Master 1 Econométrie, Statistique,
parcours Économétrie Appliquée

Modélisation des variables latentes

Mars 2024

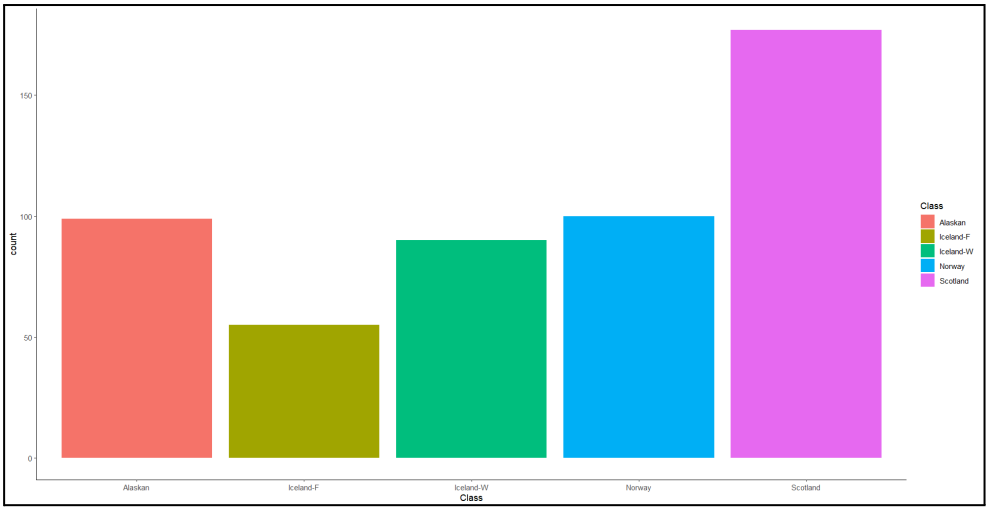
DAHMANI Amel
JAMIN Mathilde
ONNO Lilou

Résumé :

L'objectif de notre analyse est d'examiner l'authenticité du saumon en tenant compte de l'origine géographique et du mode d'élevage. Pour ce faire, nous avons réalisé une analyse descriptive pour visualiser notre jeu de données. Ensuite, nous avons mis en place un modèle PLS-DA avec 5 composantes. Cette approche nous a permis d'évaluer la qualité prédictive du modèle et d'identifier les éléments chimiques les plus pertinents en termes de discrimination. En conclusion, nous avons développé un modèle performant avec 91,26% de qualité prédictive avec plusieurs variables discriminantes telles que Fe, Ce, Cu, Cd, Ta.

1) Analyse Descriptive

Le jeu de données comporte 521 observations et 21 variables, correspondant à différentes composantes chimiques de 4 pays : Alaska, Ecosse, Norvège et Islande. L'objectif de cette étude est d'établir un modèle pour établir l'authenticité de saumon en fonction de l'origine géographique et du mode d'élevage.



Histogramme 1 : Distributions du modes de productions des saumons selon les pays.

Nous observons que l'Islande est présente deux fois, puisqu'il pratique du saumon d'élevage, ainsi que du saumon sauvage. Par ailleurs, tous les saumons de l'étude en provenance de Norvège et d'Écosse sont issus de systèmes d'élevage alors qu'ils sont issus de pêche en Alaska.

L'Ecosse présente la part la plus importante en ce qui concerne la production de saumon d'élevage avec une valeur de 177 tandis que l'Islande présente la plus petite classe pour la production de saumon d'élevage avec une valeur qui est autour de 55.

Nous avons également une distribution similaire pour l'Alaska et la Norvège, cependant l'un est issu du système sauvage tandis que l'autre est issu du système d'élevage.

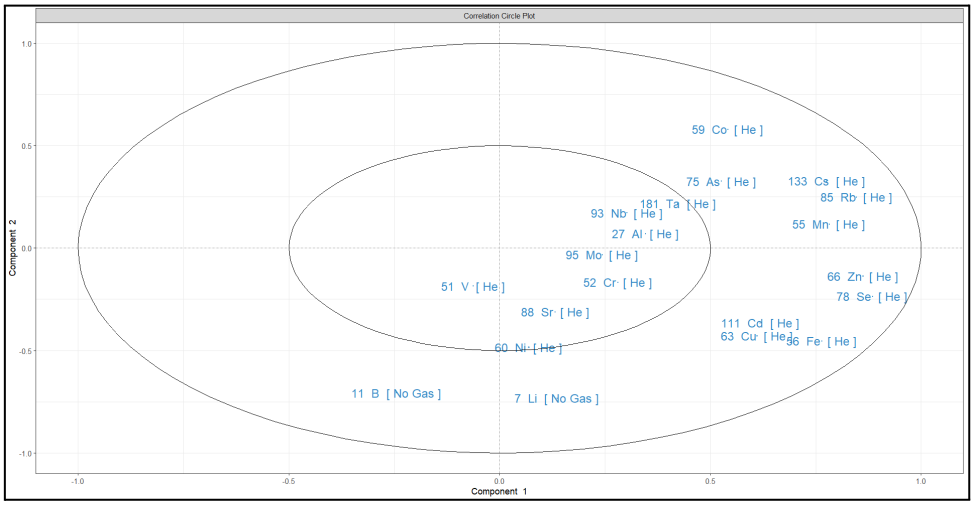
L'analyse de l'histogramme relève des tendances distinctes, en termes de distribution ou de répartition des données. La suite de notre analyse se portera avec des valeurs standardisées

2) Ajustement d'un modèle PLS - DA

Premièrement, nous avons généré un jeu d'apprentissage qui correspond à la partie de l'ensemble de données initiales que l'on a utilisée pour entraîner ou construire le modèle. Nous avons par ailleurs, alloué 80% des données à ce jeu là.

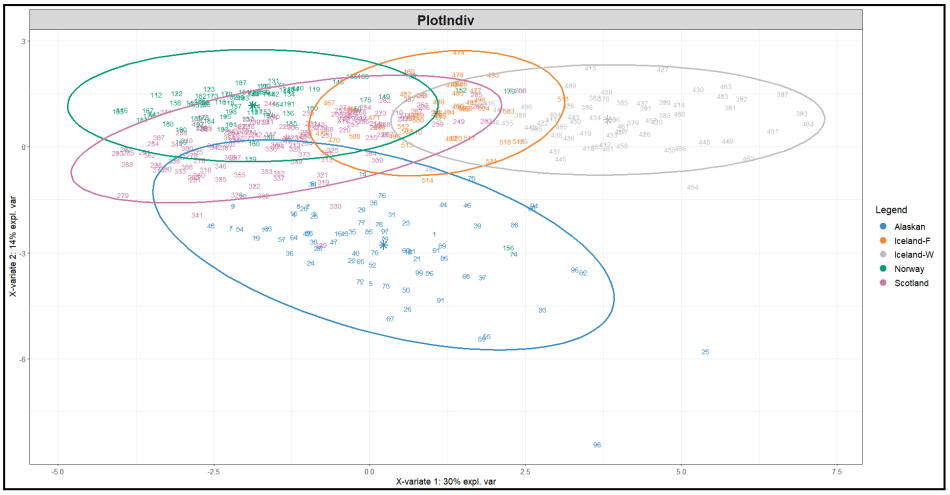
Deuxièmement, nous avons généré un jeu test qui lui correspond à la partie distincte faisant référence à l'ensemble des données initiales n'ayant pas été utilisées lors de la phase d'apprentissage du modèle.

Une fois nos jeux créés, nous avons pu ajuster notre modèle avec 10 composantes, ce qui permet d'avoir une visualisation globale de notre étude.



Graphique 1 : Projection des variables sur les composantes 1 et 2

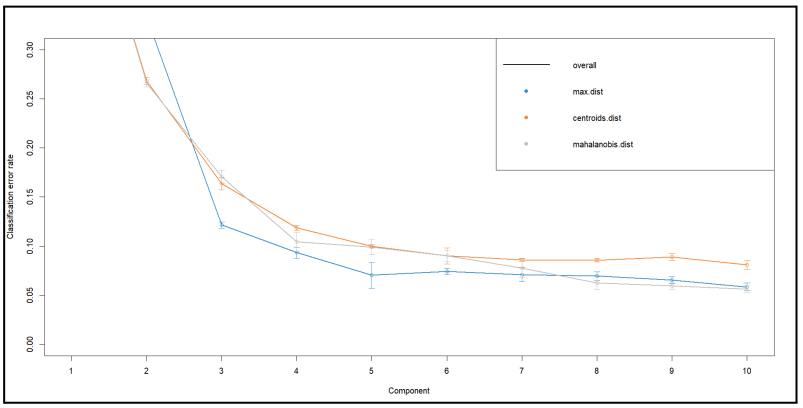
La majeure partie de nos variables se trouve dans la partie droite de la projection, seuls deux éléments se trouvent à gauche du graphique (11 et 51), et proche du centre (51). Nous pouvons aussi constater que certaines variables ont une influence positive sur l'axe 1, tandis que d'autres ont une influence négative sur l'axe 2.



Graphique 2 : Projection des individus sous forme d'ellipse

Le graphique ci-dessus, nous permet de visualiser les individus en fonction de 5 classes, qui représente leurs pays respectifs. Nous observons que certaines classes se rapprochent, notamment la Norvège, L'Islande-F et l'Ecosse, ce qui peut être expliqué par leurs systèmes d'élevage commun.

Pour la classe Alaska, nous pouvons voir que les individus sont assez éloignés, cela peut s'expliquer par le fait que nous avons essentiellement des saumons de pêche et donc des éléments chimiques diversifiés.

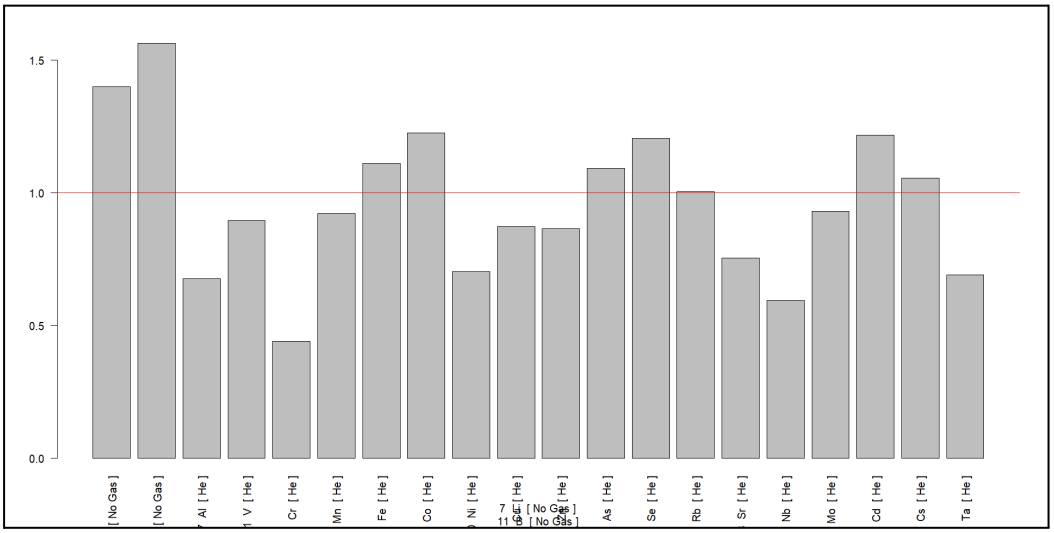


Graphique 3 : Performance globale en fonction du nombres de composantes

L'analyse du graphique nous a aidé à déterminer le nombre optimal de composantes. À première vue, il semble que la distance de Mahalanobis ne soit pas la plus optimale. Cependant, étant donné qu'elle propose 8 composantes, conformément à la règle de parcimonie, il est approprié de prendre en considération les 5 composantes en se basant sur les deux autres critères.

Nous avons fait ce choix en prenant en compte le point le plus bas de chacun des critères.

3) Bilan de la qualité prédictive

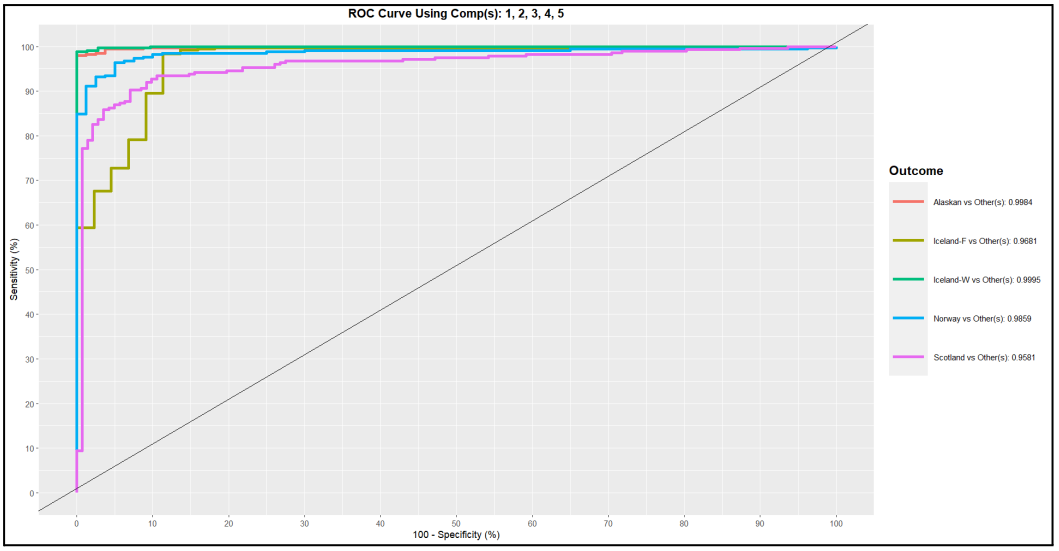


Graphique 4 : Identification des variables dites importantes

Nous pouvons grâce au graphique déterminer les variables qui impacteraient le plus notre modèle, nous avons ainsi : 7 Li ; 11 B ; 56 Fe ; 59 Co ; 75 As ; 78 Se ; 111 Cd ; 133 Cs.
Ces différents symboles chimiques correspondent respectivement à :

- 7 Li : Lithium (Li)
- 11 B : Bore (B)
- 56 Fe : Fer (Fe)
- 59 Co : Cobalt (Co)
- 75 As : Arsenic (As)
- 78 Se : Sélénium (Se)
- 111 Cd : Cadmium (Cd)
- 133 Cs : Césium (Cs)

Nous pouvons mettre en relation ces résultats, avec ceux observés lors du graphique 1, en effet toutes ces variables se retrouvent proche des axes, ce qui renforce notre analyse et ainsi leurs degrés d'importance dans l'étude.



Graphique 5 : Evaluation de performance discriminative du modèle

À la lumière du graphique présenté ci-dessus, nous observons que la performance discriminative de notre modèle PLS-DA est satisfaisante, car chacune des courbes est relativement proche du coin situé à gauche. En particulier, les performances d'Alaska et d'Islande-W semblent être les plus élevées.

4) Interprétation du modèle

Nous avons calculé le taux de classification en utilisant la diagonale de la matrice de confusion, celui-ci étant de 91,26%, cela permet de conclure que notre modèle prédit correctement 91,26% des observations. Ce résultat reste élevé, et donc implique qu'il est performant.

De plus, les variables qui discriminent le plus notre étude pour Df1 sont le fer (Fe), le sélénium (Se), le cuivre (Cu), le cadmium (Cd) et le tantale (Ta). Ces variables nous signalent quant à leur importance dans la construction dans la dimension 1.