Food Quality and Preference
March 2014, Volume 32, Part A, Pages 56–64
http://dx.doi.org/10.1016/j.foodqual.2013.06.002

Multiblock modeling for complex preference study. Application to European preferences for smoked salmon

Stéphanie Bougeard[a,] *, Mireille Cardinal[b]

[a] Anses (French Agency for Food, Environmental and Occupational Health Safety), Department of Epidemiology, F-22440 Ploufragan, France
[b] Ifremer (French Research Institute for Exploration of the Sea), Laboratoire Science et Technologie de la Biomasse Marine, F-44311 Nantes 03, France

*: Corresponding author : Stéphanie Bougeard, tel.: +33 296 010 150 ;
email address : stephanie.bougeard@anses.fr

Abstract:

The aim of the paper is to propose an alternative method to external preference mapping for the case of complex data where explanatory variables are organized in meaningful blocks. We propose an innovative method in the multiblock modeling framework, called multiblock Redundancy Analysis. The interest and relevance of this method is illustrated on the basis of a European consumer preference study for cold-smoked salmon. The study aims at explaining six homogeneous clusters of preference with explanatory parameters organized in five thematic blocks related to physico-chemical measurements, microbiological characterization, appearance attributes, odor/flavor characterization and texture descriptors. Overall indexes and graphical displays associated with different interpretation levels are proposed to sort the key drivers of preference by order of priority at the variables and at the block level. On the basis of these data, multiblock Redundancy Analysis is also compared to standard preference mapping in terms of model quality; the best model is here associated with the multiblock method.

Highlights

► We propose to apply an orginal multiblock method to external preference mapping. ► We propose original interpretation tools at the block level. ► We compare the performance of multiblock method in regards with standard external preference mapping. ► We apply this multiblock method to European preferences for smoked salmon.

Keywords: Multiblock modeling ; Multiblock redundancy analysis ; External preference mapping ; Smoked salmon

1. Context

Product development is often based on external preference mapping (prefmap) where sensory profiles are used to model consumer likings, all the variables being measured on the same products. This approach gives a reliable basis for creating products which correspond to consumer expectations. In this framework, we mainly focus on identifying the key drivers which impact the consumer preferences. External preference mapping assumes that consumers have a common perceptual space and that it can be modeled with sensory data (Jaeger, Wakeling, & Macfie, 2000). It is worth noting that other parameters are usually measured on products such as physical and chemical measurements, price and packaging descriptions. In order to improve the preference modeling and then get an overall vision of the preferred products or of the products to be developed, it is of paramount importance to explain consumer preferences not only with sensory attributes but also with these additional parameters. This could improve one of the main criticisms of prefmap, namely the poor model quality due to a product attribute space inadequate to the preference one. As a way to enhance modeling quality, we focus on external preference mapping applied not only to sensory but also to external attributes. This problematic is related to the explanation of a composite dataset, i.e., the consumer preferences (Y) with explanatory variables organized in several meaningful blocks, e.g., sensory attributes (X1), physico-chemical parameters (X2) and packaging description (X3). All these variables are measured on the same observations, i.e., the products under study, as illustrated in Fig. 1.

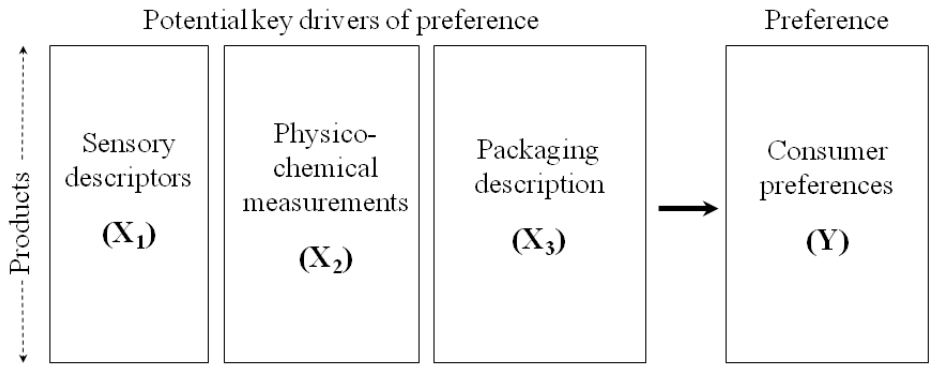

Fig.1. Example of multiblock explanatory data which aims at explaining consumer preferences.

For the time being, three data processing solutions remain for the user to take account of the explanatory block structure. (i) The widely used solution consists in linking at first sensory attributes to preferences with a two-block method such as Partial Least Squares (PLS). In a further stage, other measurements are linked to preferences to get a more accurate characterization (Semenou, Courcoux, Cardinal, Nicod, & Ouisse, 2007). But this leads to a sequential resolution where preferences are actually only explained with sensory attributes. (ii) The second kinds of methods pertain to the field of Structural Equation Modeling–PLS Path Modeling (PLS-PM) being the well-known in sensometrics–initially developed for more complex data (Pagès & Tenenhaus, 2001). This method brings information on links between blocks achieved with the inner model but these coefficients can only be given separately for each dimension under study. Nevertheless, models are usually multidimensional, especially in biological fields. In addition, PLS-PM is based on a complicated iterative algorithm with any formal convergence proofs (Henseler, 2010). (iii) Finally, some multiblock methods, such as Parallel Orthogonalized PLS (Måge, Menichelli, & Naes, 2012) are proposed. The PO-PLS method is especially developed for external preference mapping and focuses on common and unique information in each block. But the iterative algorithm and its complexity restrict the use of this method for more than two blocks (Måge et al., 2012). In this paper, we will stand in this interesting latter framework of multiblock methods while proposing an alternative approach with a direct eigensolution.

Among methods pertaining to the multiblock (K+1) setting, we single out those which are based on an optimization criterion that reflects the objectives to be addressed and leads to a direct eigensolution. Three methods which meet these constraints are available: Generalized Canonical Analysis with a Reference Table, GCA-RT (Kissita, 2003), multiblock Redundancy Analysis, mbRA (Bougeard, Qannari, & Rose, 2011) and multiblock Partial Least Squares, mbPLS (Wold, 1984). The method GCA-RT, is interesting from a theoretical point of view but may lead in practice to unstable model in case of quasi-collinear variables. The method mbPLS is a helpful and popular method in regards with its stability in case of multicollinearity but leads to solutions often not much linked to the dependent dataset. In addition, for our particular case of a single dataset to be explained, mbPLS leads to a simple PLS of **Y** and the merged dataset **X** (Westerhuis, Kourti, & MacGregor, 1998). We focus afterward on multiblock Redundancy Analysis which appears to take account of the multiblock structure of data and to lead to a model with a good fitting ability in spite of its lack of stability in case of high quasi-collinear variables (Bougeard & Qannari, 2011). Our purpose is to apply this original multiblock method to external preference mapping. This can be viewed as an extension of external preference mapping to the multiblock framework. Several interpretation tools pertaining to the field of factorial analysis and modeling are provided to further investigate the relationships among variables and datasets (Bougeard et al., 2011). All these methodological contributions are presented in Section 2. The interest of multiblock modeling analysis is illustrated on the basis of a European preference study of smoked salmons in Section 3. A discussion both on method and application is proposed in Section 4.

2. Material and method

2.1. Multiblock data and aims: European preferences for smoked salmon

The interest of multiblock modeling is illustrated on data from a European project (Adriant, Ifremer, IMR, & Matra, 2004). A preference study is conducted on thirty smoked salmon, representatives of the market range (Cardinal et al., 2004) tested by 1063 consumers. As the individual preferences of the 1063 consumers are not uniform, homogeneous clusters of hedonic assessments are provided through a latent class vector model (Semenou et al., 2007). Six clusters, respectively containing 121, 74, 349, 78, 404 and 37 consumers, are highlighted. For simplicity sake and as often in external preference mapping, we take account of homogeneous clusters instead of individual likings. It follows that the quantitative dependent dataset **Y** involves thirty observations (salmons) described by six variables (clusters of preferences), each salmon being described by the associated cluster preference average. Physical, chemical, microbiological and sensorial measurements are carried out on the same salmons. We choose to organize these forty-four explanatory parameters in five thematic blocks related to the physico-chemical measurements ($\mathbf{X_1}$ dataset, 13 variables), the microbiological characterization ($\mathbf{X_2}$, 6 variables), the appearance attributes ($\mathbf{X_3}$, 6 variables), the odor and flavor characterization ($\mathbf{X_4}$, 14 variables) and the texture attributes ($\mathbf{X_5}$, 5 variables) (see description in appendix). As all variables are expressed in non-comparable range of measurements, they are column centered and scaled to unit variance. However, it is worth noting that as the variables have been standardized, the total variance in each block is equal to the number of variables in this block. This motivates the block weighting to put the blocks to the same footing (Westerhuis & Coenegracht, 1997). Each of the (K=5) explanatory block is accommodated with an isotropic scaling factor to set them to the same total variance, chosen equal to 1/K. Therefore the merged explanatory dataset **X** (resp. **Y**) has a total variance equal to one.

These data have already been processed from many different ways (Cardinal et al., 2004; Courcoux, Qannari, & Schlich, 2006; Semenou et al., 2007). The latter authors propose internal and external preference mapping, the physico-chemical variables being related to preferences in a second step. We propose to consider the whole data, namely preferences, sensory analysis, physico-chemical and microbiological measurements in a single analysis. The first aim is descriptive and consists in explaining the consumer preferences in relation to all the explanatory variables organized in thematic blocks. This leads to two main questions:

Q1. Are there any relationships between the smoked salmon clusters of preferences $\mathbf{Y}=(y_1...y_6)$ and the external salmon attributes $\mathbf{X}=(x_1...x_{44})$?

Q2. Do the thirty smoked salmons have the same features in terms of their external description (**X**) in relation to preferences (**Y**)?

The second and pivotal aim is devoted to assess the key drivers of the salmon preferences from the forty-four external attributes. Three questions pertaining to the modeling framework can be asked:

Q3. Are there significant links between all the variables describing the external attributes $\mathbf{X}=(x_1...x_{44})$ and each clusters of preferences $\mathbf{Y}=(y_1...y_6)$?

Q4. Is it possible to sort by order of priority all the external variables describing the smoked salmons $\mathbf{X}=(x_1...x_{44})$ in relation to the overall preferences (**Y**)?

Q5. Is it possible to sort by order of priority the various external blocks ($\mathbf{X_1}...\mathbf{X_5}$) in relation to the overall preferences (**Y**)?

Multiblock Redundancy Analysis aims at answering these five questions. The first four questions can be answered with standard methods although the multiblock data structure is not taken into account for the calculations. But the latter one, as it gives information at the block level, is specific to multiblock analysis.

2.2. External preference mapping with multiblock modeling

Consider the multiblock setting where we have (K+1) datasets, *i.e.*, a dataset **Y** to be predicted from K other ones ($\mathbf{X_1}$...$\mathbf{X_K}$). The **Y** dataset (preferences) contains Q variables and each table $\mathbf{X_k}$ (sensory, physico-chemical, etc.) contains P_k variables. The merged dataset **X** related to all the explanatory variables contains $P=\sum_k P_k$ variables. All these quantitative variables are measured on the same N observations–the products under study–and are supposed to be column centered. Multiblock Redundancy Analysis is a latent variable based technique where the key idea is that each of the (K+1) datasets, *i.e.*, **Y** and ($\mathbf{X_1}$...$\mathbf{X_K}$), is summed up with a latent variable, respectively called u or (t_1...t_K), linear combination of the associated variables. Using latent variables instead of datasets allows handling more explanatory variables than in standard two-block analysis and restricting the problem of quasi-colinearity within explanatory datasets. The method derives a global latent variable t related to all the explanatory variables which is as close as possible to a latent variable u, linear combination of the variables to be explained. In addition, the global latent variable t sums up the partial ones (t_1...t_K) respectively associated with the partial datasets ($\mathbf{X_1}$...$\mathbf{X_K}$) as illustrated in Fig. 2.

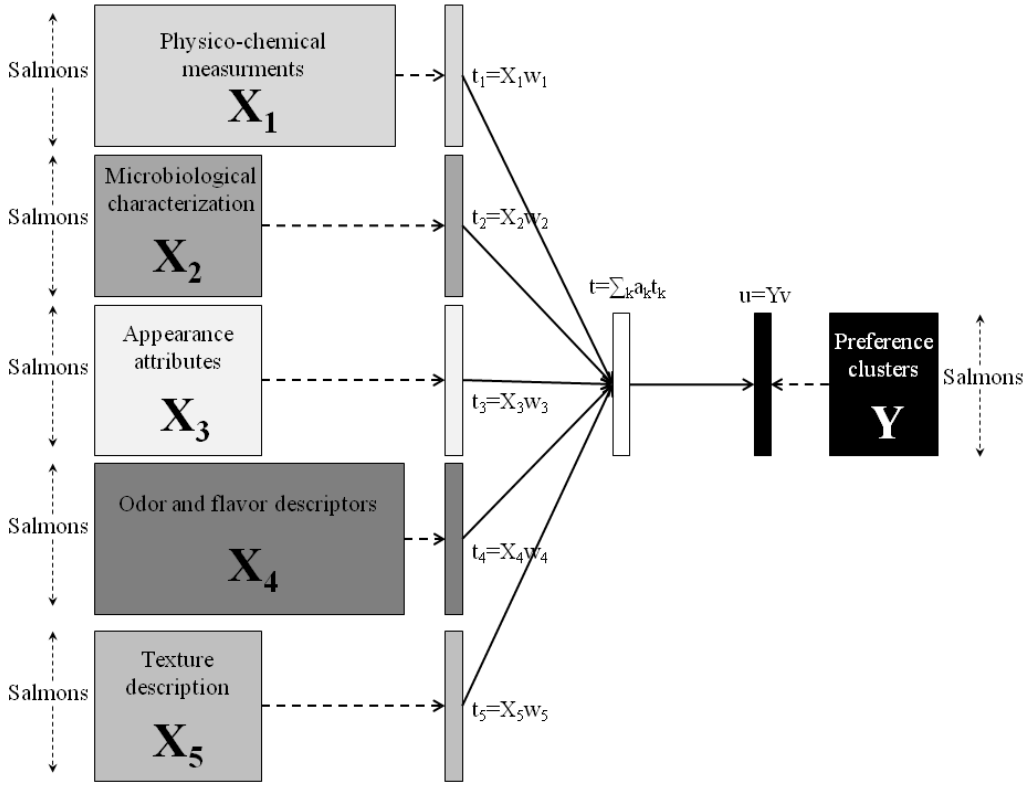

Fig.2. Graphical display of the relationships between datasets through their associated latent variables (first order dimension). Illustration on the smoked salmon data.

All these constraints can be summed up in a single criterion to be maximized based on the squared covariance between the latent variables $u^{(1)}=\mathbf{Y}v^{(1)}$ and $t^{(1)}=\mathbf{X}w^{(1)}$ where $t^{(1)}$ is defined as a synthesis of the partial datasets, namely $t^{(1)}=\sum_k a_k^{(1)} t_k^{(1)}$. Non symmetrical norm constraints, namely $||t_k^{(1)}||=||v^{(1)}||=1$, are chosen to improve the prediction ability of the associated model. We prove that the first order solution is given by the weight vector $v^{(1)}$, as the first eigenvector of $\sum_k \mathbf{Y'X_k(X_k'X_k)^{-1}X_k'Y}$ which allows taking the multiblock structure into account. Thereafter, the partial components ($t_1^{(1)}$... $t_K^{(1)}$) are given by the normalized projection of $u^{(1)}=\mathbf{Y}v^{(1)}$ on each subspace spanned by variables in blocks ($\mathbf{X_1}$...$\mathbf{X_K}$). Finally, the global latent variable $t^{(1)}$ is a weighted sum of the partial latent variables $t_k^{(1)}$. It appears that the more the components $u^{(1)}$ (preferences) and $t_k^{(1)}$ (external

attributes) are linked, the more they build the global component $t^{(1)}$. In order to extract more information from the explanatory variables and improve the preference prediction, higher order solutions are sought. It follows that the same criterion is maximized by replacing the datasets ($\mathbf{X_1}$...$\mathbf{X_K}$) with their residuals of regression onto the subspace spanned by $t^{(1)}$. Subsequent latent variables ($t^{(2)}$...$t^{(H)}$), H being the **X** rank, are sought by reiterating this process. An account of the method is detailed in (Bougeard et al., 2011).

To answer to the descriptive aim, the global latent variables t are used to highlight the common structure between the variables from at once explanatory and dependent blocks. If needed, the partial structures of each explanatory blocks may also be explored thanks to the partial components t_k. To better understand the links between blocks, correlations between global and partial components can be given. Global, but also partial, score and weight plots are interpreted in the same way as for standard two-block factorial analysis such as PLS.

To answer to the predictive aim of external preference mapping and achieve the key drivers of preference, a model between explanatory and dependent variable is provided. To avoid integrating too many variables and manage with multicollinearity, global latent variables ($t^{(1)}$...$t^{(H)}$) are used instead of original numerous explanatory **X** data. They are sought to be orthogonal by construction and ranked by order of importance in the explanation of the preferences **Y**. The optimal number of latent variables h.opt is selected thanks to a two-fold cross-validation procedure (Stone, 1974). It leads to $Y = \sum_{l=1}^{h.opt} t^{(l)} c^{(l)'} + Y^{(h.opt)}$, the vector of loadings c being the regression coefficients of **Y** onto the latent variables ($t^{(1)}$...$t^{(h.opt)}$) and $\mathbf{Y^{(h.opt)}}$ the residual matrix. Afterward, the original explanatory variables are found to provide a stable model liking preferences (**Y**) with external attributes (**X**) while using the property that components are linear combinations of **X**, *i.e.*, $t^{(h)} = \mathbf{X}w^{*(h)}$ (Wold, Martens, & Wold, 1983). It leads to the prefmap model: $Y = X[w^{*(1)}c^{(1)'} + ... + w^{*(h.opt)}c^{(h.opt)'}] + Y^{(h.opt)}$. This model can directly be compared with the vector model originally proposed by (Carroll, 1972) with the further advantages that latent variables are on the one hand linked not only with external attributes but also with preferences, and on the other hand take account of the underlying multiblock structure of external attributes.

2.3. Multiblock interpretation tools

Besides the standard regression coefficients between explanatory and dependent variables previously given, the sensometrician needs to sort explanatory variables by order of priority when the number of variables in **Y** is large, *e.g.*, several preference classes to be explained. An extension of the Variable Importance in the Prediction index (VIP), developed for PLS (Chong & Jun, 2005; Wold, 1994), to the multiblock framework is proposed. It sums up the overall contribution of each explanatory variable to the explanation of the whole preference. This new index, called VarImp, is based on the weighted squared weights w^{*2} for a model based on h.opt latent variables. Unlike the standard VIP, the VarImp index is clearly related to the multiblock framework both from the processing of w* and from the weighting by the block importance a_k. For simplicity sake, this index is expressed as percentage. Associated tolerance intervals are computed using bootstrapped simulations performed as for regression purpose (Freedman, 1981; Gosselin, Rodrigue, & Duchesne, 2010). As the indices verify the property $\sum_p \text{VarImp}_{\%}=100\%$, the threshold of 1/P is adopted, P being the number of explanatory variables. Then, each explanatory variable is considered to be a significant key driver if its 95% tolerance interval does not contain the threshold value of 100*1/P.

Finally, the sensometrician is also interested in assessing the contributions of the explanatory blocks in the overall preference explanation. We propose a specific multiblock index called the Block Importance index (short name: BlockImp) derived from the BIP index proposed by (Vivien, Verron, & Sabatier, 2005). It is based on the weighted a_k^2 coefficients, which reflect the link between each dataset $\mathbf{X_k}$ and the preferences **Y**, for a model based on h.opt latent variables. It can also be expressed as percentages. As previously, associated tolerance intervals are also computed thanks to bootstrapped simulations. As the indices verify the property $\sum_k \text{BlockImp}_{\%}=100\%$, the threshold of 1/K

is adopted, K being the number of explanatory datasets. Each explanatory block is considered to be a significant key block driver of preference if its 95% tolerance interval does not contain the threshold value of 100*1/K.

A detailed account of these tools is given in (Bougeard et al., 2011).

2.4. Comparison of multiblock modeling versus standard prefmap

It can be interesting to compare the model performance of multiblock modeling with the one of standard external preference mapping. We choose to focus on two features. (i) The first one aims at highlighting the influence of taking into account the multiblock structure. This requires the comparison of multiblock Redundancy Analysis applied to 25 explanatory variables organized in three sensory blocks (mbRA_ksenso), *i.e*., appearance, odor-flavor and texture, and the standard two-block Redundancy Analysis on 25 explanatory sensory variables (RA_senso). (ii) The second feature is to pinpoint the effects of the introduction of external variables in addition to sensory ones. This requires a comparison of multiblock Redundancy Analysis on 44 explanatory variables organized in five blocks (mbRA_ktab), *i.e.*, physico-chemical, microbiological, appearance, odor-flavor, texture, and of multiblock Redundancy Analysis applied to 25 explanatory variables organized in the three sensory latter ones (mbRA_ksenso). The method (RA_senso) is here viewed as the standard external preference mapping with a vector model based on Redundancy Analysis components, chosen instead of PCA or PLS components to get comparable results with multiblock methods. All these three models are compared on the basis of their average fitting ability (RMSEc comparable with RSS) and predictive ability (RMSEv comparable with PRESS) as functions of the number of components introduced in the model. The two-fold cross-validation procedure is repeated 500 times by setting one third of the observations out.

3. Results

3.1. Comparison of multiblock modeling versus standard prefmap

The results of the cross-validation procedure are displayed in Fig.3. The best model is the one which minimize the errors RMSEc and/or RMSEv. For the salmon data, it turns out that the standard prefmap method (RA_senso) has the best fitting ability but also the worst predictive one. In comparison, the proposed method mbRA_ktab has a correct fitting ability and the best prediction ability. In the following, this latter method will be interpreted with a model involving three components to get at once good fitting and prediction abilities.

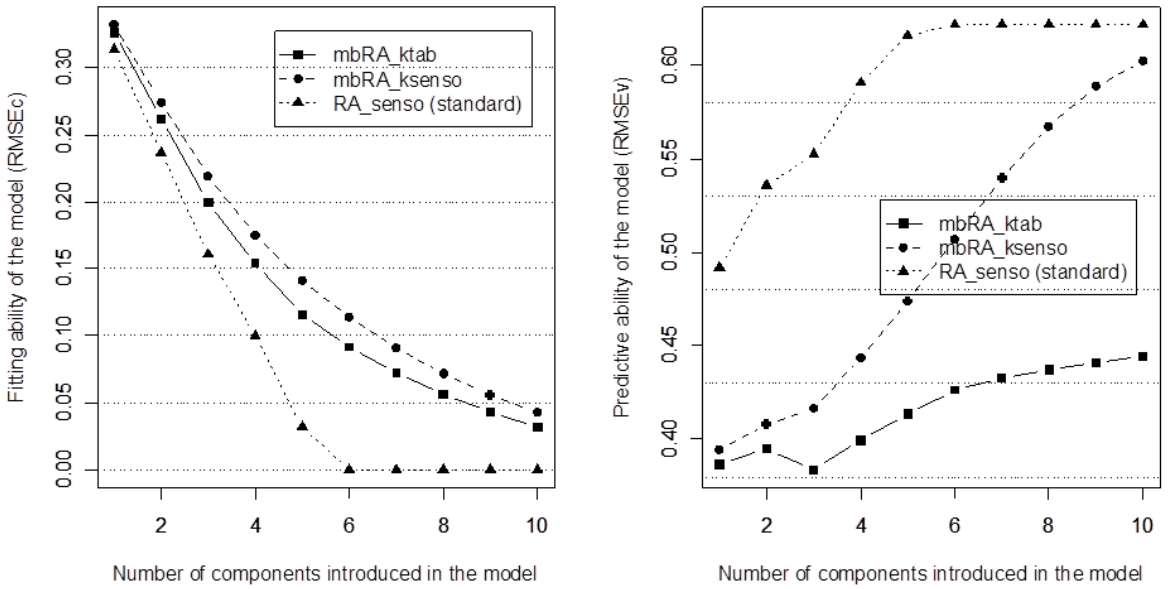

Fig.3. Fitting and prediction abilities of models as functions of the number of components introduced in the model (the best ability corresponds to the lower error value). Comparison of external preference mapping based on [square] multiblock Redundancy Analysis on 44 physico-chemical, microbiological and sensory data organized in 5 blocks (mbRA_ktab), [dot] multiblock Redundancy Analysis on 25 sensory data organized in 3 blocks (mbRA_ksenso) [triangle] and Redundancy Analysis on 25 sensory data organized in a single block (RA_senso). Illustration on the smoked salmon data for the first ten dimensions.

3.2. Descriptive interpretation

To explore the common structure of the data and the links between all the variables from ($\mathbf{X_1}$, .., $\mathbf{X_5}$) and **Y**, scores t and weights w* associated with the global latent variables orthogonal by construction are depicted. To get partial view of the links between $\mathbf{X_k}$ and **Y**, scores t_k and weights w_k associated with the partial latent variables can also be plotted (not presented here). As for two-block methods such as Redundancy Analysis or PLS, the global components t allow to describe the relationships between the explanatory variables (**X**) oriented towards the explanation of the dependent ones (**Y**).

The relationships between the 44 external attributes (**X**) and the 6 preference clusters (**Y**) are investigated from a descriptive point of view (question Q1 from section 2.1). In addition, the associated features of the thirty smoked salmons both on terms of external description and preferences are also given (question Q2). The optimal space to be interpreted is made of three dimensions which sum up 73.1% of the total inertia (respectively 34.1%, 20.2% and 18.8%). For simplicity sake, the third dimension is not interpreted in this section. The two graphical displays, *i.e.*, variable weights and individual scores, are given in Fig.4 for the first two global components.

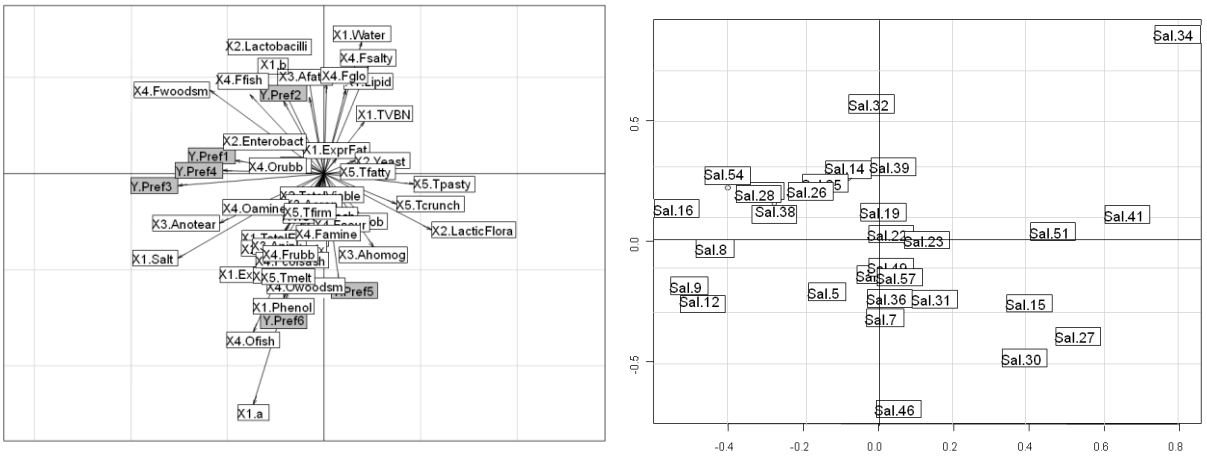

Fig.4. Graphical display of the variable weights (left) and individual scores (right) for the first two dimensions. Multiblock Redundancy Analysis of 6 clusters of preference (**Y**) explained with 44 variables organized in 5 blocks: physico-chemical ($\mathbf{X_1}$), microbiological ($\mathbf{X_2}$), appearance ($\mathbf{X_3}$), odor-flavor ($\mathbf{X_4}$) and texture ($\mathbf{X_5}$). Illustration on the smoked salmon data.

As for standard factorial analysis, inertia and explained variance of each dataset, *i.e.*, **Y**, **X** and ($\mathbf{X_1}$, ..., $\mathbf{X_K}$), with the global components, and eventually with the partial ones, can be processed (Bougeard et al., 2011). For the salmon data, the first two dimensions explain 54.3% of the overall inertia, 49.5% of the **Y** variance and 37.5% of the **X** variance. Thanks to the global latent variables, descriptive results allow describing the relationships between (i) the preference clusters (**Y** variables), (ii) the explanatory variables from all blocks (**X** variables) and above all (iii) between the preference clusters (**Y**) and the potential preference drivers (**X**). It turns out that three main directions of preference are highlighted: the preference clusters 1 (N_1=121), 3 (N_3=349) and 4 (N_4=78) can be compared in terms of preference, as well as the clusters 5 (N_5=404) and 6 (N_6=37), these latter clusters being on the opposite side from the cluster 2 (N_2=74). As a remark, the cluster 1 appears to be grouped together with the cluster 2 for the third dimension. It seems that each of these clusters is associated with specific preferred and disliked salmons and with specific preference drivers. The preferred salmons of clusters 1, 3 and 4 (average preference higher than 6) are Sal.38, Sal.28 and Sal.9 and the disliked ones (average preference lower than 4) are Sal.34, Sal.30 and Sal.31. These three clusters mainly like salty salmons (physico-chemical block, $\mathbf{X_1}$) with a high wood smoke flavor (odor-flavor block, $\mathbf{X_4}$) and mainly dislike salmons with a high level of lactic flora (microbiological block, $\mathbf{X_2}$) and a crunchy texture (texture block, $\mathbf{X_5}$). Then, the preferred salmons of clusters 5 and 6 are Sal.15, Sal.54 and Sal.27 and the disliked one is Sal.39. These two clusters mainly like rather red salmons with high phenol content (physico-chemical block, $\mathbf{X_1}$) and intense wood smoke odor (odor-flavor block, $\mathbf{X_4}$) and dislike salty, intense global flavor (odor-flavor block, $\mathbf{X_4}$) and high water content (physico-chemical block, $\mathbf{X_1}$) salmons. Finally, the preferred salmons of clusters 2 are Sal.28, Sal.25, Sal.8, Sal.2 and Sal.16 and the disliked one is Sal.27. This latter cluster mainly likes salmons where a high count of *Lactobacilli* is measured (microbiological block, $\mathbf{X_2}$), an intense fish flavor (odor-flavor block, X_4) and a rather yellow color (physico-chemical block, $\mathbf{X_1}$). This cluster also dislikes salmons with a crunchy texture (texture block, $\mathbf{X_5}$) as the clusters 1, 3 and 4.

3.3. Predictive interpretation: key drivers of preference at the variable level

Thereafter, it is of paramount importance to assess the key drivers of the smoked salmon preferences from the forty-four external attributes organized in five meaningful blocks. The optimal model involves three latent variables and explains 68.8% of the variation in the preference dataset. The regression coefficients of the model allow finding out the significant links between all the 44 external attributes and each of the six clusters of preferences (question Q3 from section 2.1; detailed results not presented here). It follows that each cluster is related with specific key preference drivers but this information are too scattered. The variable importance index allows sorting by order of

priority the 44 explanatory variables describing the smoked salmons in relation to the overall preferences (question Q4 from section 2.1). Fig.5 gives the Variable Importance index of the 14 most significant explanatory variables on the whole preferences (VarImp>1/44%=2.3%). Associated tolerance intervals are computed using results from 500 bootstrapped samples. For each of these 14 variables, is additionally given the importance of each preference cluster obtained from the absolute value of the six associated regression coefficients. Although these results are akin to those obtained from the standard VIP, the processing of the VarImp index is specific to the multiblock framework.

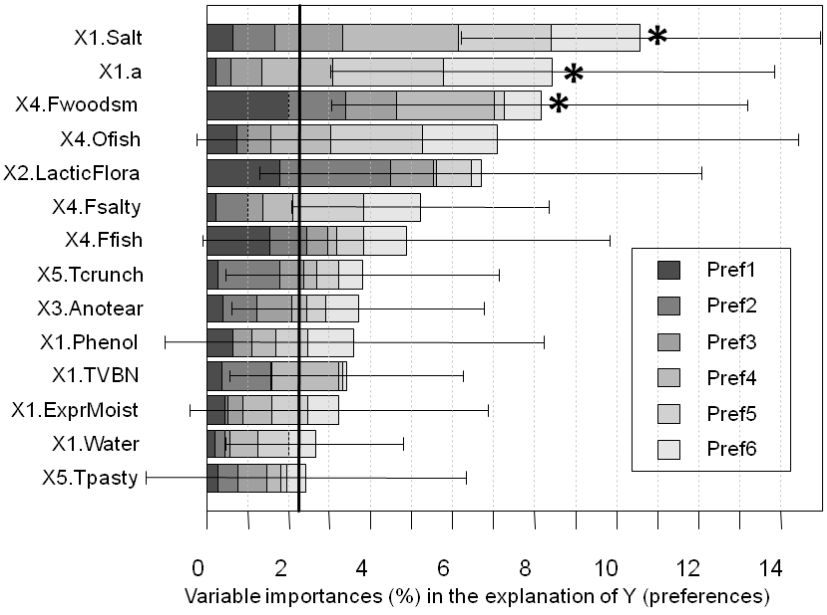

Fig.5. Graphical display of the VarImp index of the most important explanatory variables associated with their 95% tolerance interval for a model involving three latent variables. The threshold value is represented at 1/44%=2.27%. Multiblock Redundancy Analysis of 6 clusters of preference (**Y**) explained with 44 variables organized in 5 blocks: physico-chemical ($\mathbf{X_1}$), microbiological ($\mathbf{X_2}$), appearance ($\mathbf{X_3}$), odor-flavor ($\mathbf{X_4}$), texture ($\mathbf{X_5}$). Illustration on the smoked salmon data.

It follows that only 14 external variables, among 44, explain 73.8% of the preference variation. Among these 14 explanatory variables, three have a significant impact on the whole preference: the salt content ($VarImp_{Ssalt}$=10.5% $[6.2;15.0]_{95\%}$, X_1), the a measurement which represent the red dimension of color ($VarImp_a$=8.4% $[3.0;13.8]_{95\%}$, X_1) and the wood smoke flavor ($VarImp_{Fwoodsm}$=8.1% $[3.0;13.2]_{95\%}$, X_4). These three drivers of preference explain 27.1% of the overall salmon preference and it is important to notice that three out of these four variables are easy-to-get physico-chemical parameters Furthermore, interpretation of the variable importance can also be related with preference. For example, the a measurement is more important for the clusters 4, 5 and 6 ($Beta_{Pref4}$=1.03; $Beta_{Pref5}$=-1.60; $Beta_{Pref6}$=1.58) which like red salmons, rather than for the clusters 1, 2, and 3 ($Beta_{Pref1}$=0.15; $Beta_{Pref2}$=-0.21; $Beta_{Pref3}$=0.45), where beta stands for the regression coefficients.

3.4. Predictive interpretation: key drivers of preference at the block level

In the same vein, the key drivers of preference can be viewed at the block level with the block importance index. It reflects the contribution of the five external attribute blocks to the explanation of the whole preferences (question Q5 from section 2.1). It allows sorting external blocks by order of priority. Fig.6 gives the relative importance of each explanatory block $\mathbf{X_k}$ in the explanation of **Y**. The threshold value for the block significance is set to 1/K=20%. It follows that the overall preferences are

mainly explained by the odor and flavor attributes ($BlockImp_{X4}$=31.6% $[28.4;34.9]_{95\%}$), and by the physico-chemical measurements ($BlockImp_{X1}$=27.4% $[24.4;30.4]_{95\%}$).

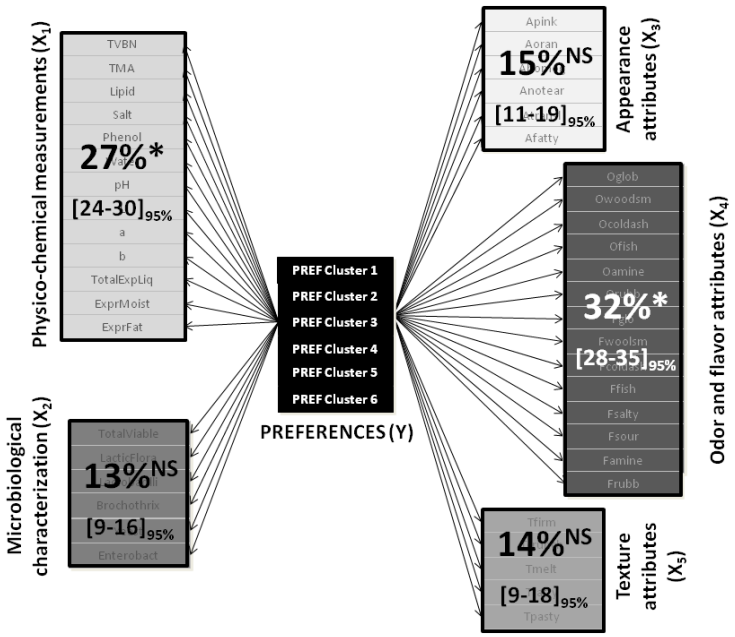

Fig.6. Graphical display of the BlockImp index of the explanatory blocks associated with their 95% tolerance interval for a model involving three latent variables. Multiblock Redundancy Analysis of 6 clusters of preference (**Y**) explained with 44 variables organized in 5 blocks: physico-chemical ($\mathbf{X_1}$), microbiological ($\mathbf{X_2}$), appearance ($\mathbf{X_3}$), odor-flavor ($\mathbf{X_4}$) and texture ($\mathbf{X_5}$). Illustration on the smoked salmon data.

4. Discussion

4.1. Discussion on method

The main point to discuss is the contribution of multiblock modeling to the framework of external preference mapping while taking account both the block structure and possibly new external variables in addition to sensory ones.

The first point to consider is the interest of blocking explanatory variables in a preference mapping. The first way may consist in splitting up sensory variables into several blocks, such as appearance, smelling, texture and flavors, and assessing the block weights in the preference explanation. In terms of broad results (only), this can be compared to conjoint analysis for trade-off in choice, as blocks can be compared with each other in regard with preference (Hair, Anderson, Tatham, & Black, 1998). The second way of blocking explanatory variables is to add external ones, such as instrumental, price or packaging properties, to sensory ones. Indeed, the presence of external data may contribute to improve the preference modeling as these variables either emphasize (if they are linked to sensory ones) or enhance (if they are not linked) the model. This can solve the problem of the possible shortcomings of the attribute list and the inadequacy of product attribute and preference spaces, and then expect to improve the model quality. To avoid overestimating the model, a cross-validation procedure such as the two-fold one presented in section 3.1 is highly recommended with special attention to the prediction ability. As all the tools, including cross-validation, are available in a free R package soon included within the ade4 software (Dray & Dufour, 2007), the same kind of study can be assessed on every data to decide the interest of adding external variables and arranging them into blocks.

It is worth noting that including external variables in a preference mapping can be solved in some other (simpler) ways. The first approach is to assess a standard prefmap where external and sensory variables are mixed in a single explanatory block. The first criticism is that too many variables in an only block may lead to an unstable model as they are often plagued by colinearity. In addition, block scaling is here not possible and a high variable number in a block, *e.g.*, near-infrared spectral data in comparison to sensory data, may influence the selection of key preference drivers (Derksen & Keselman, 1992). The second (most used) method is to linked external variables with sensory and preference attributes in an additional step (*e.g.*, (Semenou et al., 2007)). Conversely, multiblock analysis allows exploring relationships between instrumental, sensory and preference variables at the same time in a single analysis, in the common space of the global latent variables in regard with the products under study. A third existing approach consists in using a standard two-block method where block information are added afterwards. This solution could be the one of multiblock PLS which leads to a standard PLS for the case of a single dependent dataset. In this case, the latent variables associated with **X** and **Y**, resp. t and u, are the same as in standard PLS, but partial latent variables t_k and multiblock interpretation tools can also be computed. This solution leads to the same criticism as for the first reported approach.

In the framework of external preference mapping, multiblock Redundancy Analysis provides some specific benefits for users. First of all, depending on the data processing aim, the user may decide to scale all blocks to the same footing or not. Then, this method is based on latent variables instead of manifest ones; this leads to more stable results especially when variables are numerous as it restricts the problem of quasi-colinearity. In addition, the multiblock structure of explanatory variables is taken into account: each dataset $\mathbf{X_k}$ is summed up with a partial latent variable t_k, these partial latent variables being linked with the dependent one in an overall criterion to maximize; the direct eigensolution involves the $\mathbf{X_k}$ and **Y** datasets. Thereafter, common or partial latent variables can be used either through factorial graphical displays or models; to get an overall interpretation of preference key drivers, we choose here to only present results from the common structure. Finally in terms of interpretation, in comparison with the results obtained from a standard external preference mapping, *i.e.*, factorial graphical displays of variables and products, regression coefficients between preferences and external attributes, multiblock Redundancy Analysis brings both the same kinds of results and supplementary ones at the block level. This facilitates the interpretation of the added information.

Standing in the field of multiblock analysis with a prediction purpose, other related aims can be attempted. While using the PO-PLS method proposed in the same field, one can go further within the descriptive purpose and split up common and specific information in blocks (Måge et al., 2012). While using the PLS-PM methodology proposed in a comparable field (Pagès & Tenenhaus, 2001), one can additionally explore relationships between explanatory blocks with the limit of taking account a single dimension within each block for the prediction purpose. Closest to multiblock Redundancy Analysis in terms of aims and interpretation, multiblock PLS is a better tool in case of strong multicollinearity within explanatory blocks (*e.g.*, spectroscopic data) with the limit that for the case of a single dataset to explain, the method amounts to a standard PLS.

4.2. Discussion on application conditions

The manifest variables being summed up with latent variables, they can be numerous in a block. The dependent variables can be either the whole preference dataset (variables are consumers) or the preference clusters. For reasons of clarity in interpretation, clusters are more often chosen. In terms of variable number, two limits must be highlighted. The first one concerns the ratio of the number of products to the number of explanatory variables, as it is well-known that the result stability is impaired when the ratio value really decreases. The prediction ability of the cross-validation procedure is a helpful and available tool to avoid over fitting model. The second one is

about strong multicollinearity within explanatory blocks; in such a case, we recommend to use multiblock PLS (Bougeard & Qannari, 2011). In regard with its eigensolution, multiblock Redundancy Analysis is sensitive to multicollinearity within but not between blocks.

Multiblock Redundancy Analysis, as well as multiblock PLS and generally as factorial analysis, is not invariant with respect to both variable and block scale. This brings users to answer to the question on the relative importance of variables and blocks in the analysis beforehand. One main advantage of the block scaling is to allow unbalanced number of variables in each block without any trouble of down or under weight. An important assertion for users is that multiblock Redundancy Analysis, as well as multiblock PLS, has no limit in terms of block number.

4.3. Discussion on the smoked salmon case study

Following the block importance interpretation (section 3.3), the overall preferences for cold smoked salmon could be efficiently improved while focusing on odor and flavor attributes (32% of the preference variation) and on physico-chemical measurements (27% of the preference variation) relatively easy to get. It follows that a sensory panel specialized in assessing salmon odors and flavors could be an efficient tool to understand preferences. The microbiological characterization is not useful to understand consumer preference but necessary in a hygienic quality perspective.

To be more accurate, multiblock Redundancy Analysis leads to consider only 14, among 44, important drivers of preference to explain 73.8% of the preference variation (section 3.3). We can notice that only three of them are significant. These drivers mainly pertain to the physico-chemical block, namely the salt content, the a dimension of color and in a lesser extent the total phenol, the total volatile basic nitrogen content, the expressible moisture and the water contents. These drivers also pertain to the odor and flavor block, namely the wood smoke flavor, the raw salmon odor and flavor and the salty taste. To go further in the external attribute selection, it is important to notice that the wood smoke flavor is linked with the total phenol content and that the salty taste can be explained with the water content and the salt content (Cardinal et al., 2004). From all these results, R&D and marketing teams can directly focus on about ten relevant preference drivers to adapt the salmon product to the consumer requests. Sensometricians may also focus on some restricted sensory attributes to specialize the trained judges for further studies.

Multiblock Redundancy Analysis also brings specific key preference drivers for each homogeneous cluster of smoked salmon preference (section 3.2). It follows that three main directions of preference can be highlighted following clusters 1, 3 and 4, clusters 5 and 6, and finally cluster 2. Each of these directions can be associated with specific preferred and disliked salmons and with specific preference drivers. It is important to notice that this interpretation is coherent with the ones from (Cardinal et al., 2004; Semenou et al., 2007). After dealing with this overall preference study, it could be of paramount importance to focus on each preference cluster request. This further step could be carried out while processing six multiblock analyses applied to each variable in **Y**, *i.e.*, the preferences for each cluster. It leads to more accurate responses at the cluster level.

5. Conclusion and perspectives

In this paper, we propose to apply an innovative method, called multiblock Redundancy Analysis, to the framework of external preference mapping. The proposed method models the preference space with the product attribute space organized in thematic blocks. It increases the amount of extracted knowledge from the data in comparison with standard prefmap. Although standard preference mapping gives complete, accurate, sensible, but usually too scattered results, the sensometrician needs to get overall results to pass them on to R&D and marketing teams. Multiblock Redundancy Analysis makes possible to shed light on significant external variables affecting a composite dependent one, *i.e.*, the preferences, and to pinpoint key preference drivers within the various blocks. It is well-adapted to complex issues and may be a relevant decision-

support aid in preference management. This meets the specific expectations of the various actors in the food development of products confronted with preference complexity. Furthermore, the method allows many possibilities of graphical displays and combines tools from factor analysis with tools pertaining to regression. Multiblock Redundancy Analysis, together with multiblock PLS and all the associated interpretation tools, are freely available for users thanks to an R package soon included within the ade4 software (Dray & Dufour, 2007).

Still in the field of external preference mapping, multiblock methods may also be applied to sensory external data only, while taking account the underlying structure of the sensory profiles, namely appearance ($\mathbf{X_1}$), smelling ($\mathbf{X_2}$), tastes ($\mathbf{X_3}$), flavors ($\mathbf{X_4}$) and texture ($\mathbf{X_5}$). It leads to assess and compare the weights of all these blocks in the preference explanation. To go further in the product development field, multiblock methods can also be implemented to study product quality, where the quality (**Y** dataset, described with several variables) could be explained with consumer preferences ($\mathbf{X_1}$), sensory attributes ($\mathbf{X_2}$) and spectrometric measurements ($\mathbf{X_3}$) among others. In the more specific framework of control quality, the final product quality (**Y**) could also be explained with instrumental measurements carried out at K different steps of the process ($\mathbf{X_1}$...$\mathbf{X_K}$). Multiblock methods appear to be useful for various sectors in the food industry.

However, the multiblock approach presents some limitations and further investigations will be undertaken to handle more data specificities. For instance, the drawback of the vector model used at present in multiblock methods neglects that for some external attributes, preference can increase until an optimal value and then decrease. As for standard external preference mapping, the vector model could be improved while using a quadratic one. Following the ideas of (Verdun, Hanafi, Cariou, & Qannari, 2012) for PLS, both linear and quadratic terms could be involved in the maximization problem. This will allow providing ideal-point models in multiblock external preference mapping. In addition, multiblock modeling does not efficiently handle the information from hierarchical-structured data, *e.g*., a design matrix applied to products frequently met in food industry. Following the ideas of (Eslami, Qannari, Kohler, & Bougeard, In Press) for PLS, taking account for multi-group structure of products will be soon developed for multiblock methods.

Acknowledgements

The data used in this paper come from the Eurosalmon project (QLK1-2000-01575) with the financial support of the European Union. We would also like to thank the two anonymous referees for their relevant comments that improve the paper.

Appendix

Block	Variable code	Variable description
Y (Preferences)	Pref1	Preferences of cluster 1
	Pref2	Preferences of cluster 2
	Pref3	Preferences of cluster 3
	Pref4	Preferences of cluster 4
	Pref5	Preferences of cluster 5
	Pref6	Preferences of cluster 6
X1 (Physico-chemical measurements)	TVBN	Total Volatile Basic Nitrogen
	TMA	Trimethylamine content
	Lipid	Total fat content
	Salt	Salt level
	Phenol	Total phenol level
	Water	Water content

	pH	pH
	L	Lightness
	a	Hue parameter (red)
	b	Hue parameter (yellow)
	TotalExpLiq	Expressible liquid
	ExprMoist	Expressible moisture
	ExprFat	Expressible fat
X2 (Microbiologial characterization)	TotalViable	Total psychrotrophic count
	LacticFlora	Lactic flora
	Lactobacilli	Lactobacillus bacteria
	Brochothrix	Brochothrix bacteria
	Yeast	Yeast
	Enterobact	Enterobacter bacteria
X3 (Appearance attributes)	Apink	Pink color
	Aoran	Orange color
	Ahomog	Color homogeneity
	Anotear	Degree of slice tearing
	Atransl	Translucent appearance
	Afatty	Fatty aspect
X4 (Odor and flavor descriptors)	Oglob	Global odour intensity
	Owoodsm	Wood smoke odor
	Ocoldash	Cold ash odor
	Ofish	Raw salmon odor
	Oamine	Amine odor
	Orubb	Rubber odor
	Fglo	Global flavour intensity
	Fwoodsm	Wood smoke flavour
	Fcolsash	Cold ash flavour
	Ffish	Raw salmon flavour
	Fsalty	Salty taste
	Fsour	Sour taste
	Famine	Amine flavour
	Frubb	Rubber flavour
X5 (Texture attributes)	Tfirm	Firm texture
	Tcrunch	Crunchy texture
	Tmelt	Melting texture
	Tfatty	Fatty texture
	Tpasty	Pasty texture

References

Adriant, Ifremer, IMR, & Matra. (2004). Smoked salmon : Mapping of European Consumer Preferences. Final report of European project QLK1-2000-01575, Improved quality of smoked salmon for the European consumer (Eurosalmon).

Bougeard, S., & Qannari, E. M. (2011). Continuum Approach for Multiblock Methods: Overview and Regularization Purpose. In, *14th Conference of the Applied Stochastic Models and Data Analysis International Society*. Roma, Italy.

Bougeard, S., Qannari, E. M., & Rose, N. (2011). Multiblock Redundancy Analysis: interpretation tools and application in epidemiology. *Journal of chemometrics, 25*, 467-475.

Cardinal, M., Gunnlaugsdottir, H., Bjoernevik, M., Ouisse, A., Vallet, J. L., & Leroi, F. (2004). Sensory characteristics of cold-smoked Atlantic salmon (*Salmo salar*) from European market and relationships with chemical, physical and microbiological measurements. *Food Research International, 37*(2), 181-193.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, Romney, A.K., Nerlove, S.B., *Multidimensional scaling: theory and applications in the behavioral sciences*. New York: Seminar Press.

Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems, 78*, 103-112.

Courcoux, P., Qannari, E. M., & Schlich, P. (2006). Sensometrics workshop: Segmentation of consumers and characterization of cross-cultural diffrences. *Food quality and preference, 17*, 658-668.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms : frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*, 265-282.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software, 22*, 1-20.

Eslami, A., Qannari, E. M., Kohler, A., & Bougeard, S. (In Press). Two-block multi-group data analysis. Application to epidemiology. In G. Russolillo, *New perspectives in Partial Least Squares and Related Methods*: Springer Verlag.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics, 9*, 1218–1228.

Gosselin, R., Rodrigue, D., & Duchesne, C. (2010). A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometrics and intelligent laboratory systems, 100*(1), 12-21.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Conjoint analysis. In J. F. Hair, Anderson ,R.E., Tatham, R.L., Black, W.C. , *Multivariate data analysis*. Prentice Hall, New Jersey.

Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm *Computational statistics, 25*(1), 107-120.

Jaeger, S. R., Wakeling, I. N., & Macfie, H. J. H. (2000). Behavioural extensions to preference mapping: the role of synthesis. *Food quality and preference, 11*, 349-359.

Kissita, G. (2003). *Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués (PhD thesis).* University of Paris Dauphine IX, Paris.

Måge, I., Menichelli, E., & Naes, T. (2012). Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food quality and preference, 24*, 8-16.

Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and intelligent laboratory systems, 58*(2), 261-273.

Semenou, M., Courcoux, P., Cardinal, M., Nicod, H., & Ouisse, A. (2007). Preference study using a latent class approach. Analysis of European preferences for smoked salmon. *Food quality and preference, 18*, 720-728.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, 36*(1), 111-147.

Verdun, S., Hanafi, M., Cariou, V., & Qannari, E. M. (2012). Quadratic PLS1 regression revisited. *Journal of chemometrics, 26*, 384-389.

Vivien, M., Verron, T., & Sabatier, R. (2005). Comparing and predicting sensory profiles by NIRS: Use of the GOMCIA and GOMCIA-PLS multi-block methods. *Journal of chemometrics, 19*, 162-170.

Westerhuis, J. A., & Coenegracht, P. M. J. (1997). Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of chemometrics, 11*(5), 379-392.

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS model. *Journal of chemometrics, 12*, 301-321.

Wold, S. (1984). Three PLS algorithms according to SW. In S. Wold, *Symposium MULDAST (multivariate analysis in science and technology)*. Umea University, Sweden.

Wold, S. (1994). PLS for multivariate linear modeling. In H. van der Waterbeemd, *QSAR: Chemometric methods in molecular design: Methods and principles in medicinal chemistry*. Weinheim, Germany: Verlag-Chemie.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In A. Ruhe, Kastrom, B., *Proceedings of the Conference on Matrix Pencils*: Springer Verlag, Heidelberg.