Discovery and characterisation of dietary patterns in two Nordic countries. Using non-supervised and supervised multivariate statistical techniques to analyse dietary survey data

Edberg, Anna; Freyhult, Eva; Sand, Salomon; Fagt, Sisse; Knudsen, Vibeke Kildegaard; Frost Andersen, Lene; Lindroos, Anna Karin; Soeria-Atmadja, Daniel Soeria; Gustafsson , Mats G.; Hammerling, Ulf

Link to article, DOI:
http://dx.doi.org/10.6027/TN2013-548

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Discovery and characterisation of dietary patterns in two Nordic countries

Using non-supervised and supervised multivariate statistical techniques to analyse dietary survey data


norden

Discovery and characterisation of dietary patterns in two Nordic countries

Using non-supervised and supervised multivariate statistical techniques to analyse dietary survey data

Anna Edberg,[1,$] Eva Freyhult,[2,3,4] Salomon Sand,[5] Sisse Fagt,[6] Vibeke Kildegaard Knudsen,[6] Lene Frost Andersen,[7] Anna Karin Lindroos,[1] Daniel Soeria-Atmadja,[2] Mats G. Gustafsson[2] and Ulf Hammerling[2]

TemaNord 2013:548

[1] Division of Food Data, Dept of Science, National Food Agency, SE-75126 Uppsala, Sweden
[2] Cancer Pharmacology and Computational Medicine, Dept of Medical Sciences, Uppsala University, Uppsala Academic Hospital, SE-75185 Uppsala, Sweden
[3] Bioinformatics Infrastructure for Life Sciences (BILS)
[4] Science for Life Laboratory, Uppsala
[5] Dept of Risk and Benefit Assessment, National Food Agency, SE-75126 Uppsala, Sweden
[6] Dept of Nutrition, National Food Institute, Technical University of Denmark, DK-2860 Søborg, Denmark
[7] Dept of Nutrition, University of Oslo, NO-0316, Norway
[$] Present affiliation: Råd & Rön, P.O. Box 38001, SE-10064 Stockholm, Sweden

Discovery and characterisation of dietary patterns in two Nordic countries
Using non-supervised and supervised multivariate statistical techniques to analyse dietary survey data

Anna Edberg, Eva Freyhult, Salomon Sand, Sisse Fagt, Vibeke Kildegaard Knudsen, Lene Frost Andersen, Anna Karin Lindroos, Daniel Soeria-Atmadja, Mats G. Gustafsson and Ulf Hammerling

ISBN 978-92-893-2581-3
http://dx.doi.org/10.6027/TN2013-548
TemaNord 2013:548

Layout: Hanne Lebech
Cover photo: ImageSelect

Print: Rosendahls-Schultz Grafisk
Copies: 76

Printed in Denmark

This publication has been published with financial support by the Nordic Council of Ministers. However, the contents of this publication do not necessarily reflect the views, policies or recommendations of the Nordic Council of Ministers.

www.norden.org/en/publications

Nordic co-operation
Nordic co-operation is one of the world's most extensive forms of regional collaboration, involving Denmark, Finland, Iceland, Norway, Sweden, and the Faroe Islands, Greenland, and Åland.

Nordic co-operation has firm traditions in politics, the economy, and culture. It plays an important role in European and international collaboration, and aims at creating a strong Nordic community in a strong Europe.

Nordic co-operation seeks to safeguard Nordic and regional interests and principles in the global community. Common Nordic values help the region solidify its position as one of the world's most innovative and competitive.

Nordic Council of Ministers
Ved Stranden 18
DK-1061 Copenhagen K
Phone (+45) 3396 0200

www.norden.org

Content

Preface

Multivariate data analysis (MDA) has an established and acknowledged position in diverse fundamental and applied sciences, especially those of engineering but likewise various biology and biomedical disciplines. Since decades back MDA is an integral part of nutritional epidemiology and has, more recently, even found inroads to dietary surveys. Nonetheless, the latter field has hitherto witnessed but scattered MDA application and very little, if any, of advanced and exploratory nature. This situation contrasts sharply to the quite extensive accumulated resources allocated to such surveys globally, including those accomplished in the Nordic countries. Actually, dietary surveys – commonly conducted periodically within the respective national Nordic food regulatory agencies – are typically designed to capture eating habits at comparatively high levels of differentiation, thereby comprising rich data sets. Without taking advantage of appropriate computational technology, however, much information embedded in these compilations will remain curtailed. Consequently, a more regular implementation of relevant MDA within the dietary survey area will undoubtedly render intricate eating patterns, within and across countries, accessible to inspection and construal.

A truly multidisciplinary project team of experts from Norway, Denmark and Sweden was thus created to embark on the application of a carefully selected array of MDA technologies to the interrogation of dietary survey data of two Nordic countries. The team encapsulated expertise in nutrition and toxicology as well as – with a special relevance to the core incentive of this undertaking – proficiency in engineering and computational sciences. Throughout the course of work within the project – *Discovery and characterisation of dietary patterns in two Nordic countries by means of multivariate data analysis* – several diverse computational techniques were applied to data of two national polls: an excerpt of the *Danish National Survey of Diet and Physical Activity* as well as the entire (Swedish) *Riksmaten – barn 2003*.

Broadly speaking, the techniques applied here largely fall in two distinct categories: pattern recognition and predictive modelling, the latter type – also referred to as machine learning – being new to the field. A wealth of intriguing relationships was accordingly identified, of which the most pertinent are further elaborated on in a discussion section.

Moreover, the report places data, methods and findings alike in a contextual framework, aiming at supporting the reader with a view of both dietary surveys and essentials of the computational techniques used. Hopefully, this report will entice to more exhaustive computational analysis of dietary survey data, relative to that of contemporary praxis, thereby enabling the disclosure of important but entrenched patterns and consequently taking much better advantage of study investment.

Abbreviations

- MDA – Multivariate Data Analysis
- OMB – DHC Omnibus Multi-Branching Divisive Hierarchical Clustering
- PCA – Principal Component Analysis
- HPBCA – Hierarchical Prototype Bi-Cluster Analysis
- CMDS – Classical Multi-Dimensional Scaling
- RF – Random Forest
- NSC – Nearest Shrunken Centroid
- RRR – Reduced Rank Regression
- MM – Mixture Models

Executive summary

The study outlined in this report strived at disclosing pertinent patterns in dietary surveys by means of an array of multivariate data analysis (MDA) techniques. The overall purpose was thus to unveil embedded patterns in selected data material, but also to generally demonstrate feasibility of new computational technology in this area. The material selected for this purpose encompasses food consumption survey data from Sweden and Denmark. The first among those compilations is known as *Riksmaten – barn 2003*, harbouring children of three age groups (four, eight and eleven years of age), whereas the latter data set is an excerpt – holding preschool children (four to five years of age) – of the *Danish National Survey of Diet and Physical Activity*, compiled over several years until 2008. These sets of food consumption data have previously been subjected to classical statistical analysis, but were – prior to embarking on this exercise – devoid of scrutiny by means of more advanced computational techniques. The analytical exercises described in this report encompass two major fields of MDA, which can be summarised as Unsupervised Learning/Descriptive modelling, on the one hand, and Supervised Learning/Predictive Modelling, on the other.

The first among the unsupervised analyses involved inspection largely by, but not restricted to, an in-house implemented multi-branching hierarchical clustering algorithm (OMB-DHC), thereby revealing various aggregations of reasonably coherent consumers in unabridged and age-defined sub-populations. Notably, a hierarchical OMB-DHC design of operation tied to a palatable output display, unlike earlier reports in the dietary survey area, helped identifying the degree of heterogeneity of clusters appearing at several segregation levels, thereby also supporting the judicious selection of aggregations for further compilation and scrutiny. Numbers and salient features of such dietary sub-populations were found to largely, but not exactly, commensurate with those of various scientific reports in the area. Thus, 4–5 dietary clusters – in this report also referred to as dietary prototypes – emerged from our data sets at the highest hierarchical level and three among them – *Traditional*, *Soft beverages/Buns & cakes* and *Varied (healthy)* – roughly match those commonly reported elsewhere. Accordingly identified aggregations underwent further processing, i.e. the prototypes were used as input to

either of two distinct downstream (of OMB-DHC) clustering algorithms. The first among these composite procedures, here designated *Hierarchical Prototype Bi-Cluster Analysis* (HPBCA), enabled creation of an indeed very instructive two-dimensional display of pertinent dissimilarities between Danish and Swedish age-matched consumption data as well as across the Swedish preschool and elementary school consumers. As anticipated, overall dietary patterns of the two oldest age categories of *Riksmaten – barn 2003* were mutually closer, relative to those of four-year old children. More intriguingly, however, the analysis revealed rather drastic disparity between consumption patterns of Danish and Swedish preschool children. The second composite technique, here referred to as *Dietary Prototype CMDS Analysis* (DPCA), enabled the delineation and visualization of multidimensional distances across the various dietary prototypes and thus helped identifying overarching interrelationships between aggregated consumer groups. Furthermore, Principal Component Analysis (PCA) provided support to the hierarchical cluster analysis so as to explain major direct and inverse relationships between key food groups in the several intra- and inter-national data excerpts. For example, major PCA loadings helped deciphering both shared and disparate features, relating to food groups, across Danish and Swedish preschool consumers.

Data interrogation, reliant on the above-mentioned composite techniques, disclosed one outlier dietary prototype in each of the two Swedish elementary school children data subsets. This pair of group-detached prototypes showed, however, notable mutual resemblance and featured consumption of low-fat foods (largely with respect to dairy products) and besides quite healthy eating patterns. Moreover, these exercises unveiled another set of interrelated dietary prototypes, one in each of all Swedish age categories, but mutually most similar in the two older age groups. Common features are relatively low intake of *Vegetables* and *Fruit & berries* likewise fairly high consumption of *Soft beverages (sweetened)*. A dietary prototype with the latter property was identified also in the Danish data material, but without low consumption of *Vegetables* or *Fruit & berries*.

The second MDA-type of data interrogation involved Supervised Learning, also known as Predictive Modelling. These exercises involved the Random Forest (RF) and Nearest Shrunken Centroid (NSC) classification algorithms. Briefly, collections of classifiers were created to predict low and high consumers of each among a wide excerpt of food groups, subsequent to elimination of that particular food. Frequency histograms of the remaining foods (in each case) were accordingly de-

rived from these elaborations, displaying patterns of key food groups that thus jointly are indicative of discriminating such bi-partite (low/high) categories, in the absence of the targeted (outstanding) food. Very instructing displays of deeply embedded relationships inherent to the survey data emerged from these procedures, in many cases also enhancing findings derived from the unsupervised MDA work. Actually, intriguing frequency pattern similarities and discrepancies were also seen across the respective national consumption data subsets among preschool children. For example, *Potato* is firmly connected with *Rice* in the Danish data set, but rather associated with *Sausage* and *Fish* in that of Sweden. Unlike Swedish preschool children, who show tight linkage between *Bread* and both *Cheese* and *Cereals*, Danish age-matched consumers of *Bread* are tethered to *Sugar* (marmalade) and *Vegetables*. Marked trans-national disparity was also seen in dietary habits associated with *Milk* and *Meat & poultry*.

Some overarching observations are: i) certain healthy and less healthy foods tend to appear in disjoint clusters, ii) two (mutually similar and relatively prudent) dietary prototypes, one in each of the two Swedish elementary school consumer data sets, appear quite remote from those of the remaining age-matched consumers, iii) Danish and Swedish preschool consumers show notable trans-national disparity, for example the *Milk* food group as well as that of *Bread* are tethered to quite distinct (nationality-specific) consumption patterns, iv) among the several dietary prototypes identified across the trans-national data set, including age-matched excerpts of Swedish data, prototypes with the shared feature of being high in the *Soft beverages (sweetened)* food group emerged, and v) although not elaborated on in-depth, output from several analyses suggests a preference for energy-based consumption data for Cluster Analysis and Predictive Modelling, over those appearing as weight.

1. Introduction/background

1.1 Dietary surveys – consumption patterns and health

In Western societies, there is – since decades back – a trend towards higher caloric intake and overweight, but this predicament is increasingly common also in developing countries.[1] Especially in tandem with a sedentary lifestyle, overweight and obesity predisposes to an array of cardiovascular disorders, type II diabetes and other ailments.[2,3] Notably, the metabolic syndrome – a constellation of cardiovascular disease risk factors, e.g insulin resistance, abdominal obesity, hypertension and atherogenic dyslipidemia – shows high dietary association.[4–6] Although many factors, including those beyond dietary habits, seemingly contribute to this unhealthy condition, it is worth mentioning that the global increase of sugar-sweetened beverages, as seen over the past several decades, has been firmly associated with mounted risk of developing both the metabolic syndrome and overt type 2 diabetes mellitus.[7] Conversely, high intake of whole grain is connected with reduced risk of developing glucose tolerance typical of pre-diabetic conditions.[8] Moreover, dietary habits can also contribute to the risk of contracting colorectal malignancy.[9,10] Actually, there is concern among health professionals of a likely connection between high consumption of red and processed meat and colorectal carcinoma incidence, although no mechanistic model as yet has been demonstrated.[11–14] Nutritional deficiencies – typified by those tied to an array of micronutrients such as (pro-) vitamins, essential fatty acids, iodine and selenium – are, naturally, equally important to confront.[15] However, certain diets, e.g. those rich in vegetables, fruits, beans, whole-grain cereals, olive oil and certain fish species, especially in relation to typical consumption in many industrialised areas, have proven able to promote cardiovascular and overall health in the population.[16–19] Notably, adherence to the Mediterranean diet, or an appropriate geographical derivative, seems conducive to good health status.[20–22] Thus, it is imperative to create a supporting atmosphere for healthy eating, as vividly outlined in a review on prospects for governmental bodies to promoting healthy food and eating environments.[23]

Collective determinants of eating behaviour include a broad range of contextual factors which are, however, only partially understood.[24]

Food consumption surveys are conducted on a regular basis in many developed countries. The overall purpose of such exercises is to acquire knowledge on food habits among the general national population, including various key segments thereof, e.g. women, men, children, elderly, educational and/or otherwise social categories etc. Nutritional investigations conducted in Sweden and elsewhere are typically designed as either a 24-h recall or a 4-d dietary record scheme, "The European Prospective Investigation into Cancer and Nutrition" (EPIC) being an example of the former type.[25] Other schemes have, however, also found application. Notably, guidelines developed by the EFSA Expert Group on Food Consumption Data advocates that surveys should cover two non-consecutive days and use the dietary record method for infants and children and a 24-h recall method for adults.[26] Part of the Safe Foods project, funded by the European Commission's 7th Research Programme, was devoted to harmonisation of European national food consumption surveys.[27] A core part of this initiative involved the conversion of Food and Ingredients to Raw Agriculture Commodities, aiming at enabling the accurate determination of exposures to food-borne toxicants and toxins, as accomplished by the Monte Carlo Risk Assessment procedure (MCRA). Apart from the several European repositories on dietary intakes the US Dept of Agriculture has compiled a database on consumption habits, which is freely available for download.[28] Data on eating habits also provide an insight in a population's adherence to current knowledge on healthy food consumption.

Modern approaches to data analysis are instrumental to the identification of pertinent dietary patterns from available data records, regardless of whether such data are directly associated with health/disease parameters. The traditional food-based approach, oriented towards single foods and nutrients, has lost much credence over the past several years as a basis for the assessment of relationships between diet and health/disease, but much research in the nutrition area is still centred on specific foods.[29] Although recent advancement in the area has shifted the gravity some distance from single foods towards dietary patterns more advanced MDA, i.e. scrutiny beyond regular cluster and principal components analysis, has as yet found but little application within the dietary survey area.

1.2 Established and emerging statistical methodology as applied to food consumption data

Data derived from dietary surveys, on the premise of prudent questionnaire design, are helpful to several specific applications. Notably, such data can embody rich information about the food and occasionally also eating environments and are thus instrumental to dietary guidelines and even policy intervention.[23,30] Moreover, these surveys are also essential to exposure assessment of food contaminants, thereby representing a key requisite in risk assessment of regulated and non-regulated toxins/toxicants occurring in foods. Actually, unwanted substances of various sorts attached to foods are perceived as a potential health risk and their respective contribution to potential hazard is subject to considerable research.[31,32] Such substances include natural and anthropogenic contaminants, typified by various mycotoxins and an array of organohalogens, respectively.[33,34]

Clearly, adequately defined and reasonably well demarcated dietary clusters would principally open prospects for cluster specific risk assessment involving any sort of environmental contaminant, pesticide or otherwise unhealthy food-borne items. Moreover, identified dietary prototypes might facilitate planning, design and interpretation of future experimental nutritional studies, e.g. by enabling selection of volunteer individuals from within anyone or several among the accordingly defined categories. It is our notion that such information holds strong promise to supporting for example physiological mapping of selected consumer populations. Employed as prior knowledge, it would help designing and enhancing output value of experimental interrogation studies, such as quantitative determination of physiological components in blood sera using either a set of clinical chemistry parameters or broad profiling campaigns, i.e. metabolomic or proteomic scanning. Lastly, dissecting key consumption patterns among sub-populations featuring either very low or markedly high intake of a targeted food group would yield more specific dietary prototypes of outstanding interest when designing new nutritional studies as well as for epidemiological investigations. Several additional applications can be envisaged, including a new approach to compare food consumption survey data across time and region.

1.3 Applying multivariate data analysis to unveil dietary patterns

Traditionally, statistical analysis of food consumption data has been focused on intake levels of single nutrients or foods at a time, i.e. using classical univariate statistical analyses. This approach is, however, not quite appropriate for revealing consumption patterns across several dimensions, the latter direction being more aligned with the accepted concept that people eat foods rather than nutrients. Thus, the last decade has witnessed a gradual transition in research from a nutrient to a food level, translating to an emphasis on dietary patterns.[29,35,36]

MDA techniques are specifically designed to identify meaningful patterns in complex multi-dimensional data sets. This framework is thus suitable to deciphering embedded features, such as aggregations and correlations, within dietary patterns, either in isolation or merged with various health outcomes. Obviously, since foods are almost invariably consumed together, the concurrent analysis of many foods/food categories simultaneously is instrumental to the adequate identification of major and more subtle dietary prototype structures in data masses. Indeed, MDA is simply a first choice technique for this purpose. A few seminal papers on food patterns appeared already in the 1980s and 1990s, but the last decade has witnessed considerable growth in the popularity of MDA in nutritional epidemiology investigations.[35–39] Typically, patterns of food intake are discovered by means of exploratory dimension-reducing techniques, which fall into two separate categories: Cluster Analysis, on the one hand, and Factor Analysis – typically appearing as Principal Component Analysis (PCA) – on the other.

Cluster Analysis assigns consumers into discrete assemblies, each featuring a reasonably coherent eating pattern, while PCA allocating foods/nutrients into patterns based on their inter-correlations, as emerging from underlying latent variables (factors). More specifically, Cluster Analysis involves the reduction of dietary data patterns, based on individual differences, in mean dietary intakes. Although certain special algorithms allow entities to appear in several clusters, most techniques create mutually exclusive (dichotomous) patterns, i.e. each subject can belong to only one aggregation. Among the various available cluster algorithms the non-hierarchical k-means clustering has found wide application in nutrition research.[35,40] In PCA, orthogonal vectors designed to point in the directions of maximal variance are derived via the covariance matrix of the dataset employed. These vectors are commonly referred to as principal components or loading vectors and are

used to represent each cohort subject as a point in the new coordinate system, as defined by these vectors. Thus, each cohort subject is represented by coordinate values (scores) in this space, which usually is of relatively low dimension compared to the original space. Studies that apply both of these quite disparate statistical techniques in parallel are also reported.[41–44] Concisely, cluster analysis helps identifying distinct aggregations in the data, i.e. mostly consumers with similar features, whereas PCA provids deepened insight in relationships between the various food groups.

In supervised learning (SL), prediction models are designed based on observed examples that include a response variable that may take either a discrete or a continuous value. Depending on the kind of response variable, discrete or continuous, the prediction models obtained in MDA are called classification models and regression models, respectively. Common model families employed within SL are Linear Models, Regression and Classification Trees, Artificial Neural Networks, Support Vector Machines, Prototype Models (e.g. Nearest Neighbour), Random Forests and Nearest Shrunken Centroid.

Seemingly, nutritional epidemiology, let alone the more specialised area of dietary surveys, rarely incorporates SL applications, including those oriented towards classification issues. For linear regression problems, though, the Reduced Rank Regression (RRR) method has lately found application in nutrition science. A typical application of this algorithm in a nutritional epidemiology context involves assignment of food groups as predictors and an *ad hoc* selection of nutrients as responses.[45] RRR has proven useful in describing dietary patterns and their association with health and disease.[46,47] However, one should note that RRR is limited in the sense that it only allows design of linear regression models and is therefore not at all suitable when the functional relationship to model between the input variables and the response/target is non-linear. Another set of methods, collectively referred to as Mixture Models (MMs), have lately also found usage in nutritional epidemiology. MMs can help identifying latent sub-populations within an overall population without a need for data to directly recognise the underlying categories or to which an entity belongs. Certain variants of MMs can be perceived as clustering procedures. Intricate dietary patterns within a large fraction of the EPIC cohort, by means of a variant Finite Mixture Model, have recently been reported.[48] Moreover, Latent Class Analysis – closely related to Finite Mixture Model – identifies unobservable subgroups within a population typically based on categorical latent variables, but can also accommodate continuous data. Constructs are subsequently used

for regression analysis.[49,50] By means of Latent Class Analysis dietary prototypes were identified in cohorts of modest and large sizes.[51,52] Although MMs often have been used for exploratory (unsupervised) data analysis, the general MM framework is also applicable to either regression or classification problems. The underlying MM principle to fit/model mixture distributions, however, is well known to be an ill-posed problem for high dimensional data and this fact severely limits the applicability of MM to this kind of data sets.

In the work presented in this report, two MDA methods based on Statistical Learning for design of classification models were employed. These undertakings were based on implementations in R[53] and involved the classification algorithms Random Forest (RF) (implemented in the R package randomForest)[54] and Nearest Shrunken Centroid (NSC) (implemented in the R package pamr).[55] The RF technique allows the creation of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of input values.[54] Actually, RF is proven highly versatile and has demonstrated high classification accuracy in numerous cases.[56,57] Because of its popularity in many areas special and/or enhanced varieties of RF have appeared.[58–60] The second of the two families of predictive modelling, brought into play in the study outlined here, is known as Nearest Shrunken Centroid (NSC) classifier.[55] It is widely appreciated due to its relatively low computational complexity during design/learning and execution as well as its inherent ability to perform variable subset selection, as part of the design/learning procedure.

This report outlines and discusses findings, as revealed by applying MDA to food consumption data, in a Swedish national dietary survey of preschool (4-years) and elementary school (8 and 11 years old) children as well as a Danish survey encompassing 4–5 year old subjects. This undertaking was largely based on clustering, PCA, HPBCA and CMDS (see Appendix C) but also involved predictive modelling in the form of classifiers designed to predict high and low consumers of a particular food group based on an input pattern consisting of the consumption levels of all other food groups. According to the literature available, this seems to be the first time this kind of technique has been applied to food consumption data.

2. Aims and overall design of the study outlined in this report

The overall objective of the study reported here was to apply modern exploratory and prediction based MDA techniques to data on food consumption habits in order to glean deepened insight in various intra- and inter-population relationships within data sets formerly processed by classical statistical methods only. Thus, we wanted to unveil and decipher embedded multidimensional properties among and within young consumers that cannot be found using conventional univariate analyses.

Food consumption data, appearing as two overarching sets, were addressed in this study. The first among those is derived from a national food consumption survey, *Riksmaten – barn 2003*, conducted in 2003 and encompassing Swedish children, 4, 8 and 11 years old.[61] The second set involves consumption data on Danish 4–5 year old children, based on an excerpt (completed in 2008) of the *Danish National Survey of Diet and Physical Activity*. To enable trans-national consistency in statistical data interrogation the Danish data set and its Swedish counterpart (both sets encapsulating preschool children only) were harmonised to both encompass 25 food groups, rather than 35 as applied by default to the Swedish age classes. Dietary patterns observed in the Swedish data material disclosed several shared outstanding food groups, but global dietary habits were, as anticipated, not consistent across age groups. Clearly, the consumption pattern of 4-year old children was most remote, relative to those of older age. Intriguing dissimilarities between Danish and Swedish consumption patterns were found, which were subjected to deepened inspection by means of more advanced techniques, i.e. Predictive Modelling.

Hierarchical Clustering Analysis (HCA) and two distinct bi-clustering techniques were applied to the data sets in order to separate children into distinct populations, and to reveal pertinent pattern relationships, according to dietary habits. PCA was performed largely to complement cluster analysis, i.e. to retrieve information about the food groups contributing to the largest variations within the data set and to identify distinctive patterns among them. Cluster Analysis and PCA were performed on the entire Swedish data set, on different age groups separately as well

as on the Danish data set. Thus, these exploratory techniques were applied on both national data sets to identify similarities and differences in consumption patterns across Danish and Swedish subjects. Modern predictive modelling, i.e. RF and NSC classification models, were however leveraged to the deciphering of more deeply entrenched relationships. Briefly, these techniques were unleashed to identifying an embedded food consumption pattern tethered to a query food, among the remaining food groups, which thus – in many instances – enabled the successful discrimination between low and high intake consumers of the target food. Because the exploratory non-supervised techniques do not lend themselves well to disclosing such inherent patterns, supervised learning methods were deemed necessary. Indeed many additional insights were revealed indicating outstanding prospects for expanded application of predictive modelling strategies within this field.

3. Data sets

3.1 Sets of food data

Consumption data were from two distinct national surveys, as specified below:

- *Riksmaten – barn 2003 – Swedish Children* (4, 8 and 11 years old)
  - 4 years (590 children)
  - 8 years (890 children)
  - 11 years (1,016 children)
- *Danish National Survey of Dietary Habits and Physical Activity – Danish children* (age 4–5)
  - 4–5 years (122 4-year olds + 196 5-year olds = 318 children)

Riksmaten – barn 2003 is an open and estimated food diary across four consecutive days, conducted by the (Swedish) National Food Agency. Serving sizes and amounts of all foods and drinks consume by subjects were estimated by means of a picture book ("Livsmedels- och näringsintag bland barn i Sverige", *Riksmaten – barn 2003*).[61]

The Danish data set is based on data collected in 2000 through 2008, representing part of the *Danish National Survey of Diet and Physical Activity*, an ongoing project at the National Food Institute, Technical University of Denmark. The participants kept a food record for seven consecutive days in a pre-coded questionnaire with answering categories for the most commonly consumed foods and dishes in the Danish diet. The amounts of food eaten were given in household measures and estimated from different portion sizes.

3.2 Data pre-processing and uni-variate statistical analysis

Pre-processing and subsequent unsupervised statistical analyses were performed in the Matlab computer programming environment, version 7.11 (Mathworks, Inc., Natick, USA). To identify significant differences between food groups among the delineated dietary patterns non-

parametric two-sided Kruskal Wallis and Mann-Whitney U-test (performed by kruskalwallis() and ranksum() in Matlab) test were used for overall and pair wise comparisons, respectively.

For each analysis performed one of the following methods for standardization (normalisation) of the data was employed:

- Weight percent, i.e. consumption of a certain food group divided by the total consumption across all food groups.
- Energy percent, i.e. energy contribution of a certain food group divided by the total energy intake.
- Z-score normalisation of data by zscore in the Matlab environment, yielding all food group values to have zero mean and unit variance.
- Normalisation of each food group value by dividing with its arithmetic mean value.

Data were prepared for computational analysis by defining a suitable number of food group variables. Analyses of the Swedish consumption data, either in its entirety or within age-matched consumer groups, were based on 35 food groups as input variables to the various multivariate analyses. To enable adequate comparison of dietary patterns across Swedish (4 years old) and Danish children (4–5 years old), however, food groups of the Swedish preschool data were transformed into a reduced set (25), consistent with those inherent to the Danish data set. The rationale behind this variable reduction, which translates to more general food groups, is to avoid spurious identification of dietary patterns owing to dissimilarities in variable composition across the two national surveys. Analysis of a further reduced data set, wherein variables that describe the majority of all variation were excluded, was also conducted. More detailed information about food groups included in the variables appears in Appendix A.

Preparations for supervised MDA involved the following major operations: For each food group, subjects were partitioned into two categories: low and high consumers. Children were assigned to either a low- or a high consumer category, based on whether figures were below the 20^{th} percentile or above the 80^{th} percentile, respectively. Individuals with values in between were not included in the supervised MDA. Certain food groups were not attached to any consumption within a substantial fraction of the children; when the fraction of non-consumers was greater than 20% another definition of low and high consumers was adopted. The variables were dichotomised into users (intake greater than zero) and non-users (no intake). Hence, for these food groups all children were included in the supervised MDA.

4. Results

4.1 Swedish data set, all ages

4.1.1 Consumption profiles (food intake: mean relative weight)

Compliant with most studies of selected parts of the data material, e.g. defined by age or nation, an initial inspection of the unabridged Swedish dietary survey data encompassed top-down Cluster Analysis by means of the OMB-DHC algorithm, with each food group intake appearing as a weight fraction (percent) of the total intake. This provided a multi-branching display, which did not halt at any pre-defined hierarchical level, but proceeded until singlet entities were attained. The upper part of the accordingly derived dendrogram chart – built on the entire Swedish data set – discloses five main aggregations, also referred to as dietary prototypes in this report, with notable dissimilarity in size and complexity across them (Figure 1). As viewed from the highest hierarchical stage, cluster RA I (*Riksmaten – barn 2003*, All subjects, sub-population I), also labelled the *Cereals* cluster, is markedly small and relatively homogeneous. Actually, it is also most remotely situated, in relation to the remaining data (note edge lengths at the top level of Figure 1). The next assembly in line – RA II (the *Milk* cluster) – is largest of all and also features the highest heterogeneity, as revealed by many downstream segregation points. Aggregation RA III is also large, but slightly less scattered; the remaining two clusters – RA IV and RA V – show more intermediate properties with respect to size and diversion (Figure 1). The entire set of alternate cluster designations, as dictated by salient food group(s) (or other relevant pattern), appear as follows (RA I through RA V): i) *Cereals*, ii) *Milk* (low fat), iii) *Traditional*, iv) *Soft beverages (sweetened)/Buns & cakes* and v) *Varied/Water* (Figure 1 and Table 1). For clarity, outstanding food groups of each such top-level aggregation were depicted as bar charts, which can be perceived as concise graphical dietary prototypes (Figure 2). Notably, *Cereals* and *Milk* are pertinent features of clusters RA I and RA II, respectively. Cluster RA III shows a rather even distribution across many food groups, but is particularly rich in *Fruits*, *Juice*, *Milk (low-fat), Soft drinks (light)*, *Rice*, *Meat &*

poultry and *Desserts*. We have thus chosen to tentatively label this prototype *Traditional*. Conversely, Figure 2 highlights *Soft drinks (sweetened)* as a major property of prototype RA IV, but *Buns & cakes* as well as *Snacks* are additional salient features of this aggregation (Table 1). Clearly, *Milk* (RA II) and *Traditional* (RA III) dietary prototypes stand out as the two largest populations within the ensemble. Seemingly, the *Cereals* (RA I) prototype is mostly built of preschool children, whereas those of *Soft beverages (sweetened)/Buns & cakes* (RA IV) and *Varied/Water* (RA V) are largely restricted to elementary school (8- or 11 years of age) consumer categories (Figure 3A and B).

A further segmentation with regard to gender shows a slightly more even cluster size distribution among girls, compared with that of boys, translating to a modest but nonetheless clearly identifiable gender difference (Figures 3C and 3D). Moreover, the *Milk* cluster (RA II) is predominant among male subjects, whereas that of *Traditional* (RA III) is the single largest aggregation among females (Figure 3C and 3D). Main features of the five core prototypes, as identified at the highest level of segregation, are summarized in Table 1.

4.1.2 Closer inspection of a single major cluster

Partly owing to its size and heterogeneity, but also to demonstrate the enhanced potential of hierarchical cluster analysis over more simplistic protocols, the largest aggregation (RA II; 3rd in line from left end in Figure 1) underwent deepened scrutiny at two subordinate hierarchical levels. As schematically depicted (Figure 4) two assemblages appear at level 2, one (RA II/I) holding five groups and the other (RA II/II) linked to a subcluster pair (at level 3). Although the RA II/I assembly shows some preference for *Soft beverages (sweetened)*, *Buns & cakes* and *Fruit*, it is quite heterogeneous, whereas RA II/II emerged as an unequivocal *Milk* subcluster and was thus not inspected at deeper level (Appendix B, Table A 1A). The five level 3-aggregations attached to RA II/I, on the other hand, resolved into several rather distinct formations of varied characteristics (Appendix B, Table A 1B). Two among them (RA II/I/II *versus* RA II/I/V) appear as mutually quite contrasting; one featuring *Soft beverages*, *Sweets* and *Buns & cakes*, whereas the other displays *Fruit & berries*, *Vegetables* and *Cereals* as major food groups (Appendix B, Table A 1B). Thus, a noticeable fraction (17%) of a cluster with an overall reasonably prudent consumption pattern displays some resemblance to prototype RA IV (*Soft beverages (sweetened)/Buns & cakes*), appearing at the highest hierarchical level.

4.1.3 Clustering food groups

Food groups, rather than consumers, were also subjected to OMB-DHC processing in order to explore their hidden relationships. Strikingly, *Milk* segregates as the singly most distinctive class, appearing as a single major cluster in the dendrogram, but *Soft beverages (sweetened)* also takes a forward position. Other salient but less prominent groups are *Juice* and *Cereals*. The ensemble of *Fruits*, *Meat* & *poultry* and *Potato* aggregates into a rather tight subcluster (Figure 5). When analysing these results, it is important to keep in mind that the dataset was not re-normalized upon replacing consumers with food groups (here food groups were clustered according to their profiles across all individuals). Therefore, magnitudes of the individual food group profiles will depend on the total consumption level. Since *Milk* and *Soft beverages (sweetened)* occupy a large fraction of the total intake per individual, the corresponding profiles will be much larger than for the remaining food groups. Presumably, this contributes to their dominating roles (Figure 5).

4.1.4 PCA (factor) analysis of individuals as well as food groups

PCA on food group profiles across individuals was conducted. Largely in agreement with readouts from Cluster Analysis, as touched upon above and likewise performed on food categories, inspection of the two first principal components identified *Milk (fatty)/Yoghurt* and *Water/Soft beverages (sweetened)* as extreme outlier food categories (not shown). Data were, however, not normalised prior to these two analyses, which presumably enhances the extreme appearance of certain foods, e.g. *Milk*.

Principal components, derived from PCA on the entire Swedish data set and based on normalized food consumption profiles, are depicted in Figure 6. To support overview only seven predominant vector loadings of each among the first four principal components (encapsulating 67% of the variability) are shown. This output reveals, for example, that PC1 – the component of maximal variability (away from the population mean value) – is mainly dictated by high vector loadings in *Milk* consumption and – in opposite direction – in *Soft beverages (sweetened)* (appearing as *Soda* in the figure) and *Water*, or *vice versa* (the opposite principal direction is equivalent). Similarly, the next largest variability (PC2), being orthogonal to the first PC by definition, projects along a direction characterized mainly by high loadings in *Soft beverages (sweetened)* consumption and – negatively correlated – *Water* (or *vice versa*). The third principal component (PC3; orthogonal to the first two PCs), features combined

high scores for *Milk*, *Water* and *Soft beverages (sweetened)*, likewise negative loadings for *Fruit & berries*, *Juice* and *Cereals* (or *vice versa*). Finally, the fourth component (PC4) is characterized by very high vector loading on *Cereals* and, a less prominent and negatively correlated score for *Juice* (or *vice versa*). An intuitive interpretation of these results converges on subgroups of consumers that actually occupy/dominate these four directions of large variability. Then we should expect, for example, the presence of a subgroup of bi-polar consumers that either have an above average intake of *Cereals* and below average intake of *Juice*, or *vice versa*. It is, however, important to recall that the principal directions are designed in the PCA to become orthogonal to each other; caution should therefore be exercised when interpreting these results in terms of hidden consumption patterns in disparate subject groups. Nonetheless, three out of four inspected principal components seemingly commensurate reasonably well with either of three dietary prototypes, as defined by OMB-DHC analysis. Thus, the *Milk, Soft beverages (sweetened)/Buns & cakes* and *Cereals* prototypes share outstanding features of PC1, PC2 and PC4, respectively, whereas PC3 is seemingly more composite (Figure 6 and Table 1). A more revealing analysis with respect to this issue, however, would involve some sort of independent component analysis (ICA), rather than PCA. In ICA, the directions extracted are not necessarily orthogonal to each other and the corresponding score values obtained will become (approximately) independent rather than only linearly independent, as in PCA.[62,63] This direction is anticipated to open for more revealing interpretations of the underlying relationships, but remains as a topic for future work beyond the current study.

4.1.5 Consumer profiling using energy percent-based food intake

The relative intake of each among a predefined set of food groups is typically referred to as either weight or energy, both – for each individual – given in relation to the total consumption of all selected food groups together. The former parameter is most commonly used and generally perceived as adequate. Nonetheless, percentage of total energy can compensate for possible bias introduced by high- or low-density foods, and may also help removing extraneous variation due to differences in physical activity, body size or gender.[64,65] A marked dissimilarity in outcome, depending on preference for either of the aforementioned parameters, may call for further inspection to unveiling the underlying causes and to support a choice of parameter. Analysis of the data (energy per-

cent) by the OMB-DHC algorithm revealed an overall pattern almost identical to that of weight percent (not shown). Moreover, the first principal component (PC1) also resembles that based on normalization based on relative weight, but the remaining three (PC2 through PC4), were not equivalent to those derived from weight-based intake. Salient relationships noted among the dominating elements of the principal component vectors included increased *Bread* and decreased *Cake* and *Sweets* (or *vice versa*, not shown). Further downstream operation using consumption data converted to energy percent was, however, not pursued and thus all further results presented below were based on weight percent normalization.

4.2 Swedish data set, separate age groups

4.2.1 Preschool children (4 years of age)

Analysis of consumption data of Swedish 4 year old children by means of OMB-DHC revealed two main clusters, but only one among them is well separated from data centre. Moreover, the largest and less distinct aggregation at the top hierarchical level (C1) is quite heterogeneous, whereas the second assembly (C2) attains a relatively coherent shape (Figure 7). Nonetheless, these two top-level clusters clearly show distinct properties: C1 is characterized by *Milk*/*Yoghurt* foods, whereas the *Cereals* counterpart is outstanding in C2 (Appendix B, Table A2). Owing to an extensive heterogeneity of cluster C1 and its further segregation into fairly distinct sub-populations, aggregations from several hierarchical cluster levels were preferred to define dietary prototypes. This selection incorporates four separate aggregations within three overarching structures: one at each of hierarchical levels 1 and 2 and two at level 3 (Figure 7). Outstanding food groups or otherwise characteristic features of each among these four prototypes were assigned as follows: R4 I *Cereals*, R4 II *Soft beverages (sweetened)/Buns & cakes*, R4 III *Milk* and – finally – R4 IV *Varied/Water* (Table 2). Notably, OMB-DHC interrogation of the original data of Swedish preschool consumers (i.e. based on 35 foods) emerged as equivalent to that outlined above (not shown).

Using instead food group profiles as input to OMB-DHC, *Milk* attained the most outstanding single food cluster (top level), followed by *Cereals* and *Soft beverages (sweetened)*. At lower hierarchical levels, *Fruit & berries* and *Yoghurt* as well as *Potato* and *Meat & poultry* appear as co-clustering pairs (Figure 8). PCA, applied to the same food group profiles, revealed four clus-

ters, which aligned reasonably well with major attributes of the aforementioned Cluster Analysis: *Milk*, *Soft beverages (sweetened)*, *Cereals*, but with some foods appearing in a rather heterogeneous assembly, with *Potato*, *Meat & poultry*, *Fruits* etc. (not shown). Food group profiles were, however, not normalized prior to PCA and caution should thus be exercised upon interpretation. The four major principal components – obtained from PCA of subject profiles across food groups – are, however, more reliable (Figure 9). Here, the first loading vector (PC1) can be summarized as an inverse relationship between *Cereals* (high loading) and *Milk* (negatively correlated). Conversely, an inverse relationship between the pairs *Cereals + Milk* (high scores) and *Soft beverages (sweetened) + Water* (negative loadings) is an outstanding feature of the second principal direction PC2. In other words, the direction of large variability follows from increasing *Cereals + Milk* while decreasing *Soft beverages (sweetened) + Water* (or *vice versa*). The third principal component – PC3 – involves an inverse relationship between *Water* (high vector loading) and *Soft beverages (sweetened)* (negatively correlated). Finally, the fourth component – PC4 – roughly describes an inverse relationship between *Yoghurt* (high scores) and the triplet *Soft beverages (sweetened) + Milk + Water* (negative scores) (Figure 9).

Some resemblance of the accordingly obtained PCs to dietary prototypes, as defined by OMB-DHC analysis for this age category, is discernible. For example, PC1 aligns reasonably well with the *Cereals* prototype and PC2 shares some outstanding features with that labelled *Milk*, whereas PC3 may be construed as a composite of the *Soft beverages (sweetened)* and *Varied/Water* prototypes. Due to inherent properties of PCA, however, as already discussed in the context of findings summarised in Figure 6, it is not possible to conclude that each of the four orthogonal principal directions correspond to distinct subgroups of consumers that actually exist, i.e. carrying the corresponding salient characteristics described here. Together the major principal components span the space in which most of the consumption variability exists and it is likely that a set of relatively independent/isolated subgroups may be characterized best by a set of non-orthogonal directions in this space (that might be identified by means of independent component analysis). Based on this insight, it is not conflicting to find, for example, that the food group *Cereals* contributes to the variance in the form of an inverse relationship to *Milk* along the first principal direction (PC1), while at the same time contributing with the same sign as *Milk* along the second principal direction (PC2), in which *Soft beverages (sweetened) + Water* have negative loadings. Indeed, subgroups of consumers attached to all of the two inverse relationships of PC1 and PC2 might exist, but only at a conjectural level at this stage of analysis.

4.2.2 Eight-year children

This population segregates into two major clusters of dissimilar sizes and four aggregations at the ensuing hierarchical level. Especially the largest top-level cluster, featuring a short edge at the first branching point, is wide and encompasses several quite distinct assemblies at subordinate levels (Figure 10). Four reasonably coherent subclusters appear at hierarchical level two and were therefore selected to further inspection and characterisation as dietary prototypes (Table 3). As above, each among the accordingly defined four dietary prototypes was referred to as either a single or several outstanding food groups: R8 I *Varied/Water*, R8 II *Pizza/Soft beverages (sweetened)*, R8 III *Milk (fatty)* and R8 IV *Milk (low fat)/Soft beverages (light)/Juice.*

4.2.3 Eleven-year children

This data material is the most diversified at the top hierarchical level, appearing as five distinct aggregations (Figure 11). Key food groups (or overall consumption profile) of the five top-level aggregations, operationally defined as dietary prototypes, are as follows: R11 I *Milk (fatty)*, R11 II *Milk (fatty)/Varied*, R11 III *Soft beverages (sweetened)*, R11 IV *Milk (low fat)/Soft beverages (light)/Juice* and R11 V *Varied/Water* (Table 4).

4.3 Comparing Swedish age groups

As already touched upon above, four dietary prototypes were found to describe consumers of either 4- or 8-years age, whereas those of 11 years fell into five prototypes. These 13 prototype vectors, thus defining 13 subgroups, were used as input to CMDS as well as to Bi-Cluster Analysis (a likewise composite technique, here labelled HPBCA) for deepened analysis of their relationships. CMDS strives at preserving mutual distances among the inputs while decreasing the number of dimensions; for details see Appendix C. Using Euclidean distances CMDS was conducted subsequent to variable normalisation, based on z-score normalization. The observations were compressed into two dimensions and subsequently displayed by means of a 2-dimensional scatterplot (Figure 12). As shown in this display preschool children are comparatively well aggregated, whereas the elementary school consumers are more scattered. Notably, one outlier appears in each of the two elementary school age groups (prototype R8 IV among the eight years group and prototype R11 IV among the eleven years group). Closer inspection of the outlier

prototypes revealed fairly high mutual similarity and relatively high scoring for the *Fruit*, *Juice*, *Milk (low-fat)* and *Meat & poultry* food groups (Compare Tables 3 and 4).

Upon subjecting the 13 Swedish dietary prototypes to HPBCA, each of them re-appeared as a column in a two-dimensional hierarchical cluster dendrogram. This output reveals several immediately evident associations and distinctions. For example, preschool consumers appear as a reasonably coherent group (the leftmost four columns), while dietary prototypes of eight and eleven year old consumers are quite interlaced (Figure 13). Moreover, the two outlier prototypes, already displayed in Figure 12 and further commented upon above, appear as exclusive members of a common cluster in the 2-dimensional dendrogram (the two rightmost columns of Figure 13). Obviously, these two dietary prototypes share many features, such as comparatively high intake of an array of healthy foods: *Flavoured yoghurt (low fat)*, *Margarines (low fat)*, *Soft beverages (light)*, *Juice*, *Milk (low fat)* and *Rye bread* (Figure 13; Tables 3 and 4). There is, however, also some mutual disparity; e.g. the eight year outlier prototype shows higher consumption of *Eggs*, *Blood food, Ice cream* and *Desserts*, compared with that of the eleven-year outlier.

Furthermore, the RA IV prototype (*Soft beverages, sweetened/Buns & cakes*), disclosed by clustering analysis of in the entire Swedish data set (Figure 1), likewise re-appear in a quite similar manifestation in each of the age-matched data excerpts of elementary school children, i.e. across the 8- and 11-year old consumers. Notably, this pair of dietary prototypes – R8 II and R11 III – comprises a separate cluster in the HPBCA output, in analogy with a likewise joint appearance of the prudent R8 IV and R11 IV counterparts (Figure 13). Constraining the similarity requirement to *Soft beverages (sweetened)* only allows the incorporation of a third dietary prototype, i.e. the RA II of preschool children, into this encapsulation of health-compromising prototypes (compare Tables 2 through 4).

Lastly, several reasonably coherent food clusters emerged, as typified by that composed of *Sausages*, *Bread*, *Meat & poultry* and *Potato*. Finally, four pairs of foods were found to co-aggregate tightly in the chart: a) *Bread* and *Meat & poultry*, b) *Flavoured yoghurt (fat)* and *Desserts* c) *Flavoured yoghurt (low fat)* and *Margarines (low fat)*, as well as d) *Soft beverages (light)* and *Juice*. Arguably, these particular food group intakes follow each other relatively closely, regardless of consumer subtype and age.

4.4 Danish data set (preschool children, 4–5 years of age)

Contrary to the Swedish data on preschool children, as revealed by OMB-DHC analysis, Danish children of roughly equivalent age segregate into four distinct categories at the top hierarchical level (Figure 14). Three among the four top clusters lie within a rather similar size range, while the fourth cluster clearly falls outside the frame, encompassing only 8.5% of the total dataset and being most remotely situated from data centre, i.e. sitting on the most extended edge as projected from the top-level intersection. Based on prominent food groups of each main cluster centre, the dietary prototypes were labelled according to the following simplified characteristics: DNS I *Soft beverages (sweetened)/Juice*, DNS II *Milk*, DNS III *Soft beverages (light)* and DNS IV *Yoghurt/Fruit & Berries/Water* (Table 5). Although no significant statistical differences in energy intake across major clusters were identified, the *Soft beverages (light)* cluster seems to lie slightly below average (9%).

Using instead food group profiles as input to the OMB-DHC algorithm, two main structures appear at the highest hierarchical level: *Water*, on the one hand, and *Milk/Soft beverages (sweetened)*, on the other. Other notable formations, appearing at the second and third levels of the dendrogram, are *Soft beverages (light)* as well as *Bread* and *Fruit & berries*, respectively (Figure 15). As already explained above, these results should be interpreted with care, since the food group profiles were not normalized prior to PCA. However, using normalised consumption profiles as input to the PCA analysis, high vector loadings of the four main principal components are as follows: PC1 has high scores for *Water* and *Milk* + *Soft beverages (sweetened)* (last two with negative loadings), PC2 for *Milk* + *Water* and *Soft beverages (sweetened)* (the last with negative score), PC3 for *Soft drinks (sweetened)* + *Water* (both with negative loadings) and PC4 for *Fruit & berries* and *Soft beverages (sweetened)* + *Milk* + *Water* (last three with negative scores) (Figure 16). Thus, in analogy with principal components derived from accordingly processed Swedish data sets, e.g. those displayed in Figures 6 and 9, also here the PC directions of maximal variability appear in terms of a few predominant food groups. Some similarity of the accordingly obtained PCs to dietary prototypes, as defined by Cluster Analysis of the Danish children, surfaces but it is more remote relative to that between Swedish OMB-DHC/PCA counterparts.

4.5 Comparing Danish and Swedish data sets: uni-variate analysis

Spearman's correlations between all pairs of food groups identified the strongest associations between *Vegetables* and *Fruit & berries* as well as between *Bread* and *Sugar* among Danish children (Figure 17). Swedish subjects showed the strongest associations involving *Bread* and *Cheese* as well as *Bread* and *Gravy & dressing* (spreads and butter are included in *Gravy & dressing*).

4.6 Comparing Danish and Swedish data sets: non-supervised MDA

On an overall basis Danish preschool children feature high consumption in *Water*, *Soft beverages (light)*, *Soft beverages (sweetened)*, *Bread* and *Vegetables*, in relation to age-matched counterparts of *Riksmaten – barn 2003*. Outstanding features of Swedish children of this age, as appearing in the data set and in relation to those of Denmark, involve four groups: *Cereals*, *Yoghurt*, *Potato* and *Meat & poultry*. Strikingly, Swedish preschool children are known as high consumers of gruel, in some contrast to Danish counterparts, which is a likely cause of the major trans-national difference in *Cereals* intake, as revealed by inspection of data sets addressed here.

Dietary prototypes of Danish and Swedish preschool consumers (four of each nation, Figures 7 and 14; Tables 2 and 5), were subjected to deepened scrutiny. This data interrogation was equivalent to that across Swedish age groups described above and thus involved CMDS and HPBCA. As for the Swedish data, variable normalisation (z-score), Euclidian distances and reduction into two dimensions by CMDS were employed. According to results shown in Figure 18, CMDS revealed that Swedish preschool dietary prototypes are distributed over a slightly larger distance, relative to those of Denmark, but national clusters are consistently more remotely situated to any trans-national counterpart than to anyone domestic, i.e. national proximity is evident. Based on the dimension-reduced representation, the two closest inter-national clusters are roughly 57% and 26% more remote, relative to distances separating the widest apart situated intra-Danish and intra-Swedish subpopulations, respectively. This measure is, however, very approximate because of an extensive dimension reduction needed to present the output as a 2D display. Furthermore, the HPBCA disclosed a quite remarka-

ble pattern, which confirms and extends on the CMDS result (Figure 19). Clearly, *Pizza*, *Bread*, *Snacks* and *Vegetables* are more prominent among the Danish children, whereas intake of *Potato*, *Pasta*, *Rice* and *Yoghurt* are higher in the Swedish corresponding data set. Actually, the overall image is not altogether far from an inverse Danish-Swedish consumption pattern. Some among the closely co-clustering foods were: *Bread* and *Sugar* as well as *Yoghurt* and *Sausage*. Those pairs differed markedly across nations, so as the Danish preschool consumers had high intake of the former foods and low consumption of the latter and *vice versa* for the Swedish counterparts. Lastly, the PCA-analysis (based on normalised data) confirm prominent trans-national dissimilarities among preschool children, but one major principal component – PC2 – is strikingly similar across the Danish and Swedish data sets and indicates a shared tendency for reciprocal consumption of *Milk* and *Soft drinks (sweetened)* (Figures 9 and 16).

4.7 Comparing Danish and Swedish data sets: Supervised MDA

Classification models were designed to predict high/low consumers of a specific food group from the consumption patterns of all other food groups, using both Random Forest (RF) and nearest shrunken centroids (NSC). Models were built to predict high/low consumers of each among the array of food groups common to the Danish and Swedish preschool subjects. This sort of MDA disclosed exciting patterns for many high/low single food group consumers across the bi-national data set, encompassing preschool children. Here, we choose to focus on the most clear-cut outputs, thus presenting only a fraction of all RF- and NSC-based results. The RF exercises revealed certain intriguing dissimilarities – and some resemblance – across Danish (4–5 year old) and Swedish (4 year old) children, with regard to consumption patterns connected with low/high intake of selected food groups, see Figure 20 panels A and B. Notably, certain remarkable bi-national dissimilarities in residual determinates of overall eating patterns among extreme (20% lowest and 80% highest) consumers of *Bread*, *Meat & poultry*, *Potato*, *Buns & cakes* and *Soft beverages (sweetened)* were seen. By contrast, the predictive models of *Vegetables* and *Fruit & berries* clearly showed more shared properties across Danish and Swedish consumers. Moreover, some food groups, typified by *Cereals*, *Milk*, *Pasta*, *Rice* and *Juice* did not produce predictive RF models for either of the two national cohorts, i.e. high/low

consumption of each of these food groups could not always be predicted by the consumption of other food groups.

To drastically reduce the risk of creating over-fit models an alternative and markedly stricter route was chosen, involving a distinct but related algorithmic technique based on the NSC classifier, which – in addition – is computationally less demanding. Contrary to the RF based analyses described above, those incorporating NSC encompassed not only classification, but also variable selection and (likewise) partitioning of data into training (80%) and test (20%) sets (see Appendix C for details). Each accordingly built NSC model was attached to quality estimates, in this case the Area-Under-the receiver operator characteristic-Curve (AUC). Accordingly derived results showed considerable resemblance to those obtained by means of the simplified RF analyses, but the two output sets were nonetheless far from identical. The NSC models were each built on bipartite intra-national sub-populations based on single food groups; the results are outlined in Figure 20, panels C and D (for details, see Appendix C).

As inherent to the *Bread* models, *Sugar, Gravy & dressing* and *Vegetables* emerge as remarkably strong variables to discern Danish low-level consumers from those of the high intake fraction, whereas *Cheese*, *Gravy & dressing*, *Cereals* and *Water* are strong contributors to Swedish counterparts. Obviously, *Juice* and *Soft beverages (light)* have great impact on the Danish NSC *Milk* models, whereas *Cereals* and *Soft beverages (sweetened)* are essential to those of Swedish children, but also other dissimilarities are evident across these builds (Figure 24). Moreover, *Potato* NSC models revealed *Rice* as strongly supportive to the accurate segregation of Danish children in the pertinent low/high consumer type, but *Sausage* and *Fish* – both totally absent in the Danish models – are clearly essential to those of Swedish subjects. Analogously, the *Meat & poultry* NSC models of Swedish children are strongly impacted by *Eggs*, *Pasta* and *Pizza*, entirely dispensable to accurate segregation of Danish subjects within this bipartite subgroup, which rather depend on *Potato*, *Rice* and *Fruit & berries*. Clearly, *Soft beverages (sweetened)* models share some features across nations, but *Buns & cakes* clearly help separating the Swedish low/high consumers, whereas *Soft beverages (light)* and *Ice cream* are important to those of Denmark. The *Fruit & berries* builds display considerable trans-national resemblance, but *Sausage* (albeit contributing weakly) stands out only in those of Denmark. Lastly, NSC models built on the *Vegetables* food group also show trans-national similarity, but *Pizza* and *Soft beverages (light)* appear as a nation-specific impact variable for Danish and Swedish low/high separation, respectively (Figure 20).

5. Discussion

The study presented here shows selected findings and interpretations based on MDA of two dietary surveys – one conducted in Sweden and one in Denmark – focused on preschool and elementary school children. In this undertaking both non-supervised and supervised MDA techniques have found application. Thus, the focus has been strictly laid on dietary patterns, which are inherently complex and seemingly need more than a single MDA technique to become satisfactory deciphered. For example, most dietary patterns typically feature low consumption of some food groups, jointly with high consumption of other foods. Cluster Analysis, typically based on the K-means algorithm, and Factor Analysis, commonly in the form of PCA, have found increasing application in the area over the last decade.[29] In a dietary survey context, Cluster Analysis gathers consumers into non-overlapping groups based on dietary similarity, whereas PCA identifies linear combinations of foods that are frequently consumed in combination. Thus, these two statistical techniques describe diets from different perspectives. A major technique applied to the study outlined here is an in-house built development of the K-means algorithm, featuring divisive-type multi-furcating clustering operation as well as output display of several hierarchical levels. This hierarchical design proved indeed very helpful in identifying and selecting relevant populations to dietary prototypes. From this inroad, two separate extensions – CMDS and HPBCA – were designed and likewise allowed to operate on the Danish and Swedish data sets. The PCA statistical technique also found application here, but mostly to support results derived by other methods. Moreover, predictive modelling – based on the widely acknowledged RF algorithm as well as the computational-efficient NSC – was applied to selected excerpts of the entire data set, i.e. Danish and Swedish preschool children.

5.1 The use of hierarchical clustering seems to be new in this field

Contrary to many other sciences the MDA approach has made a rather late entrance in nutrition areas; although initially reported about twenty-five years ago, only the last decade has witnessed more substantial contribution of MDA to nutrition epidemiology.[29,37,38] To the best of our knowledge, however, hierarchical cluster analysis (HCA) has not hitherto been reported in this particular field. We choose to implement a largely in-house built algorithm, designated OMB-DHC and featuring multifurcating display of aggregations at several levels of hierarchy, in the MDA-oriented investigation of dietary survey data. Notably, its divisive-operating design ensures good accuracy at the highest hierarchical levels, contrary to agglomerative counterparts. It has proven highly versatile to disclosing pertinent patterns embedded in various data types.[66–68] This enhancement over more commonly employed simplistic clustering operations, devoid of such output diversification, provided valuable information on intra-cluster homogeneity as well as the degree of segregation of such aggregations, i.e. dietary prototypes, from their respective data centra. Altogether, this allowed us to inspect and assess each cluster formation for relevance prior to further description and construal as well as downstream processing. In the material addressed here, this particularly applies to Swedish four and eight year old consumers, featuring intricate multi-level aggregations and thus needing careful inspection prior to prototype definition (Figures 7 and 10). It is clearly our notion that single-level k-means clustering only, in such instances, would entail high risk of creating formations of modest or even poor adequacy with respect to relationships among the children. To this may be added that such ambiguity is not fully disclosed to users of simplistic clustering algorithms, unless tied to validation of cluster number by means of further data processing, e.g. linear discriminant analysis.[44,69] A rather wide numeral range of dietary clusters, from two to at least six, have been reported.[43,70] Although data sets can differ substantially across such studies, some uncertainty nonetheless remains as to whether cluster output, in some instances, were sufficiently inspected for relevance. Obviously, a truly hierarchical operation of clustering algorithms can confer support in this matter.

Moreover, the dendrogram structure of clustering analysis outputs enabled us to pick certain subclusters of data excerpts for deepened inspection of heterogeneity in this study material. As further discussed below a set of pre-processed data, i.e. prototypes already defined by

OMB-DHC, were also subjected to further scrutiny involving secondary clustering analysis, as enabled by two distinct algorithms, each fed with OMB-DHC output.

5.2 Energy based normalization

The food consumption survey material underwent global as well as excerpt-based (age- or national) data analysis. The entire Swedish material, encompassing three ages of young consumers (4, 8 and 11 years), revealed five populations at the highest hierarchical level, each being quite well separated from the remaining data. This output was created by feeding the OMB-DHC algorithm with weight-based data (% weight), but an almost equivalent pattern, also featuring five major clusters, emerged upon re-expressing data as relative energy (% caloric intake). Certain minor dissimilarities, typified by disappearance of *Water* as a key component of one assembly, were identified but deemed as being of rather subordinate significance to further exploration and interpretation. This outcome, i.e. slight differences, aligns well with findings in an earlier report, involving much higher subject numbers.[65] Nonetheless, certain OMB-DHC results generated from relative energy intake were not fully equivalent to those of relative weight. For example, a weight-based dendrogram representation highlights *Milk*, *Soft beverages (sweetened)* and *Fruit/Meat & poultry/Potato* as prominent food assemblies, whereas *Milk/Meat & poultry/Bread*, *Buns & cakes/Potato/Sweets* and *Sausages/Soft beverages (sweetened)* appear as outstanding aggregations in energy-based clustering analysis of food groups. Thus, it is particularly clear from the latter output that certain healthy foods, on the one hand, and some much less healthy counterparts, on the other, tend to appear in disjoint consumption patterns (not shown). Moreover, certain challenges confronted in the predictive modelling exercises also suggest a preference for data relation to energy intake. Although not pursued further on in this study, several observations nonetheless indicate an advantage of energy-related data as input to Cluster Analysis and Predictive Modelling. Actually, preference for % total energy as input to Cluster Algorithms and % weight for PCA has been suggested by others.[71]

5.3 Dietary prototypes identified

Two, four or five overarching dietary populations were identified in various fractions of the data material or in the entire set. Subsequent to further inspection of branching characteristics of the respective data excerpts, however, the aforementioned range was narrowed to 4–5 major clusters. These numbers are in reasonably good compliance with those of several earlier reports on dietary patterns for children of various ages, arriving at three to four dietary populations.[44,69,72–74] Actually, a recent report on dietary patterns in consumption data on Irish adolescents identifies five clusters, thereby featuring numeral equivalence to dietary prototypes found among 11-year old children of *Riksmaten – barn 2003* as well as the entire Swedish data set.[75] Moreover, interrogation of a British cohort of children, by means of both PCA and Cluster Analysis techniques, revealed high agreement between clusters and major principal components, largely with the following appearance of segregation: Health conscious/Plant-based, Junk/Processed food and Traditional British.[44,76] Actually, consumption populations, largely equivalent to the above-mentioned triad (with category "Traditional" being adapted to the respective cohort nationality), are rather widely documented.[44,69,73,74,76] Although data addressed and processed in this report typically segregated in four or five rather than three main clusters, the aforementioned archetypal classes were nonetheless distinguished also in the Swedish and Danish data sets.

5.4 No major gender difference

As revealed by global inspection of Swedish data, differences across genders were not dramatic, but notwithstanding revealing some pertinent characteristics. On an overall basis, female consumers displayed higher similarity in consumption patterns, relative to those of male counterparts, i.e. clusters were more evenly sized. Notably, the *Milk* cluster (also referred to as RA II; being the overall largest) clearly dominated young male consumers, whereas that of *Traditional* (RA III) emerged as predominant among young females, closely followed by the *Milk* aggregation. Actually, the *Soft beverages (sweetened)/Buns & cakes* (RA IV) cluster was almost equally sized across genders, a finding not quite anticipated (Figure 2). Several earlier reports have documented specific gender dissimilarities in dietary habit of young consumers.[44,69,72,73] Indeed, boys in Western societies seemingly have slightly higher tendency to become assigned to clus-

ters designated “traditional”, “convenience” or “fast food”, i.e. relatively unwise patterns, compared with those of girls.[44,69] Apparently, the Swedish traditional dietary prototype does not quite follow suit, but a marked preference for *Milk* among many – but not all – Swedish children may possibly introduce a slight distortion in the row.

5.5 Comparing Swedish Subgroups within and across age categories

As briefly touched upon above, the three separate age groups (4, 8 and 11 years of age) of *Riksmaten – barn 2003* were subjected to individual inspection, thereby enabling comparison within and across these categories. As derived by OMB-DHC only the oldest consumer class consistently segregated into relatively homogeneous prototypes at the highest hierarchical clustering level, but closer dendrogram inspection nonetheless allowed identification of meaningful and reasonably coherent aggregations also for the remaining age groups, but – in some instances – appearing at lower hierarchical levels. Moreover, global scrutiny across Swedish age groups by means of CMDS and HPBCA – two related but nonetheless distinct computational techniques, each allowed to operate downstream of OMB-DHC – disclosed certain pertinent features of the data material. Clearly, the dietary prototypes of preschool children appeared as the most tightly assembled constellation, which also sits most remote to those of older age groups. This observation was reinforced by co-clustering of prototypes of 4-year old consumers in the HPBCA chart (Figures 12 and 13).

Another observation, certainly not easily disclosed by conventional statistical analysis, pertains to two outlier dietary prototypes of each among the two elementary school children: Seemingly, they are major contributors to a much wider data distribution of the two older age groups, in relation to that of preschool children, and also appear in a common HPBCA dendrogram cluster. Although not identical, these two dietary prototypes nonetheless share many pertinent characteristics, notably higher preference for low-fat foods, typified by *Fruits*, *Juice*, *Milk (low-fat)*, *Flavoured yoghurt (low fat)*, *Margarines (low fat)* and *Soft beverages (light)*, relative to the remaining prototypes. Thus, these two outliers, showing some resemblance to the (tentatively labelled) *Traditional* dietary cluster as appearing in OMB-DHC output of the aggregated Swedish data set, can be perceived as reasonably prudent dietary prototypes. They represent around 20% of the consumers in each age catego-

ry. It is tempting to speculate on whether these clusters would re-appear in a prospective nutritional investigation and, in that event, be attributed to largely the same consumers over time.

Conversely, each of all age groups included a single dietary prototype attached to less healthy eating pattern (R4 II, R8 II and R11 III; see Tables 2 through 4). Although not drastically diverging from average consumption of the respective age groups, these three clusters consistently showed the lowest intake of *Vegetables* and *Fruit & berries*, relative to other age-matched counterparts. Moreover, an about three-fold higher intake of *Soft beverages (sweetened)* was a salient shared feature among these prototypes. This overall dietary pattern, as confined by the above-mentioned prototypes, encapsulates nearly 20% of all preschool children in the data set, but mounted to around 25% of elementary school consumers. Analogous to the outlier (and prudent) prototypes, discussed above, the two relatively imprudent prototypes, one in each of the elementary school age categories, clustered together in the HPBCA analysis. Notably, there are well documented relationships between consumption of *Soft beverages (sweetened)* among children and cardiovascular risk factors as well as to markers of inflammation (likewise determined in children).[77,78] Other studies, largely encompassing adult consumers, provide substantial accumulated support of a connection between such intake pattern and risk of developing the metabolic syndrome as well as type-2 diabetes mellitus,[79,80,81] Such compelling relationships provide a strong incentive to advance consumption patterns of the aforementioned dietary prototypes, as identified in the data set, towards considerate eating habits.

5.6 Comparing Swedish and Danish Consumer Patterns

An important incentive to the current study lied in comparison of Danish and Swedish children, with regard to food consumption. The respective survey protocols of the two Nordic data sets were similar and ages (4 and 4–5 years, respectively) close to identical. Yet, a superficial inspection of the respective dendrograms reveals marked dissimilarity: Unlike Swedish age-matched subjects Danish children are all distributed in very distinct groups, i.e. four well separated main clusters appear, an overall image much more alike that of Swedish 11-year old consumers. Closer scrutiny of Swedish 4-year old children also disclosed four assemblies, a numeral equivalence to the Danish data segregation pattern, but at least

two among them being vaguely separated and much closer to data centre (Figures 7 and 14). Actually, the top hierarchical level houses only two large populations. These patterns suggest higher intra prototype consumption coherence among Danish subjects, relative to those of Sweden, already at 4–5 years of age. Main aggregations of Danish children, as assigned by key food groups of each cluster, appear as *Milk*, *Soft beverages (light)*, *Soft beverages (sweetened)/Juice* and *Yoghurt/Water*, whereas those of Swedish subjects emerge as *Milk*, *Soft drinks (sweetened)/Buns & cakes*, *Cereals* and *Varied/Water*. Thus, two among the four main dietary prototypes (clusters) of each nationality displayed some resemblance at this low resolution, yet with differences also across them. Analogous with findings among the Swedish consumers (of all three age classes), a dietary prototype featuring relatively high consumption of *Soft beverages (sweetened)* appeared among the Danish consumers, amounting to 27% of the studied population. It showed, however, tighter connection to *Fruit & berries*, relative to that of Swedish age-matched children.

Overarching divergence between the two national cohorts include higher consumption of *Bread*, *Vegetables* and *Soft beverages (sweetened)* as well as lower intake of *Yoghurt*, *Cereals*, *Potato* among Danish children, compared with Swedish counterparts. Furthermore, a Finnish dietary survey, encompassing adults and three ages of children (1, 3 and 6 year old), identified three separate clusters, referred to as *Healthy*, *Traditional* and *Fast food/sweets*, in all probed age groups.[74] Contrary to Swedish and Danish subjects of about similar ages, Finnish children (6 years old) consistently consumed large and roughly equal amount of *Milk* and had similar intake of *Fruit*. *Porridge*, *Potato (fried)*, *Fat spread* and *Soft drinks (sweetened)*, on the other hand, were all major contributors to dissimilarity across the several clusters.[74] Seemingly, food consumption among Finnish subjects of this age range (4–6 years old) shows some appreciable difference from those of both Danish and Swedish counterparts. Furthermore, it should be mentioned in this context that although designs of the two national surveys elaborated on here align reasonably well there are two distinctions, which each may have some – presumably low – impact on the overall outcome: Cohort numbers differed (318 Danish *versus* 590 Swedish children) and the Danish survey extended several days beyond that of Sweden. Thus, each Danish consumer was more carefully scrutinized, whereas the Swedish cohort is larger. Intuitively, these discrepancies may cause a slightly more even distribution of Swedish subjects, in relation to Danish children. Both

designs, however, adhere to standard statistical norm, thereby safeguarding results against major bias.

Although much pertinent information could be extracted from the several HCA dendrograms and commensurate tables, further agglomeration of results output was, already at study outset, anticipated as highly conducive to enlightening intra- and trans-national pattern relationship. Thus, the main four dietary prototypes of each national preschool consumption data were further processed by means of DPCA and HPBCA. Already the first among these extended data interrogations disclosed slightly greater distance across Swedish clusters, relative to that of Danish counterparts and that three Danish sub-populations are relatively closely situated, whereas only two Swedish counterparts are mutually proximal. Moreover, all clusters are nation proximal, i.e. even widely spaced dietary prototypes are tethered to nation (Figure 18). Perhaps an even brighter illustration of intra- and inter-national relationships emerged from HPBCA: The respective consumption patterns of Danish and Swedish children displayed notable resemblance of an inverse relationship, clearly not an anticipated finding at the outset of this study (Figure 19).

5.7 The general potential of MDA in this context

Obviously, MDA is a very powerful technology to disentangling meaningful prototypes in complex multi-dimensional data sets and RF as well as NSC have wide acknowledgement of performing excellently in various advanced pattern recognition applications, including those encompassing high numbers. Thus, we have been keen on introducing this approach to deepened comparison of the roughly age-matched Danish and Swedish data sets. Although much insight in similarities and differences across consumption among Danish and Swedish preschool children was enabled by non-supervised MDA, the supervised technique – in the form of RF and NSC – clearly helped unravelling the most embedded patterns, perhaps also those of the highest relevance to dissecting trans-national consumption dissimilarities in thin slices.

5.8 Highlights of NSC modelling

As detailed in the Results section on predictive modelling, appropriate discrimination between low/high consumers – with regard to each of an array of single food groups – revealed quite different patterns across nationality, but some trans-national resemblance was also evidenced by virtue of these data interrogation techniques. As revealed by NSC predictive modelling the least divergent patterns across nations appear for classifiers within the *Vegetables* and *Fruit & berries* categories, yet with notable trans-national discrepancies. For example, unlike NSC models of *Vegetables* built on Swedish data and featuring dependence (among several foods) on *Soft beverages (light)*, *Pizza* emerges as an outstanding variable in the Danish counterparts. Moreover, *Sausage* is seemingly a variable needed to accurately distinguish high from low Danish consumers of *Fruit & berries*, but arguably dispensable to classifiers built on the Swedish equivalent data. The remaining NCS classifiers, however, showed even larger trans-national divergence. Actually, a broad view revealed several remarkable dissimilarities: For example, *Bread* – in Sweden highly connected to *Cheese* and *Cereals* – being grossly insignificant parameters to discern among Danish (low/high) *Bread* consumers, who rather appear dependent on marmalade (identified as *Sugar*) and *Vegetables*. Moreover, *Sausage* and *Fish* are strong determinants of the Swedish *Potato* NSC model accuracy, sharply contrasting to *Rice* being a prominent variable in models based on equivalent Danish data. Strikingly, *Soft beverages (light)* significantly impact NSC models of *Soft beverages (sweetened)* for Danish children, but are totally expendable for the accurate bi-partite NSC-classification of Swedish children within this food group. Likewise surprisingly, the *Meat & poultry* NSC classifiers, as well as those of *Milk*, also showed drastic divergence across nations.

To the best of our knowledge this study is seminal in the discovery of highly entrenched nation-specific dietary patterns by means of Predictive Modelling. It is, however, also our notion that the supervised learning techniques presented here and applied to the bi-national data set of preschool children hold promise for further refinement and contextual optimisation, thereby advancing the deciphering of truly embedded region- or nation-specific food consumption patterns across various multinational dietary survey data sets.

Lastly, some key trans-national disparities and similarities were identified already by Spearman's correlation analysis – a comparatively simplistic statistical technique. Those observations are *Bread*-*Sugar* and *Bread*-*Cheese* connections in Danish and Swedish preschool consumers,

respectively, as well as a consistent trans-national association between *Vegetables* and *Fruit & berries*. On a general level these findings, jointly with those mentioned above, aligns well with an earlier report, showing specific patterns of trans-national similarity and disparity in consumption habits among preschool children.[74]

6. Acknowledgement

This work was funded by the Nordic Council of Ministers, Nordic Working Group for Diet, Food & Toxicology (NKMT).

References

1 Mendez, M.A., Monteiro, C.A. and Popkin, B.M. (2005) Overweight exceeds underweight among women in most developing countries. *Am J Clin Nutr* 81 (3), 714–721.

2 Yusuf, S., Hawken, S., Ounpuu, S. *et al.* (2005) Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. *Lancet* 366 (9497), 1640–1649.

3 Mente, A., de Koning, L., Shannon, H.S. and Anand, S.S. (2009) A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. *Arch Intern Med* 169 (7), 659–669.

4 Grundy, S.M., Brewer, H.B., Jr., Cleeman, J.I., Smith, S.C., Jr. and Lenfant, C. (2004) Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* 109 (3), 433–438.

5 Murtaugh, M.A., Herrick, J.S., Sweeney, C. *et al.* (2007) Diet composition and risk of overweight and obesity in women living in the southwestern United States. *J Am Diet Assoc* 107 (8), 1311–1321.

6 Cho, L.W. (2011) Metabolic syndrome. *Singapore Med J* 52 (11), 779–785.

7 Malik, V.S. and Hu, F.B. (2012) Sweeteners and Risk of Obesity and Type 2 Diabetes: The Role of Sugar-Sweetened Beverages. *Curr Diab Rep.*

8 Wirstrom, T., Hilding, A., Gu, H.F., Ostenson, C.G. and Bjorklund, A. (2013) Consumption of whole grain reduces risk of deteriorating glucose tolerance, including progression to prediabetes. *Am J Clin Nutr* 97 (1), 179–187.

9 Fung, T., Hu, F.B., Fuchs, C. *et al.* (2003) Major dietary patterns and the risk of colorectal cancer in women. *Arch Intern Med* 163 (3), 309–314.

10 Fung, T.T., Hu, F.B., Schulze, M. *et al.* (2012) A dietary pattern that is associated with C-peptide and risk of colorectal cancer in women. *Cancer Causes Control* 23 (6), 959–965.

11 Slattery, M.L., Boucher, K.M., Caan, B.J., Potter, J.D. and Ma, K.N. (1998) Eating patterns and risk of colon cancer. *Am J Epidemiol* 148 (1), 4–16.

12 Norat, T., Bingham, S., Ferrari, P. *et al.* (2005) Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition. *J Natl Cancer Inst* 97 (12), 906–916.

13 Chan, D.S., Lau, R., Aune, D. *et al.* (2011) Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PLoS ONE* 6 (6), e20456.

14 Xu, X., Yu, E., Gao, X. *et al.* (2012) Red and processed meat intake and risk of colorectal adenomas: A meta-analysis of observational studies. *Int J Cancer.*

15 Iglesia, I., Doets, E.L., Bel-Serrat, S. *et al.* (2010) Physiological and public health basis for assessing micronutrient requirements in children and adolescents. The EURRECA network. *Matern Child Nutr* 6 Suppl 2, 84–99.

16 Mendez, M.A., Popkin, B.M., Jakszyn, P. *et al.* (2006) Adherence to a Mediterranean diet is associated with reduced 3-year incidence of obesity. *J Nutr* 136 (11), 2934–2938.

17 Ornish, D., Magbanua, M.J., Weidner, G. *et al.* (2008) Changes in prostate gene expression in men undergoing an intensive nutrition and lifestyle intervention. *Proc Natl Acad Sci U S A* 105 (24), 8369–8374.

18 Liu, E., McKeown, N.M., Newby, P.K. *et al.* (2009) Cross-sectional association of dietary patterns with insulin-resistant phenotypes among adults without diabetes in the Framingham Offspring Study. *Br J Nutr* 102 (4), 576–583.

19 Parekh, N., Okada, T. and Lu-Yao, G.L. (2009) Obesity, insulin resistance, and cancer prognosis: implications for practice for providing care among cancer survivors. *J Am Diet Assoc* 109 (8), 1346–1353.

20 Sofi, F., Cesari, F., Abbate, R., Gensini, G.F. and Casini, A. (2008) Adherence to Mediterranean diet and health status: meta-analysis. *Bmj* 337, a1344.

21 Randi, G., Edefonti, V., Ferraroni, M., La Vecchia, C. and Decarli, A. (2010) Dietary patterns and the risk of colorectal cancer and adenomas. *Nutr Rev* 68 (7), 389–408.

22 Adamsson, V., Reumark, A., Fredriksson, I.B. *et al.* (2011) Effects of a healthy Nordic diet on cardiovascular risk factors in hypercholesterolaemic subjects: a randomized controlled trial (NORDIET). *J Intern Med* 269 (2), 150–159.

23 Story, M., Kaphingst, K.M., Robinson-O'Brien, R. and Glanz, K. (2008) Creating healthy food and eating environments: policy and environmental approaches. *Annu Rev Public Health* 29, 253–272.

24 Delormier, T., Frohlich, K.L. and Potvin, L. (2009) Food and eating as social practice – understanding eating patterns as social phenomena and implications for public health. *Sociol Health Illn* 31 (2), 215–228.

25 Bingham, S. and Riboli, E. (2004) Diet and cancer – the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer* 4 (3), 206–215.

26 EFSA. (2009) General principles for the collection of national food consumption data in the view of a pan-European dietary survey. *The EFSA Journal* 7 (12), 1435.

27 Boon, P.E., Ruprich, J., Petersen, A., Moussavian, S., Debegnach, F. and van Klaveren, J.D. (2009) Harmonisation of food consumption data format for dietary exposure assessments of chemicals analysed in raw agricultural commodities. *Food Chem Toxicol.*

28 FNDDS. (2008) USDA Food and Nutrient Database for Dietary Studies, 3.0. Beltsville, M.D.: Agricultural Research Service, Food Surveys Research Group.

29 Tucker, K.L. (2010) Dietary patterns, approaches, and multicultural perspective. *Appl Physiol Nutr Metab* 35 (2), 211–218.

30 Brambila-Macias, J., Shankar, B., Capacci, S. *et al.* (2011) Policy interventions to promote healthy eating: a review of what works, what does not, and what is promising. *Food Nutr Bull* 32 (4), 365–375.

31 Hammerling, U., Tallsjo, A., Grafstrom, R. and Ilback, N.G. (2009) Comparative hazard characterization in food toxicology. *Crit Rev Food Sci Nutr* 49 (7), 626–669.

32 van Klaveren, J.D. and Boon, P.E. (2009) Probabilistic risk assessment of dietary exposure to single and multiple pesticide residues or contaminants: summary of the work performed within the SAFE FOODS project. *Food Chem Toxicol* 47 (12), 2879–2882.

33 Kabak, B., Dobson, A.D. and Var, I. (2006) Strategies to prevent mycotoxin contamination of food and animal feed: a review. *Crit Rev Food Sci Nutr* 46 (8), 593–619.

34 Darnerud, P.O., Atuma, S., Aune, M. *et al.* (2006) Dietary intake estimations of organohalogen contaminants (dioxins, PCB, PBDE and chlorinated pesticides, e.g. DDT) based on Swedish market basket data. *Food Chem Toxicol* 44 (9), 1597–1606.

35 Moeller, S.M., Reedy, J., Millen, A.E. *et al.* (2007) Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. *J Am Diet Assoc* 107 (7), 1233–1239.

36 Reedy, J., Wirfalt, E., Flood, A. *et al.* (2010) Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – With colorectal cancer risk: The NIH-AARP Diet and Health Study. *Am J Epidemiol* 171 (4), 479–487.

37 Akin, J.S., Guilkey, D.K., Popkin, B.M. and Fanelli, M.T. (1986) Cluster analysis of food consumption patterns of older Americans. *J Am Diet Assoc* 86 (5), 616–624.

38 Wirfalt, A.K. and Jeffery, R.W. (1997) Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. *J Am Diet Assoc* 97 (3), 272–279.

39 Newby, P.K. and Tucker, K.L. (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* 62 (5), 177–203.

40 Chen, H., Ward, M.H., Graubard, B.I. *et al.* (2002) Dietary patterns and adenocarcinoma of the esophagus and distal stomach. *Am J Clin Nutr* 75 (1), 137–144.

41 Costacou, T., Bamia, C., Ferrari, P., Riboli, E., Trichopoulos, D. and Trichopoulou, A. (2003) Tracing the Mediterranean diet through principal components and cluster analyses in the Greek population. *Eur J Clin Nutr* 57 (11), 1378–1385.

42 Newby, P.K., Muller, D. and Tucker, K.L. (2004) Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin Nutr* 80 (3), 759–767.

43 Crozier, S.R., Robinson, S.M., Borland, S.E. and Inskip, H.M. (2006) Dietary patterns in the Southampton Women's Survey. *Eur J Clin Nutr* 60 (12), 1391–1399.

44 Smith, A.D., Emmett, P.M., Newby, P.K. and Northstone, K. (2011) A comparison of dietary patterns derived by cluster and principal components analysis in a UK cohort of children. *Eur J Clin Nutr* 65 (10), 1102–1109.

45 Hoffmann, K., Schulze, M.B., Schienkiewitz, A., Nothlings, U. and Boeing, H. (2004) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* 159 (10), 935–944.

46 Hoffmann, K., Zyriax, B.C., Boeing, H. and Windler, E. (2004) A dietary pattern derived to explain biomarker variation is strongly associated with the risk of coronary artery disease. *Am J Clin Nutr* 80 (3), 633–640.

47 Kroger, J., Ferrari, P., Jenab, M. *et al.* (2009) Specific food group combinations explaining the variation in intakes of nutrients and other important food components in the European Prospective Investigation into Cancer and Nutrition: an application of the reduced rank regression method. *Eur J Clin Nutr* 63 Suppl 4, S263–274.

48 Fahey, M.T., Ferrari, P., Slimani, N. *et al.* (2012) Identifying dietary patterns using a normal mixture model: application to the EPIC study. *J Epidemiol Community Health* 66 (1), 89–94.

49 Patterson, B.H., Dayton, C.M. and Graubard, B.I. (2002) Latent class analysis of complex sample survey data: Application to dietary data. *J Am Stat Assoc* 97 (459), 721–729.

50 Vermunt, J.K. and Magidson, J. (2003) Latent class models for classification. *Comput Stat & Data Anal* 41, 531–537.

51 Padmadas, S.S., Dias, J.G. and Willekens, F.J. (2006) Disentangling women's responses on complex dietary intake patterns from an Indian cross-sectional survey: a latent class analysis. *Public Health Nutr* 9 (2), 204–211.

52 Sotres-Alvarez, D., Herring, A.H. and Siega-Riz, A.M. (2010) Latent class analysis is useful to classify pregnant women into dietary patterns. *J Nutr* 140 (12), 2253–2259.

53 R. (2012) R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical computing. . (http://www.R-project.org/)

54 Breiman, L. (2001) Random forests. *Machine Learning* 45, 5–32.

55 Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 99 (10), 6567–6572.

56 Khalilia, M., Chakraborty, S. and Popescu, M. (2011) Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak* 11, 51.

57 Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44, 330–349.

58 Prinzie, A. and Van den Poel, D. (2008) Random Forests for multiclass classifications: Random MultiNomial Logit. *ESWA* 34 (2), 1721–1732.

59 Dutkowski, J. and Ideker, T. (2011) Protein networks as logic functions in development and cancer. *PLoS Comput Biol* 7 (9), e1002180.

60 Bernard, S., Adam, S. and Heutte, L. (2012) Dynamic random forests. *Pattern Recognition Lett* 33 (12), 1580–1586.

61 Enghardt Barbieri, H., Pearson, M. and Becker, W. (2006) *Riksmaten – barn 2003. Livsmedels- och näringsintag bland barn i Sverige.* Livsmedelsverket.

62 Hyvarinen, A. and Oja, E. (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13 (4–5), 411–430.

63 Gustafsson, M.G. (2005) Independent component analysis yields chemically interpretable latent variables in multivariate regression. *J Chem Inf Model* 45 (5), 1244–1255.

64 Willett, W.C., Howe, G.R. and Kushi, L.H. (1997) Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* 65 (4 Suppl), 1220S–1228S; discussion 1229S–1231S.

65 Northstone, K., Ness, A.R., Emmett, P.M. and Rogers, I.S. (2008) Adjusting for energy intake in dietary pattern investigations using principal components analysis. *Eur J Clin Nutr* 62 (7), 931–938.

66 Soeria-Atmadja, D., Onell, A. and Borga, A. (2010) IgE sensitization to fungi mirrors fungal phylogenetic systematics. *J Allergy Clin Immunol* 125 (6), 1379–1386 e1371.

67 Maddah, F., Soeria-Atmadja, D., Malm, P., Gustafsson, M.G. and Hammerling, U. (2011) Interrogating health-related public databases from a food toxicology perspective: computational analysis of scoring data. *Food Chem Toxicol* 49 (11), 2830–2840.

68 Edberg, A., Soeria-Atmadja, D., Bergman Laurila, J., Johansson, F., Gustafsson, M.G. and Hammerling, U. (2012) Assessing relative bioactivity of chemical substances using quantitative molecular network topology analysis. *J Chem Inf Model* 52 (5), 1238–1249.

69 Pryer, J.A. and Rogers, S. (2009) Dietary patterns among a national sample of British children aged 1 1/2–4 1/2 years. *Public Health Nutr* 12 (7), 957–966.

70 James, D.C. (2009) Cluster analysis defines distinct dietary patterns for African-American men and women. *J Am Diet Assoc* 109 (2), 255–262.

71 Hearty, A.P. and Gibney, M.J. (2009) Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. *Br J Nutr* 101 (4), 598–608.

72 Rasanen, M., Lehtinen, J.C., Niinikoski, H. *et al.* (2002) Dietary patterns and nutrient intakes of 7-year-old children taking part in an atherosclerosis prevention project in Finland. *J Am Diet Assoc* 102 (4), 518–524.

73 Lee, J.W., Hwang, J. and Cho, H.S. (2007) Dietary patterns of children and adolescents analyzed from 2001 Korea National Health and Nutrition Survey. *Nutr Res Pract* 1 (2), 84–88.

74 Ovaskainen, M.L., Nevalainen, J., Uusitalo, L. *et al.* (2009) Some similarities in dietary clusters of pre-school children and their mothers. *Br J Nutr* 102 (3), 443–452.

75 Hearty, A.P. and Gibney, M.J. (2011) Dietary patterns in Irish adolescents: a comparison of cluster and principal component analyses. *Public Health Nutr* 13, 1–10.

76 Northstone, K. and Emmett, P. (2005) Multivariate analysis of diet in children at four and seven years of age and associations with socio-demographic characteristics. *Eur J Clin Nutr* 59 (6), 751–760.

77 Bel-Serrat, S., Mouratidou, T., Bornhorst, C. *et al.* (2012) Food consumption and cardiovascular risk factors in European children: the IDEFICS study. *Pediatr Obes* 2012 (6), 2047–6310.

78 Kosova, E.C., Auinger, P. and Bremer, A.A. (2013) The relationships between sugar-sweetened beverage intake and cardiometabolic markers in young children. *J Acad Nutr Diet* 113 (2), 219–227.

79 Palmer, J.R., Boggs, D.A., Krishnan, S., Hu, F.B., Singer, M. and Rosenberg, L. (2008) Sugar-sweetened beverages and incidence of type 2 diabetes mellitus in African American women. *Arch Intern Med* 168 (14), 1487–1492.

80 Nettleton, J.A., Lutsey, P.L., Wang, Y., Lima, J.A., Michos, E.D. and Jacobs, D.R., Jr. (2009) Diet soda intake and risk of incident metabolic syndrome and type 2 diabetes in the Multi-Ethnic Study of Atherosclerosis (MESA). *Diabetes Care* 32 (4), 688–694.

81 de Koning, L., Malik, V.S., Rimm, E.B., Willett, W.C. and Hu, F.B. (2011) Sugar-sweetened and artificially sweetened beverage consumption and risk of type 2 diabetes in men. *Am J Clin Nutr* 93 (6), 1321–1327.

82 Edwards, A.W. and Cavalli-Sforza, L.L. (1965) A Method for Cluster Analysis. *Biometrics* 21, 362–375.

83 Bocker, A., Derksen, S., Schmidt, E., Teckentrup, A. and Schneider, G. (2005) A hierarchical clustering approach for large compound libraries. *J Chem Inf Model* 45 (4), 807–815.

84 Varshavsky, R., Horn, D. and Linial, M. (2008) Global considerations in hierarchical clustering reveal meaningful patterns in data. *PLoS ONE* 3 (5), e2247.

85 Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag.

86 Lamrous, S. and Taileb, M. (2006) Divisive Hierarchical K-Means. In *Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, pp. 18–18.

Svensk sammanfattning

Det övergripande syftet med studien, vilken utmynnat i föreliggande rapport, bestod i att identifiera, tolka och beskriva konsumtionsmönster hos barn i två matvaneundersökningar, med användning av multivariata statistiska tekniker (*Multivariate Data Analysis*; MDA). Ansatsen var alltså att lyfta fram inbäddade mönster i datamaterialen, men också att generellt tydliggöra genomförbarhet och mervärde av i sammanhanget ny beräkningsteknik. Det till studien valda materialet omfattar data från två matvane-undersökningar med inriktning mot barn – en bland svenska och en bland danska konsumenter. Den förstnämnda hyser barn i tre åldersgrupper (fyra, åtta och elva år) och har tilldelats benämningen *Riksmaten – barn 2003*, emedan den andra är mer omfattande och känd som *Den nationale undersøgelse af danskernes kost og fysiske aktivitet*. Emellertid bearbetades endast ett utdrag av det danska datamaterialet, omfattande fyra till fem år gamla barn, i den här presenterade studien. Båda ovan nämnda data-uppsättningar har tidigare utsatts för klassisk statistisk bearbetning, dock ej mer avancerade beräkningstekniker. De dominerande analytiska teknikerna som nyttjats här är kända under benämningarna icke övervakad MDA (deskriptiv modellering) samt övervakad MDA (prediktiv modellering).

Betydande delar av de sammantagna analyserna utfördes med en i viktiga stycken egenutvecklad (icke övervakad) algoritm, vilken förmår gruppera multivariata data efter likhet (klustrings-analys) samt presentera resultaten överskådligt. Algoritmens särskilda utformning möjliggjorde dessutom inspektion och analys av identifierade konsumtionsgrupper på flera hierarkiska nivåer. Det danska materialet utföll därvid i fyra övergripande grupper och det svenska i fyra eller fem, beroende på hur data selekterades. Vi har valt termen dietiska prototyper för sålunda identifierade grupper; tre bland dem – *Traditionell*, *Läsk (med socker)/Godis & snacks* samt *Varierad (hälsosam)* – sammanfaller i stora drag med dito rapporterade i den vetenskapliga litteraturen. Dessa initiala resultat togs dock vidare till fortsatt bearbetning, bl a innebärande att dietiska prototyper utsattes för en efterföljande två-dimensionell klustrings-analys. Härvid framkom, föga överraskande, störst konsumtions-likhet mellan de två högre åldersgrupperna i det svenska materialet, men också avsevärda skillnader mellan danska och svenska barns

konsumtionsmönster: jämförelsen visade ett bi-nationellt kostbeteende som inte är alldeles olikt en invers relation. Vidare tilläts de initialt uttagna prototyperna att undergå ytterligare en sekundär klustringsanalys, särskilt utformad till att – på mer generaliserad nivå – tydliggöra deras inbördes (multi-dimensionella) avstånd. Härvid observerades bl a att en dietisk prototyp i vardera högre åldersgrupp, bland svenska barn, avviker kraftigt från de övriga i respektive ålderskategori. De två utstående prototyperna företer uppenbar ömsesidig likhet och karakteriseras av livsmedel med låg fetthalt (ffa m.a.p. mejeri-produkter) och i övrigt överlag hälsosamma kostprofiler. Vidare identifierades, i vardera av de svenska åldersgrupperna, enstaka tämligen väl avgränsade dietiska prototyper av annat slag – ömsesidigt mest likartade bland de två högre ålderskategorierna – vilka uppvisar relativt lågt intag av *Grönsaker* samt *Frukt & bär* jämväl ganska hög konsumtion av *Läsk (med socker)*. En dietisk prototyp med den sistnämnda egenskapen identifierades likaledes i det danska materialet, dock utan låg konsumtion av *Grönsaker* eller *Frukt & bär*. För övrigt framkom att spridningen bland de danska barnen är något lägre än hos de svenska, i motsvarande åldersgrupp.

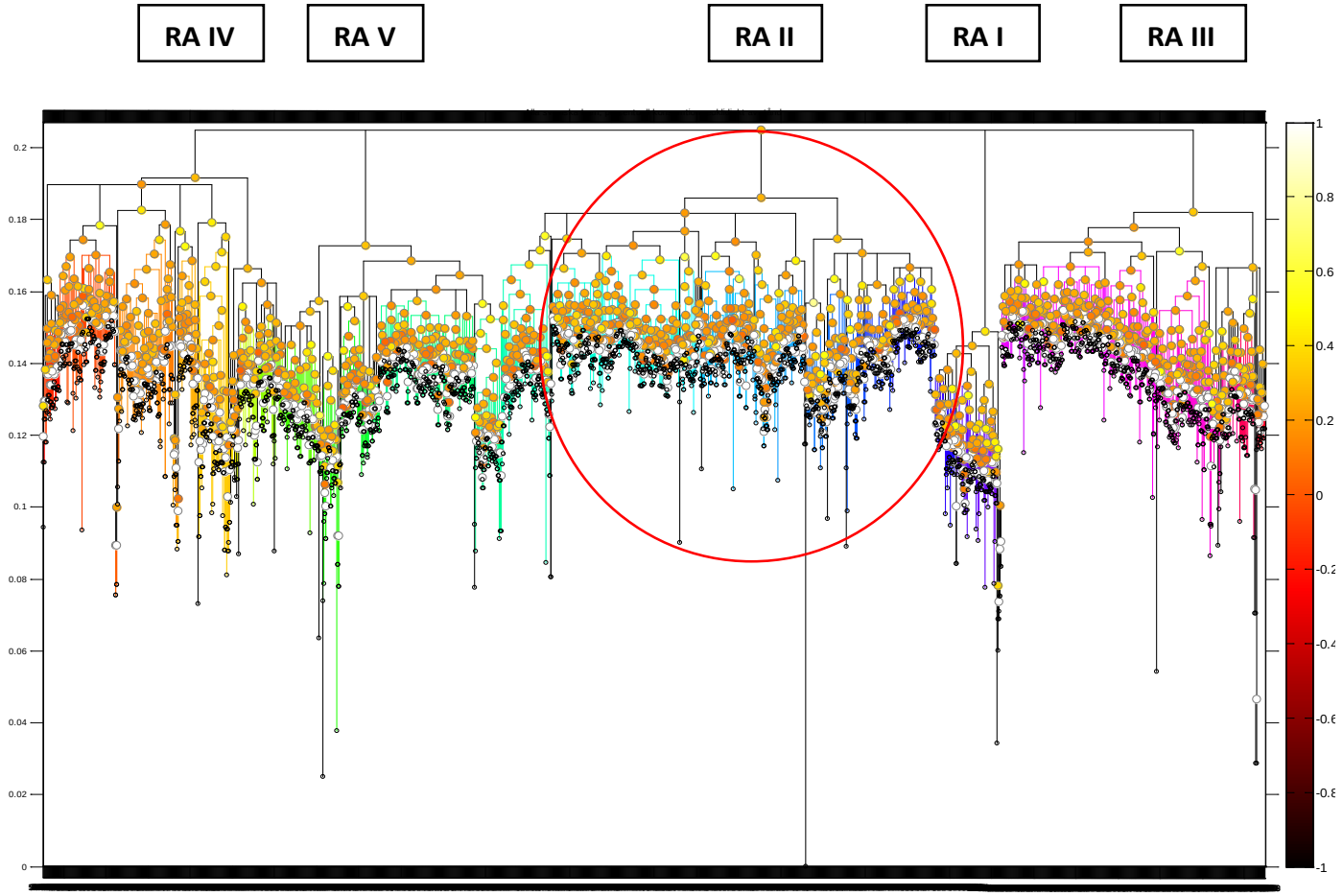
De övervakade MDA-prövningarna baserades på tränade (via de aktuella kostdata-materialen) algoritmer inom kategorierna *Random Forest* (RF) och *Nearest Shrunken Centroid* (NSC). I korthet innebär dessa analyser att låg- och högkonsumenter – för varje enskild livsmedelskategori – modellerades för korrekt klasstillhörighet, i avsaknad av det aktuella livsmedlet. Detta förfarande tydliggjorde alltså konsumtionsassociationer till respektive livsmedel. Denna delstudie gav mycket intressanta utfall som också fördjupade inledande observationer, förvärvade via icke-övervakade MDA-analyser. Några axplock bland sålunda framtagna resultat (4–5 år gamla barn) är att livsmedels-kategorierna *Grönsaker* samt *Frukt & bär* företer jämförelsevis måttliga (men inte obetydliga) skillnader mellan danska och svenska barn. Emellertid är t ex NSC-modeller för *Potatis* starkt beroende av *Ris* hos danska barn, men i stället knutna till *Korv* och *Fisk* i Sverige. På motsvarande sätt noterades stark association mellan *Bröd* och *Socker* (dvs marmelad) samt *Grönsaker* i Danmark, emedan stark koppling av *Bröd* till *Ost* och *Ceralier* framstår såsom typiskt för Sverige. Överraskande nog framkom dessutom stora skillnader över nationsgränsen med avseende på associationer till *Mjölk* samt *Kött & kyckling*.

En övergripande summering av utstående observationer, tillgängliggjorda via övervakad och icke övervakad MDA av data från ovan beskrivna matvane-undersökningar, kan formuleras enligt följande: i) En rad livsmedel tenderar grupperas på så sätt att tämligen hälsosamma,

respektive mindre nyttiga livsmedel, ofta hamnar i skilda kluster, ii) två ömsesidigt likartade och överlag hälsosamma dietiska prototyper, placerade långt från de övriga i respektive åldersgrupp, framträdde bland skolbarn i det svenska materialet, iii) danska och svenska barns (4–5 år) konsumtionsmönster är påfallande olika, bl a innebärande att flera enskilda livsmedel är knutna till skilda matvanor, iv) dietiska prototyper med betydande inslag av *Läsk (med socker)* framträdde i det danska såväl som det svenska materialet, samt v) erfarenheter från studien pekar i riktning mot att energi-baserade intags-data utgör generellt bättre underlag till klustrings-analytisk bearbetning och prediktiv modellering av nutritionella data, jämfört med dito vikts-baserade.

Figures and Tables

Figure 1. Hierarchical multi-branching cluster dendrogram, as derived by OMB-DHC, of consumption patterns within 2,495 Swedish consumers – 4, 8 and 11 years old – and encompassing the entire data set of Riksmaten – barn 2003



Consumption was expressed as weight percent and Euclidian distances were applied in the computational (clustering) process. Major (highest hierarchical level) clusters are indicated on panel top and appear accordingly in the text. Sub-populations RA I through RA V are also referred to as *Cereals*, *Milk*, *Traditional*, *Soft beverages (sweetened)/Buns & Cakes* and *Varied/Water*, respectively. The encircled aggregation holds the largest single assembly of subjects (RA II) in *Riksmaten – barn 2003*, and was targeted for inspection at several subordinate levels. See also Table 1 for mean consumption patterns (and SD) of each major sub-population. Nodes in the dendrogram ar coloured according to the average silhouette width (ASW) of the clusters appearing below, the colourbar to the right shows the scale. The y-axis shows the distance scaling.

Figure 2. Bar representations showing consumption profiles (mean intake of each food group) of the main (top level) clusters, as derived by OMB-DHC analysis of the unabridged Riksmaten – barn 2003 data set

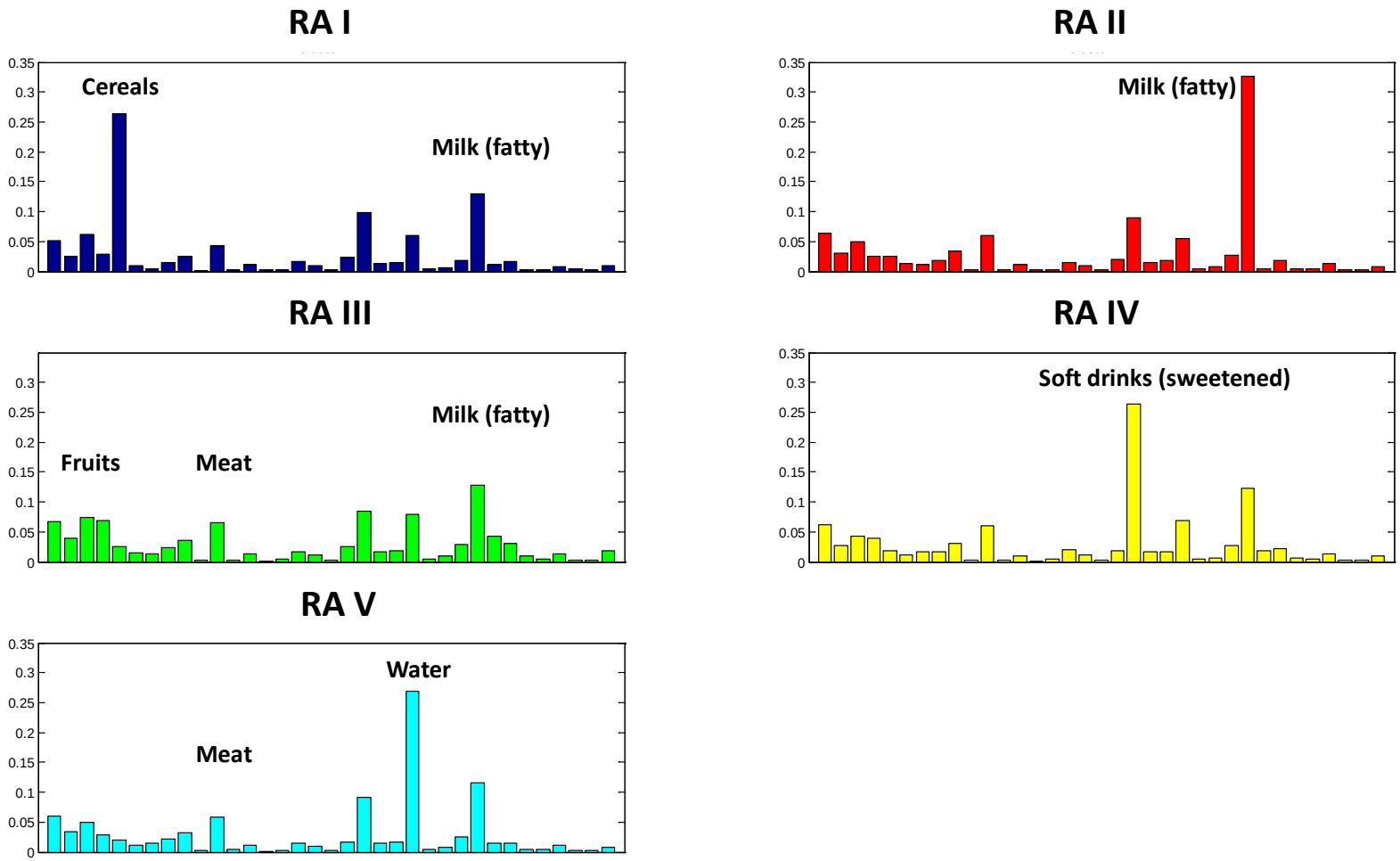
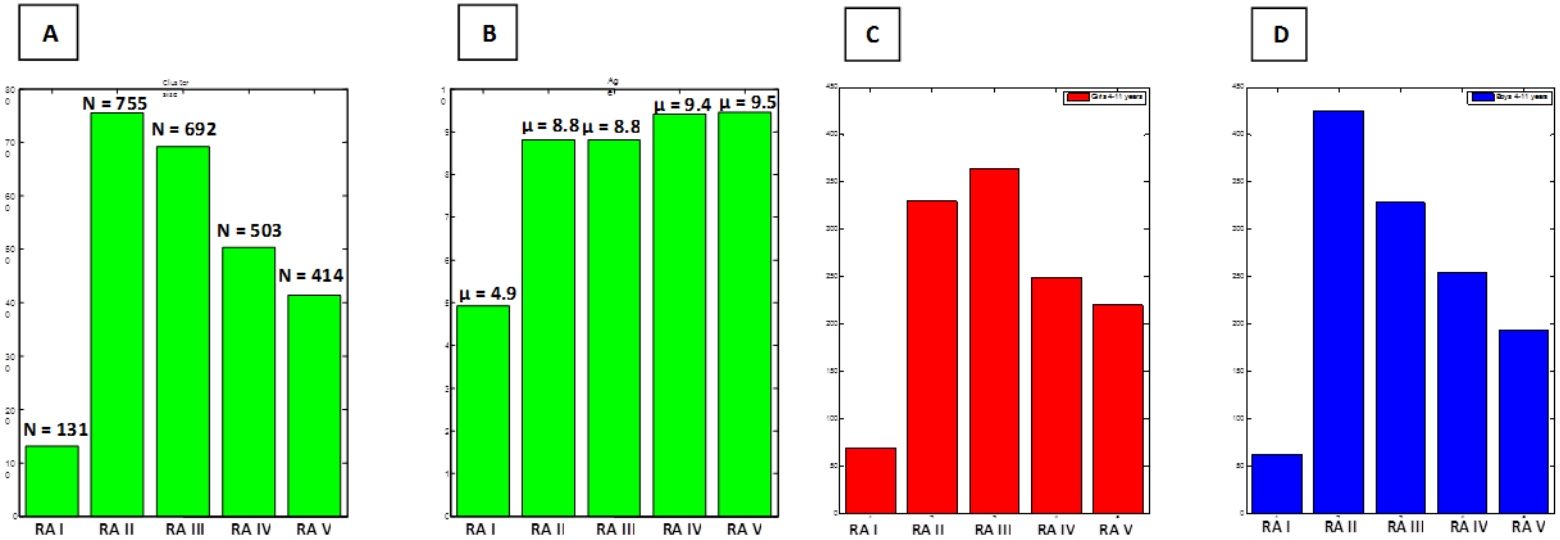
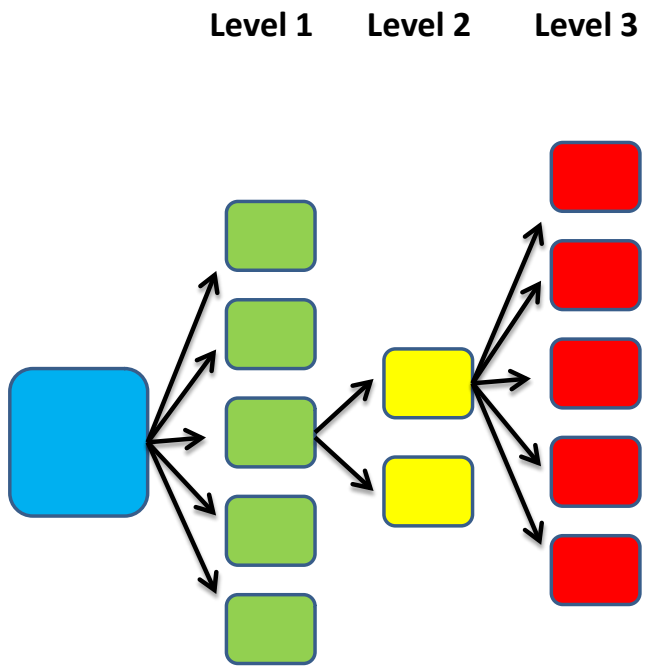


Figure 3. Distribution of age and gender across major clusters in the Swedish data set



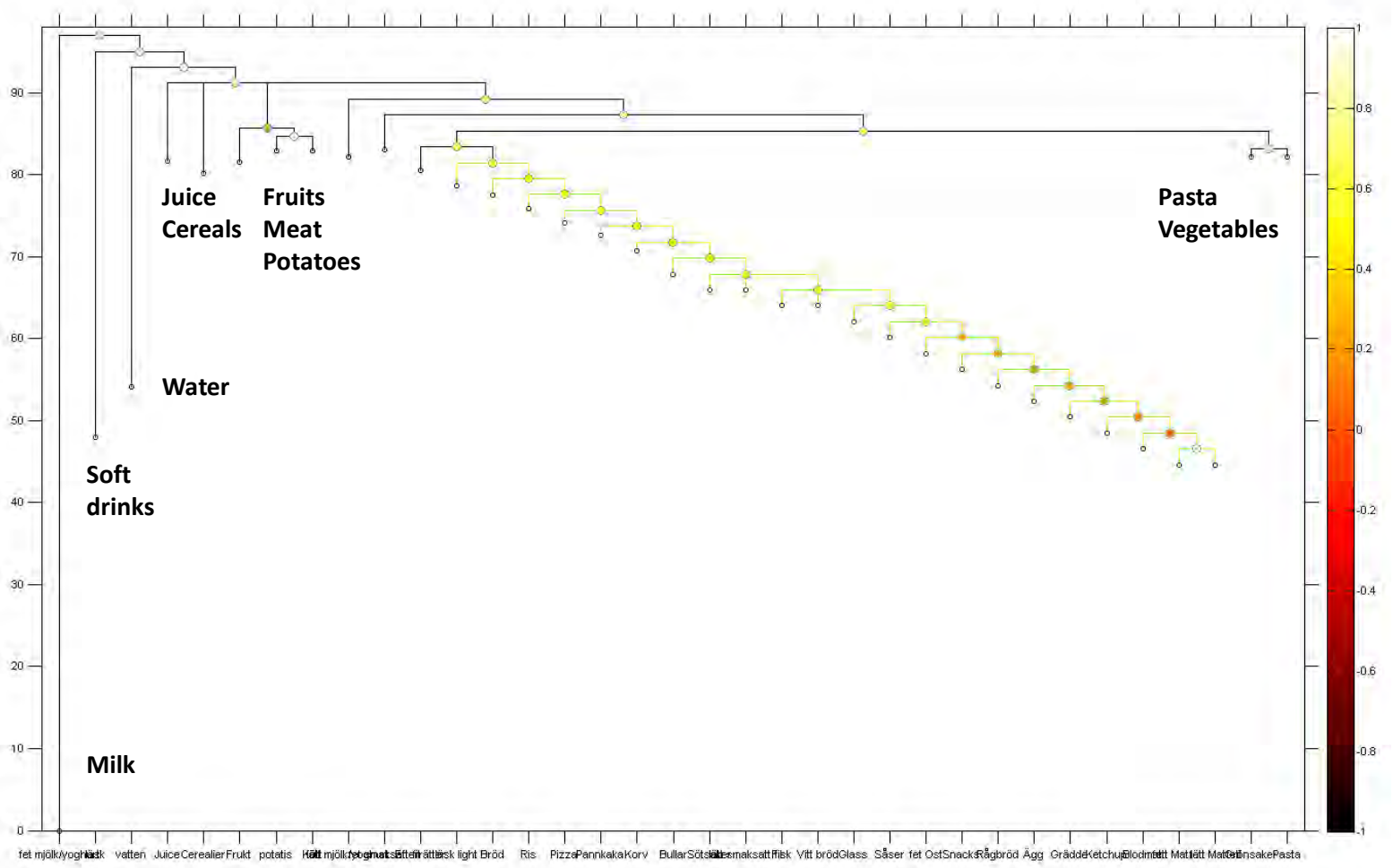
Distribution of subject numbers (A), mean ages (B) and gender (C and D) across major (RA I through RA V) clusters. The smallest cluster (RA I) is composed largely of young (4-years) consumers, whereas RA IV and RA V chiefly hold 8- and 11-year old subjects. Apart from a more even distribution across main clusters among female consumers, cluster predominance also differs slightly between genders.

Figure 4. Inspection of a single major cluster at several hierarchical levels



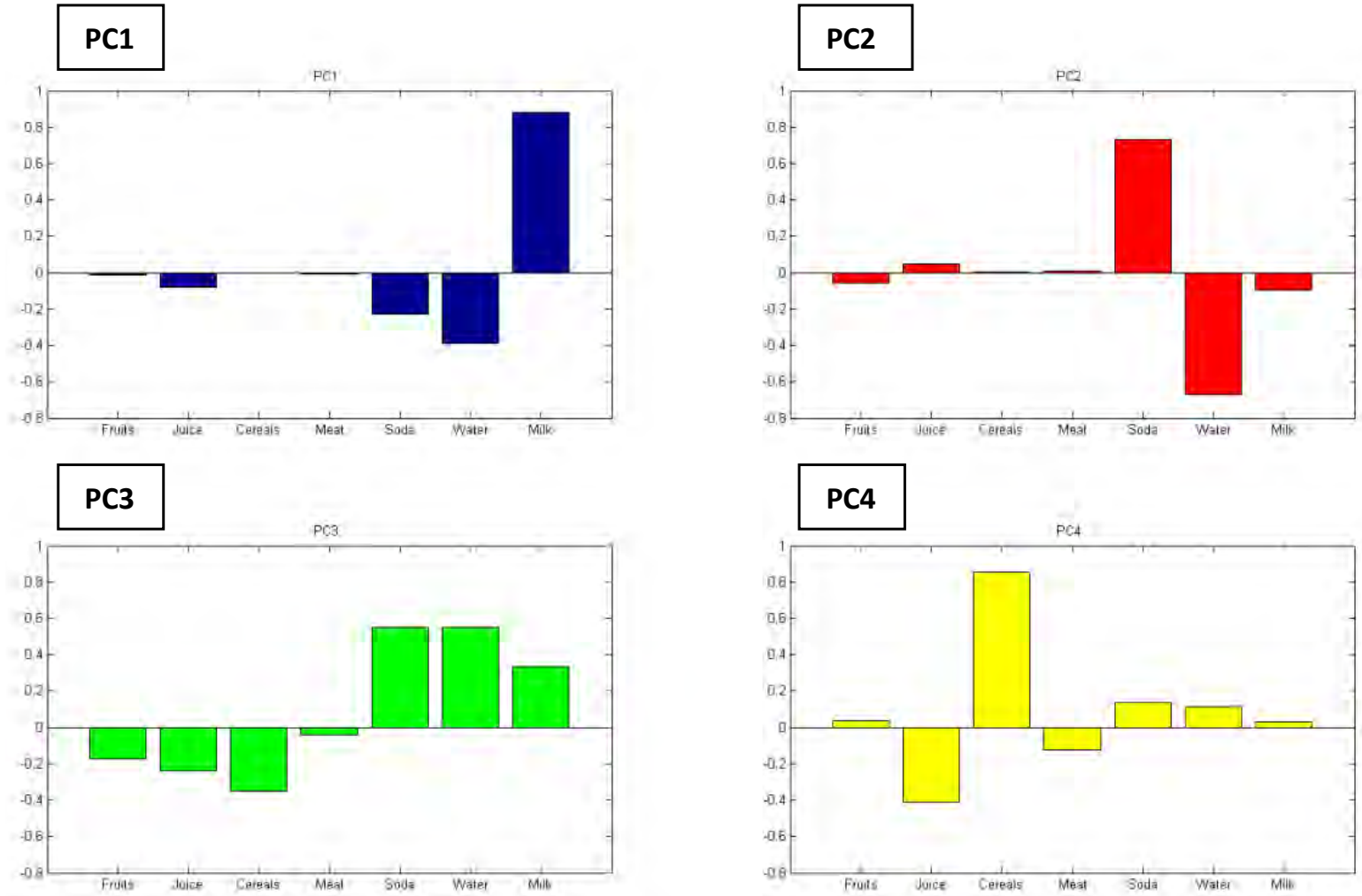
Schematic outline of targeted multi-level inspection of *Riksmaten – barn 2003* – subsequent to OMB-DHC processing - involving scrutiny of subclusters at two subordinate branching levels. Further segregation and other hierarchical splits are not shown.

Figure 5. OMB-DHC analysis of food groups – the entire Swedish data set



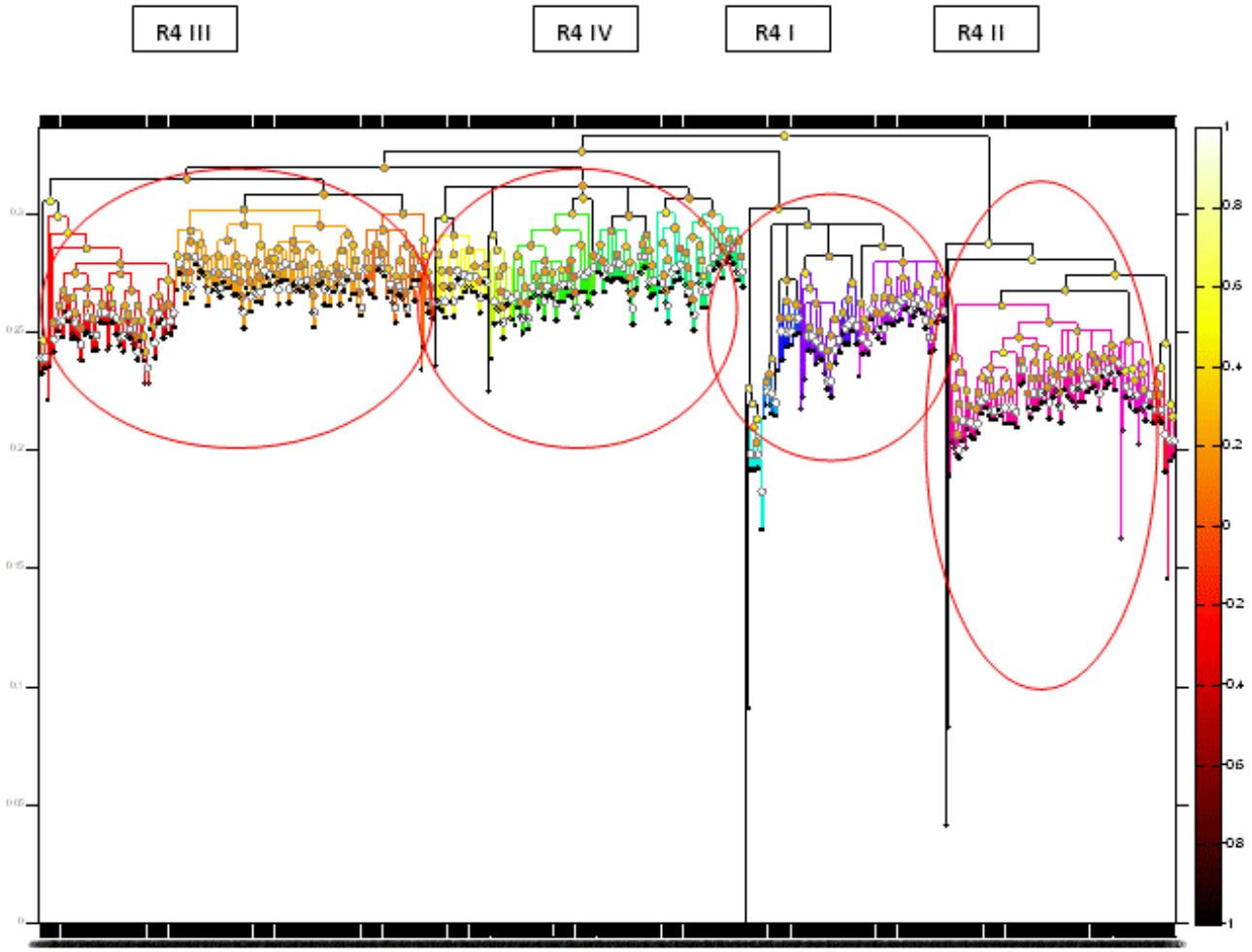
OMB-DHC output showing hierarchical structures of 35 food groups, expressed as weight percent, based on the entire *Riksmaten – barn 2003* data, encompassing children of ages 4, 8 and 11 years. Euclidian distance measure was applied in the OMB-DHC processing. Clearly, *Milk (fatty)* is the single largest food group and thus shared by most consumers, followed by *Soft beverages (sweetened)*. Moreover, *Fruits*, *Meat & poultry* and *Potato* all appear in a common cluster.

Figure 6. PCA of the entire Swedish data set



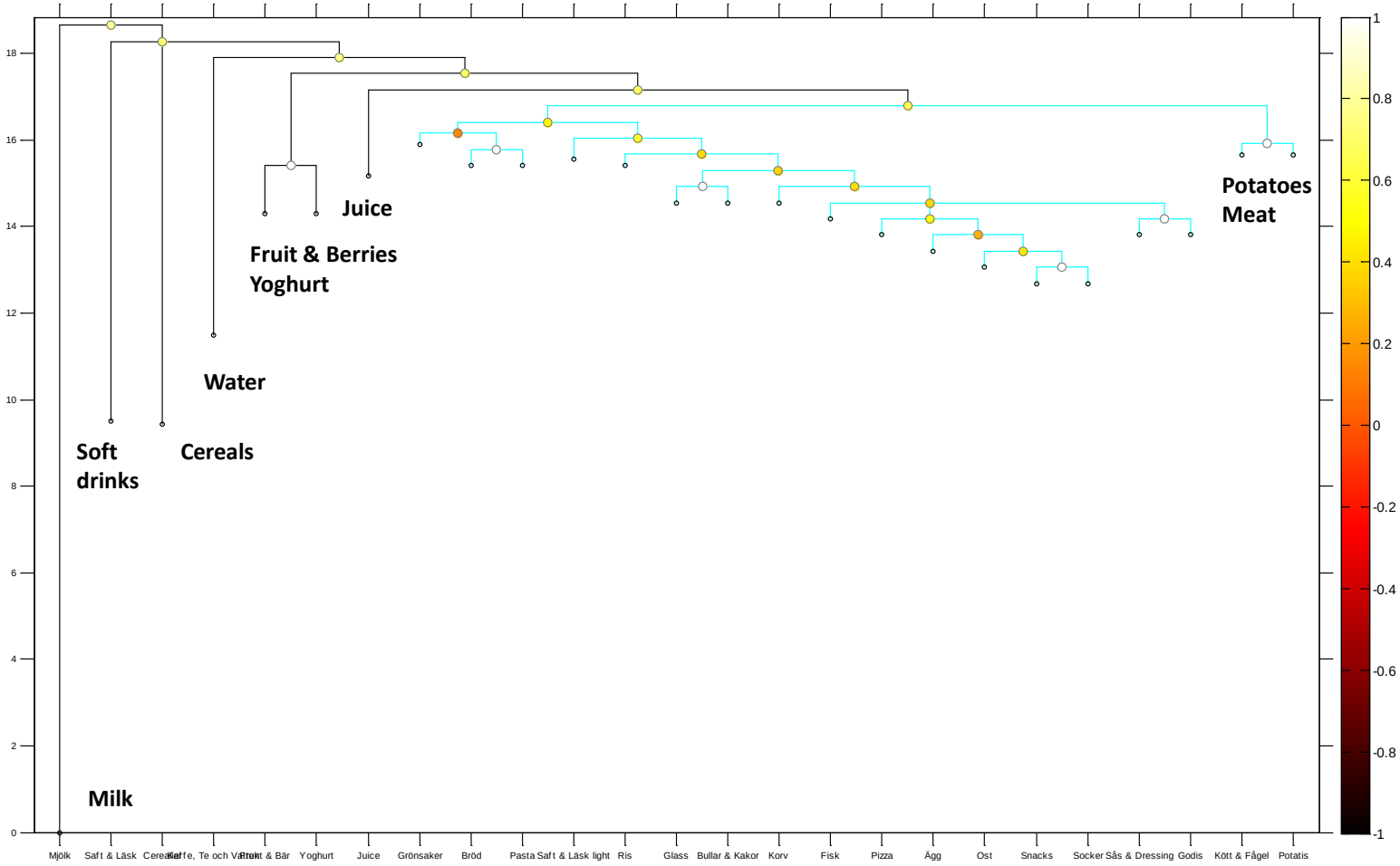
Four major principal components (PC1 through PC4), as identified by PCA of the food consumption profiles in the entire *Riksmaten – barn 2003*. For convenience, here only the seven food groups with the largest magnitudes are presented. Only a few food groups characterise each component thereby rendering them relatively easy to summarise, but caution should be exercised upon construal as to their meaning in terms of underlying consumer groups. The *Soda* category, appearing in the charts, is equivalent to *Soft beverages (sweetened)*.

Figure 7. Hierarchical cluster dendrogram – Swedish 4 year old consumers



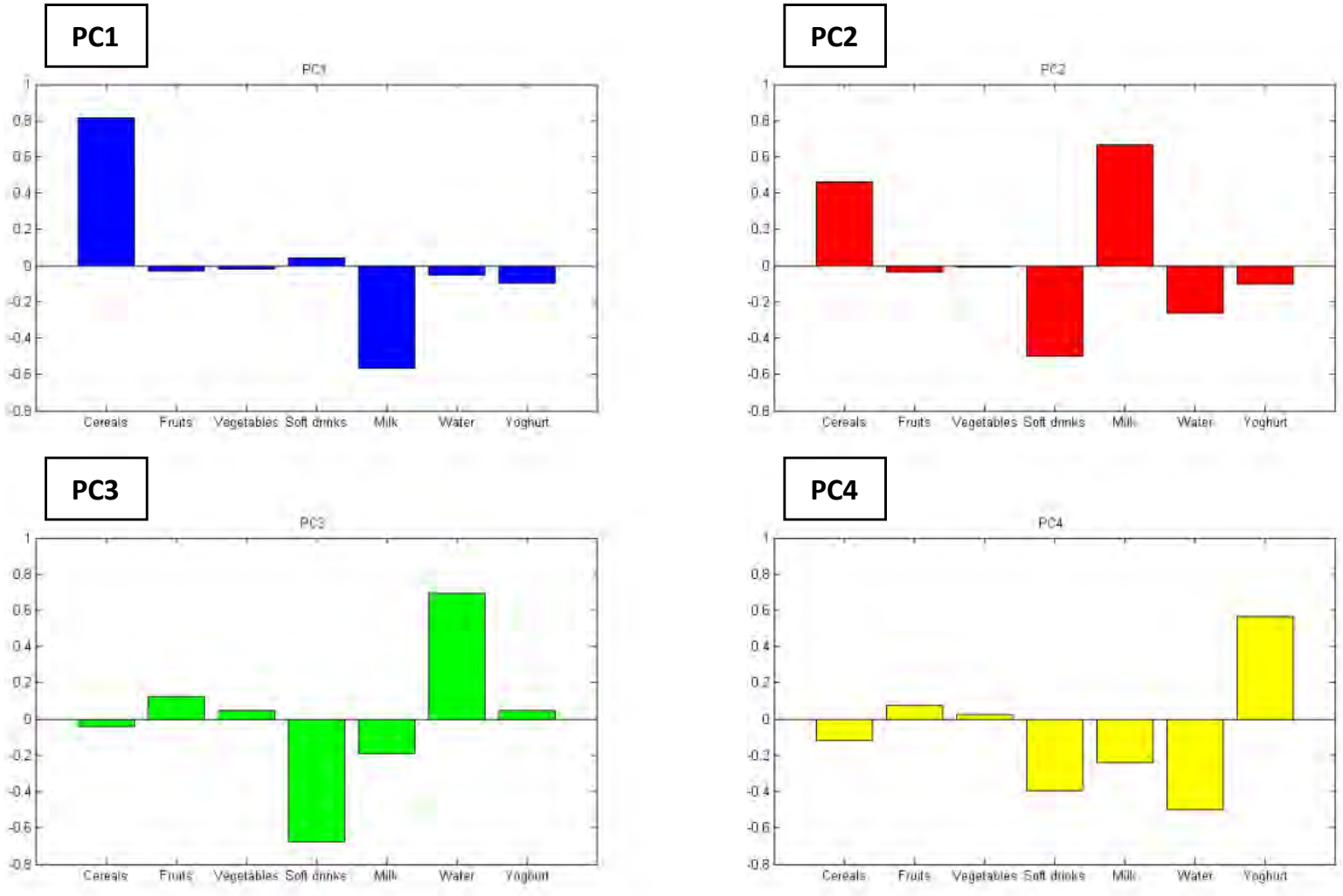
Cluster dendrogram output from OMB-DHC, based on an excerpt of *Riksmaten – barn 2003* encompassing 4 year old subjects. The data precipitate in two overarching clusters. Owing to pronounced heterogeneity of one such cluster (left-centre part of the image), three aggregations at subordinate hierarchical levels – jointly with the rightmost top level cluster – were labelled dietary prototypes (R4 I through R4 IV) and underwent subsequent analysis and comparison. The rightmost encircled cluster (R4 II) holds consumers with high intake of gruel and porridge. Alternative prototype labels, based on prominent food(s) of the respective aggregations, are as follows: *Cereals* (R4 I), *Soft beverages (sweetened)* (R4 II), *Milk* (R4 III) and *Varied/Water* (R4 IV). Mean consumption (and SD) across the various food groups of each among the aforementioned aggregations appear in Table 2.

Figure 8. OMB-DHC analysis of food groups – 4 year old Swedish consumers



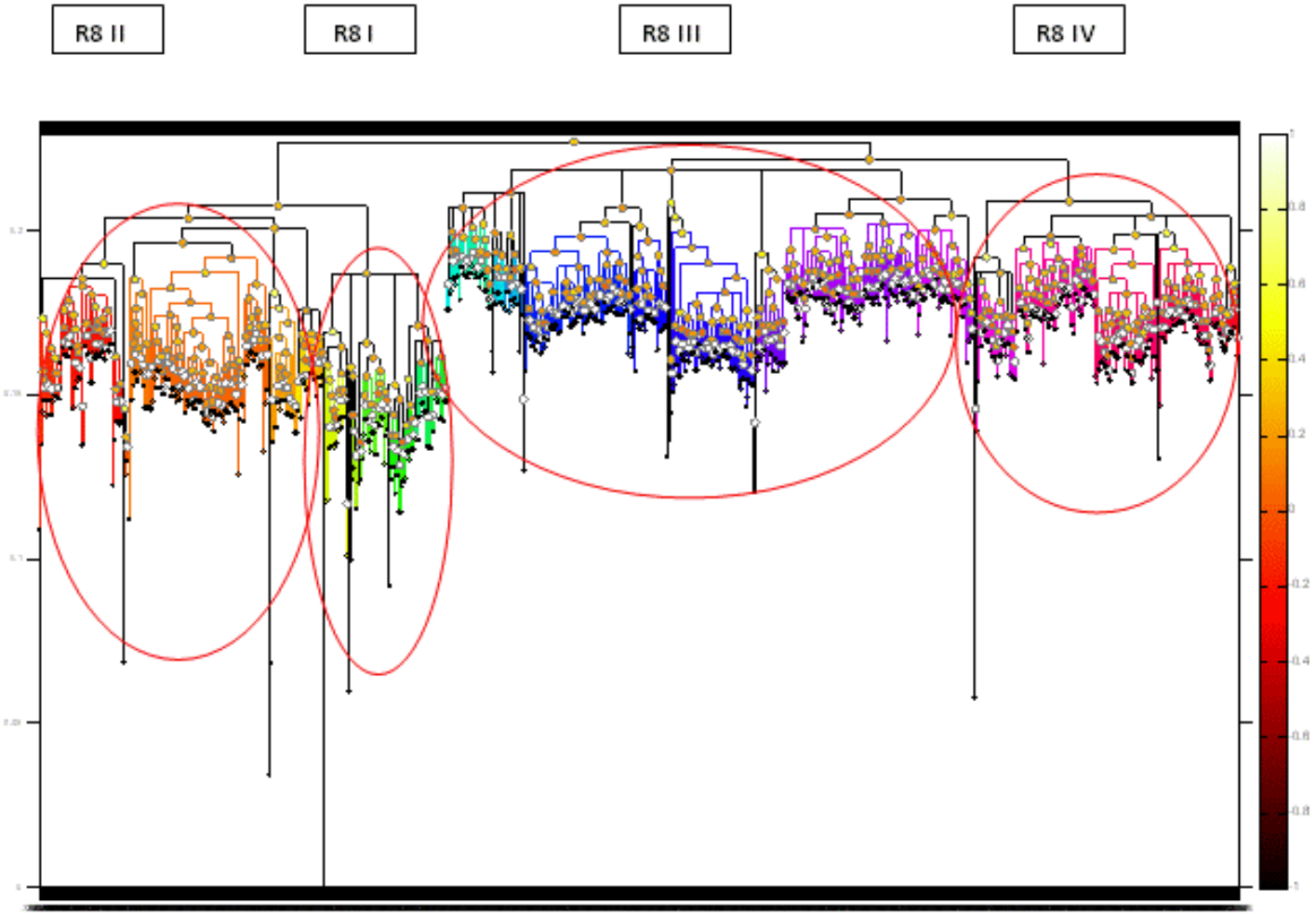
Output of OMB-DHC analysis of 25 food groups using Euclidian distance measure and based on an excerpt, 4 year old consumers, of *Riksmaten – barn 2003*. Clearly, *Milk (fatty)* is the single largest food group and thus shared by most consumers, followed by *Soft beverages (sweetened)*, *Cereals* and *Water*.

Figure 9. PCA of Swedish data - 4-year old consumers



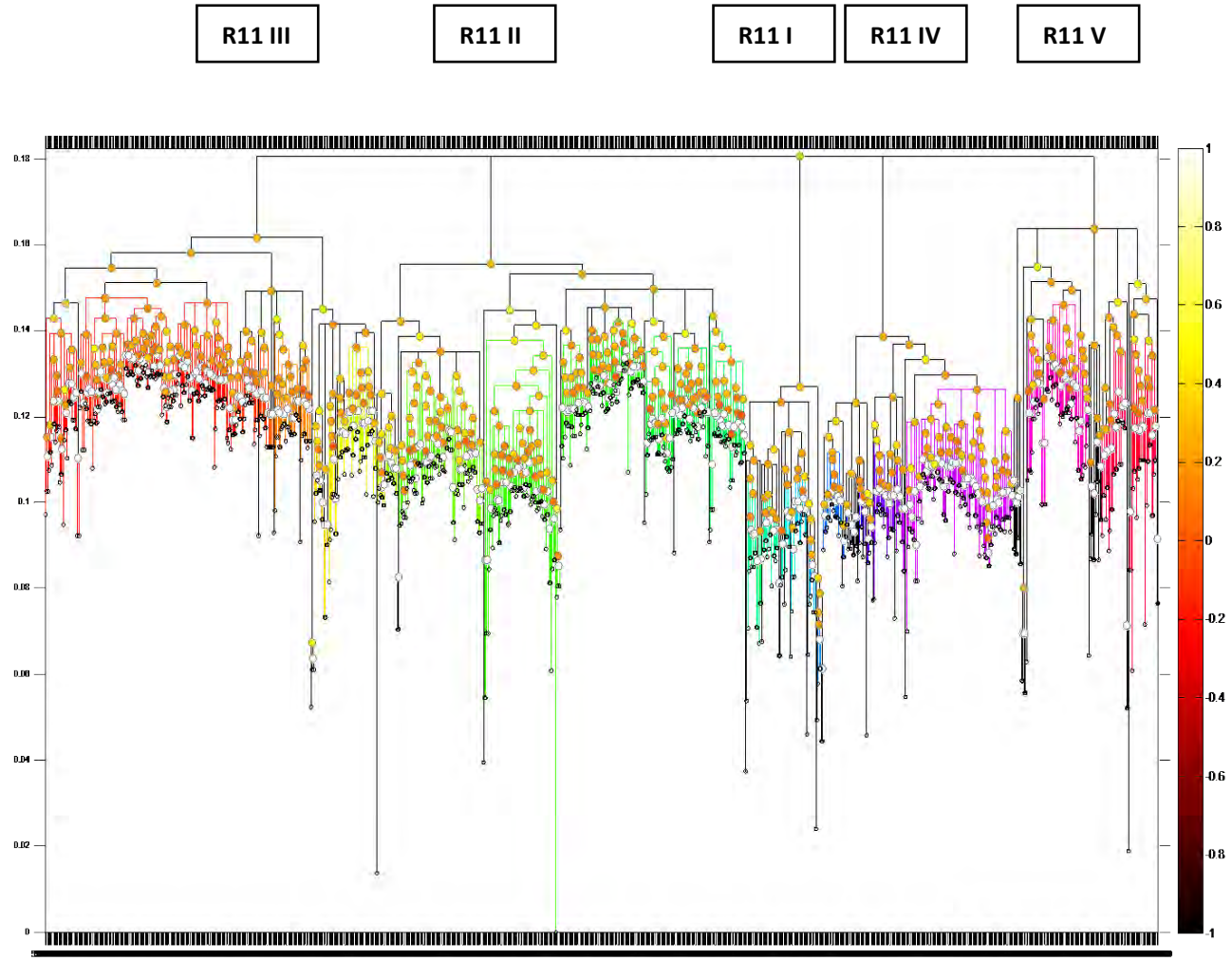
The four first principal components (PC1 through PC4), as derived from PCA of an excerpt (4 year old consumers) of *Riksmaten – barn 2003*, into food groups. Jointly these PCs carry 67% of the total variance. For convenience, here only the seven food groups with the largest magnitudes are presented. As discussed in the text, only a few food groups characterise each component thereby rendering them relatively easy to summarise, but caution should be exercised upon construal as to their meaning in terms of underlying consumer groups.

Figure 10. Hierarchical cluster dendrogram – 8 year old consumers



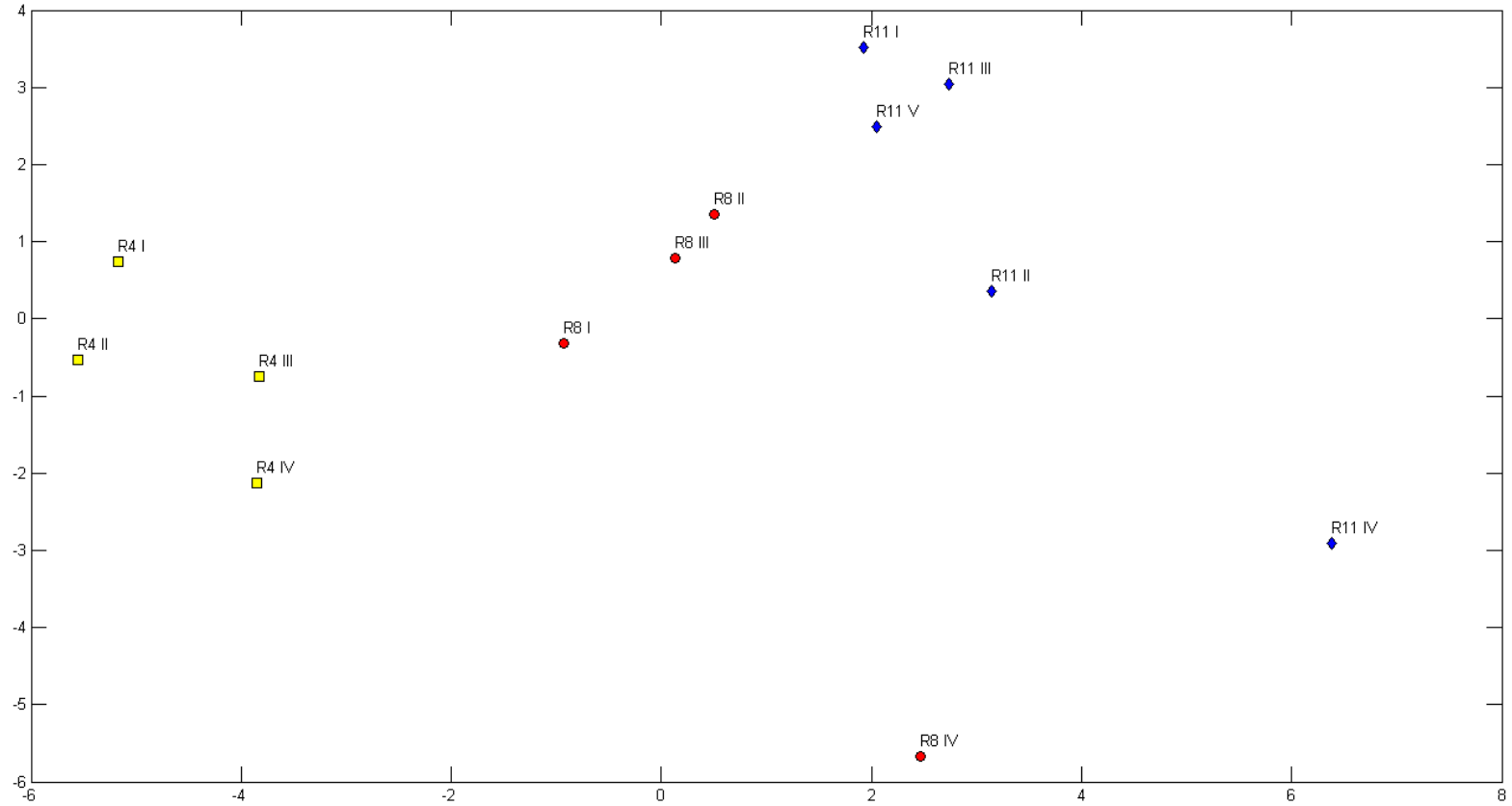
Output from OMB-DHC based on an excerpt (8 year old consumers) of *Riksmaten – barn 2003*. The data appear in two major clusters, both showing appreciable heterogeneity. Four aggregations at the second branching level (encircled) were labelled dietary prototypes, as indicated, and underwent further inspection and comparison. Alternate prototype labels – commensurate with R8 I through R8 IV – are as follows: *Varied/Water*, *Pizza/Soft beverages (sweetened)*, *Milk (fatty)* and *Milk (low-fat)/Soft beverages (light)/Juice*. Mean consumption patterns (and SD) for each accordingly defined sub-population appear in Table 3.

Figure 11. Hierarchical cluster dendrogram – 11 year old consumers



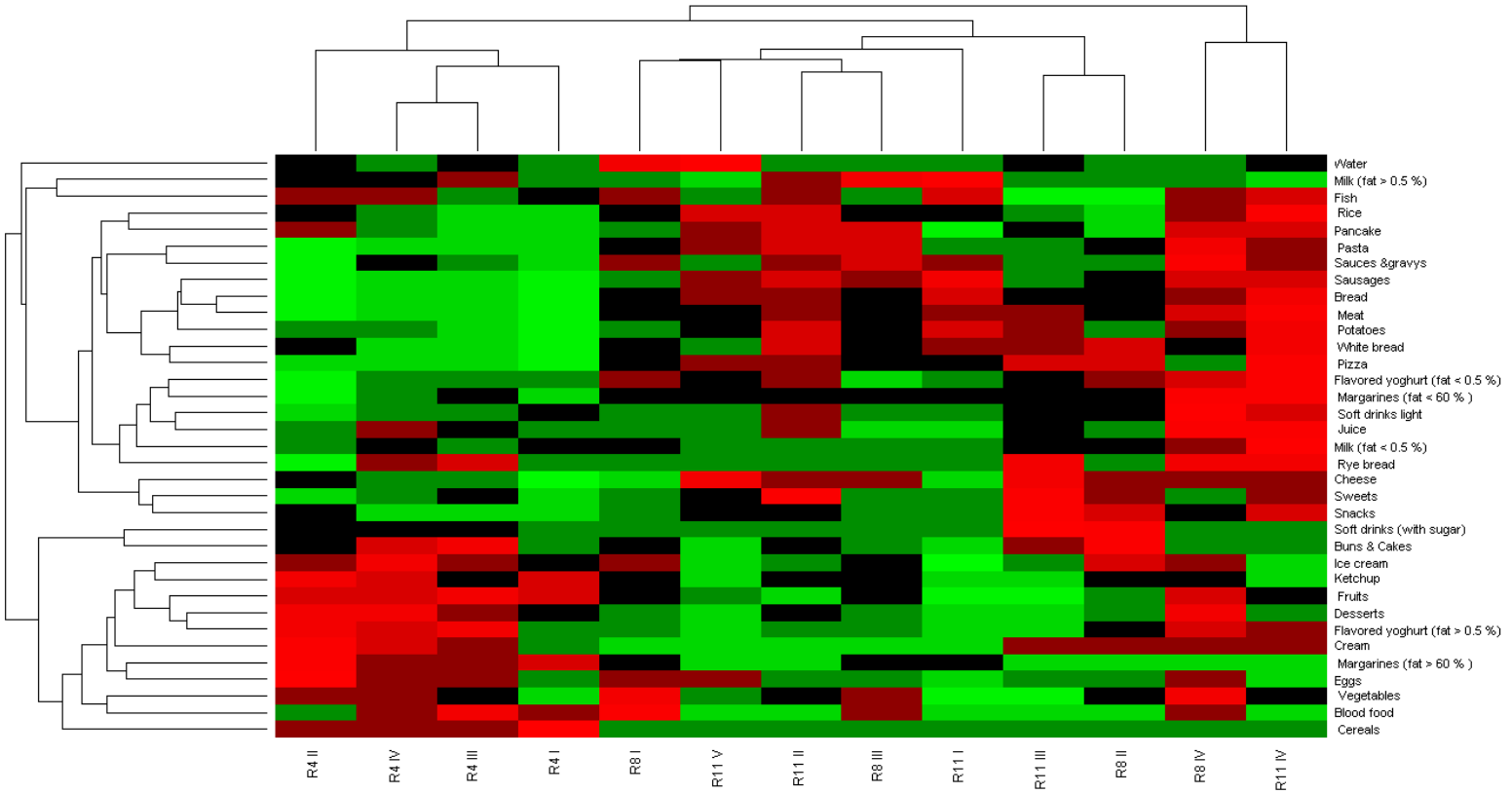
Data appear in five distinct aggregations at the highest hierarchical tier. The five clusters, referred to as dietary prototypes (R11 I through R11 V), have the following alternative labels: *Milk (fatty)*, *Milk (fatty)/Varied*, *Soft beverages (sweetened)*, *Milk (low fat)/Soft beverages (light)/Juice* and *Varied/Water*. Table 4 holds data on mean consumption (and SD) of each among the five sub-populations.

Figure 12. Classical multi-dimensional scaling of dietary prototypes – the entire Swedish data set



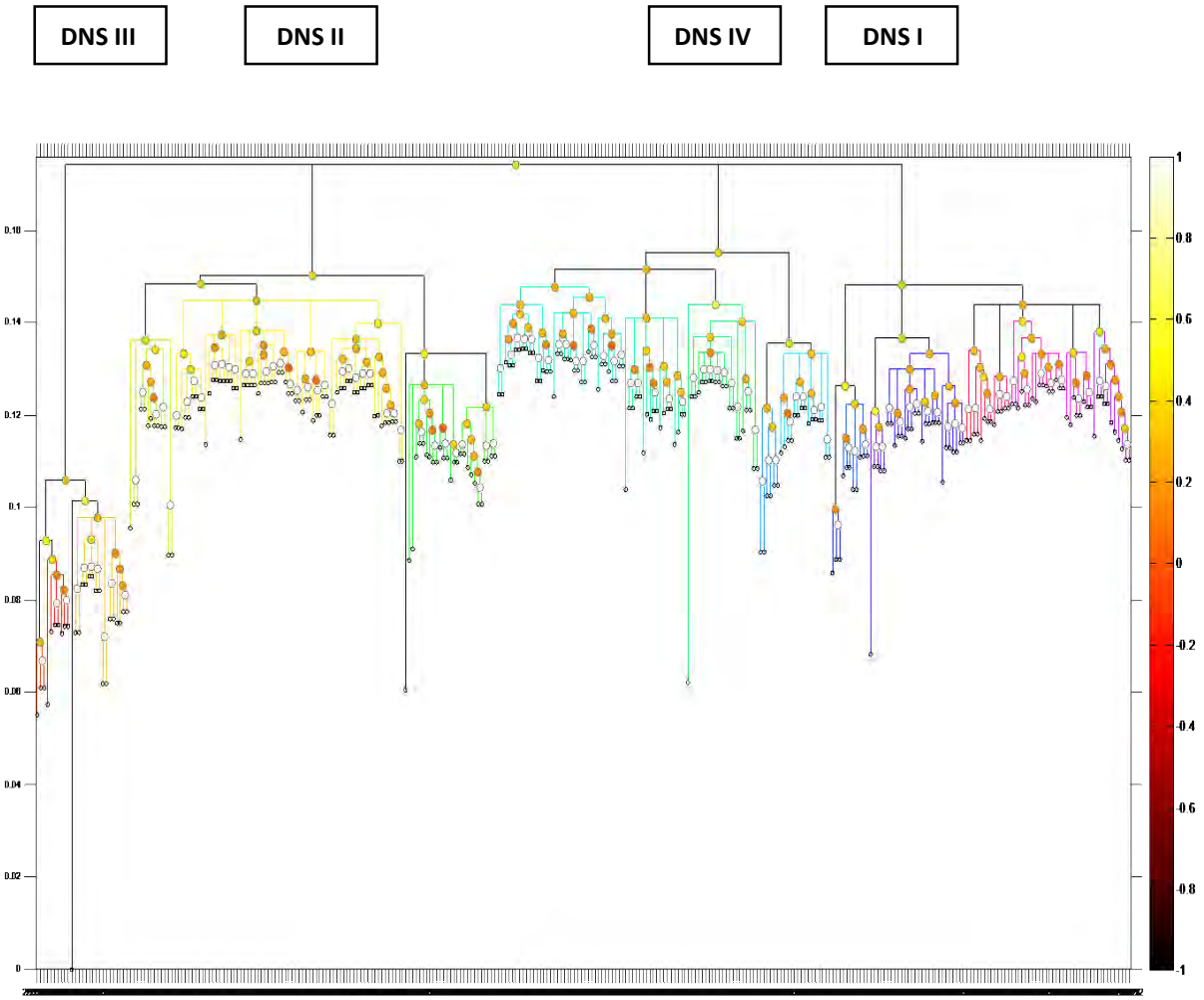
Dietary Prototype CMDS Analysis (DPCA): All major dietary prototypes (13), as derived from each of the three age groups of Swedish preschool and elementary school children (4, 8 and 11 years of age) were subjected to classical multi-dimensional scaling (CMDS). Food groups were normalised to zero mean unit variance before the analysis and Euclidian distances were employed. The display was created by reducing the accordingly derived output into two dimensions.Squares (yellow), circles (red) and diamonds (blue) refer to consumers of four, eight and eleven years, respectively. See also Figure 13, which provides additional information on prototype relationships.

Figure 13. Cluster analysis of dietary prototypes – the entire Swedish data set



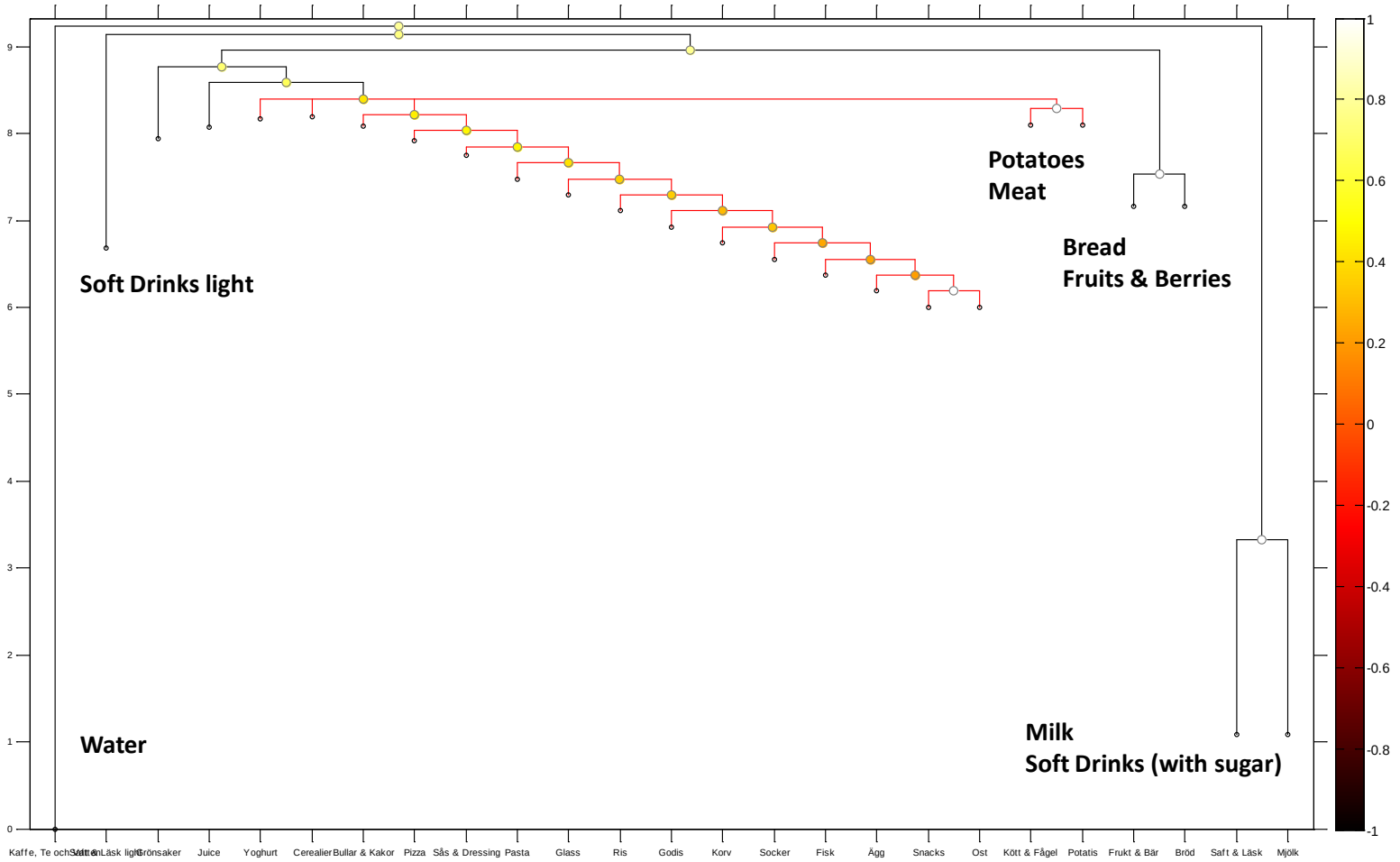
Hierarchical Prototype Bi-Cluster Analysis (HPBCA) of 13 dietary prototypes, as defined by OMB-DHC and encompassing the Swedish consumers of four, eight and eleven years of age. The lower age classes are represented by four prototypes each and the oldest category by five. Z-score data normalisation and Euclidian distances were employed. Heatmap scaling indicates high (red), low (green) or intermediate (black) consumption. Prefixes R4, R8 and R11 refer to dietary prototypes within the respective age groups. Notably, prototypes R8 IV and R11 IV, appearing here as a common cluster and sitting quite remote from age-matched counterparts (see Figure 12), have relatively healthy consumption patterns, whereas those of R4 II, R8 II and R11 III stand out (although less prominently) as comparatively unhealthy. Note that prototypes R8 II and R 11 III also comprise a distinct cluster.

Figure 14. Hierarchical cluster dendrogram – the Danish data set (4-5 year old consumers)



Multi-branching cluster dendrogram, produced by means of the OMB-DHC algorithm, fed with food survey data on Danish consumers (4–5 years old). Four distinct aggregations (DNS I through DNS IV) appear at the highest hierarchical level and were labelled dietary prototypes. Based on outstanding food group(s) inherent to each such prototype, the following alternate labels were chosen: *Soft beverages (sweetened)/Juice* (DNS I), *Milk* (DNS II), *Soft beverages (light)* (DNS III) and *Yoghurt/Fruit & berries/Water* (DNS IV). Data on mean consumption (and SD) of each cluster appear in Table 5.

Figure 15. OMB-DHC analysis of food groups – the Danish data set



Dendrogram output of OMB-DHC analysis of 25 food groups, using data on Danish subjects. Clearly, *Milk (fatty)* is the single largest food group and thus shared by most consumers, followed by *Soft beverages (sweetened)*, *Cereals* and *Water*.

Figure 16. PCA of the Danish data set

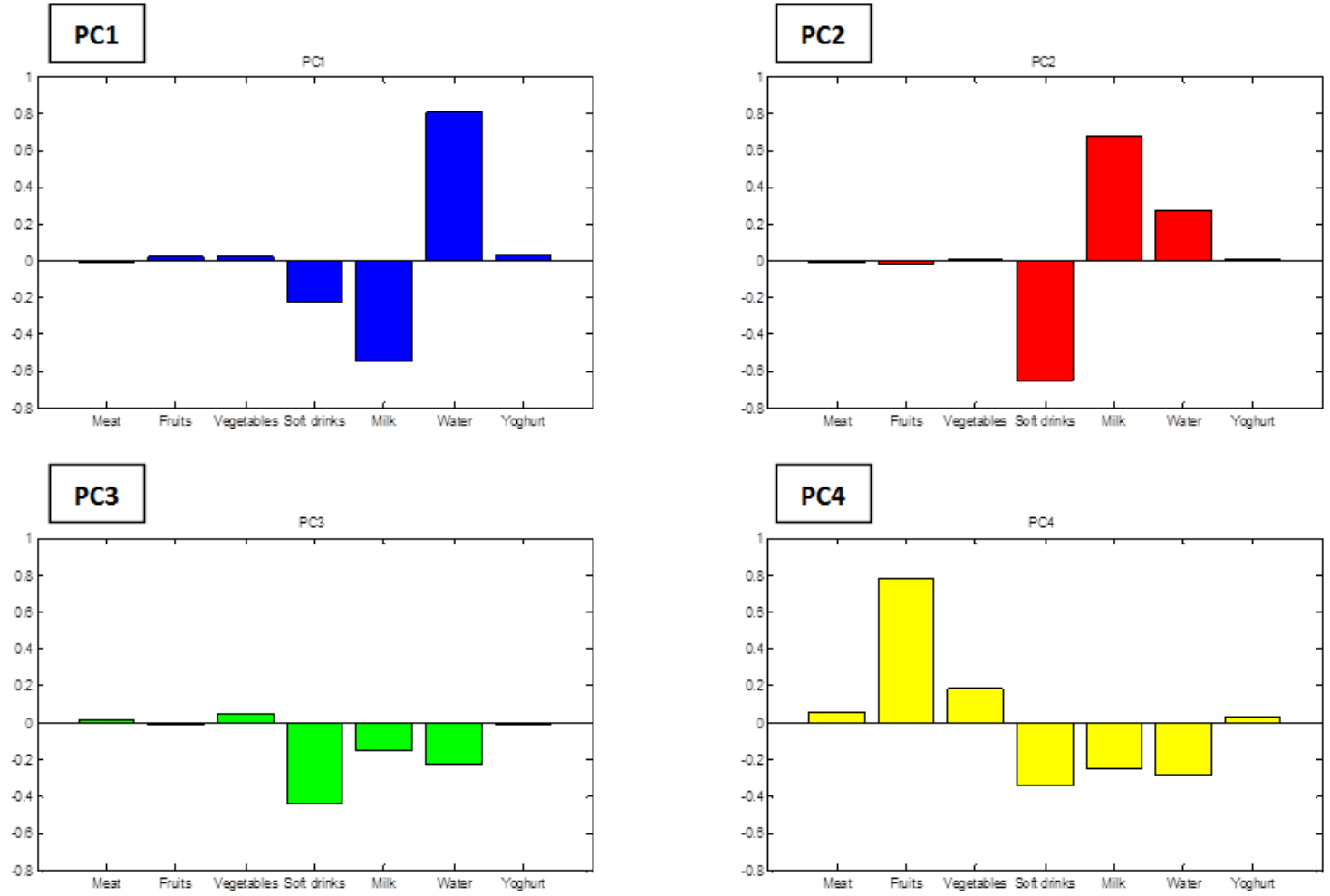
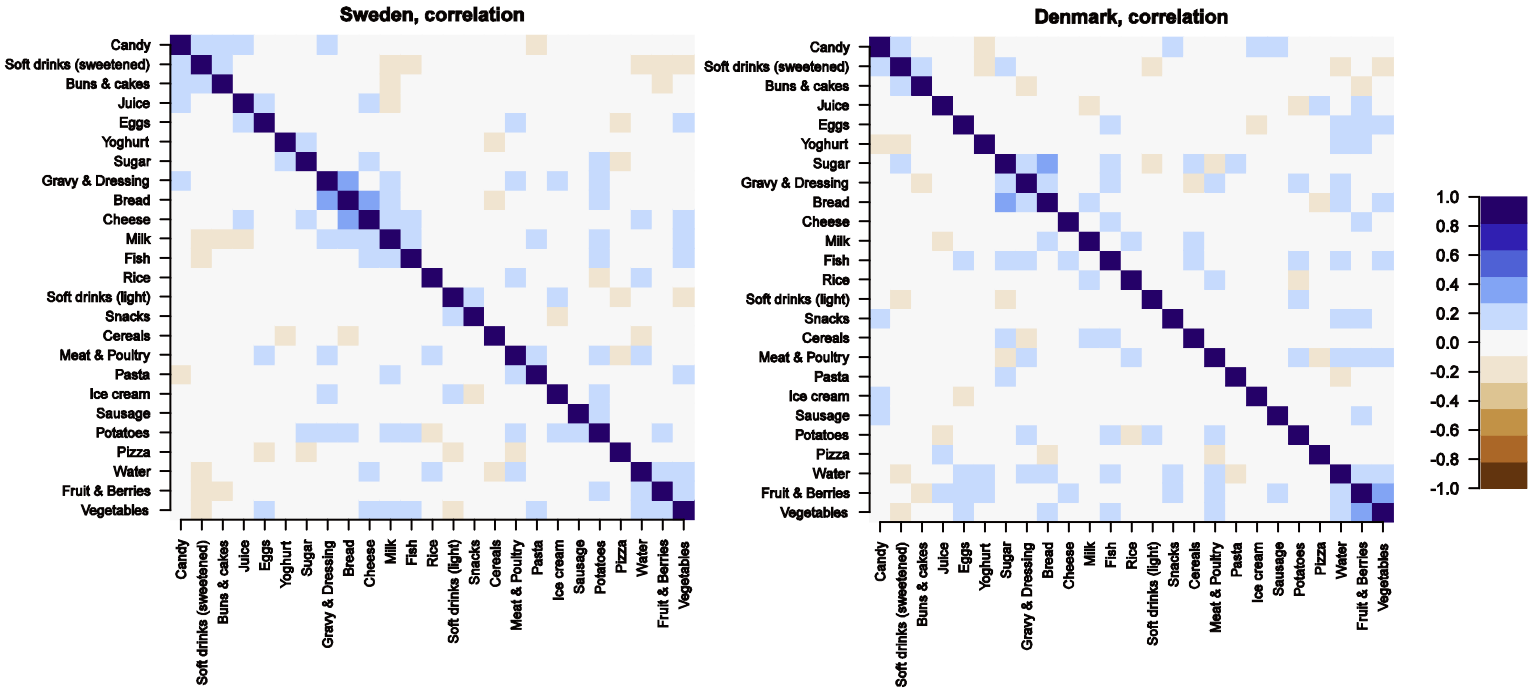


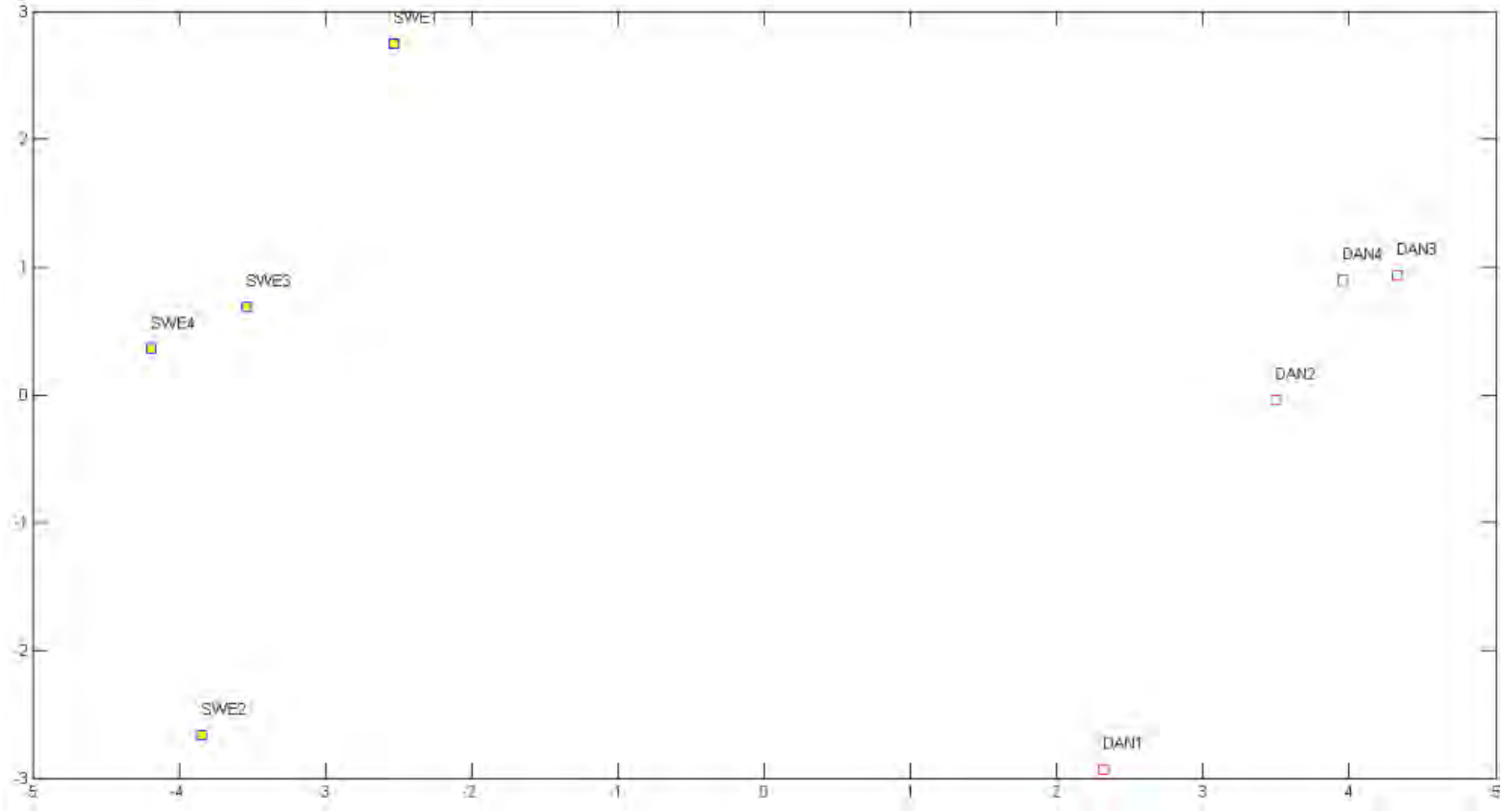
Illustration of the four first principal components (PC1 through PC4), as derived from PCA of the Danish data set (4–5 years old) into food groups. For convenience, here only the seven food groups with the largest magnitudes are presented. As discussed in the text only a few food groups build each component thereby rendering them relatively easy to summarise, but they should nonetheless be interpreted with caution with respect to underlying consumer groups.

Figure 17. Spearman correlation – Swedish and Danish data sets on preschool children



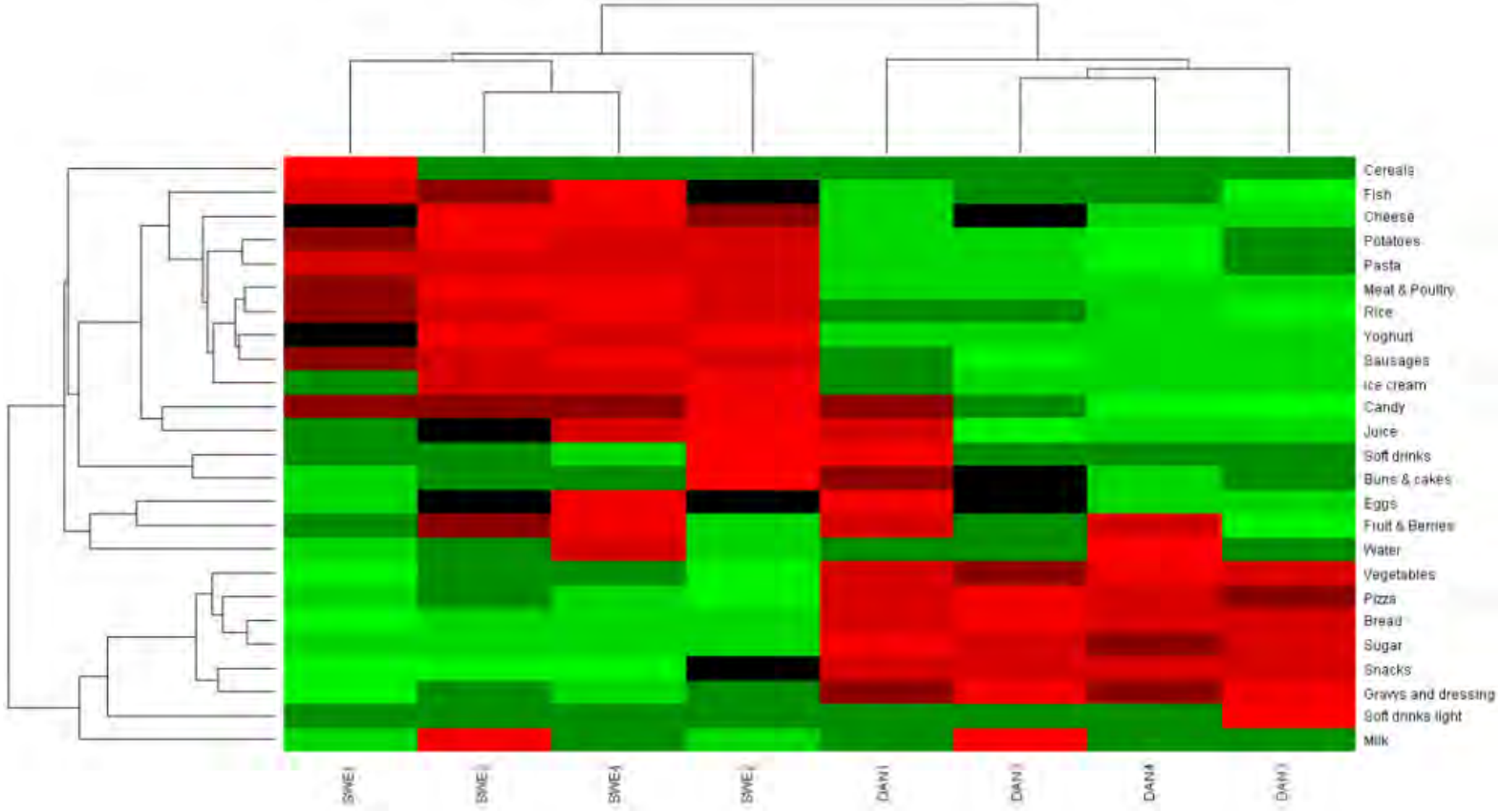
Spearman correlations between food groups for data on Swedish (left panel) and Danish (right panel) preschool children.

Figure 18. Classical multi-dimensional scaling of dietary prototypes – the Danish and Swedish data sets on preschool children



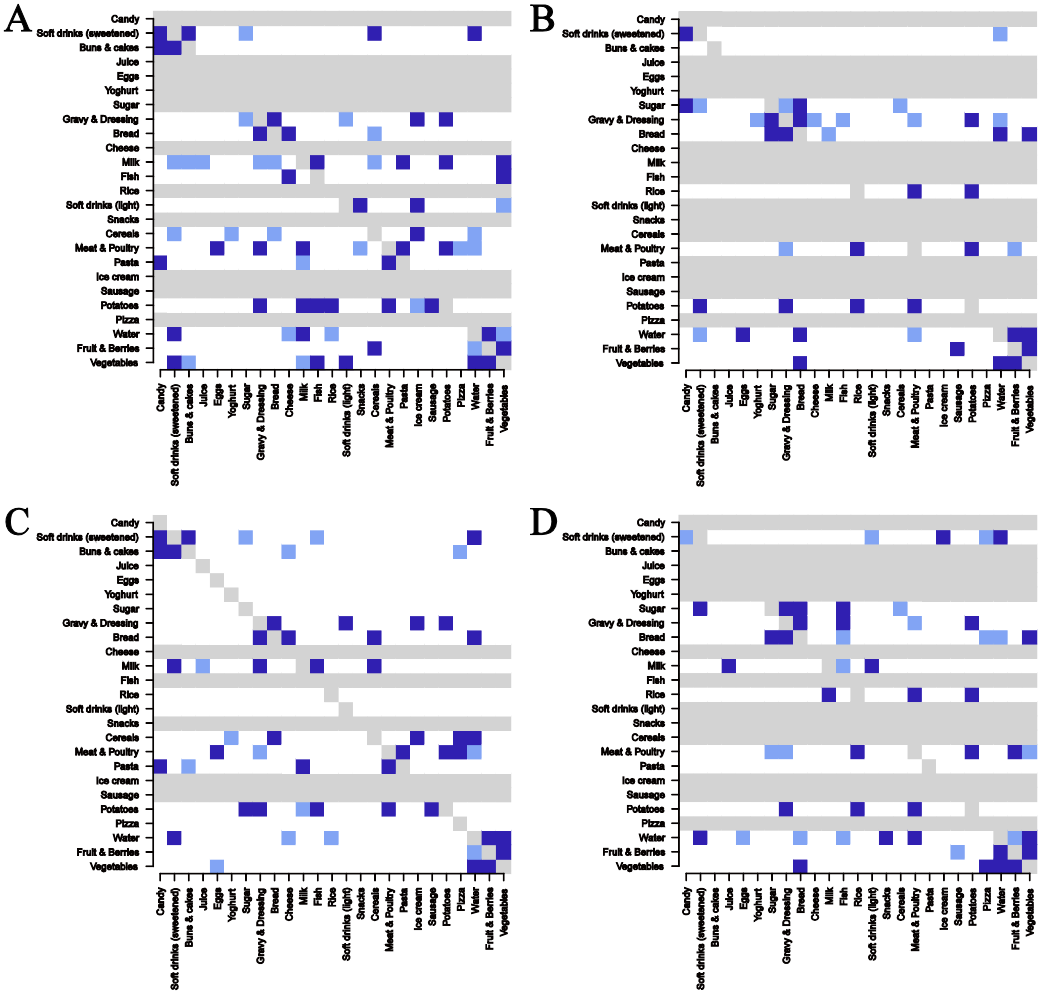
Dietary Prototype CMDS Analysis (DPCA): Classical multi-dimensional scaling (CMDS), reduced to two dimensions and based on mean consumption values of food groups (25) for each among eight sub-populations, as selected from OMB-DHC cluster analysis of Danish and Swedish preschool subjects (See Figures 7 and 14 as well as Tables 2 and 5). Variables (food groups) were normalised by means of z-score technique and Euclidian distances were employed.

Figure 19. Cluster analysis of dietary prototypes – Danish and Swedish data sets on preschool children



Hierarchical Prototype Bi-Cluster Analysis (HPBCA): Two-dimensional hierarchical clustering analysis, using eight major aggregations as input, selected from OMB-DHC cluster analysis of Danish and Swedish preschool children (See Figures 7 and 14 as well as Tables 2 and 5). Euclidian distances and z-score normalisation technique were applied. Heatmap scaling indicates high (red), low (green) or intermediate (black) consumption.

Figure 20. Predictive modelling – Danish and Swedish data sets on preschool children



Significant variables, as identified in Random Forests (RF) and Nearest Shrunken Centroid (NSC) models. A) RF, Sweden; B) RF, Denmark; C) NSC, Sweden; D) NSC, Denmark. Each horizontal line in the panels represents one model. Only models with significant OOB error rate (RF) or AUC (NSC) are considered; those being non-significant are marked grey. For a given model each variable is assigned an importance (RF) or frequency (NSC) value, if this value is significantly separated from the null model at the 0.99 level it is marked dark blue, if only at the 0.95 level light blue, non-significant values are marked white.

Table 1

Variable	RA I Mean (%)	SD (+/-)	RA II Mean (%)	SD (+/-)	RA III Mean (%)	SD (+/-)	RA IV Mean (%)	SD (+/-)	RA V Mean (%)	SD (+/-)	P- value
Cluster size (N)	**131**		**755**		**692**		**503**		**414**		
Energy (MJ)	6.3	(1.4)	7.3	(1.8)	7.2	(1.8)	7.6	(1.8)	6.9	(1.9)	
Potato	5.2	(3.6)	6.4	(4.4)	*6.7*	(4.5)	6.2	(4.3)	6.0	(4.6)	< 0.001
Vegetables	2.6	(2.4)	3.0	(2.7)	*3.9*	(3.5)	2.7	(2.6)	3.3	(3.0)	< 0.001
Fruit & berries	6.2	(5.0)	4.9	(4.3)	*7.4*	(5.6)	4.2	(4.1)	4.9	(4.6)	< 0.001
Juice	2.8	(3.8)	2.6	(4.1)	*6.9*	(7.6)	3.9	(5.6)	2.9	(4.5)	(< 0.001
Cereals	*26.3*	(8.1)	2.6	(3.4)	2.5	(3.4)	1.8	(3.1)	1.9	3.0)	< 0.001
Pancake	1.0	(1.6)	1.3	(2.0)	*1.5*	(2.5)	1.2	(2.1)	1.1	(2.2)	0.05
Pizza	0.5	(1.1)	1.2	(2.2)	1.3	(2.6)	*1.6*	(3.1)	1.5	(2.6)	< 0.001
Rice	1.5	(1.9)	1.8	(2.2)	*2.3*	(2.8)	1.7	(2.1)	2.1	(2.6)	< 0.001
Pasta	2.5	(2.5)	3.3	(3.0)	*3.6*	(3.3)	3.1	(2.9)	3.2	(2.9)	< 0.001
Rye bread	0.2	(0.5)	0.2	(0.7)	0.3	(1.0)	0.2	(0.7)	0.2	(0.5)	0.21
Meat & poultry	4.3	(2.7	6.0	(3.3)	*6.5*	(4.1)	5.9	(3.6)	5.8	(3.3)	< 0.001
Eggs	0.3	(0.5)	0.3	(0.6)	0.3	(0.8)	0.3	(0.8)	0.4	(0.9)	< 0.05
Fish	1.1	(1.4)	1.1	(1.5)	*1.3*	(1.7)	0.9	(1.3)	1.2	(1.5)	< 0.001
Blood food	0.2	(0.5)	0.2	(0.5)	0.2	(0.5)	0.1	(0.5)	0.2	(0.6)	< 0.001
Snacks	0.2	(0.5)	0.3	(0.6)	0.4	(0.9)	*0.5*	(0.8)	0.3	(0.7)	< 0.001
Buns & Cakes	1.6	(1.7)	1.4	(1.7)	1.6	(1.8)	*2.1*	(1.9)	1.5	(1.7)	< 0.001
Ice cream	0.9	(1.2)	1.0	(1.4)	1.2	(1.4)	1.1	(1.6)	0.9	(1.3)	< 0.001
Cream	0.3	(0.7)	0.2	(0.5)	0.3	(0.6)	0.3	(0.7)	0.3	(0.7)	0.39
Desserts	2.3	(2.7)	2.0	(3.1)	*2.6*	(4.2)	1.8	(3.0)	1.6	(2.6)	< 0.001
Soft beverages (sweetened)	9.8	(6.3)	9.0	(6.7)	8.4	(5.2)	*26.4*	(7.0)	9.1	(6.8)	< 0.001
Sweets	1.4	(1.3)	1.4	(1.5)	1.6	(1.9)	*1.7*	(1.6)	1.4	(1.6)	< 0.001
Sausages	1.4	(1.5)	1.9	(1.9)	1.9	(1.9)	1.7	(1.7)	1.7	(1.8)	< 0.05
Water	6.0	(5.8)	5.5	(5.8)	8.0	(5.6)	6.8	(6.2)	*26.9*	(8.0)	< 0.001
Ketchup	0.4	(0.6)	0.4	(0.6)	0.4	(0.6)	0.4	(0.6)	0.4	(0.6)	0.26
Sauces &gravys	0.6	(0.8)	0.9	(1.4)	0.9	(1.3)	0.7	v1.1)	0.7	(1.2)	< 0.05
Bread	1.7	(1.9)	2.7	(2.4)	*2.9*	(2.4)	2.7	(2.2)	2.6	(2.1)	< 0.001
Milk (fat > 0.5%)	12.9	(7.8)	*32.7*	(7.8)	12.7	(6.7)	12.3	(7.7)	11.5	(8.0)	< 0.001
Milk (fat < 0.5%)	1.2	(4.2)	0.4	(1.5)	*4.3*	(8.2)	1.8	(4.5)	1.5	(4.2)	< 0.001
Flavored yoghurt (fat > 0.5%)	1.6	(2.6)	1.8	(3.9)	3.1	(5.1)	*2.1*	(3.9)	1.5	(3.1)	< 0.001
Flavored yoghurt (fat < 0.5%)	0.3	(0.9)	0.4	(1.6)	1.0	(2.5)	0.6	(1.7)	0.5	(1.7)	< 0.001
Cheese	0.3	(0.5)	0.5	(0.7)	0.5	(0.8)	0.5	(0.8)	0.5	(0.7)	0.35
White bread	0.8	(1.0)	1.2	(1.5)	1.3	(1.7)	1.3	(1.6)	1.2	(1.4)	< 0.05
Margarines (fat > 60%)	0.4	(0.5)	0.3	(0.4)	0.3	(0.4)	0.2	(0.4)	0.2	(0.3)	< 0.001
Margarines (fat < 60%)	0.2	(0.4)	0.3	(0.4)	0.3	(0.4)	0.3	(0.3)	0.3	(0.4)	0.10
Soft beverages (light)	0.9	(2.4)	0.8	(2.8)	*1.9*	(4.9)	0.9	(2.7)	0.7	(2.5)	< 0.001

Mean percent weight consumption, and standard deviation, of 35 food groups for each of five clusters (RA I through RA V) at top hierarchical level, as derived from OMB-DHC processing of the entire Swedish data set designated *Riksmaten – barn 2003*, encompassing three age groups. Subject numbers and energy intake are also specified, for each cluster. Key features of the respective clusters – in the above-mentioned order – are as follows (only food groups appear in italic font): *Cereals*, *Milk*, *Traditional*, *Soft beverages (sweetened)*/*Buns & cakes* as well as *Varied*/*Water*.

Table 2

	R4 I Mean (%)	SD (+/-)	R4 II Mean (%)	SD (+/-)	R4 III Mean (%)	SD (+/-)	R4 IV Mean (%)	SD (+/-)	P-value
Cluster size (N)	**89**		**115**		**233**		**153**		
Energy (MJ)	**6.2**	**(1.1)**	**6.2**	**(1.1)**	**6.3**	**(1.2)**	**6.2**	**(1.1)**	***0.81***
Meat & poultry	4.4	(2.8)	4.8	(3.1)	5.1	(2.5)	5.1	(2.9)	*< 0.05*
Cereals	*28.5*	(8.5)	3.7	(5.5)	3.1	(4.0)	3.5	(4.5)	*< 0.001*
Gravys and dressing	1.2	(0.9)	1.5	(1.1)	1.4	(0.9)	1.3	(0.9)	0.07
Fruit & berries	7.4	(5.0)	7.0	(4.7)	8.3	(4.9)	9.6	(5.6)	*< 0.001*
Vegetables	2.8	(2.4)	2.6	(2.3)	3.5	(2.8)	3.9	(3.3)	*< 0.001*
Bread	2.7	(1.7)	3.3	(2.0)	3.5	(1.8)	3.4	(1.7)	*< 0.001*
Sausages	1.3	(1.4)	1.7	(1.7)	1.7	(1.6)	1.8	(1.5)	0.06
Juice	3.2	(4.2)	5.7	(7.6)	3.4	(5.1)	4.6	(5.7)	*< 0.05*
Soft beverages (sweetened)	9.0	(5.6)	*25.3*	(8.5)	9.1	(5.8)	7.7	(5.8)	*< 0.001*
Soft beverages (light)	1.1	(2.8)	1.0	(2.8)	1.1	(3.0)	1.0	(3.6)	0.31
Milk	12.7	(7.7)	12.3	(7.6)	*27.8*	(7.9)	13.4	(6.7)	*< 0.001*
Snacks	0.2	(0.4)	0.3	(0.5)	0.2	(0.3)	0.2	(0.3)	0.28
Ice cream	1.3	(1.4)	2.0	(1.6)	1.8	(1.8)	1.8	(2.1)	*< 0.001*
Candy	1.3	(1.3)	1.4	(1.2)	1.3	(1.4)	1.3	(1.5)	0.69
Water	6.1	(5.9)	5.3	(5.1)	6.5	(5.2)	*19.3*	(8.1)	*< 0.001*
Potato	4.9	(3.1)	5.4	(3.4)	5.7	(3.3)	5.4	(3.2)	0.16
Rice	1.4	(1.7)	1.7	(2.0)	1.8	(2.1)	2.0	(2.2)	*< 0.05*
Fish	1.2	(1.5)	0.9	(1.2)	1.1	(1.2)	1.4	(1.9)	0.17
Pizza	0.6	(1.1)	0.4	(1.1)	0.7	(1.4)	0.6	(1.1)	0.44
Yoghurt	3.6	(4.2)	7.1	(6.6)	7.0	(6.8)	6.9	(6.1)	*< 0.001*
Sugar	0.1	(0.3)	0.1	(0.2)	0.1	(0.2)	0.1	(0.2)	0.06
Eggs	0.3	(0.6)	0.4	(0.7)	0.4	(0.7)	0.5	(1.0)	0.75
Pasta	2.8	(2.7)	2.7	(2.7)	3.0	(2.2)	2.9	(2.2)	0.19
Cheese	0.4	(0.5)	0.5	(0.7)	0.6	(0.7)	0.6	(0.7)	*< 0.05*
Buns & cakes	1.5	(1.3)	2.8	(2.1)	1.8	(1.6)	1.8	(1.7)	*< 0.001*

Consumption data (mean % and SD; 25 food groups) of four subclusters (R4 I through R4 IV) at various hierarchical levels, derived from OMB-DHC processing of a *Riksmaten – barn 2003* excerpt, i.e. 4 years old subjects. Numbers in italic font indicate key food group of each of the aggregations which, based on such characteristics, also appear with the following respective designations: *Cereals*, *Soft beverages (sweetened)*, *Milk* and *Varied/Water*.

Table 3

Variable	R8 I Mean (%)	SD (+/-)	R8 II Mean (%)	SD (+/-)	R8 III Mean (%)	SD (+/-)	R8 IV Mean (%)	SD (+/-)	P- value
Cluster size (N)	**138**		**259**		**285**		**207**		
Energy (MJ)	**7.3**	**(1.7)**	**7.9**	**(1.7)**	**7.7**	**(1.7)**	**7.6**	**(1.6)**	**< 0.01**
Potato	5.6	(4.4)	5.6	(4.2)	6.1	(4.2)	6.7	(4.3)	*< 0.001*
Vegetables	4.0	(3.3)	3.1	(2.5)	3.5	(3.0)	4.1	(3.5)	*< 0.001*
Fruit & berries	5.6	(4.6)	4.7	(4.0)	5.2	(4.5)	7.3	(5.2)	*< 0.001*
Juice	3.5	(4.9)	3.2	(4.7)	1.9	(3.2)	*7.7*	(7.3)	*< 0.001*
Cereals	2.0	(3.3)	2.1	(3.7)	2.1	(2.7)	2.2	(3.8)	*< 0.05*
Pancake	1.1	(1.8)	1.0	(1.8)	1.5	(2.1)	1.5	(2.4)	*< 0.05*
Pizza	1.3	(2.2)	*1.8*	(3.8)	1.2	(2.1)	0.9	(1.7)	0.09
Rice	2.0	(2.4)	1.4	(1.9)	1.9	(2.2)	2.2	(2.9)	*< 0.05*
Pasta	3.2	(3.0)	3.3	(2.9)	3.6	(3.0)	4.1	(3.5)	*< 0.05*
Rye bread	0.2	(0.4)	0.2	(0.6)	0.2	(0.6)	0.3	(1.3)	0.24
Meat & poultry	5.9	(3.0)	5.9	(3.0)	6.1	(3.5)	6.8	(3.6)	0.06
Eggs	0.4	(0.9)	0.3	(0.7)	0.3	(0.7)	0.4	(0.8)	*< 0.05*
Fish	1.2	(1.5)	0.9	(1.3)	1.1	(1.4)	1.2	(1.4)	*< 0.05*
Blood food	0.3	(0.8)	0.1	(0.3)	0.2	(0.5)	0.2	(0.5)	0.35
Snacks	0.3	(0.6)	0.5	(0.6)	0.3	(0.6)	0.4	(0.7)	*< 0.001*
Buns & Cakes	1.6	(1.5)	2.2	(2.1)	1.4	(1.6)	1.5	(1.8)	*< 0.001*
Ice cream	1.2	(1.4)	1.3	(1.8)	1.0	(1.3)	1.2	(1.4)	0.08
Cream	0.2	(0.6)	0.3	(0.7)	0.2	(0.6)	0.3	(0.6)	0.45
Desserts	1.7	(2.6)	1.8	(2.6)	1.8	(2.8)	2.7	(4.0)	0.07
Soft beverages (sweetened)	9.2	(5.9)	*24.1*	(6.2)	9.2	(6.2)	8.5	(5.0)	*< 0.001*
Sweets	1.3	(1.3)	1.6	(1.4)	1.4	(1.4)	1.4	(1.5)	0.06
Sausages	1.6	(1.6)	1.7	(1.8)	1.9	(1.8)	2.0	(1.9)	0.08
Water	*23.0*	(7.6)	5.8	(5.5)	4.6	(5.4)	5.4	(4.7)	*< 0.001*
Ketchup	0.4	(0.6)	0.4	(0.5)	0.4	(0.5)	0.4	(0.5)	0.14
Sauces & gravys	0.8	(1.3)	0.7	(1.1)	0.9	(1.5)	1.1	(1.5)	*< 0.001*
Bread	2.6	(1.9)	2.7	(2.2)	2.7	(2.2)	2.8	(2.2)	0.63
Milk (fat > 0.5%)	12.7	(7.9)	14.6	(8.2)	*33.9*	(7.4)	14.1	(6.6)	*< 0.001*
Milk (fat < 0.5%)	1.8	(4.5)	2.5	(5.4)	0.4	(1.6)	*4.0*	(7.3)	*< 0.001*
Flavored yoghurt (fat > 0.5%)	1.6	(2.9)	2.1	(3.6)	1.8	(4.2)	2.9	(4.7)	*< 0.05*
Flavored yoghurt (fat < 0.5%)	0.6	(1.9)	0.7	(1.9)	0.2	(1.0)	0.8	(2.4)	*< 0.001*
Cheese	0.4	(0.7)	0.5	(0.8)	0.5	(0.8)	0.5	(0.7)	0.74
White bread	1.2	(1.4)	1.5	(1.5)	1.2	(1.5)	1.2	(1.4)	0.06
Margarines (fat > 60%)	0.3	(0.4)	0.2	(0.3)	0.3	(0.4)	0.2	(0.3)	*< 0.05*
Margarines (fat < 60%)	0.3	(0.4)	0.3	(0.3)	0.3	(0.4)	0.4	(0.4)	0.16
Soft beverages (light)	0.7	(2.7)	1.0	(3.0)	0.7	(2.1)	*2.7*	(5.8)	*< 0.001*

Data on consumption (mean % and SD) of four subclusters (R8 I through R8 IV) at the second hierarchical level, as revealed by OMB-DHC analysis of an excerpt (8 year old subjects; 35 food groups) of *Riksmaten – barn 2003*. Clusters are also referred to as *Varied/Water*, *Pizza/Soft beverages (sweetened)*, *Milk (fatty)* and *Milk (low fat)/Soft beverages (light)/Juice*.

Table 4

Variable	R11 I Mean (%)	SD (+/-)	R11 II Mean (%)	SD (+/-)	R11 III Mean (%)	SD (+/-)	R11 IV Mean (%)	SD (+/-)	R11 V Mean (%)	SD (+/-)	P- value
Cluster size (N)	**152**		**291**		**231**		**177**		**165**		
Energy (MJ)	**7.4**	**(1.7)**	**7.4**	**(1.7)**	**7.5**	**(1.8)**	**7.5**	**(1.6)**	**7.5**	**(1.7)**	**0.89**
Potato	7.4	(4.8)	7.6	(5.0)	6.8	(4.5)	8.0	(5.5)	6.3	(4.8)	P < 0.05
Vegetables	2.5	(2.6)	3.1	(3.2)	2.5	(2.7)	3.2	(3.4)	3.0	(3.0)	P < 0.05
Fruit & berries	3.1	(3.6)	4.0	(4.0)	3.1	(3.7)	6.0	(6.1)	4.3	(4.7)	P < 0.001
Juice	2.0	(3.8)	5.2	(6.1)	3.7	(5.3)	*7.6*	(9.3)	2.5	(3.8)	P < 0.001
Cereals	2.6	(3.3)	2.5	(3.9)	1.5	(2.7)	2.0	(3.3)	1.7	(2.9)	P < 0.001
Pancake	0.9	(1.7)	1.5	(2.4)	1.2	(2.2)	1.5	(3.3)	1.3	(2.7)	0.19
Pizza	1.4	(2.3)	1.7	(2.6)	2.0	(3.0)	2.5	(3.9)	1.6	(3.0)	0.21
Rice	2.0	(2.5)	2.3	(2.8)	1.6	(2.1)	2.8	(3.0)	2.4	(3.0)	P < 0.001
Pasta	3.0	(3.4)	3.7	(3.6)	3.0	(3.0)	3.5	(3.4)	3.4	(3.0)	0.12
Rye bread	0.2	(0.7)	0.2	(0.7)	0.3	(0.8)	0.3	(1.0)	0.2	(0.6)	0.97
Meat & poultry	6.6	(3.5)	6.3	(3.5)	6.4	(3.9)	8.1	(5.5)	5.8	(3.3)	P < 0.001
Eggs	0.2	(0.5)	0.3	(0.6)	0.3	(0.9)	0.2	(0.6)	0.4	(0.8)	P < 0.05
Fish	1.3	(1.8)	1.2	(2.0)	0.9	(1.3)	1.3	(1.8)	1.1	(1.6)	0.14
Blood food	0.1	(0.4)	0.1	(0.4)	0.1	(0.3)	0.1	(0.5)	0.1	(0.4)	0.75
Snacks	0.3	(0.6)	0.4	(0.8)	0.7	(1.1)	0.5	(1.3)	0.4	(0.7)	P < 0.001
Buns & Cakes	1.3	(1.8)	1.6	(2.1)	1.8	(1.8)	1.4	(1.6)	1.3	(1.8)	P < 0.001
Ice cream	0.5	(1.0)	0.9	(1.4)	0.9	(1.5)	0.8	(1.4)	0.7	(1.2)	P < 0.001
Cream	0.2	(0.5)	0.2	(0.6)	0.3	(0.8)	0.3	(0.8)	0.2	(0.5)	0.58
Desserts	1.4	(2.5)	2.0	(3.8)	1.5	(2.8)	1.7	(4.2)	1.4	(2.4)	0.05
Soft beverages (sweetened)	7.1	(7.4)	10.0	(6.3)	*27.8*	(7.3)	7.5	(5.8)	8.4	(7.0)	P < 0.001
Sweets	1.3	(1.5)	1.9	(2.2)	1.9	(1.8)	1.6	(1.9)	1.5	(1.8)	P < 0.001
Sausages	2.1	(2.1)	2.0	(2.1)	1.6	(1.6)	2.0	(2.3)	1.8	(2.0)	0.21
Water	4.7	(6.0)	8.1	(6.5)	9.2	(7.2)	8.4	(6.3)	30.5	(8.4)	P < 0.001
Ketchup	0.3	(0.5)	0.4	(0.8)	0.3	(0.6)	0.3	(0.6)	0.3	(0.6)	0.09
Sauces &gravys	0.8	(1.3)	0.8	(1.3)	0.7	(1.2)	0.8	(1.3)	0.7	(1.5)	0.92
Bread	3.3	(3.3)	3.0	(2.5)	2.7	(2.1)	3.7	(3.4)	2.9	(2.2)	P < 0.05
Milk (fat > 0.5%)	*38.7*	(7.7)	*22.2*	(5.1)	10.6	(7.1)	6.4	(5.3)	9.7	(7.3)	P < 0.001
Milk (fat < 0.5%)	0.4	(1.5)	0.4	(1.6)	1.6	(4.1)	*9.5*	(11.6)	1.4	(3.6)	P < 0.001
Flavored yoghurt (fat > 0.5%)	1.0	(2.3)	1.8	(4.2)	1.3	(2.8)	2.4	(4.7)	1.0	(2.5)	P < 0.05
Flavored yoghurt (fat < 0.5%)	0.4	(1.7)	0.7	(2.6)	0.5	(1.4)	1.0	(2.7)	0.5	(2.0)	0.31
Cheese	0.4	(0.7)	0.5	(0.8)	0.6	(0.8)	0.5	(1.0)	0.6	(0.8)	0.35
White bread	1.3	(1.6)	1.5	(1.7)	1.4	(1.8)	1.6	(2.2)	1.1	(1.5)	0.27
Margarines (fat > 60%)	0.3	(0.4)	0.2	(0.4)	0.2	(0.4)	0.2	(0.4)	0.2	(0.3)	0.07
Margarines (fat < 60%)	0.3	(0.4)	0.3	0.4)	0.3	(0.3)	0.4	(0.5)	0.3	(0.3)	0.44
Soft beverages (light)	0.6	(2.1)	1.2	(4.2)	0.9	(2.6)	*1.8*	(5.0)	0.7	(2.2)	P < 0.05

Mean consumption (% weight) and SD of each of five clusters (R11 I through R11 V) at the top hierarchical level, based on OMB-DHC dendrogram output. The data are based on an excerpt (11 year old subjects; 35 food groups) of *Riksmaten – barn 2003*. Figures in italic font indicate one or two chief food groups of each aggregation, used to arrive at the following alternative cluster designation: *Milk (fatty)*, *Milk (fatty)/Varied*, *Soft beverages (sweetened)*, *Milk (low fat)/Soft beverages (light)/Juice* and *Varied/Water*.

Table 5

	DNS I Mean (%)	SD (+/-)	DNS II Mean (%)	SD (+/-)	DNS III Mean (%)	SD (+/-)	DNS IV Mean (%)	SD (+/-)	P- value
Cluster size (N)	**86**		**102**		**30**		**100**		
Energy (MJ)	**7.7**	**(1.9)**	**7.4**	**1.6**	**6.6**	**1.3**	**7.3**	**1.3**	**0.08**
Meat & poultry	3.0	(1.9)	2.9	(1.8)	2.6	(1.3)	2.6	(1.8)	0.18
Cereals	1.9	(2.2)	2.4	(2.4)	1.5	(1.6)	2.1	(2.2)	0.22
Gravys and dressing	1.7	(1.0)	1.9	(1.1)	1.9	(1.1)	1.7	(1.1)	0.14
Fruit & berries	*9.0*	(6.1)	7.6	(5.1)	6.3	(4.1)	*8.9*	(4.6)	P < 0.05
Vegetables	5.2	(3.5)	5.1	(3.0)	5.5	(4.6)	5.8	(3.1)	0.46
Bread	7.6	(2.9)	8.0	(2.8)	7.2	(2.3)	7.3	(2.7)	0.28
Sausages	0.9	(1.2)	0.6	(0.9)	0.7	(0.8)	0.7	(1.0)	0.08
Juice	*4.8*	(4.5)	1.7	(2.3)	2.7	(3.3)	2.7	(3.1)	P < 0.001
Soft beverages (sweetened)	*24.5*	(8.5)	9.6	(5.7)	8.6	(6.1)	*9.5*	(5.3)	P < 0.001
Soft beverages (light)	2.0	(3.1)	2.6	(3.7)	*26.4*	(9.7)	2.7	(4.2)	P < 0.001
Milk	15.9	(6.2)	*33.4*	(6.7)	14.1	(7.8)	15.7	(6.9)	P < 0.001
Snacks	0.4	(0.5)	0.4	(0.5)	0.4	(0.5)	0.4	(0.5)	0.94
Ice cream	1.3	(1.4)	1.1	(1.3)	1.1	(1.4)	1.0	(1.0)	0.70
Candy	1.3	(1.0)	1.1	(0.7)	0.9	(0.6)	1.0	(1.2)	0.11
Water	8.7	(6.0)	9.5	(6.8)	9.2	(8.1)	*27.3*	(8.3)	P < 0.001
Potato	3.0	(2.3)	3.3	(2.7)	3.9	(3.2)	2.8	(2.1)	0.34
Rice	0.9	(1.4)	0.9	(1.1)	0.5	(0.7)	0.7	(1.1)	0.30
Fish	0.7	(0.9)	0.8	(0.8)	0.6	(0.7)	0.8	(0.7)	0.31
Pizza	1.1	(2.0)	1.2	(2.1)	1.0	(1.8)	1.1	(1.6)	0.98
Yoghurt	0.8	(2.1)	0.9	(2.7)	0.0	(0.2)	*1.3*	(3.2)	P < 0.05
Sugar	0.9	(1.1)	0.8	(0.7)	0.8	(0.9)	0.7	(0.5)	0.41
Eggs	0.5	(0.7)	0.4	(0.7)	0.3	(0.4)	0.3	(0.4)	0.97
Pasta	1.5	(1.5)	1.5	(1.5)	1.7	(1.6)	1.1	(1.0)	0.07
Cheese	0.3	(0.4)	0.4	(0.6)	0.3	(0.5)	0.3	(0.4)	0.38
Buns & cakes	2.2	(2.0)	2.0	(2.2)	1.7	(1.2)	1.6	(1.7)	0.11

Mean consumption data (% weight and SD) of Danish children (4–5 years old; 25 food groups), as categorized by OMB-DHC hierarchical clustering analysis (DNS I through DNS IV). Numbers in italic font indicate one or several outstanding food groups of each cluster, used to arriving at the following alternative cluster designation: *Soft beverages (sweetened)/Juice*, *Milk*, *Soft beverages (light)* and *Yoghurt/Fruit & berries/Water*.

Appendix A

Food group variables (35) used to analyze Swedish children

- Potato
- Root crops and vegetables
- Fruits & berries
- Juice
- Cereals (Infant formula, porridge, müsli and cereals)
- Pancakes, waffles and crêpes
- Pizza, pie and pastry
- Rice and grains
- Pasta
- Meat and poultry
- Eggs
- Fish &Seafood
- Blood and offal food
- Nuts, seeds and snacks
- Buns, cakes and biscuits
- Ice cream
- Cream
- Sweet soups, creams, desserts, jam and marmalade
- Soda, lemonade, sport drinks and water ices (with sugar), typically referred to as *Soft beverages (sweetened)*
- Soda, lemonade light, commonly appearing as *Soft beverages (light)*
- Sweets and sugar
- Sausage
- Coffee, tea and water
- Ketchup, spices and salt
- Sauces
- Bread, unspecified
- Bread, white
- Bread, rye
- Milk and yoghurt fat content > 0.5%
- Milk and yoghurt fat content < 0.5%
- Milk and yoghurt sweetened fat content > 1%
- Milk and yoghurt sweetened fat content < 1%

- Cheese fat content > 17%
- Spreads & butter> 60%
- Spreads & butter < 60%

Food group variables (25) used to analyze and compare dietary patterns across Swedish and Danish children

- Meat & poultry (food courses included)
- Cereals (infant formula, porridge, Müsli and cereals)
- Gravys, dressing , spreads & butter
- Fruit and berries
- Vegetables
- Bread
- Sausages
- Juice
- Soda, lemonade, sport drinks and water ices (with sugar), mostly summarised as *Soft beverages (sweetened)*
- Soda, lemonade light, also appearing as *Soft beverages (light)*
- Milk
- Snacks
- Ice cream
- Candy
- Water
- Coffee, tea and water
- Potato
- Rice
- Fish
- Pizza
- Yoghurt
- Sugar (desserts and jam included)
- Eggs
- Pasta
- Cheese
- Buns, cakes and biscuits

Appendix B

Table A 1A

Variable	RA II/I Mean (%)	SD (+/-)	RA II/II Mean (%)	SD (+/-)	P- value
Cluster size (N)	**541**		**214**		
Potato	6.4	(4.2)	6.6	(4.8)	0.93
Vegetables	*3.1*	(2.7)	2.7	(2.8)	P < 0.05
Fruit & berries	*5.2*	(4.5)	4.0	(3.7)	P < 0.001
Juice	*3.0*	(4.3)	1.8	(3.5)	P < 0.001
Cereals	2.6	(3.4)	2.6	(3.5)	0.70
Pancake	1.3	(2.1)	1.1	(1.8)	0.19
Pizza	1.2	(2.2)	1.2	(2.1)	0.52
Rice	1.7	(2.2)	2.0	(2.2)	0.09
Pasta	3.2	(2.8)	3.5	(3.4)	0.44
Rye bread	0.2	(0.6)	0.2	(0.7)	0.69
Meat & poultry	5.9	(3.3)	6.1	(3.4)	0.63
Eggs	0.2	(0.5)	0.3	(0.7)	0.35
Fish	1.0	(1.3)	1.3	(1.7)	0.12
Blood food	0.2	(0.6)	0.1	(0.4)	0.07
Snacks	0.3	(0.6)	0.3	(0.6)	0.39
Buns & Cakes	*1.6*	(1.8)	1.0	(1.5)	P < 0.001
Ice cream	*1.1*	(1.5)	0.7	(1.1)	P < 0.05
Cream	*0.3*	(0.6)	0.1	(0.4)	P < 0.001
Desserts	2.2	(3.4)	1.4	(2.2)	P < 0.001
Soft beverages (sweetened)	*10.9*	(6.4)	4.6	(4.9)	P < 0.001
Sweets	*1.5*	(1.5)	1.3	(1.4)	P < 0.05
Sausages	1.8	(1.9)	2.0	(2.0)	0.39
Water	*5.9*	(5.8)	4.4	(5.8)	P < 0.001
Ketchup	0.4	(0.6)	0.4	(0.5)	0.18
Sauces & gravys	0.9	(1.5)	0.8	(1.3)	0.12
Bread	2.5	(1.9)	3.1	(3.2)	0.17
Milk (fat > 0.5%)	28.9	(3.9)	*41.5*	(7.5)	P < 0.001
Milk (fat < 0.5%)	0.4	(1.5)	0.3	(1.3)	0.15
Flavoured yoghurt (fat > 0.5%)	*2.1*	(4.0)	1.3	(3.6)	P < 0.001
Flavoured yoghurt (fat < 0.5%)	0.5	(1.7)	0.4	(1.4)	0.5
Cheese	0.5	(0.7)	0.5	(0.8)	0.5
White bread	1.3	(1.5)	1.1	(1.4)	1.3
Margarines (fat > 60%)	0.3	(0.4)	0.3	(0.4)	P < 0.05
Margarines (fat < 60%)	0.3	(0.4)	0.3	(0.4)	P < 0.05
Soft beverages (light)	0.9	(3.0)	0.7	(2.2)	0.9

Consumption (mean % and SD) of two subclusters at level two (RA II/I and RA II/II), both being segregations of a major cluster (RA II; see Figures 1 and 4), as identified by OMB-DHC processing of the entire *Riksmaten – barn 2003* and distributed over 35 food groups. Numbers in italic font indicate outstanding dissimilarity between the two aggregations.

Table A 1B

Variable	RA II/I/I Mean (%)	SD (+/-)	RA II/I/II Mean (%)	SD (+/-)	RA II/I/III Mean (%)	SD (+/-)	RA II/I/IV Mean (%)	SD (+/-)	RA II/I/V Mean (%)	SD (+/-)	P- value
Cluster size (N)	**72**		**131**		**142**		**98**		**98**		
Potato	*12.4*	(4.7)	5.2	(4.7)	5.3	(4.7)	5.0	(4.7)	5.6	(4.7)	p < 0.001
Vegetables	2.5	(2.6)	2.2	(2.6)	3.3	(2.6)	3.0	(2.6)	*4.3*	(2.6)	p < 0.001
Fruit & berries	2.9	(2.5)	3.0	(2.5)	5.4	(2.5)	3.3	(2.5)	*11.2*	(2.5)	p < 0.001
Juice	1.6	(2.5)	1.2	(2.5)	2.2	(2.5)	*8.6*	(2.5)	1.4	(2.5)	p < 0.001
Cereals	2.5	(3.6)	2.4	(3.6)	2.4	(3.6)	2.0	(3.6)	*3.5*	(3.6)	0.10
Pancake	1.8	(2.8)	1.3	(2.8)	1.3	(2.8)	1.2	(2.8)	1.0	(2.8)	0.47
Pizza	0.9	(1.9)	1.6	(1.9)	1.4	(1.9)	1.1	(1.9)	0.8	(1.9)	p < 0.05
Rice	2.5	(3.0)	1.5	(3.0)	1.6	(3.0)	1.5	(3.0)	1.8	(3.0)	0.23
Pasta	3.1	(2.9)	2.9	(2.9)	3.3	(2.9)	*3.5*	(2.9)	3.1	(2.9)	0.69
Rye bread	0.3	(0.8)	0.1	(0.8)	0.1	(0.8)	0.2	(0.8)	0.3	(0.8)	0.17
Meat & poultry	*7.3*	(4.0)	6.1	(4.0)	5.2	(4.0)	5.6	(4.0)	5.5	(4.0)	p < 0.001
Eggs	0.3	(0.7)	0.2	(0.7)	0.2	(0.7)	0.2	(0.7)	0.4	(0.7)	0.07
Fish	1.4	(1.7)	0.9	(1.7)	0.9	(1.7)	0.9	(1.7)	1.2	(1.7)	0.16
Blood food	0.1	(0.4)	0.2	(0.4)	0.2	(0.4)	0.1	(0.4)	0.3	(0.4)	0.79
Snacks	0.2	(0.5)	0.4	(0.5)	0.2	(0.5)	0.4	(0.5)	0.2	(0.5)	0.11
Buns & Cakes	1.5	(2.0)	*2.0*	(2.0)	1.4	(2.0)	1.6	(2.0)	1.3	(2.0)	p < 0.05
Ice cream	0.9	(1.4)	*1.4*	(1.4)	1.0	(1.4)	0.7	(1.4)	1.1	(1.4)	0.11
Cream	0.2	(0.7)	0.3	(0.7)	0.2	(0.7)	0.2	(0.7)	0.4	(0.7)	0.06
Desserts	3.3	(5.4)	2.0	(5.4)	2.1	(5.4)	1.7	(5.4)	2.2	(5.4)	0.85
Soft beverages (sweetened)	7.0	(4.8)	*18.9*	(4.8)	7.8	(4.8)	8.9	(4.8)	8.5	(4.8)	p < 0.001
Sweets	1.1	(1.6)	*1.7*	(1.6)	1.4	(1.6)	1.3	(1.6)	1.4	(1.6)	p < 0.05
Sausages	*2.6*	(2.6)	1.7	(2.6)	1.8	(2.6)	1.4	(2.6)	1.8	(2.6)	p < 0.05
Water	3.0	(3.4)	3.2	(3.4)	*13.4*	(3.4)	3.0	(3.4)	3.1	(3.4)	p < 0.001
Ketchup	0.4	(0.6)	0.4	(0.6)	0.4	(0.6)	0.4	(0.6)	0.3	(0.6)	0.66
Sauces &gravys	1.2	(1.8)	0.8	(1.8)	0.8	(1.8)	0.7	(1.8)	1.0	(1.8)	0.59
Bread	2.8	(2.5)	2.4	(2.5)	2.3	(2.5)	2.5	(2.5)	2.5	(2.5)	0.58
Milk (fat > 0.5%)	26.7	(3.5)	30.8	(3.5)	29.3	(3.5)	24.9	(3.5)	28.2	(3.5)	p < 0.001
Milk (fat < 0.5%)	0.7	(2.4)	0.3	(2.4)	0.5	(2.4)	0.2	(2.4)	0.6	(2.4)	p < 0.05
Flavoured yoghurt (fat > 0.5%)	3.4	(6.0)	1.6	(6.0)	1.4	(6.0)	1.3	(6.0)	3.2	(6.0)	p < 0.001
Flavoured yoghurt (fat < 0.5%)	0.7	(2.6)	0.1	(2.6)	0.4	(2.6)	0.5	(2.6)	0.7	(2.6)	0.05
Cheese	0.4	(0.6)	0.5	(0.6)	0.4	(0.6)	0.5	(0.6)	0.4	(0.6)	0.13
White bread	1.4	(1.7)	1.3	(1.7)	1.3	(1.7)	1.2	(1.7)	1.2	(1.7)	0.68
Margarines (fat > 60%)	0.3	(0.3)	0.3	(0.3)	0.3	(0.3)	0.3	(0.3)	0.4	(0.3)	0.60
Margarines (fat < 60%)	0.4	(0.4)	0.3	(0.4)	0.3	(0.4)	0.3	(0.4)	0.3	(0.4)	0.51
Soft beverages (light)	2.4	(5.8)	0.7	(5.8)	0.6	(5.8)	0.4	(5.8)	0.9	(5.8)	0.22

Consumption (mean % and SD) of five aggregations at level three (RA II/I/I through
RA II/I/V), all being braches of a single subcluster at level 2 (RA II/I), as identified by OMB-DHC processing. See Figures 1 and 4 for guidance on cluster details. Numbers in italic font indicate outstanding features of the respective clusters. The data is derived from the comprehensive *Riksmaten – barn 2003* data and based on 35 food groups.

Table A 2

	Cluster 1	SD (+/-)	Cluster 2	SD (+/-)	P- value
Cluster size (N)	**386**		**204**		
Energy (MJ)	**6.3**	**(1.3)**	**6.2**	**(1.2)**	**0.44**
Meat & poultry	4.8	(2.7)	4.2	(2.5)	P < 0.05
Cereals	3.4	(3.9)	*27.0*	(8.5)	P < 0.001
Gravys and dressing	2.0	(1.4)	1.8	(1.3)	0.06
Fruit & berries	8.1	(5.0)	6.9	(4.9)	P < 0.05
Vegetables	3.5	(3.2)	3.0	(2.6)	0.06
Bread	*3.3*	(1.7)	2.5	(1.5)	P < 0.001
Sausages	*1.6*	(1.5)	1.4	(1.4)	P < 0.05
Juice	4.2	(5.8)	3.0	(4.0)	0.16
Soft beverages (sweetened)	11.6	(8.9)	9.8	(7.6)	0.05
Soft beverages (light)	1.0	(3.0)	1.0	(2.6)	0.37
Milk	*19.2*	(10.0)	11.6	(7.5)	P < 0.001
Snacks	0.2	(0.4)	0.2	(0.4)	0.70
Ice cream	*4.2*	(4.3)	3.1	(3.0)	P < 0.05
Candy	1.3	(1.3)	1.2	(1.3)	0.57
Water	*9.7*	(8.3)	6.2	(6.2)	P < 0.001
Potato	5.2	(3.1)	4.9	(3.2)	0.15
Rice	*1.8*	(2.0)	1.4	(1.7)	P < 0.05
Fish	1.1	(1.4)	1.2	(1.5)	0.92
Pizza	0.6	(1.2)	0.5	(1.0)	0.76
Yoghurt	*6.7*	(6.2)	3.8	(4.4)	P < 0.001
Sugar	0.8	(0.8)	0.7	(0.8)	0.08
Eggs	0.4	(0.8)	0.3	(0.6)	0.52
Pasta	2.8	(2.3)	2.6	(2.4)	0.28
Ost	*0.6*	(0.6)	0.4	(0.5)	P < 0.05
Buns & cakes	*1.9*	(1.7)	1.5	(1.3)	P < 0.05

Consumption (mean % and SD) of two major clusters (C1 and C2), as produced by OMB-DHC processing of an excerpt of *Riksmaten – barn 2003*, encompassing 4 year old subjects only and 25 food groups. Numbers in italic font indicate outstanding dissimilarity between the two aggregations.

Table A 3.

Variable	Cluster 1 Mean (%)	SD (+/-)	Cluster 2 Mean (%)	SD (+/-)	P- value
Cluster size (N)	**502**		**387**		
Energy (MJ)	**7.60**	**1.70**	**7.60**	**1.70**	**0.87**
Potato	5.9	(4.5)	6.2	(4.0)	0.06
Vegetables	*3.8*	(3.2)	3.4	(2.9)	P < 0.05
Fruits	6.0	(4.8)	5.3	(4.4)	0.06
Juice	4.8	(6.2)	2.9	(4.4)	0.05
Cereals	2.0	(3.6)	*2.2*	(3.0)	P < 0.05
Pancake	1.3	(2.1)	1.3	(2.1)	P < 0.05
Pizza	*1.4*	(2.9)	1.1	(2.0)	P < 0.05
Rice	*1.9*	(2.6)	1.8	(2.2)	P < 0.05
Pasta	*3.6*	(3.2)	3.5	(3.1)	P < 0.05
Rye bread	0.2	(1.0)	0.2	(0.6)	P < 0.001
Meat & poultry	6.4	(3.3)	6.0	(3.3)	0.06
Eggs	*0.4*	(0.8)	0.3	(0.6)	P < 0.001
Fish	1.1	(1.4)	1.1	(1.4)	P < 0.05
Blood food	0.2	(0.6)	0.2	(0.5)	P < 0.001
Snacks	*0.4*	(0.7)	0.3	(0.6)	P < 0.001
Buns & Cakes	*1.8*	(1.8)	1.5	(1.8)	P < 0.05
Ice cream	*1.2*	(1.6)	1.1	(1.3)	P < 0.05
Cream	0.2	(0.6)	*0.3*	(0.6)	P < 0.001
Desserts	*2.1*	(3.3)	1.9	(2.8)	P < 0.05
Soft beverages (sweetened)	13.7	(9.1)	10.6	(7.5)	0.14
Sweets	*1.5*	(1.5)	1.3	(1.4)	P < 0.05
Sausages	1.7	(1.8)	*1.9*	(1.8)	P < 0.05
Water	12.5	(10.2)	4.5	(5.5)	0.13
Ketchup	0.4	(0.5)	0.4	(0.6)	P < 0.001
Sauces & gravys	0.8	(1.3)	*1.0*	(1.5)	P < 0.05
Bread	*2.7*	(2.1)	2.6	(2.2)	P < 0.05
Milk (fat > 0.5%)	11.8	(6.7)	30.8	(8.1)	0.12
Milk (fat < 0.5%)	*3.1*	(6.4)	0.7	(2.4)	P < 0.05
Flavoured yoghurt (fat > 0.5%)	2.1	(3.7)	2.1	(4.3)	P < 0.05
Flavoured yoghurt (fat < 0.5%)	*0.7*	(2.0)	0.4	(1.5)	P < 0.05
Cheese	0.4	(0.7)	*0.5*	(0.8)	P < 0.001
White bread	*1.3*	(1.5)	1.2	(1.4)	P < 0.05
Margarines (fat > 60%)	0.2	(0.3)	0.3	(0.4)	P < 0.001
Margarines (fat < 60%)	0.3	(0.4)	0.3	(0.4)	P < 0.001
Soft beverages (light)	*1.7*	(4.6)	0.7	(2.1)	P < 0.05

Consumption patterns of two major clusters (top hierarchical level), using an excerpt (8 year old subjects) of *Riksmaten – barn 2003* (35 food groups) as data input to the OMB-DHC algorithm.

Appendix C: Details about MDA techniques applied to the study

Multivariate versus classical data analysis to food consumption data

A data sample, that may represent a physical object, is described by a set of recorded properties x1, x2, …xN. Thus, a vector that is composed of values of these properties, X = [x1, x2, ...,xN], termed pattern, describes an object. Properties of elements in the vector are denoted features. For example, a fruit may be described by the two features weight and colour. MDA refers to techniques used to the simultaneous observation and analysis of more than one statistical variable. At the highest level of segregation, two subtopics – exploratory methods and predictive methods – can be identified, essentially representing the following two overarching areas:

- Pattern Discovery/Unsupervised Learning/Descriptive modelling
- Predictive Modelling/Supervised Learning

Pattern discovery

Exploratory methods deal with finding structures, or groupings, within data. Typical examples are clustering and self-organising maps. The aim of predictive learning systems, on the other hand, is to predict the value of a defined feature attached to given object.

Briefly, Cluster Analysis resides in the domain of non-supervised statistical techniques, which allow division of arrays of objects into distinct categories based on the degree of shared characteristics.[82–84] Thus, entities within an accordingly identified cluster are more similar, relative to those appearing in other aggregations. Such clusters may e.g. hold entities attached to chemical or biological properties or data linked to various physiological variables, including bio-monitored parameters correlating to health or disease. Typical of many algorithms in this domain is the segregation principle, allowing categorisation of data on several hierarchical levels, which can readily be visualised in "family-tree" (dendrogram) displays. Generally, those outputs are readily interpretable since all subjects – as defined by most clustering algorithms – belong to a single population only. Although not further elaborated on here, fuzzy clustering techniques – allowing assignment of entities to two or more

groups – are also available. Algorithms designed for hierarchical clustering (HC) generally operate according to either of two distinct principles to split data: agglomerative (bottom-up) or divisive (top-down). The former design, being by far most commonly employed for this purpose, involves pair-wise merging of subclusters at each branch (data fusion) point. A major constraint of this principle, however, lies in that data fusions made at early stages of the procedure may not reveal the most adequate aggregations at higher hierarchical levels. Algorithms for divisive (top-down) hierarchical clustering (DHC), on the other hand, which iteratively split all the N samples into a hierarchy of subclusters, are more computation-intensive but typically create more reliable groups at higher branching levels.[83–85] In analogy with standard agglomerative clustering, most DHC methods still create binary (or alternatively a fixed predefined value of) divisions at each branching point. Flexible DHC algorithms allowing multiple branches at each branching level have, however, greater potential to reveal more meaningful groupings that are also easier to interpret. As outlined below, an in house-implemented multi-branching K-means-based clustering algorithm, designated OMB-DHC, was applied to the data sets addressed in this study.

PCA (factor analysis) is another non-supervised technique for the identification of embedded data structures. Interpretation of outputs from PCA is, however, less straightforward compared with those of cluster analysis. For example, an inherent complication of the former general technique, as applied to reported studies on dietary patterns, pertains to the fact that delineated relationships – e.g. principal components derived from PCA – do not necessarily refer to identifiable groups of subjects within a population, but rather to food patterns. Hence, a conceptual meaning of factor loadings cannot be guaranteed.[35] Nonetheless, the aforementioned two methods are commonly used in combination to acquire a comprehensive picture of underlying structures within the data sets. Notably, PCA helps determining the minimum number of factors that will account for the maximum variance in a data set and is thus a common technique for searching patterns in data of high dimensionality. PCA has found extensive application in a broad range of areas, including nutritional epidemiology.

Predicitive modelling

In supervised learning, a predictive model is designed with ability to predict the expected values of q response variables collected in a vector $\mathbf{y}$ from a set of p input (predictor) variables collected in a vector $\mathbf{x}$. One of the simplest forms of such models can be written $\mathbf{y}=W\mathbf{x}$, where W is a coefficient matrix. Thus, the aim of predictive learning systems is to pre-

dict the value of some property y of a given object represented by x. The property of a data sample that is subject to predicting, based on the measured features, is referred to as target. Typical targets to predict, e.g. for a chemical molecule, are receptor affinity or phenotypic effect. The domain of the target can either be discrete or continuous. In the former case, possible values assigned to the target are named classes or labels and the prediction is called classification. Some examples of common classifiers and regression models used in supervised learning are Neural Network, Random Forest, Support Vector Machine, K-Nearest Neighbour, Naive Bayes Classifier and Decision Tree.

Hierachical clustering: divisive omb-dhc versus agglomerative 2-way
The in house-implemented (in Matlab) clustering algorithm – designated OMB-DHC and applied to the bi-national dietary survey data addressed in this report – is based on the classical clustering K-means principle and using the likewise classical average silhouette width (ASW) to select the most suitable number of main and subordinate clusters at each branching point. It produces a flexible multi-branching output and is equipped with a tailor-made dendrogram display, to visualize accordingly identified hierarchical structure of input data.[66,67,86] OMB-DHC features several enhancements of a previously reported scheme, i.e. it incorporates a tuned amalgamation of the K-means algorithm and ASW thereby enabling their sequential operation in an automated mode. Moreover, edges (branches) appearing in dendrogram representations are directly proportional to distances from their respective parent cluster centra.[66,67] Analysis by means of OMB-DHC encompassed the entire Swedish dataset, the three Swedish age groups separately as well as the Danish data set. The algorithm was applied to these data sets in two different ways, i.e. based on variations among food groups with respect to children as well as variations among the children according to food groups, i.e. the transposed data matrix. Moreover, dietary prototypes defined by means of OMB-DHC also underwent subsequent MDA inspection, using classical multidimensional scaling (see also below) or a two-dimensional agglomerative hierarchical cluster analysis. The latter technique involved feeding an agglomerative hierarchical bi-clustering algorithm, readily available in the Matlab programming environment, with mean consumption data of dietary prototypes. This process involved iterative joining by similarity, initially of dietary prototypes (columns) and subsequently food categories (rows). Accordingly derived results were ultimately displayed as charts, illustrating dendrograms in two dimensions and a heatmap over input data. This composite data interrogation technique is here referred to as *Hierarchical Prototype Bi-Cluster Analysis* (HPBCA).

Classical multi-dimensional Scaling

Classical multidimensional scaling (CMDS) is a MDA technique that aims to find a low dimension representation of data that preserve inter-point distances in the original space, in order to visualize their reciprocal similarities.[85] Thus, similarity or dissimilarity among pairs of objects, expressed as distance between points, is projected in a low dimensional space. In this study CMDS-analysis, based on Euclidian distances of normalised data, was applied to dietary prototypes, as identified by OMB-DHA (see also above). We have coined the term *Dietary Prototype CMDS Analysis* (DPCA) for this particular composite technique.

PCA (factor analysis)

PCA (factor) analysis was conducted in the Matlab environment, using the princomp() function. The relative (percentage of total) variation explained by each principal component was computed and visualized as pareto-plots, to determine the number of principal components needed to show enough variation, without over-fitting the data. Moreover, some PCA outputs are displayed subsequent to reprocessing and transfer to a new coordinate system, which facilitates identification of key patterns.

Spearman's correlation coefficient

The rank based Spearman's correlation between the various food groups was computed for data in Danish (4–5 year old children) and Swedish data set (4-year old children).

Comparisons across age-specific dietary patterns and two national data sets

Two separate but related techniques – DPCA and HPBCA (outlined above) – were applied to the unabridged Swedish data (*Riksmaten – barn 2003*) and likewise to the Danish-Swedish excerpt of 4–5 year old consumers, to derive quantitative estimates of relationships across dietary patterns and to portray the associations. Thus, dietary prototypes, identified by OMB-DHC were input data and each age group was accordingly represented by either four or five such prototypes. They appear in tables as strings of mean intakes of the various food groups, 35 across the Swedish data and 25 for the Danish set; data entering the bi-national comparison were, however, harmonised to 25 foods. First, these compacted characteristics were analysed by DPCA. Such output data allow depiction of interrelationships across age- or nation-specific dietary prototypes in a reduced dimensional display as well as quantitative determination of overall and prototype-specific dietary distances across and within categories. Second, the aforementioned set of dietary proto-

types were processed by means of two way agglomerative hierarchical clustering, likewise outlined above, thereby exercising HPBCA. Such output charts illustrate relationships across dietary prototypes and foods alike.

Prediction of dietary patterns

The Random Forest (RF) machine learning technique is an ensemble classifier composed of many decision trees and outputs the class that is the mode of the classes output by individual trees. It is widely acknowledged as one of the most accurate learning algorithms currently available and produces very truthful classification for many data sets. Moreover, it provides estimates of important variables in the classification. Predictive modelling designs developed in this study were all performed in R. RF was applied to discovering pertinent patterns among low- and high consumers of each among the various food groups, apart from the single outstanding food group, i.e. to design models which, irrespectively of the targeted food, can reliably classify the respective (low/high) populations. In other words, each model was built – by means of the RF/NSC techniques – to predicting class Y based on consumption pattern of the remaining food groups X.

Danish (4–5 years old) and Swedish (4 years old) subjects were partitioned in low and high consumers, for each food group, and a RF model was build based on the remaining food groups. To be able to assess the validity of the received model, 100 RF models were built on permuted data, i.e. involving randomly shuffled data order of Y. Each RF model is associated with an OOB (out-of-bag) error rate and a VIMP (variable importance, here mean decrease accuracy) value for each variable. The permutations give us a null distribution both of OOB error rates and VIMP values. A model is considered significant if its OOB error rate is below that of the null model on the 0.95 significance level. For a significant model, the VIMP value for a specific variable can be compared to the null distribution of the VIMP values for that same variable to detect variables that are important to the model.

A more comprehensive modelling dividing the data into training and test sets was performed, briefly according to the following main steps:

- Partitioning the data (for each analysis, i.e. food group) in training (80%) and test (20%) sets
- Applying variable selection to the training set
- Model construction based on training data and selected variables
- Prediction of values for test data and estimation of model quality

The above-listed steps were iterated 100 times, ultimately generating the selection frequency of each variable (thus ranging from 0 to 100). To assess whether a particular variable has been selected more frequently than expected by chance, the entire procedure (as outlined above) was iterated 200 times, but with permuted class allocations, i.e. involving disruption of the relationship between variables and class assignment. The performance measure employed was AUC, the area under the empirical (estimated) receiver operator characteristics curve. A mean AUC (mean of 100 holdouts) and a vector composed of frequencies of the selected variables were computed for each such data output (original and permuted alike). In analogy with the RF analysis described above, the permutations give us null distributions of mean AUC and frequency values that can be used to determine significance of models and also the importance of individual variables to the models. The Nearest Shrunken Centroid (NSC) technique was applied to both variable selection and model building. To avoid bias, with regard to the NSC design (computes distances across subjects), intake values for each food group were divided by the mean consumption of this food group. It is of course also possible to use other classification methods, e.g. RF, for model building and variable selection, but RF requires more computation time and memory than NSC.

TemaNord 2013:548

Nordic Council of Ministers

Ved Stranden 18
DK-1061 Copenhagen K
www.norden.org

Discovery and characterisation of dietary patterns in two Nordic countries

This Nordic study encompasses multivariate data analysis (MDA) of preschool Danish as well as pre- and elementary school Swedish consumers. Contrary to other counterparts the study incorporates two separate MDA varieties - Pattern discovery (PD) and predictive modelling (PM). PD, i.e. hierarchical cluster analysis (HCA) and factor analysis (using PCA), helped identifying distinct consumer aggregations and relationships across food groups, respectively, whereas PM enabled the disclosure of deeply entrenched associations. 17 clusters - here defined as dietary prototypes - were identified by means of HCA in the entire bi-national data set. These prototypes underwent further processing, which disclosed several intriguing consumption data relationships: Striking disparity between consumption patterns of Danish and Swedish preschool children was unveiled and further dissected by PM. Two prudent and mutually similar dietary prototypes appeared among each of two Swedish elementary school children data subsets. Dietary prototypes rich in sweetened soft beverages appeared among Danish and Swedish children alike. The results suggest prototype-specific risk assessment and study design.

TemaNord 2013:548
ISBN 978-92-893-2581-3