British Journal of Nutrition (2022), 128, 2249–2257 doi:10.1017/S0007114522000253

Different statistical methods identify similar population-specific dietary patterns: an analysis of Longitudinal Study of Adult Health (ELSA-Brasil)

Mariane de Almeida Alves[1]*, Maria del Carmen Bisi Molina[2,3], Maria de Jesus Mendes da Fonseca[4], Paulo Andrade Lotufo[5], Isabela Martins Benseñor[5] and Dirce Maria Lobo Marchioni[1]

[1]*Department of Nutrition, School of Public Health, University of São Paulo, São Paulo 01246-904, Brazil*
[2]*Federal University of Ouro Preto, Minas Gerais 35400-000, Brazil*
[3]*Federal University of Espírito Santo, Espírito Santo 29040-090, Brazil*
[4]*National School of Public Health, Oswaldo Cruz Foundation, Rio de Janeiro 21041-210, Brazil*
[5]*Center for Clinical and Epidemiological Research, University Hospital, University of São Paulo, São Paulo 05508-000, Brazil*

(Submitted 15 September 2021 – Final revision received 13 January 2022 – Accepted 19 January 2022 – First published online 28 January 2022)

Abstract

In recent decades, different data-driven approaches have emerged to identify dietary patterns (DP) and little is discussed about how these methods are able to capture diet complexity within the same population. This study aimed to apply three statistical methods to identify the DP of the Longitudinal Study of Adult Health (ELSA-Brasil) population and evaluate the similarities and differences between them. Dietary data were assessed at baseline in the ELSA-Brasil study using a FFQ. DP were identified by applying three statistical methods: (1) factor analysis (FA), (2) treelet transform (TT) and (3) reduced rank regression (RRR). The characteristics of individuals classified in the last tertile of each DP were compared. Cross-classification and Pearson's correlation coefficients were assessed to evaluate the agreement between individuals' adherence to DP of the three methods. A similar convenience DP was identified for all three methods. FA and TT also identified a similar prudent DP and a DP highly loaded for the food groups rice and beans. Individuals classified in the third tertile of similar DP of each method presented similar socio-demographic and nutrient intake characteristics. Regarding the cross-classification, prudent DP from FA and TT presented a higher level of agreement (75 %), while convenience DP from TT and RRR presented the lowest agreement (44·8 %). The different statistical methods were able to capture the populations' DP in a similar way while highlighting the particularities of each method.

Key words: **Nutritional epidemiology: Dietary patterns: Factor analysis: Reduced rank regression: Treelet transform**

The study of the relationship between diet and health outcomes is a central issue in nutritional epidemiology research. Traditionally, these studies have focused on specific foods and nutrients and, although they have brought important contributions, these studies present the limitation of not considering diet complexity[1]. In real life, people eat meals with a variety of foods and nutrients that may be interactive or synergistic[2,3]. In this context, dietary patterns (DP) analysis has emerged as a complementary method[2], where food consumption is characterised in a holistic way and may better inform the comprehensive effect of diet on health outcomes[1,4].

DP studies have been conducted using three different approaches: *a priori* methods, such as indexes and scores, which use prior scientific knowledge on diet–disease associations, *a posteriori* methods that are entirely based on dietary data within a certain study population and hybrid methods, which combine both data on food intake in a population and pre-existing knowledge on diet–disease relationships[5,6]. The last two approaches are considered data-driven methods, because they entirely depend on the data at hand and identify population-specific DP.

Regarding data-driven methods, a range of statistical analyses can be applied to identify DP in a population. Factor analysis (FA) is the most widely applied technique[4], which evaluates the correlation matrix of food consumption data and identifies the latent factors that most explain the original data variance[7]. FA requires some subjective decisions throughout the analytical process, and interpretation of the final factors can be complicated because they are a linear combination of all original dietary data[8]. To address this limitation, a new statistical method, called treelet transform (TT), was proposed by Gorst-Rasmussen[9].

Abbreviations: DP, dietary pattern; FA, factor analysis; RRR, reduced rank regression; TT, treelet transform; FFQ, food frequency questionnaire.

* **Corresponding author:** Mariane de Almeida Alves, email marianealves@usp.br

TT has been recently applied in nutritional epidemiology to identify DP as a method that combines the strengths of factor and hierarchical cluster analyses[9,10]. Similar to FA, TT is also based on the correlation matrix of food items; however, the constructs identified are composed of a small number of food items, adding sparsity, and substantive meaning and interpretation to the DP[9,10]. To date, very few studies have applied this method to study DP[3,8,9,11,12].

The main hybrid approach applied in nutritional epidemiology is reduced rank regression (RRR). This analysis aims to directly relate the data-driven steps of pattern identification to an outcome of interest by identifying linear functions of food groups that can explain as much variation as possible of a set of outcome related variables (intermediate variables). These intermediate variables must be related to an outcome of interest, and it may be nutrients or biomarkers, both commonly applied in these studies[13,14].

As is well known, data-driven approaches differ substantially depending on the country, culture or ethnicity of different study populations[15–17]. However, little is known about how these different methods behave when applied to the same population. Considering that the data-driven approach is applied to capture diet complexity, can different methods describe the population's diet complexity similarly? Thus, the aim of this study was to compare these three different statistical methods to identify DP in the Longitudinal Study of Adult Health (ELSA-Brasil) population and evaluate their similarities and differences in describing population-specific DP.

Methods

Study participants

ELSA-Brasil is an ongoing cohort study that recruited 15 105 active and retired civil servants, aged 35–74 years, from five universities and one research institute located in three Brazilian macro-regions (southeast, northeast and south). Baseline examinations were performed in 2008–2010. Detailed information regarding the study sample and design has been described previously[18]. This study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the all institutional review boards involved (Fundação Oswaldo Cruz, Universidade Federal da Bahia, Universidade Federal do Espírito Santo, Universidade Federal de Minas Gerais, Universidade Federal do Rio Grande do Sul and Universidade de São Paulo). Written informed consent was obtained from all subjects.

For this study, only participants who had complete dietary data and complete anthropometric measurements at baseline and follow-up were considered. Individuals with misreported energy intake were excluded. Misreporting of energy intake was defined following the procedures proposed by McCroy *et al.*[19], where the agreement between the reported energy intake and predicted total energy expenditure was evaluated for each individual considering age, weight, height and sex. To calculate predicted total energy expenditure, the equation proposed by Vinken *et al.*[20] was applied. This validated equation was developed based on data from free-living individuals aged 18–81 years evaluated by the double-labelled water method. The cut-off point was defined as the ±2 SD of the agreement between the reported energy intake and predicted total energy expenditure based on the proposal of Black *et al.*[21]. This cut-off point takes into account the within-person coefficient of variation (CV) of energy intake (23 %), the CV of the technical measurement error of the double-labelled water plus the biological variation in total energy expenditure (8·2 %) and the CV of the prediction error of the total energy expenditure by Vinken's equation (17·7 %). In this sample, the ±2 SD was ±60 %; therefore, individuals were classified as under-reporters if they presented an energy intake agreement of less than 40% (*n* 137; 52·6% BMI < 30 kg/m^2; 50·4 % male) and over-reporters if they presented an energy intake agreement greater than 160 % (*n* 934; 88·65 % BMI < 30 kg/m^2; 43·0 % male). The final sample comprised 12 816 participants (Fig. 1).

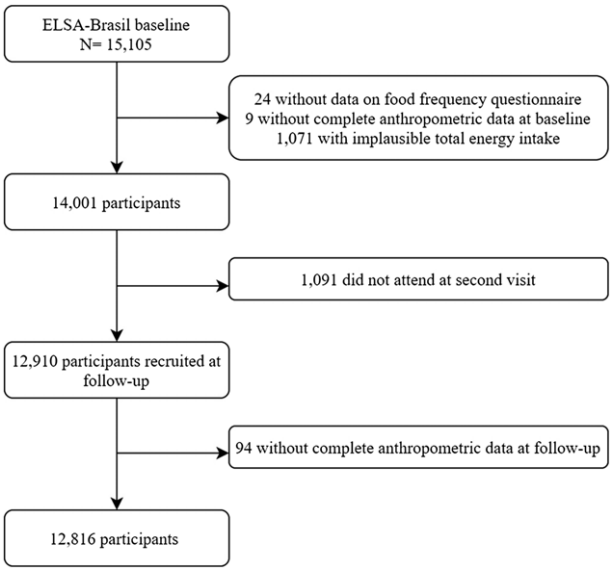

Fig. 1. Flow chart of final sample size in analysis. ELSA-Brasil, 2021.

Dietary data

Dietary data were assessed at baseline using a validated food frequency questionnaire (FFQ) composed of 114 food items. The FFQ was validated in relation to the nutrient intake of three food records. The intraclass correlation coefficient ranged from 0·22 to 0·72 for Se and Ca, respectively[22]. The FFQ was applied by trained interviewers to evaluate the participant's diet in the last 12 months regarding three sections: (1) food products/food preparations; (2) measures of consumed products and (3) consumption frequencies with eight response options including 'more than 3 times a day', '2 to 3 times a day', 'once a day', '5 to 6 times a week', '2 to 4 times a week', 'once a week', '1 to 3 times a month' and 'never/rarely'[22,23].

The food items in the FFQ were classified into twenty-five food groups: rice, cereals, bread, fruits, vegetables, beans, milk, nuts, sweets and desserts, tubers, pasta, snacks, eggs, cheese, butter/margarine, red meat, poultry, processed meat, fish, soft drinks, juice, coffee, beer, wine and distilled beverages (online

Supplementary Table S1). The food group classification was based on the similarity of nutrient profiles and eating occasion.

Dietary patterns

DP at baseline were identified using three statistical methods: FA, TT and RRR. All of these three dimension-reduction methods are comparable in identifying a population's DP because they work by aggregating food groups based on a correlation matrix; however, they present different aims, assumptions and decisions to take into account. The consumption in grams of the twenty-five food groups was the input variables for all three methods.

Factor analysis

FA is a widely applied method in nutritional epidemiology studies for identifying DP. This dimension-reduction method aims to explain as much variance as possible of the original data through latent variables (factors) that reveal the intrinsic structure of the data[1]. Principal component FA was applied in this study. Kaiser–Meyer–Olkin and Bartlett's tests were performed to verify whether our data were suitable for FA. The identified factors were orthogonally rotated using the *varimax* procedure to achieve a simpler structure with greater interpretability. The criteria used to retain the factors were eigenvalues > 1, identification of a breakpoint in the scree plot and interpretability. Food groups with factor loadings ≥ 0·30 or ≤ −0·30 were considered relevant and characteristic of the DP, and factors were labelled according to these food groups.

Treelet transform

A TT analysis was conducted as proposed by Gorst-Rasmussen[9,10]. TT combines the statistical principles of cluster and principal component analysis, leading to sparsity in factor loadings by creating components with food groups that present factor loadings exactly zero. This procedure results in fewer food items included in each factor when compared with FA, thereby facilitating the researcher's interpretation of each DP. The treelets are constructed as follows: between all of the original variables, the algorithm identifies the two variables with the largest correlation and then performs a principal component analysis on them. These two variables are replaced by a sum factor, and this procedure is repeated until all original variables have joined the hierarchical cluster tree. To select the treelets that provide greater variance, 10-fold cross-validation was applied to find the optimal cut level of the cluster tree, as proposed by Lee *et al.*[24]. To assess the sensitivity of the selected cut level, TT analysis was repeated at ± 3 levels of the optimal level. To verify the stability of the identified treelets, the *ttstab* procedure was performed through 100 bootstrap replications in sub-samples of 80 % of the original data[10].

Reduced rank regression

RRR analysis combines a data-driven approach with prior knowledge related to an outcome. In this model, the food groups are entered as predictor variables. A crucial step of this method is the selection of intermediate variables that are known to be related to the outcome of interest[13,14]. Unlike FA, RRR aims to explain as much variance as possible of these chosen intermediate variables. In this study, we built *a priori* knowledge based on obesity as an outcome of interest, selecting as intermediate variables the following nutrients or derived from nutrients variables: energy density (kcal/100 g), fibre density (g/1000 kcal) and total fat (g/d) adjusted for total energy intake using the residual method[25]. These intermediate variables were chosen based on the WHO report, which brings the evidence that energy density, fibre density and fat intake are linked with obesity risk[26] and previous published studies that have used these intermediate variables when assessing DP related to obesity[27–29]. In RRR analyses, food groups with factor loadings ≥ 0·20 or ≤ −0·20 were considered relevant and characteristic of the DP. In RRR analysis, the number of intermediate variables is a condition for the number of derived DP, and the interpretability was the criterion applied to retain the DP.

A DP score was calculated for all DP identified for each method. This variable indicates the individual-level weights associated with each DP and theoretically represents the adherence to a DP.

Statistical analyses

To enable the comparison between individuals' characteristics according to the level of adherence to each DP of the three methods, the individual's DP scores were classified into tertiles. Sociodemographic characteristics and nutritional profile of the individuals classified in the third tertile were described as means or proportions. To evaluate the agreement of individuals' classification on tertiles in similar DP of different methods, we performed a cross-classification analysis, and Pearson's correlation coefficients between DP scores were also obtained.

Statistical analyses were performed using Stata® (Statistical Software for Professionals) version 14.2[30], and only the RRR analyses were performed using SAS® Studio version 3.8 (SAS Institute Inc.).

Results

Three DP were identified in the FA (Table 1). The first DP, labelled as convenience, was highly positively loaded with the food groups sweets and desserts, pasta, snacks, eggs, cheese, butter/margarine, red meat, processed meat and soft drinks. The second DP, labelled as Brazilian traditional, was highly positively loaded with rice, beans and poultry and was highly negatively loaded with cereals, nuts, cheese and wine. The third DP, labelled as prudent, was highly positively loaded with cereals, fruits, vegetables, tubers, fish and juice. All three FA-derived DP explained 26·9 % of the variance in the original dietary data. The results for the Kaiser–Meyer–Olkin and Bartlett's tests were 0·73 and $P < 0·001$, respectively, indicating that the sample was suitable for FA.

The 10-fold cross-validation performed in the TT analysis indicated that 19 was the optimal cut level of cluster three. Repeated analysis at ± 3 levels resulted in DP with similar characteristics. The same was observed using the *ttstab* procedure, in which the DP obtained for sub-samples were similar to the total sample, indicating good stability in the TT analysis. The first

Table 1. Food groups' factor loadings, eigenvalues and explained variance among dietary patterns (retaining three factors)

Food Group	Convenience	Brazilian Traditional	Prudent
Rice	0·08	0·70	−0·02
Cereals	−0·11	−0·32	0·36
Bread	0·28	0·04	0·03
Fruits	−0·16	−0·03	0·60
Vegetables	−0·05	−0·01	0·69
Beans	0·09	0·66	0·06
Milk	−0·04	−0·11	0·16
Nuts	0·12	−0·41	0·29
Sweets and desserts	0·53	−0·12	−0·02
Tubers	0·20	0·25	0·42
Pasta	0·46	0·09	0·14
Snacks	0·69	0·04	−0·06
Eggs	0·39	0·24	0·16
Cheese	0·33	−0·40	0·20
Butter/margarine	0·32	0·18	−0·07
Red meat	0·31	0·29	0·03
Poultry	0·15	0·44	0·27
Processed meat	0·55	0·22	−0·03
Fish	0·10	0·09	0·50
Soft drinks	0·45	0·08	−0·23
Juice	0·08	−0·12	0·36
Coffee	0·15	−0·04	0·03
Beer	0·23	0·23	0·03
Wine	0·25	−0·33	0·18
Distilled	0·18	0·06	0·05
Eigenvalues	2·68	1·95	1·56
% explained variance	9·66	8·99	8·22
% cumulative explained variance	9·66	18·65	26·87

Table 2. Food groups' loadings, eigenvalues and explained variance among dietary patterns derived by Treelet transform

Food groups	Convenience	Prudent	Rice and beans
Rice	0·28		0·64
Cereals		0·34	
Bread	0·21		
Fruits		0·49	
Vegetables		0·49	
Beans	0·28		0·64
Milk			
Nuts		0·34	
Sweets and desserts	0·37		−0·20
Tubers	0·21		
Pasta	0·27		−0·14
Snacks	0·37		−0·20
Eggs	0·26		−0·14
Cheese		0·29	
Butter/margarine	0·21		
Red meat	0·23		−0·12
Poultry	0·24		
Processed meat	0·32		−0·17
Fish		0·36	
Soft drinks	0·30		−0·16
Juice		0·30	
Coffee			
Beer			
Wine			
Distilled			
Eigenvalue	2·51	1·77	1·27
% explained variance	10·10	7·03	5·07
% cumulative explained variance	10·10	17·10	22·15

TT-derived DP, labelled as convenience, was positively loaded with rice, bread, beans, sweets and desserts, tubers, pasta, snacks, eggs, butter/margarine, red meat, poultry, processed meat and soft drinks. The second DP, labelled as prudent, was positively loaded with cereals, fruits, vegetables, nuts, cheese, fish and juice. The third DP, labelled as rice and beans, was positively loaded with rice and beans and negatively loaded with sweets and desserts, pasta, snacks, eggs, red meat, processed meat and soft drinks. These three TT-derived DP explained 22·2 % of the variance in the original dietary data (Table 2).

RRR analysis derived three DP (online Supplementary Table S2) and according to interpretability, only the first DP was retained. The RRR-derived DP, labelled as convenience, was positively loaded with sweets and desserts, snacks, butter/margarine, red meat, processed meat and soft drinks and was negatively loaded with fruits, vegetables and beans. This DP explained 5·8 % of the variance in the original dietary data and 48·7 % of the variance in the intermediate variables (Table 3).

The characteristics of the total population and individuals classified in the third tertile of each DP are presented in Table 4. Most of the individuals with high adherence to the convenience DP of the three methods were younger, mostly men and smokers when compared with the total population. These individuals presented the highest mean for energy density and percentage of kilocalories from total fat and saturated fat, while presenting the lower mean of fibre density. Individuals classified in the third tertile of the prudent DP of the FA and TT methods were slightly older, most of them were females and non-smokers and presented a lower mean for energy density. Most of the individuals classified in the third tertile of the traditional Brazilian and rice and beans DP were men and non-white individuals when compared with the total population. These individuals had the lowest consumption of kilocalories from animal protein and saturated fat. High adherence to the traditional Brazilian DP was characterised by the largest percentage of smokers, and individuals with the highest adherence to rice and beans DP had the largest mean for fibre density and percentage of kilocalories from carbohydrates. The food groups' mean consumption of the individuals classified on the third tertile of each DP is presented in online Supplementary Table S3.

The agreement of individuals' classification in a similar DP of the three different methods is presented in Table 5. The prudent DP from FA and TT presented the highest level of agreement and the opposite was observed between the convenience DP from TT and RRR, with the lowest level of agreement. Pearson's correlation coefficients for convenience DP scores from FA and TT, FA and RRR, and TT and RRR were r 0·83; $P < 0·001$, r 0·63; $P < 0·001$ and r 0·39; $P < 0·001$, respectively. The correlation between the prudent DP scores from FA and TT was r 0·90; $P < 0·001$, and the traditional Brazilian and rice beans DP scores had a correlation coefficient of r 0·66, $P < 0·001$.

Table 3. Food groups' factor loadings, eigenvalues and explained variance among dietary patterns through reduced rank regression

Food group	Convenience
Rice	0·01
Cereals	−0·13
Bread	0·05
Fruits	−0·47
Vegetables	−0·30
Beans	−0·24
Milk	−0·10
Nuts	0·12
Sweets and desserts	0·26
Tubers	−0·05
Pasta	0·11
Snacks	0·34
Eggs	0·15
Cheese	0·18
Butter/margarine	0·24
Red meat	0·30
Poultry	0·08
Processed meat	0·27
Fish	0·09
Soft drinks	0·20
Juice	0·03
Coffee	−0·15
Beer	0·14
Wine	0·08
Distilled	0·08
% explained variance of predictors variables	5·84
% ED explained variance	42·86
% FD explained variance	60·14
% Total fat explained variance	43·22
% explained variance of response variables	48·74

ED, energy density; FD, fibre density.

Discussion

None of the three different statistical methods identified an identical DP; however, a similar convenience DP was observed in the three methods as the first one, for example, the one that most explained the variance in the original data. The convenience DP shared the food groups sweets and desserts, snacks, butter and margarine, red meat, processed meat and soft drinks. The FA and TT identified a prudent DP with cereals, fruits, vegetables, fish and fruit juice as common food groups, and a traditional Brazilian (FA) and rice and beans (TT) DP highly loaded for the food groups' rice and beans. Even though there were differences in food groups and factor loadings within the identified DP, reflecting the particularities of each method, we could see that, independent of the method applied, these analyses were able to capture the population's diet in a similar way.

There are some peculiarities worthy of highlighting. The RRR-convenience DP was not only related to the higher consumption of unhealthy food groups but was also related to the lower consumption of the food groups fruits, vegetables and beans – representing the opposite of prudent and traditional Brazilian/rice and beans DP identified by the other two methods. Similarly, the TT-convenience DP presented low but positive loadings for the food groups rice and beans, indicating the influence of traditional Brazilian foods on this DP. These particularities may explain the slight differences in the fibre density and the percentage of kilocalories from carbohydrate, protein, total fat and saturated fat observed in the individuals classified in the third tertile of the convenience DP from TT and RRR analyses.

Other studies have applied DP analysis to the ELSA-Brasil dataset or subsets with diverse aims. Bezerra *et al.*[31] applied latent class analysis and identified DP labelled as prudent and processed, which shared similar characteristics with the prudent and convenience DP identified in our study, respectively. Gorgulho *et al.*[32] applied FA in a subset of ELSA-Brasil (only participants from São Paulo) and also found a convenience DP, as the one that most explained the original data variance, which shared the same food groups in the convenience DP of our study. They also found DP labelled plant-based and dairy products that were highly loaded for similar food groups (fruits, vegetables and cereals) of our prudent DP identified by factor and TT analyses. The main difference was that we observed that the food group fish was highly loaded in our prudent DP, which could reflect some regional characteristics of the total Brazilian population.

A Brazilian traditional DP sharing the characteristics with the Brazilian traditional (FA) and rice and beans (TT) DP was also observed in four studies conducted with the ELSA-Brasil population[32–35] that included a range of four different statistical analyses (principal component analysis, FA, cluster analysis and multiple correspondence analysis). A DP highly loaded with the food groups rice and beans is commonly identified in studies regarding all ages and sex of the Brazilian population[36–38].

As expected, the RRR analysis explained a smaller proportion of original dietary data variance (5·8 %) than FA and TT, since RRR focuses on identifying DP that most explain the variation of the intermediate variables. Despite this methodological difference, RRR analysis was able to identify not only a DP related to energy density, fibre density and total fat but also a DP that is in fact present in the population, as a similar convenience DP was identified through FA and TT. This result was also observed by Batis *et al.* and Cunha *et al.* when comparing RRR with principal component analysis and FA, respectively[39,40]. The food group differences observed between the RRR-convenience DP and convenience DP of other methods may be relevant and bring new insights to understand the associations between this DP and health outcomes in further studies.

FA explained a higher proportion of original dietary data variance when compared with TT (26·9 and 22·2 %, respectively). TT only loads the most expressive food groups for a DP, and those that are not relevant receive a loading equal to zero, leading to a lower number of food groups contributing to a DP. This sparsity created by TT is considered as an advantage over FA, which produces a complex factor loading matrix making the interpretation of DP more susceptible to researcher assumptions. Also, because of the sparsity of TT analysis, a trade-off between the explained variance of the DP and the interpretability is inevitable[3]. Schoenaker *et al.*[8] introduced a relevant issue for TT analysis: whether it is able to capture the overall diet, as only specific food items are taken into account to predict the DP score at the individual level, some of the synergic aspects of food may be lost in this process and need to be considered when applying this approach. It is important to mention that in both methods, the overall proportion of the variance explained the DP is not large,

Table 4. Baseline characteristic and nutritional profile of total population and of individuals classified on the third tertile of each dietary pattern (mean values and standard deviations; numbers and percentages)

			Convenience						Prudent				Rice and beans/traditional			
	Total population		FA		TT		RRR		FA		TT		FA		TT	
	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %	Mean or *n*	SD or %
Age (years)	51·6	8·9	49·9	8·6	49·4	8·3	49·8	8·7	53·1	8·9	53·2	9·0	50·4	8·3	51·3	8·6
Female	6997	54·4	1532	35·7	1345	31,5	1899	44·5	2181	51,1	2346	54·9	1395	32·6	1725	40·4
BMI (kg/m^2)	27·1	4·7	27·6	4·6	27·7	4·8	27·3	4·7	27·3	4·7	27·3	4·7	27·5	4·6	27·0	4·7
Smoker*	1559	12·2	618	14·7	631	14·8	644	15·1	399	9·3	373	8·7	702	16·4	574	13·4
Educational level																
Until 8 years	1428	11,1	396	9·3	589	13·8	294	6·9	479	11·2	378	8·8	912	21·3	786	18·4
9 years or more	11 388	88,9	3876	90·7	3683	86·2	3978	93·1	3793	88·8	3894	91·2	3360	78·7	3486	81·6
Self-reported skin colour*																
White	6815	53·8	2502	59·3	2141	50·6	2505	59·4	2107	50,1	2283	54·3	1663	39·3	1821	43·1
Non-white	5858	46·2	1720	40·7	2091	49·4	1709	40,6	2100	49,9	1921	45·7	2572	60,7	2403	56·9
Energy density (kcal/g)	1·4	0·3	1·6	0·3	1·6	0·3	1·7	0·3	1·3	0·3	1·3	0·3	1·5	0·3	1·4	0·3
Fibre density (g/1000 kcal)	16·0	5·0	14·0	4·6	15·2	4·6	11·9	2·8	17·3	4·8	17·2	4·8	17·1	5·0	18·5	4·9
%kcal from carbohydrate	56·6	7·3	53·5	6·8	55·3	6·5	51·1	6·4	57·7	7·4	57·4	7·6	57·0	6·5	58·6	6·2
%kcal from protein	16·8	2·7	16·4	2·6	16·5	2·5	17·1	3·0	17·0	2·6	17·0	2·7	16·8	2·5	16·6	2·4
%kcal from animal protein	9·1	3·2	9·3	3·1	8·8	3·0	10·4	3·4	9·3	3·1	9·3	3·1	8·4	3·1	7·7	2·7
%kcal from vegetable protein	7·6	1·7	7·1	1·5	7·6	1·7	6·6	1·3	7·6	1·6	7·6	1·6	8·3	1·8	8·9	1·7
%kcal from fat	26·4	5·0	28·7	4·7	27·5	4·5	30·1	4·6	25·6	5·0	26·2	5·3	25·2	4·3	24·5	4·2
%kcal from saturated fat	8·6	2·4	9·6	2·3	8·9	2·3	10·2	2·3	8·1	2·2	8·4	2·4	7·6	1·9	7·4	1·9

FA, factor analysis; TT, treelet transform; RRR, reduced rank regression.
* Variable with missing data.

Table 5. Cross-classification of individuals' adherence to dietary patterns, according to each statistical method

FA and TT-dietary patterns	Total agreement	Opposite tertiles
Convenience	66·1	2·0
Brazilian traditional/rice and beans	58·5	5·3
Prudent	75·9	0·5
FA and RRR-convenience dietary pattern	56·3	6·0
TT and RRR-convenience dietary pattern	44·8	12·3

FA, factor analysis; TT, treelet transform; RRR, reduced rank regression.

which means that only a limited portion of the diet variance is considered when investigating DP[41].

Several studies have compared DP identified by RRR with other statistical methods and how they are associated with different health outcomes[42–47]. The findings presented in these studies did not allow us to affirm whether some of these methods are superior in estimating the association with a specific outcome. While some studies suggested that RRR analysis provided better results when investigating the association between DP and the metabolic syndrome in adults[43], obesity in preschool children[45] and bone mass in an elderly population[46], other studies found similar and consistent results independent of the method applied[39,44,47]. Considering the particularities of the RRR analysis, this may be a promising method when the study goal is exploring the combination of foods that are mediated by specific variables (intermediate variables), also adding the possibility of using metabolome and/or microbiome information[3].

The same controversial results were found when comparing DP from FA and TT analysis and their association with health outcomes. Schoenaker *et al.* compared DP derived both from TT and FA and, even though they have found similar DP, only those identified by FA were associated with incident diabetes in a middle-aged women's population[8]. Whereas Gorst-Rasmussen *et al.* obtained similar results in estimating the relative risk of myocardial infraction with DP identified by TT or FA in a Danish cohort study[9]. Since there is not robust evidence of a superior DP method to predict the relationship between diet and health outcomes, the researchers need to keep in mind the research question of their study and then select which method is more appropriate. Also, comparisons between different methods in the same study can bring new insights and complementary results to better understand each statistical method[48].

Our study had some limitations. First, all three statistical techniques applied to identify the DP require arbitrary decisions and subjective interpretations. In these data-driven approaches, the researcher defines food grouping and the label of each DP. Specifically in FA and RRR, the researcher defines the number of factors to retain and the cut-off points that define which food groups are relevant to the DP. Second, the dietary consumption data were assessed using a FFQ, a self-reported method that has some inherent bias, such as memory or social desirability. Also, in the FFQ the food items are pre-grouped, which made it not possible to have more distinctive food groups (e.g. unhealthy *v.* healthy foods groups) and it may have an impact on DP meaningfulness. The strengths of this study are the use of a validated FFQ, the large sample size and the application of the TT analysis, a novel method in nutritional epidemiology to identify populations' DP.

In conclusion, our results showed that three different statistical methods were able to capture the populations' DP in a similar way while highlighting the importance of the particularities of each method. The different aims and procedures of each method may play a relevant role in identifying associations between DP and health outcomes, and comparing these results can bring new perspectives to understand this relationship.

Acknowledgements

The ELSA-Brasil study was supported by the Brazilian Ministry of Health, the Brazilian Ministry of Science and Technology and the Brazilian National Council for Scientific and Technological Development-CNPq. The research centre of São Paulo was also supported by the São Paulo Research Foundation (FAPESP) (grant number 2011/12256-4). The Graduate Program of Public Health Nutrition is supported by the Coordination of Superior Level Staff Improvement (CAPES). M. A. A. received a scholarship from the São Paulo Research Foundation (FAPESP) (grant number 2019/13486-5). The funding agencies that supported the study had no role in the design, analysis or writing of this article.

M. A. A. and D. M. L. M. were responsible for the study concept. D. M. L. M. was also responsible for the supervision of all stages of this study. M. A. A. conducted the data analysis, interpreted the results, wrote the manuscript and had primary responsibility for final content. M. C. B. M., M. J. M. F., P. A. L. and I. M. B. contributed to critical review of the manuscript. All authors have read and approved the final manuscript.

The authors declare no conflict of interest.

Supplementary material

For supplementary materials referred to in this article, please visit https://doi.org/10.1017/S0007114522000253

References

1. Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**, 3–9.
2. Mozaffarian D, Rosenberg I & Uauy R (2018) History of modern nutrition science-implications for current research, dietary guidelines, and food policy. *BMJ* **361**, k2392.
3. Schulz CA, Oluwagbemigun K & Nothlings U (2021) Advances in dietary pattern analysis in nutritional epidemiology. *Eur J Nutr* **60**, 4115–4130.
4. Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, 177–203.
5. Jannasch F, Riordan F, Andersen LF, *et al.* (2018) Exploratory dietary patterns: a systematic review of methods applied in pan-European studies and of validation studies. *Br J Nutr* **120**, 601–611.

6. Krebs-Smith SM, Subar AF & Reedy J (2015) Examining dietary patterns in relation to chronic disease: matching measures and methods to questions of interest. *Circulation* **132**, 790–793.
7. Reedy J, Wirfalt E, Flood A, *et al.* (2010) Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: the NIH-AARP Diet and Health Study. *Am J Epidemiol* **171**, 479–487.
8. Schoenaker DAJM, Dobson AJ, Soedamah-Muthu SS, *et al.* (2013) Factor analysis is more appropriate to identify overall dietary patterns associated with diabetes when compared with Treelet transform analysis. *J Nutr* **143**, 392–398.
9. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, *et al.* (2011) Exploring dietary patterns by using the Treelet transform. *Am J Epidemiol* **173**, 1097–1104.
10. Gorst-Rasmussen A (2012) tt: Treelet transform with Stata. *Stata J* **12**, 130–146.
11. Frederiksen SB, Themsen HH, Overvad K, *et al.* (2021) Dietary patterns generated by the Treelet Transform and risk of stroke: a Danish cohort study. *Public Health Nutr* **24**, 84–94.
12. Oluwagbemigun K, Foerster J, Watkins C, *et al.* (2020) Dietary patterns are associated with serum metabolite patterns and their association is influenced by gut bacteria among Older German Adults. *J Nutr* **150**, 149–158.
13. Weikert C & Schulze MB (2016) Evaluating dietary patterns: the role of reduced rank regression. *Curr Opin Clin Nutr Metab Care* **19**, 341–346.
14. Hoffmann K, Schulze MB, Schienkiewitz A, *et al.* (2004) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* **159**, 935–944.
15. Lin H, Bermudez OI & Tucker KL (2003) Dietary patterns of Hispanic elders are associated with acculturation and obesity. *J Nutr* **133**, 3651–3657.
16. Park SY, Murphy SP, Wilkens LR, *et al.* (2005) Dietary patterns using the food guide pyramid groups are associated with sociodemographic and lifestyle factors: the multiethnic cohort study. *J Nutr* **135**, 843–849.
17. Carioca AAF, Gorgulho B, Teixeira JA, *et al.* (2017) Dietary patterns in internal migrants in a continental country: a population-based study. *PLOS ONE* **12**, e0185882.
18. Aquino EM, Barreto SM, Bensenor IM, *et al.* (2012) Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design. *Am J Epidemiol* **175**, 315–324.
19. McCrory MA, McCrory MA, Hajduk CL, *et al.* (2002) Procedures for screening out inaccurate reports of dietary energy intake. *Public Health Nutr* **5**, 873–882.
20. Vinken AG, Bathalon GP, Sawaya AL, *et al.* (1999) Equations for predicting the energy requirements of healthy adults aged 18–81 years. *Am J Clin Nutr* **69**, 920–926.
21. Black AE (2000) Critical evaluation of energy intake using the Goldberg cut-off for energy intake : basal metabolic rate. A practical guide to its calculation, use and limitations. *Int J Obesity* **24**, 1119–1130.
22. Benseñor IM, Velasquez-Melendez G, Drehmer M, *et al.* (2013) Reproducibility and relative validity of the Food Frequency Questionnaire used in the ELSA-Brasil. *Cadernos Saúde Pública* **29**, 379–389.
23. Molina MDB, Bensenor IM, Cardoso LD, *et al.* (2013) Reproducibility and relative validity of the Food Frequency Questionnaire used in the ELSA-Brasil. *Cad Saude Publica* **29**, 380–390.
24. Lee AB, Nadler B & Wasserman L (2008) Treelets – an adaptive multi-scale basis for sparse unordered data. *Ann Appl Statistics* **2**, 435–471.
25. Willett WC, Howe GR & Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* **65**, 1220S–1228S.
26. World Health Organization (2003) *Diet, Nutrition and Prevention of Chronic Disease: Report of a Joint WHO/FAO Expert Consultation*. Geneva: WHO.
27. Livingstone KM, Sexton-Dhamu MJ, Pendergast FJ, *et al.* (2021) Energy-dense dietary patterns high in free sugars and saturated fat and associations with obesity in young adults. *Eur J Nutr* **60**, 4115–4130.
28. Huybrechts I, Lioret S, Mouratidou T, *et al.* (2017) Using reduced rank regression methods to identify dietary patterns associated with obesity: a cross-country study among European and Australian adolescents. *Br J Nutr* **117**, 295–305.
29. Livingstone KM & McNaughton SA (2017) Dietary patterns by reduced rank regression are associated with obesity and hypertension in Australian adults. *Br J Nutr* **117**, 248–259.
30. StataCorp (2015) *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.
31. Bezerra IN, Bahamonde NMSG, Marchioni DML, *et al.* (2018) Generational differences in dietary pattern among Brazilian adults born between 1934 and 1975: a latent class analysis. *Public Health Nutr* **21**, 2929–2940.
32. Gorgulho B, Alves MA, Teixeira JA, *et al.* (2021) Dietary patterns associated with subclinical atherosclerosis: a cross-sectional analysis of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) study. *Public Health Nutr* **24**, 5006–5014.
33. Cardoso Lde O, Carvalho MS, Cruz OG, *et al.* (2016) Eating patterns in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): an exploratory analysis. *Cad Saude Publica* **32**, e00066215.
34. Silva GB, Fraser SDS, Neri AKM, *et al.* (2020) Association between dietary patterns and renal function in a cross-sectional study using baseline data from the ELSA-Brasil cohort. *Braz J Med Biol Res* **53**, e10230.
35. Drehmer M, Odegaard AO, Schmidt MI, *et al.* (2017) Brazilian dietary patterns and the dietary approaches to stop hypertension (DASH) diet-relationship with metabolic syndrome and newly diagnosed diabetes in the ELSA-Brasil study. *Diabetol Metab Syndr* **9**, 13.
36. Cunha DB, de Almeida RM, Sichieri R, *et al.* (2010) Association of dietary patterns with BMI and waist circumference in a low-income neighbourhood in Brazil. *Br J Nutr* **104**, 908–913.
37. Cunha DB, Bezerra IN, Pereira RA, *et al.* (2018) At-home and away-from-home dietary patterns and BMI z-scores in Brazilian adolescents. *Appetite* **120**, 374–380.
38. Cezimbra VG, Assis MAA, de Oliveira MT, *et al.* (2021) Meal and snack patterns of 7–13-year-old schoolchildren in southern Brazil. *Public Health Nutr* **24**, 2542–2553.
39. Batis C, Mendez MA, Gordon-Larsen P, *et al.* (2015) Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults. *Public Health Nutr* **19**, 195–203.
40. Cunha DB, Almeida RM & Pereira RA (2010) A comparison of three statistical methods applied in the identification of eating patterns. *Cad Saude Publica* **26**, 2138–2148.
41. Martinez ME, Marshall JR & Sechrest L (1998) Invited commentary: factor analysis and the search for objectivity. *Am J Epidemiol* **148**, 17–19.
42. Batis C, Mendez MA, Gordon-Larsen P, *et al.* (2016) Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults. *Public Health Nutr* **19**, 195–203.
43. Kurniawan AL, Hsu C-Y, Lee H-A, *et al.* (2020) Comparing two methods for deriving dietary patterns associated with risk of metabolic syndrome among middle-aged and elderly Taiwanese adults with impaired kidney function. *BMC Med Res Methodol* **20**, 255–255.

44. Barbaresko J, Siegert S, Koch M, *et al.* (2014) Comparison of two exploratory dietary patterns in association with the metabolic syndrome in a Northern German population. *Br J Nutr* **112**, 1364–1372.
45. Manios Y, Kourlaba G, Grammatikaki E, *et al.* (2010) Comparison of two methods for identifying dietary patterns associated with obesity in preschool children: the GENESIS study. *Eur J Clin Nutr* **64**, 1407–1414.
46. Melaku YA, Gill TK, Taylor AW, *et al.* (2018) A comparison of principal component analysis, partial least-squares and reduced-rank regressions in the identification of dietary patterns associated with bone mass in ageing Australians. *Eur J Nutr* **57**, 1969–1983.
47. Shakya PR, Melaku YA, Page A, *et al.* (2020) Association between dietary patterns and adult depression symptoms based on principal component analysis, reduced-rank regression and partial least-squares. *Clin Nutr* **39**, 2811–2823.
48. Zhao J, Li Z, Gao Q, *et al.* (2021) A review of statistical methods for dietary pattern analysis. *Nutr J* **20**, 37.