, page 1 of 7 doi:10.1017/jns.2021.93

Abstract

When evaluating the impact of macronutrient intakes on health outcomes, researchers in nutritional epidemiology are mostly interested in two types of information: the relative importance of the individual macronutrients and the absolute effect of total energy intake. However, the usual substitution models do not allow these separate effects to be disentangled. Dietary data are typical examples of compositional data, which convey relative information and are, therefore, meaningfully expressed in the form of ratios. Various formulations of log-ratios have been proposed as a means of analysing compositional data, and their interrelationships when they are used as predictors in regression models have been previously reported. This note describes the application of distinct log-ratio transformations to the composition of dietary macronutrients and discusses the interpretative implications of using them as explanatory variables in regression models together with a term for the total composition (total energy intake). It also provides examples that consider serum glucose levels as the health outcome and are based on data coming from an Italian population-based study. The log-ratio transformation of dietary data has both numerical and conceptual advantages, and overcomes the drawbacks of traditional substitution models.

Key words: **Compositional data: Dietary data: Energy intake: Log-ratio transformation: Macronutrients: Nutrient balances**

Introduction

Isocaloric substitution analysis has been considered the gold standard for nutritional studies aimed at evaluating the relationships between macronutrient intake and the risk of disease because it provides a means of assessing the effects of replacing specified nutrients on a health outcome while adjusting for total energy intake.

For example, taking y as a health outcome, EC, EP and EF as respectively representing the dietary intakes of energy coming from carbohydrates, proteins and fats, and TE as their sum (total energy intake), the isocaloric substitution model that leaves out EF would be:

$$f(y) = B_{s1}EC + B_{s2}EP + B_{s3}TE,$$

in which B_{s1} is interpreted as expressing the specific effect of replacing fats with carbohydrates while keeping total energy and protein intakes constant. Alternatively, the energy partition model:

$$f(y) = B_{p1}EC + B_{p2}EP + B_{p3}EF,$$

can be used to obtain replacement effects by estimating them as the difference between B_p coefficients. These approaches are equivalent, as it can be easily seen that $B_{s1}EC = B_{p1}EC - B_{p3}EF$, $B_{s2}EP = B_{p2}EP - B_{p3}EF$ and $B_{s3}TE = B_{p3}EF$.

Arnold *et al.*[1] have recently provided a conceptual description of two potential types of causal effects for compositional data such as dietary data. In their case, $B_{p(s)}$ represents the *total effects* (not conditioned on TE) of adding each macronutrient to the diet regardless of the intake of all of the other macronutrients, consequently increasing total energy 'without altering other consumption behaviours', while $B_{s(s)}$ represents the *relative effects* arising from the isocaloric replacement of fat with

* **Corresponding author:** Maria Léa Corrêa Leite, email lea.correa@itb.cnr.it

carbohydrate (B_{s1}) or protein calories (B_{s2}). These are, therefore, the joint effects of reducing fat consumption and increasing the consumption of other macronutrients. In brief, the authors assume that the interdependence of macronutrient effects arises from the conditioning on total energy intake.

However, these models cannot unequivocally disentangle the different effects of macronutrients and total energy intake because they are unsuitable for describing the intrinsic nature of variations in dietary components. Although they overcome the numerical problem of perfect data collinearity, they raise a conceptual issue: the compositional nature of the data means that the information conveyed by the compositional parts is inherently relative regardless of the inclusion of their total in the model, and so any attempt to estimate the effects of isolated variations can lead to misleading results.

In order to illustrate some concepts intuitively and schematically, let us consider a typical daily diet of 2000 calories, of which 55 % come from carbohydrates 15 % from proteins and 30 % from fats (diet A in Table 1). The macronutrient composition represents the qualitative aspect of the diets and the total calories (last column) its quantitative aspect.

The only difference between diet B and diet A is the increase in the number of calories coming from carbohydrates. In the framework of an energy partition model, interpreting the effect of adding EC as the only variation is somewhat misleading because the consequent increase in total calories (in italics to indicate its exclusion from the model) leads to changes in the 'weights' of the components of the composition: for example, the 26 % of calories coming from fat in diet B is qualitatively different from the 30 % in diet A. In diet C, calories from carbohydrates isocalorically replace those coming from fats in diet A, the effect of which would be estimated in a substitution model by including the term for total energy intake and excluding that of EF. However, this is not the only qualitative change because the relationship between protein and fat content substantially changes from 1:2 in diet A to 1:1 in diet C, thus raising the question of whether it is metabolically irrelevant. Finally, diet D illustrates how the variation in TE coincides entirely with the variation in the component left out of the model: this is actually the only component free to vary, and therefore, the only component contributing to the increases in total energy intake. It is, thus, clear that substitution methods cannot provide appropriate estimates of the effect of total energy intake.

Nutrition researchers may not only be interested in obtaining isocaloric estimates of nutrient effects but also in achieving a meaningful estimate of the effect of total energy intake regardless of the qualitative nature of a diet (particularly, in the context of obesity and related metabolic disorders). In order to illustrate this intuitively, let us consider three diets (E, A and F in Table 2) in which the absolute (quantitative) aspect increases and these changes are fully represented by the total number of calories, while their qualitative aspect (macronutrient composition) remains unchanged. Although these diets are qualitatively similar, it can be seen that the differences between the amounts of macronutrient increase together with the total number of calories, for example, the difference between the calories coming from fat and the calories coming from proteins is 150 in diet E, 300 in diet A and 450 in diet F. It is, therefore, clearly impossible to disentangle the quantitative and qualitative aspects of the diets on the basis of the differences between their compositional parts. On the other hand, the relationships (ratios) between the parts remain constant in all three diets, and it would, therefore, seem to be logical that ratios are more suitable than differences when it comes to describing the characteristics of a composition.

The basic principle of compositional data analysis is that 'any meaningful function of a composition can be expressed in terms of ratios of the components of the composition'[2] insofar as ratios are the natural means of describing variations that are intrinsically relative.

Log-ratio transformations of compositional data

Compositional data can be defined as positive vectors of parts of a whole that convey relative information[3]. Given that interest lies in the relative amounts of compositional components and that logging the ratios is a convenient means of making them more easily manageable mathematically, various expressions of log-ratios have been proposed for the analysis of compositional data.

Log-ratio transformations produce new variables that are amenable to being analysed using standard statistical methods, and a family of transformations has been introduced within the framework of log-ratio methodology. In the case of a D-part composition $(x_1, x_2, \ldots, x_D)$, the additive log-ratio (alr) transformation involves the division of each $D-1$ component by one that is arbitrarily chosen, for example, the last:

$$\mathrm{alr}(\mathbf{x}) = \left[\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \ \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right],$$

while the centred log-ratio (clr) transformation is defined by

$$\mathrm{clr}(\mathbf{x}) = \left[\ln\left(\frac{x_1}{g(\mathbf{x})}\right), \ln\left(\frac{x_2}{g(\mathbf{x})}\right), \ \ldots, \ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right],$$

where $g(\mathbf{x})$ is the geometric mean of the components of the composition[4].

Table 1. Characteristics of four fictitious diets

Diet	Macronutrient composition: kcal (%) EC	EP	EF	Total calories (%) TE
A	1100 (55·0)	300 (15·0)	600 (30·0)	2000 (100·0)
B	1400 (61·0)	300 (13·0)	600 (26·0)	*2300 (100·0)*
C	1400 (70·0)	300 (15·0)	*300 (15·0)*	2000 (100·0)
D	1100 (48·0)	300 (13·0)	*900 (39·0)*	2300 (100·0)

Table 2. Characteristics of three fictitious diets

Diet	Macronutrient composition: kcal (%) EC	EP	EF	Total calories (%) TE
E	550 (55·0)	150 (15·0)	300 (30·0)	1000 (100·0)
A	1100 (55·0)	300 (15·0)	600 (30·0)	2000 (100·0)
F	1650 (55·0)	450 (15·0)	900 (30·0)	3000 (100·0)

Subsequently, Egozcue *et al.*[5] proposed the isometric log-ratio (ilr) transformation which, in its particular expression of *balances*, consists of orthogonally decomposing the parts of a whole into non-overlapping subgroups and representing their relationships. One way of constructing orthonormal balances is to use sequential binary partition (SBP) as a bifurcating tree: the parts of a composition are successively and hierarchically split into two groups until all of the groups have a single part[6]. At each of the $D-1$ steps required to complete the partition, a generic balance is defined as the orthonormal log-ratio of the geometric mean of each group of parts:

$$\mathrm{ilr_i} = \frac{\sqrt{r_i \cdot s_i}}{\sqrt{r_i + s_i}} \cdot \ln\left(\frac{(x_1^+ \cdot x_2^+ \cdot \ldots \cdot x_r^+)^{1/r_i}}{(x_1^- \cdot x_2^- \cdot \ldots \cdot x_s^-)^{1/s_i}}\right), \quad \text{for } i = 1 \text{ to } D-1$$

where the square root coefficient is the normalising constant, and r and s are respectively the number of parts in the numerator (x^+) and the number of parts in the denominator (x^-).

One particular choice of partition has been proposed[7] in which the ilr transformation is defined as:

$$\mathrm{ilr}_i = \frac{\sqrt{(D-i)}}{\sqrt{(D-i+1)}} \cdot \ln\left(\frac{x_i}{g(x_j)}\right), \quad \text{for } i = 1 \text{ to } D-1$$

where $g(x_j)$ is the geometric mean of the compositional parts for $j = i+1$ to D. The particular characteristic of this approach is that the first new variable (ilr_1) captures all of the relevant information about the compositional part x_1, and is, thus, called the *pivot balance*[8]. A number of D sets of $D-1$ ilr(s) are defined, each of which has a different compositional part in the numerator of the pivot balance.

Macronutrient log-ratios as explanatory variables

This section describes how these different log-ratio transformations can be applied to dietary macronutrient compositions and examines the interpretative implications of their use as explanatory variables in regression models, on the basis of some elements that have recently been discussed[9]. To this end, imagine a five-part composition of dietary proteins (PR), starches (ST), simple sugars (SS), unsaturated fats (UF) and saturated fats (SF) as sources of energy, and then consider the definition of balances we have previously suggested as a means of characterising dietary exposure[10–12]. The partitioning procedure can be carried out using the following sign matrix, where the plus sign indicates that the part is assigned to the numerator of the balance, the minus sign indicates that it is assigned to the denominator, and 0 indicates that it is not involved in that particular balance:

PR	ST	SS	UF	SF
−1	−1	−1	+1	+1
+1	−1	−1	0	0
0	+1	−1	0	0
0	0	0	+1	−1

The corresponding balances can be represented as explanatory variables as follows:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_1 \ln\left(\frac{\sqrt[2]{\mathrm{UF} \cdot \mathrm{SF}}}{\sqrt[3]{\mathrm{PR} \cdot \mathrm{ST} \cdot \mathrm{SS}}}\right) + \mathrm{B}_2 \ln\left(\frac{\mathrm{PR}}{\sqrt[2]{\mathrm{ST} \cdot \mathrm{SS}}}\right) + \mathrm{B}_3 \ln\left(\frac{\mathrm{ST}}{\mathrm{SS}}\right) + \mathrm{B}_4 \ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right)$$

but note that the normalising constant has been removed, consequently reducing the orthonormality of the log-contrasts to simple orthogonality. This simplification has been previously proposed by Muller *et al.*[13], in order to enhance the interpretability of the results of regression analyses involving ilr coordinates as explanatory variables.

As in any multiple regression framework, one B-coefficient represents the change in response associated with a one unit change in the corresponding log-ratio, while keeping all of the other log-ratios constant. Adding on the log scale is equivalent to multiplying on the natural scale and since $1 = \ln(e)$, B_1 therefore represents the expected change in the dependent variable when the ratio between the geometric mean of the dietary fats and the geometric mean of the other macronutrients (proteins and carbohydrates) is multiplied by e (approximately 2·7). It is worth noting that using the binary logarithm (base 2) may often be useful as it corresponds to the effect of doubling the ratio. This effect may be subject to intrinsic confounding because an increase in fat consumption may variously involve saturated and/or unsaturated fats and, similarly, a decrease in the denominator may be variously related to proteins and/or carbohydrates. However, these confounding factors can be appropriately controlled by including in the model the other balances, whose embedded ratios are kept constant. This ensures that variations in the B_1-related ratio are made in such a way as to increase UF and SF by a common factor and decrease PR, ST and SS by a common factor. As another example, B_4 expresses the effect of multiplying the UF:SF ratio by 2·7 without modifying the relationships between the other macronutrients as represented by the ratios in the equation.

Using the simplified (orthogonal rather than orthonormal) pivot balance approach, five regression equations can be drawn up by defining the pivot balance for each of the five components and using five runs. The first is:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_{1(\mathrm{PPR})} \ln\left(\frac{\mathrm{PR}}{\sqrt[4]{\mathrm{ST} \cdot \mathrm{SS} \cdot \mathrm{UF} \cdot \mathrm{SF}}}\right) + \mathrm{B}_2 \ln\left(\frac{\mathrm{ST}}{\sqrt[3]{\mathrm{SS} \cdot \mathrm{UF} \cdot \mathrm{SF}}}\right) + \mathrm{B}_3 \ln\left(\frac{\mathrm{SS}}{\sqrt[2]{\mathrm{UF} \cdot \mathrm{SF}}}\right) + \mathrm{B}_4 \ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right),$$

the second is:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_{1(\mathrm{PST})}\ln\left(\frac{\mathrm{ST}}{\sqrt[4]{\mathrm{PR}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_2\ln\left(\frac{\mathrm{PR}}{\sqrt[3]{\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_3\ln\left(\frac{\mathrm{SS}}{\sqrt[2]{\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_4\ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right),$$

and so on. At each run, we are interested in inferences about the first balance which captures all of the available information about the part in its numerator. Thus, $\mathrm{B}_{1(\mathrm{PPR})}$ represents the expected change in the dependent variable when the balance between proteins against all of the other macronutrients is multiplied by an *e*-factor of 2·7. This variation is the result of increasing the protein component while reducing all of the other parts by a common factor as is ensured by the model's inclusion of the other balances. The $\mathrm{B}_{1(\mathrm{PST})}$ coefficient is interpreted in the same way but concerns starches.

Let us now consider the model that includes four alr(s) as predictors:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_{1_a}\ln\left(\frac{\mathrm{PR}}{\mathrm{SF}}\right) + \mathrm{B}_{2_a}\ln\left(\frac{\mathrm{ST}}{\mathrm{SF}}\right) + \mathrm{B}_{3_a}\ln\left(\frac{\mathrm{SS}}{\mathrm{SF}}\right) + \mathrm{B}_{4_a}\ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right)$$

The coefficient B_{1_a} represents the expected change in outcome when PR increases and SF decreases in such a way that the PR:SF ratio increases 2·7 times, while keeping all of the other terms in the model constant. This necessarily means that ST, SS and UF decrease by the same factor as SF. Thus, although the other parts are not explicitly present in the denominator of the ratio, B_{1_a} measures the same effect as $\mathrm{B}_{1(\mathrm{PPR})}$ and B_{2_a} measures the same effect as $\mathrm{B}_{1(\mathrm{PST})}$ and so on.

Finally, consider the case of a model including clr(s) as explanatory variables. Because of the perfect collinearity $\left(\sum_i \mathrm{clr}_i = 0\right)$, one clr (for example, the one with SF in the numerator) is left out of the equation:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_{1_c}\ln\left(\frac{\mathrm{PR}}{\sqrt[5]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_{2_c}\ln\left(\frac{\mathrm{ST}}{\sqrt[5]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_{3_c}\ln\left(\frac{\mathrm{SS}}{\sqrt[5]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right) + \mathrm{B}_{4_c}\ln\left(\frac{\mathrm{UF}}{\sqrt[5]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}}}\right)$$

Increasing a given clr while keeping the remaining log-ratios in the equation constant requires a decrease in the omitted clr by the same amount: for example, adding one unit [ln(*e*)] to ln(PR/gm), where gm is the geometric mean of the five components, implies reducing ln(SF/gm) by the same amount, and as a result, the difference [ln(PR/gm) – ln (SF/gm)] = ln(PR/SF) increases by 2 ln(*e*). That is to say, B_{1_c} represents the expected change in the dependent variable when the ratio between PR and SF is multiplied by $2{\cdot}7^2$, and this also applies to the other regression coefficients in the equation. Curiously, in this case, although the denominator of the ratios includes all of the compositional parts, the estimated effects are related to pair-wise ratios.

Accounting for total energy intake

As said in the introduction, when investigating the relationship between dietary macronutrient composition and health outcome, nutritional researchers may be interested not only in examining the isocaloric effects of relative variations in macronutrient intakes, but also in obtaining a meaningful estimate of the effect of total energy intake. In other words, not only the relative size of the compositional parts but also their absolute variable size may be relevant.

The statistical properties of alternative ways of computing a total of compositional observations have been described by Pawlowsky-Glahn *et al.*[14], who have stated that a product-based total such as $T_{\mathrm{p}} = \sqrt{D}\,\ln(\sqrt[D]{x_1\cdot x_2\cdot\ldots\cdot x_D})$ is isometric (which means that it preserves the same distances as in the logarithm value space), but a sum-based total such as $T_{\mathrm{s}} = \ln(x_1 + x_2 + \cdots + x_D)$, which looks more like what is usually understood as a total, can also be used.

Since then, Coenders *et al.*[15] have discussed cases in which a composition and its total act as explanatory variables and described the interpretative implications of using these different formulations of the total. Going back to our five-part macronutrient composition, the alternative models containing four balances and a total are:

$$f(y) = \mathrm{B}_0 + \mathrm{B}_1\ln\left(\frac{\sqrt[2]{\mathrm{UF}\cdot\mathrm{SF}}}{\sqrt[3]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}}}\right) + \mathrm{B}_2\ln\left(\frac{\mathrm{PR}}{\sqrt[2]{\mathrm{ST}\cdot\mathrm{SS}}}\right) + \mathrm{B}_3\ln\left(\frac{\mathrm{ST}}{\mathrm{SS}}\right) + \mathrm{B}_4\ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right) + \mathrm{B}_{T\mathrm{p}}(T_{\mathrm{p}}),$$

where $T_{\mathrm{p}} = \sqrt{5}\,\ln(\sqrt[5]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}\cdot\mathrm{UF}\cdot\mathrm{SF}})$, and

$$f(y) = \mathrm{B}_0 + \mathrm{B}_1\ln\left(\frac{\sqrt[2]{\mathrm{UF}\cdot\mathrm{SF}}}{\sqrt[3]{\mathrm{PR}\cdot\mathrm{ST}\cdot\mathrm{SS}}}\right) + \mathrm{B}_2\ln\left(\frac{\mathrm{PR}}{\sqrt[2]{\mathrm{ST}\cdot\mathrm{SS}}}\right) + \mathrm{B}_3\ln\left(\frac{\mathrm{ST}}{\mathrm{SS}}\right) + \mathrm{B}_4\ln\left(\frac{\mathrm{UF}}{\mathrm{SF}}\right) + \mathrm{B}_{T\mathrm{s}}(T_{\mathrm{s}}),$$

where $T_s = \ln(\mathrm{PR} + \mathrm{ST} + \mathrm{SS} + \mathrm{UF} + \mathrm{SF}) = \ln(\text{total energy})$

Both $\mathrm{B}_{T\mathrm{p}}$ and $\mathrm{B}_{T\mathrm{s}}$ express the effect of increasing the overall size of the macronutrient composition while keeping the relative importance of its component parts constant: i.e., they increase in the same proportion. However, while $\mathrm{B}_{T\mathrm{p}}$ is related to multiplicative changes in the not very intuitive quantity given by $(\mathrm{PR}{\cdot}\mathrm{ST}{\cdot}\mathrm{SS}{\cdot}\mathrm{UF}{\cdot}\mathrm{SF})^{1/\sqrt{5}}$, $\mathrm{B}_{T\mathrm{s}}$ is more simply related to multiplicative changes in total energy intake. The interpretation of the balance-related coefficients B_1 to B_4 changes only slightly depending on which *total* is held constant. As the coefficients refer to the effect of increasing the relative balance while keeping the remaining terms constant, it implies that all of the parts in the numerator of the balance increase in

the same proportion and those in the denominator decrease in the same proportion. Furthermore, if the total included in the model is B_{T_P}, the increase in the numerator of the balance is exactly counteracted by the decrease in the denominator. This perfect offset is not assured when the sum-based total (B_{T_S}) is used, but the fact that B_{T_P} can be mathematically considered the 'true total' does not mean that using B_{T_S} is a worse choice. In the particular case of dietary data, total energy intake as a sum of the calories from the different macronutrient sources is a significant characteristic of study subjects. Moreover, isometry is not a requirement when using compositions as explanatory variables.

Example: dietary macronutrients and serum glucose levels in non-diabetic subjects

This example uses data from the Italian Bollate Eye Study[16], a population-based study involving participants aged 40–74 years that was carried out in 1992–1993. The participants' dietary habits were assessed by means of a food-frequency questionnaire, and their mean daily nutrient intakes were calculated using the food compositional database compiled for epidemiological studies in Italy[17]. The study was approved by the Ethics Committee of the Italian National Research Council (CNR).

Various linear regression models with serum glucose level as the dependent variable were fitted and, as the concentration units (mg/ml) indicate compositional information, their values were logged in order to allow an estimate of relative changes. In addition to the dietary variables (total energy and fibre intake, and macronutrient log-ratios), all of the models included terms for gender, age, practising sport, television watching time, smoking and alcohol consumption. In order to simplify the interpretation of effect size, logarithms of the independent dietary variables were computed using base 2 [remember that $\log_2(z) = \ln(z)/\ln(2)$], and so the coefficients represent the effect of doubling the amounts of calories, fibre and the ratio in question. Tables 3–6 show the results of running the models, including a different formulation of the log-ratios or the total each time. The examination of the residual plots did not show any particular trend in their distribution and did not reveal deviations from linearity or homocedasticity, thus indicating that the models are adequately specified (data not show). As has been previously shown[9], the different formulations of the log-ratios represent simple reparametrisations of the same model and, as a result, the models performed equally and the estimates related to covariates common to the different models were the same.

Table 3 shows the results for macronutrient balances constructed as described above. It can be seen that the inclusion of the different formulations of the compositional total (T_P in model A and T_S in model B) had little effect on the fibre- and the balance-related coefficients. A two-fold increase in the protein:carbohydrate balance (shown in such a way that proteins increase and the two types of carbohydrate decrease by a common factor, while all of the other terms remain constant) is related to a expected multiplicative change in the outcome of $e(0{\cdot}0435) = 1{\cdot}0445$, where 0·0435 is the regression coefficient estimated for the protein:carbohydrate balance in model B). This means that the serum glucose level is expected to increase by 4·45 % (95 % confidence interval 1·79, 7·17 %).

Table 3. Results of the linear regression analysis of serum glucose levels [ln(mg/ml)] in relation to total energy and fibre intake and macronutrient balances (orthogonal coordinates)

	Coefficient	Standard error	*P*-value
Model A			
Total energy (T_p)	−0·0038	0·0078	0·627
Total fibre	0·0191	0·0144	0·186
Macronutrient balances:			
Fats *v.* proteins, carbohydrates	0·0195	0·0131	0·135
Proteins *v.* carbohydrates	0·0433	0·0130	0·001
Starches *v.* simple sugars	−0·0069	0·0061	0·256
Unsaturated *v.* saturated fats	−0·0105	0·0136	0·442
Model B			
Total energy (T_s)	−0·0030	0·0173	0·860
Total fibre	0·0154	0·0145	0·288
Macronutrient balances:			
Fats *v.* proteins, carbohydrates	0·0169	0·0120	0·158
Proteins *v.* carbohydrates	0·0435	0·0131	0·001
Starches *v.* simple sugars	−0·0075	0·0063	0·236
Unsaturated *v.* saturated fats	−0·0083	0·0133	0·533

Tables 4 and 5 show exactly the same results as the differences between them mainly regard the amount of work required. Table 4 shows the coefficients of only the first balance (pivot) that emerged from the five runs, at each of which a set of four balances (corresponding to a distinct partition) was included in the regression equation. Table 5 shows the results of the analysis of the alr transformations. We first included the four alr(s) calculated by dividing each macronutrient by saturated fats and then, in order to obtain estimates for this last, the second run included the four alr(s) with unsaturated fats as the denominator. The 'starches *v.* other macronutrients' coefficient in Table 4 indicates that increasing starches against a reduction in the other components by a common factor in such a way that the ratio is doubled leads to a decrease of 3·43 % in the serum glucose level. The identical effect in Table 5 seems to be attributable to the pair-wise ratio 'starches *v.* saturated fats', but it is important to note that the inclusion of the other alr-coordinates

Table 4. Results of the linear regression analysis of serum glucose levels [ln(mg/ml)] in relation to total energy and fibre intake and macronutrient balances (simplified pivot coordinates)

	Coefficient	Standard error	*P*-value
Total energy (T_s)	−0·0030	0·0173	0·860
Total fibre	0·0154	0·0145	0·288
Pivot balance (the first coordinate in each of the five runs)[a]:			
Proteins *v.* other macronutrients	0·0378	0·0155	0·015
Starches *v.* other macronutrients	−0·0349	0·0090	0·000
Simple sugars *v.* other macronutrients	−0·0199	0·0080	0·014
Unsaturated fats *v.* other macrontrients	0·0002	0·0134	0·990
Saturated fats *v.* other macronutrients	0·0167	0·0156	0·284

[a] Each run involves four balances, of which the first is the pivot.

Table 5. Results of the linear regression analysis of serum glucose levels [ln(mg/ml)] in relation to total energy and fibre intake and macronutrient additive log-ratio (alr) coordinates

	Coefficient	Standard error	*P*-value
Total energy (T_s)	−0·0030	0·0173	0·860
Total fibre	0·0154	0·0145	0·288
Alr-coordinates:			
Proteins *v.* saturated fats	0·0378	0·0155	0·015
Starches *v.* saturated fats	−0·0349	0·0090	0·000
Simple sugars *v.* saturated fats	−0·0199	0·0080	0·014
Unsaturated fats *v.* saturated fats	0·0002	0·0134	0·990
Saturated fats *v.* unsaturated fats[a]	0·0167	0·0156	0·284

[a] Obtained in a different run that included in the model the four alr(s) with unsaturated fats as the denominator.

held constant in the model implies that the other components should decrease in the same proportion as saturated fats; and this also applies to the other alr(s). As would be expected, the estimates related to total energy (T_s) and fibre intake are the same as those shown in Table 3.

Table 6 shows the results of the regression analysis including the clr(s) coordinates. In this case, the effects should be interpreted as relating to variations in the ratios of each macronutrient against saturated fats, which is in the numerator of the omitted clr. The coefficient for 'starches *v.* *g*(.)' should be interpreted as the effect related to a four-fold (2^2) increase in the starches:saturated fats ratio. As would be expected, the coefficient relating to a four-fold increase in the unsaturated:saturated fats ratio (−0·0209) is double that of the same ratio estimated in the ilr model (Table 3, model A). Note that the estimates related to total energy (T_p) in Table 6 are the same as those in Table 3.

Discussion

On the basis of the modern definition of compositional data, whenever researchers address the relative importance of data components, they are dealing with a compositional problem[3]. The compositional nature of dietary data does not arise from the constraint strictly imposed by conditioning on totals, but is inherent to the dynamics of our way of eating. We do not eat the component parts separately but our food consists of nutrient compositions, and diets rich in some components tend to be rich or poor in others depending on the manner we combine the foods we eat.

Table 6. Results of the linear regression analysis of serum glucose levels [ln(mg/ml)] in relation to total energy and fibre intake and macronutrient centred log-ratio (clr) coordinates

	Coefficient	Standard error	*P*-value
Total energy (T_p)	−0·0038	0·0078	0·627
Total fibre	0·0191	0·0144	0·186
Clr-coordinates:			
Proteins *v.* *g*(.)[a]	0·0166	0·0274	0·546
Starches *v.* *g*(.)	−0·0553	0·0175	0·002
Simple sugars *v.* *g*(.)	−0·0415	0·0208	0·047
Unsaturated fats *v.* *g*(.)	−0·0209	0·0272	0·442
Saturated fats *v.* *g*(.) (omitted)			

[a] *g*(.): geometric mean of all of the components of the macronutrient composition.

Although the compositional nature of dietary data is widely acknowledged, the proposed approaches to compositional data analysis continue to be largely ignored by nutritional analysts. By their very nature, compositional data convey relative information and their expression as ratios is the basic principle of compositional data analysis. Working with ratios is the key difference between compositional data analysis and standard isocaloric substitution modelling which, as it is based on differences and does not allow the relative and absolute aspects of the data to be disentangled, cannot provide a meaningful estimate of the effect of total energy intake.

In this note, we describe the use of different log-ratio transformations of dietary macronutrient composition and discuss the interpretative implications of using them as explanatory variables. As models including different log-ratio formulations provide comparable goodness-of-fit parameters, the choice of which transformation to use should be based on the subject of interest and the interpretative elements. We have previously shown the usefulness of using ilr coordinates in their particular expression as balances as a means of characterising dietary exposure[12]. This procedure has the great advantage of being flexibly adaptable to different research questions of interest. Orthogonal balances are new variables that convey non-redundant information and, in the particular case of macronutrient compositions, balances can be easily defined on the basis of sequential binary partitions that follow the components' natural clustering.

For reasons of simplicity, we have only considered macronutrient compositions but, of course, the questions raised also apply to compositions of micronutrients (vitamins and minerals): for example, researchers may be interested in evaluating the impact of the relative dominance of one vitamin over all of the others. However, although a pivot balance approach may be appropriate, it has the inconvenience of requiring a number of runs that is equal to the number of components in the composition. Nevertheless, as recently pointed out by Coenders[9], alr and simplified pivot coordinates are explanatory equivalents, and as shown in the example, the use of alr transformations as predictors and only two runs lead to the same result as the pivot approach, and so the larger the size of the composition, the greater the amount of work that is saved.

Using dietary log-ratios as explanatory variables not only makes it possible to obtain a meaningful description of the isocaloric interdependence of the components of a dietary composition, but also and simultaneously provides an interpretable estimate of the effect of total energy intake. It is, therefore, possible to evaluate the relative importance of the different parts of the composition while avoiding the fallacious interpretations that may emerge from the raw component analysis. As suggested in the introduction, some attempts to estimate absolute and unconfounded effects may be illusory because changing the concentration of one of the components clearly alters the relationships of the entire composition, while working with ratios provides a means of governing proportional relationships and allows confounding to be finely controlled.

The control provided by the inclusion in the model of the complete set of log-ratios is a key point when examining the explanatory role of a composition. The interpretation of

effects may not directly correspond to the way in which the log-ratios are constructed: although their formulation may suggest increasing one component in relation to another (alr) or all of the other components (clr), when they are included as predictors in a regression equation, the related coefficients express the effects of pair-wise ratios (clr) or those in relative terms to all of the components (alr).

In addition to numerical factors, the interest in estimates based on relative variations in dietary elements may also be due to conceptual considerations. In this regard, Kelly *et al.*[18] have suggested a convincing rationale concerning the suitability of using nutrient ratios in nutritional research on the basis of the physiological and metabolic properties of the nutrients themselves. They have also pointed out that evaluating inter- and intra-macronutrient ratios may provide a measure of dietary macronutrient quality and how proteins, carbohydrates and fats affect health outcomes[19]. Improving our knowledge of the relationships between specific macronutrient intake ratios and outcomes may help establish benchmarks for better macronutrient quality and shape future guidelines concerning the prevention and treatment of diseases[19].

In brief, using log-ratio transformations of dietary data seems to be an appropriate approach because it is consistent with the compositional nature of the data themselves, and because the formulation of log-ratios as balances may capture the interdependent dynamics of dietary components. Furthermore, the compositional procedure provides a means of meeting the three important goals of (1) disentangling the relative and absolute aspects of the data; (2) obtaining meaningful estimate of the effects of total energy intake and (3) better controlling confounding factors.

Acknowledgements

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

The author declared that she has no conflict of interest.

References

1. Arnold KF, Berrie L, Tennant PWG, *et al.* (2020) A causal inference perspective on the analysis of compositional data. *Int J Epidemiol* **49**, 1307–1313.
2. Aitchison J (2005) *A Concise Guide to Compositional Data Analysis*. http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf (accessed April 2021).
3. Pawlowsky-Glahn V, Egozcue JJ & Tolosana-Delgado R (2015) *Modeling and Analysis of Compositional Data*, 1st ed. Chichester: John Wiley & Sons.
4. Aitchison J (1986) *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. London: Chapman & Hall Ltd.
5. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, *et al.* (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* **35**, 279–300.
6. Egozcue JJ & Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* **37**, 795–828.
7. Hron K, Filzmoser P & Thompson K (2012) Linear regression with compositional explanatory variables. *J Appl Stat* **39**, 1115–1128.
8. Filzmoser P, Hron K & Templ M (2018) *Applied Compositional Data Analysis*, 1st ed. [P Diggle and ZS Gather Ursula, editors]. Cham, Switzerland: Springer Series in Statistics.
9. Coenders G & Pawlowsky-Glahn V (2020) On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT* **44**, 201–220.
10. Correa Leite ML (2016) Applying compositional data methodology to nutritional epidemiology. *Stat Methods Med Res* **25**, 3057–3065.
11. Correa Leite ML (2019) Compositional data analysis as an alternative paradigm for nutritional studies. *Clin Nutr ESPEN* **33**, 207–212.
12. Correa Leite ML (2020) Orthonormal balances as a means of characterizing dietary exposure. *Nutr Res* **81**, 90–96.
13. Muller I, Hron K, Fiserova E, *et al.* (2018) Interpretation of compositional regression with application to time budget analysis. *Austrian J Stat* **47**, 3–19.
14. Pawlowsky-Glahn V, Egozcue JJ & Lovell D (2015) Tools for compositional data with a total. *Stat Model* **15**, 175–190.
15. Coenders G, Martin-Fernandez JA & Ferrer-Rosell B (2017) When relative and absolute information matter: compositional predictor with a total in generalized linear models. *Stat Model* **17**, 494–512.
16. Correa Leite ML & Nicolosi A (2009) Dietary patterns and metabolic syndrome factors in a non-diabetic Italian population. *Public Health Nutr* **12**, 1494–1503.
17. Salvini S, Parpinel M, Gnagnarella P, *et al.* (1998) *Banca Dati di Composizione Degli Alimenti per Studi Epidemiologici in Italia*, 1st ed. Milan: Istituto Europeo di Oncologia.
18. Kelly OJ, Gilman JC & Ilich JZ (2018) Utilizing dietary micronutrient ratios in nutritional research may be more informative than focusing on single nutrients. *Nutrients* **10**. doi:10.3390/nu10010107.
19. Kelly OJ, Gilman JC & Ilich JZ (2019) Utilizing dietary nutrient ratios in nutritional research: expanding the concept of nutrient ratios to macronutrients. *Nutrients* **11**, doi:10.3390/nu11020282.