Exploring Dietary Intake Data collected by FPQ using Unsupervised Learning

1st Martin Gjoreski
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
martin.gjoreski@ijs.si

2nd Stefan Kochev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
stefan.kochev@gmail.com

3rd Nina Reščič
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
nina.rescic@ijs.si

4th Matej Gregorič
National Institute of Public Health
Ljubljana, Slovenia
matej.gregoric@nijz.si

5th Tome Eftimov
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

6th Barbara Koroušić Seljak
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

***Abstract*—Populations in countries undergoing rapid transition are experiencing food- and nutrition-related problems. To acquire high-quality nutrition information, we need beside adequate data about food consumption, also efficient methods for the extraction of information from the collected data.**

Our aim was to develop a methodology for analyzing and reasoning about dietary intake data collected by a food propensity questionnaire (FPQ) and dependent 24-hour recalls (24HRs).

We analysed a subset of data (about 197 participants) in the SI.Menu survey carried out in 2016/17 in Slovenia. The participants completed FPQs and 24HRs.

We were able to identify four clusters. Two clusters represented participants with more healthy habits, e.g., low intake of animal fats, high breakfast frequency, and high intake of fruits and vegetables. The other two clusters represented participants with less healthy habits, e.g., high intake of animal fats, low breakfast frequency and increased BMI.

The four clusters can be well separated by only four variables. This interesting discovery could lead to simplified FFQ questionnaires, which could significantly decrease the participants' burden and could ensure participant compliance in similar studies. Having big national data set related to nutrition should ease the process of creating sustainable policies that will ultimately benefit agriculture, human health and the environment.

***Index Terms*—FPQ, 24HRs, clustering, healthy habits**

I. Introduction

Information about food consumption at the country level is required for many reasons. For example, shifts in dietary patterns may cause considerable health consequences [1]. However, to acquire high-quality information, we need beside adequate data about food consumption also efficient methods for the extraction of information from the collected data. Food consumption data can be either derived from household budget surveys or food balance sheets or collected using dietary surveys, such as open-ended surveys (e.g. 24-hour recalls - 24HRs, food records) and closed-ended surveys (e.g. food frequency and food propensity questionnaires - FFQs and FPQs).

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

As each of these methods has inherent strengths and limitations [2], there is a strong need for an advanced method for extracting information from linked data sets acquired using different collection methods. In this paper, we present such a method that considers dependent data from FPQs and multiple 24HRs covering two non-consecutive days. The main reason for this selection is that these two types of methods are recommended by the European Food Safety Agency (EFSA) for the creation of the EFSA Comprehensive European Food Consumption Database [3]. While multiple 24HRs enable capturing of participants' occasional food consumption in detail, the day-to-day variation in their food intake cannot be identified. To address this limitation, there is a recommendation to collect data missing from 24HRs using a FPQs, which aims to record the frequency of food consumption per a specified time period (often, ”in the last year”). When the question about the amounts consumed is omitted, the FPQs is called Food Propensity Questionnaire (FPQ). The main idea is to estimate the probability of food consumption by analysing data from FPQs and to combine this estimate with the amount consumed (from 24HRs) to describe the distribution of usual food intake. However, there may appear a problem of missing data. For instance, in one or more 24HRs of a selected participant there is no data for one or more specific foods, which have been reported in the FPQ. As food consumption is analysed at the country level, we can impute the missing information from other participants' data. Which participants could be considered in this process of data imputation, depends on their characteristics and dietary habits that need to be defined. Another problem is related to errors, which may be systematic or random. Such errors needs to be detected and controlled so that the information extracted from the collected data is correct. In this paper, we focus on the problem of discovering subgroups (clusters) of the dietary survey participants considering their characteristics and dietary habits. To solve the problem, we consider data sets collected by an FPQ and multiple 24HRs. Once this problem is resolved, information about food consumption at the country level can be extracted from the collected data even in case of missing or erroneous data for some participants.

In the continuation of this paper, we first describe how data to be used for our research was collected. Then, a method for clustering participants into groups according to their characteristics and dietary habits is introduced. We continue presenting a method for extracting information about food consumption at the country level by analysing dependent data collected from multiple 24HRs and FPQs. Next, results of the evaluation of both methods are provided. We finalise the paper with the discussion and conclusions, where directions for our future work are also indicated.

II. Related Work

Unsupervised learning is one of the main three categories of machine learning, along with supervised and reinforcement learning. Unsupervised learning is typically used for finding patterns in a data set without pre-existing labels. One of the main approaches for unsupervised learning is clustering [4]. Clustering is a process of discovering clusters (groups) of instances (data points). For example, such clusters can be discovered by iteratively selecting clusters that minimise intra-cluster distances and maximise inter-cluster distances. In other words, the instances in the same cluster should be more similar to each other than to those in other clusters. There are different types of clustering algorithms, e.g., connectivity models, centroid models, distribution models, etc. [5]. Usually, the most appropriate clustering algorithm is chosen experimentally, depending on the data set' characteristics.

Clustering data sets collected using questionnaires have been proposed in previous studies [6]. For example, clustering has been applied to identify patterns of diet, physical activities and sedentary behavior among children or adolescents and their associations with socio-demographic indicators, and overweight and obesity [7], [8], [9]. Although 24HRs provide quantitative and detailed dietary data and perform as well as more complicated methods, such as weighted records [10], [11], it doesn't give information on usual food intake. FPQs provide exactly the missing information (frequency of consumption of specific food) but doesn't provide the information about the consumed amounts. Combining both approaches seems like an obvious step and has been done before [12] and strong and consistent relationships between reported FFQ frequency of food and food-group consumption and the probability of consumption on 24-hour recalls was found [13]. We perform clustering analysis and we interpret the results which leads to some new insights.

III. Data collection

A. FPQ

The questionnaire consists of several categories. Participants were answering questions regarding their dietary habits, such as frequency of meals (breakfast, lunch, dinner, snacks etc.), dietary habits (current appetite, adding salt at the table, eating out vs. preparing food at home) and the frequency of consumption food from specific food group (FFQ). FPQ consists of 78 questions split into 9 categories - cereals and cereal products; milk, milk products and supplements; fruit; vegetables and potato; meat, meat products, fish and supplements; fats; drinks; other.

B. 24HRs

Participants reported about their dietary habits completing 24HRs for two non-consecutive days by the help of trained interviewers. Data was collected during working and weekend days for participants selected from different Slovenian geographical regions in all four seasons using statistically correct data collection method. Once collected, 24HRs were validated and analysed so that food consumption data was matched with food composition data to provide nutrient intake of the selected participants. Nutrient values were calculated for energy, protein, fat, carbs, alcohol and water. Also food groups, such as grains, fruits etc., and their subgroups, such as white bread etc., were identified.

There was also a small subset (approx. 1%) of participants for whom only one 24HR was collected.

IV. Data preprocessing

A. Data normalization

The normalization of the data was done with regard to the preferred "healthy answers" for each FPQ question. The "healthy answers" were provided by a nutrition specialist.

For example, let us assume the question:

Q1: How often do you eat individual meals during the week (Monday through Friday)?

- 1 - Every day;
- 2 - 4 times per week;
- 3 - 3 times per week;
- 4 - 2 times per week;
- 5 - Once per week;
- 6 - Never.

For it, the preferred "healthy answer" provided by the nutrition specialist is 1 (every day). Let us assume, that v_i is the answer provided by the user for the i-th FPQ question. Then, the normalized value is calculated as:

$$v_{normalized} = \mid v_{preferred} - v_{user} \mid \quad (1)$$

Performing this transformation of the data, the normalized value represents how much the participant's answer deviates from the preferred "healthy answer".

B. Meta-variables calculation

After the data normalization process, the next step is to define the meta-variables (i.e. meta-features) that will be used in the analysis. The FPQ questions were grouped into nine subgroups by the nutrition specialist, and for each one he/she provided an "healthy answer". To calculate the meta-variables, for each subgroup, based on the participant's answer, the

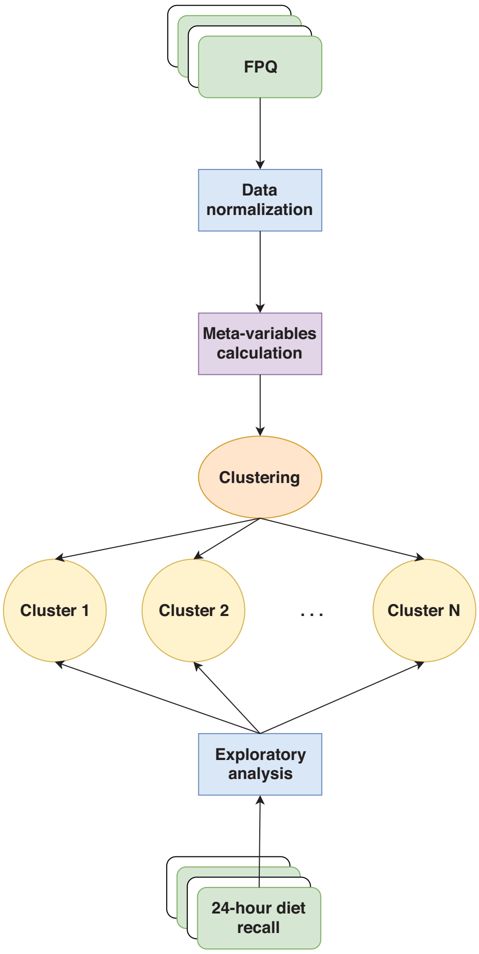

Fig. 1. Methodology pipeline.

average values were calculated. The nine subgroups (i.e. meta-variables) are:

- Breakfast consumption frequency;
- Wheat intake;
- Milk intake;
- Fruits intake;
- Vegetables intake;
- Meat intake;
- Fats intake;
- Drinks intake;
- Other (e.g., fast food intake).

For example, the meta-variable “Wheat intake” is calculated based on the average answers of seven wheat-related questions. Thus, the meta-variable “Wheat intake” represents how good is the participant’s wheat intake.

In our study, we also considered the information about the nutrient intake provided by the 24HRs to define the meta-variables E_kCal, Protein, Carbs and Fat, which represent energy expressed in kilocalories and the intake of protein, carbs and fat in grams, respectively.

By calculating the meta-variables values for each participant, the lower the value is, the closer the participant’s answers are to the ”healthy answers” that are provided by the nutrition specialist.

Finally, one additional meta-variable was calculated as an average over all other meta-variables. We called this variable ”Overall healthy answers score”, since it represents how close are the participant’s answers to the ”healthy answers” in general.

V. Clustering

After calculating the meta-variables for each participant, where each participant is represented by 10 meta-variables, the next step is to run a clustering algorithm. By applying clustering, we can follow the food behaviour of different participants.

In our experiments, we used the k-means clustering algorithm because it is simple, suitable both for small and large data sets, and most importantly it is easy to interpret. More specifically, we used the k-means algorithm, which is a centroid-based algorithm, where each cluster is associated with a centroid (i.e., an instance at center of the cluster). The algorithm has four steps: choose the number of clusters k; select k random instances as centroids; assign all the instances to the closest centroids using Eucledan distance; and, recompute the centroids of the new clusters. The steps three and four are repeated iteratively. This process is finished when: the centroids do not change; or the instances remain in the same cluster; or the maximum number of iterations is reached. We used maximum number of iterations 100. The number of clusters k was chosen experimentally by inspecting the Silhouette score [14]. It represents a ration between the average distance to elements in the same cluster with the average distance to elements in other clusters. The Silhouette score ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. We calculate the Silhouette score for different number of clusters (see Table I). Using the table, we can set the number of clusters at four, since it is a good trade-off between the granularity (number of clusters) and their purity (silhouette score). Figure 2, depicts the Silhouette score for each cluster (i.e. C1, C2, C3 and c4), when the number of clusters is set to four. On the y-axis is the Silhouette score, and on the x-axis are the clusters. It can be seen that the first three clusters (C1, C2 and C3) are cleaner and have an average score over 0.200, and the last cluster is mixed. The number of samples per clusters are: 61 in C1, 62 in C2, 37 in C3 and 36 in C4.

VI. Results and Discussion

Table 2 presents descriptive statistics of the meta-variables per clusters. The meta-variable ”Overall score” shows that the clusters C2 and C3 have lower Overall scores compared to the clusters C1 and C4. Since this variable represents how far are the participant’s answers to the “healthy answers” in general, the clusters C2 and C3 can be considered as “more healthy” clusters and the clusters C1 and C4 as “less healthy” clusters.

TABLE I
AVERAGE SILHOUETTE SCORE WITH RESPECT TO THE NUMBER OF CLUSTERS.

# Clusters	2	3	4	5	6	7	8	9	10
Average silhouette scores	0.322	0.211	0.210	0.193	0.178	0.137	0.156	0.143	0.143

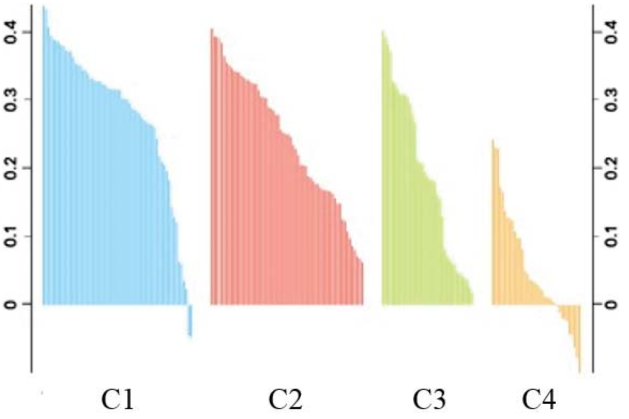

Fig. 2. Silhouette scores for each sample. The number of clusters is set to 4.

This intuition is also confirmed by the meta-variables: “Fruit intake”, “Fat intake” and “Vegetable intake”. For the variable “Milk intake” the cluster C3 has quite high values, indicating that the answers of the people in this cluster significantly differ from the recommended “healthy answers”. For the meta-variable “Breakfast consumption frequency” it can be seen that the people in cluster C1 have a quite high value, indicating that the people in this cluster do not prefer having a breakfast, since their answers differ from the recommended answer, i.e., ”People should consume breakfast”.

The variable BMI in Table 2 represents the body mass index (BMI) of the people in the specific clusters. The people in the cluster C1 have in general higher BMI compared to the rest of the clusters. The BMI was not used as input for the clustering method, it is presented only for exploratory purposes. Thus, the increased BMI of the people in the first cluster supports the intuition that the cluster C1 is a less healthy one.

Additionally, Figure 3 represents scatter plots of the clusters. It can be seen that clusters are quite well separated when using the following variables: ”Overall score”, ”Milk intake” , ”Breakfast intake”, and ”Breakfast frequency”. This may indicate that there are four clusters of people identified in our data set, which can be described with only these four variables. This interesting discovery could lead to simplified FPQ questionnaires, which could significantly decrease the participants’ burden and could ensure participant compliance in similar studies. Having big national data set related to nutrition should ease the process of creation sustainable policies that will ultimately benefit agriculture, human health and the environment.

VII. CONCLUSION

In this paper, we analyzed dietary intake data collected by a FPQ and two dependent 24HRs using unsupervised learning.

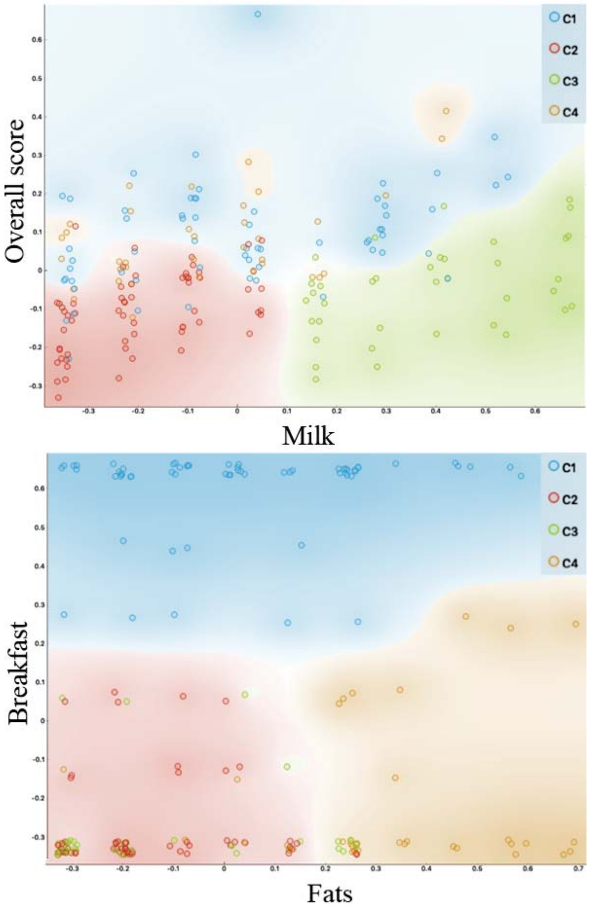

Fig. 3. Scatter plots visualization of the clusters. For each variable, the lower the values is, the closer the participant’s answers are to the “healthy answers”

We analyzed a subset of collected data for 197 participants from the SI.Menu survey carried out in 2016/2017 in Slovenia. Clustering data about the participants, they were split into four different clusters, where two clusters represent participants with more healthy habits, while the other two clusters present participants with less healthy habits. We also identify that these four clusters can be obtained only with four variables, which further indicates that this kind of information can be used to simplify FPQ questionnaires. However, to prove this observation, the complete set of data needs to be analysed considering also other questions and meta-variables.

Splitting data about participants into clusters also makes possible to overcome the problem of missing data. For example, if a participant lacks in answers to FPQ or in one or both 24HR, missing data can be borrowed from the participants from the cluster the participant belongs to.

TABLE II

STATISTICAL ANALYSIS PER CLUSTERS FOR EACH META-VARIABLE THAT WAS USED AS INPUT TO THE CLUSTERING ALGORITHM.

	C1				C2				C3				C4			
	Percentile			Std	Percentile			Std	Percentile			Std	Percentile			Std
	5	50	95		5	50	95		5	50	95		5	50	95	
Overall	0.1	0.07	0.25	0.13	0.26	0.1	0.06	0.1	0.24	0.04	0.16	0.12	0.06	0.05	0.31	0.12
Processed	0.06	0.06	0.28	0.17	0.06	0.06	0.28	0.12	0.06	0.06	0.01	0.08	0.06	0.06	0.61	0.25
Drinks	0.15	0.03	0.72	0.28	0.15	0.15	0.34	0.16	0.15	0.15	0.12	0.12	0.14	0.03	0.44	0.25
Fats	0.31	0.02	0.47	0.24	0.31	0.2	0.13	0.16	0.31	0.2	0.25	0.22	0	0.25	0.69	0.23
Meats	0.23	0.03	0.37	0.18	0.23	0.13	0.17	0.14	0.25	0.03	0.17	0.14	0.15	0.07	0.47	0.17
Vegetables	0.28	0.01	0.36	0.19	0.34	0.05	0.3	0.19	0.34	0.05	0.21	0.18	0.18	0.01	0.26	0.18
Fruit	0.22	0.08	0.28	0.16	0.42	0.02	0.18	0.19	0.42	0.02	0.2	0.21	0.12	0.08	0.48	0.18
Milk	0.34	0.03	0.41	0.27	0.34	0.22	0.03	0.13	0.16	0.41	0.66	0.19	0.34	0.09	0.31	0.21
Wheat	0.16	0.03	0.37	0.18	0.16	0.03	0.3	0.15	0.16	0.1	0.21	0.14	0.16	0	0.44	0.21
Breakfast	0.26	0.66	0.66	0.12	0.34	0.34	0.06	0.12	0.34	0.34	0.06	0.11	0.34	0.34	0.26	0.2
BMI	20.9	28.02	35.2	4.74	18.94	25.4	32.11	5.34	20.94	24.9	35.73	5.39	21.16	26.47	35.43	4.6
E_kCal	1924	2550	3249	457.6	1650.15	2137.5	2817	451.26	1602	2220	3108	451.55	1572	2305.5	2808.75	461.7
Protein	66.88	92.84	119.13	16.82	60.51	87.38	103.29	16.23	58.74	81.4	113.96	16.56	57.64	84.18	102.99	16.15
Carbs	250.8	350.63	446.74	62.92	226.9	293.91	387.34	62.05	220.28	305.25	427.35	62.09	216.15	317.01	386.2	63.48
Fat	61.48	85.94	109.5	15.42	55.62	72.04	94.94	15.3	53.99	74.82	104.75	15.22	52.98	77.7	94.67	15.62

VIII. ACKNOWLEDGMENTS

This work received funding from the European Food Safety Agency (EFSA) - ”EU Menu Slovenia LOT1 and LOT2” under Grant Agreements No. OC/EFSA/DATA/2014/02-LOT1-CT02 and No. OC/EFSA/DATA/2014/02-LOT2-CT03. It has also been supported by the Slovenian Research Agency Program P2-0098 and under Grant U2-AG-16/0672, 0287.

Information and the views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the EFSA. Neither the EFSA institution and bodies nor any person acting on their behalf may be held responsible for the use that may be made of the information contained herein.

REFERENCES

[1] J. Kearney, “Food consumption trends and drivers.” *Philos Trans R Soc Lond B Biol Sci*, vol. 3657, no. 1554, pp. 2793–2807, 2010.

[2] C. R. M. D. S. A. B. S. S. D. T. R. Schatzkin A, Kipnis V and F. LS, “A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based observing protein and energy nutrition (open) study.” *International Journal of Epidemiology*, vol. 32, no. 6, pp. 1054–62, 2003.

[3] W. D. Van Klaveren JD, Goedhart PW and V. Hvd, “A european tool for usual intake distribution estimation in relation to data collection by efsa.” 2012.

[4] D. Greene, P. Cunningham, and R. Mayer, “Unsupervised learning and clustering,” in *Machine learning techniques for multimedia*. Springer, 2008, pp. 51–90.

[5] V. Estivill-Castro, “Why so many clustering algorithms: a position paper.” *SIGKDD explorations*, vol. 4, no. 1, pp. 65–75, 2002.

[6] R. M. Leech, S. A. McNaughton, and A. Timperio, “The clustering of diet, physical activity and sedentary behavior in children and adolescents: a review,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 11, no. 1, p. 4, 2014.

[7] T. Gorely, S. J. Marshall, S. J. Biddle, and N. Cameron, “Patterns of sedentary behaviour and physical activity among adolescents in the united kingdom: Project stil,” *Journal of behavioral medicine*, vol. 30, no. 6, p. 521, 2007.

[8] B. Landsberg, S. Plachta-Danielzik, D. Lange, M. Johannsen, J. Seiberl, and M. J. Müller, “Clustering of lifestyle factors and association with overweight in adolescents of the kiel obesity prevention study,” *Public Health Nutrition*, vol. 13, no. 10A, pp. 1708–1715, 2010.

[9] R. Mistry, W. J. McCarthy, A. K. Yancey, Y. Lu, and M. Patel, “Resilience and patterns of health risk behaviors in california adolescents,” *Preventive medicine*, vol. 48, no. 3, pp. 291–297, 2009.

[10] S. A. Bingham, C. Gill, A. Welch, K. Day, A. Cassidy, K. T. Khaw, M. J. Sneyd, T. J. A. Key, L. Roe, N. E. Day, and et al., “Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records,” *British Journal of Nutrition*, vol. 72, p. 619–643, 1994.

[11] G. Johansson, “Comparison of nutrient intake between different dietary assessment methods in elderly male volunteers,” *Nutrition & Dietetics*, vol. 65, no. 4, pp. 266–271, 2008.

[12] V. Fulgoni, J. Nicholls, A. Reed, R. Buckley, K. Kafer, P. Huth, D. Dirienzo, and G. D. Miller, “Dairy consumption and related nutrient intake in african-american adults and children in the united states: Continuing survey of food intakes by individuals 1994-1996, 1998, and the national health and nutrition examination survey 1999-2000,” *Journal of the American Dietetic Association*, vol. 107, no. 2, pp. 256 – 264, 2007.

[13] A. F. Subar, K. W. Dodd, P. M. Guenther, V. Kipnis, D. Midthune, M. McDowell, J. A. Tooze, L. S. Freedman, and S. M. Krebs-Smith, “The food propensity questionnaire: Concept, development, and validation for use as a covariate in a model to estimate usual food intake,” *Journal of the American Dietetic Association*, vol. 106, no. 10, pp. 1556 – 1563, 2006.

[14] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.