CONTAMINATION PAR LE MERCURE ET CLASSIFICATION D'ESPECES EN ECOTOXICOLOGIE : APPROCHE CLASSIQUE, APPROCHE SYMBOLIQUE

Marie Chavent, Chantal Lacomblez

Mathématiques Appliquées de Bordeaux (MAB), UMR 5466
Université Bordeaux 1, 351 cours de la libération, 33405 Talence cedex
e-mail : *chavent@math.u-bordeaux.fr, lacomble@mi2s.u-bordeaux2.fr*

Alain Boudou, Régine Maury-Brachet

Laboratoire d'Ecophysiologie et d'Ecotoxicologie des Systèmes aquatiques (LEESA)
UMR EPOC 5805
Université Bordeaux 1, 351 cours de la libération, 33405 Talence cedex
e-mail : *a.boudou@ecotox.u-bordeaux.fr, r.maury-brachet@ecotox.u-bordeaux.fr*

Résumé : *l'objectif de cette étude est d'acquérir une meilleure connaissance de la contamination par le mercure des communautés de poissons du Haut-Maroni en Guyane Française, afin de prévenir les risques liés à la consommation de certaines espèces. Pour cela deux approches, l'une classique et l'autre symbolique, sont proposées pour la classification des espèces en fonction de la concentration en mercure dosée dans certains organes.*

Mots-clés *: Classification automatique, Analyse des Données Symboliques, Classification descendante hiérarchique, Segmentation non supervisée, Généralisation/Spécialisation.*

1. Introduction

Des études réalisées sur l'ensemble du territoire de la Guyane Française ont mis en évidence des imprégnations par le mercure supérieures à la norme OMS dans les cheveux des populations amérindiennes vivant dans la zone du Haut - Maroni. Cette contamination a été attribuée à la forte consommation par ces populations de poissons des rivières, ces derniers étant considérés comme les vecteurs privilégiés du métal [COR88], [BOU98], [BOU99]. Afin d'acquérir une meilleure connaissance de la contamination des poissons par le mercure, 265 poissons appartenant à 36 espèces différentes ont été péché en 1997 par les chercheurs du LEESA, en collaboration avec l'ORSTOM de Cayenne. L'étude qui suit porte sur un échantillon de 67 poissons pour lesquels on a répertorié l'espèce, le régime alimentaire et dosé la concentration en mercure dans les différents organes (en ng par gramme de poids sec).

D'une manière générale, on dispose d'un tableau de données où les n individus sont décrits sur p variables, une des variables étant qualitative à m modalités, les $p-1$ autres variables étant quantitatives. Le problème est de trouver, non pas une partition des n individus, mais une partition en K classes de l'ensemble des m modalités. Dans cette application, il s'agit d'un tableau de données où 67 poissons de 10 espèces différentes (variable qualitative à 10 modalités) sont décrits par 5 variables quantitatives indiquant la répartition de la concentration en mercure dans 5 organes du poisson (foie, branchies....).

Nous proposons deux approches afin de trouver une classification de ces 10 espèces de poissons.

La première approche cherche des classes de modalités telles que les individus possédant les modalités d'une même classe soient homogènes sur les $p-1$ autres variables quantitatives. Elle est donc basée sur une méthodologie classique de classification des individus, ici les 67 poissons, suivie d'une interprétation des classes en fonction de la variable qualitative, ici la variable espèce. On en déduit alors que deux modalités sont proches si une majorité des individus qui les possèdent sont dans une même classe. L'étude des 67 poissons et la classification des 10 espèces a été réalisée avec le logiciel SPAD3.5 [LEB96] et le logiciel SIMCA-P [UME96].

La deuxième approche est basée sur une méthodologie dite symbolique, où chaque modalité est décrite sur chaque variable quantitative par un intervalle de valeurs traduisant la variation sur cette variable des individus possédant cette modalité. Ces intervalles sont obtenus grâce à

un algorithme, DB2SO, de Généralisation/Spécialisation [STE96], [STE97]. La classification se fait donc sur un tableau de données à m lignes et $p-1$ colonnes, chaque case contenant un intervalle. La classification de ces m lignes (les 10 espèces) est obtenue grâce à un algorithme monothétique de classification hiérarchique descendante, DIV, que l'on peut également envisager comme un algorithme de segmentation non supervisé [CHA97], [CHA00]. Enfin, chaque classe de la partition ainsi obtenue est interprétée visuellement grâce à un outil, SOE (Symbolic Object Editor), permettant de représenter graphiquement les variations de certaines variables dans chaque classe [NOI97]. Ces trois méthodes sont implémentés dans le logiciel SODAS (Symbolic Official Data Analysis System) issu d'un projet européen concrétisant les différentes avancées récentes de plusieurs équipes européennes en Analyse de Données Symboliques [DID99].

2. Les données

Les données ont été collectées par les chercheurs du LEESA, en collaboration avec l'ORSTOM de Cayenne. Parmi plusieurs centaines de poissons péchés, appartenant à 36 espèces différentes et consommées par la population locale, 67 poissons appartenant à 10 espèces et 3 régimes alimentaires ont été sélectionnés pour l'étude qui va suivre (voir Tab. 1)

Carnivores	Omnivores	Détritivores
Ageneiosus brevifilis (7)	Leporinus fasciatus (3)	Doras micropoeus (8)
Cynodon gibbus (7)	Leporinus frederici (3)	Platydoras costatus (10)
Hoplias aïmara (10)		Pseudoancistrus barbatus (7)
Potamotrygon hystrix (4)		Semaprochilodus varii (8)

Tab. 1 : Tableau des effectifs des poissons de 10 espèces et 3 régimes différents

Les 5 variables quantitatives utilisées (voir Tab. 2) sont basées sur le rapport entre la concentration en mercure (µg/g pds sec) dans 5 organes (branchies, foie, intestins, estomac, reins) et la concentration dans le muscle. L'asymétrie des distributions de ces cinq rapports a motivé leur transformation logarithmique.

	ESPECE	REGIME	LN(Foie/Muscle)	....	LN(Estomac/Muscle)
1	Ageneiosus brevifili	Carnivore	-0,12	....	NA
	Cynodon gibbus	Carnivore	1,59	....	0,22
....	....	....	....	....	....
	Leporinus frederici	Omnivore	-0,04	....	-1,77
....	....	....	....	....	....
	Doras micropoeus	Détritivores	0,8	....	-0,89
67	Doras micropoeus	Détritivores	1,34	....	-1,45

Tab. 2 : Extrait du tableau de données

3. Approche classique : étude des 67 poissons

Dans un premier temps, une Analyse en Composantes Principales (ACP) suivie d'une classification ascendante hiérarchique de Ward sur les coordonnées factorielles a permis de mettre en évidence 4 classes de poissons. Dans un deuxième temps, l'ACP a été remplacée par une régression PLS et la partition en 4 classes de la classification hiérarchique sur les composantes PLS a été comparée aux premiers résultats.

3.1. ACP et classification hiérarchique

La projection sur le premier plan factoriel des centres de gravité des 10 espèces et des 4 classes (Fig. 1) nous donne une première intuition de la classification des 10 espèces. On note déjà que deux espèces sont plus difficiles à apparier.

Cette première impression est confirmée par l'interprétation des classes qui nous permet de trouver une classification de 8 espèces sur 10

- classe 1/4 : constituée à 86% de carnivores (41% de carnivores dans la population totale) et contient 92% des carnivores. Elle contient 100% des Ageneiosus brevifilis, des Cynodon Gibbus et des Hoplias aîmara.

- Classe 2/4 : constituée à 100% d'omnivores et contient 83% des omnivores. Elle contient 100% des Leporinus fasciatus (3 sur 3) et 2 Leporinus frederici sur 3.

- Classe 3/4 : constituée à 92% de détritivores et contient 36% des détritivores. Elle contient 100% des Doras micropoeus, ceux ci formant 61% de la classe.

- Classe 4/4 : constituée à 94% de détritivores et contient 54% des détritivores. Elle contient 100% des Semaprochilodus varii et 85% des Pseudoancistrus barbatus.

Un doute persiste sur la classification de deux espèces

- Les Potamotrygons Hystrix (qui ne sont pas strictement carnivores) sont répartis dans plusieurs classes.
- De même, on ne peut pas, d'après cette classification, affecter les Platydoras Costatus à une classe en particulier.

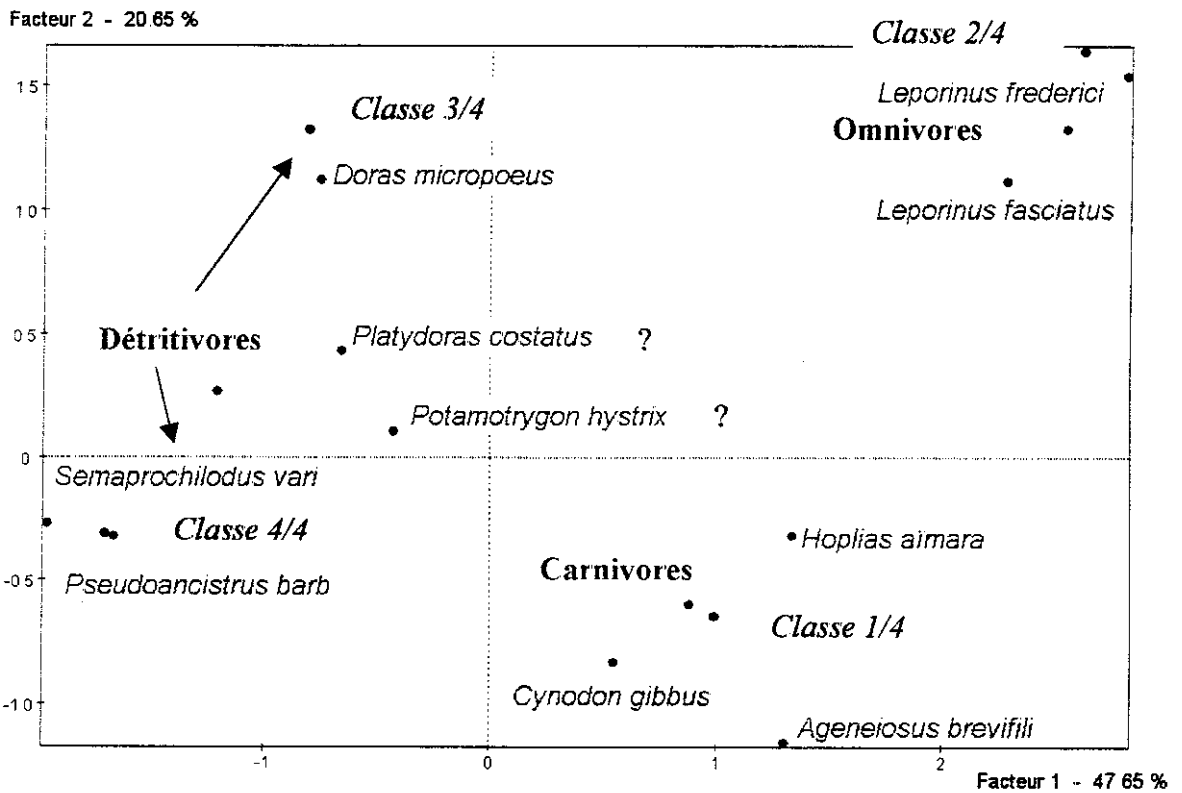


Fig. 1 : Représentation des centres de gravité des modalités de la variable « espèce » et partition en 4 classes.

3.2. Régression PLS et classification hiérarchique

On peut reprocher à cette approche de ne pas tenir compte de l'objectif de l'étude, à savoir classer les modalités de la variable « espèce ». Nous avons donc appliqué la même démarche en remplaçant l'étape d'analyse factorielle par une régression PLS multivariée des 10

indicatrices des modalités de la variable « espèce », sur les 5 variables quantitatives de concentration [WOL83], [TEN98]. On espère ainsi mieux classer les espèces puisque les composantes PLS, calculées avec le logiciel SIMCA-P, ne sont pas construites uniquement pour « bien » résumer les 5 variables quantitatives, mais également pour « bien » expliquer la variable « espèce ». La classification hiérarchique de Ward a ensuite été appliquée avec le logiciel SPAD3.5 sur les composantes PLS et une nouvelle partition en 4 classes des 67 poissons a été obtenue.

La répartition des espèces de poissons dans les classes de cette nouvelle partition étant sensiblement la même que celle de la partition précédente, le doute persiste quand à la classification des espèces Potamotrygon Hystrix et Platydoras Costatus.

Cette approche « classique » sur les poissons a donc été complétée par une approche « symbolique » sur les espèces.

4. Approche symbolique : étude des 10 espèces

Cette approche comprend trois étapes :

- Définition d'un nouveau tableau de données où chaque espèce est décrite en fonction de la concentration dans les organes des poissons de cette espèce. Une approche classique consiste à décrire chaque espèce sur chaque variable par la concentration moyenne des poissons de cette espèce. L'approche symbolique consiste à décrire chaque espèce par un intervalle de valeurs de concentration.
- Classification descendante monothétique de ces 10 espèces.
- Représentation et interprétation graphique des classes obtenues.

Ces trois étapes utilisent trois méthodes, DB2SO, DIV et SOE implémentées dans le logiciel d'Analyse des Données Symboliques développées dans le cadre du projet européen SODAS.

4.1. Description des 10 espèces par des intervalles

La méthode de Généralisation/Spécialisation utilisée pour obtenir la description des 10 classes d'espèces par des intervalles est la méthode DB2SO [STE96], [STE97]. Chaque espèce est alors décrite sur chaque variable par un intervalle de valeur (Tab. 3). Cette méthode permet de réduire l'intervalle le plus généralisant, c'est à dire l'intervalle entre la plus petite valeur

observée et la plus grande, et ce afin d'éliminer les valeurs atypiques ou aberrantes. Afin de laisser libre le choix du niveau de spécialisation, un seuil de recouvrement minimum est fixé par l'utilisateur. Ici, le seuil de recouvrement a été fixé à 80%.

	LN(Foie/muscle)	....	LN(Intestin/muscle)	LN(estomac/muscle
Ageneiosus brevifilis	[-0.8 ; -0 08]	....	[-1.45 ; -0.46]	[-1.48 ; -0.59]
	....	....	....	....
Cynodon Gibbus	[0.12 ; 1.42]	....	[-1.67 ; -0.68]	[-1.61 ; -0.1]

Tab. 3 : Extrait du tableau de données intervalles

4.2. Classification des espèces

La méthode de classification utilisée pour obtenir une partition en 4 classes de ces 10 espèces est une méthode de classification hiérarchique descendante [CHA97], [CHA00]. Elle prend en entrée des données classiques ou des données symboliques. Nous avons appliqué cette méthode sur le tableau de données intervalles (voir Tab. 3) et sur le tableau où les 10 espèces sont décrites par leurs moyennes sur chaque variable de concentration.

Chaque division se fait en fonction d'une variable et d'une dichotomie du domaine d'observation de cette variable. On cherche ainsi parmi toutes les transformations binaires possibles de toutes les variables de l'analyse, celle qui induit la meilleure bipartition au sens d'un critère de type inertie. Le critère utilisé pour mesurer l'homogénéité des classes est une extension du critère d'inertie calculé sur tableau de dissimilarités. La méthode donne en sortie un arbre hiérarchique qui est également un arbre de décision (Figures 2 et 3).

4.2.1. Partition en 4 classes du tableau d'intervalles

Les trois étapes ayant permis la construction de la partition en 4 classes de la figure 2 sont les suivantes :

Étape 1 : séparation des espèces carnivores (sauf Potamotrygon hystrix) et omnivores des espèces détritivores en fonction de la concentration dans le foie par rapport à la concentration dans le muscle.

Étape 2 : séparation des espèces détritivores en deux classes en fonction de la concentration dans l 'estomac par rapport à la concentration dans le muscle.

Étape 3 : séparation des espèces carnivores et omnivores en fonction de la concentration dans les branchies par rapport à la concentration dans le muscle.

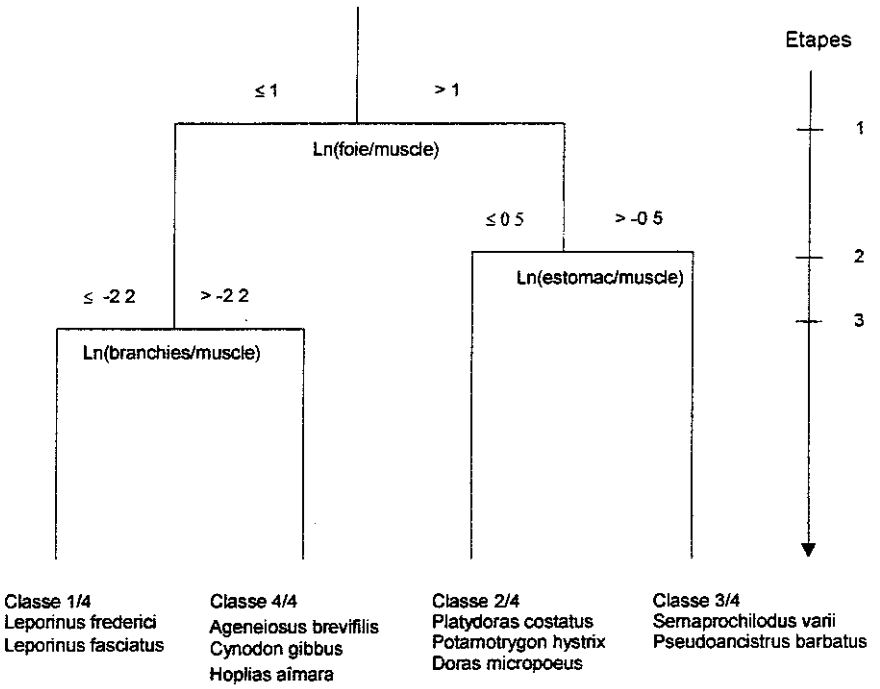


Fig. 2 : Arbre hiérarchique résumant les trois divisions successives ayant induit les partitions en 2, 3 et 4 classes des 10 espèces décrites par des intervalles.

On retrouve ici la classification obtenue dans l'approche classique mais les deux espèces pour lesquelles on avait un doute sont maintenant dans une même classe. Les variables de coupure de l'arbre donné Fig. 2 permettent d'interpréter les classes des partitions en 2 à 4 classes. On peut noter de plus que la première variable de coupure, ln(foie/muscle), est fortement corrélée au premier axe factoriel de l'ACP effectuée dans l'approche classique, tandis que la seconde variable de coupure, ln(branchies/muscle), explique bien le second axe factoriel.

4.2.2. Partition en 4 classes du tableau des moyennes

On a appliqué la méthode DIV au tableau des moyennes c'est à dire le tableau où les 10 espèces sont décrites par les moyennes de la concentration des poissons sur les 5 variables de concentration. La partition en 4 classes, et l'arbre hiérarchique obtenu sont représentés sur la Fig. 3.

On note que la partition obtenue avec les moyennes semble moins cohérente avec les résultats obtenus dans l'approche classique que la partition obtenue avec les intervalles. La première division est identique à celle de la Fig. 2, mais l'espèce Pseudoancistrus barbatus, qui était classée avec l'espèce Semaprochilodus varii dans l'approche classique et dans l'approche divisive sur les données intervalles, est ici séparée des autres espèces dès la seconde division. De même, l'espèce Leporinus fasciatus est séparée lors de la troisème division de l'espèce Leporinus frederici, ce qui n'est pas en accord avec les résultats de l'approche classique.

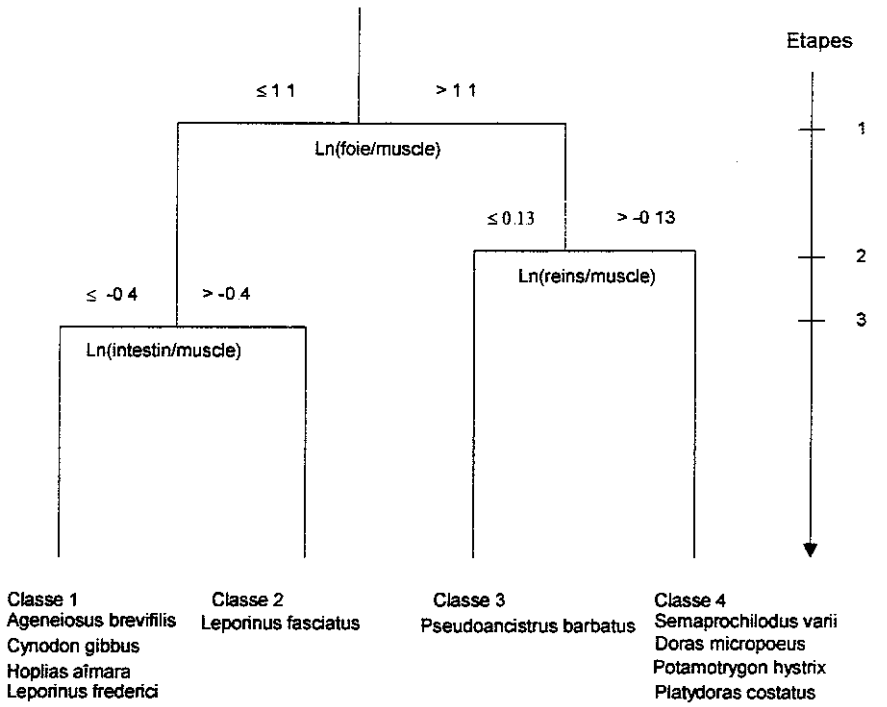


Fig. 3 : Arbre hiérarchique résumant les trois divisions successives ayant induit les partitions en 2, 3 et 4 classes des 10 espèces décrites par les moyennes.

4.3. Interprétation graphique des 4 classes

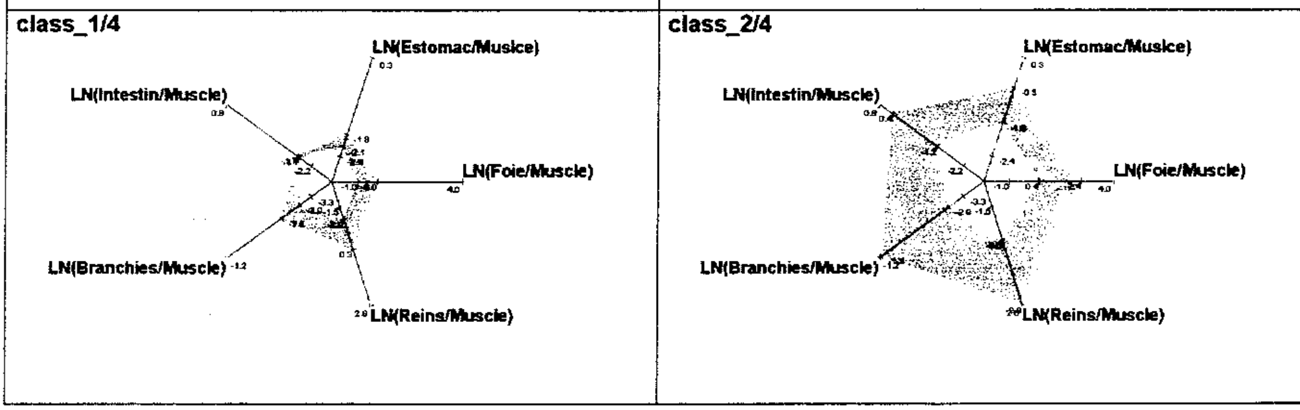
La méthode DB2SO a permis d'obtenir pour chaque classe d'espèces de la partition obtenue avec le tableau des intervalles, une description des classes, résumé dans le tableau suivant :

	LN(Foie/muscle)	....	LN(estomac/muscle)
Classe 1/4	[-0.98 ; -0.04]	....	[-2.11 ; -1.76]
Classe 2/4	[0.41; 2.42]	....	[-1.45 ; -0.49]
Classe 3/4	[1.26 ;3.81]	....	[-0.71 ;0.31]
Classe 4/4	[-0.8 ;1.37]	....	[-2.36; -0.59]

Tab. 4 : Description des classes de la Fig. 2 par des intervalles

A partir de ce tableau, chaque classe est visualisée (Fig 4.) à l'aide de l'éditeur d'objets symboliques SOE [NOI97]. Chaque classe est ainsi représentée par une « étoile », dont les branches correspondent aux variables de l'analyse, et les zones sombres relient les intervalles de variations sur chaque variable. Cette représentation en forme d'étoile, permet entre autre de comparer les classes sur plusieurs variables simultanément. On note par exemple que la classe 1/4 composée des deux espèces de Leporinus, qui était caractérisée d'après l'arbre hiérarchique de la Fig. 2 par de faible concentration dans le foie et dans les branchies, est également caractérisée par de faibles valeurs de concentration dans les autres organes. Ces représentations graphiques permettent également de comparer la variabilité des classes.

Leporinus fasciatus Leporinus frederici	Platydoras costatus Potamotrygon hystrix Doras micropoeus



Semaprochilodus varii Pseudoancistrus barbatus	Ageneiosus brevifilis Cynodon Gibbus Hoplias aimara

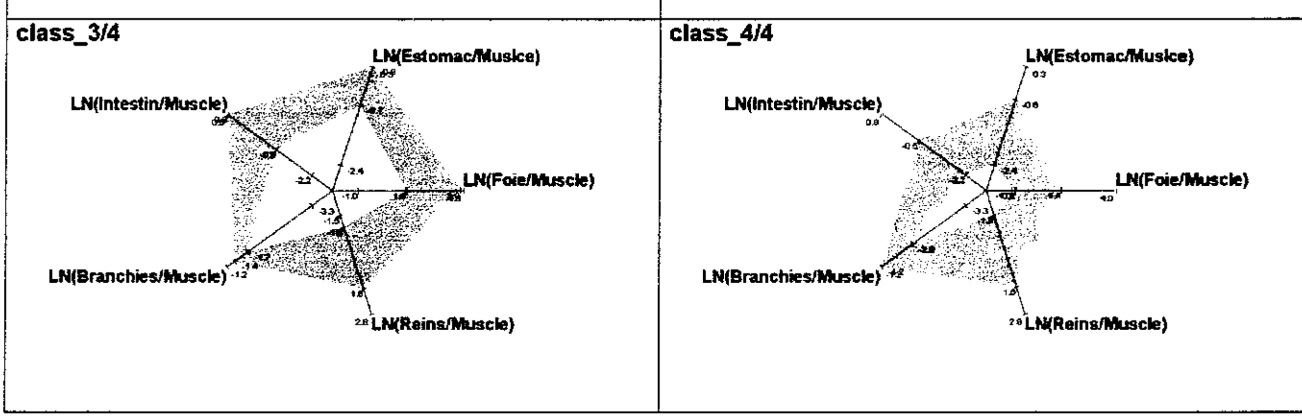


Fig. 4 : Représentation graphique des descriptions des quatre classes d'espèces

5. Conclusion

Cette étude est un préliminaire à une recherche plus approfondie et plus explicative sur les transformations et les transferts du mercure le long des chaînes alimentaires aquatiques.

Elle permet déjà cependant grâce aux techniques de classification utilisées ici de comparer et de mieux connaître les différentes espèces étudiées en terme de répartition de concentration en mercure dans les différents organes.

De nouvelles campagnes de pêche doivent nous fournir des échantillons de plus grande taille incluant :

- les herbivores pour lesquels nous n'avons pas encore obtenu de résultats probants
- des données supplémentaires concernant les différentes formes du métal dans les tissus prélevés.

L'utilisation de méthodes de classification analogues à celles présentées devrait permettre :

- De déterminer l'importance et la localisation au sein des biotopes des réactions de transformation de la forme élémentaire du métal ainsi que les voies de transfert des formes chimiques entre les biotopes (colonnes d'eau, sédiments) et les êtres vivants (bio-accumulation et bio- amplification).
- D'étudier, selon les espèces, au sein d'un même régime alimentaire, les disparités de contaminations liées aux types d'alimentation ; en particulier pour les herbivores.
- D'étudier au sein d'une même espèce les disparités de contamination liées aux types de croissance des individus et pouvoir appréhender ainsi l'âge des poissons tropicaux.
- D' expliquer des comportements alimentaires encore mal connus de certaines espèces.
- D'étudier l'influence des lieux et des saisons de capture.

RÉFÉRENCES BIBLIOGRAPHIQUES

[BOU98] Boudou A. & Ribeyre F. Mercury in the food web: accumulation and transfer mechanisms. In "Mercury and its effects on Environment and Biology". Metal Ions in biological Systems, Sigrel A. and Sigel H. edits, M. Dekker (NY) : 289-319, 1998.

[BOU99] Boudou A., Deheeger M., Frery N., Maillot E., Maury-Brachet R. & Merona (de) B. Relationships between mercury levels in fishs and humans in french Guyana (Wayanas natives communities), International Congress "Mercury as a global polluant" Rio (Brésil) Mai 1999.

[CHA97] Chavent M. Analyse des données symboliques, une méthode divisive de classification, Thèse de l'Université Paris IX-Dauphine, 1997.

[CHA00] Chavent M. Criterion-Based Divisive Clustering for Symbolic Objects, Analysis of symbolic data, eds. H.H.Bock, E. Diday, Springer, 2000.

[COR88] Cordier, S & all. Mercury exposure in French Guiana: levels and determinant . Archives of environmental Health 53 (4) : 299-303, 1988.

[DID99] E. Diday, An introduction to symbolic data analysis and its application to the SODAS project , Cahiers du CEREMADE, N° 9914, Mars 1999.

[LEB96] Lebart L.,Morineau A., Lambert T.& Pleuvret P. SPAD version 3. Système pour l'Analyse des données . CISIA, Montreuil, 1996.

[NOI97] Noirhomme M., Rouard M., Computer Graphics for Symbolic Objects, VIII International Symposium on Applied Stochastic Models and Data Analysis, Anacapri, , p. 325 :331, June 1997.

[STE96] Stéphan V., Exctracting Symbolic Objects from Relational Databases, 7th international workshop on Database and Expert Systems Applications (DEXA), éd. WagnerR.R., Zurich, p. 414:419, sept 1996.

[STE97] Stéphan V., Construction d'objets symboliques par synthèse des résultats de requêtes SQL, Thèse de l'Université Paris IX-Dauphine, 1997.

[TEN98] Tenenhaus M. La régression PLS, Théorie et pratique, TECHNIP, 1998.

[UME96] Umetri AB SIMCA-P for Windows, Graphical Software Process Modeling, Umetri AB, Box 7960, S-90719 Umea, Sweden, 1996.

[WOL83] Wold et coll. Pattern Recognition : Finding and using Regularities in Multivariate Data, in "Food Research and Data Analysis", ed. Martens J., Applied Science Publications, London, 1983.