Rapport Méthodologique

Manuel RTBfoods - Partie 3 - Tutoriel

Analyses Statistiques (ACP et régression multiple) pour Visualiser les Données de l'Analyse Sensorielle et les Relier aux Données Instrumentales

Biophysical Characterization of Quality Traits, WP2

Montpellier, France, Juillet 2021

Christophe BUGAUD, Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, France

Isabelle MARAVAL, CIRAD, Montpellier, France

Karima MEGHAR, CIRAD, Montpellier, France

Ce rapport a été rédigé dans le cadre du projet RTBfoods.

Peut être cité comme suit :

Christophe BUGAUD, Isabelle MARAVAL, Karima MEGHAR, (2022). Manuel RTBfoods - Partie 3 - *Tutoriel : Analyses Statistiques (ACP et régression multiples) pour Visualiser les Données de l'Analyse Sensorielle et les Relier aux Données Instrumentales. Biophysical Characterization of Quality Traits, WP2.* Montpellier, France : Rapport Méthodologique RTBfoods, 26 p. https://doi.org/10.18167/agritrop/00738

Éthique : Les activités, qui ont conduit à la production de ce document, ont été évaluées et approuvées par le Comité d'éthique du CIRAD (procédure d'auto-évaluation éthique H2020). Le cas échéant, les échantillons ont été préparés selon les bonnes pratiques d'hygiène et de fabrication. Lorsque des participants externes ont été impliqués dans une activité, ils ont été préalablement informés de l'objectif de l'activité et il leur a été expliqué que leur participation était entièrement volontaire, qu'ils pouvaient arrêter l'entretien à tout moment et que leurs réponses seraient anonymes et stockées de manière sécurisée par l'équipe de recherche à des fins de recherche. Un consentement écrit (signature) a été systématiquement demandé aux panélistes sensoriels et aux consommateurs participant aux activités.

Remerciements : Ce travail a été soutenu par le projet RTBfoods, https://rtbfoods.cirad.fr, à travers une subvention OPP1178942 : Breeding RTB products for end user preferences (RTBfoods), attribuée au Centre français de coopération internationale en recherche agronomique pour le développement (CIRAD), Montpellier, France, par la Fondation Bill & Melinda Gates (BMGF).

Image cover page © LAJOUS P. for RTBfoods.

Ce document a été revu par :	
Isabelle MARAVAL (CIRAD) Karima MEGHAR (CIRAD)	15/06/2021 04/06/2021
Validation finale par :	
Christophe BUGAUD (CIRAD)	02/07/2021

SOMMAIRE

Table of Contents

1 Objectif 6

2 Analyse en composantes principales 6

2.1 Définitions - principes 6

2.2 Mise en oeuvre d'une analyse en composantes principales avec XLSTAT 7

2.3 Interprétation des résultats 10

2.3.1 Comment interpréter les valeurs propres de l'ACP ? 10

2.3.2 Comment interpréter le cercle des corrélations des variables de l'ACP? 10

2.3.3 Comment interpréter les résultats associés au graphique des observations ? 12

3 Régressions linéaires 13

3.1 Definitions - principles 14

3.2 Mise en œuvre d'une régression linéaire avec XLSTAT 14

3.3 Interprétation des résultats de la Régression Linéaire 17

3.3.1 Statistiques descriptives 17

3.3.2 Comment interpréter les statistiques de multicolinéarité ? 17

3.4 Questions diverses 23

3.4.1 Peut-on se passer d'une validation ? 23

3.4.2 Combien d'observations sont-elles nécessaires pour établir les prédictions ? 24

RÉSUMÉ

Après l'évaluation sensorielle par un panel entrainé et validé, et biophysique par des mesures instrumentales des différents produits, il est alors possible de mener des traitements statistiques pour interpréter les résultats. L'objectif de ce tutoriel, utilisant le logiciel XLSTAT, est de réaliser et d'interpréter 2 types de traitements statistiques : (1) l'analyse en composante principale (ACP) qui permet de visualiser rapidement les corrélations entre les attributs sensoriels et de visualiser les produits entre eux, et (2) la régression linéaire qui permet de prédire les attributs sensoriels à partir des paramètres biophysiques (texturaux, biochimiques). Au préalable, les performances du panel ont été contrôlées et les données sensorielles ont été préparées pour l'analyse statistique (voir Manuel RTBfoods - Partie 2 - Tutoriel contrôler les performances du panel et préparer les données en analyses sensorielles avant les traitements statistiques, DOI: 10.18167/agritrop/00583). Dans tous les cas, le tutoriel s'appuie sur un exemple présenté dans un fichier Excel permettant de franchir toutes les étapes et qui a fait l'objet d'une publication. L'ACP a été centrée sur les données sensorielles pour visualiser les tendances majeures et la diversité sensorielle entre groupes de produits et entre produits. Elle permet aussi de mesurer les écarts entre produits répétés qui symbolisent la performance du panel (si les produits sont bien identiques). La régression linéaire multiple a été utilisée pour prédire les attributs sensoriels à partir de paramètres biophysiques. Pour cela, le jeu de données a été scindé en 2 jeux de données : un jeu de calibration représentant ¾ des données et un jeu de validation. Trois indicateurs de prédiction ont été calculés permettant d'évaluer à la fois la précision et la robustesse de la prédiction : le coefficient de détermination (R^2) et l'écart moyen entre les valeurs observées et prédites (RMSEC) sur le jeu de calibration et l'écart moyen entre les valeurs observées et prédites sur le jeu de validation (RMSEV). La pertinence de la validation et le nombre minimum d'observations nécessaire à la construction de modèles prédictifs ont été discutés.

Mots Clés : profil sensoriel – mesure instrumentale - analyse en composante principale – régression linéaire multiple – calibration - validation – RMSE

1 OBJECTIF

Dans le précédent tutoriel (RTBfoods_F.2.4B_Tutoriel Contrôle de la Performance & Nettoyage des Données Sensorielles Avant Traitement Statistique.pdf), nous avions présenté la méthodologie pour contrôler les performances du panel (répétabilité et homogénéité avec le panel) à la fin des analyses et pour préparer les données pour des traitements statistiques. Ce tutoriel avait permis en particulier d'éliminer des panélistes non performants ou des données non acceptables. Finalement, vous disposiez d'un tableau final qui affichait les valeurs moyennes obtenues par le panel pour chaque produit, chaque attribut sur une échelle de 0 à 10. Ce tableau peut être complété par des données instrumentales (paramètres texturaux, biochimiques) obtenues sur les mêmes échantillons.

En s'appuyant sur notre propre expérience, nous vous proposons un tutoriel utilisant le logiciel XLSTAT pour réaliser et interpréter 2 types de traitements statistiques importants pour le WP2 en analyse sensorielle :

1- l'analyse en composante principale qui permet de visualiser rapidement les corrélations entre les attributs sensoriels et de visualiser les produits entre eux,
2- et la régression multiple qui permet de prédire les attributs sensoriels à partir des paramètres instrumentaux (texturaux, biochimiques).

Rappelons que ce tutoriel n'a pas pour vocation d'expliquer les bases statistiques de l'ACP ou de la régression multiple, aussi le lecteur est invité à consulter des ouvrages et publications dédiées à ces analyses statistiques pour en savoir plus. Par ailleurs, nous ne sommes pas statisticiens, mais des utilisateurs qui avons notre propre vision de l'usage et de l'interprétation de ces outils statistiques.

Jeu de données utilisé pour les traitements statistiques

Pour illustrer les traitements statistiques, nous avons utilisé un jeu de données extrait d'une étude menée sur la banane plantain bouillie à différents stades de murissage (Exemple ACP + regression.xls). Les résultats et leurs interprétations ont été valorisés dans un article scientifique (doi:10.1111/ijfs.14765). Dans notre exemple, les données sont les valeurs moyennes de 35 observations (produits cuits issus de 13 cultivars et 3 stades de murissage) pour les variables suivantes :

- Variables sensorielles : 6 attributs (firmness, chewiness, stickiness, mealiness, moist and sweetness) ont été évalués sur une échelle de 0 à 10 + 2 attributs (sourness, astringent) ont été évalués sur une échelle binomiale (yes/no). Pour ces 2 attributs, une fréquence de "yes" a été calculée et réduite sur une échelle de 0 à 10 pour permettre une comparaison avec les autres attributs évalués sur une échelle d'intensité de 0 à 10.
- Variables instrumentales : 7 paramètres texturaux mesurés par un test de double compression (paramètre_TPA) et de pénétrométrie (puncture force) et 3 paramètres biochimiques (teneur en matière sèche, extrait sec soluble et acidité titrable)

2 ANALYSE EN COMPOSANTES PRINCIPALES

2.1 Définitions - principes

L'analyse en composantes principales (ACP) est l'une des analyses de données multivariées la plus fréquemment utilisée. Elle étudie des ensembles de données multidimensionnelles avec des variables quantitatives.

L'analyse en composantes principales est une méthode de projection très utile pour analyser des données numériques structurées en n observations / p variables. L'ACP projette les observations d'un espace à P dimensions avec p variables vers un espace à K dimensions (où K<P) de façon à conserver le maximum d'information (l'information est mesurée ici par la variance totale des nuages de points) des dimensions initiales. Il permet de :

- Visualiser et analyser rapidement les corrélations entre les p variables,

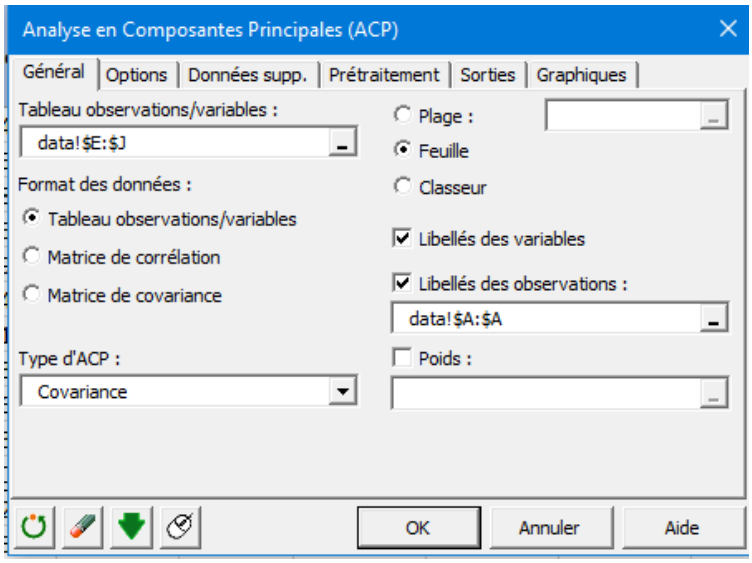
- Visualiser et analyser les n observations (initialement décrites par les p variables) dans une carte à 2 ou 3 dimensions, afin d'identifier des groupes d'observations uniformes ou atypiques.

Les dimensions de l'ACP sont aussi appelés axes ou facteurs.

Dans notre exemple, n = 35, p = 6 (seulement les attributs sensoriels évalués sur une échelle de 0 à 10) ou p = 8 (tous les attributs sensoriels).

2.2 Mise en œuvre d'une analyse en composantes principales avec XLSTAT

- Ouvrir le fichier « Exemple ACP + regression.xls».
- Après avoir activé XLSTAT, sélectionner la commande **XLSTAT / Analyse des données / Analyse en Composantes Principales**.
- Dans la fenêtre "Général", sélectionner les 6 attributs sensoriels (colonne E à J) de la feuille "data" dans "Tableau Observations/variables".
- Sélectionner la colonne A de la feuille "data" dans "Libellé des Observations".
- Activer l'option "feuille" pour visualiser les résultats dans une nouvelle feuille et "Libellé des variables".
- Dans « type d'ACP », sélectionner "Covariance".

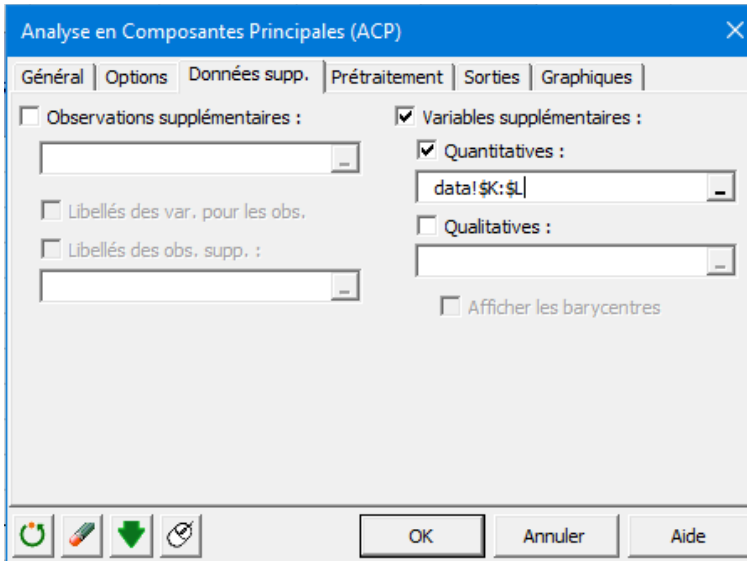


Pourquoi choisir l'option 'Covariance' ?

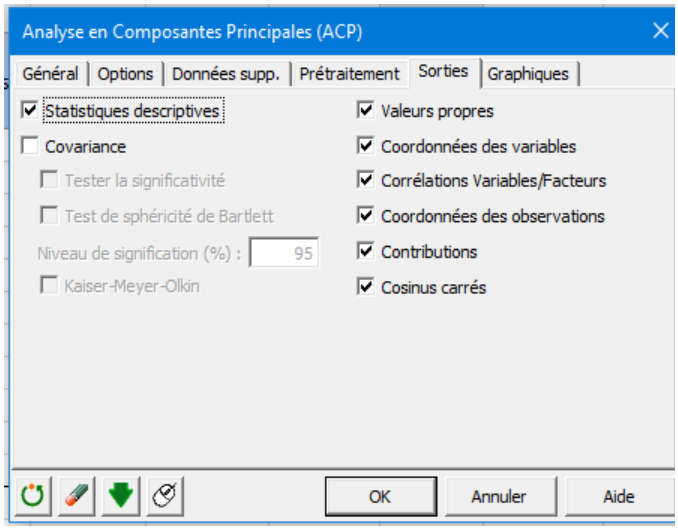
Lorsque les variables sont supposées être sur des échelles identiques (ce qui est le cas ici), ou lorsque l'on veut que la variance des variables influe sur la construction des facteurs, on utilise la covariance. Avec cette option, un attribut sensoriel utilisant pleinement l'échelle d'intensité (0 – 10) influera la construction des facteurs beaucoup plus qu'un attribut cantonné à une partie de l'échelle (par exemple entre 0 et 3).

On utilise l'option « corrélation (Pearson) » dans le cas où les variables ne sont pas toutes à la même échelle (notamment avec les données instrumentales), permettant ainsi de supprimer les effets d'échelle : ainsi une variable variant entre 0 et 1000 ne pèse pas plus dans la projection qu'une variable variant entre 0 et 1. Dans ce cas, les données ont été préalablement centrées et réduites (par Xlstat) pour chaque variable.

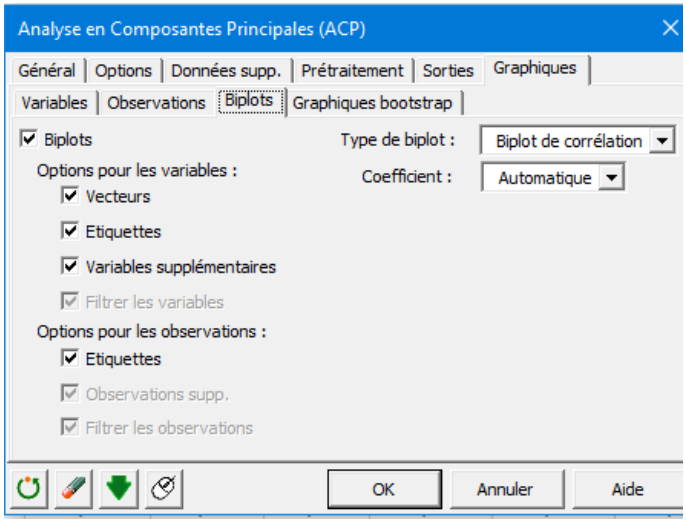
- Dans la fenêtre "Options", "n" est choisi par défaut. L'option "Rotation" peut être choisi si vous voulez appliquer une rotation à la matrice des coordonnées factorielles. Cela pourrait permettre dans certains cas exceptionnels de mieux positionner les variables sur les axes 1 et 2 et améliorer l'interprétation des résultats. Nous ne l'avons personnellement jamais utilisé.
- Dans la fenêtre "Données supplémentaires", XLSTAT permet de rajouter des observations ou des variables supplémentaires. Ces observations ou variables ne sont pas pris en compte pour le calcul des facteurs (observations ou variables passives, par opposition à observations ou variables actives). Dans notre exemple, les deux attributs notés sur une échelle binomiale (y/n), qui étaient moins importants que les 6 attributs notés sur une échelle de 0 à 10, ont été sélectionnés comme variables supplémentaires (colonnes K et L).



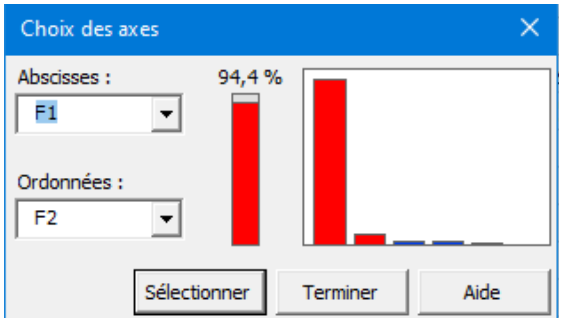
- Dans la fenêtre "Prétraitement", pour les données manquantes, sélectionner "supprimer les observations".
- Dans la fenêtre "Sortie", sélectionner "statistiques descriptives", et cocher toutes les options à droites (valeurs propres, etc…).



- Configurer les graphiques dans la fenêtre “Graphiques” :
  - Dans la fenêtre “Variable”, cocher “graphique de corrélation” et “vecteurs”,
  - Dans la fenêtre “Observations”, cocher “Graphique des observations” and “étiquettes”,
  - Dans la fenêtre “Biplots”, tout cocher et cocher « Biplot de corrélation » dans « type de biplot ».



- Les calculs commencent une fois que vous avez clique sur **OK**. Il vous est demandé de sélectionner les facteurs que vous souhaitez représenter. Par défaut, Xlstat propose les 2 premiers facteurs. Ici ces 2 premiers facteurs couvrent 94,4% de la variabilité, il ne sera alors pas nécessaire de regarder les facteurs suivants. Cliquer sur « Terminer » et les calculs se terminent.



2.3 Interprétation des résultats

2.3.1 Comment interpréter les valeurs propres de l'ACP ?

Le tableau et le graphique correspondant sont liés à un objet mathématique, les valeurs propres, qui reflètent la qualité de la projection des données. Chaque valeur propre correspond à un facteur. Les valeurs propres et les facteurs correspondants sont classés par ordre décroissant en termes de variabilité qu'ils représentent (convertie en %).

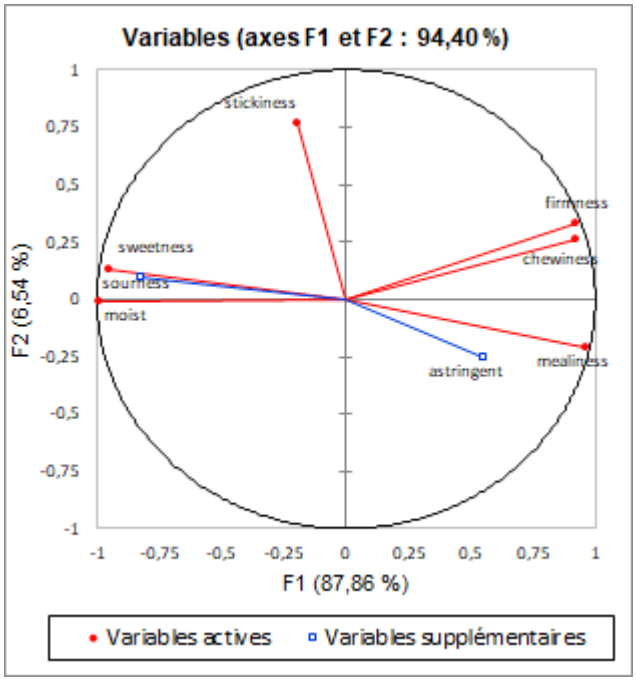
Dans notre exemple, nous pouvons voir que la première valeur propre représente 87,8 % de la variabilité totale. Cela signifie que si nous représentons les données sur un seul axe, nous serons toujours en mesure de voir 87,6 % de la variabilité totale des données. Avec deux axes, nous verrons 94,4 % de la variabilité totale, ce qui est très bien. Moins il y a de variables et plus il y a des corrélations entre variables, plus élevée est la représentation des deux premiers axes. Dans notre exemple, il n'est pas nécessaire de prendre en compte le 3ème axe qui ne représente que 2.5% de la variabilité totale. Mais il se peut que vous soyez amenés à le faire, notamment quand celui-ci représente plus de 10% (avis personnel).

Valeurs propres

	F1	F2	F3
Valeurs propres	16,658	1,241	0,473
Variabilité (%)	87,857	6,543	2,496
% cumulé %	87,857	94,400	96,895

2.3.2 Comment interpréter le cercle des corrélations des variables de l'ACP ?

La première figure est appelée **cercle de corrélation** ou tableau des variables. Le cercle de corrélation est utile pour interpréter la signification des axes et le graphique des observations.



Il montre une projection des variables actives sur les axes majeurs (F1 et F2). Lorsque deux variables sont proches du cercle, alors :

- si elles sont proches l'une de l'autre, elles sont significativement corrélées positivement (R proche de 1) : c'est le cas entre firmness et chewiness ou entre sweetness et moist.

- s'ils sont orthogonaux, ils ne sont pas corrélés (R proche de 0) : c'est le cas entre stickiness et firmness.

- si elles sont à l'opposé du centre, elles sont significativement corrélées négativement (R proche de -1) : c'est le cas entre firmness / chewiness / mealiness et moist / sweetness.

Les variables supplémentaires peuvent également être affichées. Curieusement, sourness est proche de sweetness, indiquant une possible corrélation positive entre ces 2 attributs.

Lorsque les variables sont proches du cercle, cela signifie que ces variables ont eu la plus grande influence dans la construction des axes PCA correspondants. A l'opposé, lorsque les variables sont proches du centre, cela signifie qu'elles n'ont pas contribué à la construction de ces axes majeurs, il faut éventuellement regarder pour ces variables sur les axes mineurs (> F2).

Pour confirmer qu'une variable est bien liée à un axe, il faut regarder le tableau des cosinus carrés des variables : plus le cosinus carré est grand, plus le lien avec l'axe correspondant est important. Firmness, chewiness, mealiness, moist et sweetness ont fortement contribué à l'axe 1 et stickiness à l'axe 2. Aucun attribut n'a contribué à l'axe 3.

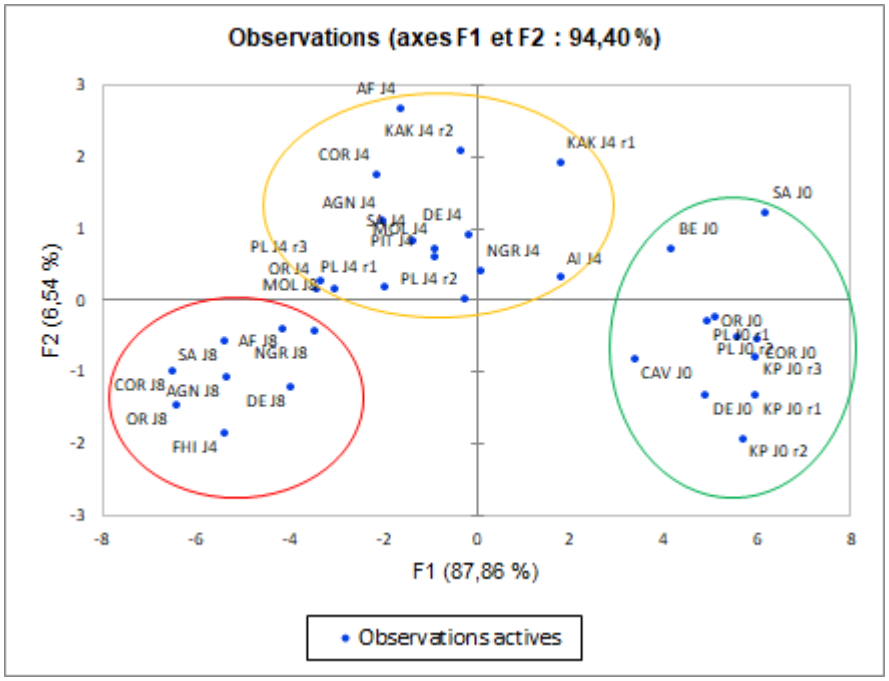
Cosinus carrés des variables :

	F1	F2	F3
firmness	**0,850**	0,111	0,022
chewiness	**0,848**	0,067	0,003
stickiness	0,040	**0,587**	0,263
mealiness	**0,925**	0,045	0,001
moist	**0,983**	0,000	0,000
sweetness	**0,909**	0,016	0,052
sourness	**0,680**	0,010	0,000
astringent	**0,302**	0,065	0,071

Les cosinus carrés reflètent la qualité de la représentation d'une variable sur un axe d'ACP. L'analyse des cosinus carrés est utilisée pour éviter les erreurs d'interprétation dues aux effets de projection. Si le cosinus carré d'une variable associée à un axe est faible, la position de la variable sur cet axe ne doit pas être interprétée.

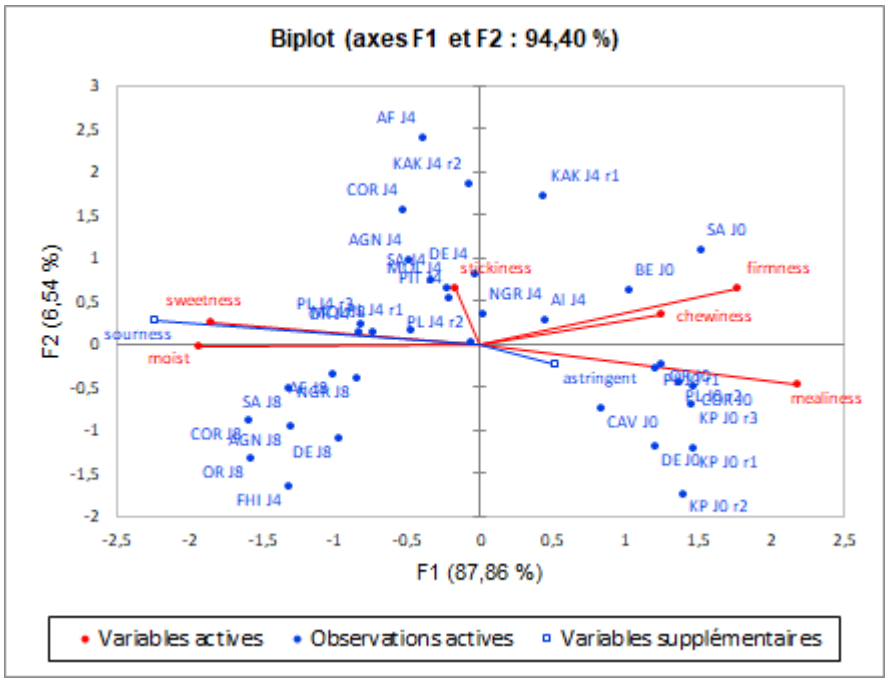
2.3.3 Comment interpréter les résultats associés au graphique des observations ?

Le **graphique des observations** représente les observations dans l'espace ACP. Il permet de regarder les observations sur une carte à deux dimensions et d'identifier des tendances.



L'axe 1 a clairement permis de séparer les variétés en fonction de leur stade de maturation (J0 - J4 - J8). Plus les plantains sont mûrs, moins les produits bouillis associés sont fermes, farineux mais plus ils sont sucrés, humides et aptes à être mastiqué. Les bananes plantains bouillies ont été perçues comme étant les plus collantes au stade intermédiaire de maturation (J4). La variabilité entre les trois stades de maturation était supérieure à la variabilité inter-cultivar. Nous pouvons voir que les répétitions des échantillons étaient proches (pour KP J0, PL J0, PL J4, et KAK J4), ce qui suggère des produits similaires et de bonnes performances du panel en termes de répétabilité. Certaines variétés comme FHI (hybride de plantain) au stade de maturation intermédiaire (J4) avait des caractéristiques sensorielles d'un plantain au stade mûr (J8).

Il est également possible d'afficher des biplots, qui sont des représentations simultanées des variables et des observations dans l'espace ACP. Il peut être intéressant pour les communications orales ou articles de visualiser les caractéristiques sensorielles des produits dans un seul graphique.



3 RÉGRESSIONS LINÉAIRES

Dans le cadre du projet RTBfoods, un des objectifs est de fournir des outils faciles et rapides à mettre en œuvre pour évaluer la qualité sensorielle des produits finis. Il est clair que les analyses sensorielles de type QDA avec un panel entrainé sont plutôt lourdes à mettre en place (organisation des séances, préparation d'un grande quantité d'échantillon, nombre limité d'échantillons à analyser, maintenir la motivation et la performance des panélistes dans le temps, etc .). Aussi pour pallier à ce problème, nous développons des méthodes instrumentales plus faciles à mettre en œuvre et plus fiables dans le temps qui pourraient remplacer les QDA. Mais pour cela, il faut disposer de modèles de prédiction de cette qualité sensorielle les plus fiables et pertinents.

L'objectif ici est donc d'utiliser la régression linéaire multiple pour prédire les attributs sensoriels à partir des paramètres instrumentaux. Comme vous allez le voir, cela va plus loin que la simple mise en relation entre des variables.

3.1 Definitions - principles

Le principe de la régression linéaire est de modéliser une variable dépendante quantitative Y (ici l'attribut sensoriel), au travers d'une combinaison linéaire de p variables explicatives quantitatives X1, X2, X3, ... (ici les paramètres biophysiques). Le modèle s'écrit pour une observation i :

$Y_i = a_0 + a_1X_{i,1} + a_2X_{i,2} + \ldots\ldots a_pX_{i,p} + \varepsilon_i,$

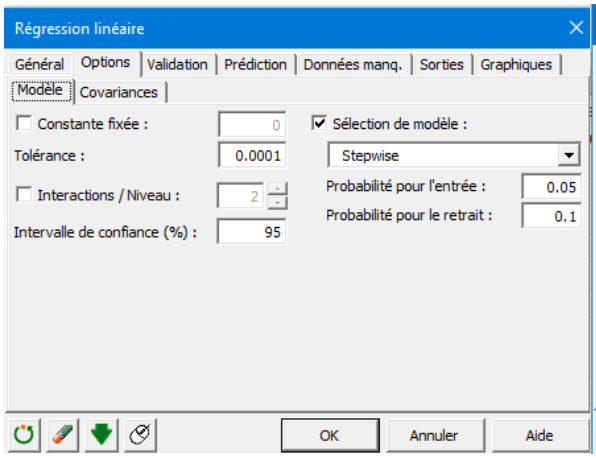
où Y_i est la valeur observée pour la variable dépendante pour l'observation i, $X_{i,j}$ est la valeur prise par la variable j pour l'observation i, et ε_i est l'erreur du modèle.

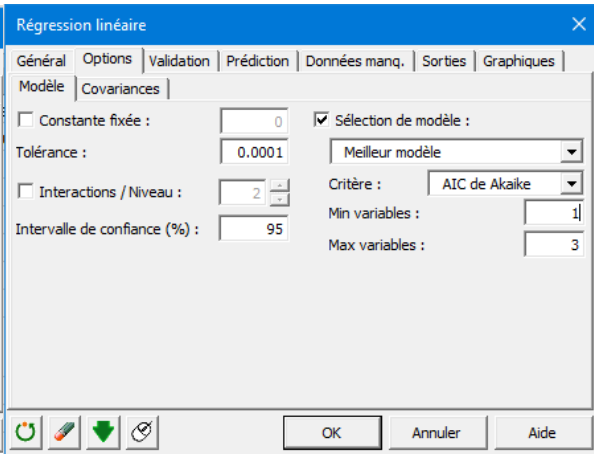
Pour la construction et l'interprétation du modèle, nous allons passer en revue plusieurs notions importantes.

Dans notre exemple, nous allons prédire les attributs sensoriels qui décrivent le plantain bouilli (travaux de thèse de Antonin Kouassi) à partir des 7 paramètres instrumentaux de texture (Puncture force, Hardness_TPA, etc …) et des 3 paramètres biochimiques (teneur en matière sèche, extrait sec soluble et acidité titrable).

3.2 Mise en œuvre d'une régression linéaire avec XLSTAT

- Ouvrir le fichier « Exemple ACP + regression.xls »
- Lorsque XLSTAT est activé, sélectionner la commande **XLSTAT / modélisation des données / Régression linéaire**.
- Dans la fenêtre "Général", sélectionner firmness (colonne E) de la feuille "Data" dans "Y / variables dépendantes".
- Sélectionner les paramètres de texture (colonnes M - S) ou biochimique (colonnes T – V) de la feuille "Data" dans "X / variables explicatives
- Sélectionner la colonne A dans "Libellé des observations"
- Dans la fenêtre "Option", la tolérance est fixée par défaut à 0.0001 et l'intervalle de confiance à 95%. Parmi les modèles proposés (meilleur modèle, stepwise, ascendant et descendant), nous vous proposons 2 options : modèle Stepwise et meilleur modèle.
- Si vous choisissez le modèle Stepwise, fixez une probabilité de 0.05 pour l'entrée d'une variable et de 0.1 pour la sortie d'une variable. Si vous choisissez le meilleur modèle, fixez 1 variable minimum, 3 variables maximum, et l'indice AIC pour déterminer le meilleur modèle.





Précisions sur la sélection du modèle

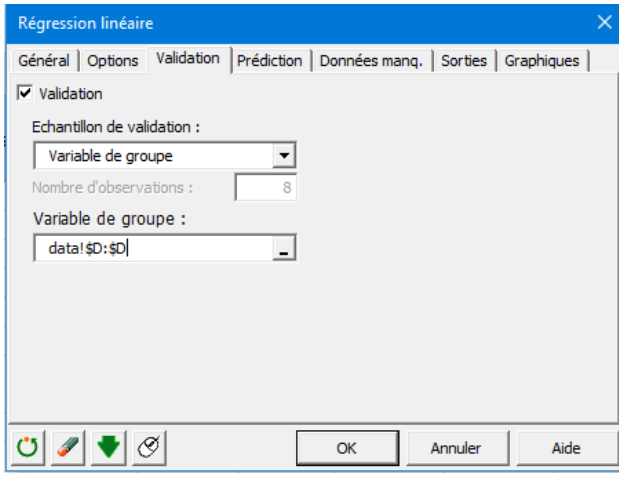
Stepwise : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle (le critère utilisé est la statistique t de Student). Si une seconde variable est telle que la probabilité associée à son t est inférieure à la « **Probabilité pour l'entrée** », elle est ajoutée au modèle. De même pour une troisième variable. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait (toujours au travers de la statistique t). Si la probabilité est supérieure à la « **Probabilité pour le retrait** », la variable est retirée. La procédure se poursuit jusqu'à ce que plus aucune variable ne puisse être ajoutée/retirée.

Ascendante : la procédure est identique à celle de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.

Descendante : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Meilleur modèle : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur. En général, nous utilisons le critère AIC de Akaike : plus ce critère est faible, meilleur est le modèle (il peut prendre des valeurs négatives).

- Dans la fenêtre “validation”, cocher « validation ». Vous pouvez alors choisir soit l’option “aléatoire”, soit l’option « variables de groupe ». Puis fixez un nombre d’observations qui constituera le jeu de validation.
    - En choisissant l’option « variables de groupe », vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour la construction du modèle (la calibration), et de 1 pour les observations à utiliser pour la validation du modèle. Dans notre exemple, c’est la colonne D qui indique les observations de calibration et de validation. Nous avons décidé dans notre exemple que les 27 échantillons de plantains ivoiriens constituaient les observations de calibration et les 8 échantillons de bananes non ivoiriennes constituaient les observations de validation. Nous aurions pu prendre par exemple les échantillons des 2 dernières séances de dégustation pour constituer le jeu de validation.



- En choisissant l'option « aléatoire », les observations sont sélectionnées de manière aléatoire. Si vous relancez la régression linéaire avec cette option, les nouvelles observations de validation seront bien sûr différentes et les résultats légèrement différents. Avec cette option, il est difficile finalement de décider du modèle définitif.

Pourquoi est-il important d'avoir des données de calibration et de validation ?

Pour que la prédiction soit robuste, le modèle à mettre en place doit être validé sérieusement. L'objectif est de prouver que le modèle génère de bonnes estimations des valeurs de la variable étudiée. Pour ce faire, il est nécessaire de travailler au moins avec un jeu de données dit de calibration (appelé aussi jeu d'apprentissage dans XLSTAT), et un jeu de données dit de validation. Assez simplement, les données de calibration sont utilisées pour construire le modèle tandis que les données de validation servent à montrer que le modèle est finalement fiable (ou robuste).

Indépendance des données de calibration et de validation - pourcentage de données calibration / validation

Pour être le plus objectif possible, les jeux de données de calibration et de validation devraient être issus de population indépendante (pour être sûr de ne pas biaiser le résultat). D'une manière plus générale, ce qui est fait est de séparer un ensemble d'échantillons initial de base en un jeu de données de calibration et un jeu de données de validation. A noter que s'il y a la possibilité d'acquérir deux jeux de données différents, un pour la calibration et un autre pour la validation du modèle, il est recommandé de le faire ! Il est souvent conseillé d'utiliser entre 60 et 80% du jeu de données initial comme jeu de calibration et les 20 à 40% restants comme jeu de validation. Néanmoins, ces pourcentages ne sont pas fixes.

***Remarque** : dans notre exemple, 27 observations ont été utilisées pour la calibration et 8 observations pour la validation, soit 23% de la totalité des observations.*

- Dans la fenêtre « données manquantes », cliquer sur "supprimer les observations », puis « vérifier pour chaque Y séparément ». L'observation sera supprimée même si la donnée manquante correspondant à une variable explicative non prise en compte par le modèle. Il faudra refaire la régression linéaire en enlevant cette variable pour que toutes les observations soient prises en compte.
- Dans la fenêtre « Sorties », cliquer sur "statistiques descriptives", "corrélations", "statistiques de multicolinéarité", "analyse de variance", « Type I/III SS », « coefficients normalisés », « Prédictions et résidus », « intervalle de confiance », et dans la fenêtre « tester les hypothèses », cliquer sur « test de normalité ».
- Dans la fenêtre « graphiques », cliquer sur tous.

Les calculs commencent une fois que vous avez cliqué **OK.**

3.3 Interprétation des résultats de la Régression Linéaire

Les résultats sur la prédiction de la fermeté à partir des mesures texturales sont présentés dans différents onglets. Dans les onglets RL0 à RL2, le jeu de validation est choisi à partir de l'option « variables de groupe ».

3.3.1 Statistiques descriptives

Les statistiques descriptives permettent de visualiser des informations sur les données issues des jeux de calibration et de validation, de comparer les valeurs (moyenne, écart type...) entre données de calibration et de validation, repérer des données aberrantes.

La matrice de corrélation permet de repérer les corrélations simples entre le descripteur sensoriel et les variables explicatives, ainsi qu'entre variables explicatives.

3.3.2 Comment interpréter les statistiques de multicolinéarité ?

Après avoir passé en revue les statistiques descriptives et la matrice de corrélation, il est important de s'attarder sur les statistiques de multicolinéarité.

Qu'est-ce que la multicolinéarité ?

On parle de colinéarité entre 2 variables explicatives lorsque qu'elles sont corrélées entre elles (R > 0.90 même si ce n'est pas une règle absolue). La colinéarité engendre plusieurs problèmes :

- les valeurs/signes des coefficients sont contradictoires, elles ne concordent pas avec les connaissances du domaine
- les résultats sont très instables, l'adjonction ou la suppression de quelques observations modifie du tout au tout les valeurs et signes des coefficients.

Le facteur d'inflation de la variance (VIF) est un indicateur de colinéarité. Il est admis qu'un VIF > 7 signale un problème de colinéarité.

Dans notre exemple, un VIF très élevé a été trouvé pour les variables Hardness, Gumminess et chewiness (onglet RL0). La matrice de corrélation montre que ces variables sont très fortement corrélées (R > 0.95). Cela peut se comprendre puisque gumminess est le produit de hardness et cohesiveness et chewiness est le produit de gumminess et springiness.

Statistiques de multicolinéarité :							
	Hardness_ TPA	Adhesiveness _TPA	Cohesiveness _TPA	Gumminess _TPA	Springiness_ TPA	Chewiness_ TPA	Puncture force
Tolérance	0,004	0,783	0,100	0,001	0,084	0,001	0,160
VIF	260,2	1,3	10,0	1415,7	11,9	772,4	6,3

En relançant la régression linéaire sans les variables Gumminess et Chewiness, les VIF sont inférieurs à 7, indiquant l'absence de colinéarité (onglet RL1).

Statistiques de multicolinéarité :					
	Hardness_ TPA	Adhesiveness _TPA	Cohesiveness _TPA	Springiness_ TPA	Puncture force
Tolérance	0,16	0,82	0,70	0,44	0,19
VIF	6,42	1,22	1,43	2,26	5,33

Dans la suite, les résultats sont ceux obtenus à partir de ces 5 variables explicatives.

Dans les tableaux suivants, XLSTAT affiche la synthèse de la sélection des variables pour le modèle.

Dans l'onglet RL1 qui propose une sélection des variables par Stepwise, une seule variable a été retenue.

Synthèse de la sélection des variables firmness par la méthode Stepwise:					
Nb. de variables	Variables	Variable IN/OUT	Statut	MCE	R^2
1	Puncture force	Puncture force	IN	0,839	0,798

Dans l'onglet RL2 qui propose une sélection des variables par le meilleur modèle (AIC), une seule variable a été retenue. Notons qu'il est proposé 2 autres modèles à 2 et 3 variables qui affichent un R^2 légèrement supérieurs. Mais comme le critère AIC est plus faible pour le premier modèle (avec une seule variable), c'est celui-ci qui a été retenu. De toute façon, les 2 modèles avec 2 et 3 variables n'améliorent pas beaucoup la prédiction : pour preuve les R^2 ajustés pour les modèles à 2 ou 3 variables sont presque égaux au modèle à une seule variable (le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle).

Synthèse de la sélection des variables firmness par la méthode du meilleur modèle :					
Nb. de variables	Variables	MCE	R^2	R^2 ajusté	AIC de Akaike
1	Puncture force	0,839	0,798	0,790	**-2,830**
2	Hardness_TPA / Puncture force	0,815	0,812	0,796	-2,699
3	Hardness_TPA / Cohesiveness_TPA / Puncture force	0,820	0,819	0,795	-1,684
Le meilleur modèle pour le critère de sélection choisi est affiché en bleu					

Nous pouvons constater dans cet exemple que finalement la sélection des variables par Stepwise ou par le meilleur modèle (AIC) conduit au même résultat : une prédiction de firmness par Puncture force avec un coefficient de détermination (R^2) de 0.798. **Ce R^2 signifie que 79.8% de la variabilité de la fermeté est expliquée par la puncture force. Ce coefficient est l'un des 3 critères importants (R^2, RMSEC, RMSEV) que nous allons calculer pour évaluer la précision et la robustesse de la prédiction.**

Le tableau « coefficients d'ajustement » affiche les statistiques relatives à l'ajustement du modèle de régression. La signification de ces coefficients est disponible sur le répertoire Aide en cliquant sur l'icone « Aide » de la fenêtre générale.

Coefficients d'ajustement (firmness) :		
Statistique	Echantillon d'apprentissage	Echantillon de validation
Observations	27,000	8,000
Somme des poids	27,000	8,000
DDL	25,000	6,000
R^2	0,798	0,907
R^2 ajusté	0,790	
MCE	0,839	0,959
RMCE	0,916	0,979
MAPE	17,988	15,416
DW	1,911	
Cp	1,623	
AIC	-2,830	
SBC	-0,238	
PC	0,234	

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Comme l'option Type I/III SS (SS : Sum of Squares) a été activée, les tableaux de Type I SS et de Type III SS sont affichés. Nous ne regardons que le tableau de Type III SS où l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Comme ici, il n'y a qu'une seule variable, il n'y a pas de différence entre les types I et III SS.

Les paramètres du modèle sont ensuite affichés. L'équation du modèle est ensuite affichée en reprenant les paramètres significatifs du modèle (en gras dans le tableau).

Paramètres du modèle (firmness) :						
Source	Valeur	Erreur standard	t	Pr > |t|	Borne inférieure (95%)	Borne supérieure (95%)
Constante	**1,459**	0,379	3,844	**0,001**	0,677	2,240
Hardness_TPA	0,000	0,000				
Adhesiveness_TPA	0,000	0,000				
Cohesiveness_TPA	0,000	0,000				
Springiness_TPA	0,000	0,000				
Puncture force	**1,106**	0,111	9,951	**< 0,0001**	0,877	1,335
Equation du modèle (firmness) :						
firmness = 1,459+1,106*Puncture force						

La fermeté est prédite par la puncture force selon l'équation suivante : Firmness = 1.459 + 1.106*Puncture force.

Le tableau des **coefficients normalisés** permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Dans le cas d'un modèle à plusieurs variables, ce tableau permet de visualiser la variable qui influe le plus sur la variable à expliquer.

Coefficients normalisés (firmness) :						
Source	Valeur	Erreur standard	t	Pr > |t|	Borne inférieure (95%)	Borne supérieure (95%)
Hardness_TPA	0,000	0,000				
Adhesiveness_TPA	0,000	0,000				
Cohesiveness_TPA	0,000	0,000				
Springiness_TPA	0,000	0,000				
Puncture force	0,894	0,090	9,951	**< 0,0001**	0,709	1,078

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation en particulier la valeur observée et la valeur prédite, ainsi que la différence entre les 2 (appelée résidu).

A partir de ce tableau, **nous allons pouvoir calculer deux critères très importants pour la prédiction, l'erreur quadratique moyenne (appelé en anglais Root Mean Square Error) sur le jeu de calibration (RMSEC) et sur le jeu de validation (RMSEV)**.

Qu'est-ce que l'erreur quadratique moyenne (RMSE) ?

Ce critère fournit une indication sur la dispersion des valeurs observées par rapport aux valeurs prédites. Il peut être relié à la variance du modèle.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Yi_p - Yi_o)^2}{n}}$$

Où Yip et Yio sont respectivement les valeurs prédites et observées pour l'échantillon i et n est le nombre d'échantillons.

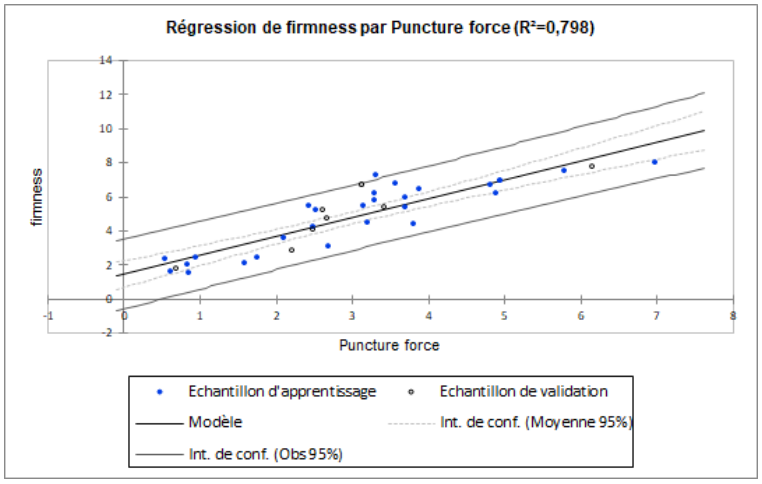
L'unité de ce critère est celui de la variable à expliquer. Dans le cas d'attributs sensoriels, le RMSE exprime une dispersion des valeurs observées par rapport aux valeurs prédites sur une échelle de 0 à 10. Dans le cas d'un RMSE proche de 1, cela signifie que le modèle est capable de prédire un attribut sensoriel avec plus ou moins 1 point d'écart sur une échelle de 0 à 10, ce qui est généralement une bonne prédiction car proche de notre niveau de répétabilité (comprise entre 1 et 2).

Le RMSEC a été calculé sur les 27 observations du jeu de calibration et le RMSEV sur les 8 observations du jeu de validation selon la formule ci-dessus.

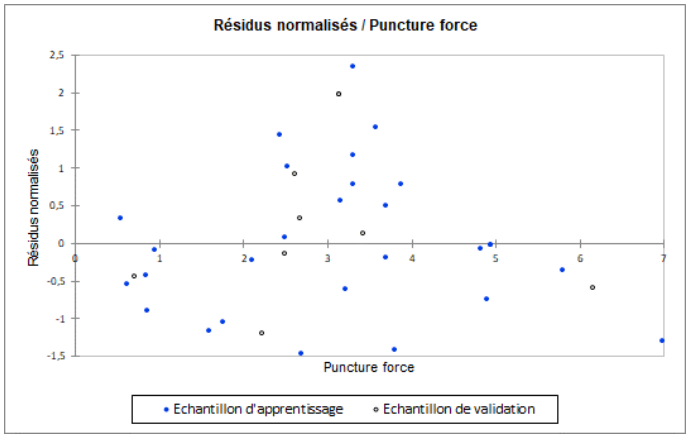
RMSEC = 0.88 (cellule O157) et RMSEV = 0.85 (cellule O193).

Ces valeurs sont inférieures à 1, ce qui signifie que la prédiction a été précise (grâce au R^2 de 0.798 et RMSEC de 0.88) et robuste (grâce au RMSEV de 0.85).

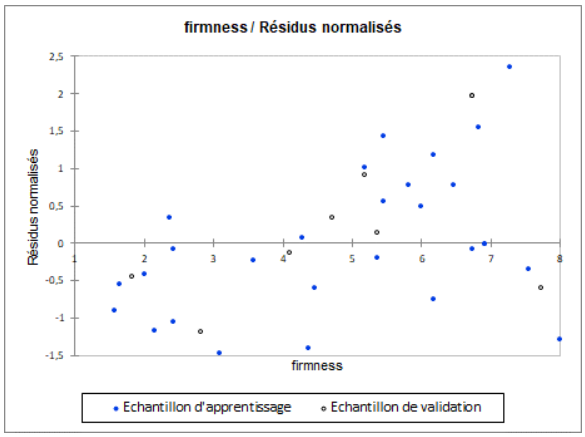
Les graphiques qui suivent permettent de visualiser les résultats mentionnés ci-dessus. Comme il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions.

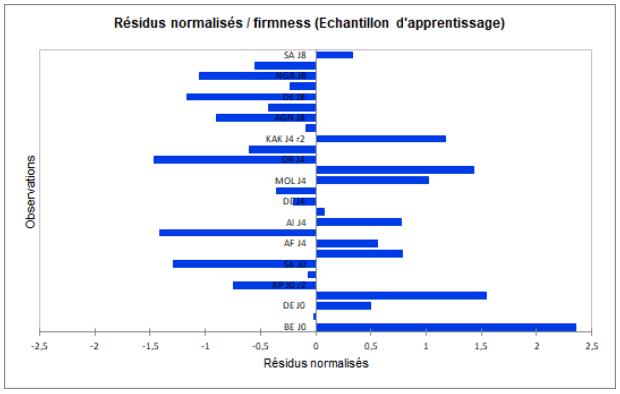


Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle, notamment l'hétéroscédasticité (ou forme en entonnoir).

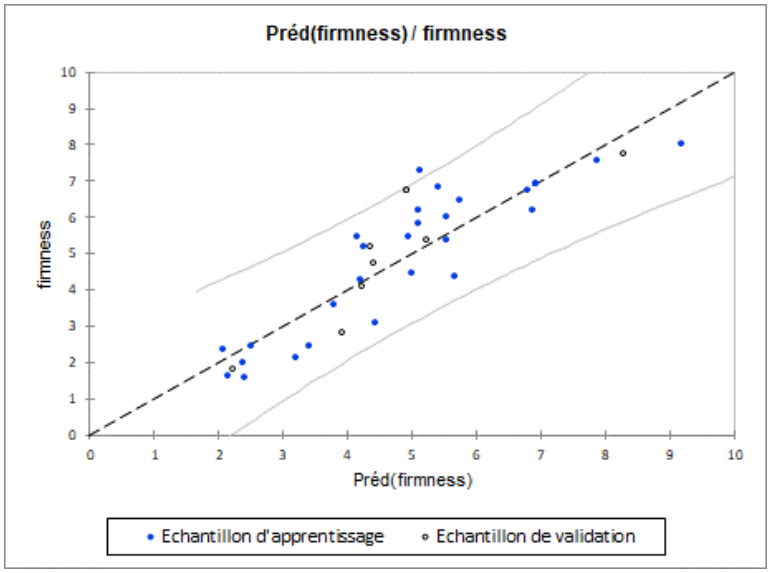


Les graphiques suivants permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons (pour le jeu de calibration et le jeu de validation). Ce dernier graphique permet de voir rapidement si un nombre anormal de données sort de l'intervalle [-2, 2] sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.





Le dernier graphique permet de visualiser la fermeté observée en fonction de la fermeté prédite avec l'intervalle de confiance. En général ce qui est présenté dans des articles ou communication est l'inverse, c'est à dire la fermeté prédite en fonction de la fermeté observée. Pour cela, en cliquant sur les points de la figure, modifiez les lettres de la série pour avoir la fermeté prédite en ordonné et la fermeté observée en abscisse.



Enfin un test de normalité des résidus est proposé. Il est important de le faire car s'il est négatif (c'est-à-dire l'hypothèse de normalité des résidus est rejetée), cela signifie que les résidus ne suivent pas une loi normale, et donc que la construction de la prédiction n'a pas de sens. Ici la p-value est supérieure à 0.05, ce qui signifie que les résidus suivent une loi normale et donc que nous pouvons accepter cette prédiction.

Test de l'hypothèse de normalité des résidus (firmness) :	
W	0,970
p-value (bilatérale)	0,590
alpha	0,05
Interprétation du test :	
H0 : Les résidus suivent une loi Normale.	
Ha : Les résidus ne suivent pas une loi Normale.	
Etant donné que la p-value calculée est supérieure au niveau de signification seuil alpha=0,05, on ne peut pas rejeter l'hypothèse nulle H0.	

3.4 Questions diverses

3.4.1 Peut-on se passer d'une validation ?

Dans beaucoup de travaux publiés, les relations entre les descripteurs sensoriels et les mesures instrumentales sont établies avec un seul jeu de données sans tester leur validation.

Dans notre exemple, nous l'avons fait sur la variable Firmness (onglet RL3) à partir des 35 observations sans jeu de validation. Firmness est prédite à la fois par puncture force et hardness avec un R^2 de 0.83 et un RMSE de 0.79. Ces résultats montrent que firmness peut être prédite par 2 variables de texture complémentaires (test de pénétrométrie et de compression) avec une erreur de moins de 0.80 point (sur une échelle de 0 à 10). Ce résultat est meilleur que celui trouvé précédemment avec les 27 observations (R^2 = 0.80, RMSEC = 0.88) mais nécessite une mesure supplémentaire (hardness mesuré par un test de compression). Remarquons que la prédiction a été guère améliorée (R^2 qui est passé de 0.80 à 0.83, RMSE de 0.88 à 0.80) malgré la variable supplémentaire. L'absence d'un jeu de validation ne nous permet pas de conclure sur la robustesse de cette prédiction. Pour valider cette prédiction, il faudrait un nouveau jeu de données à partir de nouvelles variétés ou des

variétés déjà étudiées lors d'une nouvelle saison de production (bien sûr caractérisées avec le même jury et dans les mêmes conditions).

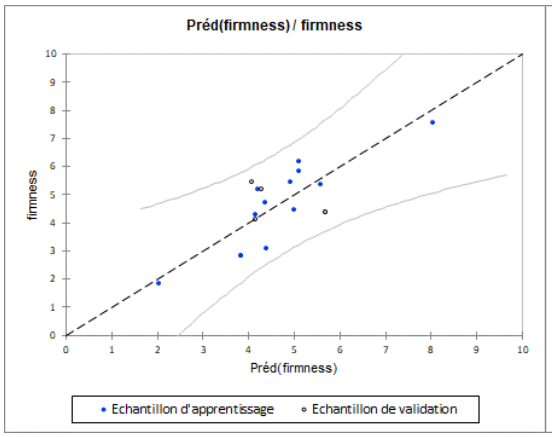
3.4.2 Combien d'observations sont-elles nécessaires pour établir les prédictions ?

Avant tout, il n'existe pas d'outil statistique qui permette de définir le nombre d'observations minimum pour établir des prédictions. Nous ne pourrons donner que quelques conseils.

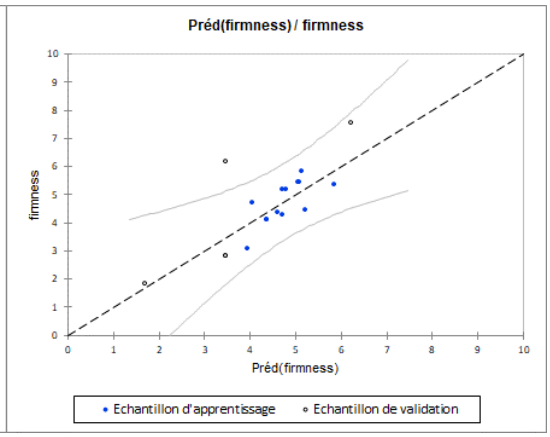
1- Compte-tenu de la lourdeur des analyses sensorielles, il est très difficile de dépasser les 50 observations (qui nécessiteraient au moins 10 séances avec le même jury), contrairement aux mesures instrumentales (plus de 100 pour prédire des paramètres chimiques à partir des mesures spectrales).

2- Privilégier des observations représentatives de toute la diversité sensorielle, c'est-à-dire des observations qui couvrent l'étendue des échelles d'intensité pour les attributs majeurs. Voici 2 exemples de construction de firmness avec 16 observations seulement extraites des plantains au stade de murissage J4. Dans les cas, 12 observations ont été choisies en calibration et 4 en validation. Dans le premier cas (onglet RL4), les observations de calibration couvraient bien la gamme de firmness (de 1.8 à 7.5), alors que dans le deuxième cas (onglet RL5), les observations de calibration couvraient une gamme de firmness plus réduite (de 3.0 à 5.8). Le tableau ci-dessous et les figures sont une synthèse des résultats de prédiction.

Données calibration	firmness	Equation du modèle	R^2	RMSEC	RMSEV
RL4	1.8 – 7.5	firmness = 1.2 + 1.2*Pf	0.77	0.73	1.06
RL5	3.0 – 5.8	firmness = 0.9 + 8.0*Hardness	0.47	0.54	1.55
RL1	*1.6 – 8.0*	*Firmness = 1.4 + 1.1*Pf*	*0.80*	*0.85*	*0.88*



Avec les données RL4



avec les données RL5

La prédiction obtenue avec les données RL4 est finalement très proche de celle obtenue avec toutes les données (RL1), alors que le nombre d'observations était de 16 (contre 35 pour RL1). Le fait d'avoir des valeurs de firmness qui couvrent bien toute l'échelle

d'intensité permet cette bonne prédiction malgré un faible nombre d'observations. Par contre, la prédiction de firmness avec les données RL5 est moins précise et diffère des précédentes. Cela s'explique par l'utilisation en calibration d'observations qui ne couvrent que partiellement l'échelle d'intensité. Remarquons qu'une observation du jeu de validation (KAK J4 r2) a été mal prédite à partir de l'équation de RL5, ce qui met en cause la robustesse de cette prédiction (d'où la pertinence d'avoir un jeu d'observation pour la validation, voir première question « peut-on se passe de validation ? »).

3. La précision et la robustesse d'une prédiction ne dépend pas seulement du nombre d'observations mais aussi de la nature des liens entre la variable à expliquer et des variables explicatives. Dans notre exemple, stickiness a été très mal prédit par les mesures texturales et ceci quel que soit le nombre d'observations. Il est clair que pour cet attribut, il sera nécessaire de rechercher d'autres types de mesures instrumentales pour espérer identifier des prédicteurs de cet attribut.

Institut : Cirad – UMR QualiSud

Addresse : C/O Cathy Méjean, TA-B95/15 - 73 rue Jean-François Breton - 34398 Montpellier Cedex 5 - France

Tel : +33 4 67 61 44 31

Email : rtbfoodspmu@cirad.fr

Site Web : https://rtbfoods.cirad.fr/