Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

FoodVis: A System for Visual Exploration of Food Composition Data

Maria Eduarda M. de Holanda

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinicius Ruela Pereira Borges

Brasília
2024

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

FoodVis: A System for Visual Exploration of Food Composition Data

Maria Eduarda M. de Holanda

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinicius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof.a Dr.a Maristela Holanda
CIC/UnB

Prof.a Dr.a Roberta Oliveira
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 18 de setembro de 2024

Dedicatória

Dedico esse trabalho a todos que me ajudaram durante a minha trajetória acadêmica. Ao meu pai, que, como formado em Ciência da Computação, me orientou e esteve sempre ao meu lado durante todo o caminho. À minha mãe, que sempre me apoiou e esteve presente em cada passo, com muito carinho. E ao meu companheiro Bruno, que me acompanhou em todos os momentos desta jornada, oferecendo sua força quando eu mais precisei e sua paciência nos desafios mais difíceis.

Agradecimentos

Meus sinceros agradecimentos a todos que contribuíram de forma significativa para a realização deste trabalho.

Primeiramente, gostaria de agradecer ao meu orientador Prof. Dr. Vinicius Ruela Pereira Borges, por toda a paciência e orientação ao longo deste trabalho, que foram essenciais para o meu desenvolvimento.

Agradeço também a toda a minha família, por sempre me apoiarem em todas as decisões e momentos, proporcionando o suporte necessário para que eu seguisse em frente.

Sou igualmente grata à Universidade de Brasília (UnB) e aos demais professores do Departamento de Ciência da Computação, pelos valiosos aprendizados e contribuições ao longo do curso.

Além disso, estendo meus agradecimentos à ProIC/UnB, à Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF), processo nº 00193-00001288/2021-16, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 143060/2023-6, por terem financiado esta pesquisa e tornado este projeto possível.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

FoodVis: A System for Visual Exploration of Food Composition Data

Maria Eduarda M. de Holanda
Dep. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
eduarda.holanda@aluno.unb.br

Vinícius R. P. Borges
Dep. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
viniciusrpb@unb.br

***Abstract*—Analyzing food composition is a challenging task mainly due to the wide diversity of ingredients and nutrients in food samples, resulting in complex data presenting both categorical and numerical attributes. Moreover, the ever-growing volume of food composition data makes data analysis difficult and time-consuming for nutritionists. This scenario motivated us to propose FoodVis, an interactive web-based tool that allows nutritionists to gain insights and extract meaningful knowledge by interpreting visualizations of food composition data. The tool was devised by taking into account activities commonly performed by nutritionists, such as identifying similar food products, analyzing nutrient distributions, exploring ingredient relationships, and labeling products based on specific attributes. In this sense, FoodVis incorporates interactive visualizations based on dimensionality reduction, graphs, and parallel coordinates, each one conveying specific patterns and properties of data in both global and local analysis. We performed experiments to validate FoodVis, focusing on the effectiveness and quality of visualizations. Moreover, interviews with nutritionists were conducted to assess the usability of FoodVis, providing valuable insights into its practical applications in nutrition.**

***Index Terms*—food composition data, data visualization, data analysis, interactive tool, point placement strategies, web application**

I. Introduction

Food is composed of various compounds essential for maintaining health and supporting bodily functions [1]. These compounds include macronutrients like proteins, fats, and carbohydrates, as well as micronutrients like calcium, potassium, and vitamin C [2]. Additionally, information regarding the nutritional composition of food can be obtained through laboratory analysis or food composition tables. However, for consumers, this information is usually available on nutritional labels, which provide details about their nutrients and ingredients [3].

In the field of nutrition, understanding the particularities of food components and the relationship between them is not an easy task. The wide diversity of foods and their constituent nutrients and ingredients presents significant challenges to dietitians, nutritionists, and other health-related professionals, here called the domain specialists. Therefore, analyzing the large amount of data associated with various food products can be overwhelming and time-consuming when performed manually by them. In this sense, technology comes in handy to provide innovative solutions to face these challenges [4].

In the last two decades, several researchers have explored food data analysis using automatic or semi-automatic intelligent strategies [5]. Some works have proposed food embeddings to represent food items as continuous vectors [6], data mining techniques in the animal and food industries [7], and a scalable platform for recognizing traditional food knowledge using deep learning models [8]. Whilst those techniques accomplished their proposal, the interpretability of machine learning models and the produced results might not be intuitive and simple to understand by nutritionists.

A promising approach involves combining visualization techniques and natural language processing (NLP) to present complex data understandably and conveniently. Visualization techniques aim at generating intuitive graphical representations of data, enabling researchers to identify trends, patterns, and connections that, otherwise, might be overlooked. NLP techniques are employed to extract meaningful information from textual data, such as food descriptions, labels, and nutritional content, enhancing the data analysis process. For instance, visualization-based approaches have been introduced to improve fraud detection in e-commerce transactions [9], explore demographic representation in clinical trials [10], and visualize high-dimensional data using bi-kernel t-SNE [11].

To the best of our knowledge, there is a lack in the literature regarding data visualization approaches designed for analyzing food composition data. Most current methods are not suitable to address the challenges posed by food datasets, which are characterized by mixed and heterogeneous attributes related to nutrients (fat, calories, iron, etc) and ingredients (lists of text spans). Nutritionists and researchers would greatly benefit from interactive visualization tools that can make the analytical process of food composition data more efficient while retaining user control. Visualization and NLP approaches also provide advantages in this regard, providing intuitive ways to explore and interpret large amounts of food-related data, enhancing the decision-making process, and facilitating the process of knowledge discovery.

Taking that into consideration, an interactive web-based tool, named FoodVis, has been developed specifically for domain specialists to address these challenges. This tool employs advanced visualization techniques such as Principal Component Analysis (PCA) [12], Uniform Manifold Approximation and Projection (UMAP) [13], a Graph-based visualization [14], and Parallel Coordinates [15]. By leveraging these methods, we provide an intuitive and interactive platform to explore food datasets. Furthermore, the tool utilizes NLP to process product descriptions and ingredient lists from the dataset, enhancing the visualizations and clustering

techniques. To ensure the tool's relevance to specialists, we formulated four tasks that guided its development: identifying similar products, finding relationships between ingredients, analyzing overall nutrient distribution, and labeling the dataset using the graphical representation according to specific ingredients.

The main contributions of this paper are described as follows:

- Development of an interactive visual exploration system that allows users to analyze food composition data using visualization techniques and participate in the process of knowledge discovery;
- Presentation of a structured representation of food data that integrates both categorical and numerical columns to be used as input to the visualizations and clustering techniques;
- Integration of NLP techniques to extract meaningful information from text-based food descriptions, labels, and nutritional content.

This paper is structured as follows. Section II presents the related works on the use of visualization and data mining strategies for food data. Section III presents the task analysis, detailing the steps considered to evaluate the visualization tool's usability. Section IV provides a thorough explanation of each visual component. Section V describes the methodology employed in the development of the tool. Section VI describes the experimental results focusing on the evaluation of multidimensional projections and expert feedback. Section VII and VIII conclude this paper discussing the findings and limitations of our tool as well as possibilities for future work.

II. Related work

Several works in the field of data analysis have attempted to develop visualization tools and techniques that would simplify the extraction of significant information from complex datasets. This section presents related works that have significantly contributed to this area, focusing on both visualization tools and techniques that are particularly relevant to food data analysis.

Visualization tools have proven to be useful in various contexts, as proposed by Maçãs et al. [9]. The authors designed ATOVis to improve the detection of Account Takeover fraud in e-commerce transactions. This specific fraud is challenging due to the existence of fraudulent activity patterns that are not identified by a machine learning algorithm when a fraudster adapts his approach. To overcome this issue, the tool employs three innovative visualization approaches to highlight the detailed fraud model. Specifically, task abstraction for ATO detection, visualization models for transactional data, and a multiscale timeline for a comprehensive overview of the data are used to develop the model. To validate the tool, a dataset with over 4 million e-commerce transactions was considered. Furthermore, the results showed that, compared to using spreadsheets alone, the tool significantly increased the speed of detecting specific fraud patterns.

Similarly, another visualization tool, developed now by Carmeli et al. [10], explores demographic representation in clinical trials of FDA-approved medicinal products between 2015 and 2021. The authors were motivated by the diverse representation of clinical trials for assessing drug safety and efficacy, using publicly available data from the FDA's Drug Trials Snapshots (DTS) and disease incidence data from the National Cancer Institute and Centers for Disease Control and Prevention. Moreover, by scraping and aggregating data from FDA sources and validating it against DTS reports, the interactive web-based tool allows users to explore 339 FDA drug and biologic approvals across different aspects such as demographics, approval year, and indication. It was observed that this proposal not only improved the understanding and communication about clinical trial diversity but also suggested improvements for data access, reporting, and stakeholder engagement to enhance trial inclusion and health outcomes.

Eftimov et al. [6] introduced a method for representing food items as vectors of continuous numbers (food embeddings) to allow advanced food data analysis using machine learning approaches. For that purpose, the authors have applied the food embeddings in four tasks: automated determination of food groups, detection of food classes (raw, derivative, or composite), identification of similar food concepts, and qualitative evaluation by an expert in the area. The results demonstrated that these vector representations significantly outperform traditional methods used in food data analysis, highlighting the potential of food embeddings to advance knowledge discovery in the field.

Dimensionality reduction techniques can be useful for visualizing high-dimensional data since they map multidimensional data into lower dimensions by preserving the relationships between data instances [16]. Zhang et al. [11] proposed a method for visualizing high-dimensional data with out-of-sample extensions using bi-kernel t-SNE. It employs Gaussian kernel functions and principal component analysis (PCA) to approximate projections from high-dimensional to low-dimensional spaces and effectively separate inliers and outliers. The results show that bi-kernel t-SNE can extend t-SNE projections to new data points with high accuracy, achieving better outlier detection and visualization compared to other dimensionality reduction methods. Finally, it demonstrates that the proposed method significantly improves visualization quality and computational efficiency for handling out-of-sample data.

Ana Belén García [7] explored the use of Data Mining (DM) techniques for obtaining valuable insights from datasets in the animal and food industries, focusing on examples within the cattle industry. This problem was addressed by a DM-based method, which involved collecting, analyzing, and interpreting data to implement effective controls in food safety, animal health, public health, and environmental programs. Furthermore, García detailed how these techniques can manage and analyze data from various sources, including animal identification systems, movement records, and health data, to enhance traceability and food supply chain management.

Mursanto et al. [8] proposed the development of a scalable platform for gathering and recognizing traditional food knowledge using deep learning models. The techniques employed include collecting high-quality images of traditional Indonesian foods, developing an automatic and scalable food recognition system, and implementing multiprocess inference services to handle efficiently simultaneous requests. The

results showed that the model with the best performance achieved an AUROC score of 0.99, and the multiprocess inference service improved the request success rate by up to 70%. These findings indicate that the platform effectively recognizes traditional foods and can be extended to include additional food types and handle higher user requests.

These related works demonstrate different applications and benefits of visualization tools and data analysis techniques shown in various contexts and for different purposes.

III. Task analysis

Nutritionists face the challenging task of analyzing the composition of food, which is essential to promoting healthy eating habits [17], formulating balanced diets [18], and even evaluating the nutritional quality of food [19]. The growing number of available food composition datasets is a concern for nutritionists once they are able to analyze smaller datasets considering the current practices. Specifically, the detailed nutritional values and ingredients contained within these data require efficient and accurate analysis methods.

Through a review of literature approaches for food composition data analysis and consultations with nutritionists collaborating with this research, we identified some practical difficulties and gaps in current analysis methods. To address these challenges, FoodVis aims to support nutritionists and researchers in four distinct tasks. These tasks are designed to make it easier to identify patterns, trends, and complex relationships within the data samples as follows:

- T1: **Identification of similar products**: The specialist should be able to identify and compare products with similar nutritional content and composition. This task provides a broad overview of the dataset, enabling users to explore and detect clusters or patterns within all the data. This global perspective is important as it is the foundation for more focused analysis.
- T2: **Analysis of nutrient distribution**: The specialist should be able to observe the distribution of nutrients among products and identify patterns or outliers. This task allows users to explore deeper into the nutritional aspects, highlight trends in nutrient concentration, detect outliers, and compare nutrient profiles within specific categories.
- T3: **Finding relations between ingredients**: The specialist should be able to explore relationships between pairs of ingredients and identify common combinations. This task aims to help the nutritionist discuss the significance of the ingredient pairings and understand how they might influence the nutritional value and composition of food items.
- T4: **Labeling and coloring products**: The specialist should be able to label and color the products based on different criteria to enhance the analysis of the other tasks. It is integral to the entire workflow, enabling users to categorize and distinguish products based on various descriptive attributes.

IV. Visualization Techniques

In this section, we describe each visual component that are incorporated into FoodVis, in which Figure 1 reveals a complete overview. The first visualization is the Global View, composed of a Point Placement strategy, which displays the entire dataset. The other two visualizations, Parallel Coordinates and Graph-based Visualization, require a filtering step to be properly used, which can be performed by cross-filtering Visualization 1 and then selecting the points based on their color or ingredients.

A. Point placement Visualization

The goal of point placement visualizations is to map high dimensional objects to lower dimensional points in two-dimensional space in such a way that similar objects are placed by nearby points and dissimilar objects are placed by distant points in the visual space [20]. As a result, the visual analysis of the generated layout takes advantage of the human perception system to interpret and identify the global and local patterns of the underlying data according to their similarity relationships. Two state-of-the-art were considered to be included in the visualization: Principal Component Analysis (PCA) [12] and Uniform Manifold Approximation and Projection (UMAP) [13].

PCA is a statistical technique that reduces data dimensionality by performing a linear mapping of the data in a high dimensional space to a lower dimensional space by taking into account the variances of each attribute and their relations to generate the transformed data. Differently, UMAP computes a weighted graph based on the similarity relations for each pairwise distance, in which the graph's weights are determined in an optimization process that defines the low dimensional embedding. Due to the distinct approaches underlying these techniques, PCA proves to be preferable when the data exhibits linear relationships, while UMAP is more suitable for datasets with complex and non-linear structures. In PCA, the axes represent the principal components, which capture the most variance in the data. In contrast, the axes in UMAP do not have a specific meaning, as the technique focuses on preserving the relative distances between points rather than aligning them to specific components.

Figure 2 (a) and (b) illustrates a scenario when the specialist selected to analyze only the macronutrients of the food products (Protein, Carbohydrate, Fats, Fiber, and Sugar) as well as the Energy (kcal) obtained [21]. It is possible to see that categories are more distinctly separated in 2(b) UMAP, "Protein Food" are mostly spread on the top of the visualization while "Mixed Dishes" can be mostly observed in the middle. Now for 2(a) PCA, the groups of points show more overlap but it is still possible to see products from the same categories tightly close together.

Furthermore, the specialist can interact with the visualization to enhance their experience. These include zooming in and out, panning across the plot, selecting groups of points, and even downloading it as an image. Figure 2 also shows a few interaction examples: 2(c) users can click on legend items to show or hide specific categories or groups of data points; 2(d) users can explore original information such as the nutrients and ingredients, which can be useful for understanding the composition of a food product.

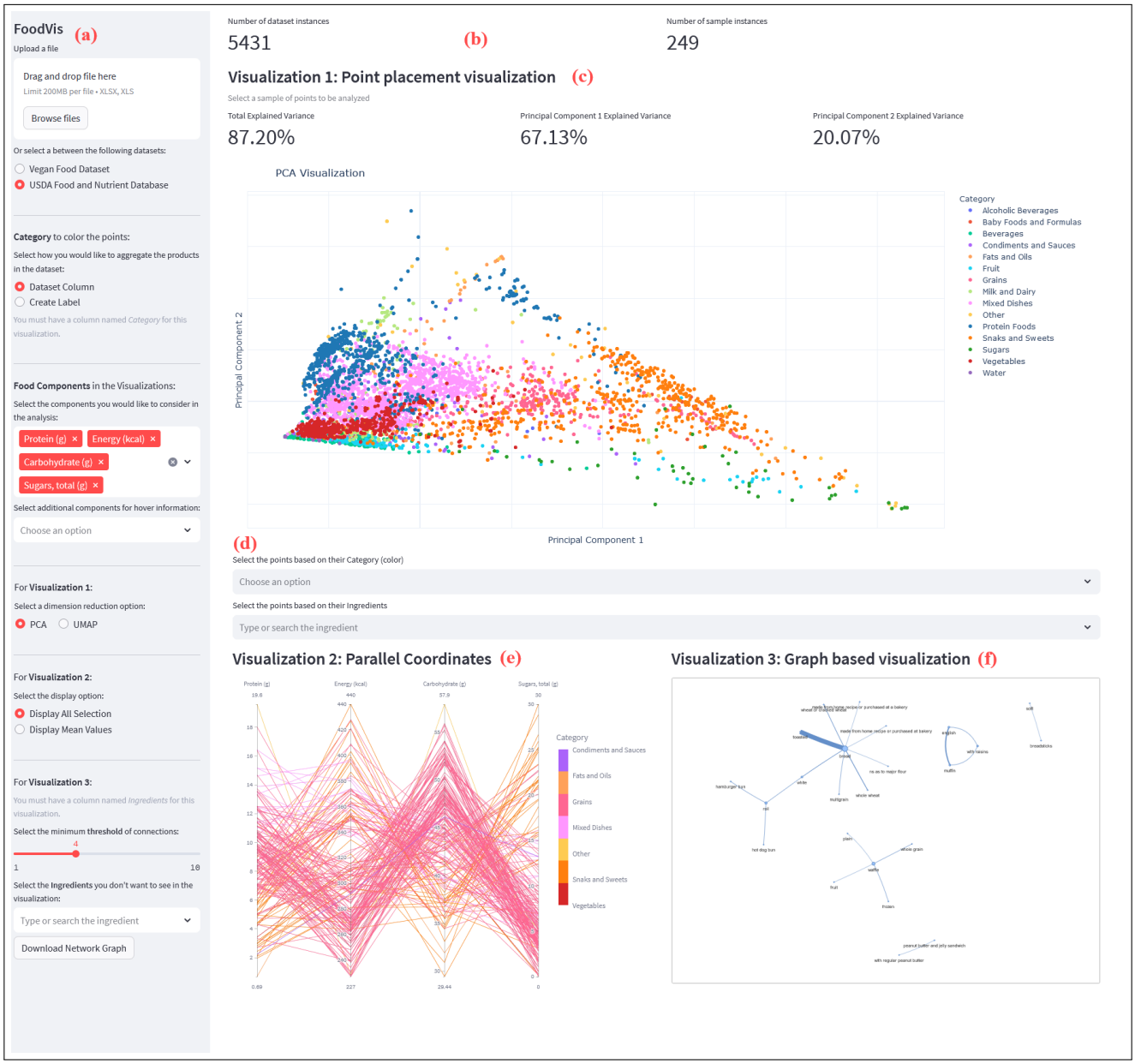


Fig. 1. The system is composed of some components: (a) the attribute selection area where users can upload datasets, select attributes, and configure visualization options; (b) overview of number of instances in the global and local view respective; (c) the point placement visualization displaying high-dimensional data in a two-dimensional space; (d) selection options for visualization 2 and 3; (e) the parallel coordinates view allowing for the comparison of nutritional attributes; and (f) the graph-based view illustrating relationships between pairs of ingredients.

B. *Parallel Coordinates*

The parallel coordinates view is a multivariate visualization technique that displays data attributes as parallel axes, with data points represented as polylines connecting across these axes. It allows for the exploration and analysis of relationships and patterns across multiple attributes simultaneously [15]. One of the benefits of this technique is the ability to see groups of polylines converging to certain value ranges on different axes. This analysis can be further enhanced by using different colors for the lines and adding interactive tools. Nonetheless, this visualization presents limitations for very large datasets, which may cause too much visual cluttering, making visual interpretation unfeasible.

In our tool, this visualization is designed to allow specialists to explore the distribution of various nutritional components across different food items. In Figure 3 we can see examples of the two different displays option discussed in Section V-C. The user can observe either (a) all selected points across parallel axes or (b) the mean of attributes categorized by color. In the first scenario all products categorized as "Fats and Oils" were selected, whereas in the second one, products from "Condiments and Sauces", "Fruit", "Milk and Dairy", "Mixed Dishes" and "Protein Foos" were also considered.

Figure 3 also shows different ways the specialist can interact with the visualization. They can 3(c) drag the polylines along the axes to filter intervals, or 3(d) drag the axis names across the plot to rearrange the order of the attributes. As we can infer from these images, the high amount of "Energy (kcal)" is directly associated with the quantity of "Total Fats" in the selected category.

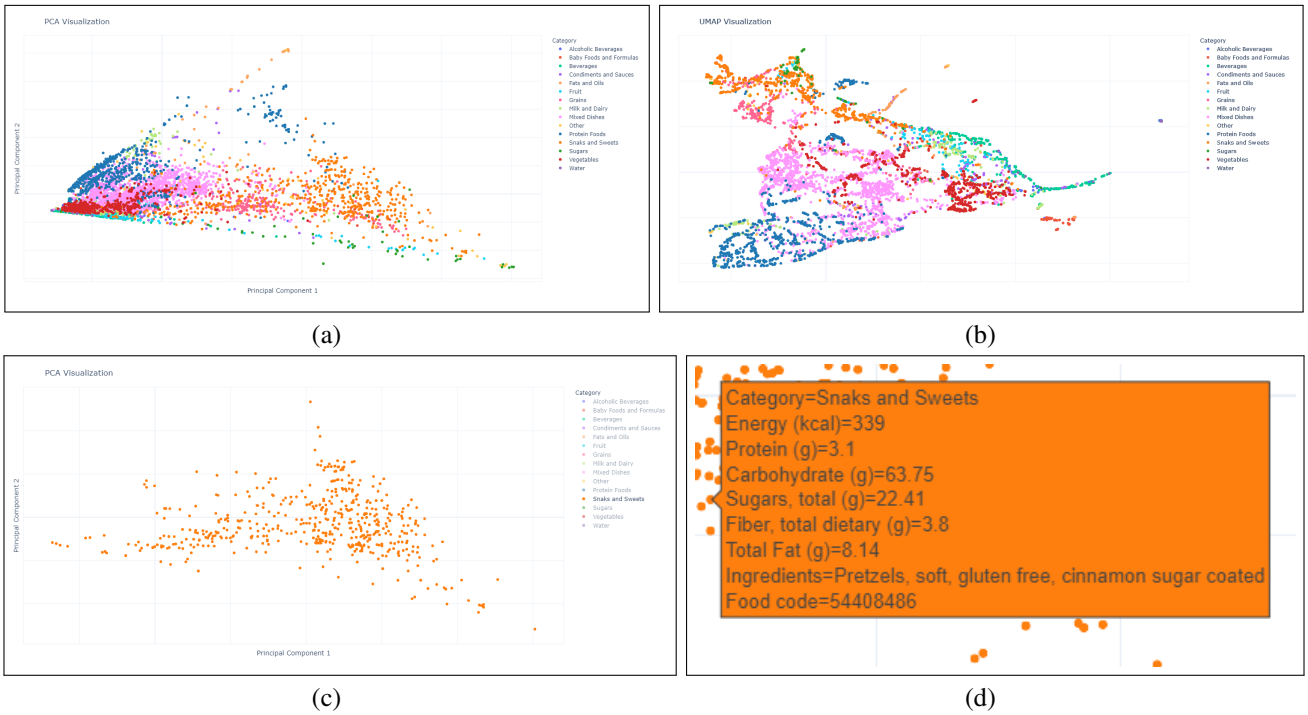


Fig. 2. Layouts produced by the point placement visualization considering only macro nutrients: (a) PCA; (b) UMAP. Interaction example on a PCA Layout: (c) legend filtering; (d) hover information capture.

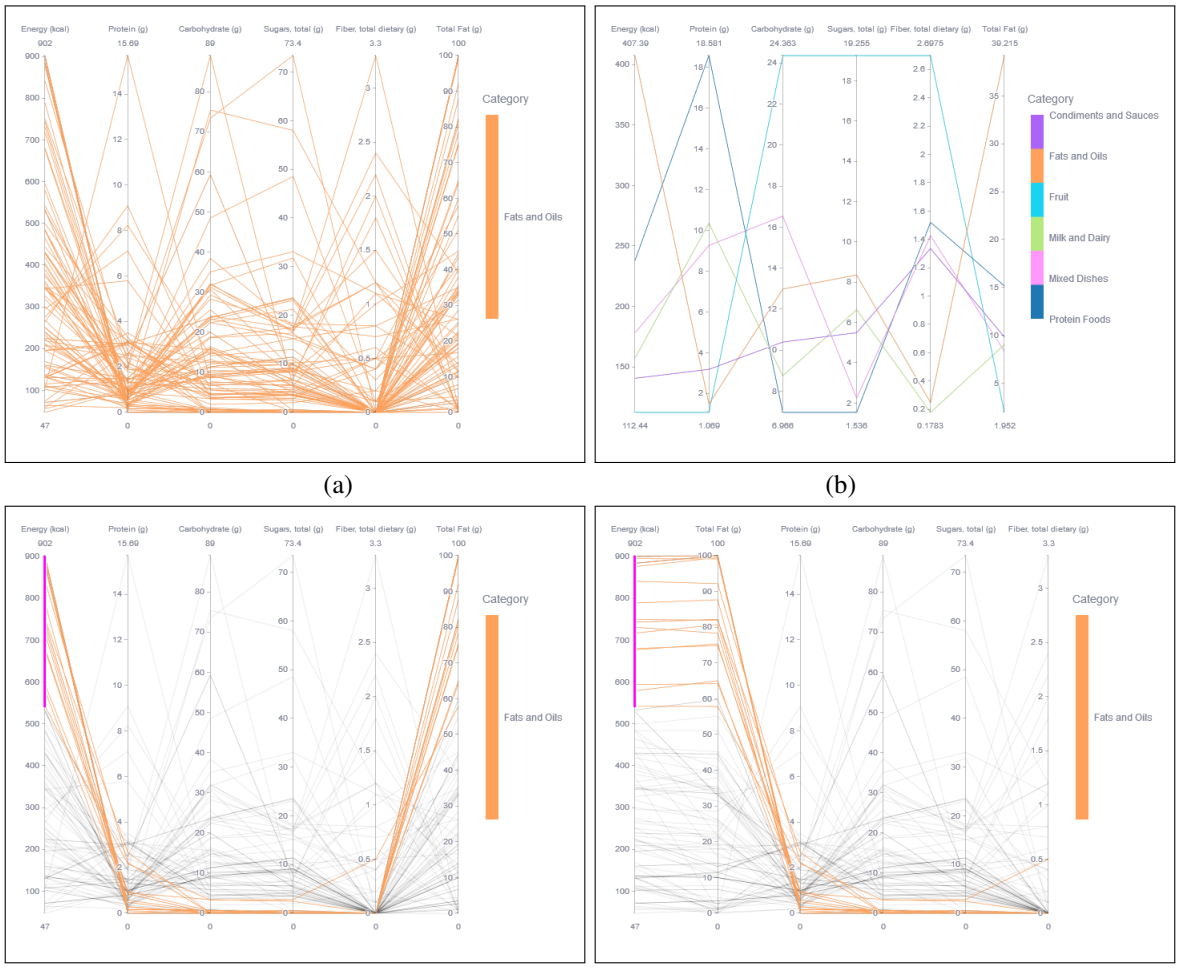
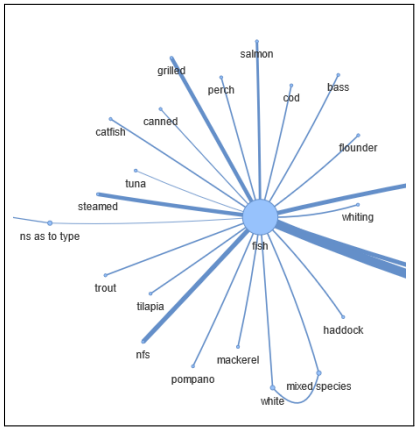
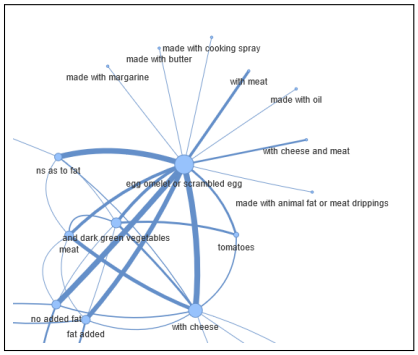


Fig. 3. Layouts produced by parallel coordinates considering display options: (a) all selection; (b) mean values. Interaction example on a parallel coordinates layout where the user can: (c) filter by the range of attribute values; (d) rearrange the the order of attributes.

C. Graph-based Visualization



(a)



(b)

Fig. 4. Examples of sub graphs that can be observed in the Graph-based Visualization when "Protein Foods" category is selected.

Graph theory is a branch of mathematics focused on the study of graphs, which are structures used to model pairwise relationships between objects [22]. A graph consists of nodes (or vertices) representing entities, and edges (or links) connecting pairs of nodes, symbolizing the relationships between them.

In the context of FoodVis, ingredients are represented as nodes, and the edges denote the co-occurrence of these ingredients in different food items. This visualization technique is useful for mapping complex networks of relationships, allowing for the identification of connections and associations that may not be easily seen through other forms of analysis.

Figure 4 illustrates a scenario where all products from "Protein Foods" were selected. Analyzing the entire visualization can be somewhat overwhelming if a great amount of products is selected, but when we focus on nodes we can note some patterns. For instance, in this selection, it is possible to observe that (a) "fish" appear several times and with different types like "tuna", "salmon" or "perch". Similarly, (b) "egg omelet or scrambled egg" is also quite common, when made with "margarine", "butter" and "oil". The user can gain this insight by increasing and decreasing the zoom in the visualization as well as clicking and dragging on the plot.

V. Methodology

In this section, we outline the methodology employed in the development of the *FoodVis* tool. We begin by detailing the proposed workflow to display the visualizations, including the steps from data upload to final output. We considered two datasets to validate the visualizations and to guide the tool's development. They were chosen due to their focus on nutrition and ingredient analysis for various food products. After that, we explore the user interface, highlighting the customization options available to enhance their experience. Also, we describe the pre-processing techniques applied to the data to improve the quality of the visualizations so that we can finally provide a comprehensive analysis of each visual component, emphasizing their unique purpose within the system.

The system was developed with Python, leveraging Matplotlib and Plotly for dynamic visualizations, while the Streamlit framework ensures a smooth interface for navigation and exploration. Furthermore, it incorporates scikit-learn [1] for handling the dataset, offering natural language processing capabilities for analyzing textual data.

A. Workflow

The workflow of FoodVis involves data loading, pre-processing, and rendering visualizations based on user selection as shown in Figure 5. Here is an overview of the different components and their respective roles within the system:

1. **Dataset Upload**: The tool enables users to upload their food composition datasets. It is designed to handle datasets that contain information on nutrients, ingredients, and product descriptions.
2. **Attribute Selection**: Once the dataset is uploaded, users select specific attributes they wish to analyze. This step allows users to focus on relevant aspects of the data, adjusting the visualizations to their analytical needs.
3. **Data Pre-processing**: The data is pre-processed to ensure it is in a suitable format for visualization. This may include handling missing values, normalizing data ranges, and transforming categorical variables into numerical representations.
4. **Visualization Generation**: The tool then generates up to three visualizations: a Global View and two Local Views that are resulted from a filtering step, all described in Section IV.

B. Datasets

Two food composition datasets presenting similar aspects regarding their attributes were considered to design the proposed method.

1) Vegan Food Composition Dataset: This dataset was built according to the nutrition facts of vegan products labels that are commercialized in Brazil. It is a small dataset which presents only 276 data instances, described by 15 attributes as shown in Table I. The attribute [Ingredients] refers to a list of ingredients (separated by comma) and the attribute {Nutrients} corresponds to a set of single attributes constituted by minerals, vitamins, protein, fat, etc. Moreover, the food products were categorized as three types of meat, three types of poultry, two types of fish, three types of dairy and two types of pork.

[1] https://scikit-learn.org/

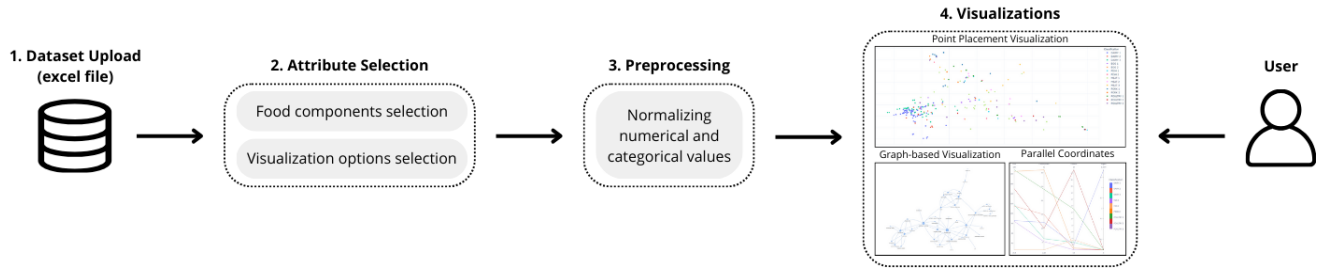


Fig. 5. Workflow of the interactive food data analysis tool. The workflow starts with (1) dataset upload from a excel file which is followed by (2) the selection of attributes, including food components and visualization options. The next step is (3) preprocessing, which involves normalizing numerical and categorical values. Finally, (4) the data is visualized through three different visual components.

TABLE I
DESCRIPTION OF THE VEGAN FOOD COMPOSITION DATASET.

Attributes	Type	Cardinality
Food category	Nominal	1
Product's name	Nominal	1
[Ingredients]	List of Nominals	1
{Nutrients}	Numerical	12

2) USDA Food Composition Dataset: The United States Department of Agriculture's (USDA) Food Composition Database[2] contains information for various types of food including the amounts of different vitamins and minerals found in the foods as well as macronutrient percentages. This dataset is a lot bigger, presenting $5,431$ data instances which are characterized by 49 attributes. Table II describes the types of all attributes. Each food code contained in the dataset is linked to one WWEIA (What We Eat In America)[3] category number. A pre-processing step was required to access the category name for each product.

TABLE II
DESCRIPTION OF THE USDA FOOD COMPOSITION DATASET.

Attributes	Type	Cardinality
Food category	Nominal	1
Food code	Nominal	1
[Ingredients]	List of Nominals	1
{Nutrients}	Numerical	46

The datasets used in this study consist of detailed nutritional information and ingredient lists for a variety of food items. These datasets were chosen to align with the tool's focus on nutrition and ingredient analysis, providing a rich source of data for visualization and exploration. However, it is possible to manually upload any dataset with similar characteristics. By allowing users to upload a personalized dataset, the tool becomes versatile and flexible to a wide range of food composition datasets with diverse goals.

C. User Selection Interface

We designed a selection interface to allow users to customize the visualizations according to their specific goals, which can be seen on Figure 1(a). This step defines the scope of the analysis, as it ensures that the generated visualizations are aligned with the user's goals, providing more relevant insights [23].

The process begins with the categorization of products. Users can assign colors to each product based on predefined categories in the dataset, such as food groups or nutrient profiles. Alternatively, users can set the colors of symbols associated with the visualizations to the data instances according to their criteria. This feature can be useful when the input data is not originally labeled, but the user needs to incorporate some type of categorization into the visualizations, such as coloring points by combining ingredients. This strategy enables nutritionists to include additional information based on data attributes in the visualization, thus creating visual distinctions to reflect their research focus.

Next, the user should select the food components to analyze within the visualization according to his particular goals. This step is essential for gradually selecting the most relevant attributes in the dataset, such as specific nutrients, ingredients, or other food characteristics. By focusing on particular components, users can interact with the visualizations to highlight the most relevant data properties to their analysis.

Finally, the tool offers multiple customization options for each type of visualization, which are further detailed. For the point placement visualization, users can choose between two dimensionality reduction techniques, PCA or UMAP, depending on which method better fits their data. In the Parallel Coordinates visualization, users can either display all points across parallel axes for detailed comparison or view the mean of attributes categorized by color, allowing for a summarized view. For the graph-based visualization, users can filter out specific ingredients they consider irrelevant and set a threshold to display only ingredient pairs that co-occur at least the defined number of times. Therefore, these three levels of customization enable the proposed web-based tool to meet a wide range of user needs, making it versatile and robust for exploring food composition data.

D. Preprocessing

As food datasets can present attributes from different types (numerical and nominal), a preprocessing is required to generate a new structured representation that properly fits as input to the visualization techniques [24].

After the dataset is uploaded, a text cleaning step is required for nominal attributes to eliminate inconsistencies that could otherwise lead to errors or misinterpretations. When it comes to the Point placement Visualization, we employed the one-hot encoding approach to transform the

[2]USDA Food and Nutrient Database for Dietary Studies
[3]What We Eat in America (WWEIA) Food Categories

"Ingredients" component to binary attributes once each food product can present a variable number of values (ingredients). The procedure consists of mapping each value of "Ingredients" to a new binary attribute indicating the presence or absence of the underlying characteristic. For example, if a food product has the "water" and "onion" as ingredients but does not have soy flour then both water and onion will be represented as 1 while soy flour 0. This step is necessary since only numerical attributes are allowed as input to this visualization. Moreover, the remaining numerical attributes presenting real values were normalized, in which all obtained values for each attribute are in the range $[0, 1]$.

The Parallel Coordinates does not require data preprocessing since this technique handles fewer attributes than the other visualization techniques considered. Moreover, Parallel Coordinates maps each attribute to a vertical axis in the visual space so that it is straightforward to draw the polylines representing each sample of the dataset.

The Graph-based Visualization demanded an additional data preprocessing step. The goal of this visualization is to create a graph where each node represents an ingredient, and the edges indicate other ingredients that co-occur in the food items, with the weight reflecting the frequency of these connections, as can be illustrated on Figure 6. To achieve this, each pair of ingredients is processed, and a threshold is applied so that only ingredient pairs that co-occur more than a user-defined number of times are displayed. Users may also filter out specific ingredients to do not appear in the visualization if they judge to be irrelevant in the context of the analysis. For example, the specialist can omit "water" as it appears in most products.

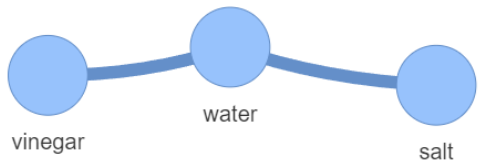


Fig. 6. Example of a generated graph representing co-occurrence of ingredients in distinct foods of the dataset.

VI. Experimental Results

The specialist's performance using the proposed tool depends on the quality of the produced layouts. Thus, we performed experiments aiming to evaluate the quality of the visualization techniques discussed previously in Section **??**. The experiments used both Vegan and USDA food composition datasets, and state-of-the-art metrics in the visualization field were considered for that purpose.

A. Evaluation of Multidimensional Projections

For the multidimensional projections introduced on Figure 1(c), four state-of-the-art and recent dimension reduction techniques were assessed: Principal Component Analysis (PCA) [12], Uniform Manifold Approximation and Projection (UMAP) [13], t-Distributed Stochastic Neighbor Embedding (t-SNE) [25] and TriMap [26]. PCA was chosen due to its popularity in data analysis tasks, while t-SNE, UMAP and TriMap were considered due to the their abilities when dealing with data that contain different types of attributes and good discriminability when forming groups in the layout [27].

In order to evaluate the quality of the selected visualization techniques, we followed two steps: finding the best choices of hyperparameters in each method and then comparing the results with two state-of-the-art evaluation metrics. Following the quality assessment strategy presented in related researches [28], we chose the Silhouette Coefficient (SC) to compute the separability between groups of points [29] and the Trustworthiness to measure the preservation of the local structure and similarity relations between neighbor data instances [30].

To compute SC, the points in the visual space must be clustered using an algorithm. Here, we chose K-Medoids algorithm to determine K clusters since medoid centers are more appropriate representatives than mean centers due to the presence of categorical values in the dataset. The Euclidean distance is considered for K-Medoids and for computing the silhouette coefficient. The analysis of the clustering results was performed by running K-Medoids and varying K within the range $[2, 30]$, in which the silhouette coefficient is computed for each obtained clustering. Finally, we compute the mean and standard deviation from those values.

Table III describes the obtained SCs for the visualization techniques on the Vegan Food Dataset. TriMap and UMAP achieved the higher silhouette scores, while PCA and t-SNE obtained lower scores. These results indicate that the clusters formed are the most distinct and well-separated for TriMap and UMAP. It is worth noting that SC is independent regarding the label "Category", thus emphasizing that the focus is to analyze the separability between the groups of points in the layout.

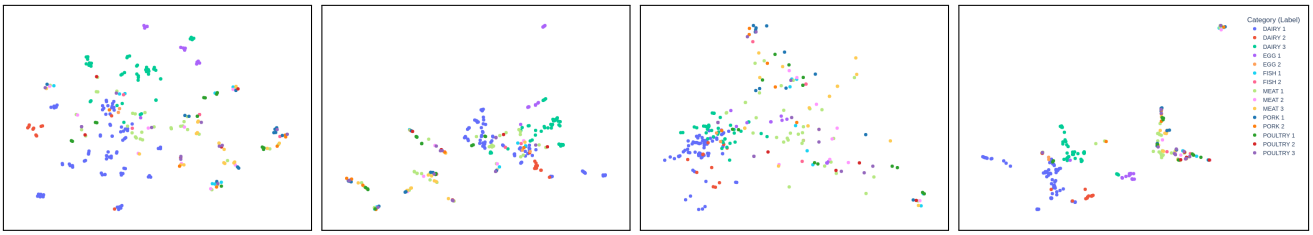
TABLE III
Evaluation Metrics for Vegan Food Composition Dataset

Visualization Techniques	Silhouette Coefficient	
	Mean	*Standard Deviation*
t-SNE	0.4407	0.0727
UMAP	0.5016	0.0668
PCA	0.4176	0.0469
TriMap	0.5513	0.0425

Table IV describes the obtained silhouette coefficients for the visualization techniques on the USDA's dataset. TriMap yielded a higher SCs value, while t-SNE achieved a poor performance. We can see from Figure 8(d) that while this tight clustering on TriMap's projection contributed to a high SC value, it could also limit the interpretability of the visualization since it requires zooming in to differentiate between the points.

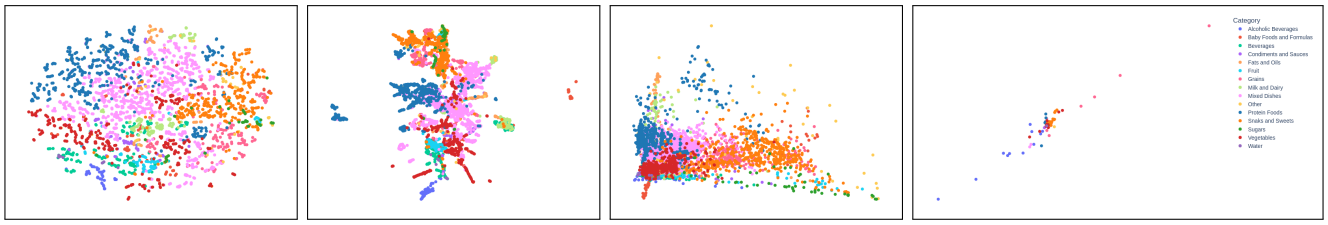
TABLE IV
Evaluation Metrics for USDA Food Composition Dataset

Visualization Techniques	Silhouette Coefficient	
	Mean	*Standard Deviation*
t-SNE	0.3298	0.0143
UMAP	0.4161	0.0215
PCA	0.3935	0.0236
TriMap	0.6080	0.1766



(a) (b) (c) (d)

Fig. 7. Layouts produced by the visualizations based on point placement by considering both ingredients and nutritional values of the Vegan Food Dataset: (a) t-SNE; (b) UMAP; (c) PCA; (d) TriMap.



(a) (b) (c) (d)

Fig. 8. Layouts produced by the visualizations based on point placement by considering both ingredients and nutritional values of the USDA Food and Nutrient Dataset: (a) t-SNE; (b) UMAP; (c) PCA; (d) TriMap.

As for the evaluation using the trustworthiness metrics, for each object in the high dimensional space, the proportion of its k-nearest neighbor is computed and compared in relation to the k-neighbor points in the visual space for the corresponding 2D point. Finally, we compute this neighborhood preservation rate for a neighborhood value k by averaging the precision for all data instances. This rate value lies in the range $[0, 1]$, in which higher values are related to better preserving the neighborhood structure of data instances in the visual space.

Figure 9 presents the results of the trustworthiness for the considered multidimensional visualizations. Figure 9(a) suggests that PCA and TriMap present better preservation of original neighbors in the low dimensional space for the Vegan Food Dataset, indicating that the local structure and similarity relations were better retained than UMAP and t-SNE. Nevertheless, Figure 9(b) indicates better results with t-SNE for a reduced number of neighbors, while UMAP proved to be a better choice for a larger number of neighbors. We can observe as well that the overall value of Trustworthiness increased for USDA Food Composition Dataset, probably due to its considerable size.

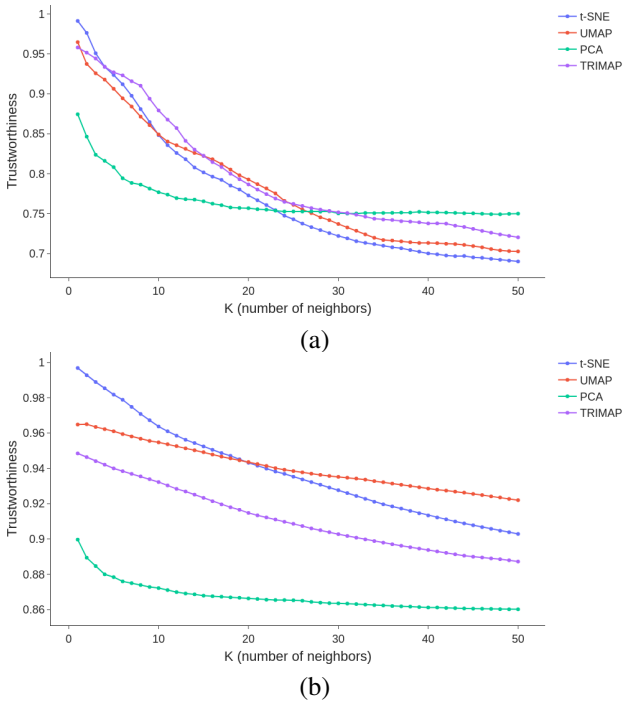


Fig. 9. Neighborhood preservation based on trustworthiness: (a) Vegan Food Dataset; (b) USDA Food and Nutrient Dataset.

B. Domain Expert Evaluation and Feedback

Much research in the literature considers expert reviews to assess the usefulness of visual analytics systems [31], [32]. Likewise, we conducted a pilot interview with three domain experts, separately, holding a PhD in Nutrition, and currently serving as professors and researchers at the University of Brasília (UnB). Before the interviews, each participant received a detailed tutorial to familiarize them with the purpose of our tool, its workflow, and the datasets to be used in the analysis. At the beginning of each interview, we dedicated a few minutes to provide a brief tutorial to address any remaining questions.

Each interview lasted approximately 30 minutes and was conducted via online video conference. The participants accessed the tool remotely through Streamlit's Community Cloud [4] platform. During the sessions, the experts were asked to perform Tasks 1-4, as described earlier in Section III, with the interviewer ensuring that all instructions were clear. After that, we opened space for participants so that they could give the tool their overall evaluation, impressions, criticisms, or suggestions.

[4] Streamlit's Community Cloud

1) Observations: To keep the analysis consistent, each participant selected the same food components for analysis: Energy, Protein, Carbohydrate, Total Sugar, and Total Fat. The ingredients were set for hover information, and the point placement visualization used PCA, as it was the most intuitive technique for some participants.

For Task 1, which involved the identification of similar products, participants were given the freedom to explore the point placement visualization independently. They found the process of identifying and comparing similar food items to be both intuitive and straightforward. One participant noted that products within the "Protein Foods" category tended to group closely due to similar values for protein and carbohydrates, but some outliers, such as canned beans, appeared more distant due to lower protein content. Another participant was satisfied to observe similarities between products from different categories that would not be easily seen without this type of visualization. Lastly, the final participant pointed out how the PCA visualization showed that lower nutritional value items were positioned to the left, with more caloric products located towards the right, corresponding to categories like "Fats and Oils, Snacks and Sweets, and Sugars" being situated at the extremes.

For Tasks 2 and 3, participants focused on analyzing the "Protein Foods" category. One participant mentioned that by selecting specific ranges in the second visualization, they could observe some patterns such as products in the range of 200 and 300 kcal containing between 10 and 30 grams of protein and low carbohydrates, while more caloric products in the range of 600 to 700 tended to have the same amount of protein although higher fat content. Some initially found it challenging to analyze the data due to its large size (867), but the tool's interactivity was highly helpful. Regarding the third visualization, all of them found it pretty intuitive to understand the connections between the ingredients, but could not discover many insights due to the intrinsic nature of the dataset. They also suggested that while the USDA dataset was large and generic, the tool would become significantly more useful when applied to more focused datasets. One even mentioned that the graph-based visualization would be useful in their research to analyze the formulation of ingredients in a specific category and verify which formulation is most common.

For Task 4, which involved labeling and coloring products, the feedback varied. One participant appreciated the concept of reducing annotation time, although they felt that the process was still somewhat manual and could benefit from further automation. Another participant found the functionality useful for categorizing products more quickly after conducting a preliminary analysis of the dataset. However, a third participant was unsure how this feature could apply to their current research but recognized its potential utility in other fields.

Throughout the interviews, the participants provided several additional comments and suggestions. All participants found PCA more intuitive and easier to interpret compared to other techniques. They also inquired whether the tool could support analysis in Portuguese, as it is their first language. Some suggested adding customizable color schemes to adapt the tool for specific purposes, such as aligning it with some journal format required for publication. Furthermore, the option to switch between light and dark modes, a feature offered by Streamlit, was well-received. Several participants suggested improving ingredient selection by allowing users to filter for key ingredients, such as "chicken" or "pork" rather than selecting each specific part (e.g., legs, wings). They also suggested adding a feature that shows which points fall within the selected filter and mentioned the potential benefit of including standard deviation when visualizing average component values, as there could be significant variations in certain categories. However, the overall feedback was very positive and they expect to utilize this tool soon in their activities.

VII. Discussion

The results obtained and the feedback from experts showed that the *FoodVis* tool can assist nutrition specialists in visually exploring food composition data. The visualizations provided intuitive insights into food composition data, enabling the comparison of products based on their nutrients and ingredients. However, some limitations were observed, and potential improvements were suggested for future work.

For the point placement visualization, even though PCA did not present the best metric results, the experts found this dimension reduction technique to be particularly intuitive, noting that it offered a clear and understandable representation of the data. As for the parallel coordinates view, the experts suggested adding a cross-filtering option to allow them to track the points selected within the corresponding range, providing more transparency in the analysis process. Additionally, experts suggested including standard deviation when visualizing average component values. This enhancement would allow for a more nuanced view, revealing any significant variations within specific categories that could otherwise be overlooked.

For the graph-based visualization, some challenges arose due to the overlapping of nodes in the layout. Adding new preprocessing steps, such as tokenization or stopword removal, could minimize this issue by reducing the number of nodes displayed. Furthermore, participants suggested improving the ingredient selection process by allowing users to filter for key ingredients, such as "chicken" or "pork" rather than requiring them to select each specific part (e.g., legs, wings).

Finally, regarding the feedback provided from the task of labeling and coloring products, a promising feature to be included in future work is incorporating machine learning techniques to classify products based on specific prompts. For instance, the tool could classify products as "vegetarian" or "non-vegetarian" based on their ingredients, or group products according to sugar content, which would greatly enhance the tool's ability to address specific analytical needs.

VIII. Conclusion

Analyzing food composition data is not a straightforward task, but it can provide new insights into nutritional and ingredient content. Our web-based tool uses a structured pipeline to assist specialists for this purpose, starting from dataset upload, preprocessing, and the generation of three visualizations, each of them addressing specific tasks: point

placement visualizations to assist identifying similarities between food products; parallel coordinates to visualize nutrient distribution patterns; and graph-based visualizations to reveal relationship between data - in our case, ingredient co-occurrences.

Both qualitative and quantitative experiments were conducted to evaluate the quality of the visualizations. The evaluation of multidimensional projections revealed better cluster separability for TriMap technique, while PCA and UMAP presented better preservation of original neighbors depending on the chosen dataset. Domain expert interviews were conducted to complement the evaluation step and assess the tool's usability and utility, providing additional areas for improvement.

Future work shall address the suggestions and comments from the experts, and incorporate machine learning techniques for product classification, aiming to further extend the tool's usefulness in nutrition science. Moreover, it would be valuable to conduct further user experiments with participants unfamiliar with the tool, to obtain impartial feedback from a more diverse audience, including those not directly involved with the tool's development, and improve its overall accessibility and usefulness.

IX. Acknowledgments

The authors of this paper would like to thank ProIC/UnB, *Fundação de Apoio a Pesquisa do Distrito Federal* (FAPDF), process number 00193-00001288/2021-16, and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), process number 143060/2023-6 for the grant that supported this research.

References

[1] S. Bo, M. Fadda, D. Fedele, M. Pellegrini, E. Ghigo, and N. Pellegrini, "A critical review on the role of food and nutrition in the energy balance," *Nutrients*, vol. 12, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2072-6643/12/4/1161

[2] D. S. Kelley and A. Bendich, "Essential nutrients and immunologic functions," *The American Journal of Clinical Nutrition*, vol. 63, no. 6, pp. 994S–996S, 06 1996.

[3] M. J. Moreira, J. García-Díez, J. M. M. M. de Almeida, and C. Saraiva, "Evaluation of food labelling usefulness for consumers," *International Journal of Consumer Studies*, vol. 43, no. 4, pp. 327–334, 2019.

[4] K. D. Lewis and B. M. Burton-Freeman, "The role of innovation and technology in meeting individual nutritional needs1,2," *The Journal of Nutrition*, vol. 140, no. 2, pp. 426S–436S, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022316622069863

[5] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[6] T. Eftimov, G. Popovski, E. Valenčič, and B. K. Seljak, "Foodex2vec: New foods' representation for advanced food data analysis," *Food and Chemical Toxicology*, vol. 138, p. 111169, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278691520300570

[7] A. B. Garcia, "The use of data mining techniques to discover knowledge from animal and food data: Examples related to the cattle industry," *Trends in Food Science & Technology*, vol. 29, no. 2, pp. 151–157, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924224412002129

[8] P. Mursanto, A. Wibisono, P. Fahira *et al.*, "In-tfk: a scalable traditional food knowledge platform, a new traditional food dataset, platform, and multiprocess inference service," *Journal of Big Data*, vol. 10, no. 47, 2023. [Online]. Available: https://doi.org/10.1186/s40537-023-00728-1

[9] C. Maçãs, E. Polisciuc, and P. Machado, "Atovis – a visualisation tool for the detection of financial fraud," *Information Visualization*, vol. 21, no. 4, pp. 371–392, 2022. [Online]. Available: https://doi.org/10.1177/14738716221098074

[10] A. B. Carmeli, L. Meloney, and B. E. Bierer, "Data visualization explorer: A tool for participant representation in pivotal trials of fda-approved medicinal products," *Patterns (New York, N.Y.)*, vol. 4, no. 5, pp. 100 713–100 713, 2023.

[11] H. Zhang, P. Wang, X. Gao, Y. Qi, and H. Gao, "Out-of-sample data visualization using bi-kernel t-sne," *Information Visualization*, vol. 20, no. 1, pp. 20–34, 2021. [Online]. Available: https://doi.org/10.1177/1473871620978209

[12] I. T. Jolliffe, "Principal component analysis and factor analysis," *Principal Component Analysis*, pp. 150–166, 1986.

[13] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[15] S. Tilouche, V. P. Nia, and S. Bassetto, "Parallel coordinate order for high-dimensional data," 2019.

[16] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 3, pp. 564–575, 2008.

[17] D. Machida and Y. Sugiura, "Relationships among local agricultural product purchases, self-cooked meal consumption, and healthy eating habits: A cross-sectional study in a town in gunma, japan," *Healthcare*, vol. 10, no. 8, 2022. [Online]. Available: https://www.mdpi.com/2227-9032/10/8/1510

[18] J. Sobal, H. L. Muncie, C. M. Valente, B. R. DeForge, and D. Levine, "Physicians' beliefs about vitamin supplements and a balanced diet," *Journal of Nutrition Education*, vol. 19, no. 4, pp. 181–185, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022318287801875

[19] X. Dai, L. Wu, and W. Hu, "Nutritional quality and consumer health perception of online delivery food in the context of china," *BMC Public Health*, vol. 22, no. 1, p. 2132, Nov 2022. [Online]. Available: https://doi.org/10.1186/s12889-022-14593-9

[20] F. V. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1229–1236, 2008.

[21] J. R. Townsend, T. O. Kirby, T. M. Marshall, D. D. Church, A. R. Jajtner, and R. Esposito, "Foundational nutrition: Implications for human health," *Nutrients*, vol. 15, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2072-6643/15/13/2837

[22] R. Diestel, *Graph Theory: 6th edition*. Springer (print edition); Reinhard Diestel (eBooks), 2024. [Online]. Available: https://books.google.com.br/books?id=52UTEQAAQBAJ

[23] C. Deval, E. S. Brooks, M. Dobre, R. Lew, P. R. Robichaud, A. Fowler, J. Boll, Z. M. Easton, and A. S. Collick, "Pi-vat: A web-based visualization tool for decision support using spatially complex water quality model outputs," *Journal of Hydrology*, vol. 607, p. 127529, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169422001044

[24] Å. Rinnan, L. Nørgaard, F. van den Berg, J. Thygesen, R. Bro, and S. B. Engelsen, "Data pre-processing," *Infrared spectroscopy for food quality analysis and control*, vol. 2009, pp. 29–50, 2009.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[26] E. Amid and M. K. Warmuth, "TriMap: Large-scale Dimensionality Reduction Using Triplets," *arXiv preprint arXiv:1910.00204*, 2019.

[27] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, pp. 1–73, 2021.

[28] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, "Graph-based measures to assist user assessment of multidimensional projections," *Neurocomputing*, vol. 150, pp. 583–598, 2015.

[29] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data," in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1345–1354.

[30] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 384–391.

[31] M. R. Islam, S. Akter, L. Islam, I. Razzak, X. Wang, and G. Xu, "Strategies for evaluating visual analytics systems: A systematic review and new perspectives," *Information Visualization*, vol. 23, no. 1, pp. 84–101, 2024.

[32] L. E. Resck, J. R. Ponciano, L. G. Nonato, and J. Poco, "Legalvis: Exploring and inferring precedent citations in legal documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 3105–3120, 2023.