# COMPARAISON DE METHODES DE DISCRIMINATION. APPLICATION A L'ETUDE D'UNE COLLECTION DE CHROMATOGRAMMES DE PROTEINES DE BLE.

P.COURCOUX [1], D.BERTRAND [2], J.C.AUTRAN [3]

1) ENITIAA, Chaire de Mathématiques et Informatique, Rue de la Géraudière 44072 NANTES

2) INRA, Laboratoire de Technologie Appliquée à la Nutrition, Rue de la Géraudière 44026 NANTES

3) INRA, Place Viala 34060 MONTPELLIER

## SUMMARY

Continuous signals such as spectra, chromatograms or electrophoregrams are very often used in quality control. They generally present a large amount of data and redundant informations and therefore require special treatments. Several procedures of classification on such signals are described. Their efficencies are compared on a collection of 180 chromatograms of wheat proteins. The discrimination of both genotypes and growing years led to about 95 per cent right classifications. The quality of reallocation scores shows that they could be used as prediction of technological quality of wheat flour.

## INTRODUCTION

La chromatographie est une méthode performante de caractérisation d'un produit et son utilisation en contrôle de la qualité est devenue courante. Les signaux enregistrés sont continus et leur digitalisation conduit à des spectres de plusieurs centaines de données dont l'exploitation est souvent délicate. La taille des signaux traités et la grande redondance des données recueillies sont le plus souvent réduites par intégration des principaux pics observés, méthode qui présente l'avantage d'être peu sensible à d'éventuels décalages de la ligne de base ou des temps de rétention. Cependant, la perte d'informations peut être conséquente et occulter dans certains cas des phénomènes importants. De nombreuses applications de méthodes d'analyse de données ont été développées pour le traitement des signaux très proches de la chromatographie (spectoscopie, électrophorèse, ...) et ont conduit à des résultats probants. L'objectif de ces travaux est souvent d'extraire du signal original les informations permettant d'expliquer ou prédire une caractéristique du produit étudié. Dans ce travail, on testera la pertinence de quelques méthodes de discrimination et de classement sur des chromatogrammes de protéines de blé. La qualité des farines, notamment boulangère, étant très liée au génotype des blés employés ainsi qu'à leurs conditions culturales, la discrimination de ces paramètres est un problème d'intérêt industriel. On comparera l'efficacité et la pertinence des différentes méthodes employées sur une collection de chromatogrammes de farines de blé.

## MATERIEL ET METHODES

Nature de la collection et des signaux traités.

La collection étudiée comporte 180 échantillons de blé provenant d'un programme de sélection de l'INRA et appartenant à 7 variétés cultivées dans des lieux différents. On dispose de lots récoltés sur deux campagnes différentes (87-88).

Les protéines extraites de la farine ont été caractérisées par gel filtration en HPCL. Les conditions opératoires de la préparation et de la chromatographie des échantillons sont décrites dans Dachkevitch, Autran (1989).

Après digitalisation, les signaux ont été tronqués entre 7 et 22 mn et chaque chromatogramme est caractérisé par une série de 151 données. La figure 1 présente 12 signaux appartenant à des génotypes et des années de cultures différents. Leur observation directe montre peu de différences visibles hormis le décalage de hauteur de la ligne de base qui ne résulte que de réglages différents du détecteur et ne possède pas de signification biochimique.

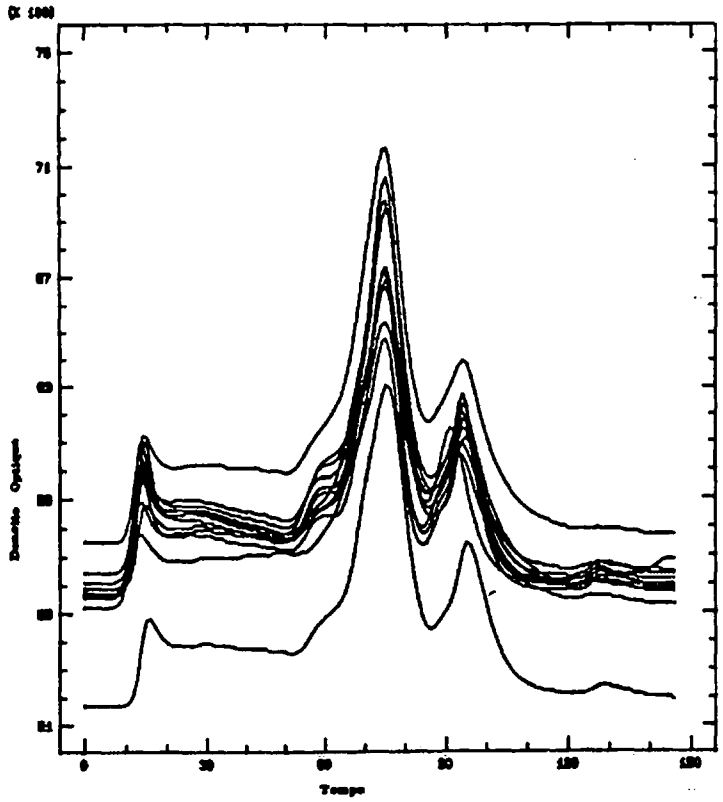

Figure1: Exemples de chromatogrammes (gel filtration en HPLC) de protéines de blé.

Etapes préliminaires.

Un traitement préalable a été appliqué aux signaux étudiés pour éliminer d'éventuels décalages des temps de rétention, le pic le plus important servant à aligner les chromatogrammes. De même, les lignes de base ont été réajustées afin de ne pas introduire de biais dans le traitement.

La collection a été divisée, par tirage au hasard, en 2 lots : l'échantillon étalon (120 chromatogrammes) permettra d'estimer les critères de décision et l'échantillon de vérification (60 chromatogrammes) servant à valider les procédures utilisées.

Avant toute procédure de discrimination, la collection a été traitée par analyse en composantes principales (ACP) afin d'éliminer la redondance des informations et réduire les données. On traitera par la suite les premières coordonnées factorielles de l'ACP plutôt que les 151 variables originales.

Procédures de discrimination.

* méthode A : on reclasse les échantillons au groupe dont le centre de gravité est le plus proche (au sens de la distance euclidienne usuelle).

Si l'on désigne par $g_i$, $i=1..Q$, les centres de gravité des différentes classes, l'observation x sera affectée à la classe q telle que :

$$d^2(x,g_q) = \min_{i=1..Q} (d^2(x,g_i))$$

* méthode B (ou méthode des K plus proches voisins)

Pour une observation x, on définit la densité du groupe q de la façon suivante:

$$d_q(x) = \frac{K_q}{n_q V_x}$$

Avec $K_q$ : nombre de points du groupe q appartenant aux K plus proches voisins de x

$n_q$ : cardinal du groupe q

$V_x$ : volume de la sphère contenant les K plus proches voisins de x

La méthode consiste ensuite à affecter un individu au groupe dont la densité parmi ses K plus proches voisins est la plus grande.

Le nombre K de voisins à calculer influant sur la qualité de la discrimination, il est impératif de tester plusieurs valeurs de ce paramètre.

* méthode C :

La collection de chromatogrammes est traitée par une analyse factorielle discriminante (AFD). L'ACP préliminaire permet de s'affranchir de la singularité de la matrice de variance-covariance des données initiales. On reclasse les observations par examen des distances à chacun des centres de gravité des groupes en utilisant la distance euclidienne usuelle sur les coordonnées issues de l'AFD. Cette technique est strictement identique à une affectation utilisant la métrique de Mahalanobis sur les données d'origine.

* méthode D :

L'utilisation de la distance euclidienne usuelle dans la méthode précédente est la plus courante mais détermine des surfaces de décision linéaires entre classes. Elle est donc mal adaptée à des groupes de formes et de tailles différentes.

Une variante peut consister à créer une métrique par groupe et à utiliser la règle d'affectation suivante :

x est affecté au groupe q si $d^2_{Mq}(x,g_q) = \min_{l=1..Q} (d^2_{Ml}(x,g_l))$

avec $d^2_{Ml}(x,g_l) = {}^t(x-g_l)\, M_l\, (x-g_l)$

et $M_l$ la matrice de terme général $1/var_{lj}$ ($var_{lj}$ étant la variance de la variable j dans le groupe l).

D'autres types de métriques locales sont utilisables, notamment des métriques non diagonales, mais demandent des collections de données très grandes pour estimer les nombreux paramètres nécessaires à leur mise en oeuvre.

**RESULTATS - DISCUSSION**

La figure 2 décrit les valeurs propres de l'ACP préliminaire.

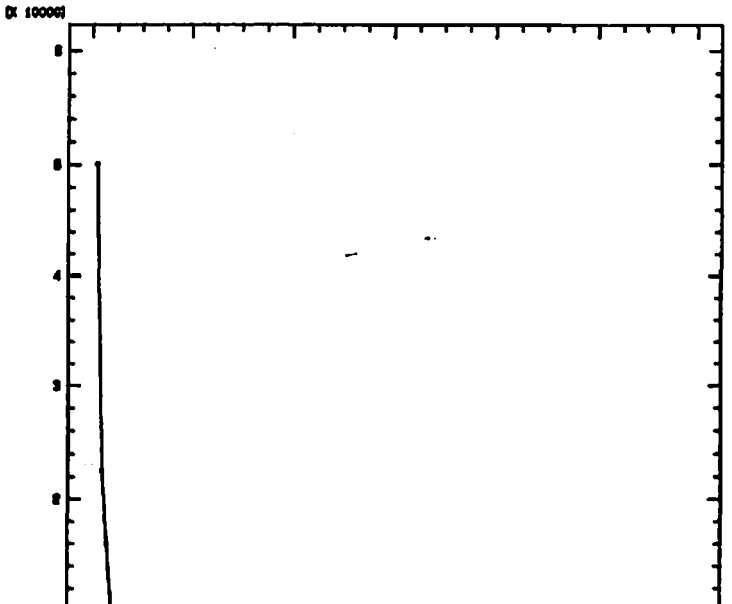


Figure 2: Valeurs propres de l'ACP préliminaire sur les 120 chromatogrammes étalons.

On constate une décroissance rapide puisque les 20 premiers axes factoriels expliquent plus de 99% de l'inertie totale. Par la suite, on traitera les coordonnées des observations sur ces 20 premiers axes.

L'application de la méthode A aux données issues de l'ACP montre une bonne discrimination de l'année de culture (figure 3); sur l'échantillon de vérification, seules 4 observations ne sont pas reclassées correctement dans leurs groupes d'origine (soit 6.6%). La discrimination des génotypes conduit à un taux de bien-classés de 87% sur l'échantillon étalon et de 84% sur l'échantillon de vérification.

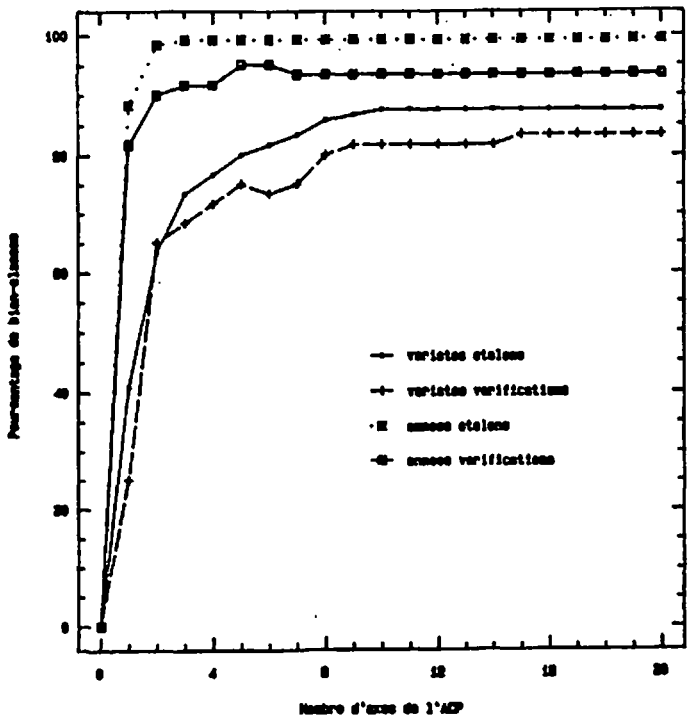

Figure 3: Méthode A : discrimination des génotypes et des années de culture. Résultats d'affectation en pourcentage de biens classés.

La méthode des K plus proches voisins (méthode B) améliore légèrement les résultats de la méthode précédente. Des essais de variation du nombre de voisins examinés montrent que les meilleurs résultats sont obtenus avec K égal à 1 (autrement dit, en observant uniquement le voisin le plus proche de chacune des observations). La qualité de ces affectations décroit lentement avec une augmentation de K. Ce constat est bien sûr lié aux données et à la bonne séparation des groupes. L'année de culture (figure 4) est très bien reconnue, les observations de vérification étant reclassés à deux exceptions près. Pour les génotypes, les observations étalons sont reclassés à 89% et les vérifications à 95%.

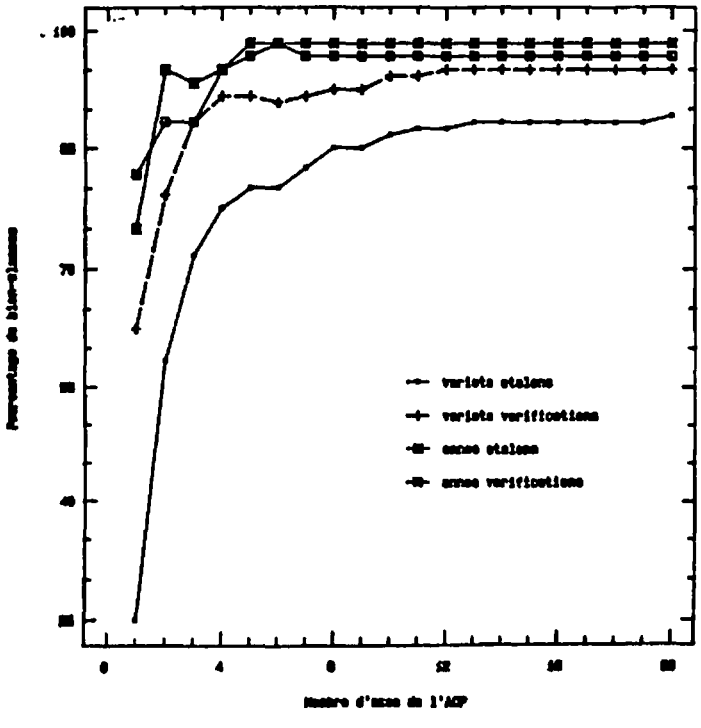

Figure 4: Méthode B: (plus proches voisins) discrimination des génotypes et des années de culture. Résultats d'affectation en pourcentage de bien classés.

Pour l'évaluation des performances de la méthode C (affectation en utilisant la distance de Mahalanobis) et D (utilisation d'une métrique par groupe), on a effectué une analyse factorielle discriminante. Le plan 1-2 de l'AFD sur les variétés est reproduit à la figure 5.

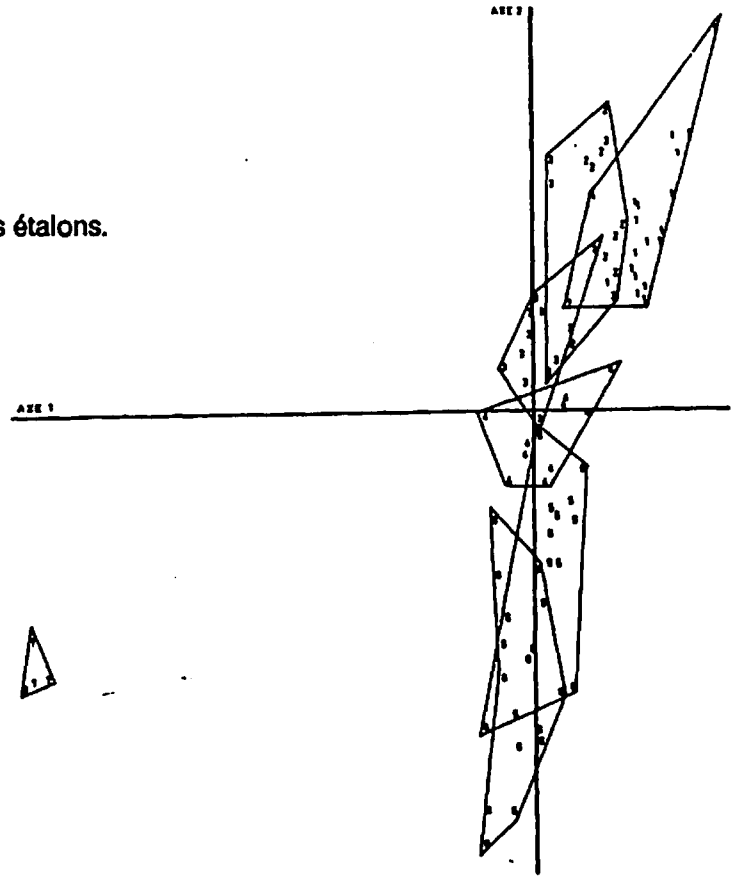


Figure 5: Discrimination des génotypes: premier plan factoriel de l'AFD sur les chromatogrammes étalons.
1:Camp Rémy 2:Festival 3:Scipion
4:Thésée 5:Fidel 6:Pernel 7:Appolo

L'axe le plus discriminant sépare parfaitement la variété Appolo qui apparait très différente des autres, confirmant l'aspect particulier de son spectre. On trouvera à la figure 6 les résultats des procédures d'affectation C et D appliquées aux coordonnées issues de l'AFD. On constate une amélioration des résultats par rapport aux procédures précédentes, avec un avantage pour la méthode D (métriques locales) qui reclasse parfaitement les individus étalons et affecte 53 observations de vérification sur 60 dans leurs groupes d'appartenance.

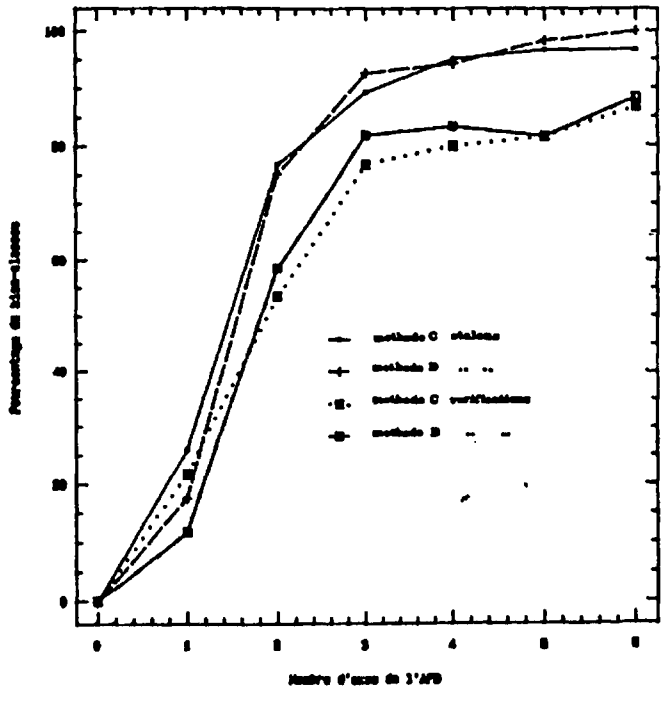

Figure 6: Affectation après AFD. Résultats des méthodes C et D appliquées à la discrimination des génotypes en pourcentage de bien classés.

## CONCLUSIONS

Les résultats présentés montrent que les travaux de reconnaissance de formes à partir d'analyse de données peuvent être appliqués à des signaux de chromatographie. Les résultats de reclassement de différentes méthodes de discrimination montrent leur efficacité pour la reconnaissance de variétés de blé et de conditions de culture à partir de chromatogrammes de protéines. Une technique d'affectation utilisant des métriques locales aprés analyse factorielle discriminante semble supérieure aux autres et conduit à des prédictions très efficaces. La qualité de ces résultats de classification et la liaison étroite entre génotype et caractéristiques physico-chimiques ou technologiques des blés devraient permettre de prédire avec précision la qualité des farines à partir de leurs chromatogrammes.

## BIBLIOGRAPHIE


D.BERTRAND, P.ROBERT, W.LOISEL, *Identification of some wheat varieties by near infrared reflectance spectroscopy*, Journal of the Science of food and Agriculture 36, 1120 (1985).

G.CELEUX, E.DIDAY, G.GOVAERT, Y.LECHEVALLIER, H.RALAMBONDRAINY, *Classification automatique des données*, Dunod (1989)

T.DACHKEVITCH, J.C.AUTRAN, *Prediction of baking quality of bread wheats in breeding programs by size-exclusion HPLC*, Cereal Chemistry, (1989,sous presse).

M.F.DEVAUX, D.BERTRAND, P.ROBERT, M.QANNARI, *Application of multidimensional analyses to the extraction of discriminant spectral patterns from NIR spectra*, Applied spectroscopy, 42, (6), 1015 (1988).

E.DIDAY, J.LEMAIRE, J.POUGET, F.TESTU, *Eléments d'analyse de données*, Dunod (1982)

J.M.ROMEDER, *Méthodes et programmes d'analyse discriminante*, Dunod, (1973)

R.TOMASSONE, M.DANZART, J.J.DAUDIN, J.P.MASSON, *Discrimination et classement*, Masson (1988)

M.C.VIRION, *Méthodologies statistiques de la discrimination: Application aux électrophorégrammes des farines de blés*, Université des sciences et techniques du Languedoc, Thèse de 3$^{e}$ cycle, Montpellier (1988)

ASU

ASSOCIATION
POUR LA STATISTIQUE
ET SES UTILISATIONS

ABK6

1ères JOURNÉES

# AGRO-INDUSTRIE et METHODES STATISTIQUES

## Angers, 14 et 15 juin 1990

**Organisées par :**

L'Ecole Supérieure d'Agriculture d'Angers ( ESA )

L'Ecole Nationale Supérieure Agronomique de Rennes ( ENSAR )

L'Ecole Nationale des Ingénieurs des Techniques
des Industries Agricoles et Alimentaires ( ENITIAA )

L'Université de Rennes II

L'Institut Universitaire de Technologie de Vannes

sous le patronage de :
**l'Association pour la Statistique et ses Utilisations (ASU)**

GROUPE ESA