Exploring Students Eating Habits through Individual Profiling and Clustering Analysis

Michela Natilli[1], Anna Monreale[1], Riccardo Guidotti[1,2], and Luca Pappalardo[2]

[1] University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy, name.surname@unipi.it,
[2] KDDLab, ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, name.surname@isti.cnr.it

Abstract. Individual well-being strongly depends on food habits, therefore it is important to educate the general population, and especially young people, to the importance of a healthy and balanced diet. To this end, understanding the real eating habits of people becomes fundamental for a better and more effective intervention to improve the students' diet. In this paper we present two exploratory analyses based on centroid-based clustering that have the goal of understanding the food habits of university students. The first clustering analysis simply exploits the information about the students' food consumption of specific food categories, while the second exploratory analysis includes the temporal dimension in order to capture the information about *when* the students consume specific foods. The second approach enables the study of the impact of the time of consumption on the choice of the food.

Keywords: food analytics, individual models, clustering analysis

1 Introduction

Nutrition is a crucial factor of an individual's lifestyle, that may influence both their physical health and subjective well-being [3, 19]. On the one hand, food is a major source of pleasure, meals are an important opportunity of social aggregation in many cultures [1, 22] and dining together reduce people's perceptions of inequality [13]. On the other hand, an excessive consumption, as well as a deficient intake, of specific aliments can lead to severe physical disorders [2]. In this regard, it has been showed that fast-food and sugar-sweetened drinks consumption is associated with risk of obesity and diabetes [20, 23], whereas adopting a high-fiber diet can improve blood-glucose regulation [29] and consuming fruit and vegetable could have a potentially large impact in reducing many noncommunicable diseases [15]. Given the strong relationship between eating habits and individual well-being, it is important to educate the general population, and especially young people, to the importance of a healthy and balanced nutrition [5, 16].

Unfortunately, publicly available datasets describing eating habits and food consumption – such as the EFSA database[3] and the the ERS database[4] – suffer

[3] https://www.efsa.europa.eu/it/food-consumption/comprehensive-database
[4] https://www.ers.usda.gov/data-products/food-consumption-and-nutrient-intakes.aspx

from several limitations, mainly consisting in the presence of biases of diverse nature, the lack of information at individual level, the short period of data collection and the limited number of individuals involved. Being based on surveys, daily journals or food diaries in which respondents write down what they eat and drink, the information on food consumption is often incomplete and can be affected by the well-known problem of memory effect related to self-report [25], or by the bias due to the tendency in survey respondents to answer questions in a way that will be viewed favorably by others, the so-called "social desirability" bias [21]. Moreover, the great effort required by food diaries can force the survey respondents to simplify the registration of food intake, hence reducing the accuracy of the registered information [14, 18].

Motivated by the criticalness of these aspects, in this paper we propose a data-driven approach to the understanding of eating habits, leveraging the access to a real-world database describing all the meals consumed by around 82,000 students at the canteen of University of Pisa during a 7-years period. This dataset provides us an unprecedented picture of the foods consumed by young people according to their gender, age, geographic origin, course of study, both at lunch and dinner, and their possible evolution over time. We highlight that the food consumption data are collected automatically by means of students electronic cards, hence overcoming the problems that afflict survey-based data collection. The analysis of our data with data mining techniques reveals interesting patterns. In particular, we present two clustering analyses with the aim of segmenting the students under observation based on their food habits. The first clustering analysis is based on a student profile called *foodprint*, which summarizes the food consumption of each student in specific food categories. It leads to the discovery of four main groups of eating habits, which are characterized by the students' propensity to follow a healthy diet. For example, the cluster of *balanced* students describe individuals with a varied diet, including both healthy dishes and junk food, while in the cluster of *voracious* students with a more manifold and fatty diet is preferred. These results show a variation of eating habits across the population, allowing for the possibility of proper interventions to improve the students' diet and subjective well-being. The second clustering analysis exploits the information about *when* the students consume their meals. To this end, we introduce the definition of *temporal foodprints* and we apply an analytical process that uses the clustering analysis for both discovering typical food habits and identifying groups of students having similar food habits. The results of this additional clustering analysis allows us to study the impact of the seasons and the time of consumption on the food choice.

The paper is organized as follows. Section 2 discusses the related work. Section 3 provides a description of real food data under analysis. Section 4 describes the clustering analysis based on student food consumption, while Section 5 shows the clustering analysis based on the temporal food habits of students. Finally, Section 6 concludes the paper.

2 Related Work

The process of construction and extraction of a personal data model is generally referred to as *user profiling*. A *personal data model* contains the systematic behaviors expressing the repetition of habitual actions, i.e., personal patterns. These patterns can be expressed as simple or complex indexes [7], behavioral rules [11], set of events [10], typical actions [27], etc. On the one hand, users profiles are employed to analyze and understand human behaviors and interactions. On the other hand, profiles are exploited by real services to make predictions, give suggestions, and group similar users [6]. Profiles can be classified as individual or collective according to the subject they refer to [12]. An *individual* profile is built considering the data of a single individual. This kind of profiling is used to discover the particular characteristics of a certain individual, to enable unique identification for the provision of personalized services. We talk about *collective* data models when personal data or individual models generated by individual profiling are aggregated without distinguishing the individuals.

In this work we propose an approach using a data modeling similar to [4,8,9], i.e., a vector-based modeling. Moreover, all these approaches and the presented one adopt clustering as methodology to extract the individual and collective patterns. In particular, in [4] the authors defined how to build individual profiles based on mobile phone calls such that the profiles are able to characterize the calling behavior of a user. By analyzing these data model three categories of users are identified: residents, commuters and visitors. Similarly, in [8] the temporal dimension of retail market data is used to discriminate between residents and tourists. A more analytical approach, similar to the one presented in this paper, is described in [9]. The authors present an individual and collective profiling of shopping customers according to their temporal preferences. However, these works focusing on the temporal dimension do not consider what the customers buy. On the other hand, in this work we also take into account the different types of food purchased per meal producing in this way more valuable profiles.

With respect to the field of food there are various work analyzing food habits, food consumption, consequences of a certain diet, etc. However, to the best of our knowledge, this project is the first attempt of using an automatic data-driven approach for extracting the groups of individuals with similar food consumption habits [17]. In the literature it has been shown various predefined groups of people having a certain relationship with certain categories of food. For example in [20, 24, 28] is shown how fast-food and sugar-sweetened drinks consumption is associated with risk of obesity among teenagers, the environmental influences of adolescent, and which is the result of a healthy behavior in school-aged children, respectively. With respect to adults, in [5] is examined the link between dietary habits and depression, in [23] the relationship between local food environments and obesity, and in [22] the role of food in life in various countries. In all these works the group of food consumers are predefined a priori while in this work our target is to extract group of students having similar patterns of food habits.

3 Food Dataset

As proxy of our study we have access to a dataset provided by the Tuscan Institute of Right to Study (DSU)[5] describing 10 millions of meals consumed by about 82,871 students at the canteen of University of Pisa during a period of seven years, from January 1, 2010 to December 26, 2016 (see Table 1-left for more detailed information). The cost of a meal at the canteen varies over the years and with the number of dishes composing the meal. The dataset contains also meals of 19,141 students (23%) who have free meals at the canteen.

meals:	10,034,413
students:	82,871
grant students:	19,141
free meals:	4,730,658
dishes:	950
food categories:	41
period:	2,551 days
from:	01/01/2010
to:	12/26/2016

student id	time	dishes
A4578A	18/04/2015 12:42:00	pasta with tomato sauce, chicken breast, fruit
G23T20	18/04/2015 12:43:00	mushroom risotto, salad, fruit
GE54Y7	18/04/2015 12:44:00	pasta with tomato, fruit
⋮	⋮	⋮

Table 1. Basic statistics of the dataset (left) and description of meals in the Mensana dataset (right). Each meal consists of a student identifier, a timestamp and the list of dishes composing the meal.

Each meal is described by a record indicating the student who consumed the meal, the type of the meal, and the list of dishes (e.g., pasta with tomato sauce, salad, apple, etc.) composing the meal (see Table 1-right for an example). While there are 950 different dishes in the dataset, they are grouped in 41 food categories, each containing dishes with homogeneous nutrients characteristics.

The composition of a meal in terms of the dish composing it changes according to the student's choice based on the menu available at that day. Figure 1-(left) reports that the most popular meal composition is that one with 4 items (i.e., first course, second course, side course and fruit or dessert). The second preferred composition is that one with 3 items (typically, first or second course, side dish and fruit or dessert). Only few students choose additional items with respect to the complete meal (6 items) or meals with only 1 or 2 items.

The total number of meals consumed at the canteen does not vary significantly over the years and it is about 1,400K each year. On the other hand, there is a slightly decrease of the number of students going to the university canteen, passing from the 30k of 2010 to the 27k of 2016. As shown in Table 2, most of the students (around 60,000) consume less than 10 meals in total, denoting

[5] https://www.dsu.toscana.it/

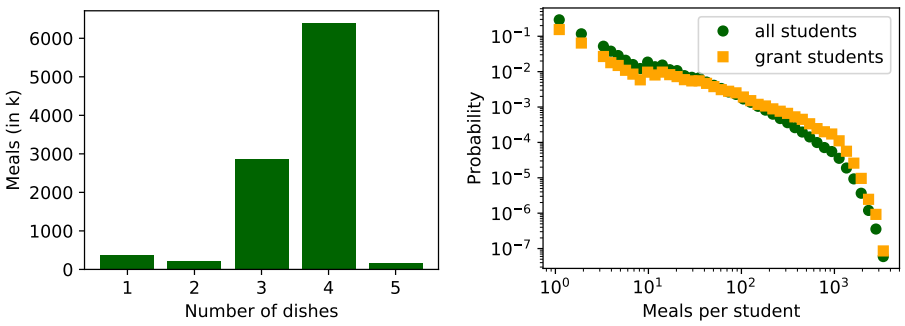

Fig. 1. Distributions: number of dishes per meal (left), meals per student (right).

All students:	
Students with at least 10 meals:	60,112
Students with at least 100 meals:	22,647
Students with at least 1000 meals:	1,448

Table 2. Number of students based on consumed meals

the presence of a heavy-tail distribution of the number of meals consumed by the students (see Figure 1-(right)). This means that the students going at the canteen with regularity are the minority. Only students with scholarships, for whom meals at the canteen are for free, show a slightly higher regularity.

4 Food Consumption Analytics

The dataset described in the previous section enables us to understand students' food habits by means of an appropriate vector-based user profiling. In this section we present a clustering analysis that aims at grouping students, who regularly eat at the canteen, by using information describing their typical food *consumption*.

We describe the student food consumption by using an individual model named *foodprint*. In practice, this model summarizes the consumption of the student in each food category of Table 3. We represent a *foodprint* f_u of a student u by using a vector of 41 attributes a_i with i=1, . . . , 41 (i.e., an attribute for each food category). Let $D_i^{(u)}$ be the number of dishes of the student u in the food category i and $M^{(u)}$ be the total number of meals of the student u, the value of the i-th attribute is $a_i^{(u)} = D_i^{(u)}/M^{(u)}$. We underline that if a student takes two dishes of the same category in a meal (e.g., potato dishes as second course and as side), then in $D_i^{(u)}$ is counted twice, while in $M^{(u)}$ we count only one meal.

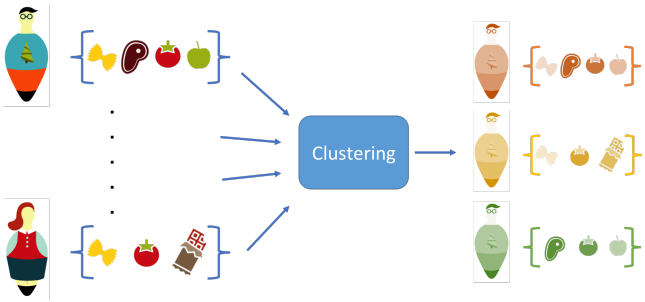

Fig. 2. Clustering Process based on student foodprints

4.1 Data Preprocessing

Since we are interested in analyzing the behavior of regular students, we select a sub-population of the whole students in the dataset. In particular, we do not include in the analysis students who use the canteen sporadically by applying to the data a filter on the student meals based on the frequency in the use of the canteen. We studied the frequency distribution of the consumption of meals, and this leaded to select only students who had consumed at least 100 meals and having a distance between two meals of a maximum of 100 days. The number of days is defined considering the long periods of summer holidays or exam sessions where students may not be in Pisa. After this filtering we obtain 1,607,993 meals related to 6,890 different students. Therefore, we build a foodprint summarizing the food consumption for each one of these students.

Before proceeding with the cluster analysis we compute the correlation between the attributes of the students' profiles to identify (if present) high correlations among food categories. We observed only two strong correlations. A correlation of 0.8 between category a_{33} (soups only vegetables) and category a_{34} (potato and vegetable soups), and a correlation of 0.68 between category a_{101} (tofu o soybeans with cheese) and category a_{102} (tofu or soy with vegetables). Therefore we kept all of the categories separated using all the 41 attributes.

4.2 Clustering Analysis

The goal of the cluster analysis on the student foodprints is to identify groups of students with similar behaviors in terms of food choices. To this aim, we adopt the K-Means [26] clustering algorithm as it is shown in the literature to have good effects in grouping profiles when a vector based model is used [4, 9]. K-Means requires to specify the number of clusters k as parameter. To identify the best value for k, we applied the standard approach (see [26]) that runs the clustering with several values for k and selects the one such that a further increase in k generates no significant improvement in the clusters compactness. This aspect is measured using the Sum of Squared Error (SSE). The optimal number of clusters should be that in which the the curve has a significant inflection (elbow). As

consequence, we selected $k = 4$. Figure 2 describes the clustering process based on student *foodprints*, that starts with the modelling of meals and dishes of a student with a foodprint and ends by extracting students food habits.

The characterization of the four clusters provides some interesting information that allows us to understand how students behave in the university canteen. We assigned a name to each cluster with respect to their characterization given by the food categories of the cluster centroids using the following four adjectives that well describe the feeding behavior of the students of these clusters:

- Cluster 0: *Balanced* (30.95%)
- Cluster 1: *Foodie* (17.43%)
- Cluster 2: *Health fanatic* (33.64%)
- Cluster 3: *Voracious* (17.98%)

Figure 3 depicts a graphical representation of the 4 clusters that summarizes the food profile of each cluster and that we discuss in detail in the following.

Balanced. The cluster called *balanced* covers 31% of the total observed students and most of them are males. Figure 3 depicts a radar chart showing the typical food choices of this group pf students. These students have a rather varied diet: they eat both complex dishes and healthy dishes. Moreover, they take almost always fruit and just in few occasions the dessert; they eat enough vegetables, but also pasta (or couscous) with meat.

Foodie. The *foodie* cluster contains 1,198 students and most of them are males. This cluster differs from the *voracious* cluster because these students insert a greater quantity of healthy food into their diet (see Figure 3). In particular, 50% of their meals contain fruit while 35% contain desserts. They often consume flour with meat, cheese or eggs, but often also eat cooked or grilled vegetables and raw vegetables. Therefore, they have a rather fatty diet, but in some cases they try to include more healthy food in their diet.

Voracious. The cluster named *voracious* involves 1,235 students and most of them are females. Observing the typical food choices of this group in Figure 3 we can see that they tend to eat mostly fatty foods. It is evident that they prefer flour with meat, they often take dessert instead of fruit and they eat a lot of dishes based on potatoes, red meat and salami. Moreover, compared to others students, voracious students have the highest consumption of fried dishes (potatoes or vegetables) and the lower consumption of cooked or grilled vegetables.

Health fanatic. The cluster called *health fanatic* includes 2,311 students with an almost equal distribution between male (58%) and female (42%), and it is characterized by a healthy diet, as shows the radar chart in Figure 3. Indeed, they eat very often fruit, raw vegetables and cooked or grilled vegetables. These students, compared to others, consume the lowest share of flours with meat, cheese or eggs, and they have a low consumption of red meats or salami, desserts,

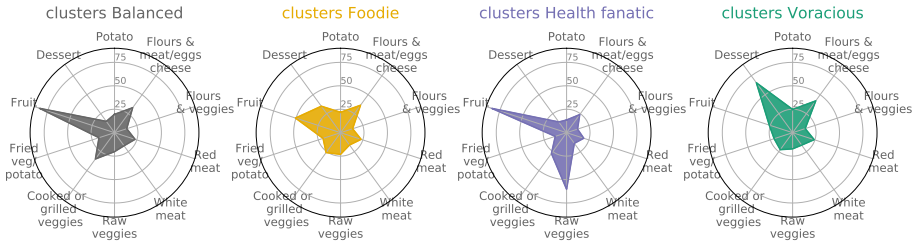

Fig. 3. Main food categories choices among clusters

white meats, potato dishes and fried foods. They are the main consumers of tofu or soy-based dishes. Although the share of consume is low, we underline that the supply of these categories in the canteen is limited with respect to the others.

We also perform a study on the clusters aiming at verifying the influence of the demographical origin of students on the food choices. We observe that, looking at the origin of the students in the various clusters, the differences are not significant. There is a tendency towards the prevalence of students from southern Italy and Islands (Sicily and Sardinia) in the *foodie* group, but in general it does not seem that geographical origin is an important feature in the definition of clusters. Moreover, we also investigate the distribution of students within cluster according to their course of study, in order to verify whether there are differences among students attending humanistic courses and students attending scientific courses. Although the majority of students attends a scientific course, we find that in the *voracious* cluster there is the largest quota of students from humanistic courses.

5 Temporal Food Habits Analytics

The profiling and the clustering analysis presented in the previous section do not consider any temporal information of *when* the student consumes the meals. In this section, we extend the previous analysis that is able to group students on the basis of their food *habits* taking into consideration the time of food consumption. This clustering process, depicted in Figure 4, is based on the definition of the *temporal foodprints* of a student, and on two tasks called *food habits discovery* and *student grouping*.

The student *temporal foodprint* is an extension of the *foodprint* f_u with the following temporal information: the *year* and the *season* of the meals consumption and the knowledge that discriminates between lunch and dinner. This extension allows us to define the typical consumption of the student in each food category during the lunch or dinner in a specific year and season (winter, spring, summer and autumn). Therefore, for each student we have the set of his temporal foodprints $F_u = \{f_u^{t_1}, \ldots, f_u^{t_n}\}$, where each $f_u^{t_i}$ is a *temporal foodprint* with $t_i{=}(y_i, s_i, \mu_i)$ representing the temporal information composed of three elements: y_i denoting the *year*, s_i denoting the *season*, and μ_i indicating lunch or dinner.

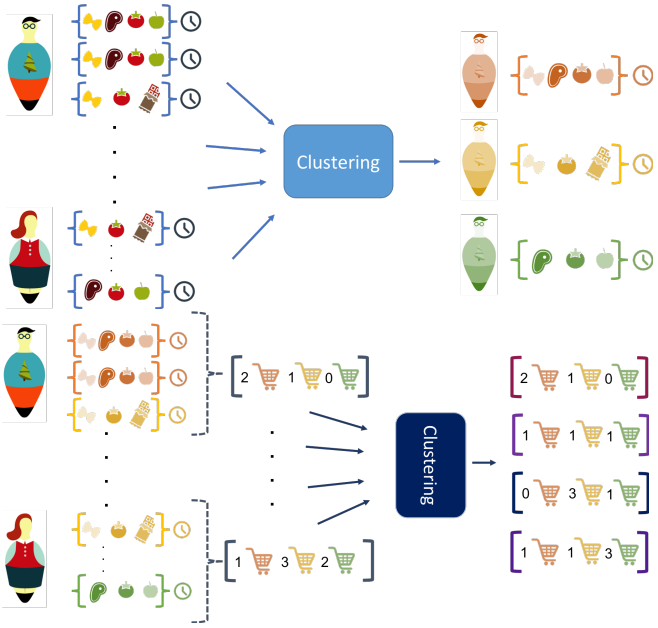

Fig. 4. Two-steps clustering process based on temporal foodprints: segmentation of the temporal foodprints *(top)* and grouping of the food habits profiles *(bottom)*

Temporal Food Habits Discovery. Similarly to the previous section, given the *temporal foodprints* of the students, we can apply a clustering algorithm to extract typical food habits by exploiting the specificity of the student consumption with respect to the temporal information t used. This clustering does not provide a students' segmentation but a segmentation of the temporal foodprints, i.e., the students temporal habits (see Figure 4-(top)). Thus, as highlighted in Figure 4-(bottom), a student can have his temporal foodprints distributed over different clusters, meaning that he is characterized by different food habits.

Student Grouping. The knowledge of how temporal foodprints are distributed for each student enables a student segmentation on the basis of the temporal food habits. To this aim, for each student we construct a *food habit profile* that describes the intensity of the student presence in each cluster (food habit). We represent the student habit profile through a vector of attributes h_j with $j = 1, \ldots k$, where each attribute h_j represents the intensity of a certain temporal food habit. Given the set of *temporal foodprints* of a student u, we denote by $N_{(u)}^{h_j}$ the number of his temporal foodprints belonging to the food habit h_j, and by N_u the total number of his temporal *foodprints*. Finally, we model the intensity of a student u in a cluster as $h_j = N_{(u)}^{h_j}/N_u$. We can now group the students according to their temporal food habits by re-applying a clustering algorithm on the student *food habits profiles* as shown in Figure 4-(bottom).

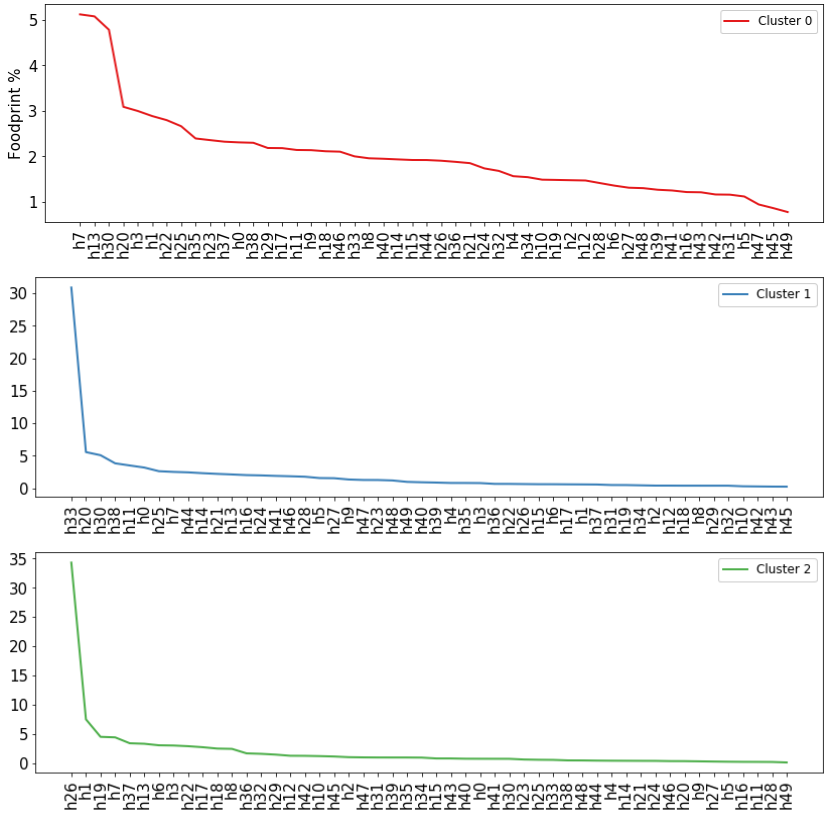

Fig. 5. Sorted levels of intensity for the temporal foodprints within the centroid of clusters 0, 1, 2. With the exception of cluster 0 which has three temporal foodprints consistently higher than the rest for all the other centroids it is possible to isolate a unique dominant temporal foodprint.

5.1 Clustering Analysis

In order to perform the double-cluster analysis described above we selected students having at least 10 meals over the whole observed period. After this filter there are 45,952 students to be analyzed. Note that, the filter is different from the previous analysis because now we are not interested only into regular students.

For discovering the temporal food habits (Figure 4-(bottom)) we adopt again the k-means algorithm. Following the same procedure described above we select k=50 as number of cluster. In other words, we fix the number of different temporal food habits to be equal to 50. The 50 different temporal habits are used to build the *food habits profiles*. For each student, his food habits profile describes how his food behavior is distributed over the different temporal food habits.

On top of the *food habits profiles* we employ k-means for grouping the students according to their food habits. We performed this analysis using differ-

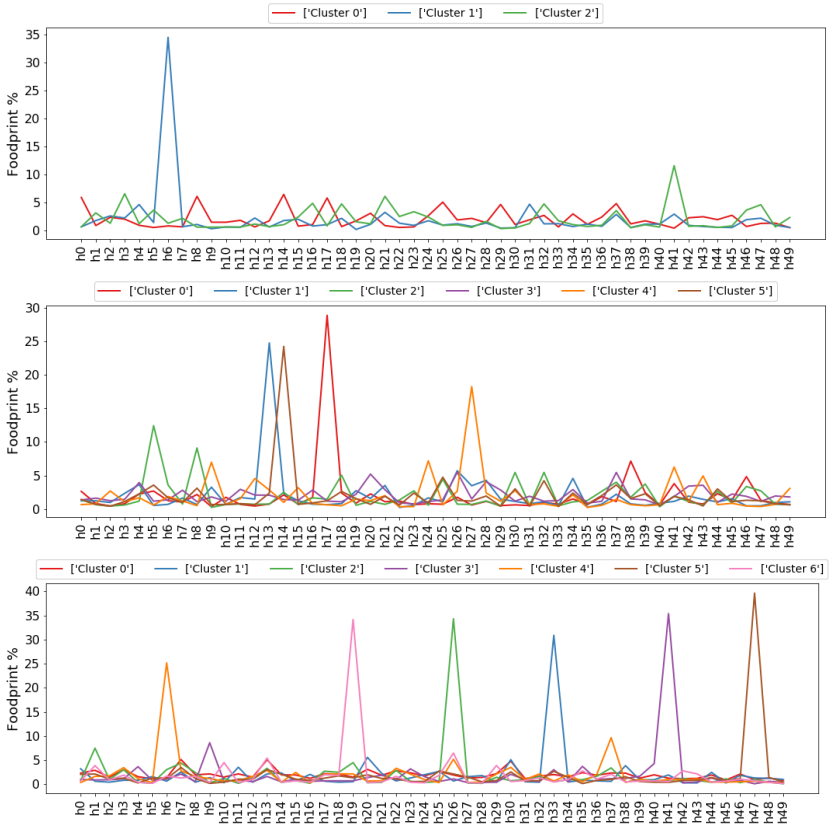

Fig. 6. Levels of intensity for the temporal foodprints using different temporal discretizations: year-lunch/dinner with three clusters, year-month-lunch/dinner with six clusters, and year-season-lunch/dinner with seven clusters, respectively.

ent temporal aggregations t: year and lunch/dinner information, month and lunch/dinner information, season and lunch/dinner information. In this paper we report only the results related to this last combination. However, similar results (in terms of the presence of the same pattern of having a unique dominant temporal foodprint) are obtained using the other temporal combinations. Using again the elbow method we selected $k = 7$ as number of different groups for the food habits profiles.

We analyze the clusters of the students' temporal food habits with respect to three dimensions: seasonality, time of the meals (lunch or dinner) and consumption of food categories. The first aspect we considered are the levels of intensity of the temporal food habits in the centroids of the clusters obtained. Figure 5 reports the sorted intensities for the clusters of food habits with identifiers 0, 1 and 2, respectively. On the y-axes is reported the intensity of the corresponding food habit that we have in the x-axis.

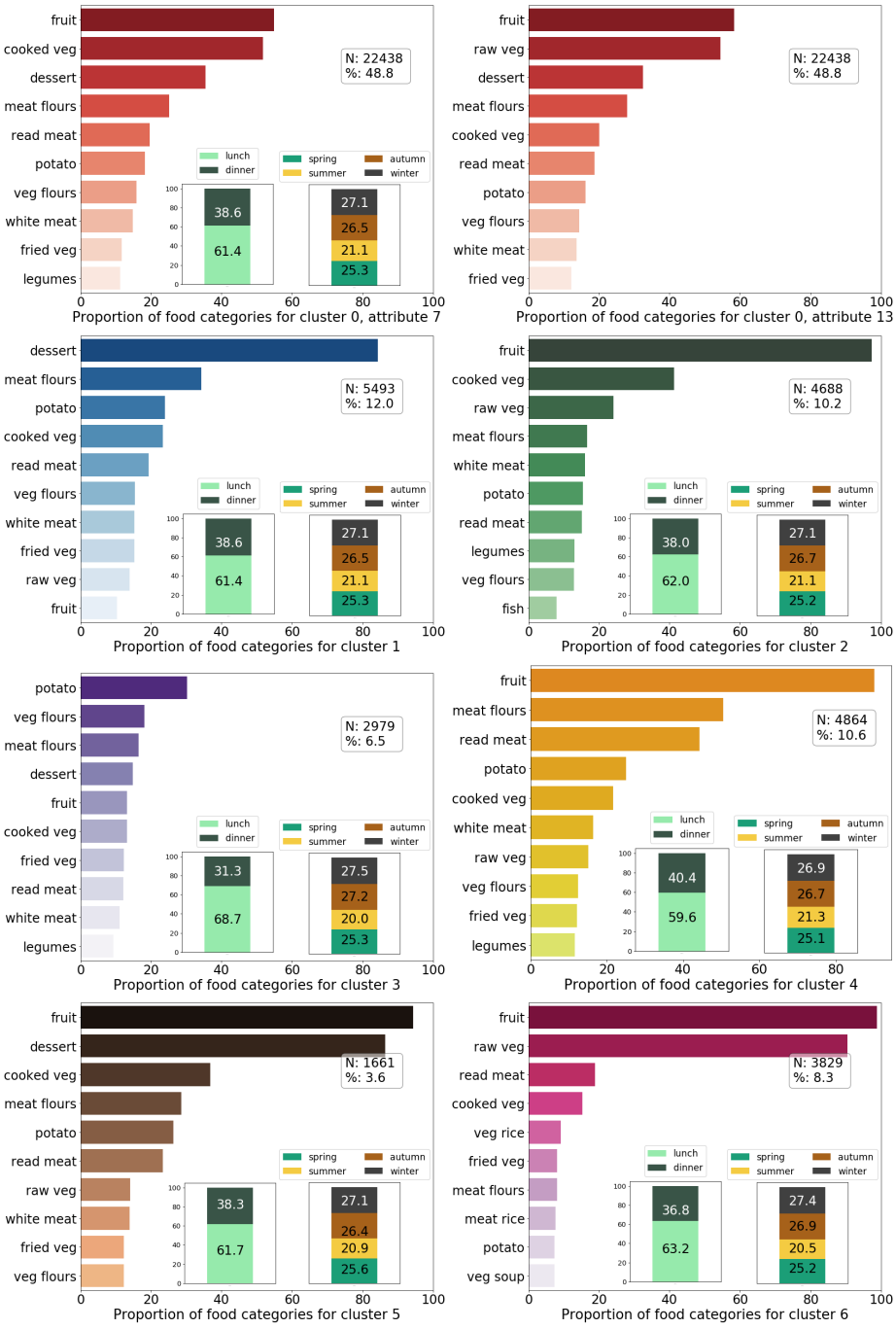

Fig. 7. Dominant attributes among clusters

We found out that, with the exception of cluster 0, which has three food habits consistently higher than the rest, for all the other centroids it is possible to isolate a unique dominant food habit. Therefore, we can characterize the seven groups by means of the dominant food habits of their centroids.

Before analyzing the dominant food habits we highlight in Figure 6 how a different choice of the temporal dimension t and consequently of the number of clusters k, lead to different centroids. However, for all these centroids, the pattern d[illegible]ving a unique dominant temporal foodprint always holds. Thus, this seems a prerogative of this kind of modeling that is independent from the choice of the temporal dimension and of the number of clusters, that can be adjust to the needs of the data analyst without loosing the possibility to describe an entire food habit profile with a unique temporal foodprint.

In the following we present the results of characterization of the 7 clusters obtained by setting as temporal information the seasons and time of the meal consumption (the third graph in Figure 6). Figure 7 shows for each cluster the information about the highest levels of food consumption over the different food categories, the impact of the season and of the time of consumption. We observe that the seasons have no impact on the food choice of the observed students. This aspect can be due to the fact that the food categories generalize too much the food consumption, i.e., it does not properly consider the recipes ingredients putting for example a light "summer pasta" with tomato and basil in the same category of a more rich "winter pasta" with cooked tomato and other vegetables. On the other hand, some differences are present with respect to the distribution between lunch and dinner. In *Cluster 4* we have the dominant food habit with the largest percentage of meals consumed during dinner time. The students in this group have a proper and nutrient dinner mainly consisting in meat (red and white) with potato or other vegetables as side dishes. We highlight that *fruit* is the top ten of all the dominant food habits, and it is almost always the most frequent, except for *Cluster 1* and *Cluster 3*. This can be due to the fact that in every possible meal composition a fruit or a dessert can be added: the fact that fruit is preferred over dessert is a good indication of the students choices. It is interesting to notice how for the students of Cluster 1 the dessert is far more important than fruit. We can finally differentiate *Cluster 1*, *Cluster 3* and *Cluster 4* from the others. Indeed, these three clusters identify a quite fat diets and diets rich of starches and animal proteins, while the other are more based on the consumption of raw and cooked vegetables.

6 Conclusion

In this paper we have presented two exploratory analyses based on centroid-based clustering aiming at understating the food habits of university students. The first clustering analysis, based on a student profile that describes the student consumption in specific food categories, allow us to discover four different clusters: *voracious*, *health fanatic*, *foddies* and *balanced*. The second analysis instead is based on a student profile that takes into account also the temporal

information describing *when* the meal is consumed. This additional information together with a double clustering analysis allows us to perform a deeper analysis of the students' food habits also studying the impact of the seasons and the time of consumption on the food choice. The results of our analyses could be useful for suggest improvements to the students diet. Clearly, individual suggests might lead to privacy concerns that should be addressed appropriately. Interesting future improvements of the work include the use of food categories that introduce a lower generalization of the recipes of the consumed dishes and the link of the students' consumption with the nutrient values of the meals.

Acknowledgments. This work is part of the project Rasupea-Mensana funded by Regione Toscana on PRAF 2012-2015 funds as part of the "Nutrafood" project. Rasupea-Mensana is promoted by the University of Pisa and the Scuola Superiore Sant'Anna in collaboration with the Regional Agency for the Right to University Study of Tuscany (Azienda Regionale per il Diritto allo Studio Universitario Toscana - DSU) and Pharmanutra. This work is also partially supported by the EU H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures", grant agreement 654024 *"SoBigData"* (http://www.sobigdata.eu). The authors thank the staff of DSU (as part of Rasupea) and of University of Pisa for providing data and support for data linkage.

References

1. The cultural dimension of food. Technical report, Barilla center for food nutrition, 2017.
2. K. R. DeVault and D. O. Castell. Updated guidelines for the diagnosis and treatment of gastroesophageal reflux disease. *The American journal of gastroenterology*, 100(1):190, 2005.
3. A. Eertmans, F. Baeyens, and O. Van Den Bergh. Food likes and their relative importance in human eating behavior: review and preliminary suggestions for health promotion. *Health Education Research*, 16(4):443–456, 2001.
4. L. Gabrielli, B. Furletti, R. Trasarti, F. Giannotti, and D. Pedreschi. City users' classification with mobile phone data. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 1007–1012. IEEE, 2015.
5. M. M. Gillen, C. N. Markey, and P. M. Markey. An examination of dieting behaviors among adults: Links with depression. *Eating behaviors*, 13(2):88–93, 2012.
6. R. Guidotti. Personal data analytics: capturing human behavior to improve self-awareness and personal services through individual and collective knowledge. 2017.
7. R. Guidotti, M. Coscia, D. Pedreschi, and D. Pennacchioli. Behavioral entropy and profitability in retail. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.
8. R. Guidotti and L. Gabrielli. Recognizing residents and tourists with retail data using shopping profiles. In *International Conference on Smart Objects and Technologies for Social Good*, pages 353–363. Springer, 2017.
9. R. Guidotti, L. Gabrielli, A. Monreale, D. Pedreschi, and F. Giannotti. Discovering temporal regularities in retail customers shopping behavior. *EPJ Data Science*, 7(1):6, 2018.

10. R. Guidotti, A. Monreale, M. Nanni, F. Giannotti, and D. Pedreschi. Clustering individual transactional data for masses of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 195–204, New York, NY, USA, 2017. ACM.
11. R. Guidotti, G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi. Market basket prediction using user-centric temporal annotated recurring sequences. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 895–900. IEEE, 2017.
12. M. Hildebrandt. Defining profiling: a new type of knowledge? In *Profiling the European citizen*, pages 17–45. Springer, 2008.
13. A. P. JULIER. *Eating Together: Food, Friendship, and Inequality*. University of Illinois Press, 2013.
14. M. Livingstone, A. Prentice, J. Strain, W. Coward, A. Black, M. Barker, P. McKenna, and R. Whitehead. Accuracy of weighed dietary records in studies of diet and health. *Bmj*, 300(6726):708–712, 1990.
15. K. Lock, J. Pomerleau, L. Causer, D. R. Altmann, and M. McKee. The global burden of disease attributable to low consumption of fruit and vegetables: implications for the global strategy on diet. *Bulletin of the World Health Organization*, 83(2), 2005.
16. M. I. Maiorino, G. Bellastella, D. Giugliano, and K. Esposito. Can diet prevent diabetes? *Journal of diabetes and its complications*, 31(1):288–290, 2017.
17. I. Martinucci, M. Natilli, V. Lorenzoni, L. Pappalardo, A. Monreale, G. Turchetti, D. Pedreschi, S. Marchi, R. Barale, and N. de Bortoli. Gastroesophageal reflux symptoms among italian university students: epidemiology and dietary correlates using automatically recorded transactions. *BMC Gastroenterology*, 18(1):116, 2018.
18. R. M. Ortega, C. Pérez-Rodrigo, and A. M. López-Sobaler. Dietary assessment methods: dietary records. *Nutricion hospitalaria*, 31(3), 2015.
19. C. M. Peat, L. Huang, L. M. Thornton, A. F. Von Holle, S. E. Trace, P. Lichtenstein, N. L. Pedersen, D. W. Overby, and C. M. Bulik. Binge eating, body mass index, and gastrointestinal symptoms. *Journal of psychosomatic research*, 75(5):456–461, 2013.
20. M. A. Pereira, A. I. Kartashov, C. B. Ebbeling, L. Van Horn, M. L. Slattery, D. R. Jacobs Jr, and D. S. Ludwig. Fast-food habits, weight gain, and insulin resistance (the cardia study): 15-year prospective analysis. *The lancet*, 365(9453):36–42, 2005.
21. D. L. Phillips and K. J. Clancy. Some effects of "social desirability" in survey studies. *American Journal of Sociology*, 77(5):921–940, 1972.
22. P. Rozin, C. Fischler, S. Imada, A. Sarubin, and A. Wrzesniewski. Attitudes to food and the role of food in life in the usa, japan, flemish belgium and france: Possible implications for the diet–health debate. *Appetite*, 33(2):163–180, 1999.
23. J. C. Spence, N. Cutumisu, J. Edwards, K. D. Raine, and K. Smoyer-Tomic. Relation between local food environments and obesity among adults. *BMC public health*, 9(1):192, 2009.
24. M. Story, D. Neumark-Sztainer, and S. French. Individual and environmental influences on adolescent eating behaviors. *Journal of the Academy of Nutrition and Dietetics*, 102(3):S40–S51, 2002.
25. S. Sudman and N. M. Bradburn. Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, 68(344):805–815, 1973.
26. P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2006.
27. R. Trasarti, R. Guidotti, A. Monreale, and F. Giannotti. Myway: Location prediction via mobility profiling. *Information Systems*, 64:350–367, 2017.

28. C. A. Vereecken, S. De Henauw, and L. Maes. Adolescents' food habits: results of the health behaviour in school-aged children survey. *British Journal of Nutrition*, 94(3):423–431, 2005.
29. L. Zhao, F. Zhang, X. Ding, G. Wu, Y. Y. Lam, X. Wang, H. Fu, X. Xue, C. Lu, J. Ma, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science*, 359(6380):1151–1156, 2018.

7 Appendix

cat	foodcat	category
a_8	potato	potato dishes
a_{10}	plain flours	plours (pasta, couscous, dumplings) in white
a_{11}	meat flours	flours (pasta, couscous, dumplings) with meat/cheese/eggs
a_{12}	fish flours	flours (pasta, couscous, dumplings) with fish
a_{13}	veg flours	flours (pasta, couscous, dumplings) with vegetables
a_{20}	plain rice	graminaceae (rice, spelled, etc.) in white
a_{21}	meat rice	graminaceae (rice, spelled, etc.) with meat/cheese/eggs
a_{22}	fish rice	graminaceae (rice, spelled, etc.) with fish
a_{23}	veg rice	graminaceae (rice, spelled, etc.) with vegetables
a_{31}	meat soup	soups with meat/ cheese/ egg
a_{32}	fish soup	soups with fish
a_{33}	veg soup	soups with vegetables
a_{34}	legumes soup	potato and legumes soups
a_{51}	read meat	red meat / salami
a_{52}	white meat	white meat
a_{53}	processed meat	meat (white/red) - processed
a_{60}	fish	fish
a_{62}	fried fish	fish - fried
a_{71}	cheese salad	cheese salad
a_{81}	raw veg	raw vegetables
a_{82}	cooked veg	cooked or grilled vegetables
a_{83}	legumes	legumes
a_{91}	meat pie	eggs, molded, pies with meat and cheeses
a_{92}	veg pie	eggs, molded, pies with vegetables
a_{93}	fried veg	vegetables / fried potatoes / other fried
a_{101}	soy cheese	tofu or soy with cheeses
a_{102}	veg soy	tofu or soy with vegetables
a_{212}	cheese	cheese
a_{213}	sandwiches	sandwiches, piadines, pizzas, bunnies
a_{415}	fruit	fruit
a_{416}	dessert	dessert

Table 3. The food categories.