BACHELOR

Towards a greener plate Optimizing food choices to minimize climate impact

Boot, Thomas H.M.

Award date:
2023

Awarding institution:
Tilburg University

Link to publication

Towards a greener plate
Optimizing food choices to minimize climate impact

Final bachelor project data science

Thomas H.M. Boot

Tue: 1604635
UvT: 2065721

Supervisor:
dr. J.C. Wagenaar

Final version

Tilburg, June 17, 2023

Department of
Mathematics & Computer Science

Department of
TiSEM

1 Abstract

Food waste is a global issue with significant economic, environmental, and social consequences. This research focuses on finding alternative food product compositions with similar nutritional values but reduced climatic impact, addressing food waste and environmental concerns. The study combines three essential elements: emissions, nutrition, and food waste, to guide customers in making more sustainable choices.

The research explores two approaches to reducing climatic impact. The first approach involves clustering techniques to identify alternative product options with lower emissions and similar nutritional profiles. Clustering will be applied for the one-to-one replacements. The second approach utilizes linear optimization to find optimal compositions of food products that minimize climatic impact while maintaining nutritional balance. Linear optimization will be applied for the one-to-many and many-to-many replacements.

On average, the clustering methods reduce around 1.4 kilograms of CO_2 per kilogram of the replaced product, while linear optimization has an average decrease of approximately four kilograms per replacement.

Keywords: *Clustering, CO_2 reduction, Food waste, linear optimization*

Contents

2 Introduction

Food waste is defined as food that is fit for consumption, but is consciously discarded at the consumption or retail stages [1]. Food that is damaged during transport and thus is thrown away should also be considered food waste.

Globally around one-third of the product food ends up as food waste[2]. The European Union (EU) produces approximately 88 million tons of food annually. 20% of the generated food in the EU ends up as food waste [3]. This is a loss of around 143 billion dollars [4].

The fact that 20% of the European-produced food ends up as food waste contrasts even more with the estimated 33 million European people who can not have a sufficient qualitative and quantitative meal every other day [5].

Food waste does not only highlight the distribution issue of food in Europe, but environmentally harmful resources that are used during production are wasted as well. Primary food production necessitates using resources such as energy, land, water, and raw materials, all of which have economic and environmental consequences. The greenhouse emissions, such as CO_2, during production and the other resources that are used during the production, are absolutely seen as not more harmful. They are relatively seen as more harmful since there is no consumption pleasure, and the waste is the same.

When food is wasted, most of the time, it is not used for compost or another functional purpose, but it almost always ends up in landfills. In the USA, around 95% of all discarded food ends up in landfills. At 21%, it is the greatest component of municipal solid waste [1]. 5% of waste from landfills is diverted for composting. CO_2, a powerful greenhouse gas contributing to global warming, is one of the greenhouse gasses produced when food waste decomposes [1].

Food waste, not taking into account damaged food, is produced throughout the whole production chain. Households account for 53% of the food waste; production and processing for 30%; retailers for 5%; and food services such as restaurants and cafeterias for 12% [6]. So, the biggest improvements in food waste reduction can be made at the household level.

Research on household food waste has found that fruits and vegetables are mostly thrown away, followed by prepared meals, bread, meat, milk, and packaged items. Surprisingly, 37% of the households said they threw out "none" or "hardly any" food in all six categories [7]. If it is true that 37% almost does not throw anything out, then the remaining 63% of the waste-creating households do throw even more away. However, it could be the case that the 37% group is unaware of their food waste behavior.

Given the substantial quantity of food waste at the household level, it should be clear that it is most likely to make a significant reduction at the household level. When it comes to strategies to decrease food waste, it appears that the majority of Europeans point to individual responsibility. 63% of the respondents of the survey agree that improved food-related activities in terms of planning and purchasing will help to reduce waste [8].

In summary, 20% of the European-produced food ends up as food waste. Most food waste is created at the household level, and 63% of people find it useful to get guided to reduce food waste[8].

This research will combine the climate impact, nutritional diet, and likelihood of ending up as food waste so that customers get suggestions for alternative products with less environmental impact. This can, for example, be used in online supermarkets. For example, when a customer has product A in its basket, and product B is less emitting and nutritionally relatively close. In that situation, the customer can be notified whether they want to replace product A with the less emitting product B. It could be the case that a product will be one-to-one replaced, but it could also be that one product

gets a suggestion to be replaced by three different products. The customers will get these suggestions so that they can make a considered choice regarding their impact on the environment.
The purpose of this research is to find out if there is an alternative composition for a bag of products with similar nutritional values and in balance with the recommended amount, but with less climatic impact.
The first step that needs to be done is to merge the three essential elements: emission, nutrition, and food waste together. After that, the reduction of climatic impact can be done.
The reduction of climatic impact will be done with two approaches. The first approach for finding an alternative composition for a bag of products is using different clustering techniques. The Second approach to finding a solution for the research problem will be done with linear optimization.

In the Netherlands, multiple guidance labels exist for customers to make more informed food choices. Food products must have the ingredients, the nutritional score, and the latest consumption date or best-before date on them [9]. There are more requirements for food packaging, but that is irrelevant for this research.

There are already multiple publications in the field of diet optimization to reduce CO_2 emission and food waste. Take, for example, the research of Van Dooren et al. [10]. This research combined a nutritional diet, climate impact, and financials to find the optimal diet.
Another research in this field is the research of Janssens et al. [11], who researched the impact of consumers' behavior on waste in daily food provisioning.
The research that will be presented in this paper will work further on the research of Van Dooren et al. [10] and the research of Janssens et al. [11].

3 Methodology

In this chapter, firstly, the used datasets will be reviewed. After that, the preparation of the datasets will be considered. This contains the preprocessing of the datasets and the merging of the datasets. Thirdly the dimensionality reduction methods are reviewed. Fourthly, the clustering that will be applied is discussed. Finally, the linear optimization method is discussed.

3.1 Datasets

In this research, primarily three datasets are used. The first dataset, the nutritional value dataset, contains the nutritional characteristics of the products that are considered in this research. The second dataset, the CO_2 dataset, contains data about how much CO_2 is produced during the cultivation of the products. The third dataset, the waste dataset, contains data regarding waste on consumer level.

3.1.1 Dataset 1: Nutritional values

The nutritional data that is used, is taken from the Dutch food composition database (NEVO). NEVO is a Dutch/Netherlands Food Information Resource (NethFIR) component. NethFIR includes multiple food databases, including food composition data for generic and branded foods with data about nutrients, allergens, and characteristics such as sustainability and portion sizes [12]. The Netherlands Nutrition Centre (NNC) and the RIVM collaborate on NethFIR. NEVO has extensive food composition data, particularly for generic products. RIVM and the NNC Institute manage the databases. The NEVO datasets are primarily for educational purposes [12].

The database contains data on the compositions of food that is frequently eaten by the Dutch population. These food products contribute significantly to the intake of nutrients for the Dutch population [13]. The NEVO database contains a dataset with more than 125 nutrient characteristics such as the amount of fat, amount of Magnesium, or amount of vitamin K for 2207 food products.

3.1.2 Dataset 2: Carbon dioxide during production

The SU-EATABLE LIFE (SEL) [14] database offers a compendium of carbon and water footprint values for 323 food items. The SEL dataset consists of 3349 extrapolated carbon footprint estimates from 841 articles published between 1998 and 2019, and 937 extrapolated water footprint values from 88 articles published between 2005 and 2018. The original data is summarized into 85 typologies, 11 sub-typologies, 323 pieces of data on carbon footprint, 72 typologies, and nine sub-typologies. Moreover, the dataset includes uncertainty and data quality assurance such as the Kurtosis value[15].

3.1.3 Dataset 3: Food waste

Data about how much waste households produce is provided by research from Wageningen University of Research (WUR) [16]. The dataset contains 150 records of how much households have thrown away. The data is from a questionnaire that participants had to fill in about what and how much they have thrown away last week.

For the questionnaire, the participants went through a four-step procedure. Firstly the participants had to go through an ethical/privacy check since the gathered data must be considered personal data. Secondly, the participants got an announcement to try to be aware of their food waste. This is done to get more reliable data since many people are not completely aware of food waste [16]. However, it should also be considered that people get motivated to change their behavior due to the awareness notification [17]. The third step is filling in the questionnaire about how much food is wasted in fifteen categories. The food has to be categorized into one of the following fifteen categories: beverages, beans, bread, candy, cereal, eggs, fruit, meat, pasta, potatoes, sauce, soup, topping for bread, vegetables, and yogurt. The quantification for each category is done in everyday units. For example, sugar is measured in the number of spoons, while meat is measured in portions. The fourth step is to recalculate everything to grams so that the dataset contains only grams per food group.

These four steps are already done, and therefore the dataset contains 150 responses about how much participants have thrown away divided into fifteen different categories measured in grams.

3.2 Data preparation

3.2.1 Preparing the datasets

After dropping redundant columns, the nutrition dataset has the following columns:

1. $Foodgroup$: which is one of 27 different food groups, such as mixed dishes, vegetables, legumes, and alcoholic beverages.
2. $NEVOcode$: a unique code representing the product.
3. $Foodname$: the name of the product in English.
4. $Synonym$: when there exists a synonym of the product name in Dutch.
5. $Quantity$: which is either 100ml or 100mg determining the quantity of the product.

Finally, the nutrition dataset has 133 columns with additional information about what nutrients are in the product.

The nutritional data is stored in different ways. There are columns saved as float numbers, integer numbers, a string with a dot indicating the decimal, and a string with a comma indicating a decimal. All these numbers are changed into float numbers.

The CO_2 dataset contains the following columns:

1. $Foodgroup$: which tells to which of the four food commodity groups the product belongs.
2. $Commodityfood$: a column with food commodity type indicating which of the 85 groups the product belongs to.
3. $Subcommodity$: a column with sub-typologies indicating an additional specification for the typologies. Indicating to which of the 85 sub-typologies the product belongs.
4. $Foodcommodityitem$: with the actual name of the product.
5. Column n: the number of times the product occurred in the researched papers.
6. $Meancolumn$: with the mean CO_2 score of all the CO_2 values of that product in the literature research.

The mean is chosen because some products occur only once in the literature study, while others occur several times, and the mean takes all the values into account.

The waste dataset contains four columns:

1. $Category$: with fifteen different category names.
2. $Absolutewaste$: with the absolute waste per week in kilograms.
3. $Absoluteintake$: with absolute intake per week in kilograms.
4. $Relativewaste$: with the relative waste per category per week.

The original waste dataset is 150 records of what people consciously have thrown away divided into fifteen categories. The mean is taken for each column of the original waste dataset, which will be used as the waste value for each category. This dataset contains the absolute waste per week in grams while the relative amount of waste is desired. The relative waste is desired so that products can get a waste probability. This probability indicates the likelihood of a product ending up as food waste.

To get the relative waste data, the consumption data from the average Dutch person is used [18]. With absolute waste and absolute consumption, the relative waste is calculated. The relative food waste will later be used to create a waste column that takes into account the waste probability.

3.2.2 Connecting the datasets

Connections are made between the three datasets to give each record in the nutrition dataset the most appropriate CO_2 value and waste probability. The data comes from three different sources with no identification column to connect easily. So, to give each nutrition record the most appropriate CO_2 value, three methods, TF-IDF, Levensthein distance, and spaCy, are tested to see which method is the best solution for the connection of the different data sources.

Firstly a match is made with the highest TF-IDF score. Term frequency-inverse document frequency (TF-IDF) is one of the most common vectorizer methods in natural language processing [19]. It determines how important a term is relative to the whole corpus. This will result in a vector that is unique for each word. TF-IDF vectorizes a word by calculating the term frequency (TF) and multiplying that with the inversed document frequency (IDF) [19]. The term frequency counts the number of times the word occurs in the document. The inverse of the document frequency is the inverse of how many documents contain a particular word. These papers [20] [21] give more in-depth information about TF-IDF.

The results are checked manually to check the performance of the TF-IDF method since there is no artificial method to evaluate it. The TF-IDF method did not give satisfying results. This is because TF-IDF is not exactly suited for matching strings. Therefore, another approach is needed.

The Levensthein distance is an algorithm to determine how similar two strings are. The higher the Levensthein score, the more different the two strings are [22]. The Levenstein distance is calculated by going through each string on a character base and checking whether the exact or next character is identical. Based on that, the Levensthein score will increase or stay the same. These [23] [24] [25] papers will give additional information about the Levensthein distance.

The Levensthein should give better results from a theoretical point of view since the Levensthein distance is more appropriate for this problem than the TF-IDF distance. Furthermore, empirically based, the Levensthein distance gives better results than TF-IDF, but there were still major mistakes in the results. The performance of the different models on 50 randomly selected products can be seen in the appendix in figure 15. The method still has major mistakes, and therefore improvements in the approach are still needed.

One improvement is to add guidance. To guide the model in the right direction, the 85 categories from the CO_2 dataset are manually divided into the 27 categories from the nutrition dataset. With this manual guidance, a sub-dataset is created for each nutrition product, and in that sub-dataset, the Levensthein distance is calculated for each of the records. When the Levensthein algorithm does not find any match at all, the spaCy library is used to find a match.
SpaCy is a freely available open-source library that contains algorithms and linguistic data that can be used for Natural language processing [26]. SpaCy is built for processes that require text understanding since it can be used for information extraction. This tool can be particularly useful when products have to be classified. The SpaCy library supports more than 72 languages and has 80 pre-trained pipelines for 24 different languages [27]. These papers or books[26] [28] [29] give additional information about spaCy and the usage of spaCy.

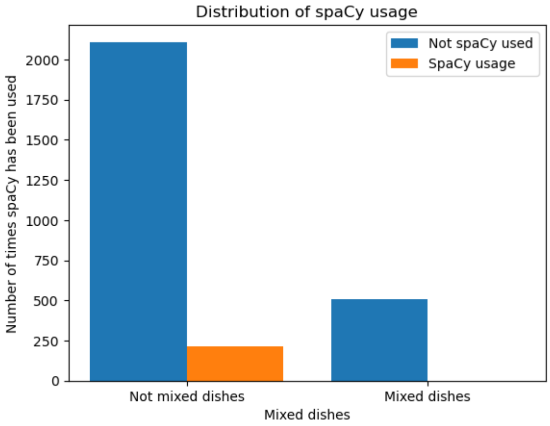


Figure 1: SpaCy usage

As shown in figure 1, spaCy is not used that much compared to the Levensthein approach. In total, spaCy is used 214 times for making a connection in not-mixed dishes, and spaCy has been used once for making a connection in mixed dishes.

With the guided approach, the model is manually guided in the right direction while being able to select all the CO_2 values. After the guidance, the model can still select the correct CO_2 value, and with all the different sub-datasets, all the different CO_2 values are available beforehand. After running the model, there are 60 different CO_2 values, which is not close to the ideal 323 different CO_2 values.

As shown in Figure 2, the distribution is not normally distributed over the values. Still, with the guided approach method, the CO_2 data density is higher than when it is done manually, and therefore this method is kept.

The waste dataset has fifteen rather general categories, and the nutrition dataset has 27 categories that do not match one on one. To connect the datasets properly, spaCy is used to find matches between the nutrition product names and the waste categories. The distribution of the nutrition products can be seen in figure 3.

As shown in figure 3, not all categories contain the same number of products. This can be caused by

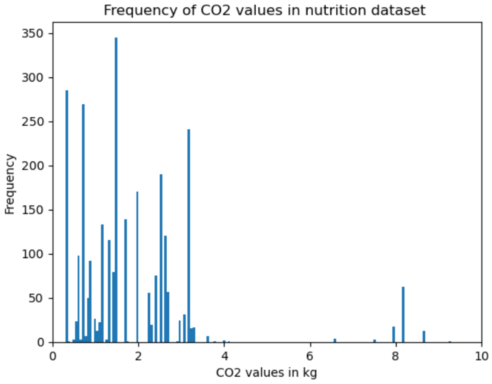


Figure 2: Distribution of CO_2 values

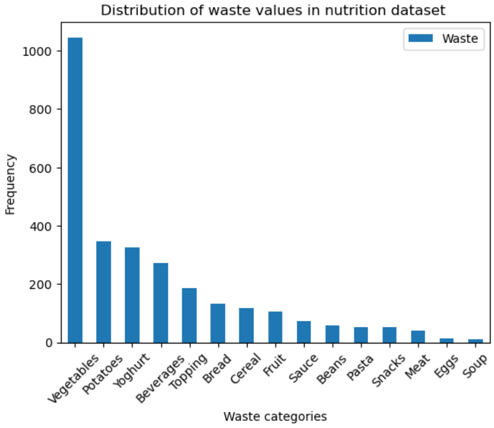


Figure 3: Distribution of waste values

misclassification or by the structure of the original data. Here it is a combination of both. There are misclassifications; take, for example, the highest peak. In the nutrition dataset, two categories can be classified as vegetables: vegetables and legumes. The Vegetable category has 253 records, and the legumes category has 36 records. This together does not come close to the over 3000 classifications with spaCy. On the other hand, it is unlikely that over 100 different products are classified as soup in the data.

Furthermore, a CO_2 column that takes into account the waste is made. This column contains a CO_2 value plus the CO_2 values multiplied by the waste probability. The waste probability is calculated by dividing the absolute waste per week by the absolute consumption of a category per week. From now on, when reference is made to the CO_2 value, the corrected CO_2 value is referred to.

Finally, the nutrition columns that contain only zero values are dropped because there is no useful data in these columns.

The final dataset has 2832 records and 132 columns containing necessary product information, the nutrition values, the CO_2 that is needed to produce one kilogram of the product, and the percentage of what is thrown away per food product.

3.3 Dimension reduction

The goal of dimension reduction is to reduce the complexity of the data while keeping as much information as possible in the data. This is done to reduce running time and to make the model as interpretable as possible for customers. This is both desired because it is desired

The dimension of a dataset can be reduced by either creating new features that combine existing features or keeping only the columns that contain the most information [30].

Both reduction methods and additional approaches are used in this project. The method that creates new features will be used to reduce the dimensionality of the data for clustering. The second method, which keeps the original values, will be used to reduce the dimensionality for linear optimization.

This section will explain three different dimensionality reduction approaches and the advantages and disadvantages of each method.

One of the most common methods when new features are created is principal component analy-

sis (PCA). PCA is an unsupervised machine-learning algorithm to reduce the dimensionality of the dataset [31], which still contains the necessary dataset information. Firstly the covariance matrix is made to identify correlations between columns. Secondly, the eigenvectors and eigenvalues are used to identify the principal components. Lastly, a feature vector is made to decide which principal components have the most variance and should be kept [32]. In conclusion, PCA trades a bit of data accuracy for simpler data. These papers [31] [33] [32] give more in-depth information about principle component analysis.
PCA is used on the 124 columns that contain nutrition data. The first eight principle components already explain 99.12% of the variance of the whole dataset. The percentage of what each of the first ten principal components explains can be seen in figure 4.

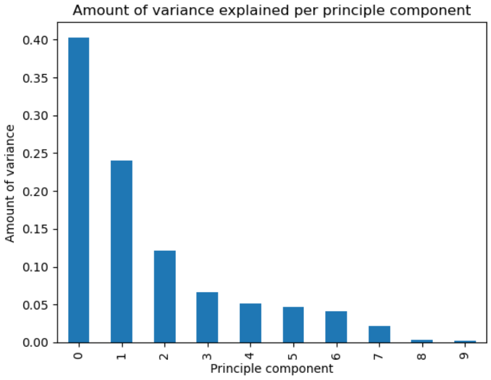


Figure 4: The ten principal components that explain the most variance

The remaining 114 principal components explain 0.88% of the variance in the data.
Two major disadvantages of PCA are that it creates new data instead of keeping the original data and that the results are hard to interpret. The creation of new values is no issue for clustering. Still, PCA is not suited for linear optimization since the linear optimization constraints are not interpretable when constraints are made for the principal components. So, PCA will only be applied to data for clustering.

The other dimension reduction method only keeps the data's most important feature columns. The data is not modified with this approach. Thus, this method can be applied before linear optimization. To achieve this, the $SelectKBest$ from sklearn is used. The $SelectKBest$ selects the most important features based on univariate statistical tests [34]. The ten most important columns are selected to keep enough information in the data while keeping it interpretable for customers. $SelectKBest$ is applied to the dataset, and the ten most important columns are: Calcium, Vitamin K2, five types of acids saturated total, one Fatty acid trans-cis, and one fatty acid unidentified. These columns are apparently important for the explainability of the CO_2 distribution, but the components are not significantly present in most of the products. However, these columns are not easily distinguishable for customers. A third dimensionality reduction method is applied to keep it as explainable as possible for customers.

The third and final method that is applied to reduce the data is based on the presence of a component in the product. Examples of components in products are: Fat, Sugar, and Water. The ten components that are most prominent in the products are selected. The sum of each component is taken, and the ten products that have the highest sums are selected. These products are Protein total, Protein animal, Sugar, Fatty acids total, Starch total, Fat total, Carbohydrate available, Water, Energy in

Kj, and energy in kcal. A correlation matrix is plotted to check if certain columns describe the same phenomenon, as shown in figure 5.

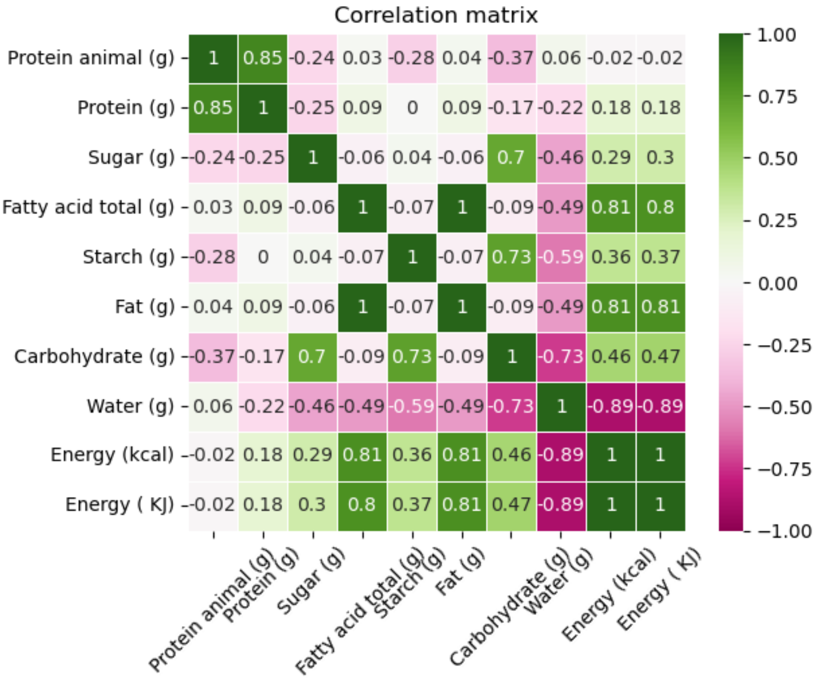


Figure 5: Correlation matrix of the ten columns

As it is known, animal-based protein is a subset of the whole category of proteins. Therefore, the animal-based category is dropped from the ten most relevant components group. As shown in figure 5, both energy columns display the same phenomenon but in a different quantity. Therefore only the energy in kcal is kept. Finally, fatty acid total is a subgroup of fats. The correlation between both columns is 1. Therefore fatty acids total is dropped, and the reduced dataset contains only seven columns that indicate the nutritional characteristics of the product. So the reduced dataset contains the following seven columns to describe the product: protein, sugar, starch, fat, carbohydrate, water, and energy.

In conclusion, the PCA method is used to reduce the dimensionality that is used in the clustering approach. This method is a better dimensionality reduction method, but the results are hard to interpret. Furthermore, the constraints for linear optimization are hard to make on the principal components. $SelectKBest$ is applied to the data to reduce the dimensionality while keeping the original columns, but this approach is not easily useable for customers and, therefore, not used.
The dataset for clustering contains eight principal components describing the products. Keeping the most present components is the method that is used for dimensionality reduction for linear optimization. After dropping the correlating columns, the dataset for linear optimization contains seven columns that describe the products.

3.4 Clustering

The idea of CO_2 reduction with clustering is to replace products with products that are nutritionally similar but less polluting. Clustering will be applied on the nutritional columns that describe the products to create clusters with products that are similar nutritional-wise. For example, when a product is selected from cluster 23, the product in that cluster with the lowest CO_2 value is suggested. The CO_2 value is not taken into account when clustering.

Three different clustering approaches are applied. Firstly the affinity propagation method, which is able to self-identify the number of clusters in the data. This method is applied to the original 124 data columns. Secondly, the K-means approach to the 124-column data. Thirdly, the K-means approach is applied to the PCA-reduced data.

One disadvantage of clustering is the running time when the replacement is not one-to-one. All three clustering methods replace one product for one product. When clustering is applied to many-to-one, one-to-many, and many-to-many replacements, the running time will be n^2.

Firstly the affinity propagation method. This clustering method aims to identify clusters and can self-determine the number of clusters present in the data [35]. It is a so-called exemplar-based clustering method [36]. The method is derived from a standard inference method, and several experiments have shown that it performs consistently [36], but research found that this method can lead to suboptimal clustering solutions [35]. However, this method is able to self-identify the number of clusters in the data. This literature [36] [37] [38] gives additional information about affinity propagation clustering.

The affinity propagation clustering has been applied to the data. It is tried to get the maximum number of clusters so that the replacements are nutritionally as close as possible to the original point. The damping parameter, which indicates the extent of the previous value being maintained compared to the new value, influences the number of clusters. The damping variable is optimized to get the maximum number of clusters.

With the damping variable corresponding to the highest number of clusters, 189 clusters are made, and the nutrition products are divided over the 189 clusters. The distribution over the clusters is displayed in figure 6.

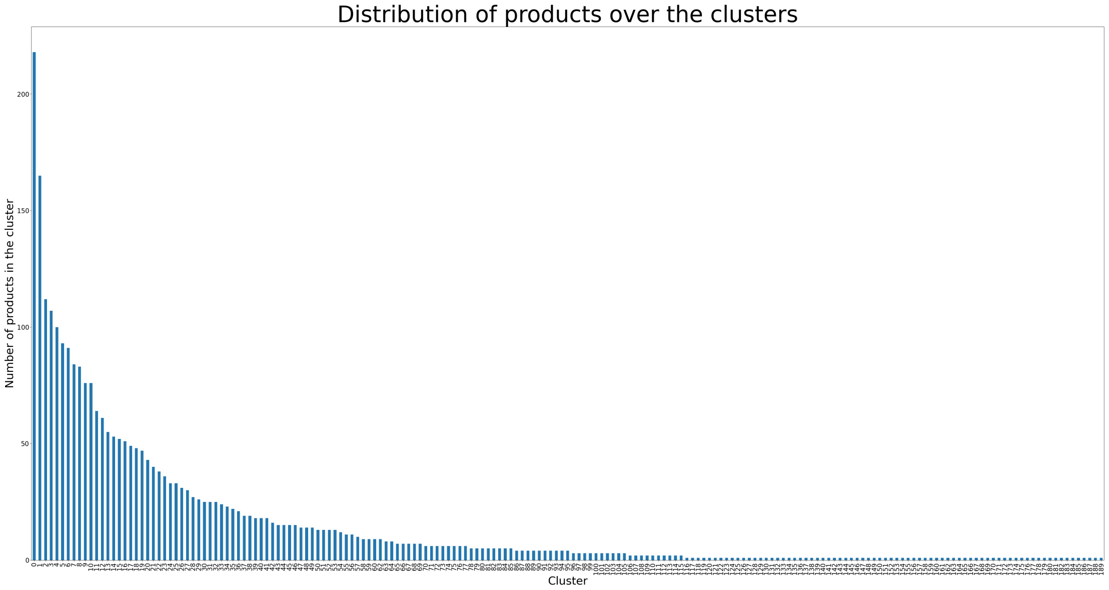


Figure 6: Distribution of products over the Affiliation clusters

As can be seen, the products are not evenly distributed over the clusters. The optimal distribution over the clusters would be that each cluster has a few assigned products, and the number of assigned products should be as equal as possible over the clusters. Currently, in the biggest cluster, over 200 products will have the same alternative. When the products are evenly distributed over the clusters, the replacements will be more diverse, given that each product has the same likelihood of being replaced.
To achieve this, K-means clustering is applied to the data. K-means is one of the most straightforward clustering algorithms. For each point in the data, it calculates the squared distance to the centroids. After that, the centroids are updated again to the middle of each cluster, and this process continues until it stabilizes or after a number of runs [39]. One major drawback of the K-means clustering algorithm is that the performance relies heavily on the initial randomly generated centroids [40]. This problem can be solved by running the algorithm over the same data multiple times. These papers [39] [41] [42] provide more information about K-means.
There are 2832 products in the dataset, and clusters of four points are created. Therefore 708 clusters are made, and each cluster contains four data points. The cluster size of four is selected to create equal cluster sizes while keeping them small. 1416 clusters also meet the requirements, but with 1416 clusters, 50% of the products are not improbable with the clusters, while only 25% of the products can not be improvable.

3.5 Linear optimization

Linear optimization is the second approach to suggest alternative products given a bag of products. One advantage of linear optimization is its running time, which is at most n^2 when replacing one product with a bag of products and vice versa.
Linear optimization, also called linear programming, is the approach of obtaining the best result by maximizing or minimizing in a mathematical model whose requirements are represented by linear

relationships [43].
So, for example, the function that needs to be minimized is the following function: $f(x_1, x_2, ..., x_n) = k_1x_1 + k_2x_2 + ... + k_nx_n$ with these constraints: $a_11x_1 + a_12x_2 \leq w_1$, $a_21x_1 + a_22x_2 \leq w_2$ and with the following non-negative variables:$x_1 \geq 0$, $x_2 \geq 0$, $x_n \geq 0$ [44].
To apply this concept to reduce the CO_2 emission of the bag of products, the goal is to minimize the CO_2 variable. The restrictions are based on the other nutrition values of the product in such a way that the suggestion should have a similar composition or it should be healthier. The linear optimization method has the following restrictions:

- **Protein:** The protein in the suggestion should have at least as many proteins as the original product. Research from Biomed Central has found that the satiety after a 60% meal was significantly higher than after a meal with 19% proteins [45], which reduces the need for unnecessary overconsumption.
- **Sugar:** The suggestion should have at most as much sugar as the original product. The WHO strongly recommended reducing the sugar consumption of both children and adults[46].
- **Starch:** Starch has a lower and upper bound, so the suggestion should contain at least 80% and at most 120% compared to the original product. This is because starch is a crucial part of a healthy diet; however, it can cause health issues if it is consumed too much [47].
- **Fat:** Fat has an upper bound, so the suggestion should contain less fat than the original product. There are several reasons why a low-fat diet is preferred. Diets heavy in fat have a weak satiating impact [48]. Furthermore, low-fat diets reduce the chances of coronary heart diseases [48].
- **Carbohydrate:** Carbohydrate has a lower and upper bound, so the suggestion should contain at least 80% and at most 120% compared to the original product. This is because carbohydrate is an important part of a healthy diet because it provides glucose to the body [49]. However, too much carbohydrate increases insulin resistance, which is adverse for type 2 diabetes [50].
- **Water:** Water has a lower and upper bound, so the suggestion should contain at least 80% and at most 120% compared to the original product. Water is essential for life and should therefore not be reduced too much, but when customers do not select a drink, a drink should not be the suggestion.
- **Energy:** Energy has a lower and upper bound, so the suggestion should contain at least 80% and at most 120% compared to the original product. Energy is essential for life, but a high-density diet is linked with an increased chance of Alzheimer's disease [51].

Given P, the set of all products, given original products $OP \subset P$, which is a subset of P, and given replacement products $RP \subset P$ which is a subset of P. The mathematical objective is:

- $min \sum_{f \in RP} CO_2[f] * x[f]$

Where $x[f]$ is the decision variable representing the quantity of food item f to be selected, each $x[f]$ is a non-negative integer indicating the amount of food item f to include in the optimal solution.
$CO_2[f]$ is the CO_2 emissions coefficient for food item f. It represents the environmental impact associated with producing one unit of food item f.

With constraints:

- Protein constraint: $\sum_{f \in RP} Protein_f * x_{ft} \geq Protein_t \forall t \in OP$
- Sugar constraint: $\sum_{f \in RP} Sugar_f * x_{ft} \leq Sugar_t \forall t \in OP$

- Starch constraint: $\sum_{f \in RP} Starch_f * x_{ft} \geq 0.8 * Starch_t \forall t \in OP \wedge \sum_{f \in RP} Starch_f * x_{ft} \leq 1.2 * Starch_t \forall t \in OP$
- Fat constraint: $\sum_{f \in RP} Fat_f * x_{ft} \leq Fat_t \forall t \in OP$
- CHO constraint: $\sum_{f \in RP} CHO_f * x_{ft} \geq 0.8 * CHO_t \forall t \in OP \wedge \sum_{f \in RP} CHO_f * x_{ft} \leq 1.2 * CHO_t \forall t \in OP$
- Water constraint: $\sum_{f \in RP} Water_f * x_{ft} \geq 0.8 * Water_t \forall t \in OP \wedge \sum_{f \in RP} Water_f * x_{ft} \leq 1.2 * Water_t \forall t \in OP$
- Energy constraint: $\sum_{f \in RP} Energy_f * x_{ft} \geq 0.8 * Energy_t \forall t \in OP \wedge \sum_{f \in RP} Energy_f * x_{ft} \leq 1.2 * Energy_t \forall t \in OP$

3.5.1 Data modification

There are two additional data modifications applied to the preprocessed data to prepare the data for linear optimization. Firstly a new dataset, a filler dataset, is created with the fifty least polluting products. Secondly, the original and the filler dataset are expanded for better results.

The filler dataset is created with the fifty least emissive products from the whole dataset. The fifty least polluting products are separated into a new dataset to see what the influence of the proportion of these fillers is in the suggestions. This will be used to see the influence on, for example, the CO_2 reduction when 10% instead of 70% of the product replacements are fillers. One disadvantage of this approach is that, for example, for herbs, 10% is already a lot. Examples from the filler dataset are: lettuce raw, and parsnip raw.

Both datasets are expanded to ten times the original size. Each record in the original dataset is present in the new dataset in 10%, 20%, till 100% of the original product. Both datasets are extended to prevent linear optimization from picking proportions that are less than 10% of the product. So, this dataset contains ten times more records than the original dataset.

During the expansion of the dataset, a new column is created, which indicates the partion of the original product. For example, it is the second part of the product, which is 20%. All the records are multiplied by 0.2, and the newly formed column will get 0.2 as a value.

With expanded data, linear optimization can be run with integers instead of percentages. This will result in partitions of products that are at least 10% of the original product instead of fractions of products.

For example, the record brandy in the original dataset is one record with its nutritional characteristics. In the extended dataset, there are ten records with brandy. The first record is brandy||0.1 with ten percent of each of the nutritional characteristics. The second record for brandy is brandy||0.2 with twenty percent of each of the nutritional characteristics. This continues until brandy||1 with the original characteristics of brandy.

The dataset without the fillers now has 21580 records, while the filler dataset has 500 records.

3.5.2 Research setup

For the linear optimization approach, three items are researched. Firstly, the influence of fillers in the suggestion is researched with a restriction of at least $m\%$ fillers. m has the values 0, 0.1 until 1 representing 0%, 10%, till 100%. Secondly, the influence of fillers in the suggestion is researched with a restriction of at most $u\%$. u has the values 0, 0.1 until 1 representing 0%, 10%, till 100%. Thirdly the influence of the number of products in a bag on CO_2 reduction is researched.

When one product is replaced using linear optimization, the linear optimization should be within the restrictions. When researching the influence of the fillers, the optimization should also take into

account at least $m\%$ of fillers or, at most, $u\%$ of fillers. For the research with at least boundary, the following constraint is added: $\sum_{f \in RP} fraction_f * x_{ft} \geq m \; \forall t \in OP$. Where m indicates the fraction of products in RP that should at least be from the same category. When the at-most boundary is needed for the experiment, the following constraint is added: $\sum_{f \in RP} fraction_f * x_{ft} \leq u \; \forall t \in OP$ Where u indicates the fraction of products in RP that should at most be from the same category.
A running time threshold is added to prevent long running times. The maximum running time is 60 seconds per optimization.
When a customer selects, for example, meat, it is assumed that the customer wants something that is at least similar to meat and not something completely different. When a product is selected from vegetables, a new dataset is made with all the products that belong to the vegetable group and the filler group so that there are products in the dataset with a similar taste and products that potentially have the highest impact on the CO_2 reduction. To keep track of the percentage of fillers in the suggestions, a new column is added with a decimal indicating the size of the product. For example, the row with medlar 10% has a 0.1 in the column and the medlar 50% has a 0.5 in the corresponding column. Another column with the same principle as the filler category is made for the products in the same category.

4 Results

To find the best alternatives for different products, multiple methods are applied. Firstly three different clustering methods are applied and evaluated.
The three clustering methods do a one-to-one replacement. This means that one product is replaced by one other product. The first two parts of linear optimization, the part with an at-least and an at-most boundary, do a one-to-many replacement. This means that one product can be replaced by multiple products. The third part of linear optimization does a many-to-many replacement. The products in the bag can be replaced by multiple products.
Finally, the best cluster method is compared to the linear optimization method.

4.1 Clustering

There are three different clustering methods applied to the data. Firstly, the Affinity propagation clustering method does not require a number of clusters. The number of clusters is automatically determined based on the data. Secondly, K-means is applied with 708 clusters. Finally, K-means is applied to the first eight principal components of the data. The methods are evaluated by calculating the Euclidean distance between the original and suggested products in the grid. The grid is based on the nutritional characteristics of the products. Each nutritional characteristic is one dimension in the grid. The Euclidean distance is plotted against the difference in CO_2 between the two products. As can be seen in the figures, the x-axis represents the nutritional similarity between the original and the alternative product. The y-axis represents the pollution difference between the products. The optimal replacement is when the products are nutritionally as similar as possible and, from a pollution point of view as different as possible.

The scatterplot, as shown in figure 7, shows the distribution of the suggestions based on the affinity propagation clustering method. The relative difference in CO_2 is visualized in figure 16, which is in the appendix. Figure 7 shows that most of the alternatives reduce at-most five kg CO_2. There are a few points with a distance smaller than 50, which means that the products are nutritionally seen close to each other. Most suggestions have a distance smaller than 800. More about the distribution of the reduction can be seen in table 1.

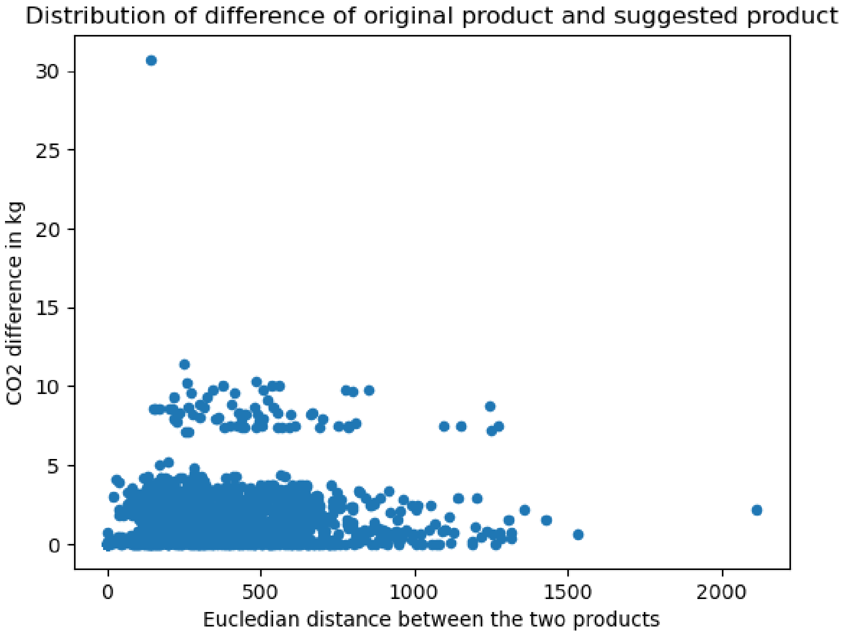


Figure 7: Affinity propagation method

Secondly, the performance of the suggestions that are made with K-means on the original data can be seen in figure 8. The relative difference in CO_2 is visualized in figure 17, which is in the appendix.

Figure 8 shows that the distance between the suggested points and the original is closer than the affinity clustering method. This is because the clusters are smaller. Most of the CO_2 reduction is fewer than five kg. This is approximately the same compared to the affinity method. More statistics about the distribution of the reduction can be found in table 1.

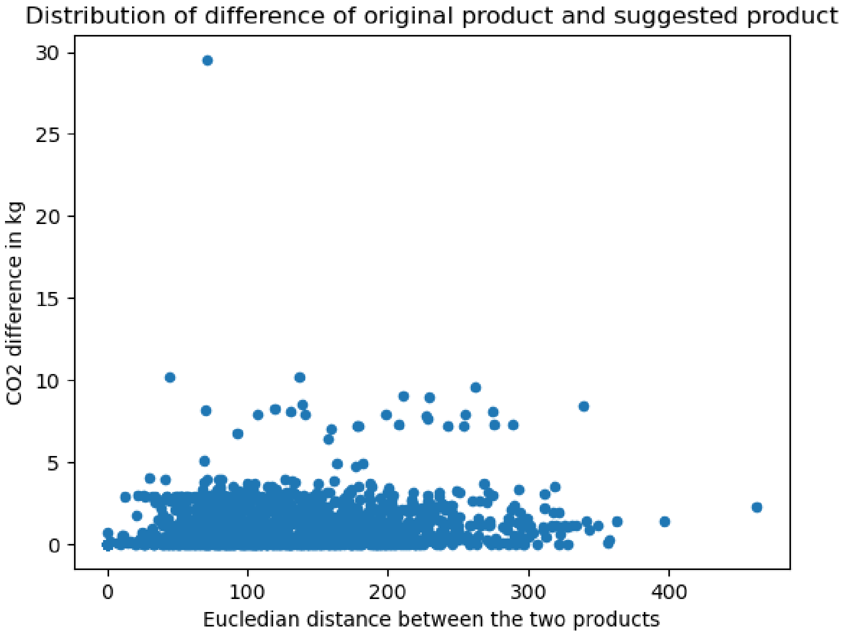


Figure 8: K-means method on original data

Thirdly, the distribution of how the clustering performs on data that is created with the first eight principal components of the data. The scatterplot can be seen in figure 9. The relative difference in CO_2 is visualized in figure 18, which is in the appendix. The calculated distance for this method is the distance between the PCA points and not the original points. Therefore, the distance results can not be compared one on one with the two other clustering methods. More statistics about the distribution can be found in table 1.

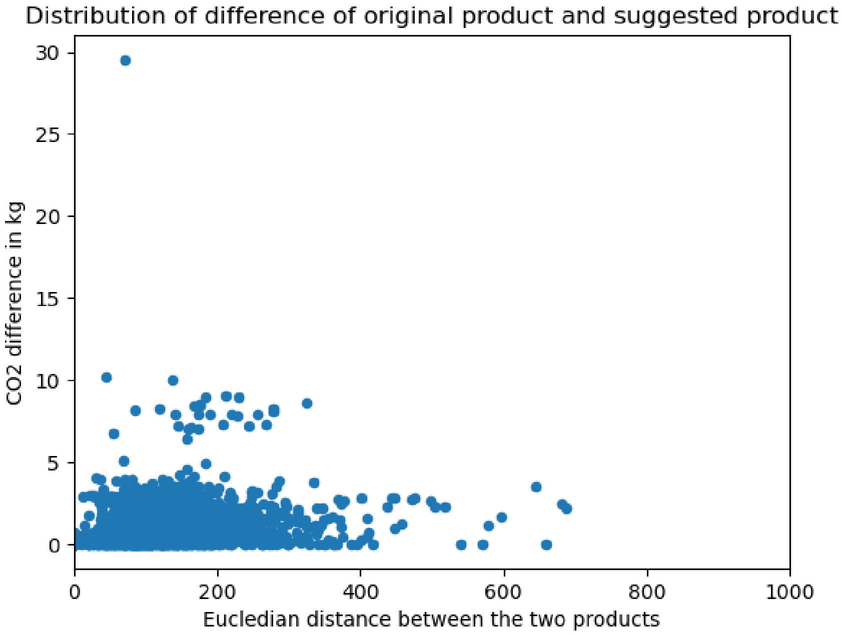


Figure 9: K-means method on the first eight principle components

When comparing figure 7, figure 8, and figure 9, it is seen that the different approaches have different performances. This is also visible in table 1.

What can be seen from the figures is that there is a bunch of points at (0,0). This is due to each least emitting product of each cluster. For that item, there is no better alternative. Thus, the distance between the original and the alternative is 0. So, when the product that is already the least emitting product in the cluster is selected to select the least emitting product in its group, it selects itself, or both products have the same CO_2 value, which is the lowest of that cluster. Therefore the distance between the two points and the CO_2 reduction is zero.

The first method, the affinity method, has been compared to the two K-means with the least points at (0,0) because that model has fewer clusters than the two K-means methods, which both have 708 clusters.

The affinity method reduces, on average, the most CO_2 per alternative. However, the mean Euclidean distance between the original product and the suggested alternative is also the biggest compared to the two k-mean approaches. So, the trade-off between the distance and the CO_2 reduction is not optimal.

The suggested nutritional characteristics from the K-means approach with the original data are the closest to the original based on the nutritional characteristics of the product. On the other hand, this method, on average, reduces the least CO_2 per alternative.

The K-means PCA approach and the affinity approach both have a few outliers in the cluster data. This influences the mean. However, it shows that the model does not perform as constant as the K-means on the original data.This can be seen both in the responsible figures and in table 1. This should be kept in mind while comparing the different approaches.

Considering the mean values of each list, the k-mean performed on the original data still performs better. Still, the difference between K-means original and K-means PCA is smaller.

	Affinity propagation		K-means original data		K-means PCA data	
	Euclidean distance	CO_2 difference	Euclidean distance	CO_2 difference	Euclidean distance	CO_2 difference
Mean	336.92	1.60	82.87	0.78	105.46	1.17
Std	234.52	1.77	86.35	1.39	128.57	1.48
Min	0.0	0.0	0.0	0.0	0.0	0.0
25%	187.53	0.40	0.0	0.0	1.53	0.0
50%	290.67	1.24	83.62	0.16	95.21	0.86
75%	447.67	2.38	153.87	1.15	159.35	1.76
Max	2116.57	30.70	462.80	29.53	3318.14	29.39

Table 1: Table with descriptive statistics about the distributionof the clustering performance

In conclusion, the three different approaches are capable of reducing CO_2 while using products that are as closely as possible related based on nutrition. As can be seen in table 1, the Affinity propagation method has the highest mean reduction compared to the other methods. However, the mean Euclidean distance is the highest as well for affinity propagation. On average, to reduce one kg of CO_2 with this method, the alternative product is at a distance of 174.4.
The K-means method on the original data has the lowest mean Euclidean distance but has the lowest CO_2 reduction compared to the other two methods. On average, to reduce one kg of CO_2 with this method, the product is at a distance of 71.4.
The K-means on the PCA data has a mean Euclidean distance of approximately 1.5 times the mean distance of K-means on the original data. In contrast, the CO_2 reduction is approximately 1.25 times higher. On average, to reduce one kg of CO_2 with this method, the product is at a distance of 90.1.
When the highest CO_2 reduction is needed, the best method is the affinity propagation method. The downside of this method is that the alternatives are the furthest away on average compared with the other methods. So, when the best trade-off is desired, under the assumption that the trade-off trend is linear, the K-means method on original data should be used.

4.2 Linear optimization

This section contains the results regarding the linear optimization approach to reduce the CO_2 emissions by replacing food products with less emitting products. The first part will contain results when one product is replaced, and the second part will contain the results when fewer emitting products replace a bag of products. It should be kept in mind that the number of replacements is not considered.

4.2.1 One product

One big consideration when replacing one product with products that either belong to the same category or are considered fillers is the balance between the two groups of products. Two different approaches are taken to see the influence of the balance between both groups. Firstly, with at-least x% of products that are in the same category, and secondly, with at-most y% fillers in the replacement. To check the performance of each concept, fifty products are randomly selected, and the CO_2 minimization is run for each of the fifty products.
The decrease of CO_2 per replacement percentage within the same category can be seen in figure 10. While running the code, not every minimization problem converted within sixty seconds. For the eleven box plots in figure 10, the number of non-convergence within sixty seconds is displayed in the following list: [2, 2, 2, 8, 10, 13, 15, 19, 24, 29, 33]. Each number of the list indicates the number of non-convergences within 60 seconds. So, the first number two, corresponds with two non-convergences when 0% of the product should be replaced within the same category. The higher the replacement percentage within the same group, the higher the number of non-convergence. This can be seen in the list. It should be kept in mind that when the number of non-convergences increases, the statistical significance of the results decreases.
As it can be seen in figure 10, the model performs the best when 0% should be replaced within the same category. This is reasonable because the product can be fully replaced with filler products. The performance decreases when the percentage of replacements within the same category increases until 80%. But 80% of the replacements within the same category, already almost 50% of the runs did not convert within 60 seconds.

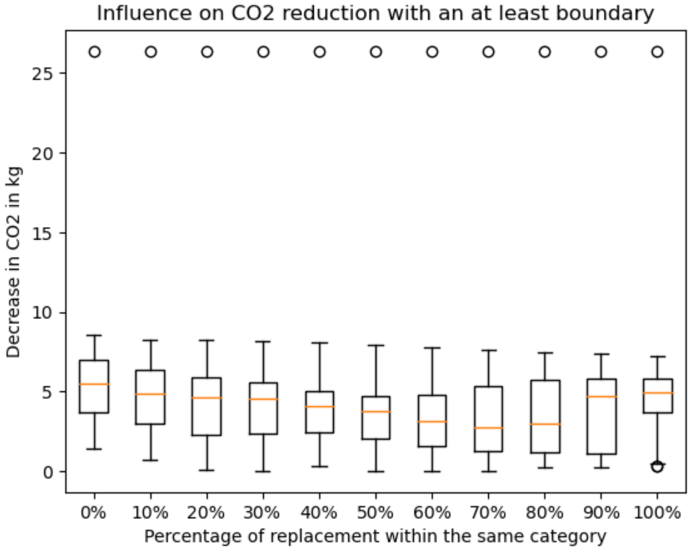


Figure 10: Distribution with at-least boundary

When working with an at-most boundary instead of an at-least boundary, more products converge within sixty seconds. 2 out of the 50 products do not converge for all the percentages. These two products are identical for all the cases. It should be remembered that the eleven different situations return similar values for all the products except for one product. The distribution is shown in figure 11. As can be seen, the at-most boundary does not influence the results significantly. This should be investigated further.

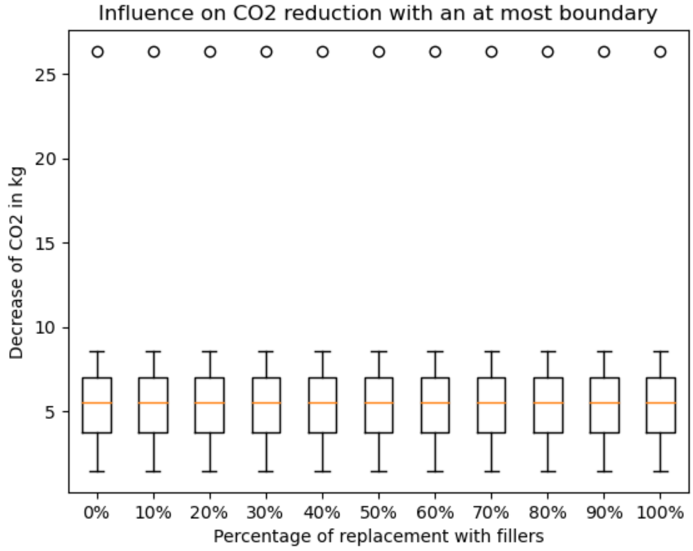


Figure 11: Distribution with an at-most boundary

4.2.2 Bag of products

The size of the bags is investigated for replacing a bag of products with a bag of new products. For a bag of products, replacing products within the same category is no longer possible since a bag of products can contain products from multiple categories. Because of that, only the bag's size is considered. The model is run on fifty bags with a size of two, fifty bags with a size of three, fifty bags with a size of four, and finally, fifty bags with a size of five. The composition of the bags is generated randomly. All the 200 bags did convert within sixty seconds except for one bag in the group of bags with bag size two. This can be explained since when there are more products in a bag, there is more space to find optimal solutions. The results can be seen in figure 12.

The size of the bags impacts the amount of CO_2 that can be reduced by replacing products. As shown in figure 12, the mean difference between bags with size two is approximately three, while the mean reduction on bags of size is approximately nine. This makes sense because, in bigger bags, there is more space to fit products in.

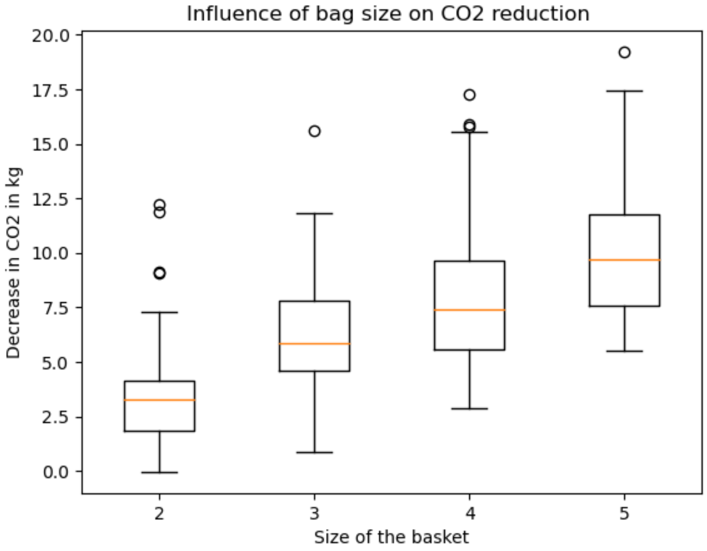


Figure 12: Distribution of CO2 reduction per bag size

4.3 Comparing the methods

In this last result section, the best-performing clustering method, the K-means with original data, and the linear optimization method are compared to each other. Even though the two methods are fundamentally incomparable, the comparison might give useful insights. Clustering performs a one-to-one replacement, while linear optimization performs a one-to-many replacement.

To compare both methods, the most polluting fifty products are replaced using the K-means without PCA. The fifty most polluting products are used as well for the evaluation of linear optimization method.

The performance of the best clustering method is visualized in figure 13. The mean reduction is 3.81 kg CO_2, and the mean distance is 123.36. The distribution of this plot has similarities compared to the distribution when the whole dataset is processed. Still, the cluster that has around nine kilograms of CO_2 reduction is relatively seen more represented in this subset.

The clustering method has similar mean CO_2 reductions compared to the replacements that are done using linear optimization. The results are similar to the results with the at-least boundary of 70% and higher.

Thus, the clustering approach has a similar mean CO_2 reduction level compared to the linear optimization approach.

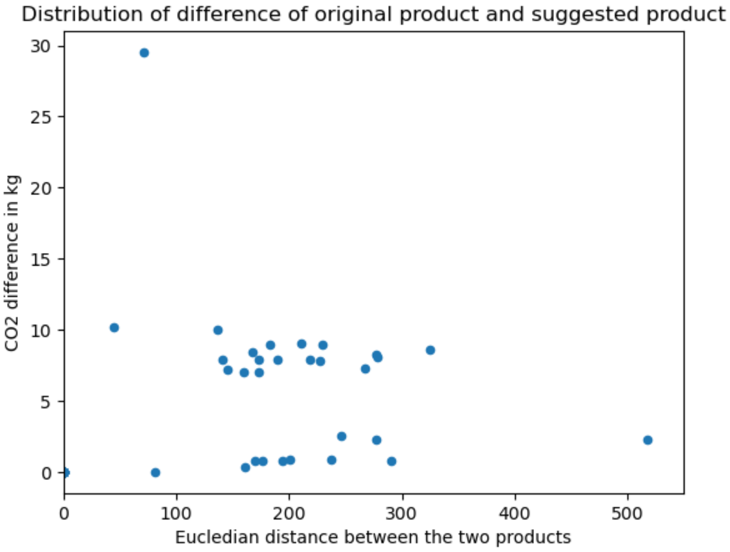


Figure 13: Fifty most polluting product replacements with clustering

5 Discussion

There are multiple considerations and assumptions that should be kept in mind while analyzing the results. The first part of the discussion contains considerations regarding the data and the connection between the data sources. The second part contains considerations regarding the clustering methods. The third part contains considerations regarding linear optimization. The fourth and final part contains general considerations about the idea as a whole.

5.1 Data

There are mainly three different issues regarding the data and its quality. The first issue is regarding the quality and dimensionality of the data. The second remark concerns the quality of the connections from the different datasets. The third and final remark on the data is creating a new data frame for linear optimization.

The CO_2 data is based on literature research and, for many products, the mean of several CO_2 values is taken. The production of certain products is not equally emissive throughout the year. Take the production of tomatoes during the summer and winter. The CO_2 pollution varies between 0.1 kg CO_2 during the summer and 10.2 kg CO_2 per kilogram of tomatoes during the winter [52]. The seasonal trend is not taken into account in this research.
One additional aspect that should be considered is the product's origin during the year. Transport pollution is considered for these CO_2 values, which can vary significantly during the year. For the UK to import products from Morocco to the UK by boat, it pollutes on average 70 grams of CO_2 per kilogram of food. When Moroccan products are imported by car, a CO_2 value of approximately 350 grams per kilogram should be considered. Local products pollute less than 10 grams CO_2 per kilogram for transport [53]. This data dimensionality is not available in the data, but it could potentially impact the results significantly. The variation of origin is not taken into account either in this research.
The waste data is based on a survey about what people consciously throw away. However, waste disposal can be considered an unconscious, habitual process that happens without much thought [54]. Even though the participants of the waste survey were encouraged to be very alert about their waste disposal, the unawareness phenomenon should be taken into account while processing the data.
The issues regarding data can be solved by using more detailed data.

The waste dataset categorized the waste into fifteen different waste probabilities while other datasets divided their products into more categories. This causes issues with the connection of the different datasets. The best and final method that is used to connect the different datasets still has some wrongly connected products. This can be seen in appendix 8.1. These mismatches between nutritional, waste, and CO_2 products severely limit the methods' performance.
This issue can be solved by using better connection methods like pre-trained neural networks. With well performing pre-trained neural network, there will be fewer misclassifications.

The new dataset for linear optimization is created to solve the issue of requiring fractions of products. One of the disadvantages of linear optimization is that the coefficients are either continuous or integers. Generally, it is not desired to require a lot of fractions of products, but this dataset is made to at least get slices of products. Now the model can select a part of the product that is not a fraction of the product the part of the product is either 10%, 20% till 100% of the product. One additional remark that should be taken into account is that all the products are now separated into ten new records, while it can differ per product what partitions are desired.

5.2 Clustering

There is no straightforward method to evaluate the performance of the different clustering methods. As already explained, the current method calculates the difference in CO_2 and measures the Euclidean distance between both points in the grid. This method relies on one important assumption which assumes that each dimension is equally important for the calculation of the Euclidean distance between the two data points. However, this is not the case because nutrients that occur in small proportions are measured in μg while more common nutrients are measured in grams, and the energy is measured in either kcal or kJ. The current measurement for performance is the best option, but this consideration should be kept in mind.

5.3 Linear optimization

Three considerations should be kept in mind while analyzing the results of the linear optimization approach. The first consideration is regarding the quality of the low CO_2 dataset. The second consideration is regarding the constraints that are made for linear optimization. The third and final consideration is regarding the limitation in the running time of the linear optimization approach.

The first consideration that should be kept in mind while interpreting the results from the linear optimization is the quality and accuracy of the fifty products with the lowest CO_2 values. This dataset contains one product that is labeled as fruit, eleven products that are labeled as legumes, and 38 product that are labeled as mixed dishes. Examples of products that are labeled as mixed dishes in the low CO_2 dataset are: babi pangang without rice, lasagna bolognese, pizza with frozen fish, and soup clear with meat. These four products have a corrected CO_2 label of 0.342 kg CO_2 per kilogram of the product. 0.342 kg CO_2 per kilogram of the product is one of the lowest ratings for CO_2 in the original CO_2 dataset. Products in the original CO_2 dataset with a CO_2 rating of 0.342 are: Soy meal, banana, pear, carrot, potato, and barley.
When the mixed dishes are removed from the data, the results show a similar trend, but not identical. The results can be seen in figure 14. The number of optimizations that are not increased within the sixty seconds gap are: [1, 1, 1, 4, 6, 8, 10, 14, 23, 29, 34]. Each number of the list indicates the number of non-convergences within 60 seconds. The first number, number one, corresponds with one non-convergences when 0% of the product should be replaced within the same category. When the replacement percentage within the same category increases, the CO_2 reduction decreases. This trend continues until 80%, and at that point, the results become unreliable because almost 50% of the products are not replaced within sixty seconds. As shown in figure 14, the results have a similar trend when the mixed dishes are left out.

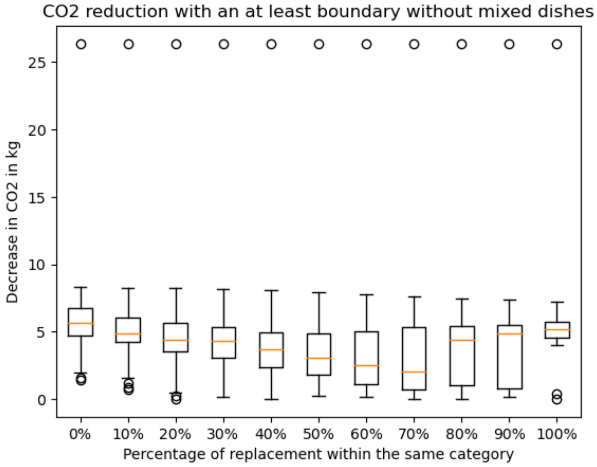


Figure 14: Distribution of CO_2 reduction when no mixed dishes are included

The second consideration that should be kept in mind is the constraints that are made for linear optimization. The current constraints are all separated per food characteristic. For example, given two products, one product that scores high on fat and low on sugar, and the second product that scores a fraction higher on sugar, but 80% lower on fat. Currently the second product is not recommended because there is a strict border to reduce sugar. The restrictions would be better if a new overall badness factor is introduced, which should also be minimized.

The third and final consideration that should be kept in mind is the limitation of the running time. This concept of replacing products or bags of products with less polluting products should be used during online shopping, so extreme running times are not desired. Therefore the running time is capped at sixty seconds. Sixty seconds is still quite high for real-life application but suited for experimenting. There is not always a CO_2 reduction achieved on the products. This is either due to the fact that the input was already optimal or due to the longer running time. The values without CO_2 reduction are not considered during the evaluation. So the results are biased toward the success results. The with no improvement are left out because it is not possible to distinguish whether they are not improving because of the running time or that they are already optimal.

5.4 General considerations

The most important reason why humans consume food is to obtain nutrition from the food for a healthy body. One often overlooked reason is that customers want to enjoy the food. Furthermore, a significant part of the weekly consumption is entirely consumed for pleasure. Products such as ice cream, alcohol, coffee, and chocolate are commonly consumed to promote a positive mental state [55]. Both methods that are presented in this work to reduce the CO_2 pollution leave the whole aspect of consuming food for pleasure out of sight. Customer satisfaction is likely to increase when the consumption for pleasure is taken into account. This can be applied when data is available about how products taste and their function.

The quality of the replacements will increase when this is taken into account. With that, there is another issue that should be taken into account, and that is that food taste preference is always closely linked to cultural development [56]. This means that every cultural background and subcultures have different preferences for replacements.

6 Conclusion

In short, the three data sources are brought together in the new dataframe. The new dataframe contains products with their respective nutrient characteristics, the waste probability, and the CO_2 emission for its production.

With the new dataframe, three different clustering methods are applied to reduce the CO_2 emission while suggesting products that are as closely related as possible. The best-performing clustering method is the K-means on the original data.

After that, linear optimization is applied to the dimensionality-reduced dataset. The linear optimization is applied with at least $m\%$ within the same category, and the linear optimization is applied with at most $u\%$ of filler products. Finally, the influence of the bag size on the CO_2 reduction is investigated.

The purpose of this research is to find out if there is an alternative composition for a bag of products with similar nutritional values and in balance with the recommended amount, but with less climatic impact. This paper shows that there is indeed an alternative composition for a bag of products with similar nutritional values, which is guided toward the recommended daily amount, with less climatic impact. This can be achieved by using the discussed clustering or linear optimization techniques.

One major issue which needs to be taken into account in this research is that people consume food not only for nutrients, but for pleasure as well. So, alternatives, where the taste is not taken into account, can be used, but it is not likely that customers will be satisfied with it.

To improve this conceptual design, the taste of the product should be taken into account. This can either be done with data about the taste of products, but a disadvantage is that taste perception differs per culture and per person. Another approach to tackle the issue of taste is to use data about how people have replaced products in the past.

7 References

References

[1] Harvard. *Food Waste, the big picture*. n.d. URL: https://www.hsph.harvard.edu/nutritionsource/sustainability/food-waste/ (visited on 02/27/2023).

[2] Esther Alvarez de Los Mozos, Fazleena Badurdeen, and Paul-Eric Dossou. "Sustainable consumption by reducing food waste: A review of the current state and directions for future research". In: *Procedia Manufacturing* 51 (2020), pp. 1791–1798.

[3] European Commission. *EU Food Loss and Waste Prevention Hub*. 2019. URL: https://ec.europa.eu/food/safety/food_waste/eu-food-loss-waste-prevention-hub/about (visited on 02/27/2023).

[4] Silvia Scherhaufer et al. "Environmental impacts of food waste in Europe". In: *Waste management* 77 (2018), pp. 98–113.

[5] Beatrice Garske et al. "Challenges of food waste governance: An assessment of European legislation on food waste and recommendations for improvement by economic instruments". In: *Land* 9.7 (2020), p. 231.

[6] Lauren G Block et al. "The squander sequence: Understanding food waste at each stage of the consumer decision-making process". In: *Journal of Public Policy & Marketing* 35.2 (2016), pp. 292–304.

[7] Roni A Neff, Marie L Spiker, and Patricia L Truant. "Wasted food: US consumers' reported awareness, attitudes, and behaviors". In: *PloS one* 10.6 (2015), e0127881.

[8] Directorate-General for Communication. *Food waste and date marking*. https://ec.europa.eu/commfrontoffice/publicopinion. DOI: 10.4232/1.12515.

[9] GOV. *Labelling of food*. 2019. URL: https://business.gov.nl/regulation/labelling-food/ (visited on 02/27/2023).

[10] Corné Van Dooren et al. "Combining low price, low climate impact and high nutritional value in one shopping basket through diet optimization by linear programming". In: *Sustainability* 7.9 (2015), pp. 12837–12855.

[11] Kim Janssens et al. "How consumer behavior in daily food provisioning affects food waste at household level in The Netherlands". In: *Foods* 8.10 (2019), p. 428.

[12] *NEVO-online 2021: background information*. Accessed: 13-3-2023.

[13] RIVM. *Dutch Food Composition Database (NEVO)*. 2021. URL: https://www.rivm.nl/en/dutch-food-composition-database (visited on 03/13/2023).

[14] *SU-EATABLE LIFE: a comprehensive database of carbon and water footprints of food commodities*. https://figshare.com/articles/dataset/SU-EATABLE_LIFE_a_comprehensive_database_of_carbon_and_water_footprints_of_food_commodities/13271111. DOI: https://doi.org/10.6084/m9.figshare.13271111.v2.

[15] Figshare. *SU-EATABLE LIFE: a comprehensive database of carbon and water footprints of food commodities*. 2021. URL: https://figshare.com/articles/dataset/SU-EATABLE_LIFE_a_comprehensive_database_of_carbon_and_water_footprints_of_food_commodities/13271111 (visited on 03/13/2023).

[16] Erica van Herpen et al. "A validated survey to measure household food waste". In: *MethodsX* 6 (2019), pp. 2767–2775.

[17] Erica Van Herpen et al. "Comparing wasted apples and oranges: An assessment of methods to measure household food waste". In: *Waste Management* 88 (2019), pp. 71–84.

[18] RIVM. *Voedingsmiddelen Consumptie*. 2023. URL: https://www.wateetnederland.nl/resultaten/voedingsmiddelen (visited on 03/30/2023).

[19] Fatih Karabiber. *TF-IDF — Term Frequency-Inverse Document Frequency*. 2023. URL: https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document%20frequency/ (visited on 03/29/2023).

[20] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[21] Thorsten Joachims. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Tech. rep. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[22] Ethan Nam. *Understanding the Levenshtein Distance Equation for Beginners*. 2019. URL: https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0 (visited on 03/29/2023).

[23] Rishin Haldar and Debajyoti Mukhopadhyay. "Levenshtein distance technique in dictionary lookup methods: An improved approach". In: *arXiv preprint arXiv:1101.1232* (2011).

[24] Li Yujian and Liu Bo. "A normalized Levenshtein distance metric". In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1091–1095.

[25] Gueddah Hicham. "Introduction of the weight edition errors in the Levenshtein distance". In: *arXiv preprint arXiv:1208.4503* (2012).

[26] Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020.

[27] spaCy. *spaCy 101: Everything you need to know*. 2023. URL: https://spacy.io/usage/spacy-101 (visited on 03/29/2023).

[28] Xavier Schmitt et al. "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate". In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE. 2019, pp. 338–343.

[29] Akhilesh Kumar Singh and Ananya Verma. "An efficient method for aspect based sentiment analysis using spacy and vader". In: *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE. 2021, pp. 130–135.

[30] Rukshan Pramoditha. *11 Dimensionality reduction techniques you should know in 2021*. 2021. URL: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b.

[31] G Thippa Reddy et al. "Analysis of dimensionality reduction techniques on big data". In: *Ieee Access* 8 (2020), pp. 54776–54788.

[32] Zakaria Jaadi. *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. 2023. URL: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (visited on 04/04/2023).

[33] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. "A review of principal component analysis algorithm for dimensionality reduction". In: *Journal of Soft Computing and Data Mining* 2.1 (2021), pp. 20–30.

[34] SKlearn. *Feature selection*. 2023. URL: https://scikit-learn.org/stable/modules/feature_selection.html.

[35] Fanhua Shang et al. "Fast affinity propagation clustering: A multilevel approach". In: *Pattern recognition* 45.1 (2012), pp. 474–486.

[36] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points". In: *science* 315.5814 (2007), pp. 972–976.

[37] Delbert Dueck. *Affinity propagation: clustering data by passing messages*. University of Toronto Toronto, ON, Canada, 2009.

[38] Kaijun Wang et al. "Adaptive affinity propagation clustering". In: *arXiv preprint arXiv:0805.1096* (2008).

[39] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. "The global k-means clustering algorithm". In: *Pattern recognition* 36.2 (2003), pp. 451–461.

[40] José M Pena, Jose Antonio Lozano, and Pedro Larranaga. "An empirical comparison of four initialization methods for the k-means algorithm". In: *Pattern recognition letters* 20.10 (1999), pp. 1027–1040.

[41] Youguo Li and Haiyan Wu. "A clustering method based on K-means algorithm". In: *Physics Procedia* 25 (2012), pp. 1104–1109.

[42] Kristina P Sinaga and Miin-Shen Yang. "Unsupervised K-means clustering algorithm". In: *IEEE access* 8 (2020), pp. 80716–80727.

[43] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Vol. 6. Athena scientific Belmont, MA, 1997.

[44] Wikipedia. *Linear programming*. 2023. URL: https://en.wikipedia.org/wiki/Linear_programming.

[45] Dominik H Pesta and Varman T Samuel. "A high-protein diet for reducing body fat: mechanisms and possible caveats". In: *Nutrition & metabolism* 11.1 (2014), pp. 1–8.

[46] Paula Moynihan et al. "Implications of WHO Guideline on Sugars for dental health professionals". In: *Community dentistry and oral epidemiology* 46.1 (2018), pp. 1–7.

[47] Christine Mikstas. *Foods High in Starch*. 2022. URL: https://www.webmd.com/diet/foods-high-in-starch (visited on 05/15/2023).

[48] Eric Jéquier and George A Bray. "Low-fat diets are preferred". In: *The American journal of medicine* 113.9 (2002), pp. 41–46.

[49] Harvard. *Carbohydrates*. 2023. URL: https://www.hsph.harvard.edu/nutritionsource/carbohydrates (visited on 05/15/2023).

[50] Asim K Duttaroy. *Evidence-Based Nutrition and Clinical Evidence of Bioactive Foods in Human Health and Disease*. Academic Press, 2021.

[51] Scott E Kanoski et al. "The effects of a high-energy diet on hippocampal function and blood-brain barrier integrity in the rat". In: *Journal of Alzheimer's Disease* 21.1 (2010), pp. 207–219.

[52] Georgios K Ntinas et al. "Carbon footprint and cumulative energy demand of greenhouse and open-field tomato cultivation systems under Southern and Central European climatic conditions". In: *Journal of cleaner production* 142 (2017), pp. 3617–3626.

[53] David Coley, Mark Howard, and Michael Winter. "Food miles: time for a re-think?" In: *British Food Journal* 113.7 (2011), pp. 919–934.

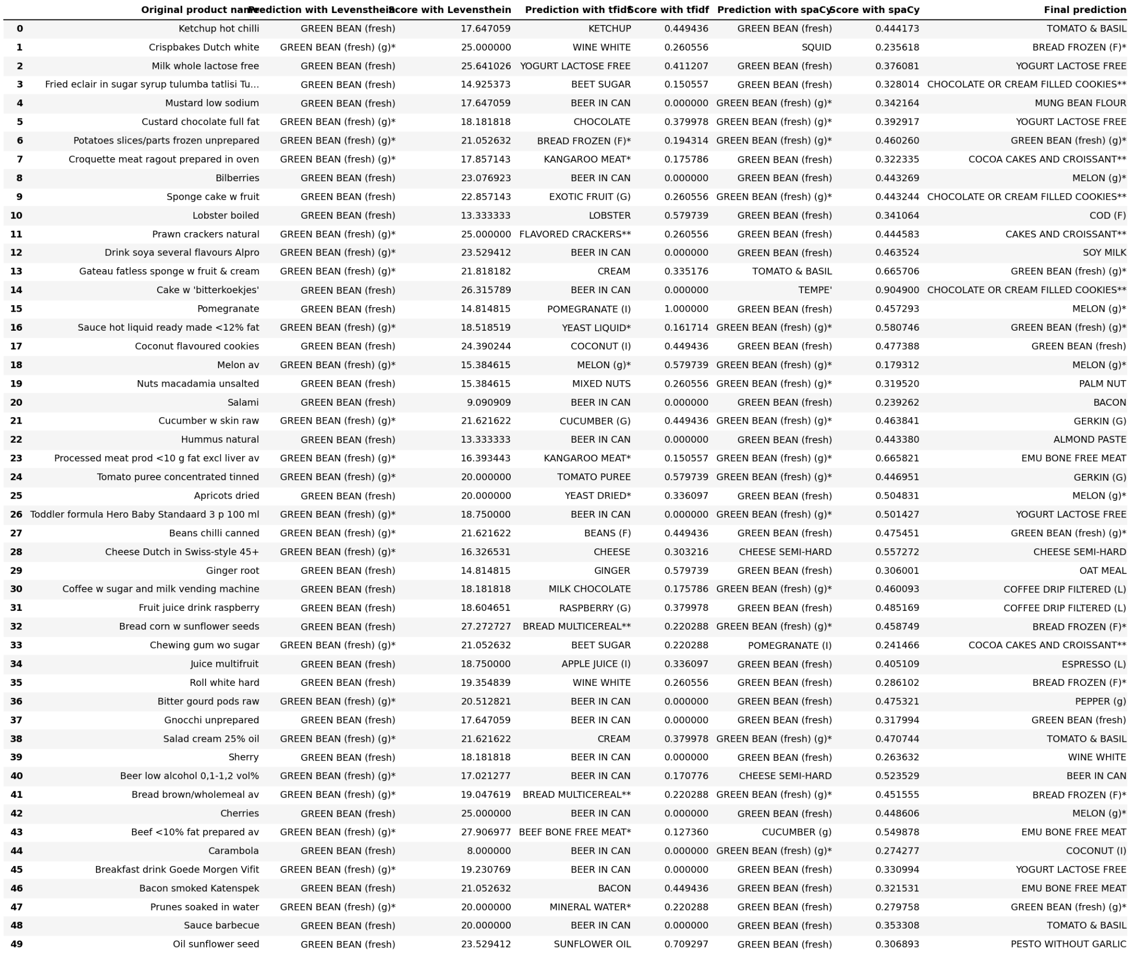
[54] Rob Comber and Anja Thieme. "Designing beyond habit: opening space for improved recycling and food waste behaviors through processes of persuasion, social influence and aversive affect". In: *Personal and ubiquitous computing* 17 (2013), pp. 1197–1210.

[55] Jane E Clark. "Taste and flavour: their importance in food choice and acceptance". In: *Proceedings of the nutrition society* 57.4 (1998), pp. 639–643.

[56] Len Tiu Wright, Clive Nancarrow, and Pamela MH Kwok. "Food taste preferences and cultural influences on consumption". In: *British food journal* 103.5 (2001), pp. 348–357.

8 Appendix

8.1 Performance comparison of different linguistic models



	Original product name	Prediction with Levensthein	Score with Levensthein	Prediction with tfidf	Score with tfidf	Prediction with spaCy	Score with spaCy	Final prediction
0	Ketchup hot chilli	GREEN BEAN (fresh)	17.647059	KETCHUP	0.449436	GREEN BEAN (fresh)	0.444173	TOMATO & BASIL
1	Crispbakes Dutch white	GREEN BEAN (fresh) (g)*	25.000000	WINE WHITE	0.260556	SQUID	0.235618	BREAD FROZEN (F)*
2	Milk whole lactose free	GREEN BEAN (fresh)	25.641026	YOGURT LACTOSE FREE	0.411207	GREEN BEAN (fresh)	0.376081	YOGURT LACTOSE FREE
3	Fried eclair in sugar syrup tulumba tatlisi Tu...	GREEN BEAN (fresh)	14.925373	BEET SUGAR	0.150557	GREEN BEAN (fresh)	0.328014	CHOCOLATE OR CREAM FILLED COOKIES**
4	Mustard low sodium	GREEN BEAN (fresh)	17.647059	BEER IN CAN	0.000000	GREEN BEAN (fresh) (g)*	0.342164	MUNG BEAN FLOUR
5	Custard chocolate full fat	GREEN BEAN (fresh) (g)*	18.181818	CHOCOLATE	0.379978	GREEN BEAN (fresh) (g)*	0.392917	YOGURT LACTOSE FREE
6	Potatoes slices/parts frozen unprepared	GREEN BEAN (fresh) (g)*	21.052632	BREAD FROZEN (F)*	0.194314	GREEN BEAN (fresh) (g)*	0.460260	GREEN BEAN (fresh) (g)*
7	Croquette meat ragout prepared in oven	GREEN BEAN (fresh) (g)*	17.857143	KANGAROO MEAT*	0.175786	GREEN BEAN (fresh)	0.322335	COCOA CAKES AND CROISSANT**
8	Bilberries	GREEN BEAN (fresh)	23.076923	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.443269	MELON (g)*
9	Sponge cake w fruit	GREEN BEAN (fresh)	22.857143	EXOTIC FRUIT (G)	0.260556	GREEN BEAN (fresh) (g)*	0.443244	CHOCOLATE OR CREAM FILLED COOKIES**
10	Lobster boiled	GREEN BEAN (fresh)	13.333333	LOBSTER	0.579739	GREEN BEAN (fresh)	0.341064	COD (F)
11	Prawn crackers natural	GREEN BEAN (fresh) (g)*	25.000000	FLAVORED CRACKERS**	0.260556	GREEN BEAN (fresh)	0.444583	CAKES AND CROISSANT**
12	Drink soya several flavours Alpro	GREEN BEAN (fresh) (g)*	23.529412	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.463524	SOY MILK
13	Gateau fatless sponge w fruit & cream	GREEN BEAN (fresh) (g)*	21.818182	CREAM	0.335176	TOMATO & BASIL	0.665706	GREEN BEAN (fresh) (g)*
14	Cake w 'bitterkoekjes'	GREEN BEAN (fresh)	26.315789	BEER IN CAN	0.000000	TEMPE'	0.904900	CHOCOLATE OR CREAM FILLED COOKIES**
15	Pomegranate	GREEN BEAN (fresh)	14.814815	POMEGRANATE (I)	1.000000	GREEN BEAN (fresh)	0.457293	MELON (g)*
16	Sauce hot liquid ready made <12% fat	GREEN BEAN (fresh) (g)*	18.518519	YEAST LIQUID*	0.161714	GREEN BEAN (fresh) (g)*	0.580746	GREEN BEAN (fresh) (g)*
17	Coconut flavoured cookies	GREEN BEAN (fresh)	24.390244	COCONUT (I)	0.449436	GREEN BEAN (fresh)	0.477388	GREEN BEAN (fresh)
18	Melon av	GREEN BEAN (fresh) (g)*	15.384615	MELON (g)*	0.579739	GREEN BEAN (fresh) (g)*	0.179312	MELON (g)*
19	Nuts macadamia unsalted	GREEN BEAN (fresh)	15.384615	MIXED NUTS	0.260556	GREEN BEAN (fresh) (g)*	0.319520	PALM NUT
20	Salami	GREEN BEAN (fresh)	9.090909	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.239262	BACON
21	Cucumber w skin raw	GREEN BEAN (fresh) (g)*	21.621622	CUCUMBER (G)	0.449436	GREEN BEAN (fresh) (g)*	0.463841	GERKIN (G)
22	Hummus natural	GREEN BEAN (fresh)	13.333333	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.443380	ALMOND PASTE
23	Processed meat prod <10 g fat excl liver av	GREEN BEAN (fresh) (g)*	16.393443	KANGAROO MEAT*	0.150557	GREEN BEAN (fresh) (g)*	0.665821	EMU BONE FREE MEAT
24	Tomato puree concentrated tinned	GREEN BEAN (fresh) (g)*	20.000000	TOMATO PUREE	0.579739	GREEN BEAN (fresh) (g)*	0.446951	GERKIN (G)
25	Apricots dried	GREEN BEAN (fresh)	20.000000	YEAST DRIED*	0.336097	GREEN BEAN (fresh)	0.504831	MELON (g)*
26	Toddler formula Hero Baby Standaard 3 p 100 ml	GREEN BEAN (fresh) (g)*	18.750000	BEER IN CAN	0.000000	GREEN BEAN (fresh) (g)*	0.501427	YOGURT LACTOSE FREE
27	Beans chilli canned	GREEN BEAN (fresh) (g)*	21.621622	BEANS (F)	0.449436	GREEN BEAN (fresh)	0.475451	GREEN BEAN (fresh) (g)*
28	Cheese Dutch in Swiss-style 45+	GREEN BEAN (fresh) (g)*	16.326531	CHEESE	0.303216	CHEESE SEMI-HARD	0.557272	CHEESE SEMI-HARD
29	Ginger root	GREEN BEAN (fresh)	14.814815	GINGER	0.579739	GREEN BEAN (fresh)	0.306001	OAT MEAL
30	Coffee w sugar and milk vending machine	GREEN BEAN (fresh)	18.181818	MILK CHOCOLATE	0.175786	GREEN BEAN (fresh) (g)*	0.460093	COFFEE DRIP FILTERED (L)
31	Fruit juice drink raspberry	GREEN BEAN (fresh)	18.604651	RASPBERRY (G)	0.379978	GREEN BEAN (fresh)	0.485169	COFFEE DRIP FILTERED (L)
32	Bread corn w sunflower seeds	GREEN BEAN (fresh)	27.272727	BREAD MULTICEREAL**	0.220288	GREEN BEAN (fresh) (g)*	0.458749	BREAD FROZEN (F)*
33	Chewing gum wo sugar	GREEN BEAN (fresh) (g)*	21.052632	BEET SUGAR	0.220288	POMEGRANATE (I)	0.241466	COCOA CAKES AND CROISSANT**
34	Juice multifruit	GREEN BEAN (fresh)	18.750000	APPLE JUICE (I)	0.336097	GREEN BEAN (fresh)	0.405109	ESPRESSO (L)
35	Roll white hard	GREEN BEAN (fresh)	19.354839	WINE WHITE	0.260556	GREEN BEAN (fresh)	0.286102	BREAD FROZEN (F)*
36	Bitter gourd pods raw	GREEN BEAN (fresh) (g)*	20.512821	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.475321	PEPPER (g)
37	Gnocchi unprepared	GREEN BEAN (fresh)	17.647059	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.317994	GREEN BEAN (fresh)
38	Salad cream 25% oil	GREEN BEAN (fresh) (g)*	21.621622	CREAM	0.379978	GREEN BEAN (fresh) (g)*	0.470744	TOMATO & BASIL
39	Sherry	GREEN BEAN (fresh)	18.181818	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.263632	WINE WHITE
40	Beer low alcohol 0,1-1,2 vol%	GREEN BEAN (fresh) (g)*	17.021277	BEER IN CAN	0.170776	CHEESE SEMI-HARD	0.523529	BEER IN CAN
41	Bread brown/wholemeal av	GREEN BEAN (fresh) (g)*	19.047619	BREAD MULTICEREAL**	0.220288	GREEN BEAN (fresh) (g)*	0.451555	BREAD FROZEN (F)*
42	Cherries	GREEN BEAN (fresh)	25.000000	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.448606	MELON (g)*
43	Beef <10% fat prepared av	GREEN BEAN (fresh) (g)*	27.906977	BEEF BONE FREE MEAT*	0.127360	CUCUMBER (g)	0.549878	EMU BONE FREE MEAT
44	Carambola	GREEN BEAN (fresh)	8.000000	BEER IN CAN	0.000000	GREEN BEAN (fresh) (g)*	0.274277	COCONUT (I)
45	Breakfast drink Goede Morgen Vifit	GREEN BEAN (fresh) (g)*	19.230769	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.330994	YOGURT LACTOSE FREE
46	Bacon smoked Katenspek	GREEN BEAN (fresh)	21.052632	BACON	0.449436	GREEN BEAN (fresh)	0.321531	EMU BONE FREE MEAT
47	Prunes soaked in water	GREEN BEAN (fresh) (g)*	20.000000	MINERAL WATER*	0.220288	GREEN BEAN (fresh)	0.279758	GREEN BEAN (fresh) (g)*
48	Sauce barbecue	GREEN BEAN (fresh)	20.000000	BEER IN CAN	0.000000	GREEN BEAN (fresh)	0.353308	TOMATO & BASIL
49	Oil sunflower seed	GREEN BEAN (fresh)	23.529412	SUNFLOWER OIL	0.709297	GREEN BEAN (fresh)	0.306893	PESTO WITHOUT GARLIC

Figure 15: Table with the comparison of the performance of the different models

8.2 Relative CO_2 difference

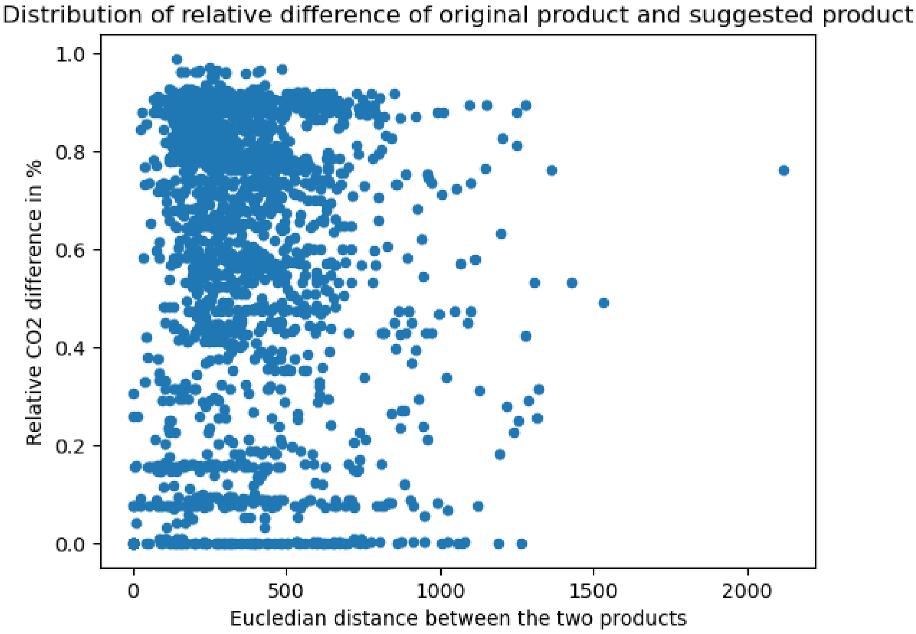


Figure 16: Relative difference Affinity propagation method

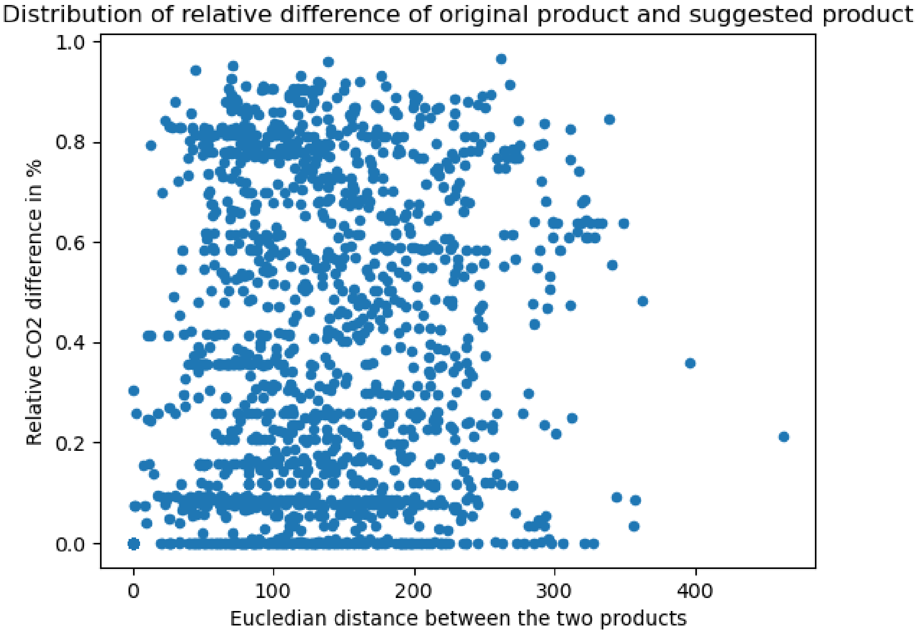


Figure 17: Relative difference with k-means method on original data

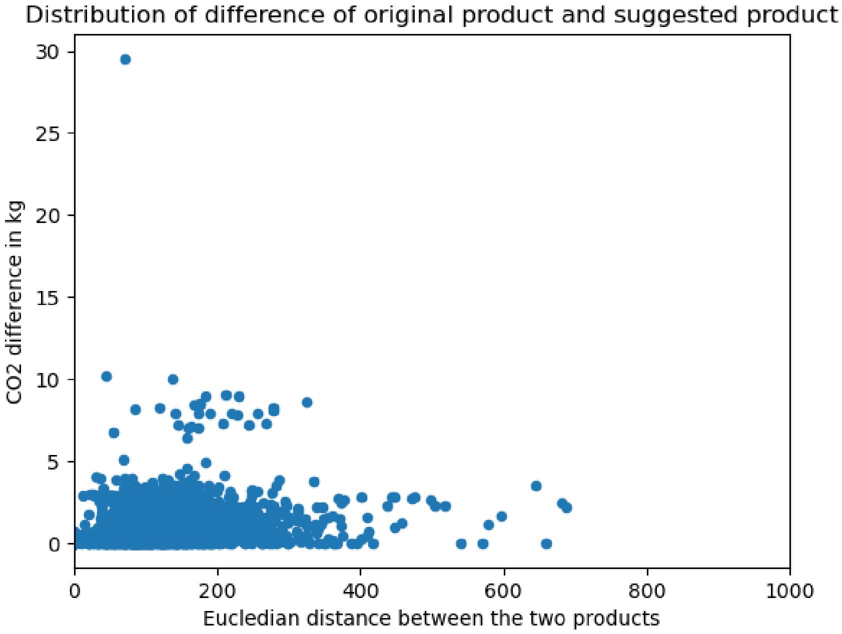


Figure 18: Relative difference with K-means method on the first eight principle components