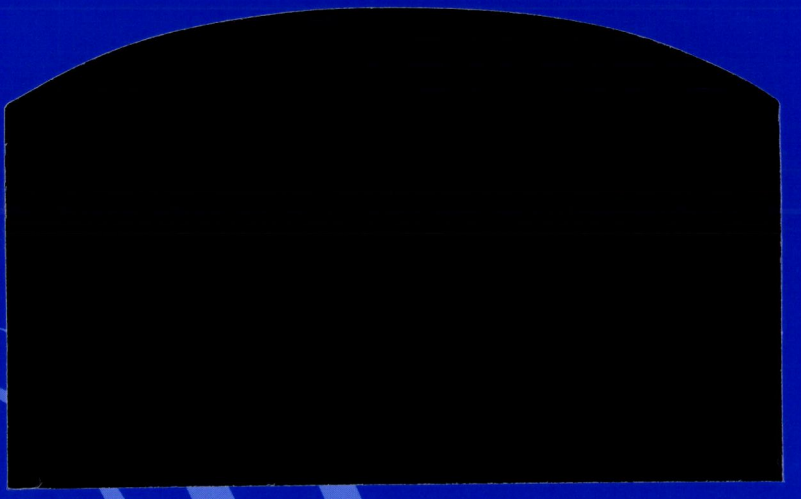
CENTRE SAINT-LAURENT
ST. LAWRENCE CENTRE



Environnement Canada
Conservation de l'environnement

Environment Canada
Environmental Conservation

RAPPORT SCIENTIFIQUE ET TECHNIQUE

DÉVELOPPEMENT D'UN INDICE INTÉGRATEUR DE DANGER POUR LA FAUNE AQUATIQUE AU MOYEN D'UNE BATTERIE DE BIOMARQUEURS

Rapport ST-217

Développement d'un indice intégrateur de danger pour la faune aquatique au moyen d'une batterie de biomarqueurs

Dr Nathalie Chèvre
Toxicité aquatique
Section biologie de l'environnement

Centre Saint-Laurent
Conservation de l'environnement
Environnement Canada

Février 2001

COMMENTAIRES DES LECTEURS

Veuillez adresser vos commentaires sur le contenu du présent rapport au Centre Saint-Laurent, Conservation de l'environnement, Environnement Canada – Région du Québec, 105, rue McGill, 7e étage, Montréal (Québec), H2Y 2E7.

On devra citer la publication comme suit :
Chèvre N. 2001. *Développement d'un indice intégrateur de danger vis-à-vis de la faune aquatique au moyen d'une batterie de biomarqueurs.* Environnement Canada - Région du Québec, Conservation de l'environnement, Centre Saint-Laurent. Rapport scientifique et technique ST-217, 84 pages.

N° de catalogue En152-1/217-2001F
ISBN 0-662-85481-0

Perspective de gestion

Ce rapport est publié dans le cadre du plan d'action fédéral-provincial Saint-Laurent Vision 2000 (SLV 2000). Le développement d'un indice intégrateur de danger pour la faune aquatique fournit un outil intéressant aux gestionnaires dans une optique de protection et de conservation du milieu aquatique. Cet indice peut en effet permettre de discriminer rapidement entre les sites les plus contaminés et ainsi de prendre des mesures correctrices appropriées.

Management Perspective

This document is published as part of the federal-provincial St. Lawrence Vision 2000 action plan. The integrative hazard index for aquatic fauna provides managers with a sound tool for protecting and conserving the aquatic environment. The index rapidly discriminates among the most contaminated sites so that appropriate remedial measures may be taken.

Remerciements

Je tiens à remercier chaleureusement les personnes que j'ai côtoyé pendant ce stage post-doctoral et qui m'ont apporté leur aide au niveau scientifique, mais aussi au niveau personnel, me permettant de découvrir un peu mieux le Québec et les québecois.

Je tiens particulièrement à remercier :

- Dr Christian Blaise qui m'a accueilli dans son équipe pendant cette année au CSL;

- Dr François Gagné, qui m'a aidé pour la partie biomarqueurs;

- Pierre Gagnon, qui m'a été d'un grand secours pour toute la partie statistique;

- Les Professeurs Roman Slowinski et Szymon Wilk, à Poznan (Pologne) qui m'ont apporté leur aide pour l'application de la théorie des ensembles approximatifs;

- Le groupe d'écotoxicologie du Centre Saint-Laurent : Manon Harwood, Sylvain Trottier, Brian Walker, Christine Girard et Geneviève Farley pour leurs conseils judicieux mais également pour leur bonne humeur qui a contribué à mon agréable séjour au Canada.

Je remercie enfin le Fonds national suisse pour la recherche qui a financé mon stage post-doctoral.

Résumé

Mots-clés: écotoxicité aquatique, indice intégrateur, danger, biomarqueurs

Les problèmes liés à la pollution de la biosphère prennent de plus en plus d'importance depuis quelques décennies. Il ne se passe en effet pas une journée sans que l'on entende parler de problématique environnementale : produits à effets œstrogéniques, gaz à effet de serre, etc. Face à ces dangers, il convient de développer des outils toujours plus performants pour assurer la protection et la conservation des écosystèmes. L'utilisation de biomarqueurs comme indicateurs d'effets écotoxiques est relativement récente mais présente l'avantage de mettre en évidence des effets avant qu'ils ne soient observables au niveau des organismes eux-mêmes. Ils peuvent ainsi être considérés comme un signal d'alarme vis-à-vis d'une pollution.

Le but de ce travail est de construire un indice qui permettrait d'intégrer les données d'une batterie de biomarqueurs afin d'estimer le potentiel toxique d'un milieu aquatique pour une population qui y vit. À cette fin, les organismes choisis sont des mollusques bivalves (*Mya arenaria*), prélevés dans sept sites le long du fjord du Saguenay (Québec, Canada) et sur lesquels sept biomarqueurs ont été mesurés (métallothionéines, activité du cytochrome P4501A1, génotoxicité, peroxydation des lipides, phagocytose, activité non spécifique des estérases).

La première partie de ce travail a consisté à évaluer la discrimination entre les sites concernés. Trois méthodes ont été utilisées : une méthode paramétrique – analyse discriminante, et deux méthodes non paramétriques – théorie des ensembles approximatifs et arbres hiérarchiques. En effet, les données ne satisfaisant pas à toutes les hypothèses pour l'application de l'analyse discriminante, il paraissait intéressant de comparer les résultats avec des méthodes non paramétriques.

Les résultats montrent que quatre sites sur sept se discriminent bien, les trois autres étant plus difficiles à distinguer. Les résultats obtenus avec l'analyse discriminante et la théorie des ensembles approximatifs sont comparables et la qualité de la classification est excellente, puisqu'elle atteint 90 %. En revanche, les arbres hiérarchiques donnent une classification de moins bonne qualité. La théorie des ensembles approximatifs présente des avantages : elle est indépendante de la distribution des données, elle permet de tenir compte de leur incertitude

(utilise des classes) ce qui permet d'inclure des jugements subjectifs, et surtout elle permet une caractérisation des différents sites étudiés. C'est donc une méthode séduisante qui peut être vue comme une alternative à l'analyse discriminante.

La deuxième partie concerne la construction de l'indice proprement dit. Deux méthodes ont été choisies pour normaliser les données : le calcul d'un ratio avec un site de référence et l'utilisation des classes définies par la théorie des ensembles approximatifs. Le calcul de l'indice proprement dit représente la somme des ratios ou des classes des différents biomarqueurs pour chaque site. Dans les deux cas, l'indice donne des résultats cohérents, les sites les plus pollués ayant des valeurs différentes (plus élevées ou plus faibles) des sites sans contamination directe. Néanmoins, le choix d'un site de référence pour le ratio pose un problème (quel site choisir? généralisation de l'indice?). L'utilisation de classes est donc préférée au ratio. Afin de généraliser l'utilisation de cet indice (pour pouvoir par exemple comparer différentes études entre elles), il conviendrait d'optimiser la batterie de biomarqueurs utilisée. Un effort devrait également être fait pour valider les classes ou pour les définir *a priori*. La question de la prise en compte de la variation temporelle des biomarqueurs se pose également.

Abstract

Key-words: aquatic ecotoxicity, integrated index, hazard, biomarkers

The problem of environmental pollution has grown considerably over the last few decades. Indeed, every day brings another environmental problem to our attention: the estrogenic effects of certain products, the impact of global climate change, etc. To meet these challenges, new, more effective tools for protecting and conserving ecosystems are required. While biomarkers have only relatively recently been employed as an indicator of environmental toxicity, they have the advantage of pointing up effects before they are seen at the level of the organism itself. In this respect, biomarkers might be considered a warning sign of pollution.

The aim of this project was to develop an index for use in integrating data from a battery of biomarkers to assess the potential toxicity of an aquatic medium to a population that lives therein. In this case, bivalve clams (*Mya arenaria*) were selected as the test organism. Specimens were collected at seven sites along the Saguenay Fjord, in Quebec, Canada, and measured for seven biomarkers: metalothionein, cytochrome P4501A1 activity, DNA strand breakage evaluation, lipid peroxidation, vitellin, phagocytosis, and activity of non-specific esterase.

The first part of the study consisted of determining discrimination between the different sites. Three methods were used to this end, one parametric (discriminant analysis), and two non-parametric (rough set analysis and classification tree analysis). Indeed, because the data did not satisfy all the hypotheses needed for a discriminant analysis, we compared the results using non-parametric methods.

The results show that four sites are well discriminated, but the three others are more difficult to distinguish. The results of discriminant and rough set analyses are comparable, and the quality of classification is good, reaching 90%. By contrast, the quality of classification tree analysis is poor. Rough set analysis has some advantages: it doesn't depend on data distribution, it takes the vagueness of the data into account by using classes — which makes subjective evaluation possible — and, most especially, it allows for the characterization of study sites. This method is therefore an appealing alternative to discriminant analysis.

The second part of the study is the development of the index proper. Two methods were chosen to standardize biomarker values: a ratio with a reference site, and the classes generated by the rough set analysis. Calculation of the index is the sum of the ratio or classes of the different biomarkers for each site. In both cases, this index provides coherent results, the most polluted sites offering relatively different values from the sites with no direct contamination. Nonetheless, the choice of a reference site for the ratio does present a few problems in terms of site selection and generalizing of the results. As such, the use of classes is preferable. To generalize use of the index (e.g. to compare different studies), a battery of biomarkers should be optimized. Further, an attempt should be made to validate the classes or define them *a priori*. The issue of temporal variation of biomarkers should also be taken into account.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ADN	biomarqueur mesurant les dommages à l'ADN
ASE	site Anse de Saint-Etienne
ASJ	site Anse Saint-Jean
Barq	site Anse à la Barq
Baude	site Baie-du-Moulin à Baude
BE	site Baie Éternité
Era	site Anse aux Érables
EROD	biomarqueur mesurant l'activité du cytochrome P4501A1
LPO	biomarqueur mesurant la peroxydation des lipides
mg	milligramme
µg	microgramme
min	minute
MT	biomarqueur mesurant les métallothionéines
nmole	nanomole
NspE	biomarqueur mesurant l'activité non spécifique des estérases
PHAG	biomarqueur mesurant la phagocytose
PS	site Petit Saguenay
Vn	biomarqueur mesurant la vitélline
TEA	théorie des ensembles approximatifs
σ	écart type
$\bar{x}$	moyenne
<	plus petit que
>	plus grand que

1 Introduction

L'urbanisation et le développement industriel des zones se trouvant le long des cours d'eau ont introduit des centaines de produits toxiques dans le milieu aquatique au cours du XX^{ème} siècle. Avec le temps, il s'est avéré que ces polluants induisaient des effets nocifs sur l'écosystème aquatique, entraînant, à plus ou moins long terme, la diminution de la diversité biologique occasionnée par la diminution des populations les plus sensibles. Afin de protéger notre environnement, des tests écotoxicologiques sont appliqués depuis quelques années afin de déterminer, entre autres, la toxicité d'eaux résiduaires. Complémentaires à l'analyse chimique, ces tests permettent de mettre rapidement en évidence un danger pour la faune et la flore d'un milieu donné.

L'intérêt de ce type d'information est particulièrement mis en évidence lorsque l'on est capable d'intégrer l'ensemble des mesures écotoxicologiques en un tout cohérent. La complexité de l'interprétation des données individuelles se traduit alors par une seule valeur qui s'avère utile pour la compréhension du danger d'une pollution à l'égard d'un écosystème. Cette valeur peut ensuite être utilisée pour faciliter la prise de décision quant aux mesures correctrices à prendre pour améliorer cet écosystème. Ainsi, l'indice PEEP permet d'estimer le potentiel toxique associé aux effluents industriels (Costan *et al.*, 1993) et l'indice SED-TOX, celui associé aux sédiments dulçaquicoles (Bombardier et Bermingham, 1999).

L'utilisation de biomarqueurs comme indicateurs d'effets écotoxicologiques est relativement récente. L'avantage de ces mesures est qu'elles permettent de mettre en évidence des effets au niveau suborganismal, soit bien avant que ceux-ci ne soient observables au niveau de l'organisme lui-même. L'utilisation d'une batterie de biomarqueurs, regroupant des effets de génotoxicité, d'immunotoxicité ou encore d'endocrinotoxicité, permet d'avoir une idée globale de la toxicité d'un milieu contaminé. Blaise *et al.* (à publier) et Bresler *et al.* (1999) ont appliqué une telle approche, les premiers sur le fjord du Saguenay, les deuxièmes sur différents littorals européens. Les données de ces études sont cependant assez complexes à interpréter. Il nous paraissait donc intéressant d'essayer de développer un indice qui permettrait de regrouper toutes

les mesures en une seule valeur qui pourrait alors être utilisée pour évaluer le danger du milieu étudié.

Les données utilisées pour la construction de l'indice ont été recueillies par Blaise *et al.* (à publier). La première partie du travail a consisté à évaluer la discrimination entre les sites concernés. En effet, si les sites ne se distinguaient pas bien, il était vain de chercher à construire un indice. Dans cette optique, différentes méthodes ont été considérées. L'analyse discriminante est la plus classique mais elle est limitée par des hypothèses telles que la normalité des données ou l'homogénéité des variances/covariances. Deux méthodes non paramétriques – la théorie des ensembles approximatifs et les arbres hiérarchiques – ont également été utilisées et les résultats comparés avec ceux de l'analyse discriminante.

La deuxième partie du travail concerne la construction de l'indice. Différents essais ont été effectués, notamment pour la normalisation des valeurs et pour la pondération.

2 Données de biomarqueurs

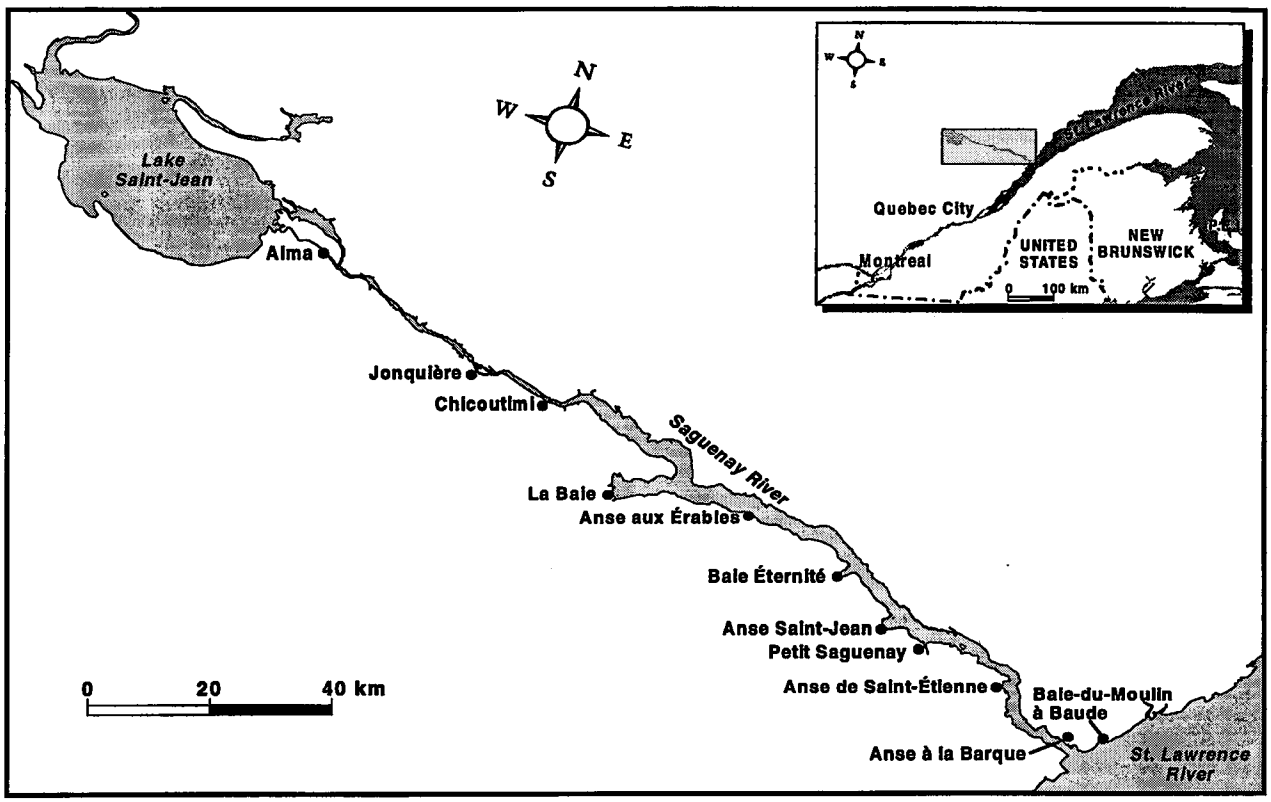
2.1 DONNÉES À DISPOSITION

Les données à disposition pour ce travail ont été mesurées sur des myes (*Mya arenaria*) prélevées dans la rivière Saguenay (Québec, Canada). Les myes sont des mollusques bivalves qui filtrent l'eau pour se nourrir et respirer. Elles sont donc particulièrement susceptibles d'être exposées aux polluants présents dans les eaux de surface et les matières en suspension. De plus, ces organismes sont en contact constant avec les sédiments, milieu dans lequel les contaminants peu solubles se concentrent. Les mesures d'effets précoces des contaminants sur ces individus pourraient donc permettre de mettre en évidence des dangers toxicologiques bien avant que l'effet ne puisse être observé au niveau de l'organisme lui-même, ceci afin d'identifier les sites qui constitueraient un risque pour la survie de la population.

Sept sites de prélèvements ont été choisis, six le long du fjord du Saguenay et un en dehors de cet écosystème, sur le Saint-Laurent (figure 1). Les quatre sites en amont – Anse aux Érables (région industrielle), Baie Éternité (quai), Anse Saint-Jean (rejets urbains), Petit Saguenay (quai) – se trouvaient près de sources de contamination connues. Les deux sites en aval, soit ceux de l'Anse Saint-Étienne et de l'Anse à la Barque n'étaient soumis à aucune contamination directe. Le site Baie-du-Moulin, à Baude, est quant à lui situé sur le Saint-Laurent et ne présente pas de sources de contamination spécifique. À chacun de ces sites, 15 myes ont été prélevées, ce qui représente un total de 105 échantillons.

Sept biomarqueurs ont été choisis pour être mesurés sur les myes prélevées dans les différents sites. Il s'agit tout d'abord des métallothionéines (MT) et de l'activité de 7-éthoxyrésorufine-O-déséthylase (EROD) qui sont des biomarqueurs de défense. Puis le dommage à l'ADN (ADN), la peroxydation des lipides (LPO), la vitélline (Vn), la phagocytose (PHAG) et l'activité non spécifique des estérases (NspE), qui peuvent être considérés comme des biomarqueurs d'effet sous l'action de contaminants chimiques ou biologiques.

Le tableau 1 présente les moyennes et les erreurs standards des différents biomarqueurs pour chaques sites. Les résultats complets sont rapportés à l'annexe 1.



Remarque : Quatre sites sont soumis à une contamination : Anse aux Érables (région industrielle), Baie Éternité (quai), Anse Saint-Jean (rejets urbains), Petit Saguenay (quai) et trois sites ne subissent pas de contamination directe : Anse de Saint-Étienne, Anse à la Barque, Baie-du-Moulin à Baude. Ce dernier se situe sur le Saint-Laurent.

Figure 1 Sites de prélèvements le long du fjord du Saguenay

Tableau 1
Moyennes et erreurs standards des différents biomarqueurs pour chaque site étudié

Sites	MT[1]	EROD[2]	ADN[3]	LPO[4]	Vn[5]	PHAG[6]	NspE[7]
Era	2,13	2,40	633	56,0	380	1,3	7,8
	± 0,08	± 0,55	±33	± 5,0	± 82	± 0,1	± 0,6
BE	0,85	0,52	1185	7,0	253	2,5	4,6
	± 0,07	± 0,12	±70	± 0,5	± 46	± 0,7	± 1,0
ASJ	1,82	0,02	331	10,1	761	1,3	15,0
	± 0,07	± 0,01	±20	± 0,9	± 80	± 0,1	± 1,0
PS	1,36	0,80	308	3,6	219	0,9	1,5
	± 0,07	± 0,10	± 24	± 0,3	± 79	± 0,1	± 0,1
ASE	1,60	1,50	1300	6,0	292	0,7	1,9
	± 0,12	± 0,30	± 41	± 0,2	± 13	± 0,1	± 0,1
Barq	0,58	1,10	636	9,5*	250	0,7	4,5
	± 0,06	± 0,23	± 172	± 0,7	± 17	± 0,1	± 0,3
Baude	0,61	1,14	1258	11,0	288	0,7	2,3
	± 0,04	± 0,55	± 47	± 0,6	± 24	± 0,1	± 0,2

1. MT : nmole de MT/mg protéines.
2. EROD : 7-hydroxyrésorufine formé min/mg protéines.
3. ADN : µg d'ADN dans le surnageant/mg protéines.
4. LPO : µg TBA réagissant/mg protéines.
5. Vn : µg phosphate dans la phase organique/mg protéines.
6. PHAG : µg fluorescéine (bactéries ingérées)/mg protéines.
7. NspE : µg fluorescéine/min/mg protéines.

Métallothionéines

L'exposition d'organismes tels que les poissons à des métaux lourds comme le cuivre ou le cadmium donne lieu à la synthèse de métallothioénines (MT) dans les tissus hépatiques, rénaux et des branchies. Les MT sont des protéines de faible poids moléculaire, riches en cystéine et omniprésentes dans les organismes vivants. Elles exercent une fonction défensive lors d'exposition à des métaux lourds car elles peuvent se lier à ceux-ci. Elles préviennent ainsi les liaisons avec des groupes sulfhydryles d'autres protéines importantes et limitent ainsi la toxicité des métaux.

L'affinité des MT pour les métaux varie, le mercure, le cuivre et le cadmium ayant l'affinité la plus élevée. Il faut cependant noter que des effets pathologiques des métaux peuvent être détectés avant que les MT ne soient excédées (Thomas, 1990). Notons également que les MT peuvent être induites par d'autres stress tels que l'inflammation et le stress oxydatif (Blaise *et al.*, à publier).

Activité du cytochrome P4501A1

Le cytochrome P4501A1 (EROD) est un enzyme majeur dans les processus de biotransformation (détoxification et toxification) des xénobiotiques organiques tels que les hydrocarbures aromatiques polycycliques (comme les HAPs et les biphényles polychlorés ou BPC) et les pesticides. Sa présence est donc généralement liée à une pollution par un composé organique mais il faut cependant noter qu'il peut y avoir des changements liés à la saison et au sexe (Thomas, 1990). L'enzyme étudié ici est le cytochrome P4501A1 qui catalyse les réactions d"hydroxylation des HAP.

Génotoxicité ou dommages à l'ADN

Les dommages à l'ADN surviennent lorsque des adduits de xénobiotiques se forment sur l'ADN et lorsque les mécanismes de réparation sont altérés. Les dommages peuvent être liés à différents types de pollution (i,e, non spécifique à une classe de contaminants comme dans le cas de la MT et du cytochrome P4501A1) et peuvent conduire à des mutations, au cancer ou encore à la mort de la cellule (Blaise *et al.*, à publier). L'effet mesuré ici est en fait le nombre de cassures (brins simples) d'ADN qui augmente généralement avec la réparation de l'ADN endommagé.

Peroxydation des lipides

La peroxydation des lipides (LPO) est associée à l'oxydation des lipides polyinsaturés par les radicaux libres tels que l'ion superoxyde dans les membranes biologiques. Il en résulte une perturbation de la membrane cellulaire et une perte de l'activité des enzymes qui y sont associés, Le dommage oxydatif pourrait être un mécanisme universel de la toxicité des xénobiotiques, mais plusieurs facteurs, tels que la nutrition et l'âge, peuvent également influencer ce paramètre (Thomas, 1990).

Vitélline

La vitélline (Vn) est une glycoliphosphoprotéine majeure chez les organismes ovipares. Elle constitue la principale source d'énergie pour le développement de l'embryon. Sa synthèse est sous le contrôle des récepteurs estrogéniques qui peuvent être modulés par des contaminants environnementaux. La baisse ou l'augmentation de la Vn peut induire des effets sur la mère et l'embryon. La vitellogénèse a normalement lieu durant le développement saisonnier des oocytes qui précède généralement la fertilisation et la ponte. Cependant, ce processus peut être activé autant chez les mâles que les femelles par des substances ayant une affinité pour le récepteur aux oestrogènes. Par exemple, le nonylphénol est un produit de dégradation des surfactants appartenant à la classe des polyéthoxylates et peut activer la synthèse de la Vn chez le poisson et la mye.

Phagocytose

Chez les myes, la phagocytose (PHAG) est assurée par les hémocytes de l'hémolymphe, cellules qui occupent les principales fonctions immunitaires de ces invertébrés. Une baisse de la phagocytose peut être provoquée par un xénobiotique et suggère donc une atteinte au système immunologique (Blaise *et al.*, à publier).

Activité non spécifique des estérases

L'activité non spécifique des estérases (NspE) augmente lors d'infections bactériennes ou en présence de parasites dans les hémocytes et dans l'hémolymphe. Cette mesure représente donc l'état métabolique de ces tissus et donne une indication sur l'état du système immunologique (Blaise *et al.*, à publier).

Les biomarqueurs PHAG, ADN et NspE ont été déterminés dans les hémocytes et le biomarqueur Vn dans l'hémolymphe. Les biomarqueurs MT et EROD l'ont été dans la glande digestive.

2.2 DONNÉES UTILISÉES

Le tableau 2 présente le début de la table de décision des biomarqueurs mesurés sur les différents sites de la rivière Saguenay (la table complète se trouve à l'annexe 1). Sept biomarqueurs ont ainsi été mesurés sur des organismes provenant de sept sites différents. Quinze organismes ont été mesurés par site pour un total de 105 individus.

Tableau 2
Table de décision des biomarqueurs mesurés sur différents sites du Saguenay*

MT	EROD	ADN	LPO	Vn	PHAG	NspE	SITES
2,07	1,89	1513,0	5,65	194,60	0,46	1,68	ASE
2,23	?	1091,2	7,17	305,20	0,69	1,74	ASE
0,91	?	1493,00	6,33	282,50	1,07	2,04	ASE
1,62	0,67	1241,38	7,05	214,06	1,00	2,64	ASE
1,16	0,23	1398,34	5,94	208,18	0,82	2,59	ASE
...	...	...	...	...	...	...	...

* Seuls les cinq premiers éléments sont décrits ici, la table complète se trouve à l'annexe 1.

Le but de l'analyse était d'évaluer comment les sites se discriminaient entre eux et s'il était possible de trouver la provenance d'un individu sur la base des biomarqueurs mesurés. Dans cette optique, les hypothèses de travail suivantes ont été formulées :

- Dans le cas de l'EROD, il manque 45 données sur 105, du fait que de nombreux échantillons présentaient des mesures non détectables. L'analyse a été faite de deux façons, soit avec le biomarqueur EROD avec seulement les 60 données connues et sans les valeurs du biomarqueur EROD. En effet, le fait de remplacer les valeurs manquantes introduit un biais, tout particulièrement dans le cas du site ASJ pour lequel on n'a que trois valeurs connues sur 15.
- Les données manquantes isolées (une pour Vn et une pour ADN) sont remplacées par la valeur médiane du site auquel elles appartiennent.

- Les valeurs aberrantes ont été déterminées sur la base du calcul suivant :

  Soit : x : valeur testée

  MED : valeur médiane du site

  MAD : mesure de dispersion robuste (Median Absolute Deviation)

  x est considérée comme valeur aberrante si :

  $$|(x\text{-}MED)/MAD| > 6 \qquad (1)$$

  Six valeurs aberrantes ont été mises en évidence. Suivant les données listées à l'annexe 1, elles se rapportent aux :

  18^{e} échantillon : site BE, valeur aberrante pour Vn
  42^{e} échantillon : site Baude, valeur aberrante pour ADN
  80^{e} échantillon : site Barq, valeur aberrante ADN
  90^{e} échantillon : site Barq, valeur aberrante MT
  92^{e} échantillon : site PS, valeur aberrante LPO
  99^{e} échantillon : site PS, valeur aberrante Vn.

  Il reste donc 99 données pour effectuer l'analyse discriminante

- Les différents sites ont tout d'abord été considérés séparément pour l'analyse, puis les sites supposés « peu pollués » ont été groupés.

3 Méthodes statistiques

3.1 THÉORIE DES ENSEMBLES APPROXIMATIFS

La théorie des ensembles approximatifs (TEA) a été introduite par Pawlak (1982). Cette théorie permet d'analyser et de classer des données en tenant compte de leur imprécision. Elle présente également l'avantage d'être indépendante de la distribution des données. Proposée comme alternative aux méthodes classiques telles que l'analyse discriminante (Krusinska *et al.*, 1992), la TEA a été utilisée avec succès pour différents ensembles de données, notamment dans les domaines médical (Fibrak *et al.*, 1986; Pawlak *et al.*, 1986; Slowinski *et al.*, 1989), financier (Slowinski *et al,* 1994) et des eaux usées (Rossi *et al.*, 1999).

Les caractéristiques de cette approche la rendent très intéressante pour l'analyse des données de biomarqueurs puisque les études sur le terrain sont susceptibles de générer des résultats qui ne sont pas normalement distribués (i,e, non paramétriques).

Le logiciel ROSE2, un logiciel très convivial, téléchargeable sur internet (http://www-idss.cs.put.poznan.pl/software/rose/) permet l'application de cette théorie.

3.1.1 Terminologie de la théorie des ensembles approximatifs

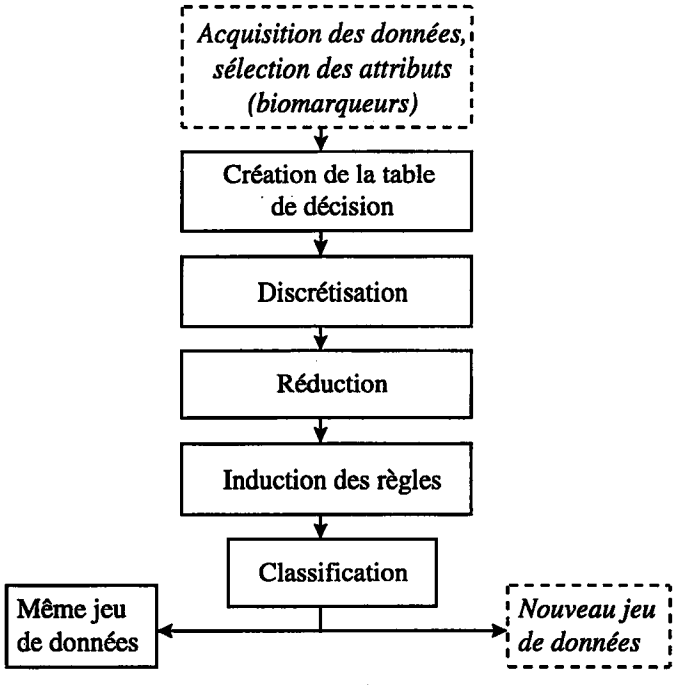
La TEA utilise un vocabulaire spécifique qu'il convient de connaître avant son application. Les données de base utilisées dans le cadre de cette approche se présentent sous la forme d'une ***table d'information***, parfois appelée « table de décision ». Un exemple est donné au tableau 3.

Les lignes de la table correspondent aux différents mollusques sur lesquels les biomarqueurs ont été mesurés et qui sont appelés ***objets***. Les propriétés de chaque objet sont perçues par le biais d'***attributs conditionnels*** et d'un ou de plusieurs ***attributs décisionnels***. Dans l'exemple ci-dessus, les attributs conditionnels représentent les mesures des différents biomarqueurs des organismes et l'attribut décisionnel correspond au site d'où ils proviennent.

Tableau 3
Exemple de table d'information

Biomarqueurs							Site
MT	EROD	ADN	LPO	Vn	PHAG	NspE	
2,07	1,89	1513	5,65	194,6	0,46	1,68	ASE
0,66	0,28	1575,86	8,01	262,2	0,33	4,25	BE
0,52	0,34	675,89	10,53	269,79	0,88	3,25	Baude
1,06	0,01	306,25	7,85	812,95	1,36	12,89	ASJ
Attributs conditionnels							Décision

L'application de la TEA nécessite plusieurs étapes : discrétisation, formation des atomes, recherche des redondances, génération des règles, classification et validation (figure 2).



Remarque : Les étapes entourées de traitillés ne sont pas traités dans ce travail.

Figure 2 Étapes relatives à l'application de la théorie des ensembles approximatifs (d'après Rossi *et al.* 1999)

3.1.2 Discrétisation

La théorie des ensembles approximatifs nécessite que toutes les données (appelées attributs) soient sous forme discrète. Les données continues doivent donc être discrétisées avant le processus d'induction des règles.

Le processus de discrétisation ne fait pas partie de la théorie des ensembles approximatifs. Cependant, il s'agit d'une étape essentielle qui peut faire varier les résultats de l'analyse suivant la manière dont elle est réalisée (Dougherty *et al.*, 1995). La discrétisation « granule » le domaine continu des différents attributs, ce qui est bénéfique pour l'induction de règles courtes et fortes. La discrétisation ne doit cependant pas être vue comme une perte de précision. En effet, les données sont en général assez variables et la discrétisation permet d'appliquer l'analyse sans être trop influencée par ces dernières.

Il existe plusieurs manières de discrétiser des données continues. La première, et la plus évidente, est de créer soi-même les classes sur la base de ses connaissances, par exemple : niveau faible, moyen et élevé.

Il n'est cependant pas toujours facile d'établir un tel classement. Il est alors possible d'utiliser différents algorithmes permettant la transformation des données continues en données discrètes.

Le logiciel ROSE2 permet trois types de discrétisation :

- locale;
- locale supervisée;
- globale.

Discrétisation locale et locale supervisée

Ce processus traite un attribut à la fois et doit donc être appliqué consécutivement à tous les attributs. L'algorithme utilisé est un algorithme récursif qui minimise l'entropie, celle-ci étant ici définie comme la quantité d'information nécessaire (en bits) pour déterminer une classe (Fayyad et Irani, 1993. Brièvement, l'algorithme permet de déterminer, dans un groupe d'objets, un point de coupe qui sépare ce groupe en deux parties et minimise l'information nécessaire pour former ces deux parties. Cet algorithme a été proposé en 1993 par Fayyad et Irani (1993). Une

étude comparative entre différents algorithmes de discrétisation menée par Dougherty *et al.* (1995) montre qu'il donne de bons résultats.

La discrétisation locale supervisée permet d'inclure des conditions d'arrêt, comme un nombre de classes maximum.

Discrétisation globale

Ce processus effectue une discrétisation simultanée sur tous les attributs de la table contrairement à la discrétisation locale qui traite chaque attribut séparément. Ce type d'approche suppose donc qu'il existe une certaine concordance entre les attributs qui varient « ensemble » pour les différentes classes.

3.1.3 Formation des atomes

Le principal concept de la TEA est la **relation d'indiscernabilité**, normalement associée à un ensemble d'attributs. Si on considère le tableau 4, les trois exemples (ou objets {O1, O3, O4}) sont semblables pour les attributs ADN, LPO et PHAG. Les trois derniers objets {O2, O3, O4} sont semblables pour les attributs MT, Vn et PHAG et les deux derniers objets {O3, O4} sont semblables pour tous les attributs sauf l'attribut décisionnel.

Tableau 4
Exemple de table d'information discrétisée

Biomarqueurs							Site
MT	EROD	ADN	LPO	Vn	PHAG	NspE	
2	1	1	1	3	0	1	ASE
0	1	2	3	2	0	1	BE
0	0	1	1	2	0	0	BE
0	0	1	1	2	0	0	ASJ
Attributs conditionnels							Décision

Les ensembles qui sont indiscernables sont appelés ***ensembles élémentaires***. Ainsi l'ensemble des attributs ADN et PHAG définissent l'ensemble élémentaire {O1, O3, O4}. Lorsque tous les attributs sont pris en compte, les ensembles élémentaires formés par les objets sont appelés ***atomes***. Dans le tableau 4, on peut compter 3 atomes {O1}, {O2}, {O3, O4}. Les attributs décisionnels peuvent être exprimés de la même façon et on parle alors de ***concept***. Dans notre exemple, les concepts sont formés par les différents sites car il n'y a qu'un attribut décisionnel.

Sur la base des atomes, il est possible d'évaluer comment les différents objets se placent par rapport aux concepts. En effet, dans un cas idéal, chaque atome correspondrait à un concept, ou plus simplement, chaque critère de décision aurait ses propres caractéristiques. Dans le cas présent, chaque site devrait être défini par des mesures de biomarqueurs différentes des autres sites. Or, ce n'est pas le cas dans notre exemple (tableau 4) car les objets 3 et 4, qui appartiennent au même atome, n'ont pas le même attribut décisionnel. Ces objets sont donc en conflit et on parle d'inconsistance des données.

La TEA offre un moyen pour tenir compte de ces inconsistances. Pour chaque concept X, on calcule le plus grand et le plus petit nombre d'objets contenus dans X sur la base des atomes. On parle d'approximation supérieure et inférieure de X. Ainsi, dans l'exemple ci-dessus, le site BE contiendrait au maximum deux atomes (0 0 2 1 2 0 0 et 0 0 1 1 2 0 0) et au minimum un (0 0 2 1 2 0 0). Si on compte les objets, l'approximation supérieure donne trois objets {O2, O3, O4} et l'approximation inférieure un seul {O2}. L'ensemble {O3, O4}, qui contient les éléments de l'approximation supérieure qui ne font pas partie de l'approximation inférieure, est appelé ***région frontière***. Les éléments de la région frontière ne peuvent pas être classés avec certitude comme membre de l'un ou l'autre concept.

Une mesure de l'inconsistance des données est fournie par l'***exactitude*** de l'approximation calculée comme l'approximation inférieure sur l'approximation supérieure d'un concept. Dans l'exemple ci-dessus, pour le site BE, l'exactitude vaut 1/3 = 0,3333.

Il est également possible de calculer la ***qualité*** de l'approximation des données. Cette qualité est définie comme la somme des objets des approximations inférieures pour tous les concepts sur le nombre d'objets totaux. Dans le cas ci-dessus, la qualité vaut: (0+1+1)/4 = 0,5.

3.1.4 Recherche des redondances

Par le concept d'indiscernabilité, il est aisé de définir les attributs redondants. En effet, si un ensemble d'attributs et sous-ensemble d'attributs définissent la même relation d'indiscernabilité, alors chaque attribut qui appartient au sous-ensemble et non a l'ensemble est redondant.

Dans le tableau 4, l'ensemble défini par Vn et PHAG définit les ensembles élémentaires {O1, O4} {O2, O3}. C'est également le cas du sous-ensemble défini par Vn seul. Ainsi, l'attribut PHAG est redondant.

3.1.5 Génération de règles

Les règles générées par la TEA sont de type « si...alors ». Ces règles se basent sur les atomes et concepts formés précédemment. Pour chaque concept, les règles induites par son approximation inférieure sont ***certainement valables***. Les règles générées par son approximation supérieure sont ***possiblement valables*** et leur application résulte en une indécision quant à la classification.

À titre d'exemple, dans le tableau 4, les règles suivantes peuvent être énoncées :

Règles certaines :	MT et Vn élevé	⇒	site ASE
	ADN et LPO élevés	⇒	site BE
Règles possibles :	EROD et NspE très faibles	⇒	site BE
	EROD et NspE très faibles	⇒	site ASJ

Chaque règle de décision est caractérisée par une ***force de suggestion*** qui prend en compte le nombre d'objets satisfaisant les conditions sur les règles et appartenant à la classe de décision suggérée. Dans le cas des règles approximatives, la force est calculée séparément pour chaque classe possible de décision. Les règles les plus fortes sont habituellement plus générales, Cela signifie que la partie conditionnelle est plus courte et moins spécialisée.

La procédure de génération des règles se base sur un algorithme qui est une version modifiée d'un algorithme appelé LM2 (Grzymala-Busse, 1992; Skowron,1993). Cet algorithme est constitué d'un groupe d'algorithmes d'induction qui se concentrent sur l'induction de toutes les descriptions discriminantes des classes de décision ou sur les descriptions qui leur sont proches.

On dit qu'une ***description*** est ***discriminante*** si elle est complète (chaque exemple positif - i.e., qui appartient à la classe de décision - doit être reconnu comme appartenant à cette classe) et consistante (chaque exemple négatif - i.e., appartenant à une autre classe - ne doit pas être reconnu comme appartenant à cette classe). Les descriptions discriminantes définies pour une classe sont assumées comme étant minimales, i.e., en enlever une signifie que la description de la classe n'est plus complète.

3.1.6 Classification

L'une des attentes formulées à l'égard de la TEA consiste à établir des classifications prédictives des observations. Une fois les règles formulées, pouvons-nous prévoir à quel groupe particulier une observation appartient? L'avantage de cette étape est qu'elle permet également de juger de la qualité de l'analyse. En effet, si la méthode discrimine bien entre les groupes, la classification sera bonne, ou mauvaise si tel n'est pas le cas.

Deux types de classifications peuvent être effectuées :

- classification *a priori;*
- classification *post-hoc.*

La classification *post-hoc* se base sur les mêmes données qui ont servi à construire les règles. Son utilisation conduit presque toujours à une bonne classification, mais les résultats n'ont pas de véritable valeur prédictive. En revanche, la classification *a priori*, basée sur un autre lot de données disponibles, donne une bonne estimation de la qualité de l'analyse effectuée.

Comme il n'est pas toujours évident d'avoir un deuxième lot de données, on peut utiliser la technique qui consiste à laisser de côté une partie des données (k sur N). Les règles sont alors générées sur les N-k objets et les k données sont classées sur la base de ces règles. Répétée plusieurs fois, cette technique, appelée ***validation croisée***, permet d'avoir une idée de la qualité de la classification sur la base des règles générées par la TEA. Le résultat final se présente sous la forme d'une ***matrice de confusion*** représentant les différents objets et leur classification (juste ou fausse). Le logiciel ROSE2 permet le choix du nombre k qui, ici, a été choisi égal à 1. Ce cas particulier de validation se nomme « leaving-one-out test ». Notons que le seuil de décision des éléments non classables a été fixé à 30. En effet, lorsqu'un objet peut potentiellement appartenir à

deux classes, on regarde les règles auxquelles il correspond. Si la force des règles d'appartenance à un site dépasse un certain seuil, on attribue l'élément à ce site.

3.2 ANALYSE DISCRIMINANTE

L'analyse discriminante a été effectuée avec le logiciel SAS, version 8. La normalité de la distribution des données a été vérifiée préalablement à l'analyse (test de Kolmogorov-Smirnov, $\alpha = 0,05$). L'homogénéité de variances/covariances a également été testée sur la base d'un test utilisant une distribution Chi-carré (Morrison 1976).

Pour établir une comparaison avec la TEA, la qualité de la classification est évaluée sur la base d'un test de validation croisée avec un élément, soit un « leaving-one-out test ». Notons que pour la classification d'une mye dans un site, la probabilité a été fixée à 10 %.

3.3 ARBRES DE DÉCISION

3.3.1 À propos des arbres de décision

Les arbres de décision sont utilisés pour prévoir l'affectation d'observations ou d'objets à des classes de variables dépendantes catégorielles à partir de leurs mesures sur une ou plusieurs variables prédictives. La flexibilité des arbres de décision en font une analyse très attrayante car elle ne dépend pas de la distribution des données. Elle peut donc être utilisée lorsque les techniques traditionnelles échouent ou encore comme technique exploratoire.

Il existe différents type de construction d'arbre. L'un des plus connus est la méthode CART (Classification And Regression Trees) qui utilise une recherche de grille exhaustive de toutes les segmentations univariées possibles pour trouver les segmentations d'un arbre de décision (Breiman *et al.*, 1984).

3.3.2 Construction des arbres de décision

La construction des arbres de décision a été effectuée avec le logiciel STATISTICA, version 5.5. Les données utilisées sont les mêmes que pour l'analyse discriminante, c'est-à-dire que les données aberrantes ont été supprimées.

Deux méthodes ont été utilisées pour la construction des arbres, une méthode de *recherche exhaustive de segmentation univariée style-CART* et une méthode de *segmentation*

univariée basée sur une méthode discriminante, méthode rapide et non biaisée, mais aussi plus technique que la méthode CART. Cette dernière méthode est basée sur un algorithme de type QUEST, utilisant une modification d'analyse discriminante récursive (Loh et Shih 1997). En général, ces deux approches sont considérées comme complémentaires.

4 Résultats et discussion des méthodes d'analyse

4.1 RÉSULTATS DE LA THÉORIE DES ENSEMBLES APPROXIMATIFS

4.1.1 Discrétisation

La discrétisation choisie ici est une discrétisation locale non supervisée. Les essais avec une discrétisation globale montrent que cette dernière donne de moins bons résultats de classification. De plus, intuitivement, une discrétisation globale ne nous semble pas justifiée du fait que les différents biomarqueurs sont totalement indépendants les uns des autres (un biomarqueur peut répondre pour un site pollué et pas du tout pour un autre, alors que c'est le contraire pour un autre biomarqueur).

La table de discrétisation complète se trouve à l'annexe 2. Les classes formées par la discrétisation pour les différents biomarqueurs sont représentées au tableau 5.

Tableau 5
Classes générées par la discrétisation locale effectuée à l'aide du logiciel ROSE2

Classe	MT [nmoles MT/mg protéines]	ADN [µg supernageant DNA/mg protéines]	LPO [µg acid thiobarbiturique réagissant/mg protéines]
0	0,4< x < 0,8	100< x <500	2,2< x <4,4
1	0,8< x < 1,7	500< x < 900	4,4< x <7
2	1,7< x <2,4	900< x <1700	7< x <20
3			20< x <100
Classe	**Vn [µg phosphate dans la phase organique/mg protéines]**	**PHAG [µg florescéine (bactéries ingérées)/ mg protéines]**	**NspE [µg florescéine / mg protéines]**
0	90< x <350	tout	0,6< x <1
1	350< x <400		1< x <4
2	400< x <550		4< x <10
3	550< x <1250		10> x >24

Il y a un facteur 4 entre la classe la plus faible et la plus forte pour les biomarqueurs MT, ADN et Vn. Le facteur est de 20 pour les biomarqueurs LPO et NspE.

L'étude des classes montre qu'elles sont cohérentes pour les MT. En effet, les valeurs de la classe la plus haute (>1,7) sont des valeurs que l'on retrouve chez les invertébrés vivant dans des eaux contaminées (Malley *et al.*, 1993; Couillard *et al.*, 1995a et 1995b; De Lafontaine *et al.*, 1999). Pour les autres biomarqueurs, on peut penser qu'un facteur 4 (et à plus forte raison 20) montre une différence notable avec le niveau faible. Cependant, il est très difficile de faire des comparaisons avec les données de la littérature qui souvent ne sont pas recueillies sur le même organisme ni avec les mêmes méthodes. Le fait que la classification soit bonne encourage cependant à penser que les classes sont cohérentes.

Pour valider véritablement les classes, il faudrait pouvoir les comparer avec des données sur des myes contaminées à divers degrés, par exemple en laboratoire. Ces classes seraient alors réellement représentatives de valeurs faibles, moyennes ou élevées des différents biomarqueurs, Elles pourraient être utilisées pour comparer différents sites, ce qui n'est pas le cas actuellement. En effet, dans cette étude, les classes sont déterminées sur la base des valeurs existantes. Si les sites supposés « sans contamination directe » sont pollués, cela introduirait un biais. Il serait donc plus judicieux d'avoir des classes absolues.

Le cas de la phagocytose est intéressant car le logiciel n'arrive pas à trouver de classes. Or, si on observe la distribution des données (figure 3), on s'aperçoit qu'elles sont effectivement très hétérogènes et que l'on arrive pas vraiment à déterminer une structure. Un essai a été fait en ajoutant artificiellement une classe à la phagocytose (classe 0 : $0< x <1,1$, et classe 1 : $x > 1,1$). L'ajout d'une classe améliore la qualité et l'exactitude de l'approximation et diminue le nombre de règles approximatives. En revanche, l'exactitude de la classification finale diminue avec un pourcentage qui passe de 90,5 % pour la discrétisation classique à 83 % pour la discrétisation avec ajout d'une classe. Le but de cette étude étant justement la classification des sites, la discrétisation effectuée par le logiciel semble donc bien appropriée. Il est intéressant de noter qu'un essai effectué avec des classes créées subjectivement donne des résultats nettement moins bons que ceux trouvés ici (62 % d'exactitude pour la classification). Ceci est certainement dû au fait que l'on a tendance à créer trop de classes.

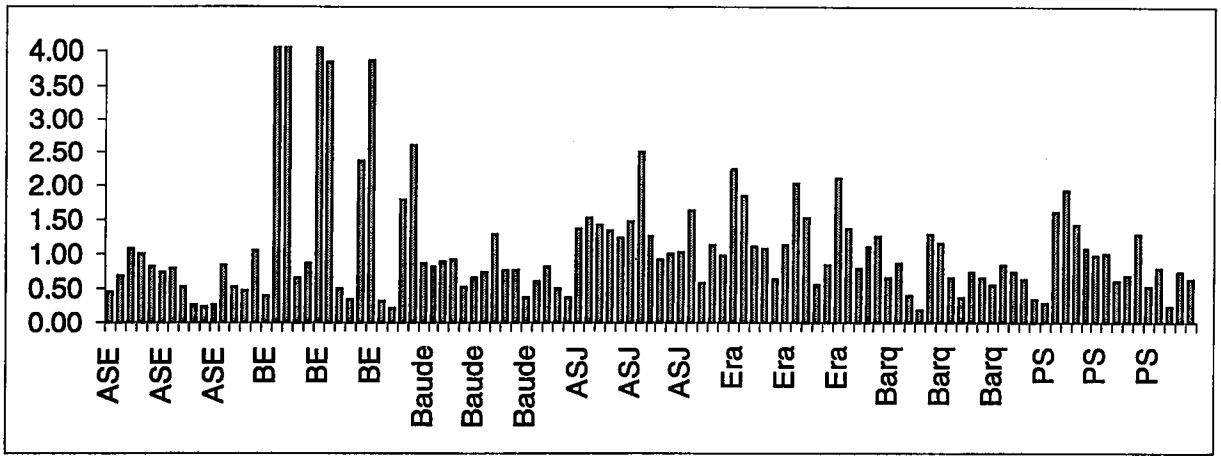


Figure 3 Distribution des données de phagocytose pour les différents sites de prélèvement

4.1.2 Approximation et redondance

L'approximation effectuée sur les données discrétisées donne un total de 33 atomes avec une exactitude de 90 % et une qualité de 95 %. Les approximations inférieures et supérieures pour les différents sites de même que l'exactitude sont résumées à l'annexe 3.

Le biomarqueur phagocytose est redondant, ce qui n'est pas étonnant puisque la discrétisation n'a pas distingué de classes pour cette mesure.

4.1.3 Règles

Seize règles, dont une approximative, ont été générées sur la base des données. Les règles les plus fortes sont présentées au tableau 6. L'ensemble des règles est présenté à l'annexe 3.

Un site se décrit ou se distingue très bien s'il a peu de règles (grande force) et qu'il n'appartient pas à une règle approximative.

Il y a deux règles très fortes (100 % d'objets appartenant au site satisfont la règle). Il s'agit du site Era (région industrielle) qui se décrit par un niveau élevé de peroxydation des lipides et du site PS (quai) qui se décrit par un faible niveau de peroxydation des lipides et un faible niveau de dommages à l'ADN. Notons que dans ce dernier cas, il n'y a que 13 objets pour ce site, la règle a donc bien une force de 100 %. Le site ASJ (rejets urbains) se décrit également très bien avec deux règles. Il est caractérisé principalement par un niveau faible de dommages à l'ADN et un niveau élevé d'activité non spécifique des estérases. Cependant, une grande partie des myes

montrent également un niveau élevé de vitélline. Le site ASE (sans contamination directe) se décrit complètement par trois règles et il est principalement caractérisé par un niveau de peroxydation des lipides assez faible, une activité non spécifique des estérases faible et un niveau moyen de métallothionines. Un nombre élevé de myes ont cependant un niveau élevé de métallothionines.

Tableau 6
Règles de force supérieure à 6*

15	LPO=3					-> site Era
13	ADN=0	*et*	LPO=0			-> site PS
14	ADN=0	*et*	NspE=3			-> site ASJ
13	ADN=0	*et*	Vn=3			-> site ASJ
11	ADN=2	*et*	LPO=2	*et*	NspE=1	-> site Baude
10	LPO=1	*et*	NspE=1	*et*	MT=1	-> site ASE
6	ADN=2	*et*	NspE=2			-> site BE
Règle approximative :						
5	ADN=1	*et*	NspE=1			-> site Baude *ou* site Barq

* La force de la règle est dénotée par le chiffre au début de chaque ligne. Les règles se lisent : « si...et...et...alors ».

Les confusions se situent au niveau des sites BE (quai), Barq (sans contamination directe) et Baude (sans contamination directe). Le site Baude est décrit par une règle qui a une force de 79 %. Il est principalement caractérisé par un niveau élevé de dommages à l'ADN et de peroxydation des lipides, avec un faible niveau de métallothionines, mais il peut être confondu avec le site Barq.

Les sites BE et Barq sont les plus difficiles à décrire, avec respectivement quatre règles pour le site BE et trois règles pour le site Barq (force totale 85 %) et une règle approximative.

Si les règles permettent de discriminer entre les différents sites, elles permettent également une caractérisation de ces derniers. Par exemple, le site Era se distingue par un niveau

très élevé de peroxydation des lipides. Or, il s'agit d'un dommage important causé aux tissus par un bon nombre de polluants. Le site Era se trouve effectivement dans une zone sensible à la contamination puisque proche de la zone industrielle. Il n'est donc pas étonnant de constater un tel type de problème dans cette région. De même, le site ASJ se caractérise, entre autres, par un niveau élevé de vitélline et d'estérases non spécifiques. Or, ce site se trouve près d'une zone de rejets d'eaux usées dans lesquelles il y a souvent un taux élevé de molécules à effet œstrogénique, ce qui pourrait expliquer le taux élevé de vitélline. De même, ces eaux contiennent beaucoup de microorganismes, ce qui pourrait être une hypothèse pour expliquer le taux élevé d'estérases.

En revanche, le site BE est caractérisé par plusieurs règles, ce qui pourrait s'expliquer par une contamination hétérogène. Ceci paraît réaliste puisque ce site est situé près d'un quai et subit donc d'autres influences que celles associées aux bateaux.

Il est également intéressant de noter que les sites Barq et Baude, deux sites sans contamination directe qui se confondent, sont aussi des sites géographiquement proches, soit en amont et en aval de l'embouchure du fjord du Saguenay.

4.1.4 Matrice de confusion

La matrice de confusion calculée sur la base d'un test de validation croisée avec un élément est présentée au tableau 7. La qualité de la classification globale est très bonne puisqu'elle atteint quasiment 91 %.

Découlant des remarques précédentes, on constate que les sites pour lesquels la classification se fait le mieux sont les sites Era, PS, ASE et ASJ, soit ceux pour lesquels les règles sont les plus fortes. Les sites BE et Barq ont en revanche de moins bons résultats. Fait étonnant, le site Baude, qui peut se confondre avec le site Barq, a un pourcentage d'exactitude de classification de 100 %. En fait, si on observe le tableau, on voit que ce sont les éléments du site Barq pour lequel certains objets se classent dans le site Baude.

Tableau 7
Matrice de confusion calculée sur la base d'un test de validation croisée avec un seul élément

	ASE	BE	Baude	ASJ	Era	Barq	PS	Aucun	% exact
ASE	**14**	0	0	0	0	0	0	1	93
BE	2	**10**	0	0	0	1	0	1	71
Baude	0	0	**14**	0	0	0	0	0	100
ASJ	0	0	0	**14**	0	0	0	1	93
Era	0	0	0	0	**15**	0	0	0	100
Barq	0	0	2	0	0	**10**	0	1	77
PS	0	0	0	0	0	0	**13**	0	100

Exactitude totale ($\bar{x} \pm \sigma$): **90,9 ± 11,9 %**.

* Données sans EROD et sans valeurs aberrantes.

4.1.5 Autres essais

On a déjà parlé du fait de créer les classes subjectivement, ce qui diminue la qualité de la classification. D'autres essais ont également été effectués en prenant en compte les valeurs du biomarqueur EROD (ce qui réduit le nombre de données à 60), en laissant les valeurs aberrantes ou en groupant les sites sans contamination directe.

Les résultats montrent que l'ajout du biomarqueur EROD n'améliore pas véritablement la classification et il est donc raisonnable de supposer qu'il n'est pas nécessaire à la classification. Il faut cependant noter qu'il y a peu de valeurs EROD disponibles et surtout qu'elles sont mal réparties (trois seulement pour ASJ, par exemple). Pour pouvoir réellement tirer des conclusions sur la redondance de l'EROD, il faudrait avoir plus de valeurs.

Le fait d'enlever les valeurs aberrantes améliore la classification puisqu'elle atteint une exactitude de 91 % (contre 85 %). La nécessité d'enlever les valeurs dites aberrantes peut cependant être discutée car ces valeurs peuvent également apporter de l'information. Dans le cas où on décide de les supprimer, il sera judicieux de choisir une méthode appropriée, par exemple une méthode non paramétrique.

Le fait de grouper les sites « sans contamination directe » améliore la classification mais les données ne sont alors plus les mêmes que lorsque les sites sont pris séparément. Il faut dire qu'il n'y a, a priori, aucune raison pour grouper les sites puisqu'il est impossible d'affirmer que les

trois sites groupés sont vraiment non pollués. La différence la plus importante lorsque les sites « non pollués » sont groupés est liée à la discrétisation de la phagocytose. Celle-ci devient en effet significative et passe de zéro classe à quatre classes. Cependant, si la classification finale s'en trouve améliorée par rapport à la classification avec les sites non groupés, l'ajout de classes pour la phagocytose dans l'analyse des sites non groupés diminue légèrement la qualité de la classification. Le problème est donc différent d'un cas à l'autre et il n'y a aucune raison de remettre en cause la discrétisation effectuée pour les sites non aggrégés.

4.2 RÉSULTATS DE L'APPLICATION DE L'ANALYSE DISCRIMINANTE

4.2.1 Analyse discriminante

L'analyse discriminante a été effectuée sur les données non transformées. En effet, le test de Kolmogorov-Smirnov montre que, pour trois biomarqueurs seulement, la normalité n'est pas respectée pour un site. Lorsque les données sont transformées en log, il s'y ajoute un biomarqueur pour lequel la normalité des données n'est pas respectée dans deux sites. Le test d'homogénéité des covariances effectué préalablement à l'analyse discriminante montre que l'hypothèse nulle, i.e. les covariances sont égales, peut être rejetée (α = 0,1). L'analyse discriminante a donc été appliquée en supposant des données multi-normales hétérogènes.

Le tableau 8 montre la matrice de confusion obtenue avec une validation croisée à un élément.

Tableau 8
Matrice de confusion calculée sur la base d'un test de validation croisée avec un seul élément

	ASE	BE	Baude	ASJ	Era	Barq	PS	% exact
ASE	**15**	0	0	0	0	0	0	100
BE	0	**12**	1	1	0	0	0	86
Baude	0	3	**9**	0	0	2	0	64
ASJ	0	0	0	**15**	0	0	0	100
Era	0	0	0	0	**15**	0	0	100
Barq	0	2	2	1	0	**8**	1	62
PS	0	0	0	0	0	0	**13**	100

Exactitude totale ($\bar{x} \pm \sigma$): **87,4 ± 17,5 %**.

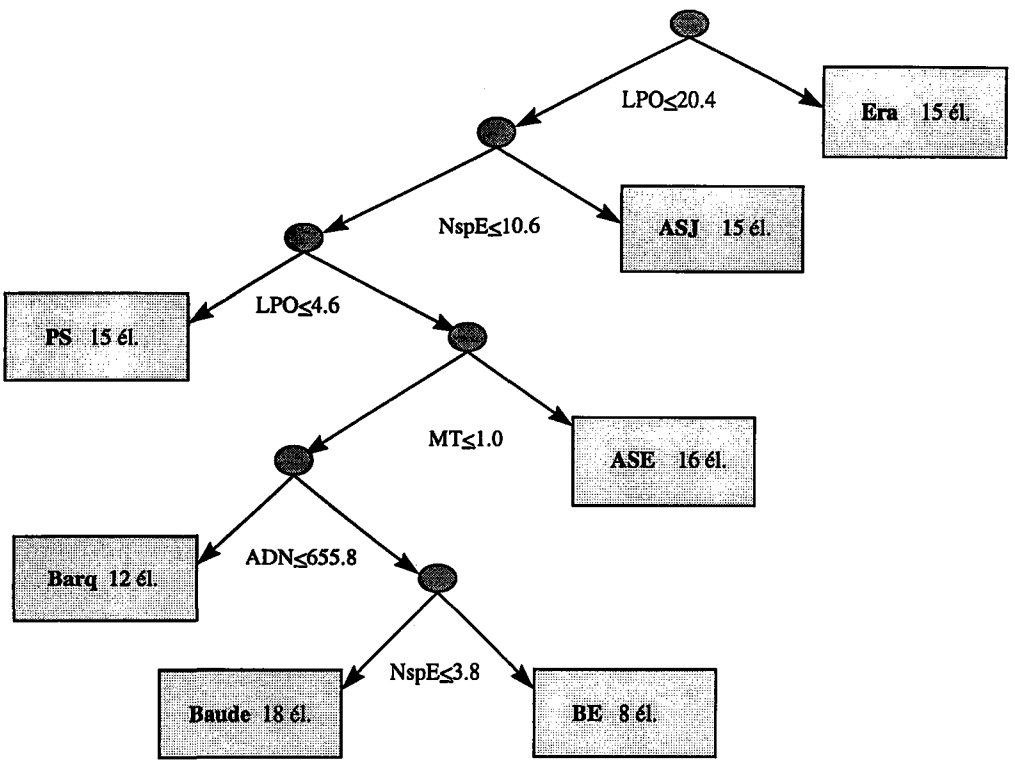
La qualité de la classification est ici aussi très bonne puisque l'on atteint 87 % d'exactitude. À nouveau, les sites Era, ASJ, PS et ASE se classent très bien. Le site BE obtient par contre une meilleure classification que précédemment, alors que les sites Barq et Baude obtiennent une classification nettement moins bonne. La différence la plus évidente se rapporte au site Baude qui obtient un score relativement bas alors qu'avec la TEA, il avait un score de 100 %. Ces différences ne sont pas étonnantes puisqu'il s'agit de deux méthodes différentes. Lorsque les sites se discriminent bien, on obtient les mêmes résultats, mais lorsqu'il y a confusion, le résultat dépend beaucoup plus de la méthode utilisée. Notons que la TEA donne un résultat global de classification plus homogène que l'analyse discriminante (écart type passant de 12 à 18 %).

4.3 RÉSULTATS DE L'APPLICATION DE LA MÉTHODE DES ARBRES DE DÉCISION

Construire un arbre de décision pose le problème de savoir où arrêter la segmentation. En effet, plusieurs choix sont possibles, allant de la segmentation complète à un arrêt relativement rapide basé sur un nombre d'erreurs minimales dans le classement.

La segmentation complète effectuée avec un algorithme de type CART montre un arbre relativement complexe comprenant 29 nœuds. Une validation croisée effectuée 15 fois montre que l'erreur de mauvais classement se situe autour de 19 % (soit 81 % de bon classement). Cet arbre, trop complexe, n'est pas satisfaisant. Il s'agit donc de réduire le nombre de segmentations sans augmenter de manière trop importante le pourcentage de mauvais classement. Quelques essais ont montré qu'en fixant la déviance à 5, i.e. mesure basée sur le maximum de vraisemblance et servant à minimiser les erreurs de prévision, l'arbre de décision construit est satisfaisant. En effet, le nombre de nœuds est réduit à 12 et l'erreur de mauvais classement est de 28 % (soit 72 % de bon classement). Cette augmentation peut sembler importante, mais il faut noter que le premier choix (arbre complet) tient compte de chaque particularité de chaque élément. Il donne donc de bons résultats dans le cas d'une validation avec les mêmes données, mais risque d'être difficile à valider avec un autre ensemble de données. Le deuxième cas est plus intéressant car plus global. Notons encore que l'essai avec la méthode de segmentation univariée basée sur une méthode discriminante donne le même arbre et le même type de résultats.

La figure 4 montre le schéma de l'arbre de décision construit avec un algorithme de type CART, la déviance étant fixée à 5.



*Remarque : D*éviance fixée à 5.

Figure 4 Arbre de décision de type CART

Ici, comme pour la TEA, la discrimination est interprétable. Le site Era se distingue par un niveau élevé de peroxydation des lipides alors que le site ASJ se distingue par un niveau élevé d'estérases non spécifiques. Les sites les plus difficiles à discriminer se situent au bas de l'arbre; il s'agit à nouveau des sites Baude, BE et Barq. Ces caractéristiques peuvent se comparer à celles des règles de la TEA, mais elles semblent moins précises. Pour le site ASJ par exemple, il n'est pas question de la vitélline qui était utilisée dans la TEA. A priori on s'attend donc à une moins bonne classification. Notons que déjà au niveau de l'arbre lui-même, on peut voir que certaines myes ne se classent pas bien puisque, par exemple, le site Baude contient 18 éléments (soit 5 de plus que le nombre réel) et le site BE n'en contient que 8 (soit 6 de moins que le nombre réel).

Le tableau 9 donne la matrice de confusion obtenue avec une validation croisée à un élément. La qualité de la classification est nettement moins bonne que pour la TEA ou pour l'analyse discriminante. Elle est également beaucoup plus hétérogène.

Tableau 9
Matrice de confusion de l'arbre de décision de type CART calculée sur la base d'un test de validation croisée avec un seul élément

	ASE	BE	Baude	ASJ	Era	Barq	PS	% exact
ASE	**13**	1	1	0	0	0	0	87
BE	2	**5**	4	2	0	0	1	36
Baude	0	2	**7**	2	0	3	0	50
ASJ	0	2	0	**12**	0	1	0	80
Era	0	1	0	0	**14**	0	0	93
Barq	0	2	0	0	0	**11**	0	85
PS	1	0	0	0	0	2	**10**	77

Exactitude totale ($\bar{x} \pm \sigma$): **72,6 ± 21,2 %**.

Le site pour lequel la classification est la meilleure est le site Era avec 93 %. Notons qu'aucun site n'obtient 100 %. Les sites ASE, ASJ obtiennent une bonne classification. C'est également le cas pour le site Barq, ce qui a de quoi surprendre car il s'agit d'un des sites qui se classent le moins bien pour ce qui est de la TEA et de l'analyse discriminante. Les sites les moins bien classés sont les sites Baude et BE. Comme souligné précédemment, il faut s'attendre à ce que le choix de la méthode influence la classification des sites qui se classent les moins bien. Ainsi le site BE se classe le mieux avec l'analyse discriminante, un peu moins bien avec la TEA et encore moins bien avec l'arbre de décision.

4.4 COMPARAISON DE LA TEA, DE L'ANALYSE DISCRIMINANTE ET DES ARBRES DE DÉCISION

La théorie des ensembles approximatifs et l'analyse discriminante appliquées aux données de biomarqueurs donnent une exactitude de classification assez similaire (91 % et 87 % respectivement). Ce résultat va dans le même sens que la comparaison effectuée par Krusińska *et al.* (1992). Par rapport à l'analyse discriminante, la TEA présente cependant certains avantages. Le premier est que son application ne dépend pas de la distribution des données. Cette spécificité est particulièrement intéressante dans le cas de données d'écotoxicité car, bien souvent, leur distribution n'est pas classique. Des transformations de variables, plus ou moins heureuses, sont donc le plus souvent nécessaires pour pouvoir appliquer une méthode paramétrique.

Le deuxième avantage de la TEA repose sur le fait qu'elle génère des règles qui caractérisent les différents objets, comme ici les sites. En effet, on apprend que le site Era, situé dans une zone industrielle, peut être distingué des autres sites uniquement par son niveau élevé de peroxydation des lipides. C'est une indication importante car elle peut nous fournir des indices quant au danger d'un tel site. Ces remarques ne sont pas applicables à l'analyse discriminante qui, elle, génère des équations difficilement interprétables.

D'autre part, la TEA permet également d'inclure des données subjectives puisqu'elle travaille avec des classes et non avec des chiffres. Ainsi, on aurait pu inclure dans ces données la notion de santé de la population qui pourrait être évaluée lors de l'échantillonnage au niveau, par exemple, de sa densité : population élevée, population moyenne, faible population.

Le principal désavantage de la TEA est essentiellement lié au fait qu'elle utilise des classes, ce qui oblige à commencer par une discrétisation des données. Cependant, les tests effectués ici montrent que les résultats de la TEA sont équivalents à ceux de l'analyse discriminante, ce qui semble indiquer que les algorithmes de discrétisation sont très performants et que cette étape n'engendre pas de problèmes. Il convient cependant de toujours évaluer la pertinence des classes formées, par exemple en les comparant à des valeurs de laboratoire.

La construction d'arbres de décision est l'une des méthode qui présentent les mêmes avantages que la TEA au niveau de la caractérisation des objets et de la distribution des données, sans comporter le problème de la classification. Cependant, les résultats obtenus ici semblent montrer que la classification effectuée par cette méthode est moins bonne qu'avec les deux autres

méthodes (73 % contre 87 et 91 %). Si l'on compare l'arbre de décision (figure 4) avec les règles générées par la TEA (tableau 6), on s'aperçoit que les règles les plus fortes se retrouvent entièrement ou partiellement dans les critères de sélection de l'arbre. Ainsi, le site Era se caractérise par un niveau élevé de peroxydation des lipides dans les deux cas. Pour le site ASJ, on retrouve le niveau élevé de l'activité non spécifique des estérases et pour le site PS, le niveau très bas de peroxydation des lipides. Cependant, dans la plupart des cas, les règles générées par la TEA sont plus précises que les critères de choix des arbres de décision, ce qui pourrait expliquer la meilleure classification de la TEA.

L'analyse discriminante et la TEA, qui donnent des résultats semblables, peuvent être vues comme des méthodes complémentaires. En effet, si les données suivent une distribution sous-jacente, l'analyse discriminante est plus intéressante car elle ne passe pas par une étape de discrétisation. Par contre, si les données ne suivent pas de distribution, si l'on veut inclure des aspects subjectifs ou encore si l'on veut caractériser des objets, la TEA est une méthode particulièrement efficace et valide.

5 Création d'un indice de danger

5.1 CONSIDÉRATIONS PRÉLIMINAIRES

Deux types d'indices ont déjà été développés au Centre Saint-Laurent : l'indice BEEP pour l'évaluation des rejets industriels (Costan *et al.*, 1993) et l'indice SED-TOX pour l'évaluation des sédiments contaminés (Bombardier et Bermingham, 1999). Ces indices sont de nature déterministe et sont basés sur la moyenne des effets pour une batterie multitrophique de bioessais. Un indice a également été développé en France pour l'évaluation de la toxicité des effluents (Vindimian *et al.*, 1999).

Il existe par contre très peu de documentation sur un indice développé à partir de biomarqueurs. Narbonne *et al.*, (1999) ont construit un indice basé sur l'addition des différents biomarqueurs pondérés en fonction de leur potentiel discriminant. Calabrese (1997) propose également un indice, mais basé sur des effets pouvant aller des mesures de biomarqueurs à des pathologies cliniques. Bien que nous n'ayons pas de telles données, une approche hiérarchique paraît intéressante.

5.2 MÉTHODOLOGIE

5.2.1 Addition des biomarqueurs

La solution choisie pour la construction de l'indice est l'addition des différents biomarqueurs. En effet, les valeurs sont mesurées sur le même organisme. Intuitivement, plus le nombre de mécanismes affectés est élevé, plus l'organisme risque de développer une pathologie. Par contre, dans le BEEP ou le SED-TOX, les tests sont effectués à plusieurs niveaux trophiques. L'utilisation de la moyenne se justifie donc pour donner une idée globale de la toxicité. Notons que l'addition des biomarqueurs est également la méthode utilisée par Narbonne *et al.* (1999).

Il faut néanmoins garder en mémoire que l'utilisation de l'addition ou de la moyenne présente des biais. En effet, ces deux méthodes compensent les chiffres élevés par les chiffres faibles. En additionnant des biomarqueurs, par exemple, il peut donc arriver qu'un site avec un niveau très élevé pour la peroxydation des lipides mais faible pour les autres biomarqueurs, présente le même indice qu'un site qui a des valeurs moyennes partout. Il en va de même pour le

calcul de la moyenne des effets. L'introduction de ce biais est difficilement évitable à partir du moment où l'on cherche à intégrer toutes les données d'une étude en une seule valeur. Cela met en évidence les limites d'un indice et souligne le fait que l'indice en soi n'apporte que peu d'information sans une interprétation des données par un expert.

5.2.2 Normalisation des valeurs

Pour pouvoir additionner les biomarqueurs, il s'agit de transformer les valeurs de biomarqueurs pour les amener à des valeurs comparables. Pour cela, deux solutions s'offrent à nous : la **normalisation par rapport à une site « contrôle »** et l'**utilisation de classe** (faible, moyen, élevé).

- La **normalisation par rapport à une site « contrôle »** est sans doute la méthode la plus connue et la plus utilisée pour normaliser des valeurs. Il s'agit de calculer un ratio par rapport à une valeur connue, par exemple un site non contaminé. Le problème est de choisir la valeur de référence car les résultats peuvent changer en fonction de ce choix.

  La référence choisie ici est le site Baude, car il est présumé sans contamination directe et il est situé hors du fjord du Saguenay, Le calcul se fait de la manière suivante :

  Pour chaque individu :

  $$ratio = \frac{vmesurée}{vmédiane} \qquad (2)$$

  avec : *vmesurée* : valeur de biomarqueur mesurée sur l'individu
  vmédiane : valeur médiane du site Baude pour ce biomarqueur

  $$indice = \sum ratios \qquad (3)$$

- La deuxième méthode consiste à **utiliser des classes,** par exemple de type niveau faible, moyen ou élevé. Comme discuté dans la première partie pour la discrétisation de la théorie des ensembles approximatifs, l'idéal serait de former des classes sur la base du jugement d'un expert. Ceci n'étant malheureusement pas possible ici, nous avons choisi de reprendre les classes formées par la TEA (tableau 5). L'avantage est que la détermination de ces classes utilise une méthode non paramétrique moins contraignante que la méthode choisie par Narbonne *et al.* (1999) qui utilise l'étendue et les intervalles de confiance supposant une distribution normale des données. Notons que pour l'addition, les classes du tableau 5 ont été augmentées d'un chiffre pour éviter d'utiliser le nombre 0 qui a des propriétés particulières. La classe 0 devient donc la classe 1, la classe 1 la classe 2 et ainsi de suite.

5.2.3 Pondération

Nous avons choisi de ne pas introduire de pondération pour la construction de l'indice. En effet, bien que l'on puisse supposer que certains biomarqueurs montrent des effets plus graves que d'autres, il est très difficile de décider d'un poids et de la valeur de ce poids *a priori*. Nous sommes donc conscients que le fait de ne pas introduire de poids peut amener à une surestimation de certains effets, mais ce désavantage semble préférable à l'introduction de poids arbitraire pouvant mener à des résultats incohérents.

L'indice intégrateur se calcule donc de la manière suivante :

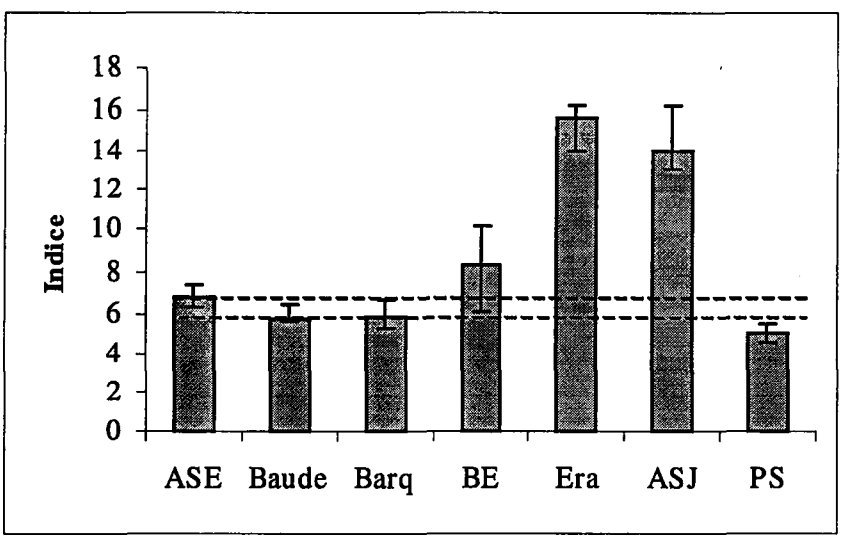
$$\textbf{Indice} = \textbf{MT} + \textbf{ADN} + \textbf{LPO} + \textbf{Vn} + \textbf{PHAG} + \textbf{NspE} \qquad (4)$$

6 Résultats et discussion pour l'application de l'indice

6.1 RÉSULTATS

6.1.1 Normalisation par ratio

La figure 5 présente les résultats de l'indice calculé pour chaque site avec comme référence la médiane du site Baude. Les résultats complets sont présentés à l'annexe 4.



Remarque : La valeur de référence est la médiane du site Baude pour chaque biomarqueur. Les barres sur les graphiques représentent les 1er et 3^{e} quartiles pour les 15 individus de chaque site. Les traitillés horizontaux présentent la valeur minimum et la valeur maximum des médianes de sites sans contamination directe.

Figure 5 Indice médian calculé pour chaque site sur la base d'un ratio non pondéré.

Les sites sans contamination directe (ASE, Baude, Barq) ont des valeurs d'indice entre 5,8 et 6,8. Les sites ASJ et Era se distinguent nettement de cet intervalle avec des valeurs médianes de 14 et 15,6 respectivement. Le site BE se distingue également avec une valeur médiane de 8,3, mais on constate que la distribution des valeurs est assez étendue et que le premier quartile se trouve dans l'intervalle des sites sans contamination directe. Le site PS se distingue également de l'intervalle, mais cette fois, la valeur médiane, valant 5, est plus basse. Notons qu'aussi bien les différences positives que négatives expriment un effet.

La figure 6 montre l'apport des différents biomarqueurs pour le calcul de l'indice dans le cas de chaque site. On constate que si les trois premiers sites ont un indice de valeur très proche, l'importance des différents biomarqueurs n'est pas la même. Le site Baude, qui sert de référence, a une valeur d'environ 1 pour chaque biomarqueur. Le site ASE, comparativement au site Baude, a un niveau plus élevé de MT et un niveau plus bas de LPO. Il en résulte un indice un peu plus élevé. Le site Barq a un niveau plus élevé de NspE, mais plus bas d'ADN et traduit finalement la même valeur d'indice que le site Baude. Le site BE a un niveau plus élevé de PHAG et NspE (toujours comparativement à Baude) ce qui augmente la valeur de l'indice. Les bas résultats du site PS sont dus à un niveau très bas d'ADN et de LPO qui compensent une valeur assez élevée de MT. Enfin pour les deux sites les plus contaminés, les valeurs élevées de l'indice sont dues à un haut niveau de LPO et de MT pour le site et Era, et à un haut niveau de Vn et de NspE pour le site ASJ.

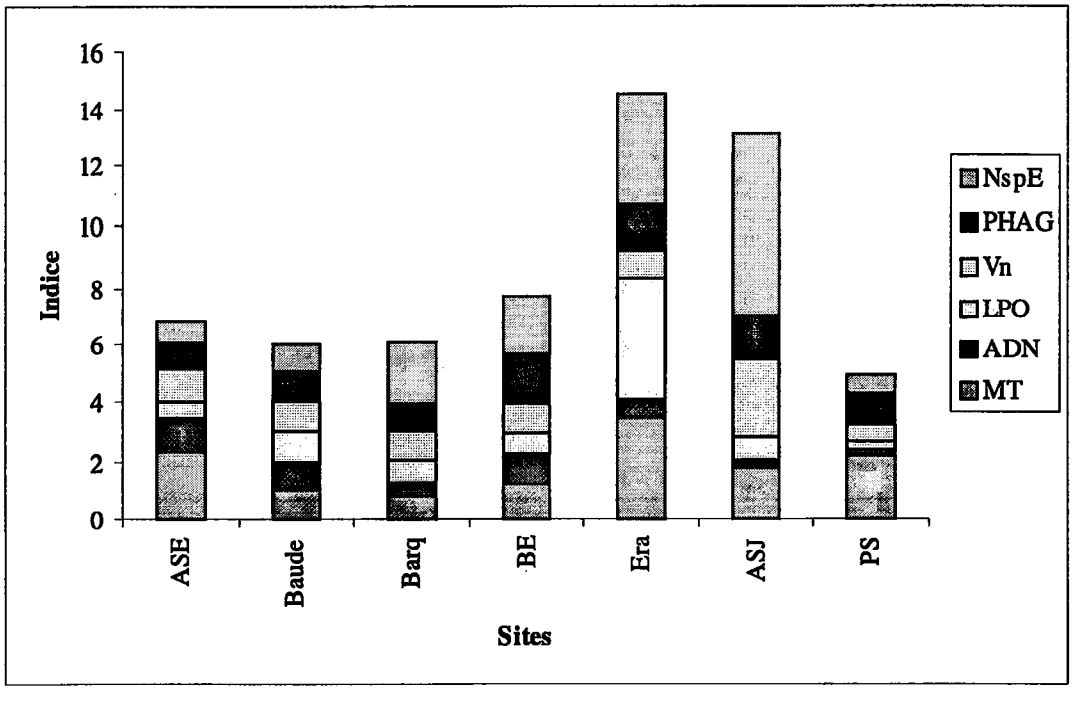
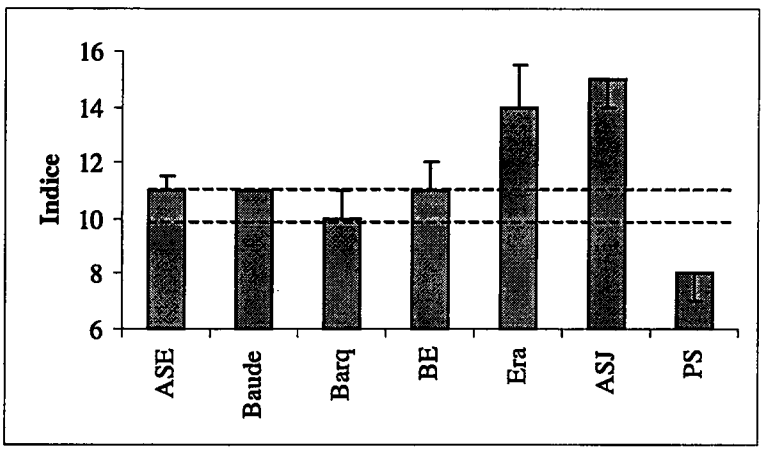


Figure 6 Importance des différents biomarqueurs pour le calcul de l'indice comme somme des ratios

6.1.2 Classes

La figure 7 présente les résultats avec l'utilisation des classes formées par discrétisation lors de la TEA. Les résultats complet sont fournis à l'annexe 5.



Remarque :Les barres représentent les 1er et 3e quartiles pour les 15 individus de chaque site. Les traitillés horizontaux présentent la valeur minimum et la valeur maximum des médianes de sites sans contamination directe.

Figure 7 Indice médian par site calculé en additionnant les classes non pondérées de chaque biomarqueur

Les sites sans contamination directe (ASE, Baude, Barq) ont des valeurs d'indice entre 10 et 11. Les sites ASJ et Era se distinguent nettement de cet intervalle avec des valeurs médianes de 15 et 14 respectivement. Le site BE, par contre, ne se distingue pas bien avec une valeur médiane de 11. À nouveau, le site PS se distingue de l'intervalle avec une valeur médiane plus basse valant 8.

La figure 8 montre l'apport des différents biomarqueurs pour le calcul de l'indice avec les classes. Les valeurs des trois premiers sites (sans contamination directe) sont très proches. Si les biomarqueurs Vn, PHAG et NspE sont les mêmes pour les trois sites, le site ASE a plus de MT que les deux autres, le site Baude plus de LPO et le site Barq plus de NspE. Les différences se compensent pour donner un indice de même valeur. Le site BE est comparable avec les sites Baude et Barq avec des valeurs semblables pour les différents biomarqueurs. Rappelons que ce site était l'un de ceux qui se classait le plus difficilement quelle que soit la méthode choisie.

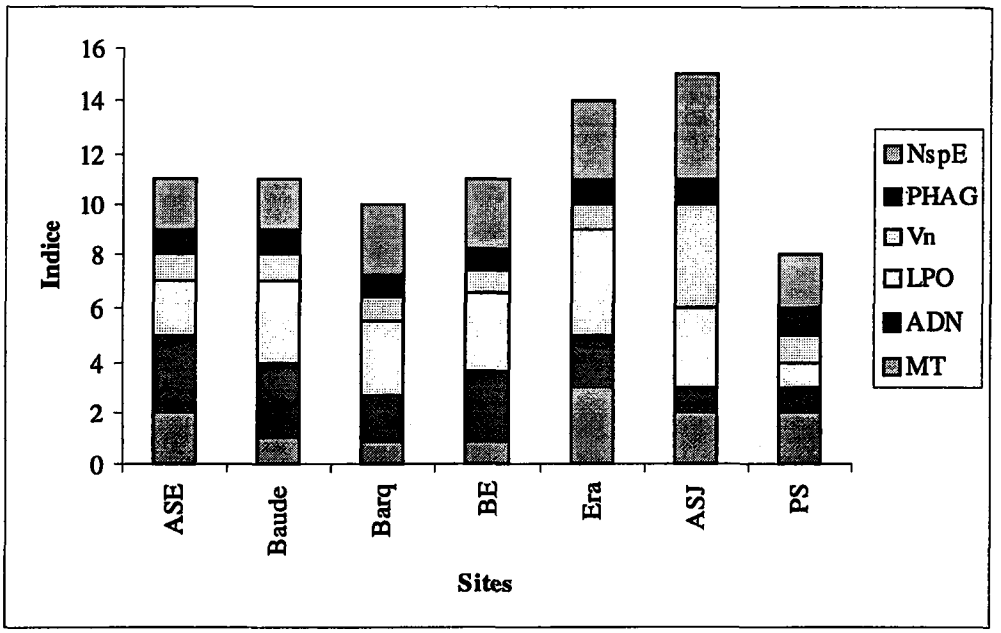


Figure 8 Importance des différents biomarqueurs pour le calcul de l'indice comme somme des classes

6.2 DISCUSSION DE L'INDICE

6.2.1 Ratio versus classes

Globalement, l'indice basé sur les ratios ou sur les classes donne des résultats comparables et cohérents. Les sites très pollués (Era, ASJ) se distinguent toujours des sites sans contamination directe (ASE, Baude, Barq). Pour les sites PS et BE, l'indice avec les classes correspond mieux aux résultats des analyses de discrimination entre sites. En effet, ces analyses montraient que le site PS se discriminait très bien alors que le site BE était plus difficile à classer. Avec les classes, l'indice montre un site PS qui se distingue bien des sites sans contamination directe et un site BE confondu. Avec le ratio, c'est le contraire qui se produit. Cette remarque irait donc dans le sens d'une préférence de l'indice par classe.

Le ratio présente encore d'autres inconvénients, comme celui du choix du site de référence. Le fait de choisir un site qui serait contaminé peut entraîner des erreurs au niveau de l'interprétation des résultats. À titre d'exemple, des essais ont été faits en prenant comme référence le site ASJ (que l'on sait contaminé). Si la distribution des autres sites est a peu près

pareille, le site ASJ se voit attribuer un indice de 6, c'est-à-dire exactement dans la zone des sites sans contamination directe.

D'autre part, l'interprétation des résultats se fera toujours à partir de ce site de référence. Sur la figure 6, le site ASE présente un niveau de MT plus élevé que le site Baude et un niveau de LPO plus bas. Il ne s'agit pas de valeurs absolues, mais de valeurs relatives, ce qui pose un problème pour la généralisation de l'indice. En effet, des indices calculés pour différentes études, et donc avec des sites de référence différents, ne seront pas comparables.

Le fait d'utiliser des classes peut résoudre ce problème. Actuellement, l'indice a été calculé sur la base de classes générées par la discrétisation de la TEA. Il serait possible d'envisager la création de classes absolues (présentant des niveaux faible, moyen, fort par exemple) qui pourraient être utilisées dans des études différentes et ensuite comparées.

L'indice par classe nous semble donc plus avantageux que l'utilisation du ratio puisqu'il permet de s'affranchir du problème du site de référence, mais également d'envisager une généralisation de cet indice et donc des comparaisons possibles avec d'autres études. Nous y reviendrons au point 6.2.2.

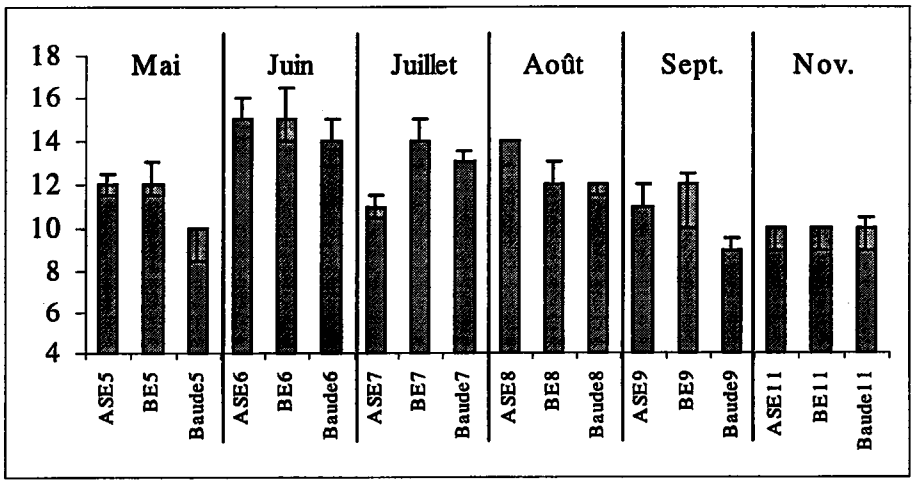
Le fait d'utiliser des classes non paramétriques permet de s'affranchir des hypothèses sur la distribution des données tout en donnant des résultats cohérents. Il serait intéressant de pouvoir comparer notre indice avec celui de Narbonne *et al.* (1999) lors d'une prochaine étude.

6.2.2 Généralisation de cet indice

L'indice par classes, développé dans le cadre de ce travail, est particulièrement intéressant s'il peut être utilisé dans d'autres études et servir pour des comparaisons de pollution entre sites. Nous avons déjà souligné que la détermination *a priori* de classes pourrait permettre de telles comparaisons.

Dans le cadre d'une généralisation, il s'agirait également d'optimiser la batterie de biomarqueurs utilisés, comme c'est le cas pour l'indice BEEP. Dans cette étude, des essais effectués avec une sélection de biomarqueurs (prendre seulement MT, ADN et NspE par exemple) montre que les résultats changent beaucoup suivant cette sélection, Il serait donc important de choisir une batterie et de la fixer pour la suite des études,

Enfin, il faut également noter que, dans cette étude, les mesures ont été prise à une période donnée. Or, certains biomarqueurs fluctuent dans le temps, tels la vitélline liée à la reproduction. De tels changements peuvent fortement perturber le calcul d'un indice comme le montre les résultats de Bresler *et al.* (1999). À titre d'exemple, la figure 9 donne les indices calculés pour trois sites (ASE, BE, Baude), en fonction du mois de prélèvement.



Remarque : L'indice est ici basé sur cinq biomarqueurs seulement (MT, ADN, Vn, PHAG, NspE), les autres données n'étant pas disponibles.

Figure 9 Fluctuation de l'indice en fonction du temps pour trois sites de prélèvements (ASE, BE, Baude)

7 Synthèse et perspectives

La première partie de ce travail visait à évaluer la discrimination entre les différents sites de prélèvement en utilisant différentes méthodes d'analyse. Les analyses effectuées montrent qu'ils se distinguent bien les uns des autres (exactitude de classification près de 90 %).

Parmi les méthodes utilisées, la théorie des ensembles approximatifs (TEA) présente des avantages intéressants. Contrairement à l'analyse discriminante classique, elle ne dépend pas du type de distribution des données, tout en fournissant de bons résultats (plus de 90 % d'exactitude de reclassification). D'autres méthodes non paramétriques comme la construction d'arbres hiérarchiques donnent des résultats nettement moins bons.

La TEA permet également de tenir compte de l'incertitude des données au travers de la création de classes par discrétisation. Ces classes ne sont pas simplement des classes mathématiques mais peuvent également être interprétées comme des niveaux d'effets différents des biomarqueurs (effets faibles, moyens, forts). L'utilisation de classes laisse également la possibilité d'inclure des mesures subjectives liées à des observations de terrain, par exemple, une observation de la population d'organismes qui paraît en santé ou malade.

La formation des règles de la TEA est également un avantage car elle permet une interprétation de la discrimination. Mentionnons, à titre d'exemple, le site Era dont toutes les myes prélevées se caractérisent par un niveau élevé de LPO. Ces règles sont cohérentes (le site Era est un site très pollué et donc susceptible d'induire un haut taux de LPO) et fournissent un outil supplémentaire pour l'interprétation des résultats.

Au vu de ces résultats, il nous semble intéressant de présenter la TEA comme une alternative à l'analyse discriminante, notamment lorsque la distribution des données n'est pas normale ou lorsqu'il existe des mesures non chiffrables.

La deuxième partie de ce travail visait à construire un indice qui permette de mettre en évidence les sites pollués par rapport aux sites non pollués. L'indice basé sur l'addition des classes (non paramétriques) des différents biomarqueurs pour chaque site présente des résultats intéressants et permet de mettre en évidence les sites « à problème ». Le fait d'utiliser des classes permet d'éliminer le problème du choix du site de référence, alors que l'utilisation de classes non

paramétriques permet de travailler avec des données dont la distribution n'est pas forcément normale.

Pour une utilisation plus générale de l'indice, il restera plusieurs paramètres à optimiser. Par exemple, la batterie de biomarqueurs devrait être fixée, de même que les classes qui devraient être déterminées *a priori*. Il restera également à intégrer le facteur temps et à mieux comprendre sa signification dans le contexte de l'indice. Il pourra, en effet, refléter des réponses d'effets naturels ou anthropiques en fonction de la période de collecte des spécimens fauniques.

Références

Blaise, C., P. Gagnon, J. Pellerin, P.D. Hansen et S. Trottier. « Spatial and temporal variation of biomarkers in clam *Mya arenaria* obtained from Saguenay fjord ». À publier.

Bombardier, M. et N. Bermingham (1999). « The SED-TOX index: toxicity-directed management tool to assess and rank sediments based on their hazard-concept and application ». *Environmental Toxicology and Chemistry, 18* (4) : 685-698.

Breiman, L., J.H. Friedman, R.A. Olshen et S.J. Stone (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Bresler, V., V. Bissinger, A. Abelson, H. Dizer, A. Sturm, R. Kratke, L. Fishelson, P.D. Hansen (1999). « Marine molluscs and fish as biomarkers of pollution stress in littoral regions of the Red Sea, Mediterranean Sea and North Sea ». *Helgoland Marine Research, 53* : 219-243.

Calabrese, E.J., L.A. Baldwin, P.T. Kostecki et T.L. Potter (1997). « A toxicologically based weight-of-evidence methodology for the relative ranking of chemicals of endocrine disruption potential ». *Regulatory Toxicology and Pharmacology, 26* : 36-40.

Costan, G,, N. Bermingham, C. Blaise et J.F. Férard (1993). « Potential ecotoxic effects probe (PEEP): A novel index to assess and compare the toxic potential of industrial effluents ». *Environ. Toxicol. Water Qual., 8* : 115-140.

Couillard, Y., P.G.C. Campbell, A. Tessier, J. Pellerin-Massicotte, J.C. Auclair (1995a). « Field transplantation of a freshwater bivalve, *Pyganodon grandis*, across a metal contamination gradient. I. Temporal changes in metallothionein and metal (Cd, Cu, and Zn) concentrations in soft tissues ». *Can. J. Fish. Aquat.* Sci. *52* : 690-702.

Couillard, Y., P.G.C. Campbell, A. Tessier, J. Pellerin-Massicotte, J.C. Auclair (1995b). « Field transplantation of a freshwater bivalve, *Pyganodon grandis*, across a metal contamination gradient. II. Metallothionein response to Cd and Zn exposure, evidence for cytotoxicity, and links to effects at higher levels of biological organization ». *Can. J. Fish. Aquat. Sci. 52* : 703-715.

DeLafontaine, Y., F. Gagné, C. Blaise, G. Costan, P. Gagnon (2000). « Biomarkers in zebra mussels (*dreissena polymorpha*) for the assessment and monitoring of water quality of the St Lawrence River (Canada) »,*Aquatic Toxicol. 50* : 51-71.

Dougherty, J., R. Kohavi et M. Sahami (1995). « Supervised and unsupervised discretization of continuous features ». In: *Machine Learning: Proceedings of the Twelfth International Conference*. Prieditis A. & Russell S., Morgan Kaufmann Publishers, San Francisco, CA.

Fayyad, U.M. et K.B. Irani (1993). « Multi-interval discretization of the continuous-valued attributes for classification learning », In *Proceedings of the Thirteenth International Joint Conference on the Artificial Intelligence.* Chambery, France, 28 août au 3 septembre, Morgan Kaufmann, pp. 1022-1027.

Fibrak, J., K. Pawlak, K. Slowinski et R. Slowinski (1986). « Rough sets based decision algorithm for treatment of duodenal ulcer by HSV », *Bull, PAS, Biological Series, 34* (10-12) : 227-246.

Grzymala-Busse, J.W. (1992). « LERS: a system for learning from examples based on rough sets ». In *Intelligent Decision Support, Handbook of Application and Advances of Rough Sets Theory.* Slowinski (ed.), Kluwer Academic Publishers, Dordrecht, pp. 3-18.

Krusińska, E., R. Slowinski, J. Stefanowski (1992). « Discriminant versus rough sets approach to vague data analysis ». *Applied stochastic models and data analysis, 8* : 43-56.

Loh, W.-Y. et Y.-S. Shih (1997). « Split selection methods for classification trees ». *Statistica Sinica, 7* :815-840.

Morrison, D.F. (1976). « Multivariate Statistical Methods », 2nd ed. McGraw-Hill Back Compagny, NY, USA.

Malley, D.F., J.F. Klaverkam, S.B. Brown et P.S.S. Chang (1993). « Increase in metallothionein in freshwater mussels *Anodonta grandis grandis* exposed to cadmium in the laboratory and the field ». *Water Poll. Res. J. Canada, 28* : 253-273.

Narbonne, J.M., M. Daubèze, C. Clérandeau et P. Garrigues (1999). « Scale of classification based on biochemical markers in mussels: application to pollution monitoring in European coasts ». *Biomarkers, 4* (6) : 415-424.

Pawlak, Z. (1982). « Rough sets ». *International Journal of Information and Computer Sciences, 11* : 341-356.

Pawlak, K., K. Slowinski, R. Slowinski (1986). « Rough sets classification of patients after highly selective vagotomy for duodenal ulcer ». *International Journal of Man-Machine Studies, 24* : 413-433.

ROSE2 (Rough Set Data Explorer):, http://www-idss.cs.put.poznan.pl/software/rose/

Rossi, L., R. Slowinski et R. Susmaga (1999). « Rough set approach to the evaluation of stormwater pollution ». Accepté pour publication dans : *8th International Conference on Urban Storm Drainage (ICUSD),* Sydney, Australie.

Skowron, A. (1993). « Boolean reasoning for decision rules generation ». In *Methodologies for Intelligent Systems., Lectures Notes in Artificial Intelligence,* Vol. 689, J. Komorowski, Z. W. Ras (eds), Springer-Verlag, Berlin, Allemagne, pp. 295-305.

Slowinski, K., R. Slowinski et J. Stefanowski (1989). « Rough sets approach to analysis of data from peritoneal lavage in acute pancreatitis ». *Medical Informatics, 13* (3) : 143-159.

Slowinski, R. et C. Zopounidis (1994). « Rough set sorting of firms according to bankruptcy risk ». In *Applying Multiple Criteria Aid for Decision to Environmental Management,* M. Paruccini. Kluwer (eds), Dordrecht, Pays-Bas, pp. 339-357.

Thomas, P. (1990). « Molecular and biochemical responses of fish to stressors and their potential use in environmental monitoring ». *American Fisheries Society Symposium, 8* : 9-28.

Vindimian, E., J. Garric, P. Flammarion, E. Thybaud et M Babut (1999). « An index of effluent aquatic toxicity designed by partial least squares regression, using acute and chronic toxicity tests and expert judgements ». *Environmental Toxicology and Chemistry, 18* (10) : 2386-2391.

Annexes

1 Données de biomarqueurs (Blaise *et al.*, à publier)

7 sites : ASE, BE, Baude, ASJ, Era, Barq, PS

7 biomarqueurs : MT, EROD, ADN, LPO, Vn, PHAG, NspE

MT	EROD	ADN	LPO	Vn	PHAG	NspE	SITES
2,07	1,89	1513,00	5,65	194,60	0,46	1,68	ASE
2,23	?	1091,92	7,17	305,20	0,69	1,74	ASE
0,91	?	1493,00	6,33	282,50	1,07	2,04	ASE
1,62	0,67	1241,38	7,05	214,06	1,00	2,64	ASE
1,16	0,23	1398,34	5,94	208,18	0,82	2,59	ASE
1,21	2,05	1269,72	5,20	294,74	0,73	2,17	ASE
1,29	0,62	1303,03	6,60	293,10	0,80	2,57	ASE
1,05	?	1511,76	6,45	301,89	0,54	1,60	ASE
1,32	1,15	1366,80	6,45	307,18	0,26	0,97	ASE
1,72	2,66	997,50	6,03	307,10	0,23	1,34	ASE
1,47	2,55	1122,00	6,50	275,90	0,25	1,51	ASE
1,24	0,04	1409,84	5,70	352,90	0,84	1,78	ASE
2,09	?	1337,90	5,39	355,26	0,53	1,89	ASE
1,59	1,88	1293,86	4,83	354,09	0,48	1,67	ASE
2,42	2,75	1129,44	5,72	328,49	1,05	2,23	ASE
0,69	?	576,80	2,62	94,59	0,39	9,34	BE
0,94	?	971,43	9,83	296,00	7,21	13,57	BE
1,00	?	915,34	8,28	771,33	9,18	2,34	BE
0,93	?	1544,44	9,26	300,00	0,65	0,82	BE
0,76	?	1099,76	5,99	279,07	0,87	9,14	BE
0,81	?	1389,89	8,14	0,00	4,12	7,49	BE
1,17	?	1320,75	9,55	265,73	3,85	1,06	BE
0,91	?	1091,27	8,73	194,74	0,51	0,71	BE
0,66	0,28	1575,86	8,01	262,20	0,33	4,25	BE
0,79	0,28	1084,80	9,93	259,49	2,36	5,59	BE
1,44	0,72	1043,81	7,89	363,64	3,86	1,10	BE
0,71	?	1130,16	4,14	234,15	0,33	1,17	BE
0,22	0,88	1113,45	7,50	0,00	0,22	5,60	BE
0,99	0,43	1510,87	6,00	278,48	1,79	4,59	BE
0,66	?	1400,00	5,59	201,06	2,61	1,94	BE

MT	EROD	ADN	LPO	Vn	PHAG	NspE	SITES
0,52	0,34	675,89	10,53	269,79	0,88	3,25	Baude
0,45	?	1185,81	13,09	261,99	0,81	1,54	Baude
0,38	0,85	1265,40	9,45	167,79	0,90	2,87	Baude
0,62	0,60	1153,33	7,35	267,49	0,91	1,32	Baude
0,66	?	1299,01	14,00	318,81	0,52	1,99	Baude
0,83	0,65	989,90	13,57	185,29	0,66	1,57	Baude
0,39	0,54	1143,25	14,17	244,68	0,74	2,51	Baude
0,65	?	1587,55	16,09	258,06	1,29	3,56	Baude
0,47	?	807,69	9,49	201,42	0,77	2,44	Baude
0,49	0,55	780,35	9,45	310,34	0,77	2,40	Baude
0,74	?	1495,48	12,76	431,19	0,37	1,92	Baude
0,53	?	2694,12	8,04	267,22	0,61	3,19	Baude
0,75	?	1570,18	9,13	350,00	0,82	2,83	Baude
0,69	4,44	1115,12	11,75	253,52	0,51	1,90	Baude
0,93	?	1109,22	11,52	527,93	0,36	1,75	Baude
1,06	0,01	306,25	7,85	812,95	1,36	12,89	ASJ
1,47	0,03	341,46	9,58	597,77	1,52	18,68	ASJ
1,78	0,02	248,37	6,14	798,56	1,42	24,38	ASJ
1,29	?	353,95	6,14	711,86	1,34	19,42	ASJ
0,90	?	403,05	7,46	635,90	1,23	12,31	ASJ
1,02	?	354,10	12,14	401,83	1,48	17,45	ASJ
1,10	?	302,70	9,15	680,63	2,49	21,60	ASJ
1,11	?	299,77	15,54	564,10	1,25	13,30	ASJ
0,84	?	306,12	7,91	659,46	0,92	8,05	ASJ
0,98	?	360,66	9,12	929,91	1,00	12,83	ASJ
1,19	?	324,86	12,70	550,51	1,02	12,02	ASJ
1,18	?	418,67	17,05	328,21	1,64	12,86	ASJ
1,06	?	471,94	6,21	1104,90	0,58	14,18	ASJ
1,11	?	335,46	15,03	1541,67	1,13	18,97	ASJ
1,64	?	136,84	9,44	1100,92	0,97	11,79	ASJ

MT	EROD	ADN	LPO	Vn	PHAG	NspE	SITES
2,08	4,23	802,85	23,81	578,31	2,24	9,90	Era
2,12	2,13	751,63	31,50	175,82	1,84	8,62	Era
2,20	1,56	425,74	71,67	0,00	1,10	10,78	Era
2,30	1,50	537,50	40,38	223,88	1,08	5,67	Era
2,37	1,41	495,05	60,78	1178,57	0,64	7,83	Era
1,93	?	560,34	106,55	217,11	1,13	9,44	Era
1,91	6,21	718,65	46,34	83,72	2,04	12,29	Era
2,31	0,79	490,20	49,23	257,14	1,53	5,74	Era
2,05	?	870,86	46,13	900,83	0,54	4,66	Era
2,12	0,24	634,24	63,30	372,97	0,85	8,50	Era
2,20	?	524,93	49,19	248,45	2,11	8,69	Era
2,23	0,91	679,20	46,01	232,39	1,37	8,92	Era
2,54	4,07	668,46	50,26	220,78	0,78	4,34	Era
2,37	3,25	667,71	76,05	428,57	1,11	7,34	Era
1,17	?	674,33	76,76	589,74	1,27	4,27	Era
0,51	0,70	272,11	9,77	308,82	0,65	6,27	Barq
0,70	0,08	569,06	5,84	246,45	0,87	5,26	Barq
0,81	?	825,87	17,09	297,30	0,41	6,12	Barq
0,56	?	?	11,42	323,53	0,19	3,16	Barq
0,38	?	1784,81	9,53	254,62	1,30	5,47	Barq
0,53	2,01	519,91	7,81	253,97	1,16	4,95	Barq
0,58	0,89	444,44	10,08	390,35	0,65	5,76	Barq
0,48	?	631,74	8,29	133,33	0,37	4,73	Barq
0,41	0,62	461,35	8,14	241,53	0,74	3,27	Barq
0,46	1,88	468,14	7,45	273,08	0,65	4,12	Barq
0,54	?	426,33	6,67	266,95	0,57	4,67	Barq
0,53	?	635,61	10,19	203,19	0,83	2,45	Barq
0,42	1,94	865,89	13,00	156,57	0,74	4,13	Barq
0,47	1,27	609,30	8,80	?	0,62	4,34	Barq
1,40	0,69	392,22	8,37	281,50	0,33	2,27	Barq

MT	EROD	ADN	LIPO	Vn	PHAG	NspE	SITES
1,16	?	221,45	2,45	0,00	0,30	0,66	PS
1,37	?	393,00	7,52	65,87	1,60	1,42	PS
1,84	1,63	208,14	2,87	50,93	1,93	2,69	PS
0,90	?	323,72	3,04	65,87	1,43	2,05	PS
1,07	0,27	274,59	4,06	137,25	1,07	1,25	PS
1,02	0,27	293,76	3,14	293,19	0,98	2,15	PS
1,35	0,24	256,04	3,69	129,31	1,01	2,02	PS
1,45	?	188,73	4,44	0,00	0,61	1,46	PS
1,21	0,49	326,86	3,83	1256,54	0,69	1,45	PS
1,64	0,81	478,62	3,01	180,77	1,30	1,41	PS
1,38	1,80	483,01	3,38	106,06	0,52	1,38	PS
1,56	?	301,12	2,23	150,26	0,79	1,12	PS
1,08	?	210,19	2,84	385,19	0,24	0,81	PS
1,78	0,85	383,65	3,45	178,77	0,73	1,59	PS
1,55	?	285,21	3,46	280,99	0,62	1,26	PS

2 Données de biomarqueurs discrétisées

Discrétisation locale non supervisée
7 sites : ASE, BE, Baude, ASJ, Era, Barq, PS
6 biomarqueurs : MT, ADN, LPO, Vn, PHAG, NspE

MT	ADN	LPO	Vn	PHAG	NspE	SITES
2	2	1	0	0	1	ASE
2	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	0	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
2	2	1	0	0	1	ASE
1	2	1	0	0	1	ASE
2	2	1	0	0	1	ASE
0	1	0	0	0	2	BE
1	2	2	0	0	3	BE
1	2	2	0	0	0	BE
0	2	1	0	0	2	BE
0	2	2	0	0	2	BE
1	2	2	0	0	0	BE
1	2	2	0	0	0	BE
0	2	2	0	0	2	BE
0	2	2	0	0	2	BE
1	2	2	0	0	0	BE
0	2	0	0	0	0	BE
0	2	2	0	0	2	BE
1	2	1	0	0	2	BE
0	2	1	0	0	1	BE

MT	APA	LPO	Vn	PHAG	EST	SITES
0	1	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
0	1	2	0	0	1	Baude
0	1	2	0	0	1	Baude
0	2	2	2	0	1	Baude
0	2	2	0	0	1	Baude
0	2	2	0	0	1	Baude
1	2	2	2	0	1	Baude
1	0	2	3	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	1	3	0	3	ASJ
1	0	1	3	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	2	1	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	2	3	0	2	ASJ
1	0	2	3	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	2	0	0	3	ASJ
1	0	1	3	0	3	ASJ
1	0	2	3	0	3	ASJ
1	0	2	3	0	3	ASJ

MT	APA	LPO	Vn	PHAG	EST	SITES
2	1	3	3	0	2	Era
2	1	3	0	0	2	Era
2	0	3	0	0	2	Era
2	1	3	0	0	2	Era
2	1	3	3	0	2	Era
2	1	3	0	0	2	Era
2	1	3	0	0	3	Era
2	1	3	0	0	2	Era
2	1	3	3	0	2	Era
2	1	3	0	0	2	Era
2	1	3	0	0	2	Era
2	1	3	0	0	2	Era
2	1	3	0	0	2	Era
2	1	3	1	0	2	Era
1	1	3	3	0	2	Era
0	0	2	0	0	2	Barq
0	1	1	0	0	2	Barq
0	1	2	0	0	2	Barq
0	1	2	0	0	1	Barq
0	1	2	0	0	2	Barq
0	0	2	0	0	2	Barq
0	1	2	0	0	2	Barq
0	0	2	0	0	1	Barq
0	0	2	0	0	2	Barq
0	0	1	0	0	2	Barq
0	1	2	0	0	1	Barq
0	1	2	0	0	2	Barq
0	1	2	0	0	2	Barq

MT	PAPA	LPO	Vn	PHAG	EST	SITES
1	0	0	0	0	0	PS
1	0	0	0	0	1	PS
1	0	0	0	0	1	PS
1	0	0	0	0	0	PS
1	0	0	0	0	1	PS
1	0	0	0	0	1	PS
1	0	0	0	0	1	PS
1	0	0	0	0	1	PS
1	0	0	0	0	1	PS
1	0	0	0	0	0	PS
1	0	0	0	0	0	PS
1	0	0	0	0	1	PS
1	0	0	0	0	0	PS

3 Atomes, approximations, règles et classification

Données sans EROD, valeurs aberrantes supprimées

1. Atomes et approximations

Atomes : 33
Qualité totale : 0,9495
Exactitude totale : 0,9038

	Site						
	ASE	**BE**	**Baude**	**ASJ**	**Era**	**Barq**	**PS**
Nb, Objets	15	14	14	15	15	13	13
App, Inf,	15	14	11	15	15	11	13
App, Sup,	15	14	16	15	15	16	13
Exactitude	1,00	1,00	0,69	1,00	1,00	0,69	1,00

2. Règles

Explications :
n° de règle. Description. [nombre d'objets appartenant à la règle, nombre d'objets qui appartiennent à la règle et au site ou **force de la règle**, pourcentage d'objets du site classés ou **force relative de la règle**, pourcentage d'objets appartenant à cette règle classés ou **niveau de discrimination**] [nombre d'objets concernés pour chacun des sites] [n° des objets concernés pour chaque site]

règle 1. (LPO = 1) & (NspE= 1) & (MT = 1) => (Site=ASE); [10, 10, 66,67%, 100,00%][10, 0, 0, 0, 0, 0, 0] [{3,4,5,6,7,8,10,11,12,14},{},{},{},{},{},{}]

règle 2. (LPO = 1) & (MT = 2) => (Site=ASE); [4, 4, 26,67%, 100,00%][4, 0, 0, 0, 0, 0, 0] [{1,2,13,15},{},{},{},{},{},{}]

règle 3. (NspE = 0) & (LPO = 1) => (Site=ASE); [1, 1, 6,67%, 100,00%][1, 0, 0, 0, 0, 0, 0] [{9},{},{},{},{},{},{}]

règle 4. (Vn = 0) & (ADN = 2) & (LPO = 2) & (MT = 1) => (Site=BE); [5, 5, 35,71%, 100,00%][0, 5, 0, 0, 0, 0, 0] [{},{17,18,21,22,25},{},{},{},{},{}]

règle 5. (ADN = 2) & (NspE = 2) => (Site=BE); [6, 6, 42,86%, 100,00%][0, 6, 0, 0, 0, 0, 0] [{},{19,20,23,24,27,28},{},{},{},{},{}]

règle 6. (MT = 0) & (LPO = 0) => (Site=BE); [2, 2, 14,29%, 100,00%][0, 2, 0, 0, 0, 0, 0] [{},{16,26},{},{},{},{},{}]

règle 7. (LPO = 1) & (MT = 0) & (ADN = 2) => (Site=BE); [2, 2, 14,29%; 100,00%][0, 2, 0, 0, 0, 0, 0] [{},{19,29},{},{},{},{},{}]

règle 8. (ADN = 2) & (NspE = 1) & (LPO = 2) => (Site=Baude); [11, 11, 78,57%, 100,00%][0, 0, 11, 0, 0, 0, 0] [{},{},{31,32,33,34,35,36,37,40,41,42,43},{},{},{},{}]

règle 9. (ADN = 0) & (NspE = 3) => (Site=ASJ); [14, 14, 93,33%, 100,00%][0, 0, 0, 14, 0, 0, 0] [{},{},{},{44,45,46,47,48,49,50,51,53,54,55,56,57,58},{},{},{}]

règle 10. (Vn = 3) & (ADN = 0) => (Site=ASJ); [13, 13, 86,67%, 100,00%][0, 0, 0, 13, 0, 0, 0] [{},{},{},{44,45,46,47,48,50,51,52,53,54,56,57,58},{},{},{}]

règle 11. (LPO = 3) => (Site=Era); [15, 15, 100,00%, 100,00%][0, 0, 0, 0, 15, 0, 0] [{},{},{},{},{59,60,61,62,63,64,65,66,67,68,69,70,71,72,73},{},{}]

règle 12. (NspE = 2) & (LPO = 2) & (ADN = 1) => (Site=Barq); [5, 5, 38,46%, 100,00%][0, 0, 0, 0, 0, 5, 0] [{},{},{},{},{},{76,78,80,85,86},{}]

règle 13. (MT = 0) & (ADN = 0) => (Site=Barq); [5, 5, 38,46%, 100,00%][0, 0, 0, 0, 0, 5, 0] [{},{},{},{},{},{74,79,81,82,83},{}]

règle 14. (LPO = 1) & (ADN = 1) => (Site=Barq); [1, 1, 7,69%, 100,00%][0, 0, 0, 0, 0, 1, 0] [{},{},{},{},{},{75},{}]

règle 15. (LPO = 0) & (ADN = 0) => (Site=PS); [13, 13, 100,00%, 100,00%][0, 0, 0, 0, 0, 0, 13] [{},{},{},{},{},{},{87,88,89,90,91,92,93,94,95,96,97,98,99}]

Règles approximatives

règle 16, (ADN = 1) & (NspE = 1) => (Site=Baude) OR (Site=Barq); [5, 5, 100,00%, 100,00%][0, 0, 3, 0, 0, 2, 0] [{},{},{30,38,39},{},{},{77,84},{}]

4 Calcul de l'indice avec ratio (référence site Baude)

Résumé

	Médiane	Quartiles	
ASE	6,8	6,3	7,4
Baude	5,8	5,6	6,5
Barq	5,8	5,3	6,7
BE	8,3	6,1	10,2
Era	15,6	14,0	16,3
ASJ	14,0	13,0	16,2
PS	5,0	4,6	5,5

Tableau complet

MT	ADN	LPO	Vn	PHAG	NspE	Sites	Indice
3,26	1,32	0,49	0,74	0,60	0,77	ASE	7,16
3,51	0,95	0,62	1,15	0,90	0,79	ASE	7,92
1,44	1,30	0,54	1,07	1,40	0,93	ASE	6,68
2,56	1,08	0,61	0,81	1,31	1,21	ASE	7,57
1,82	1,22	0,51	0,79	1,07	1,18	ASE	6,59
1,91	1,11	0,45	1,11	0,95	0,99	ASE	6,51
2,04	1,13	0,57	1,11	1,04	1,18	ASE	7,06
1,66	1,32	0,55	1,14	0,70	0,73	ASE	6,09
2,08	1,19	0,55	1,16	0,33	0,44	ASE	5,76
2,71	0,87	0,52	1,16	0,30	0,61	ASE	6,16
2,32	0,98	0,56	1,04	0,33	0,69	ASE	5,92
1,95	1,23	0,49	1,33	1,09	0,81	ASE	6,90
3,30	1,17	0,46	1,34	0,69	0,86	ASE	7,82
2,51	1,13	0,41	1,34	0,62	0,76	ASE	6,78
3,82	0,98	0,49	1,24	1,37	1,02	ASE	8,93

MT	ADN	LPO	Vn	PHAG	NspE	Sites	Indice
1,09	0,50	0,23	0,36	0,51	4,26	BE	**6,95**
1,48	0,85	0,85	1,12	9,41	6,20	BE	**19,90**
1,46	1,34	0,80	1,13	0,84	0,37	BE	**5,95**
1,20	0,96	0,51	1,05	1,14	4,17	BE	**9,04**
1,28	1,21	0,70	0,00	5,38	3,42	BE	**11,99**
1,84	1,15	0,82	1,00	5,02	0,48	BE	**10,32**
1,43	0,95	0,75	0,74	0,66	0,32	BE	**4,85**
1,04	1,37	0,69	0,99	0,43	1,94	BE	**6,47**
1,25	0,94	0,85	0,98	3,08	2,55	BE	**9,66**
2,26	0,91	0,68	1,37	5,04	0,50	BE	**10,77**
1,11	0,98	0,36	0,88	0,43	0,54	BE	**4,30**
0,35	0,97	0,64	0,00	0,28	2,56	BE	**4,80**
1,56	1,32	0,52	1,05	2,34	2,10	BE	**8,87**
1,04	1,22	0,48	0,76	3,41	0,88	BE	**7,80**
0,82	0,59	0,90	1,02	1,14	1,48	Baude	**5,96**
0,71	1,03	1,12	0,99	1,05	0,70	Baude	**5,61**
0,60	1,10	0,81	0,63	1,17	1,31	Baude	**5,63**
0,97	1,00	0,63	1,01	1,19	0,60	Baude	**5,42**
1,04	1,13	1,20	1,20	0,68	0,91	Baude	**6,17**
1,30	0,86	1,17	0,70	0,86	0,72	Baude	**5,61**
0,62	1,00	1,22	0,92	0,96	1,15	Baude	**5,86**
1,03	1,38	1,38	0,97	1,68	1,63	Baude	**8,07**
0,74	0,70	0,82	0,76	1,00	1,11	Baude	**5,14**
0,77	0,68	0,81	1,17	1,00	1,09	Baude	**5,52**
1,16	1,30	1,10	1,63	0,48	0,88	Baude	**6,54**
1,19	1,37	0,78	1,32	1,07	1,29	Baude	**7,03**
1,09	0,97	1,01	0,96	0,67	0,87	Baude	**5,56**
1,46	0,97	0,99	1,99	0,46	0,80	Baude	**6,68**

MT	ADN	LPO	Vn	PHAG	NspE	Sites	Indice
1,67	0,27	0,67	3,07	1,77	5,89	ASJ	**13,34**
2,32	0,30	0,82	2,26	1,98	8,53	ASJ	**16,21**
2,80	0,22	0,53	3,02	1,85	11,13	ASJ	**19,54**
2,04	0,31	0,53	2,69	1,75	8,87	ASJ	**16,18**
1,41	0,35	0,64	2,40	1,60	5,62	ASJ	**12,03**
1,61	0,31	1,04	1,52	1,93	7,97	ASJ	**14,38**
1,73	0,26	0,79	2,57	3,25	9,86	ASJ	**18,46**
1,75	0,26	1,34	2,13	1,63	6,07	ASJ	**13,19**
1,33	0,27	0,68	2,49	1,21	3,67	ASJ	**9,65**
1,54	0,31	0,78	3,51	1,30	5,86	ASJ	**13,31**
1,88	0,28	1,09	2,08	1,32	5,49	ASJ	**12,14**
1,85	0,36	1,46	1,24	2,13	5,87	ASJ	**12,93**
1,67	0,41	0,53	4,17	0,76	6,48	ASJ	**14,02**
1,75	0,29	1,29	5,82	1,47	8,66	ASJ	**19,29**
2,59	0,12	0,81	4,16	1,26	5,38	ASJ	**14,32**
3,27	0,70	2,05	2,18	2,92	4,52	Era	**15,63**
3,34	0,65	2,71	0,66	2,40	3,94	Era	**13,70**
3,46	0,37	6,16	0,00	1,44	4,92	Era	**16,35**
3,62	0,47	3,47	0,85	1,40	2,59	Era	**12,39**
3,74	0,43	5,22	4,45	0,83	3,58	Era	**18,25**
3,05	0,49	9,15	0,82	1,48	4,31	Era	**19,30**
3,00	0,63	3,98	0,32	2,66	5,61	Era	**16,19**
3,64	0,43	4,23	0,97	1,99	2,62	Era	**13,88**
3,24	0,76	3,96	3,40	0,71	2,13	Era	**14,19**
3,35	0,55	5,44	1,41	1,11	3,88	Era	**15,74**
3,47	0,46	4,23	0,94	2,75	3,97	Era	**15,81**
3,51	0,59	3,95	0,88	1,79	4,07	Era	**14,80**
4,00	0,58	4,32	0,83	1,02	1,98	Era	**12,73**
3,73	0,58	6,53	1,62	1,45	3,35	Era	**17,26**
1,85	0,59	6,60	2,23	1,66	1,95	Era	**14,87**

MT	ADN	LPO	Vn	PHAG	NspE	Sites	Indice
0,81	0,24	0,84	1,17	0,84	2,86	Barq	6,75
1,10	0,50	0,50	0,93	1,14	2,40	Barq	6,57
1,27	0,72	1,47	1,12	0,53	2,80	Barq	7,91
0,89	0,47	0,98	1,22	0,25	1,44	Barq	5,26
0,83	0,45	0,67	0,96	1,51	2,26	Barq	6,69
0,91	0,39	0,87	1,47	0,85	2,63	Barq	7,11
0,76	0,55	0,71	0,50	0,49	2,16	Barq	5,17
0,65	0,40	0,70	0,91	0,97	1,49	Barq	5,12
0,72	0,41	0.64	1.03	0.85	1.88	Barq	5.54
0.85	0.37	0.57	1.01	0.74	2.13	Barq	5.67
0.83	0.55	0.88	0.77	1.09	1.12	Barq	5.23
0.66	0.75	1.12	0.59	0.97	1.89	Barq	5.97
0.74	0.53	0.76	0.99	0.81	1.98	Barq	5.80
1.83	0.19	0.21	0.00	0.39	0.30	PS	2.92
2.90	0.18	0.25	0.19	2.52	1.23	PS	7.27
1.42	0.28	0.26	0.25	1.87	0.94	PS	5.02
1.69	0.24	0.35	0.52	1.40	0.57	PS	4.77
1.60	0.26	0.27	1.11	1.28	0.98	PS	5.49
2.13	0.22	0.32	0.49	1.32	0.92	PS	5.40
2.29	0.16	0.38	0.00	0.79	0.67	PS	4.30
2.58	0.42	0.26	0.68	1.70	0.64	PS	6.27
2.18	0.42	0.29	0.40	0.68	0.63	PS	4.60
2.46	0.26	0.19	0.57	1.03	0.51	PS	5.03
1.71	0.18	0.24	1.45	0.32	0.37	PS	4.28
2.81	0.33	0.30	0.68	0.95	0.72	PS	5.78
2.44	0.25	0.30	1.06	0.81	0.57	PS	5.43

5 Calcul de l'indice avec classes non paramétriques

Résumé

	Médiane	Quartiles	
ASE	11	11	11.5
Baude	11	11	11
Barq	10	10	11
BE	11	11	12
Era	14	14	15.5
ASJ	15	14	15
PS	8	7	8

Tableau complet

MT	ADN	LPO	Vn	PHAG	NspE	Site	Indice
3	3	2	1	1	2	ASE	**12**
3	3	2	1	1	2	ASE	**12**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	1	ASE	**10**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
2	3	2	1	1	2	ASE	**11**
3	3	2	1	1	2	ASE	**12**
2	3	2	1	1	2	ASE	**11**
3	3	2	1	1	2	ASE	**12**

MT	ADN	LPO	Vn	PHAG	NspE	Site	Indice
1	2	1	1	1	3	BE	9
2	3	3	1	1	4	BE	14
2	3	3	1	1	1	BE	11
1	3	2	1	1	3	BE	11
1	3	3	1	1	3	BE	12
2	3	3	1	1	1	BE	11
2	3	3	1	1	1	BE	11
1	3	3	1	1	3	BE	12
1	3	3	1	1	3	BE	12
2	3	3	1	1	1	BE	11
1	3	1	1	1	1	BE	8
1	3	3	1	1	3	BE	12
2	3	2	1	1	3	BE	12
1	3	2	1	1	2	BE	10
1	2	3	1	1	2	Baude	10
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
1	2	3	1	1	2	Baude	10
1	2	3	1	1	2	Baude	10
1	3	3	3	1	2	Baude	13
1	3	3	1	1	2	Baude	11
1	3	3	1	1	2	Baude	11
2	3	3	3	1	2	Baude	14

MT	ADN	LPO	Vn	PHAG	NspE	Site	Indice
2	1	3	4	1	4	ASJ	15
2	1	3	4	1	4	ASJ	15
2	1	2	4	1	4	ASJ	14
2	1	2	4	1	4	ASJ	14
2	1	3	4	1	4	ASJ	15
2	1	3	2	1	4	ASJ	13
2	1	3	4	1	4	ASJ	15
2	1	3	4	1	4	ASJ	15
2	1	3	4	1	3	ASJ	14
2	1	3	4	1	4	ASJ	15
2	1	3	4	1	4	ASJ	15
2	1	3	1	1	4	ASJ	12
2	1	2	4	1	4	ASJ	14
2	1	3	4	1	4	ASJ	15
2	1	3	4	1	4	ASJ	15
3	2	4	4	1	3	Era	17
3	2	4	1	1	3	Era	14
3	1	4	1	1	3	Era	13
3	2	4	1	1	3	Era	14
3	2	4	4	1	3	Era	17
3	2	4	1	1	3	Era	14
3	2	4	1	1	4	Era	15
3	2	4	1	1	3	Era	14
3	2	4	4	1	3	Era	17
3	2	4	1	1	3	Era	14
3	2	4	1	1	3	Era	14
3	2	4	1	1	3	Era	14
3	2	4	1	1	3	Era	14
3	2	4	2	1	3	Era	15
2	2	4	4	1	3	Era	16

MT	ADN	LPO	Vn	PHAG	NspE	Sites	Indice
1	1	3	1	1	3	Barq	10
1	2	2	1	1	3	Barq	10
1	2	3	1	1	3	Barq	11
1	2	3	1	1	2	Barq	10
1	2	3	1	1	3	Barq	11
1	1	3	1	1	3	Barq	10
1	2	3	1	1	3	Barq	11
1	1	3	1	1	2	Barq	9
1	1	3	1	1	3	Barq	10
1	1	2	1	1	3	Barq	9
1	2	3	1	1	2	Barq	10
1	2	3	1	1	3	Barq	11
1	2	3	1	1	3	Barq	11
2	1	1	1	1	1	PS	7
2	1	1	1	1	2	PS	8
2	1	1	1	1	2	PS	8
2	1	1	1	1	1	PS	7
2	1	1	1	1	2	PS	8
2	1	1	1	1	2	PS	8
2	1	1	1	1	2	PS	8
2	1	1	1	1	2	PS	8
2	1	1	1	1	2	PS	8
2	1	1	1	1	1	PS	7
2	1	1	1	1	1	PS	7
2	1	1	1	1	2	PS	8
2	1	1	1	1	1	PS	7