Bohn, Kristin; Amberg, Michael; Meier, Toni; Forner, Frank; Stangl, Gabriele I.; Mäder, Patrick

Estimating food ingredient compositions based on mandatory product labeling

Original published in:	Journal of food composition and analysis / International Network of Food Data Systems. - Orlando, Fla. : Academic Press. - 110 (2022), art. 104508, 9 pp.
Original published:	2022-03-12
ISSN:	0889-1575
DOI:	10.1016/j.jfca.2022.104508
[*Visited:*	2022-10-25]

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

journal homepage: www.elsevier.com/locate/jfca

Estimating food ingredient compositions based on mandatory product labeling

Kristin Bohn [a], Michael Amberg [a], Toni Meier [c,d], Frank Forner [c], Gabriele I. Stangl [c], Patrick Mäder [a,b,*]

[a] *Technische Universität Ilmenau, Max-Planck-Ring 14, 98693 Ilmenau, Germany*
[b] *Faculty of Biological Sciences, Friedrich Schiller University, 07745 Jena, Germany*
[c] *Martin-Luther-Universität Halle-Wittenberg, Institute for Agricultural and Nutritional Sciences, Von-Danckelmannplatz 2, 06120 Halle (Saale), Germany*
[d] *Competence Cluster for Nutrition and Cardiovascular Health (nutriCARD), Jena-Halle-Leipzig, Germany*

ARTICLE INFO

Keywords:
Recipe analysis
Food compositional tables
Food retail market
Nutrition label
Reformulation

ABSTRACT

Having a specific understanding of the actual ingredient composition of products helps to calculate additional nutritional information, such as containing fatty and amino acids, minerals and vitamins, as well as to determine its environmental impacts. Unfortunately, producers rarely provide information on how much of each ingredient is in a product. Food manufacturers are, however, required to declare their products in terms of a label comprising an ingredient list (in descending order) and Big7 nutrient values. In this paper, we propose an automated approach for estimating ingredient contents in food products. First, we parse product labels to extract declared ingredients. Next, we exert mathematical formulations on the assumption that the weighted sum of Big7 ingredients as available from food compositional tables should resemble the product's declared overall Big7 composition. We apply mathematical optimization techniques to find the best fitting ingredient composition estimate. We apply the proposed method to a dataset of 1804 food products spanning 11 product categories. We find that 76% of these products could be analyzed by our approach, and a composition within the prescribed nutrient tolerances could be calculated, using 20% of the allowed tolerances per Big7 ingredient on average. The remaining 24% of the food products could still be estimated when relaxing one or multiple nutrient tolerances. A study with known ingredient compositions shows that estimates are within a 0.9% difference of products' actual recipes. Hence, the automated approach presented here allows for further analysis of large product quantities and provides possibilities for more intensive nutritional and ecological evaluations of food.

1. Introduction

Taking the increasing amount of pre- and ultraprocessed foods in supermarkets and the prevailing burden of malnutrition into account (Afshin et al., 2019; Monteiro et al., 2019), more sophisticated evaluation and monitoring tools - resulting in nutritionally better balanced foods - should be developed to overcome the challenges in the food-environment-health nexus. Currently, consumers are confronted with hundreds to thousands of different foods in retail markets, which are often preprocessed and labeled with a range of mandatory and voluntary information. Unfortunately, the extent of different brands, labels, stores and marketing strategies does not help consumers make more sophisticated decisions but rather leads to increased consumer confusion (Wobker et al., 2015). Although the updated health claim regulation (European Commission, 2012a) has allowed for evidence-based health and nutrient statements on foods since 2012, only a small number of everyday foods are labeled with these. According to Bratzke et al. (2018), only 1.9% of all available meat products in food retailing markets in Germany are explicitly labeled *health* promoting. For dairy products, market penetration was slightly lower, ranging between 0.6% and 2.7% (average: 1.6%). On the other hand, the overconsumption of sugars, saturated fats and salt (and other food additives) cause tremendous disease and financial cost burdens, which undermine the viability of health care systems (Bommer et al., 2017; Meier et al., 2015, 2017; Tremmel et al., 2017; World Health Organization, 2019). Moreover, the majority of consumers have a keen interest in more transparent and more holistic labeling of food products, taking into account not only further nutritional information such as mineral content

[*] Corresponding author at: Technische Universität Ilmenau, Max-Planck-Ring 14, 98693 Ilmenau, Germany.
E-mail address: patrick.maeder@tu-ilmenau.de (P. Mäder).

https://doi.org/10.1016/j.jfca.2022.104508
Received 19 September 2021; Received in revised form 6 February 2022; Accepted 3 March 2022
Available online 12 March 2022

but also key environmental facts (Verbraucherzentrale Bundesverband e.V, 2019). A crucial piece of information in this context is the composition of ingredients (recipe) when aiming for more precise and cross-product information on health and environmental sustainability. Such information is only rarely given on food packages. In this work, we propose an automated, optimization-based approach to calculate the ingredient compositions of preprocessed food types via mathematical optimization. We evaluate the approach on a dataset consisting of 1804 real food products belonging to different food categories offered in German supermarkets in 2019 and 2020. Eventually, we evaluate whether the estimated food ingredient composition corresponds to the actual composition declared by the producer.

2. Methods

Our proposed method aims to calculate the food composition, i.e., the percentage of all ingredients in a food product. A list of ingredients and the so-called Big7, i.e., the amount of energy, fat, saturated fat, carbohydrates, sugar, protein and salt (sodium chloride) contained in the product are publicly disclosed on a product's package. We propose a six-step process (cp. Fig. 1) to estimate a product's composition based on this information. First, we extracted individual ingredients from a product's list of ingredients as well as its Big7 information. These steps involved splitting and interpreting the list of ingredients into single, common-name ingredients that are mapped to a food compositiional table (FCT). For our study, we used the German Nutrient Database (Bundeslebensmittelschlüssel) with up to 138 nutrients per ingredient as the FCT (Hartmann et al., 2014). Combining ingredient and nutrient information, we built a system of linear equations constrained by additional information regarding the regulatory allowed declaration tolerances. Eventually, we selected the best solution among all possible ingredient compositions fulfilling the system of equations. Therefore, the best solution was defined as the solution with the minimal squared error of the calculated Big7 values compared to the labeled Big7 values. Throughout the following sections, we discuss each of these steps in detail.

2.1. Data extraction

From each food product p, we acquired its Big7 B and its ingredient list L in the form of a text string (cp. Step 1a and 1b in Fig. 1) denoted as:

$$p = (B, L). \tag{1}$$

Therefore, Big7 B includes (1) energy, (2) fat, (3) saturated fat, (4) carbohydrates, (5) sugar, (6) protein, and (7) salt content and is denoted as:

$$B = (b_1, b_2, \cdots, b_7). \tag{2}$$

The European Union's food labeling regulations prescribe an ingredient list on every packaged food product (European Commission, 2011). When aiming to analyze large quantities of food products, manually acquiring this information from a product's package is time-consuming (cp. Step 1 in Fig. 1). A method that can support and speed up manual data acquisition is optical character recognition (OCR) (Lazzari et al., 2018), which retrieves textual information from product images automatically. Alternatively, commercial product data collections, e.g., GDSN data pools (GS1 Germany), or community-collected product data, e.g., open food facts (Open Food Facts association), may be used as data sources.

2.2. Split ingredient list into individual ingredients

The EU food labeling regulations also prescribe that ingredients must be listed in descending order of their contents on a product's package (European Commission, 2011). A precondition for further analysis is the separation of these lists into individual ingredients. Declared ingredients

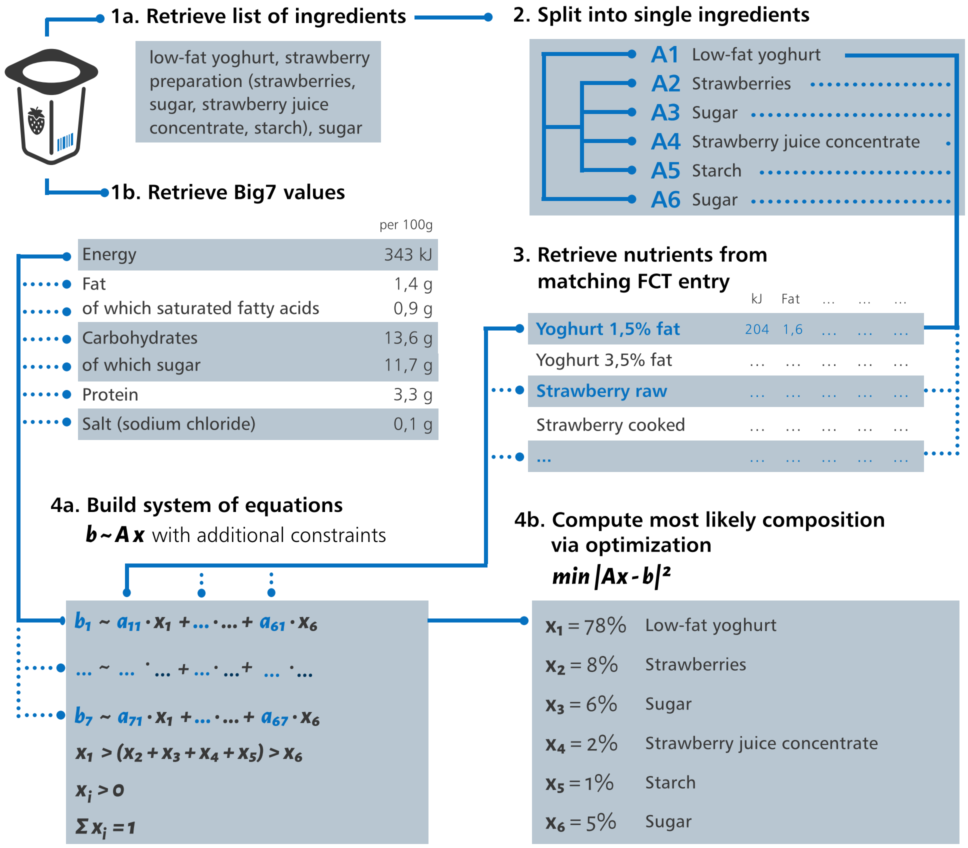
1a. Retrieve list of ingredients
low-fat yoghurt, strawberry preparation (strawberries, sugar, strawberry juice concentrate, starch), sugar
1b. Retrieve Big7 values
per 100g
Energy 343 kJ
Fat 1,4 g
of which saturated fatty acids 0,9 g
Carbohydrates 13,6 g
of which sugar 11,7 g
Protein 3,3 g
Salt (sodium chloride) 0,1 g
2. Split into single ingredients
A1 Low-fat yoghurt
A2 Strawberries
A3 Sugar
A4 Strawberry juice concentrate
A5 Starch
A6 Sugar
3. Retrieve nutrients from matching FCT entry
kJ Fat … … …
Yoghurt 1,5% fat 204 1,6 … … …
Yoghurt 3,5% fat … … … … …
Strawberry raw … … … … …
Strawberry cooked … … … … …
… … … … … …
4a. Build system of equations
b ~ A x with additional constraints
$b_1 \sim a_{11} \cdot x_1 + \ldots \cdot \ldots + a_{61} \cdot x_6$
$\ldots \sim \ldots \cdot \ldots + \ldots \cdot \ldots + \ldots \cdot \ldots$
$b_7 \sim a_{71} \cdot x_1 + \ldots \cdot \ldots + a_{67} \cdot x_6$
$x_1 > (x_2 + x_3 + x_4 + x_5) > x_6$
$x_i > 0$
$\Sigma x_i = 1$
4b. Compute most likely composition via optimization
$\min |Ax - b|^2$
$x_1 = 78\%$ Low-fat yoghurt
$x_2 = 8\%$ Strawberries
$x_3 = 6\%$ Sugar
$x_4 = 2\%$ Strawberry juice concentrate
$x_5 = 1\%$ Starch
$x_6 = 5\%$ Sugar

Fig. 1. Overview of the approach: we retrieve (1a) a list of ingredients and (1b) declared nutrition information from a product's label information. (2) The declared ingredient string is split into individual ingredients. (3) These ingredients are mapped to a food compositional table (FCT) to retrieve their Big7 nutrients. (4a) All acquired information is used to build an equation system subject to a variety of constraints. (4b) An optimization method is used to identify an optimal solution, i.e., an estimate of the product's ingredient composition.

may be accompanied by percentages referring to the relative weight of an ingredient in the composition. Additionally, ingredients may be hierarchically composed of a list of subingredients, which in extreme cases may be further composed of subingredients thereby spanning a compositional tree. For this analysis, we were mainly interested in the most detailed ingredient information. Thus, out of the ingredient list L, we extracted n most detailed ingredients or subingredients I. To utilize the dominance of ingredients as additional information, we also considered their hierarchical positions E. Furthermore, we captured compositional amounts of these ingredients if declared (cp. Step 2 in Fig. 1). We denote an ingredient list L as:

$$L = (n, I, E, R), \tag{3}$$

where n is the number of noncompositional ingredients, $I = (i_1, i_2, \cdots, i_n)$ refers to the noncompositional ingredients in the form of a text or string, $E = (e_1, e_2, \cdots, e_n)$ is the hierarchical position (number) of every ingredient, and $R = (r_1, r_2, \cdots, r_n)$ is the given percentage per ingredient ranging from 0.0 to 1.0 if available.

While the regulative framework prescribes the contents of ingredient lists, they do not detail their formatting, e.g., no rule is in place regarding the separation of ingredients, meaning that they may be separated by commas as well as by semicolons. The declaration of subingredients may be indicated with a colon or using different forms of brackets. From a technological point of view, the ingredient list is a sequence of characters, called a string, possibly annotated with additional details and clustered into a hierarchy that we aimed to decompose into a well-defined data structure. From a human perspective, this is simple due to our linguistic comprehension and acquired knowledge in performing similar tasks. However, a computer program needs precise rules to perform this separation, i.e., which character signals a delimiter between ingredients and which characters signals subingredients. For example, brackets are sometimes used to indicate subingredients, while for other products, they indicate allergens. Consider the ingredient herbs containing the allergen mustard to illustrate the variety of common ingredient formatting used on products today: "Herbs (MUSTARD)" vs. "Herbs containing MUSTARD" vs. "Herbs (contains MUSTARD)". All these representations and many more are commonly used in product declarations and need to be interpreted as the ingredient herbs. We propose a rule-based approach to cope with this problem. More specifically, we denoted grammar defining our agreed structure of an ingredient list. Additionally, we propose a set of rewriting rules that transform deviations, i.e., ingredient lists using a different formatting, into this ideal ingredient list format. We use the grammar as well as the rewriting rules as input to the well-known ANTLR parser generator (Parr, 2020), which generates a program for our given ingredient list parsing problem.

2.3. Map ingredients to food compositional table

In the mapping process, we aimed to retrieve Big7 values for all identified ingredients in the previous step (cp. Step 3 in Fig. 1). These values are commonly available in food composition tables (FCTs). An FCT contains detailed sets of information on the nutritionally important components of typical foods. The table lists not only the Big7 values, but also additional nutrients, vitamins, minerals and others. We mapped the n ingredients I elementwise to entries in the FCT FCT_{all} denoted as:

$$f_m : I^n \rightarrow FCT^n_{all}. \tag{4}$$

Once mapped, we acquired the respective Big7 information of this ingredient to be later used in our approach. Eventually, this step of the proposed process resulted in $A_i = a_{i,1}, a_{i,2}, \cdots, a_{i,7}$, with A_i referring to the overall Big7 information for ingredient i and $a_{i,1}, a_{i,2}, \cdots, a_{i,7}$ being individual Big7 values. Since the FCT typically only contains widely distributed ingredients, we allowed such entries to also be manually added FCT_{add}, i.e., $FCT_{all} = FCT \cup FCT_{add}$. For this study, we used the German nutrient database (Bundeslebensmittelschlüssel (BLS) (Hartmann et al., 2014)) containing 10,169 food entries comprising processed food products and ingredients as well as preprocessed ingredients, raw and cooked, with up to 138 nutrients documented per entry. We used fuzzy string matching to map a product's ingredient to the BLS. This method compared the name of an ingredient to all BLS entries and retrieved the most similar ingredient. We used the Levenstein distance as a similarity metric (Levenshtein, 1965), i.e., computing the minimal number of single-character edits (insertion, deletion, substitution) needed to change one word into another. However, fuzzy string matching did not always result in the most similar entry because entries of FCTs carry a standard name, while producers may choose from a rich set of synonyms or grammatically different forms to refer to the same ingredient. Thus, in this step wrongly matched ingredients can occur. Furthermore, the processing grade may not be declared appropriately in the ingredient list. Therefore, an expert manually checked and potentially corrected matches with a low similarity. We not only corrected the given mapping but also stored a kind of white list of matches to be used in further matching processes. For some food ingredients, such as goji berries, the BLS does not provide a suitable entry. For these cases, we retrieved information from other FCTs as FCT_{add}, such as the Food DataCentral Database (U.S. Department of Agriculture, Agricultural Research Service, 2019). The extend of the curated information will be discussed within Section 3.1.

2.4. Calculate the product's ingredients composition

Eventually, our goal was to calculate the unknown relative amount x per ingredient i in a given product. The information retrieved in the previous steps allowed for us to formulate a linear equation system assuming that the weighted sums of every ingredient's Big7 value $a_{i,j}$ should match the product's overall Big7 values $B = (b_1, b_2, \cdots, b_7)$ as close as possible (cp. Step 5 of Fig. 1) and denoted this relationship as

$$\begin{aligned}
b_1 &= a_{1,1}\cdot x_1 + a_{2,1}\cdot x_2 + a_{3,1}\cdot x_3 + \ldots + a_{n,1}\cdot x_n\\
b_2 &= a_{1,2}\cdot x_1 + a_{2,2}\cdot x_2 + a_{3,2}\cdot x_3 + \ldots + a_{n,2}\cdot x_n\\
&\ldots\\
b_7 &= a_{1,7}\cdot x_1 + a_{2,7}\cdot x_2 + a_{3,7}\cdot x_3 + \ldots + a_{n,7}\cdot x_n.
\end{aligned} \tag{5}$$

Thus, we argued that $b = A \cdot x$ where b refers to the product's overall Big7, A refers to the Big7 per ingredient and x denotes the unknown proportion of ingredients in the analyzed food product. Since all relevant information, i.e., declared Big7 and nutrient information retrieved from the FCT, were already normalized to 100 g or 100 ml of the product and ingredient, respectively, there was no need to specifically consider product weight or volume in our analysis. Other forms of normalization, such as per portion, were not permitted within the EU and were, therefore, beyond our scope. In cases where FCT ingredient information is only available per weight but required per volume, the ingredient's density is needed to convert between both. This equation system was subject to regulatory prescribed tolerance ranges per Big7. Table 1 shows the defined tolerances among the Big7 according to the European Commission (2012b), with most nutrients having a constant minimum

Table 1
Allowed tolerances per Big7 defined by the European Commission (2012b).

Nutrient	Content per 100 g	Tolerance
carbohydrates, sugar, protein, fiber	$< 10\,g$	$\pm 2\,g$
	$10 - 40\,g$	$\pm 20\%$
	$> 40\,g$	$\pm 8\,g$
fat	$< 10\,g$	$\pm 1.5\,g$
	$10 - 40\,g$	$\pm 20\%$
	$> 40\,g$	$\pm 8\,g$
saturated fat	$< 4\,g$	$\pm 2\,g$
	$\geq 4\,g$	$\pm 20\%$
salt	$< 1.25\,g$	$\pm 0.375\,g$
	$\geq 1.25\,g$	$\pm 20\%$

and maximum tolerance as well as a relative tolerance in between. Knowing a product's declared Big7 b, we can define their allowed tolerance $b_{tolerance}$. For example, a product with a declared protein value of 25 g would be allowed to vary by up to 20%, amounting to $b_{3,tolerance} = 5\ g$. These tolerances constrained the possible solutions to our equation system and we denoted them as:

$$b - b_{tolerance} < A\cdot x < b + b_{tolerance} \tag{6}$$

This system of equations is, additionally, subject to further constraints regarding the calculated proportions of ingredients:

$$\begin{aligned} & x_i > 0 \\ & \sum x_i = 1 \\ & x_i \geq x_{i-1} \end{aligned} \tag{7}$$

These constraints reflect that: (1) all ingredients need to contribute a minimal proportion, (2) the proportions of all ingredients need to sum up to 1, respectively and 100%, and (3) the proportions of ingredients need to decrease from the first to the last ingredient. It should be noted, that behind within ingredient x_i some aggregated subingredients can be present, which are included in the equations, but are not shown here for simplicity.

Next, our goal was to determine a solution to this system of equations. From a mathematical point of view, a system of linear equations with constraints may be either infeasible when there is no solution that satisfies all of the constraints or feasible otherwise. A feasible system of linear equations may have no solution (overdetermined), one solution (even-determined), or an infinite number of solutions (underdetermined). An overdetermined problem with no solution occurs if more linearly independent equations exist than unknowns. Even-determined is the ideal case that yields exactly one solution and results from a number of linearly independent equations, i.e., ingredients' Big7 values are linearly independent, equaling the number of unknown ingredient amounts. An underdetermined problem exists when there are more unknowns than independent equations, resulting in an infinite number of possible solutions, i.e., various alternative ingredient combinations yield the Big7 values declared on the product. Solvable, i.e., even-determined, problems will only rarely exist since products rarely consist of exactly seven ingredients. However, we can often still find a suboptimal solution to an over- or underdetermined problem by treating it as an optimization problem. Therefore, we formulated the given problem as an approximate equation system and employed optimization techniques to identify its best possible solution:

$$b \simeq A\cdot x. \tag{8}$$

An optimal solution to this problem minimizes the difference between the calculated and labeled Big7 measured in terms of squared error, i.e., a least squares problem:

$$\min\left|\left|\frac{(A\cdot x - b)}{b_{tolerance}}\right|\right|^2. \tag{9}$$

Therefore, we normalized differences by the predefined tolerance ranges $b_{tolerance}$ to facilitate their more intuitive interpretation with regard to regulatory tolerances. Solving this optimization problem yielded a proportional combination of ingredients ($x_1, x_2, \cdots, x_n$) nearest to the Big7 declared on the product's package B. We used linear programming, a mathematical technique that allowed for determining an optimal solutions that satisfy several constraints at once (Dantzig and Thapa, 1997). Linear programming has previously been employed in nutrition sciences, e.g., for solving the mink diet problem (Ben-David et al., 1997). All our analyses were implemented with R, and we used the limSolve package (Soetaert et al., 2009a; Soetaert et al., 2009b) to solve the optimization problem.

An ingredient list may numerically declare the relative amount of one or more ingredients, e.g., "10% sugar". This information is additional information toward the product's ingredient composition and should be used, if available, in the optimization problem. Therefore, we formulated these as additional conditions in the equation system. Furthermore, the success of our approach depends on our ability to precisely map product ingredients to FCT entries. This was not restricted to a single ingredient but also extended to its various degrees of processing, e.g., fresh tomato vs. single-concentrated tomato puree vs. triple-concentrated tomato puree, which greatly influenced the ingredient's Big7 values. However, these processing stages will rarely be completely covered by an FCT. We propose adding water as a virtual ingredient x_w to account for the ingredient's Big7 variations arising from ingredient processing, such as cooking, drying, and freezing, and denote the following additional constraint:

$$\begin{aligned} & x_{w,min} < x_w < x_{w,max} \\ & \sum x_i + x_w = 1. \end{aligned} \tag{10}$$

Processing water contributes to the overall sum of ingredients and can, contrary to the other ingredients, be positive or negative but does not contribute to any Big7 values. Processing water is only considered when the constraints counteract, so no solution can be found without its application.

3. Evaluation

To evaluate the proposed approach, we formulated and systematically studied the following research questions:

1. **Applicability of the approach and optimization error.** How often and with which quality can the proposed approach deliver a solution to the formulated optimization problem for a large variety of given food products?
2. **Food category-specific analysis.** Are such quality variations specific to certain food categories and if so, what are potential reasons?
3. **Viability of processing water.** Does considering water as an additional undeclared ingredient improve analytical results?
4. **Relaxing optimization constraints.** Does relaxing of individual nutrient constraints yield an accurate ingredient composition for otherwise infeasible optimization problems?
5. **Estimated vs. actual composition.** How close does the estimated ingredient composition of a product match its actual composition?

3.1. Dataset

We employed a dataset with 1804 real food products belonging to eleven food categories and consisting of twelve ingredients on average ranging from a minimum of 1 to a maximum of 68 ingredients (cp.

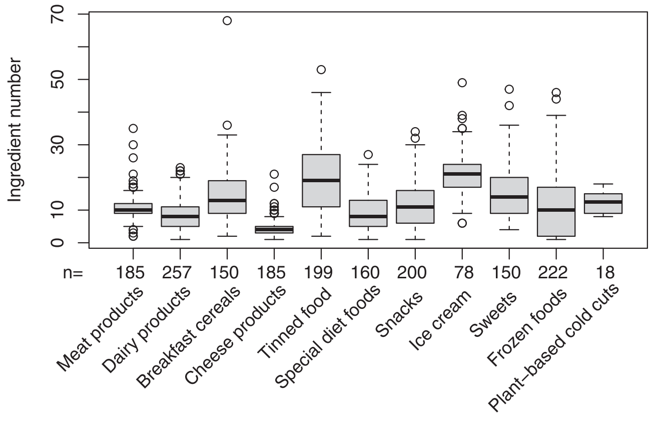

Fig. 2. Evaluated products per food category and their average number of ingredients, including additives. The number n on the x-axis refers to the count of products per category.

Fig. 2). The dataset was provided by the University Halle-Wittenberg and comprises the following categories: snacks, sweets, milk products, cheese products, cereals, tinned foods, frozen products, meat products, and special diet food products. Individual food products were collected from several supermarkets in Germany in 2019 and 2020, representing a representative sample reflecting the current market. Since vegan and vegetarian cheese and meat products substantially diverge from their nonvegetarian counterparts, we decided to consider them as the extra group plant-based cold cuts, with only a very small sample size. For the same reason, we separated ice cream from frozen foods. Special diet foods comprise products from health food shops such as suitable for diabetics or gluten-free. Dairy products do not include cheese. In terms of the average number of ingredients, cheese products were characterized by the lowest number (5 on average), while tinned food and ice cream showed the highest number of ingredients (19 and 20). The overall highest number of 68 ingredients was observed in a breakfast cereal product. For each product, graduate students of nutritional sciences collected the name, GTIN, list of ingredients, Big7 values and other nutrient information from its package. Then, the students formatted the ingredient lists into an upfront agreed style that we could accurately parse in subsequent analysis. We matched each ingredient to the FCT discussed above (cp. Sec.2.3) and manually checked and corrected ingredients with low matching scores. Therefore, we acquired a dictionary mapping 842 declared ingredients to 373 FCT entries. These mappings included synonyms but also additives, such as smoke, that were not listed in the FCTs since they contain no nutrients. In total, for the 1804 products consisting of 21,458 ingredients, 1701 (7,9%) declared ingredients were matched via the dictionary. Out of the 1804 products, merely 9 declared their Big7 normalized to 100 ml of product volume rather than 100 g of product weight. Respective FCT entries where either already available per 100 ml or could be converted via known densities, e.g., water, milk, and oil.

4. Results

4.1. RQ1: applicability of the approach and optimization error

For 64% of the food products from our dataset (1142 out of 1804), our approach was able to calculate an ingredient composition so such that all specified constraints were fulfilled. Therefore, we observed an underdetermined equation system in 731 cases (64%), an overdetermined equation system in 387 cases (34%), and an even-determined equation system in merely 24 cases (2%) (cp. Fig. 3 (left)). That is, for only 28 food products (2%), an ingredient composition can be estimated by solving the equation system, i.e., the Big7 values matched exactly. For 1114 food products (98%), an ingredient composition can only be estimated via optimization, meaning that the Big7 values did do not match exactly but were within the EU's prescribed tolerances. Considering the optimization error relative to the respective tolerance per Big7 value and averaged across those (cp. Eq.9), we can compare matching quality across products. Fig. 3 (right) shows a histogram of this error exposing a decreasing asymptotic error behavior. The error's median amounted to 0.038, i.e., the average value was within a $\sqrt{(0.038)} = 19\%$ tolerance range (cp. Eq.9). We observed a rather similar distribution of error across underdetermined and overdetermined systems of equations (cp. Fig. 3 (right)).

4.2. RQ2: Food category-specific analysis

In a more detailed analysis, we studied whether and how much the optimization error can be attributed to the category of a food product. We observed large variations in the number of products that can be analyzed, i.e., feasible equation system, ranging from 20% to 100% analyzable products per food category (cp. Fig. 4 (left)). From left to right in the figure, we observed the following amounts of analyzable products: meat products (22%), dairy products (96%), breakfast cereals (88%), cheese products (33%), tinned food (82%), special diet food (52%), snacks (62%), ice cream (84%), sweets (65%), frozen food (80%), and plant-based cold cuts (100%). When comparing the analyzable products per category in terms of optimization error (cp.

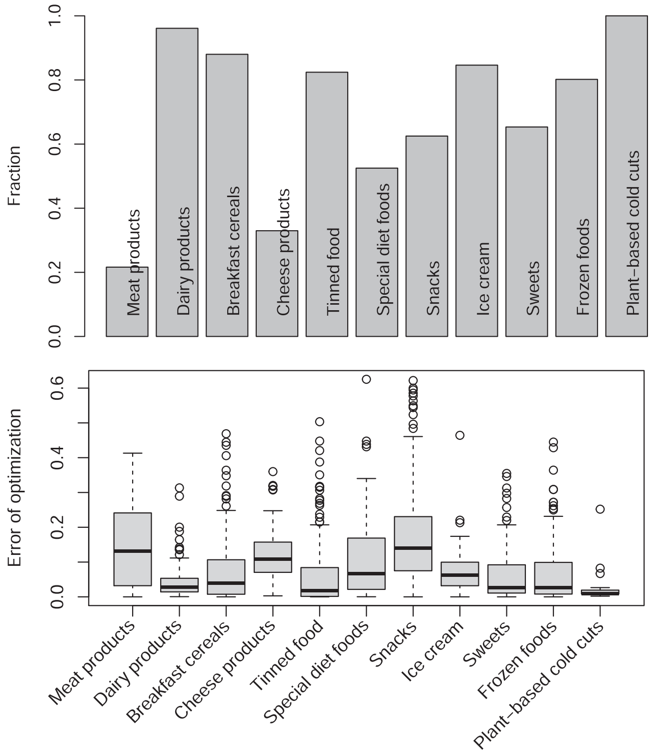

Fig. 4. Bar chart showing the proportion of feasible equation systems of food products per category (left) and box plots showing optimization error per category (right).

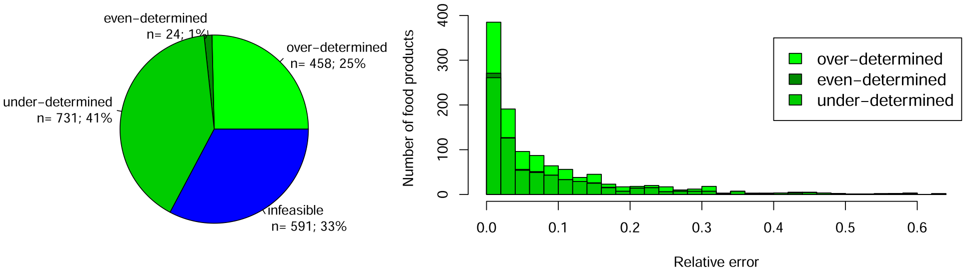

Fig. 3. Pie chart illustrating the proportions of products with solvable optimization problem (RQ1) in green and with unsolvable in blue (left) and distribution of optimization error across the products with feasible optimization problem (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 (right)), we observed a trend toward product categories with a larger share of feasible optimization problems yielding a lower, i.e., better, average optimization error. However, there were also exceptions to this observation, e.g., snack products yielded the highest average optimization error with widely distributed outliers while being roughly average among all food categories (61% vs. 64%) in terms of analyzable products.(cp. Fig. 5).

4.3. RQ3: viability of processing water

We applied the processing water concept to all food products (n = 591) with a previously infeasible optimization problem (cp. RQ1). With water as an additional ingredient, 28% (n = 165) of these products can now be analyzed (cp. Fig. 6). Therefore, we set $x_{w,min} = -3$ and $x_{w,max} = 3$, i.e., we allowed for a maximum of 300% dehydration or dilution, respectively. Across these now analyzable products, optimization delivered an averaged processing water of 116% dehydration, ranging from 300% dehydration to 70% maximum dilution. The extremes of dehydration (cp. Fig. 5) occurred for potato chips and vegetable chips in the snack category. These energy-dense, dried or fried foods are characterized by intense processing that largely removes water, which is not outlined in the ingredient lists. Extremes of dilution occurred when cereals or beans are hydrated, and Big7 value declaration may include the water that is needed for preparation. In summary, for 1378 out of 1804 food products (76%), an ingredient composition could be calculated via automated optimization. However, for 425 food products (24%), an ingredient composition cannot be computed since the given constraints were not satisfiable together.

4.4. RQ4: relaxing optimization constraints

No ingredient composition could be calculated for 24% of the selected foods. Thus, estimating an ingredient composition meeting all constraints, mainly those imposed by the EU's Big7 tolerances, was not possible for these products. Products with infeasible equation systems existed across all food categories, except for plant-based cold cuts. These systems occurred most dominantly among cheese and meat products and less dominantly among special dietary foods and sweets (cp. Fig. 6). Relaxing some of the equation system's constraints can yield a feasible equation system while violating the allowed declaration tolerances for the relaxed nutrient(s). Among the food categories that include a relatively high number of products, whose ingredient compositions were not estimable due to infeasible constraints, we observed that for meat products, special diet foods and sweets, relaxing a single nutrient constraint typically yielded a feasible equation system (cp. Fig. 7). Fig. 8 shows how often a certain Big7 nutrient needs to be relaxed to yield a feasible optimization problem per food category. We observed no typical nutrients that needed to be relaxed per category. However, for example, for sweets, we observed that carbohydrates and sugars predominantly had to be relaxed.

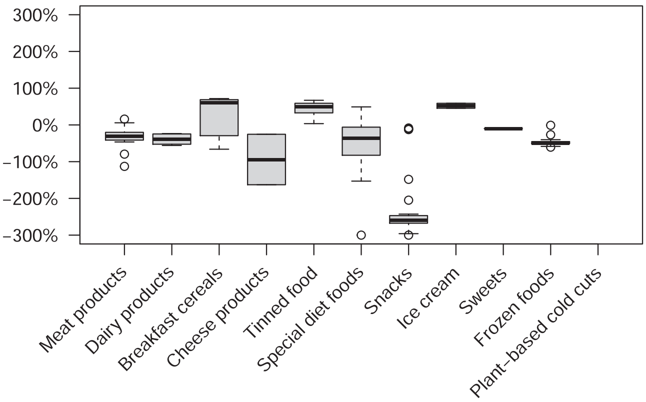

Fig. 5. Boxplot showing each of the eleven food categories across the 165 products analyzed with the processing water concept, the average estimated water content due to dehydration or dilution in the production process.

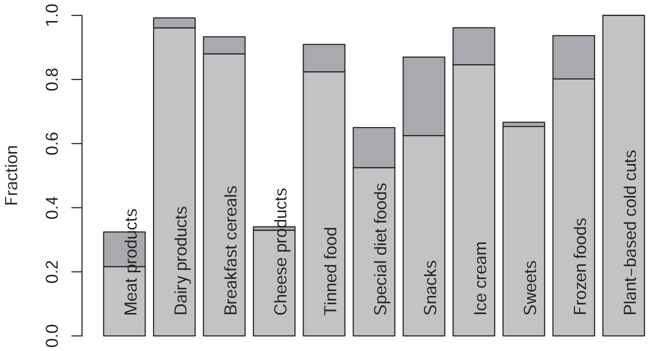

Fig. 6. Bar chart showing the proportion of feasible equation systems per product category without processing water (light gray) and the additional share that becomes feasible with processing water (dark gray).

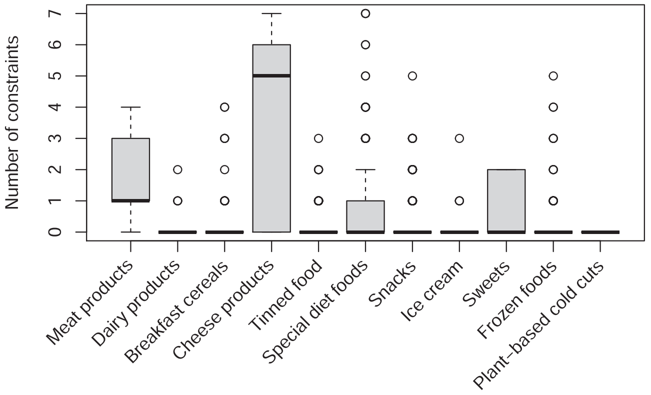

Fig. 7. Bar chart showing the number of Big7 constraints that would need to be relaxed to yield a feasible equation system and to estimate an ingredient composition for otherwise nonestimable products.

4.5. RQ5: validation of the approach on actual composition data

To evaluate the accuracy of the developed algorithm, we compared the calculated and actual ingredient compositions. This comparison was conducted for 33 food products composed of 247 ingredients, of which the manufacturers were willing to provide quantitative information on actual ingredients. These products belonged to three categories, namely, meat products, frozen foods, and tinned foods, and consisted of seven ingredients on average. We compared the quality of our approach on the level of the entire product composition in terms of the mean absolute difference between the calculated and actual ingredient compositions. Here, we found that the estimated ingredients differed on average by 2.7% from the actual ingredient contents. A deviation of more than 5% was observed for only four products (cp. Fig. 9), i.e., beyond some outliers, the estimated ingredient composition fit the actual ingredient composition very well. The standard deviation showed that there were no strong outliers diverging from the average. To obtain more insights into differences in ingredient compositions and the effect of the number of Big7 constraints, Fig. 10 shows that 206 ingredients belonging to 26 products were characterized by a feasible equation system that can be estimated with a median absolute error of 0.9%. The remaining product ingredients were only estimable by relaxing 1 (4 products), 2 (1 product), or 3 (2 products) Big7 constraints, and we observed slightly increasing median absolute errors of 0.9%, 2%, and 4%, respectively. However, we also observed outliers with large differences between the

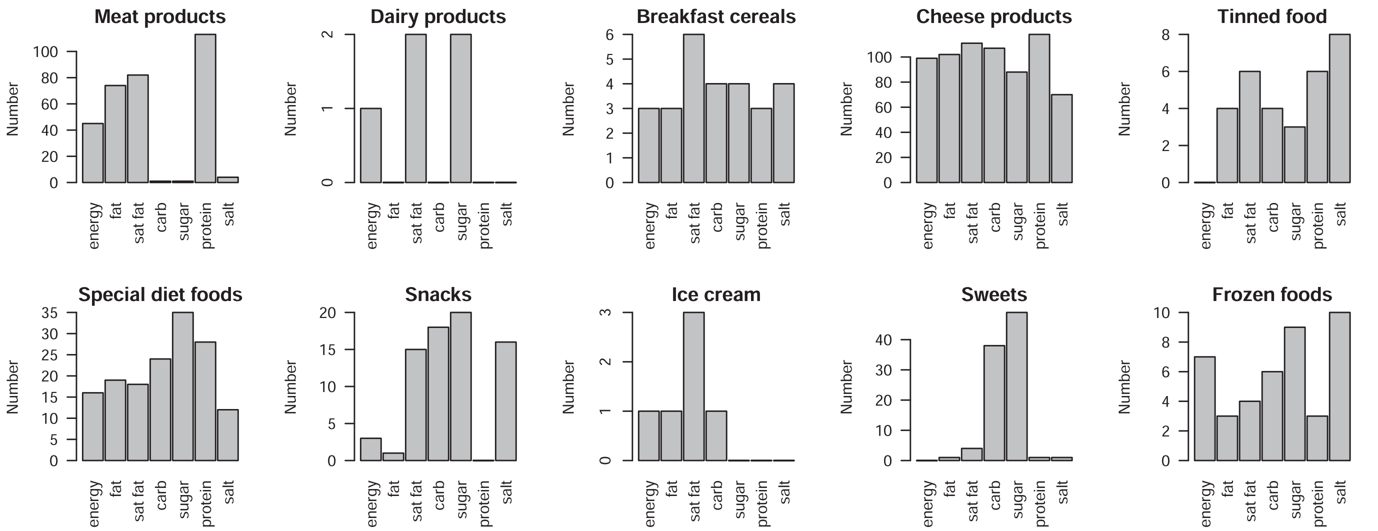

Fig. 8. Per food category, the number of times that a prescribed Big7 tolerance would need to be relaxed to yield a feasible optimization system and to estimate a product ingredient composition.

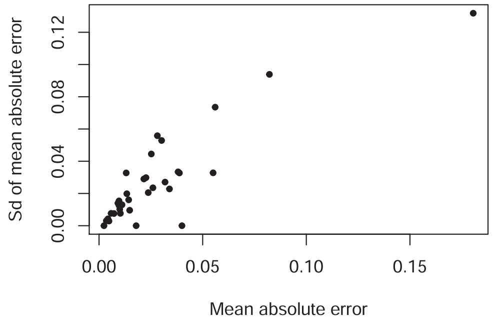

Fig. 9. Mean absolute difference and standard deviation of calculated ingredient composition in relation to the actual composition of 33 food products.

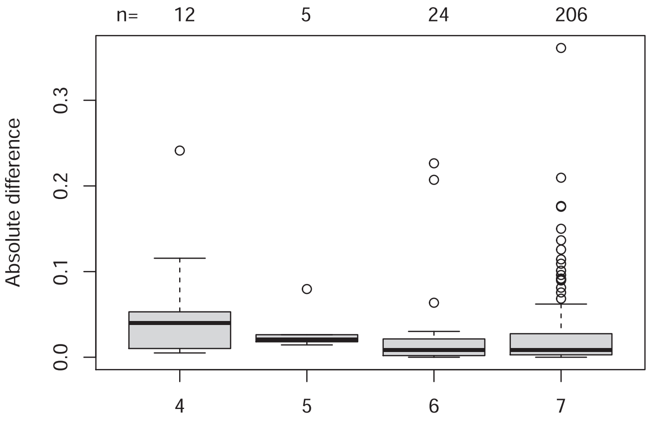

Fig. 10. Averaged absolute difference of calculated and actual ingredient amount for 33 products' ingredient compositions grouped by number of respected Big7 constraints (x-axis) with 7 meaning that all constraints were fulfilled.

actual and calculated amounts of an ingredient. These outliers included ingredients of a berry mix, pesto and pickled cabbage.

5. Discussion

The proposed approach offers a great opportunity to assess the nutritional value of foods in a much more differentiated way than previously. This approach can also be used to identify particularly desirable nutrients, such as n-3 PUFAs and iodine, which are usually not indicated on food packaging. In addition, this approach enables a calculation of life cycle assessment based on the ingredients. Furthermore, the developed approach can be used for nutritional advice in consumption surveys.

Taking the Big7-related regulatory prescribed tolerance ranges into account, our study shows that the proposed approach is able to determine the majority of the ingredient compositions of the food products analyzed. However, depending on the food category, we observed that applied constraints have to be relaxed to calculate compositions (cp. Fig. 6). In particular, cheese and meat products revealed relatively few Big7 values in the allowed ranges and thus low predictability of ingredient compositions. The nutrients of meat vary greatly depending on its fat content. While the fat content of dairy products, such as milk or yogurt, needs to be declared in the ingredient list following EU regulations, declaring the fat content of meat products is not required. As a result, it is difficult to match meat ingredients to the correct entry of an FCT. Cheese products are hard to estimate due to their processing via fermentation from milk and lactic acid bacteria. Processing water would allow for a translation from the volume of milk to the declared fat content while still not resulting in the correct sugar values. Proposing a more appropriate solution to this problem should be addressed in the future.

The quality of ingredient compositions calculated by our approach depends on several factors. First, a precise recognition and interpretation of the declared ingredients as well as a precise matching to FCT entries is a prerequisite for our approach. We propose an advanced parser approach and a mixture of partial string matching and an expert curated and incrementally enhanced dictionary supporting problematic ingredient matches. Second, nutritional values in FCTs are average values representing a common instance of an ingredient, which may vary due to natural influences and variations in the ingredient's production process. For example, tomatoes grown on open farmland may be characterized by different nutrients than those grown in a greenhouse. Overcoming this challenge requires a more detailed declaration of the product as well as richer and more up-to-date FCTs. Third, a fundamental assumption of our approach is that ingredients' Big7 values should sum up to the product's Big7 value. However, the processing of ingredients may cause substantial differences in their nutrients, e.g., a dried tomato differs greatly from a raw tomato in terms of, e.g., water content. This poses a twofold problem for our approach: (1) producers rarely declare ingredients' processing degree on the package and (2) FCT may only provide nutrients for the raw ingredient or a limited set of processing stages. We decided to match the raw ingredient when no further information was declared. However, we argue that at least the case of missing declaration (see (1) above) could be approached by analyzing product categories for common ingredients and their processing degree, e.g., the majority of potato chip products will contain fried potatoes. Such an analysis could be supported by the latest pattern matching and natural language processing methods. Fourth, two or more of a product's ingredients may be characterized by Big7 values in equal proportions, e.g., raw wheat equaling wheat flakes or cheese types with equal fat contents, making it possible to trade one ingredient for the other during optimization (linear combination). These cases complicate the optimization, and the delineation of the respective ingredients is challenging. In the extreme case, where the Big7 values of two ingredients are identical, our proposed method cannot separate them and only estimate their combined amount in a product. This problem occurs often for ingredients that are characterized by only a single nonzero Big7 value, e.g., water and salt differ solely in the amount of the Big7 value of salt, and alcohol and acid differ often solely in the amount of calories. Fifth, another assumption of the approach is that all nutrient information is normalized to the same unit and quantity. In our evaluation, 99.5% of the products were normalized to 100 *g* of product weight and for the remaining products all required information was available per 100 *ml* of product, justifying our assumption. For products declared per 100 *ml* of product volume and FCT entries solely available per 100 *g* of ingredient, density information would be required to convert between volume and weight and their acquisition remains a future excise when scaling the approach to even larger quantities of products.

Optimization and linear programming are applied to several problems in nutrition science. A prominent example is diet problems and the optimization of food costs that shall at the same time satisfy certain nutrient requirements (van Dooren, 2018). We found only a few studies that treat ingredient composition estimation as an optimization problem. As one of the first references, Marcoe and Haytowitz (1993) briefly proposed the idea of using linear optimization for estimating ingredient compositions of several convenience foods. However, the verbally described method is highly manual, and the authors did not further detail their approach. Westrich et al. (1994) studied the estimation of dietary fiber and linoleic acid amounts in 31 food products and compared them to analytically measured values. The authors compared three different nutritionists using three different methods each: (1) the common trial-and-error method, (2) linear programming, and (3) quadratic programming. Since they considered all three methods to be highly manual, they were mainly interested in whether optimization would yield an equal or more accurate estimation than the purely manual and expertise-dependent trial-and-error estimation. The authors found the trial-and-error method to be less accurate than optimization methods at estimating dietary fiber. However, only in terms of absolute error did they find no difference in terms of estimated linoleic acid. The authors observed a significantly faster computation when using the optimization techniques. Ng et al. (2015) proposed a large-scale analysis using linear programming to estimate added sugars in US-offered beverages in 2007 and 2008. Based on an estimation of intrinsic sugars derived from a beverage's nutrition facts label, the authors deduced the amount of added sugar. The authors highlighted the necessity for a differentiated treatment and discussion of estimation errors potentially suggesting the need for an error measure that is relative to the amount of an ingredient or nutrient, respectively. Lamarine et al. (2018) mapped food diaries with fuzzy string matching to FCTs. The authors reported an accuracy of 89% when mapping ready-to-eat products. We applied a similar approach in matching ingredients to the FCT but found the more general ingredient naming requiring a stricter matching to achieve sufficient accuracy.

In conclusion, in our study, we used a large dataset spanning several food categories to systematically evaluate the estimation of ingredient composition. Previously, methods were highly manual, requiring additional inputs from an expert nutritionist, such as which ingredients to exclude from the calculation, where to allow tolerances and how much. Our proposed approach is fully automated using parser technology to separate ingredient declarations and string-matching methods to find the most suitable entry within an FCT. An expert was employed one time to define general mappings that could not be performed automatically, and these mappings will benefit all further analyses. However, ingredient mapping will remain a crucial task to achieve further automation.

6. Conclusions

In this paper, we demonstrated that by using mandatory food label information and an automated optimization approach, we were able to calculate the ingredient composition of common food products. In a comparison of our estimated ingredient compositions with actual recipes from different producers, we observed an average estimation error of 2.7% per ingredient. Via automation, large quantities of products can thus be evaluated in a short time. We found that by increasing the number of calculated Big7 values that fit within allowed ranges with the labeled Big7 values, the predictability of ingredient composition increased. The approach is still limited and does not deliver satisfactory results for food that undergoes special processing, such as cheese, where utilizing the declared fat content or processing grade could help to overcome current high estimation tolerances. In its current form, our approach and additional databases quantify up to 139 nutrients per food item and corresponding environmental impacts (cp. (Meier et al., 2021) and is thereby suitable for different use cases, e.g, (1) to allow for a more comprehensive and broadened nutrient tracking – on an individual basis or in prospective cohort studies, (2) to monitor the progress of food reformulation policies beyond Big7 nutrients (Bundesministerium für Ernährung und Landwirtschaft, 2022) and by using a broader range of nutrients to allow for a more sophisticated public health evaluation, (3) to support food manufactures with additional information to optimize health and environmental profiles of their food products, and (4) to develop specific food products, which are in particular suitable for previously or chronically diseased people that need a specific diet regime, e.g., cardiovascular diseases, chronic kidney disease, or diabetes type II.

Funding

Our project was funded by the Deutsche Bundesstiftung Umwelt (DBU) grant: 33981/01.

CRediT authorship contribution statement

Conceptualization: KB, PM; Methodology: KB, PM; Software: KB, MA; Data acquisition, Curation: KB, MA, FF, TM; Investigation: KB; Visualization: KB, PM; Writing – original draft: KB, PM, GS; Writing - review & editing: KB, PM, GS, TM, MA, FF; Project administration: KB, PM, TM; Funding acquisition: KB, PM, TM.

Conflict of Interest Statement

The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

Acknowledgments

We would especially thank all the students from Martin Luther University Halle-Wittenberg performing product data acquisition.

References

Afshin, A., Sur, P.J., Fay, K.A., Cornaby, L., Ferrara, G., Salama, J.S., Mullany, E.C., Abate, K.H., Abbafati, C., Abebe, Z., Afarideh, M., Aggarwal, A., Agrawal, S., Akinyemiju, T., Alahdab, F., Bacha, U., Bachman, V.F., Badali, H., Badawi, A., Bensenor, I.M., Bernabe, E., Biadgilign, S.K.K., Biryukov, S.H., Cahill, L.E., Carrero, J.J., Cercy, K.M., Dandona, L., Dandona, R., Dang, A.K., Degefa, M.G., El Sayed Zaki, M., Esteghamati, A., Esteghamati, S., Fanzo, J., Farinha, C.S.e.S., Farvid, M.S., Farzadfar, F., Feigin, V.L., Fernandes, J.C., Flor, L.S., Foigt, N.A., Forouzanfar, M.H., Ganji, M., Geleijnse, J.M., Gillum, R.F., Goulart, A.C., Grosso, G., Guessous, I., Hamidi, S., Hankey, G.J., Harikrishnan, S., Hassen, H.Y., Hay, S.I., Hoang, C.L., Horino, M., Islami, F., Jackson, M.D., James, S.L., Johansson, L., Jonas, J.B., Kasaeian, A., Khader, Y.S., Khalil, I.A., Khang, Y.H., Kimokoti, R.W., Kokubo, Y., Kumar, G.A., Lallukka, T., Lopez, A.D., Lorkowski, S., Lotufo, P.A., Lozano, R., Malekzadeh, R., März, W., Meier, T., Melaku, Y.A., Mendoza, W., Mensink, G.B.M., Micha, R., Miller, T.R., Mirarefin, M., Mohan, V., Mokdad, A.H., Mozaffarian, D., Nagel, G., Naghavi, M., Nguyen, C.T., Nixon, M.R., Ong, K.L., Pereira, D.M., Poustchi, H., Qorbani, M., Rai, R.K., Razo-García, C., Rehm, C.D., Rivera, J.A., Rodríguez-Ramírez, S., Roshandel, G., Roth, G.A., Sanabria, J., Sánchez-Pimienta, T.G., Sartorius, B., Schmidhuber, J., Schutte, A.E., Sepanlou, S.G., Shin, M. J., Sorensen, R.J.D., Springmann, M., Szponar, L., Thorne-Lyman, A.L., Thrift, A.G., Touvier, M., Tran, B.X., Tyrovolas, S., Ukwaja, K.N., Ullah, I., Uthman, O.A., Vaezghasemi, M., Vasankari, T.J., Vollset, S.E., Vos, T., Vu, G.T., Vu, L.G., Weiderpass, E., Werdecker, A., Wijeratne, T., Willett, W.C., Wu, J.H., Xu, G., Yonemoto, N., Yu, C., Murray, C.J.L., 2019. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet 393, 1958–1972. https://doi.org/10.1016/S0140-6736(19)30041-8.

Ben-David, M., Hanley, T.A., Klein, D.R., Schell, D.M., 1997. Seasonal changes in diets of coastal and riverine mink: the role of spawning pacific salmon. Can. J. Zool. 75, 803–811. https://doi.org/10.1139/z97-102. ⟨https://doi.org/10.1139/z97-102⟩.

Bommer, C., Heesemann, E., Sagalova, V., Manne-Goehler, J., Atun, R., Bärnighausen, T., Vollmer, S., 2017. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. Lancet Diabetes Endocrinol. 5, 423–430. https://doi.org/10.1016/S2213-8587(17)30097-9.

European Commission, 2012b. Regulation (eu) no 1169/2011 of the European Parliament and of the council of 25 October 2011: Guidance with regard to the setting of tolerances for nutrient values declared on a label.

Bratzke, F., Ritschel, F., Wache, R., et al., 2018. Market analysis of potentially cardioprotective foods in context of legal health and nutrition claims. focus: meat, dairy and egg products. Ernahrungs Umschau 65, 2–11.

Dantzig, G.B., Thapa, M.N., 1997. Linear Programming 1. chapter Introduction. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York.

European Commission, 2011.Regulation (eu) no 1169/2011 of the European Parliament and of the council of 25 october 2011 on the provision of food information to consumers.Official Journal of the European Union 54, 18–61.

European Commission, 2012a. Establishing a list of permitted health claims made on foods, other than those referring to the reduction of disease risk and to childrenas development and health.Official Journal of the European Union.

Bundesministerium für Ernährung und Landwirtschaft, 2022. Das Produktmonitoring zur nationalen Reduktions- und Innovationsstrategie. ⟨https://www.bmel.de/DE/themen/ernaehrung/gesunde-ernaehrung/reduktionsstrategie/reduktionsstrategie-produktmonitoring.html⟩.

GS1 Germany. Global data synchronization network. ⟨https://www.gs1.org/services/gdsn/global-data-model⟩. Accessed: 2021–05-02.

Hartmann, B., Schmidt, C., Sandfuchs, K., 2014. Bundeslebensmittelüssel (bls) version 3.02.Max Rubner-Institut-Bundesforschungsinstitut für Ernährung und Lebensmittel.

Lamarine, M., Hager, J., Saris, W.H.M., Astrup, A., Valsesia, A., 2018. Fast and accurate approaches for large-scale, automated mapping of food diaries on food composition tables. Front. Nutr. 5, 38. https://doi.org/10.3389/fnut.2018.00038.

Lazzari, G., Jaquet, Y., Kebaili, D.J., Symul, L., Salathé, M., 2018. Foodrepo: an open food repository of barcoded food products. Front. Nutr. 5, 57. https://doi.org/10.3389/fnut.2018.00057.

Levenshtein, V.I., 1965. Binary codes capable of correcting deletions, insertions, and reversals. Dokl. Akad. Nauk SSSR 163 (4), 845–848.

Marcoe, K., Haytowitz, D., 1993. Estimating nutrient values of mixed dishes from label information. Food Technol. 47, 69–75.

Meier, T., Senftleben, K., Deumelandt, P., Christen, O., Riedel, K., Langer, M., 2015. Healthcare costs associated with an adequate intake of sugars, salt and saturated fat in germany: a health econometrical analysis. PloS One 10.

Meier, T., Deumelandt, P., Christen, O., Stangl, G., Riedel, K., Langer, M., 2017. Global burden of sugar-related dental diseases in 168 countries and corresponding health care costs. J. Dental Res. 96, 845–854.

Meier, T., vonBorstel, T., Welte, B., Hogan, B., Finn, S.M., Bonaventura, M., Friedrich, S., Weber, K., Dräger de Teran, T., 2021. Food waste in healthcare, business and hospitality catering: composition, environmental impacts and reduction potential on company and national levels. Sustainability 13. https://doi.org/10.3390/su13063288.

Monteiro, C.A., Cannon, G., Levy, R.B., Moubarac, J.C., Louzada, M.L., Rauber, F., Khandpur, N., Cediel, G., Neri, D., Martinez-Steele, E., Baraldi, L.G., Jaime, P.C., 2019. Ultra-processed foods: what they are and how to identify them. Public Health Nutr. 22, 936–941. https://doi.org/10.1017/S1368980018003762.

Ng, S.W., Bricker, G., Li, K.p., Yoon, E.F., Kang, J., Westrich, B., 2015. Estimating added sugars in us consumer packaged goods: An application to beverages in 2007-08. J. Food Comp. Anal. 43, 7–17. https://doi.org/10.1016/j.jfca.2015.04.004.

Open Food Facts association. Open food facts. Available: ⟨https://world.openfoodfacts.org⟩. Accessed: 2020–09-02.

Parr, T., 2020. Antlr - another tool for language recognition (version 4). Available: ⟨https://www.antlr.org⟩. Accessed: 2020–10-11.

Soetaert, K., Van den Meersche, K., van Oevelen, D., 2009a. limsolve: Solving linear inverse models. R package 1.5.1.

Soetaert, K., Van den Meersche, K., van Oevelen, D., 2009b. Package limsolve, solving linear inverse models in R.

Tremmel, M., Gerdtham, U.G., Nilsson, P.M., Saha, S., 2017. Economic burden of obesity: a systematic literature review. Int. J. Environ. Res. Public Health 14, 435.

U.S.Department of Agriculture, Agricultural Research Service, 2019. Fooddata central. Available: ⟨https://fdc.nal.usda.gov⟩.Accessed: 2020–010-04.

van Dooren, C., 2018. A review of the use of linear programming to optimize diets, nutritiously, economically and environmentally, 48-48 Front. Nutr. 5 (48). https://doi.org/10.3389/fnut.2018.00048.

Verbraucherzentrale Bundesverband e.V, Team Lebensmittel,. nachhaltiger lebensmittelkonsum - von der nische in die breite. Online:⟨https://www.vzbv.de/pressemitteilungen/nachhaltiger-konsum-keine-einseitige-verantwortung-der-verbraucher⟩.Accessed:2019–01-31 2019.

Westrich, B.J., Buzzard, I.M., Gatewood, L.C., McGovern, P.G., 1994. Accuracy and efficiency of estimating nutrient values in commercial food products using mathematical optimization. J. Food Comp. Anal. 7, 223–239. https://doi.org/10.1006/jfca.1994.1026.

Wobker, I., Eberhardt, T., Kenning, P., 2015. Consumer confusion in german food retailing: the moderating role of trust. Int. J. Retail Distrib. Manag.

World Health Organization, 2019. Essential nutrition actions: mainstreaming nutrition through the life-course.