Multiblock modeling for complex preference study

Application to European preferences for smoked salmon

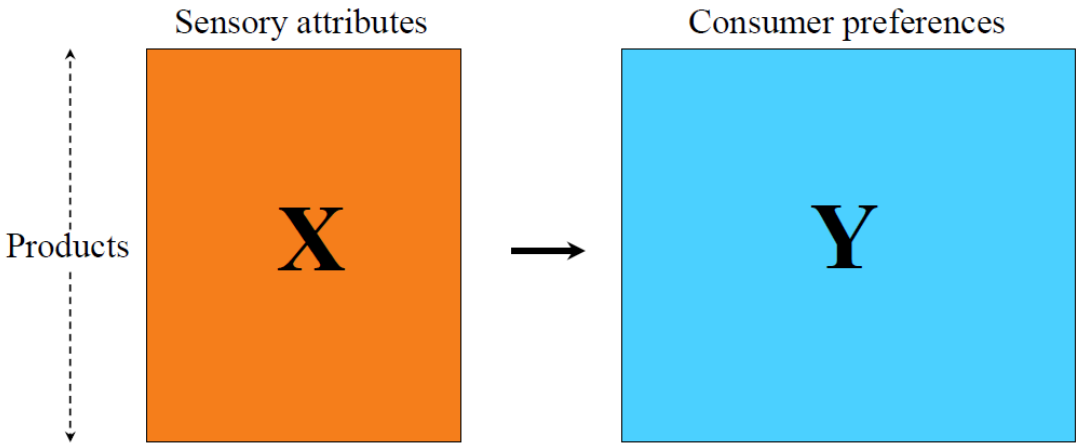
S. Bougeard[1] & M. Cardinal[2]

[1] *French Agency for food, environmental and occupational health & safety (Anses), Department of epidemiology*
[2] *French research institute for exploration of the sea (IFREMER), Department of biotechnologies and sea resources*

Table of contents

Linking sensory to consumer data

External multiblock preference mapping



Data features

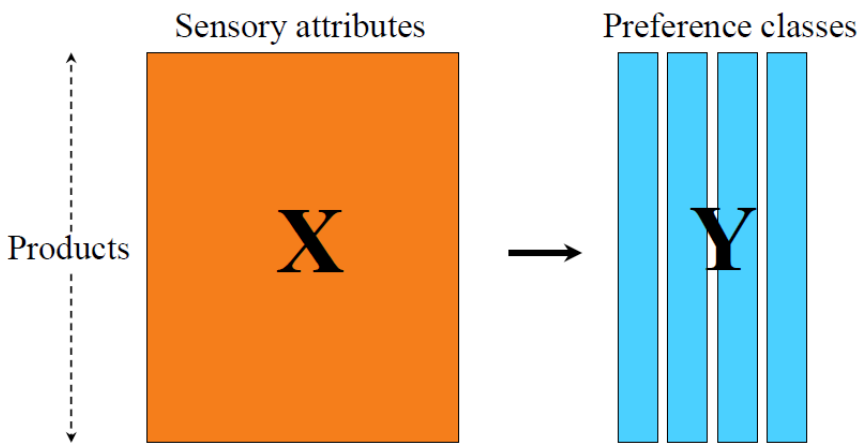
- Large number of explanatory variables (**X**) organized into meaningful blocks,
- Several variables to explain (**Y**) usually summed up in homogeneous classes.

Aims

- **Descriptive** : investigate the relationships between variables and between blocks in relation with the products,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Linking sensory to consumer data

External multiblock preference mapping



Data features

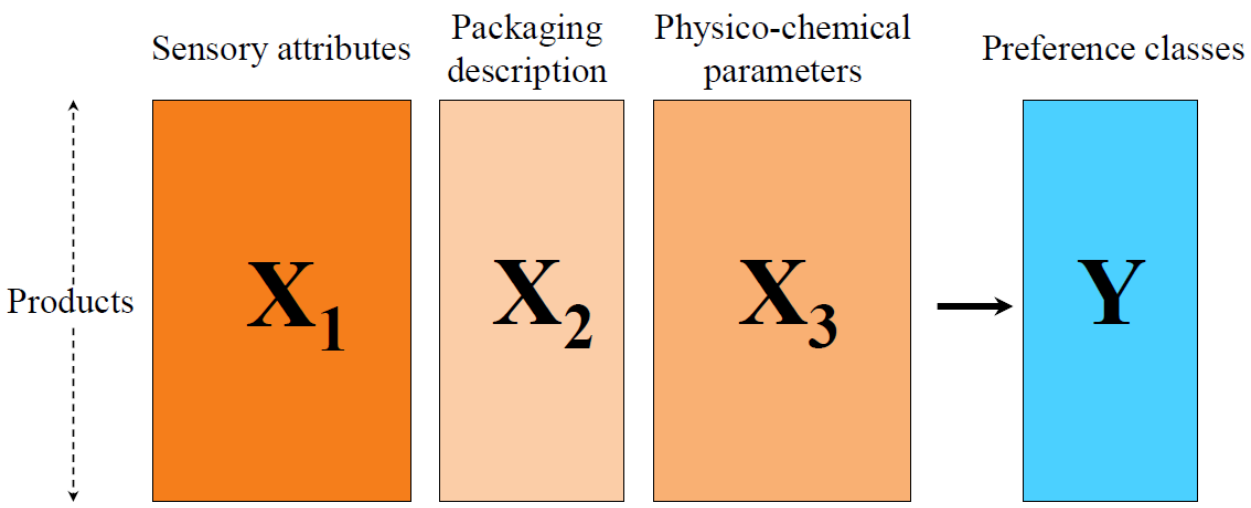
- Large number of explanatory variables (**X**) organized into meaningful blocks,
- Several variables to explain (**Y**) usually summed up in homogeneous classes.

Aims

- **Descriptive** : investigate the relationships between variables and between blocks in relation with the products,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Linking sensory to consumer data

External multiblock preference mapping



Data features

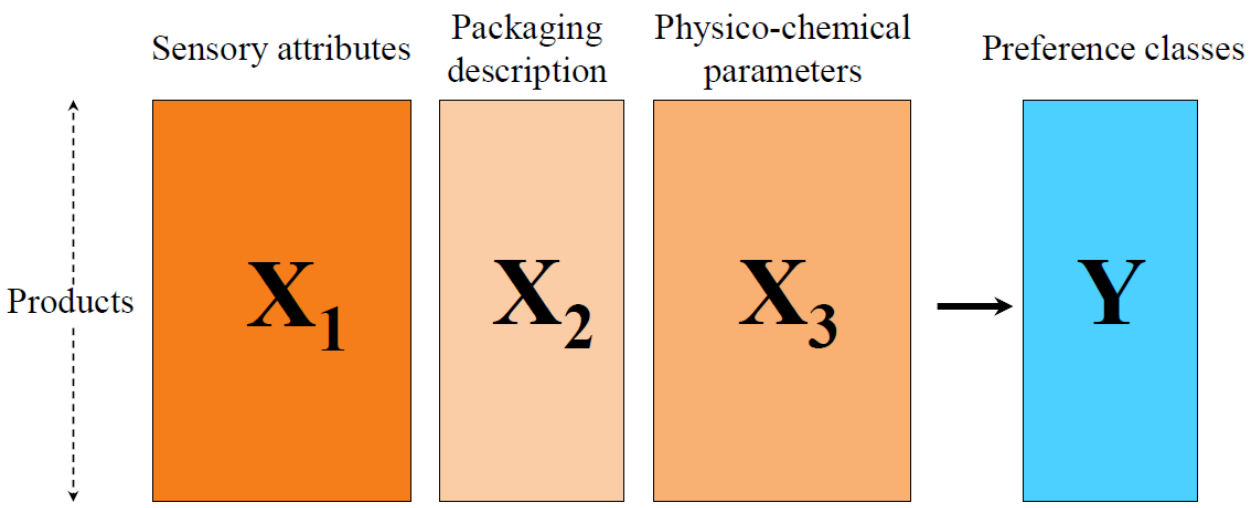
- Large number of explanatory variables (**X**) organized into meaningful blocks,
- Several variables to explain (**Y**) usually summed up in homogeneous classes.

Aims

- **Descriptive** : investigate the relationships between variables and between blocks in relation with the products,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Linking sensory to consumer data

External multiblock preference mapping



Data features

- Large number of explanatory variables (**X**) organized into meaningful blocks,
- Several variables to explain (**Y**) usually summed up in homogeneous classes.

Aims

- **Descriptive** : investigate the relationships between variables and between blocks in relation with the products,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Usual data processing for external multiblock preference mapping

Too simple data processing

- First step : link sensory attributes ($\mathbf{X_1}$) to preferences ($\mathbf{Y}$),
- Second step : link other measurements ($\mathbf{X_2}, \ldots, \mathbf{X_K}$) to preferences ($\mathbf{Y}$) to better characterize them.
- **Limit** : no overall resolution.

Too sophisticated data processing

- Use structural equation modeling (*e.g.*, PLS path modeling) to link ($\mathbf{X_1}, \ldots, \mathbf{X_K}$) to preferences $\mathbf{Y}$,
- **Limit** : too complicated iterative algorithm (no convergence proof).

Some multiblock proposals

- Multiblock PLS [Wold, 1984] : for the case of a single dataset $\mathbf{Y}$, mbPLS=PLS,
- PO-PLS [Måge et al., 2008] : complicated iterative algorithm.

Usual data processing for external multiblock preference mapping

Too simple data processing

- First step : link sensory attributes ($\mathbf{X_1}$) to preferences ($\mathbf{Y}$),
- Second step : link other measurements ($\mathbf{X_2}, \ldots, \mathbf{X_K}$) to preferences ($\mathbf{Y}$) to better characterize them.
- **Limit** : no overall resolution.

Too sophisticated data processing

- Use structural equation modeling (*e.g.*, PLS path modeling) to link ($\mathbf{X_1}, \ldots, \mathbf{X_K}$) to preferences $\mathbf{Y}$,
- **Limit** : too complicated iterative algorithm (no convergence proof).

Some multiblock proposals

- Multiblock PLS [Wold, 1984] : for the case of a single dataset $\mathbf{Y}$, mbPLS=PLS,
- PO-PLS [Måge et al., 2008] : complicated iterative algorithm.

Usual data processing for external multiblock preference mapping

Too simple data processing

- First step : link sensory attributes ($\mathbf{X_1}$) to preferences ($\mathbf{Y}$),
- Second step : link other measurements ($\mathbf{X_2}, \ldots, \mathbf{X_K}$) to preferences ($\mathbf{Y}$) to better characterize them.
- **Limit** : no overall resolution.

Too sophisticated data processing

- Use structural equation modeling (*e.g.*, PLS path modeling) to link ($\mathbf{X_1}, \ldots, \mathbf{X_K}$) to preferences $\mathbf{Y}$,
- **Limit** : too complicated iterative algorithm (no convergence proof).

Some multiblock proposals

- Multiblock PLS `[Wold, 1984]` : for the case of a single dataset $\mathbf{Y}$, mbPLS=PLS,
- PO-PLS `[Måge et al., 2008]` : complicated iterative algorithm.

Aims of the talk

Methodological contribution

- Presentation of an original multiblock modeling method,
- Development of useful and relevant associated interpretation tools.

Applicative contribution

- Interpreted example of an external multiblock preference mapping in the food field,
- Available code program in `R` : `ktab+1` package included within the `ade4` package (http ://pbil.univ-lyon1.fr/ade4/).

Aims of the talk

Methodological contribution

- Presentation of an original multiblock modeling method,
- Development of useful and relevant associated interpretation tools.

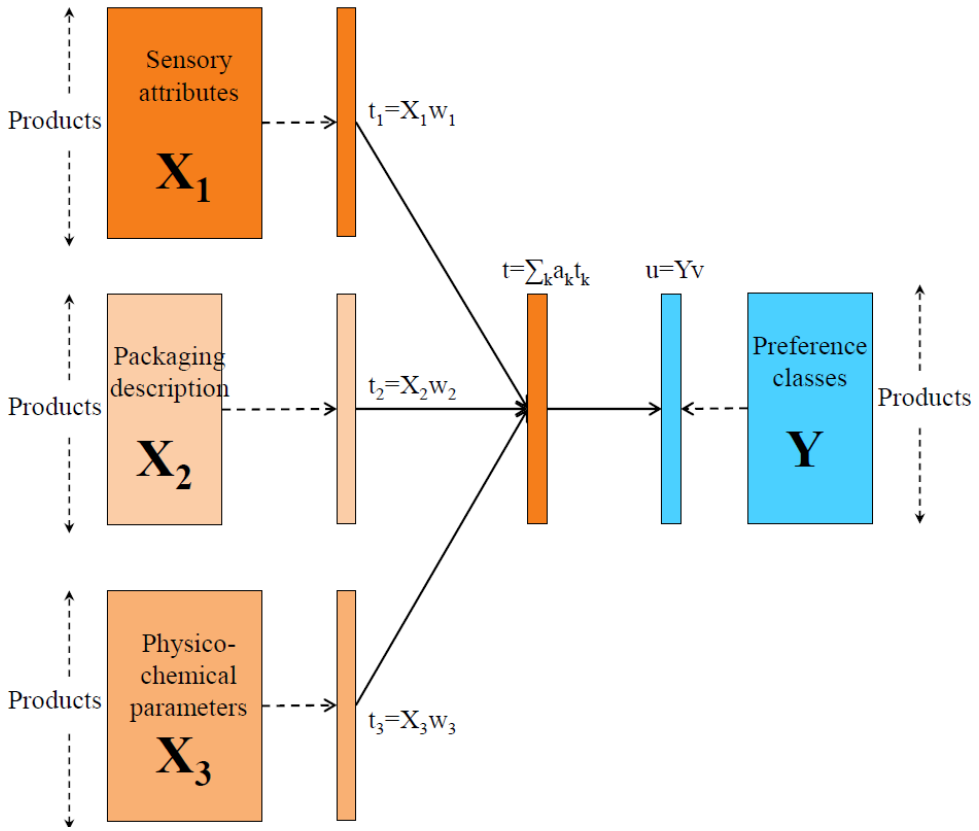
Applicative contribution

- Interpreted example of an external multiblock preference mapping in the food field,
- Available code program in `R` : `ktab+1` package included within the `ade4` package (http ://pbil.univ-lyon1.fr/ade4/).

Table of contents

Overall criterion to maximize

Multiblock redundancy analysis [Bougeard et al., 2011]



Key ideas

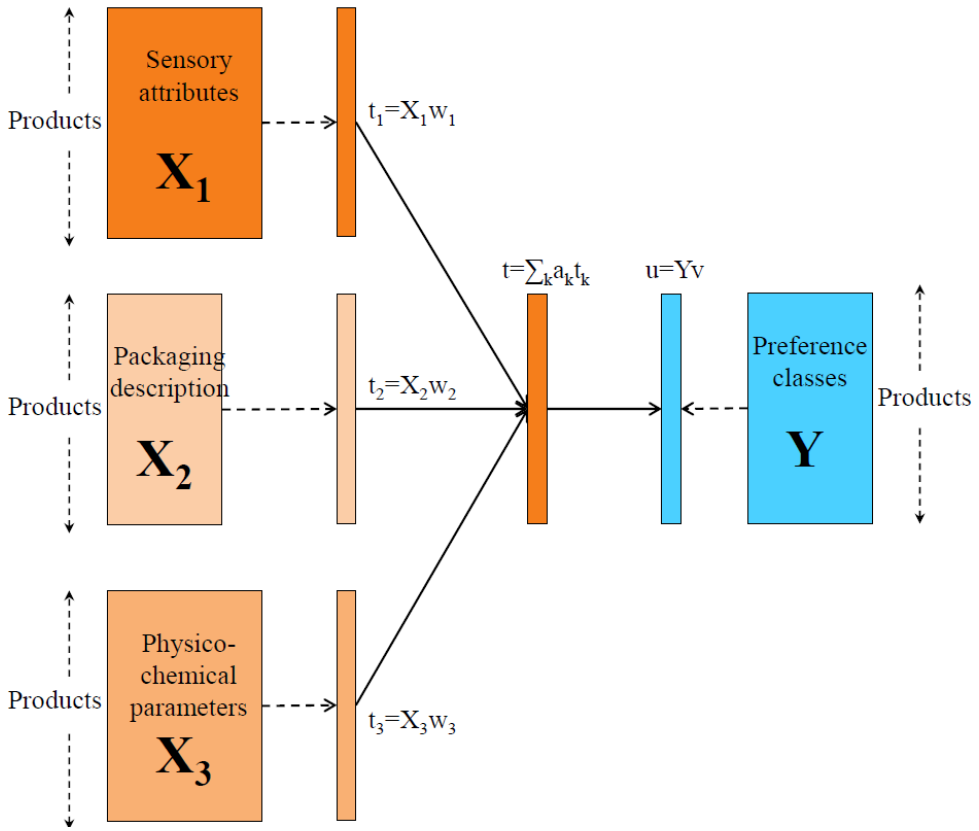
- Sum up each dataset with latent variables,
- Relate the explanatory latent variables with the dependent one,
- Seek all these latent variables by maximizing an overall criterion,
- Get a direct eigensolution.

Criterion to maximize

Max. $\sum_k cov^2(u^{(1)}, t_k^{(1)})$ with $u^{(1)} = Yv^{(1)}$, $t_k^{(1)} = X_k w_k^{(1)}$, $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$, $\sum_k a_k^{(1)^2} = 1$ and $||v^{(1)}|| = ||t_k^{(1)}|| = 1$

Overall criterion to maximize

Multiblock redundancy analysis [Bougeard et al., 2011]



Key ideas

- Sum up each dataset with latent variables,
- Relate the explanatory latent variables with the dependent one,
- Seek all these latent variables by maximizing an overall criterion,
- Get a direct eigensolution.

Criterion to maximize

Max. $\sum_k cov^2(u^{(1)}, t_k^{(1)})$ with $u^{(1)} = Yv^{(1)}$, $t_k^{(1)} = X_k w_k^{(1)}$, $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$, $\sum_k a_k^{(1)^2} = 1$ and $||v^{(1)}|| = ||t_k^{(1)}|| = 1$

Direct eigensolution

Multiblock redundancy analysis [Bougeard et al., 2011]

First order solution

- $v^{(1)}$ is the eigenvector of $\sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$
- $t_k^{(1)} = P_{X_k} u^{(1)} / ||P_{X_k} u^{(1)}||$
- $t^{(1)} = \sum_k \frac{cov(u^{(1)}, t_k^{(1)})}{\sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})}} t_k^{(1)}$

with the projector $P_{X_k} = X_k (X_k' X_k)^{-1} X_k'$

Higher order solution

Residuals of the orthogonal projections of X_k onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)})$, ...

Interpretation

- The explanatory multiblock structure is taken into account,
- The partial explanatory components t_k are derived from the projection of the dependent component u onto each X_k space,
- The more the partial components u and t_k are linked, the more they build the global component t.

Interpretation

Improve the prediction ability (orthogonalized regression).

Direct eigensolution

Multiblock redundancy analysis [Bougeard et al., 2011]

First order solution

- $v^{(1)}$ is the eigenvector of $\sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$
- $t_k^{(1)} = P_{X_k} u^{(1)} / ||P_{X_k} u^{(1)}||$
- $t^{(1)} = \sum_k \frac{cov(u^{(1)}, t_k^{(1)})}{\sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})}} t_k^{(1)}$

with the projector $P_{X_k} = X_k (X_k' X_k)^{-1} X_k'$

Higher order solution

Residuals of the orthogonal projections of X_k onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)})$, ...

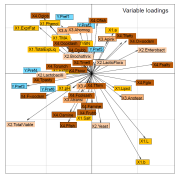
Interpretation

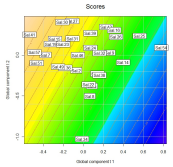
- The explanatory multiblock structure is taken into account,
- The partial explanatory components t_k are derived from the projection of the dependent component u onto each X_k space,
- The more the partial components u and t_k are linked, the more they build the global component t.

Interpretation

Improve the prediction ability (orthogonalized regression).

Main interpretation tools





[1] From a descriptive point of view : factorial graphical displays.

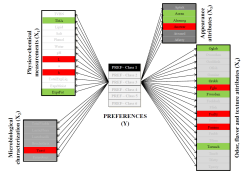
[2] Selection of the optimal model (cross-validation procedure).

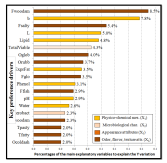
[3] Aims

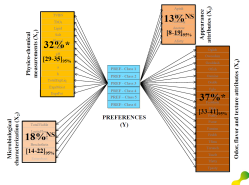
- Link all the explanatory variables with all the dependent ones,
- Sort the explanatory variables by order of priority,
- Sort the explanatory blocks by order of priority.

[3] Predictive interpretation tools

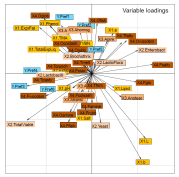
- $\beta_{p,q}^{(1\rightarrow h_{opt})} = \sum_{h=1}^{h_{opt}} w^{(h)} * c^{(h)\prime}$
- $VarImp_{p}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} \dfrac{a_k^{(h)2} w_{[p]}^{(h)2}}{\sum_{p=1}^{P} a_k^{(h)2} w_{[p]}^{(h)2}}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$
- $BlockImp_{k}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} a_k^{(h)2}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$

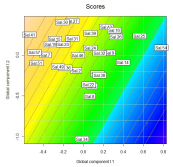






Main interpretation tools





[1] From a descriptive point of view : factorial graphical displays.

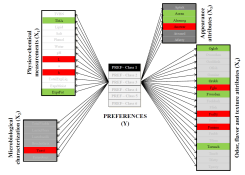
[2] Selection of the optimal model (cross-validation procedure).

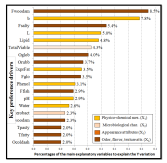
[3] Aims

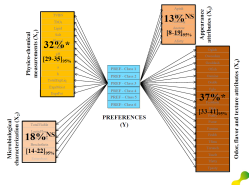
- Link all the explanatory variables with all the dependent ones,
- Sort the explanatory variables by order of priority,
- Sort the explanatory blocks by order of priority.

[3] Predictive interpretation tools

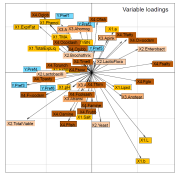
- $\beta_{p,q}^{(1\rightarrow h_{opt})} = \sum_{h=1}^{h_{opt}} w^{(h)} * c^{(h)\prime}$
- $VarImp_{p}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} \dfrac{a_k^{(h)2} \, w_{[p]}^{(h)2}}{\sum_{p=1}^{P} a_k^{(h)2} \, w_{[p]}^{(h)2}}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$
- $BlockImp_{k}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} a_k^{(h)2}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$

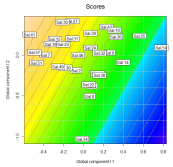






Main interpretation tools





[1] From a descriptive point of view : factorial graphical displays.

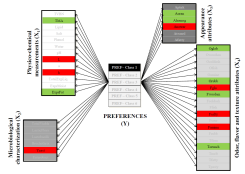
[2] Selection of the optimal model (cross-validation procedure).

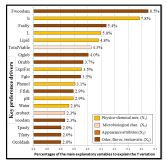
[3] Aims

- Link all the explanatory variables with all the dependent ones,
- Sort the explanatory variables by order of priority,
- Sort the explanatory blocks by order of priority.

[3] Predictive interpretation tools

- $\beta_{p,q}^{(1\rightarrow h_{opt})} = \sum_{h=1}^{h_{opt}} w^{(h)} * c^{(h)'}$
- $VarImp_{p}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} \dfrac{a_k^{(h)2}\, w_{[p]}^{(h)2}}{\sum_{p=1}^{P} a_k^{(h)2}\, w_{[p]}^{(h)2}}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$
- $BlockImp_{k}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} a_k^{(h)2}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$





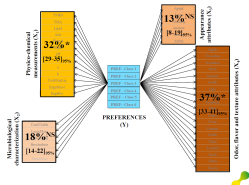
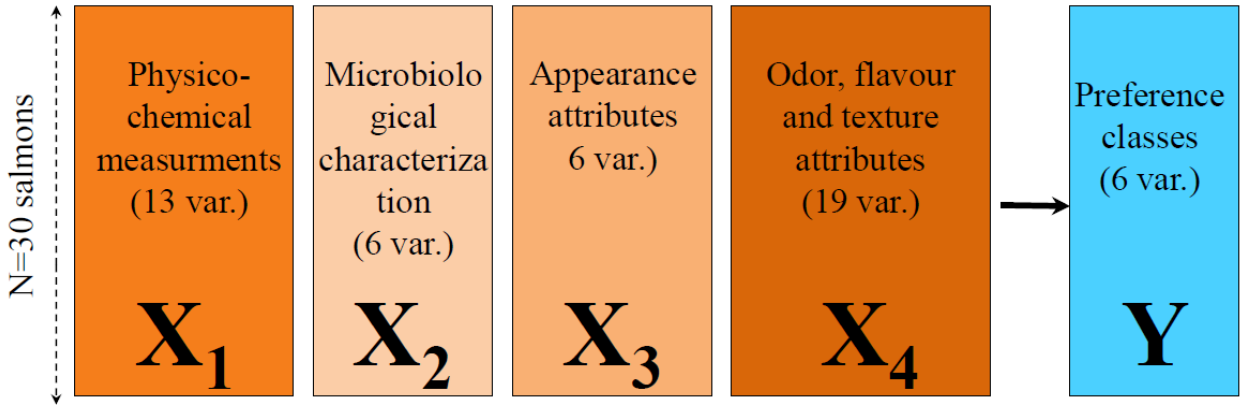


Table of contents

Multiblock external preference mapping of smoked salmons

Eurosalmon project [Cardinal et al., 2004]



Salmon data features

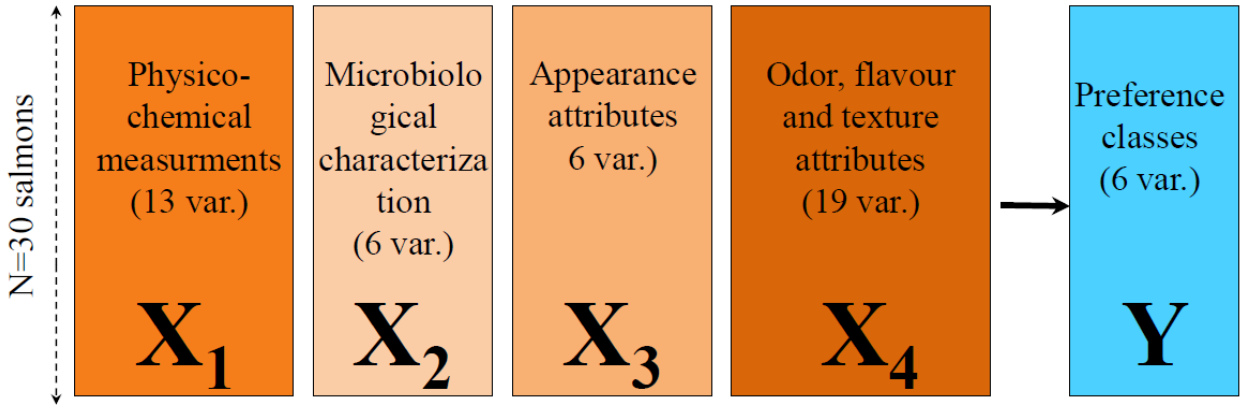
- **Y** : 6 preference classes from 1063 consumers [Semenou et al., 2007],
- **X** : 44 potential preference drivers organized into 4 blocks,

Aims

- **Descriptive** : explain the consumer preferences with the explanatory variables and blocks in relation with the tasted salmons,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Multiblock external preference mapping of smoked salmons

Eurosalmon project [Cardinal et al., 2004]



Salmon data features

- **Y** : 6 preference classes from 1063 consumers [Semenou et al., 2007],
- **X** : 44 potential preference drivers organized into 4 blocks,

Aims

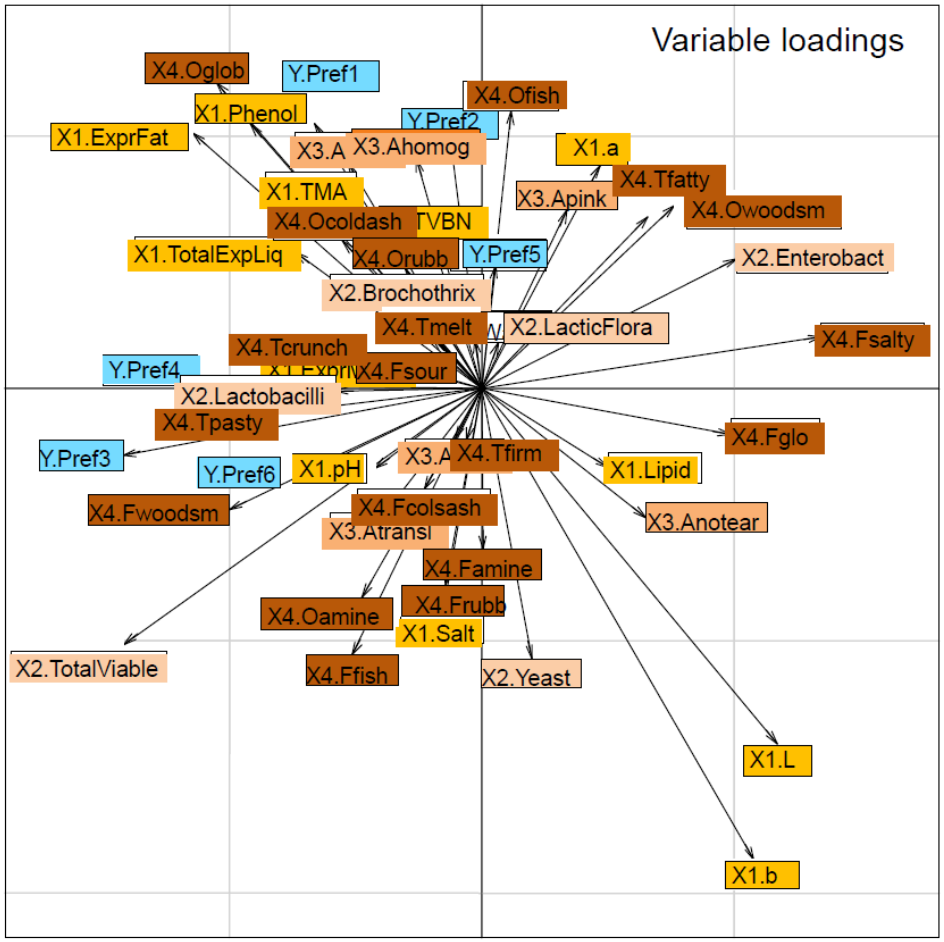
- **Descriptive** : explain the consumer preferences with the explanatory variables and blocks in relation with the tasted salmons,
- **Predictive** : assess the key drivers of preference at the variable and block levels.

Descriptive interpretation

Interpretation of the relationships between variables, blocks and products

Interpretation

- The preference classes 1 and 2, resp. 3, 4 and 6, are roughly comparable,
- The preference classes 1 and 2 like pink salmons with intense global odor.
- The tasted salmons can be placed in relation with the expected preference.

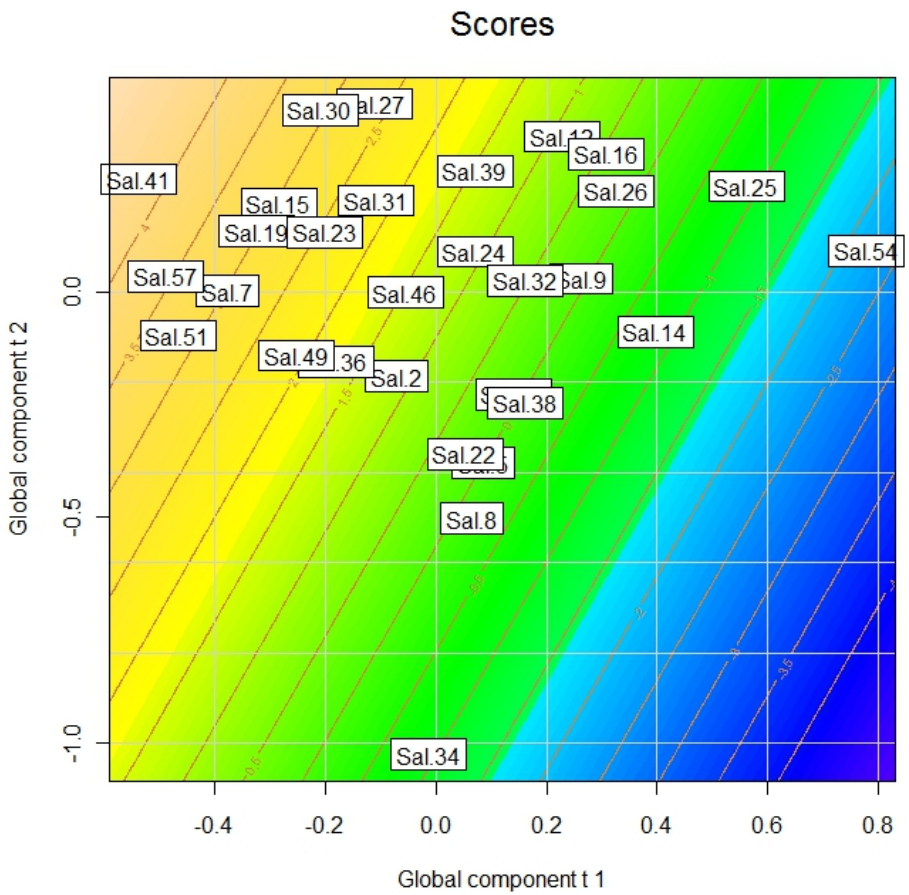


Descriptive interpretation

Interpretation of the relationships between variables, blocks and products

Interpretation

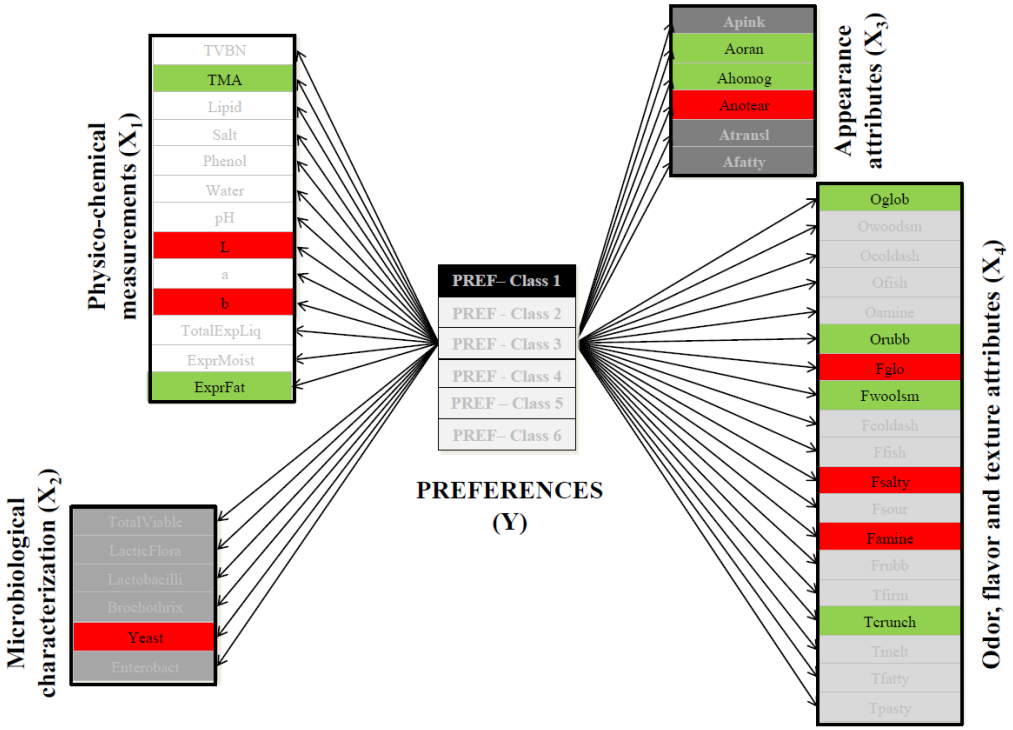
- The preference classes 1 and 2, resp. 3, 4 and 6, are roughly comparable,
- The preference classes 1 and 2 like pink salmons with intense global odor.
- The tasted salmons can be placed in relation with the expected preference.



↖ direction of the increase of the average expected preference.

Key drivers of preference at the variable level (1)

Regression coefficients and bootstraped tolerance interval. Optimal model with 4 components.



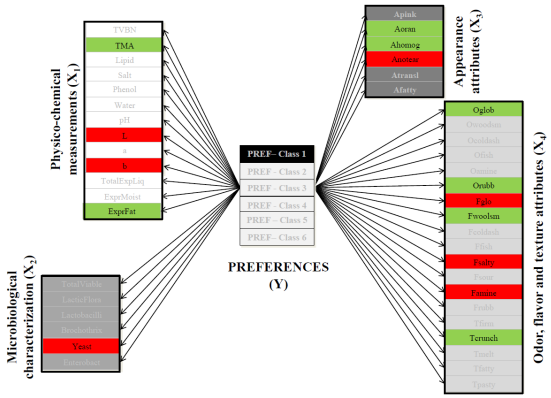
Interpretation for preference class 1 ($N = 121$ consumers)

- The consumers of class 1 are sensible to appearance (L-, b^*-, orange color+, color homogeneity+, slice tearing-)
- They also like some specific taste (salty-), flavors (Fglo-, Fwoodsm+, Famine-) and texture (Tcrunch+)

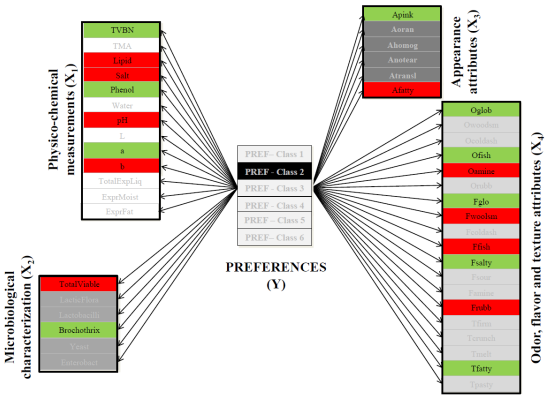
Positive significant key drivers of preference (GREEN), negative significant key drivers of preference (RED), non significant variables (WHITE-GREY)

Key drivers of preference at the variable level (2)

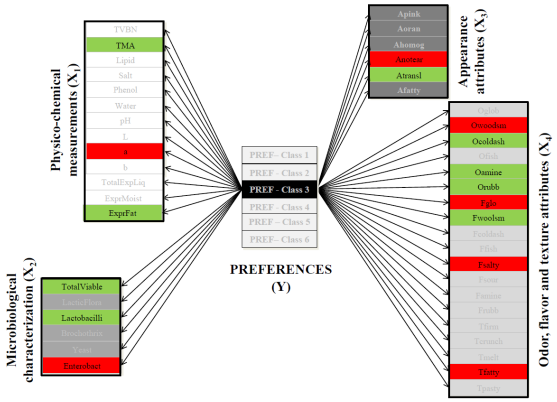
Regression coefficients and bootstraped tolerance interval. Optimal model with 4 components.



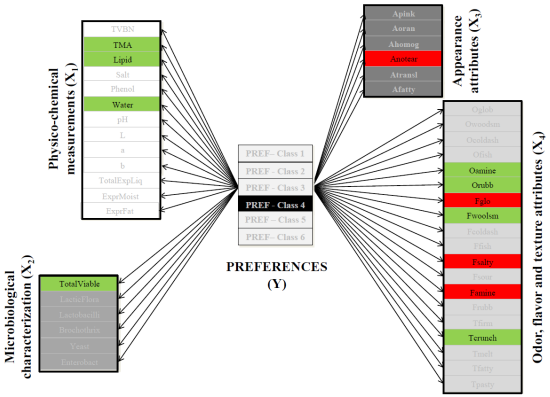
Pref 1 (N=121)



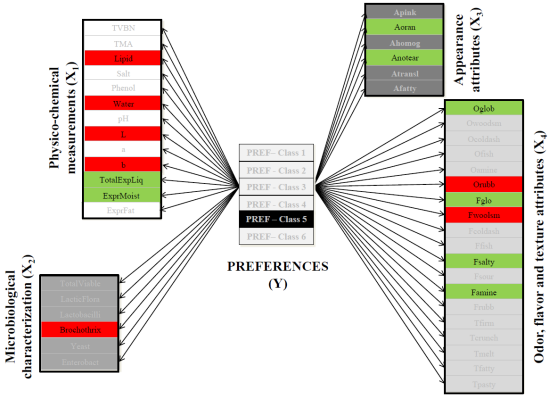
Pref 2 (N=74)



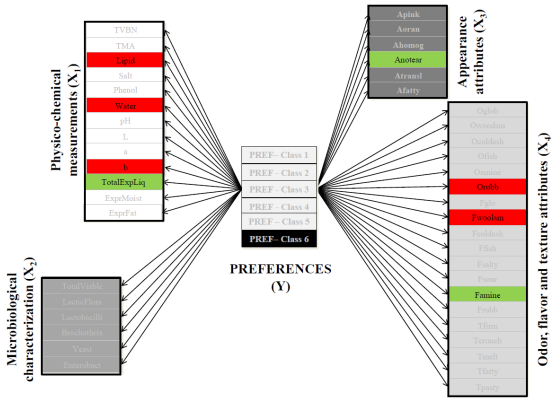
Pref 3 (N=349)



Pref 4 (N=78)



Pref 5 (N=404)

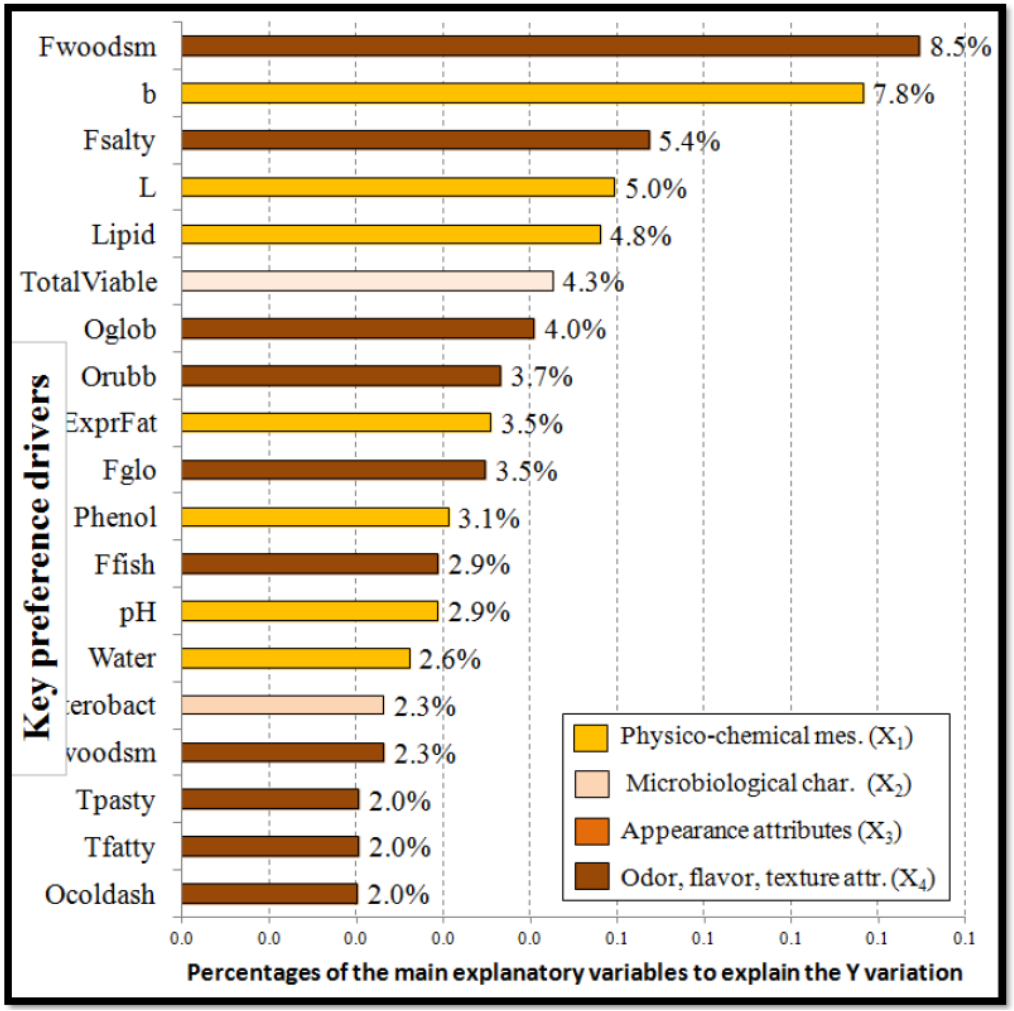


Pref 6 (N=37)

Results are difficult to sum up → Difficulties to get overall interpretation of key drivers.

Key drivers of preference at the variable level (3)

Variable Importance expressed as percentage and bootstraped tolerance interval. Optimal model with 4 components.



Interpretation for overall preference

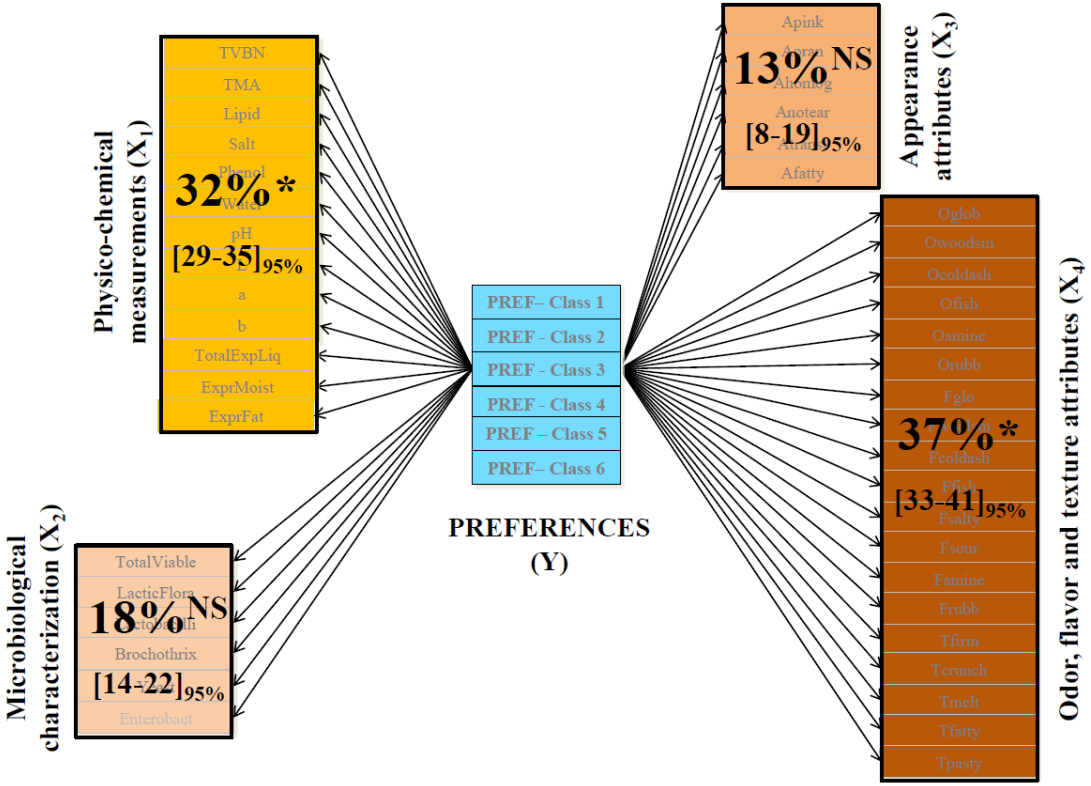
The model explains 82% of the variation in Y which is significantly explained by :

- The wood smoked flavor ("++" for classes 1, 3 and **4**, "–" for classes 2, **5** and 6),
- The hue parameter b^* (yellow) ("–" for classes 1, 2, **5** and 6),

  $\rightarrow$ Both these variables explain 14.3% of the overall preference.

Key drivers of preference at the block level

Block Importance expressed as percentage and bootstraped tolerance interval. Optimal model with 4 components.



Interpretation for overall preference

The model explains 82% of the variation in Y, which is significantly explained by :

- The odor, flavor and texture attributes (37%),
- The physico-chemical measurements (32%),

Table of contents

Conclusion & perspectives

Conclusion

- Multiblock Redundancy Analysis handles the specificity of complex data for external preference mapping,
- Increase the amount of extracted information from the data (both standard and specific results),
- Sort by order of priority key preference drivers at the variable and block level,
- Freely available method(s) and interpretation tools : `ktab+1` package integrated in the `ade4` software (http ://pbil.univ-lyon1.fr/ade4/).

Perspectives

- Direct extension to the explanation of several dependent blocks $(\mathbf{Y}_1, \ldots, \mathbf{Y}_{K'})$ or to qualitative (dummy) variables,
- Handle information on products (experimental design to control some topical parameters) / PhD thesis (A. Eslami, 2010-2013)

Conclusion & perspectives

Conclusion

- Multiblock Redundancy Analysis handles the specificity of complex data for external preference mapping,
- Increase the amount of extracted information from the data (both standard and specific results),
- Sort by order of priority key preference drivers at the variable and block level,
- Freely available method(s) and interpretation tools : `ktab+1` package integrated in the `ade4` software (http ://pbil.univ-lyon1.fr/ade4/).

Perspectives

- Direct extension to the explanation of several dependent blocks $(\mathbf{Y_1}, \ldots, \mathbf{Y_{K'}})$ or to qualitative (dummy) variables,
- Handle information on products (experimental design to control some topical parameters) / PhD thesis (A. Eslami, 2010-2013)

Multiblock modeling for complex preference study

Application to European preferences for smoked salmon

S. Bougeard[1] & M. Cardinal[2]

[1] *French Agency for food, environmental and occupational health & safety (Anses), Department of epidemiology*
[2] *French research institute for exploration of the sea (IFREMER), Department of biotechnologies and sea resources*