# Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols.

S. Taibi-Hassani(1), J.-C. Thoisy-Dur(2), P. Lepelletier(1), J. Bodin(1), N. Bennegadi-Laurent(1), J.-J. Bessoule(3), A. Bispo(4), J. Bodilis(5), R. Chaussod(6), N. Cheviron(2), J. Cortet(7), S. Criquet(8), J. Dantan(1), S. Dequiedt(9), O. Faure(10), C. Gangneux(1), J. Harris-Hellal(11), M. Hedde(2), A. Hitmi(12), M. Le Guedard(13), M. Legras(1), G. Pérès(14), C. Repinçay(2), L. Rougé(14), N. Ruiz(15), I. Trinsoutrot-Gattin(1) et C. Villenave(16)

(1) ESITPA, Agri'Terr, CS 40118, 76134 Mont Saint Aignan Cedex
(2) INRA Versailles-Grignon PESSAC UR 251, 78026 Versailles Cedex
(3) Université Victor Segalen Bordeaux 2, 146, rue Léo-Saignât, 33076 Bordeaux Cedex
(4) ADEME, 20 avenue du Grésillé, B P 90 406, 49004 Angers Cedex 01
(5) Université de Rouen, Laboratoire de Microbiologie Signaux et Microenvironnement, EA 4312, 76821 Mont Saint Aignan
(6) INRA Université de Bourgogne, UMR BP 86510, 21065 Dijon Cedex
(7) UMR UL/INRA, 1120, Vandoeuvre-les-Nancy
(8) IMBE, UMR CNRS 7263, Université Aix-Marseille, 13397 Marseille Cedex
(9) UMR Agroécologie - plateforme GenoSol, INRA de Dijon, BP 86510, 21065 Dijon Cedex
(10) Equipe Geosciences & Environnement Ecole Nationale des Mines de St Etienne, 42023 St Etienne Cedex 02
(11) BRGM, Equipe de Biogéochimie Environnementale, 3, avenue Claude Guillemin, BP36009, 45060 Orléans Cedex 02
(12) UMR 547 PIAF IUT Clermont-Ferrand Université d'Auvergne 100 rue de l'égalité, 15013 Aurillac
(13) LEB Aquitaine Transfert – ADERA UMR 5200, INRA Bordeaux Aquitaine, 33883 Villenave D'Ornon Cedex
(14) Université Rennes 1, UMR 6553 CNRS "Ecobio", Station Biologique, 35380 Paimpont
(15) IRD, Centre France-Nord, UMR 211 BIOEMCO, 93143 Bondy Cedex
(16) ELISOL environnement 34060 Montpellier Cedex 02

*: Auteur correspondant: staibi@esitpa.fr

Reçu: janvier 2013; Accepté: septembre 2013

**RÉSUMÉ**

Le volume des données définies dans le programme Bioindicateurs 2 (Ademe) et le très grand nombre de variables biologiques à tester (une centaine) nécessitent des techniques d'analyse telles que les Random Forests qui peuvent s'affranchir du problème de multi-colinéarité pour la sélection d'indicateurs sensibles aux différents facteurs étudiés.

La méthodologie des Random Forests consiste en la sélection des variables les plus discriminantes. Ainsi nous avons recherché la meilleure sélection en étudiant l'ensemble des variables biologiques représentant la Microflore et la Faune. Cette démarche a porté sur l'ensemble des indicateurs d'effet issus du programme Bio2, les indicateurs de la flore et d'accumulation (escargot) n'ayant pas été traités. Ces travaux ont été mis en œuvre sur les trois facteurs de discrimination: l'usage des sols, les niveaux de contamination en ETM, et les niveaux de contamination en polluants organiques.

Nous avons ensuite regroupé les variables les plus discriminantes issues de chaque analyse par RF. Une analyse discriminante linéaire a ensuite été mise en œuvre pour chaque facteur en vue d'élaborer un modèle prédictif.

**Mots clés**

Random Forests, Analyse discriminante, bioindicateurs, sélection, polluants organiques, ETM, occupation des sols.

***SUMMARY***

***STATISTICAL APPROACH TO SELECT BIOINDICATORS BY RANDOM FORESTS FOR SOIL MONITORING***

*The volume of data, and the large number of biological variables to be tested (one hundred), require analytical techniques, such as Random Forests, which can overcome the problem of multi-colinearity for the selection of indicators, sensitive to various factors.*

*Random Forests methodology is appropriate for the selection of the most discriminant variables. So, we searched for the best way to select them, by bringing together all biological variables, representing the Microflora and Fauna. This approach focuses on impact indicators from the Bio2 program, indicators of flora and indicators of accumulation (snails) were not included.*

*This work has been implemented on the three factors of discrimination: land use, metallic contamination levels and organic contamination levels.*

*We grouped the most discriminating variables from each RF analysis. Linear discriminant analysis was then implemented for each factor, in order to develop a predictive model.*

***Key-words***

*Random Forests, Discriminant Analysis, bioindicators, organic pollutants, metal contamination, land use.*

***RESUMEN***

***ENFOQUE ESTADÍSTICA PARA LA SELECCIÓN DE LOS INDICADORES POR RANDOM FORESTS PARA LA VIGILANCIA DE LA CALIDAD DEL SUELO***

*El volumen de datos definidos en el programa bioindicadores 2 (Ademe) y el muy grande numero de variables biológicas para probar (una centena) necesitan técnicas de análisis como los Random Forests que pueden liberarse del problema de multicolinealidad para la selección de indicadores sensibles a los diferentes factores estudiados.*

*La metodología de Random Forests consiste en la selección de variables las más discriminantes. Así buscamos la mejor selección agrupando el conjunto de las variables biológicas que representan la Microflora y la Fauna. Estos trabajos se realizaron sobre los tres factores de discriminación: el uso de los suelos, los niveles de contaminación en ETM y los niveles de contaminación en contaminantes orgánicos.*

*Luego, agrupamos las variables las más discriminantes derivadas de cada análisis por RF. Un análisis discriminante linear se realizo después para cada factor con vista a elaborar un modelo predictivo. Se observaron los indicadores del grupo Flora únicamente sobre un sub-conjunto de 47 parcelas de modalidades contrastadas, así no los incluimos en nuestro estudio. Las variables "estandarizadas" del grupo Flora podrán estar integradas en un segundo tiempo.*

***Palabras clave***

*Random Forests, análisis discriminante, bioindicadores, selección, contaminantes orgánicos, ETM, uso del suelo.*

Le programme national Bioindicateurs utilise les résultats acquis par plus de 20 laboratoires sur 47 placettes de prélèvement (Pérès *et al.,* 2011, Pérès *et al.,* 2012, http://ecobiosoil.univ-rennes1.fr/ADEME-Bioindicateur).

La stratégie du traitement des données du programme national Bioindicateurs (2008-2012) repose sur (i) des mesures expérimentales d'un grand nombre de bioindicateurs dans des contextes pédo-géo-climatiques contrastés sur l'ensemble du territoire français, en utilisant le même protocole d'échantillonnage, (ii) la gestion de l'ensemble des données biologiques, physico-chimiques et des métadonnées dans une base de données relationnelle, (iii) le développement, par l'équipe de biostatisticiens, du datamining des données collectées au cours des années 2009 à 2011.

Le principal objectif de cet article est de développer une méthodologie adaptée pour sélectionner les ensembles de bioindicateurs pertinents, au regard du paramètre à prédire tels que l'usage des sols et les diverses contaminations, pour l'évaluation de risques pour les sols dans les écosystèmes.

Compte tenu du nombre très important de données (>200.000), il s'est très vite avéré nécessaire d'harmoniser et de centraliser les résultats obtenus à travers la création d'une base de données accessible à tous. Celle-ci a ainsi permis la mise à disposition des résultats pour l'ensemble des équipes en vue de croiser les données obtenues par des laboratoires différents et de faciliter la manipulation et l'analyse des données (Cluzeau *et al.,* 2008).

L'objectif des traitements de données exposés ici est de hiérarchiser les bioindicateurs en fonction de leur sensibilité aux facteurs environnementaux et aux perturbations (contaminations et usage du sol) d'une part, et de proposer un indicateur agrégé de la réponse à ces facteurs d'autre part. Cet article décrit brièvement la méthodologie statistique proposée (par le groupe Biomath) pour atteindre ces objectifs. Notons que pour les indicateurs du groupe Flore ainsi que pour les indicateurs d'accumulation (escargot), d'autres approches ont été développées (Le Guedart *et al.* 2013, Remon *et al.* 2013, Pauget *et al.* 2013).

# MATÉRIELS ET MÉTHODES

## La méthode des Random Forests (RF)

La méthode des Random Forests est récente. Elle a été développée par Breiman (2001). Elle trouve déjà de nombreuses applications dans différents domaines telles que l'écologie (Brosteaux, 2005) ou l'agriculture (Cutler *et al.,* 2007). Parmi les performances des RF nous pouvons citer la possibilité de s'affranchir du problème de multi-colinéarité et de la présence de données manquantes, de permettre le traitement à la fois des variables qualitatives et quantitatives (Laroutis et Taibi, 2011) et surtout de gérer le cas des grandes bases de données.

L'objectif de l'étude par Random Forests est de détecter l'importance des variables dans un jeu de données. Attribuer une importance aux variables explicatives $X_i$ nous a conduits à analyser la sensibilité de ces variables pour une variable à expliquer fixée *Y*. L'importance des variables est donnée par l'indice de Gini, plus cet indice est élevé plus la variable *Xi* est importante.

Dans le cas présent *Y* décrit l'usage du sol ou le niveau de contamination en ETM ou encore le niveau de pollution des produits organiques. Dans le cas des données de l'ensemble du programme, le but de l'étude est d'affecter à la variable *Xi* une probabilité d'occurrence.

## Principe

L'analyse par RF fait partie de l'ensemble des méthodes d'apprentissage statistique. Elle consiste à (i) montrer si le prédicteur de *Y* choisi *a priori* est bon (ii) sélectionner les variables les plus discriminantes ou prédictrices. Basée sur le principe du Bagging, l'originalité de l'analyse est l'agrégation de *K* arbres construits avec un faible nombre de variables. Chaque nœud est construit avec un faible nombre de variables mais toujours constant et choisi aléatoirement.

Plus précisément, on construit plusieurs modèles indépendants pour prédire la même variable *Y*. On agrège ensuite les prédictions des modèles bâtis. Agréger les prédictions de plusieurs modèles indépendants permet de réduire la variance et donc de réduire l'erreur de prédiction. Toutefois, dans la pratique, les échantillons sont rarement indépendants. Pour pallier cet inconvénient, on remplace ces échantillons par des échantillons bootstrap obtenus par *n* tirages avec remise à partir de l'échantillon initial.

## Détails sur l'élagage dans le cadre des forêts aléatoires

On peut se limiter à des arbres de taille *q* relativement réduite, voire même triviale avec $q = 2$. La sélection aléatoire d'un nombre réduit de prédicteurs potentiels à chaque étape de construction d'un arbre accroît significativement la variabilité en mettant en avant nécessairement d'autres variables. Chaque modèle est « moins bon » mais l'agrégation conduit finalement à de meilleurs résultats. Le nombre de variables tirées aléatoirement n'est pas un paramètre auquel RF est sensible. Breiman suggère d'utiliser $q = p^{½}$ par défaut.

La méthode statistique de quantification de l'importance des variables, en d'autres termes de l'importance de leur sensibilité au facteur, suit une stratégie de recherche de toutes les variables importantes afin de diagnostiquer l'ensemble des variables importantes. Pour cela, la stratégie est de procéder

à des permutations des variables étudiées et de mesurer l'augmentation de l'erreur éventuelle.

L'interprétation est ensuite facilitée par le calcul d'un indice donnant l'importance de chaque variable dans l'agrégation de modèles. Le package R (Breiman et Cutler, 2005) utilisé est « randomForest », et nécessite de fixer 3 paramètres: ntree, mtry et maxnodes. Nous avons fixé:

• ntree: nombre d'arbres de la forêt = 1000
• mtry: nombre fixe de variables choisies aléatoirement = 10,
• maxnodes: nombre maximal de nœuds par arbres = 6.

## Méthodologie pour les variables biologiques

L'objectif étant de définir des batteries d'indicateurs sensibles à des facteurs anthropiques, un travail de définition des facteurs à tester a été mené. Les modalités des sites ont été définies selon des gradients de contaminations organique et métallique. Les classes de contaminations métalliques ont été caractérisées en combinant les valeurs des vibrisses internes et externes des différents éléments métalliques d'après les données du RMQS (Villanneau *et al.,* 2008) et les données INRA-ASPITET pour l'arsenic (Baize, 2000). Les classes de contamination organique ont été définies par une classification ascendante hiérarchique selon la méthode de Ward, appliquée sur l'ensemble des sites. Le niveau des contaminations en ETM ou en produits organiques est classé en trois groupes: Forte, Moyenne, Faible. Les usages du sol ont été déterminés sur la base du Corine Land Cover distinguant cultures (C), prairies (P), forêts (F) et bois, friches et ourlets boisés (B).

Dans le but de sélectionner des indicateurs sensibles, nous avons cherché à mettre en évidence les réponses significatives de 74 variables biologiques à une perturbation liées à la contamination chimique ou à un usage du sol. Les tests sont réalisés sur le jeu de données propre à la problématique abordée.

Afin de donner le même statut aux variables dans les analyses, nous avons mis en œuvre une méthode permettant de traiter les grandes masses de données. En effet le grand nombre d'observations indépendantes définies dans ce programme et le très grand nombre de variables biologiques à tester nécessite des techniques d'analyse telles que les Random Forests qui peuvent s'affranchir du problème de multi-colinéarité. Afin d'élaborer un modèle explicatif et prédictif de l'état d'un sol nous avons injecté la batterie d'indicateurs sensibles aux différents facteurs, sélectionnée par Random Forest dans une analyse discriminante.

La méthodologie des Random Forest consiste en la sélection des variables les plus discriminantes. Nous avons ainsi recherché la meilleure sélection en regroupant les deux groupes de variables biologiques représentant la Microflore et la Faune. Ces travaux ont été mis en œuvre sur les trois facteurs de classement: l'occupation des sols, les niveaux de contamination en ETM et en polluants organiques.

Nous avons ensuite utilisé les variables les plus discriminantes issues de chaque analyse discriminante par RF et analysé de nouveau ce groupe de variables par analyse discriminante linéaire pour chaque facteur de classement. En effet les RF nous ont permis de réduire le nombre de variables entrées initialement dans le modèle.

Il existe des travaux de synthèse de nombreux indicateurs biologiques d'état des sols (Bastida *et al.,* 2008). Les variables biologiques mesurées regroupent les mesures descriptives des communautés de la macrofaune, des lombrics, de la mésofaune, de la nématofaune, des biomasses microbiennes, des activités enzymatiques, certaines variables microbiologiques des fonctions microbiennes et de la diversité fonctionnelle.

L'IBQS (ORD_CLASSE, ORD_BRUT) est un indice de macrofaune construit selon la présence d'espèces. Les lombrics sont décrits par les abondances ou les biomasses des groupes écologiques et des indices construits sur ces groupes (BIOmasse totale, BIO_ANéciques, BIO_ENdogés, BIO_EPigés, Abondance totale, AB_ANéciques, AB_ENdogés, AB_EPigés, RICHesse_SPEcifique, DIVersité_SPEcifique, EQUItabilité_SPEcifique). La mésofaune est décrite par les variables des abondances des groupes fonctionnels des collemboles (EPIédaphiques, HEMIédaphiques, EUédaphiques, COLL_TOT, MICROARTHR_TOT, COLL_R, COLL_DIV, COLL_EQUI, AB_ACAR, AB_AR). La nématofaune est décrite par les variables des abondances des groupes fonctionnels des nématodes et des indices construits sur ces groupes (PHYTO_FALC, PHYTOPARA, BACT_TOT, FONG, OMNI, CARNI, PHYTO, TOT_ENTOM, MI, PPI, NCR, EI, SI, CI). Les variables des activités enzymatiques sont l'arylsulfatase (ARYLS), la galactosidase (GALACTO), la laccase, la lipase (LYP), l'arylaminase (ARYLN), la β-glucosidase (B_GLUCO), les phosphatases (P_AC, P_ALC), la cellulase (CELL), la n-acethyl_glucosidase, (NAG), la xylanase (XYL), la FDA. Les fonctions microbiennes mesurées sont les variables de respiration microbienne (RESP et RESP_SPE) et l'UFC. Les biomasses microbiennes sont décrites par la biomasse totale, et le ratio BIOMASSE/CT, les différents ADNr 18S, ADNr 16S, l'ADN spécifique OPRF, l'ADN_RDT_EXTRACT ((ADN total extrait du sol), les variables des ergostérols totaux et libres (ERGO_TOT, ERGO_LIBRES), les variables spécifiques des acides gras phospholipides (différents ratio de PLFA) et celles de la diversité bactérienne fonctionnelle (PP, AWCD).

# RÉSULTATS ET DISCUSSION

## Variables sensibles aux niveaux de contamination en ETM

Le *tableau 1* décrit la sélection par la méthode de Random Forests des variables biologiques pour les 47 parcelles

**Tableau 1 -** Variables obtenues par RF pour le facteur contamination en ETM, variables descriptives du groupe des biomasses microbiennes (vert), des fonctions microbiennes (gris), des lombriciens (rose), des nématodes et des collemboles (jaune).

***Table 1 -*** *Variables selected by RF for heavy metal contamination, descriptors of the group of microbial biomasses (green), microbial functions (grey), earthworms (pink), nematodes and mesofauna (yellow).*

| | |
|---|---|
| PLFA_GTP | PLFA bactéries à Gram+ Total |
| AB_AN | Abondance des Anéciques |
| ARYLS | Arylsulfatase |
| BIO | Biomasse Totale Lombriciens |
| BIO_AN | Biomasse des Anéciques |
| PLFASM | PLFA saturés/PLFA mono-insaturés |
| COLL_DIV | Indice de Diversité - Collemboles |
| RESP | Respiration microbienne |
| RICH_TAXO | Richesse Taxonomique - Lombriciens |
| XYL | Xylanase |
| PHYTOPARA | Ab. Phytophages parasitaires - Nématodes |
| Biomasse_CT | Ratio Biomasse microbienne/Carbone total |
| NCR | Nematode Channel Ratio |
| RICH_SPE | Richesse Spécifique - Lombriciens |
| PLFAFB | PLFA fongique/PLFA bacterien |
| P_AC | Phosphatase acide |
| AB | Abondance totale - Lombriciens |
| PP | Puits Positifs diversité bacterienne |
| DIV_SPE | Diversité Specifique - Lombriciens |
| ERGO_LIBRE | Ergosterol libre |
| HEMI | Hémiédaphiques - Collemboles |
| BIO_EN | Biomasse des Endogés |

expérimentales dont les modalités sont spécifiques à chaque site, de plus elles décrivent un gradient en ETM. Les parcelles expérimentales de tous les sites sont classées selon les trois niveaux de contamination en ETM préalablement définis. Nous avons retenu les 20 variables apparaissant systématiquement pour tous les sous-échantillons obtenus à partir des tirages avec remise au sein de l'échantillon d'apprentissage. En effet, certaines variables biologiques sont ensuite apparues avec une occurrence de 50 % (5 fois sur les 10 tirages) avec des indices de Gini beaucoup plus faibles.

En résumé, certaines variables descriptives des Lombriciens, dont trois indices (en rose), plusieurs biomasses bactériennes et fongiques dont trois PLFA (en vert), mais aussi plusieurs indices et abondances en microfaune (en jaune) et quelques activités enzymatiques et diversité fonctionnelle microbienne (en gris) apparaissent sensibles à la contamination par les ETM pour les sites expérimentaux considérés. Certains sites contaminés en ETM présentent aussi des niveaux de contamination organique en HAP et en produits phytosanitaires. Les sites agricoles sont de contamination faible en ETM, mais d'usage différent. Nous appliquons l'analyse discriminante linéaire en utilisant la sélection obtenue par les modèles Random Forests, et ce pour chaque cas étudié ci-dessus.

## Analyse discriminante pour le facteur de contamination métallique

En mettant en œuvre l'AFD (Analyse Factorielle Discriminante) sur les indicateurs sélectionnés par les Random Forests, les résultats montrent que la discrimination est parfaite entre les 3 niveaux de contamination. En effet, l'axe 1 a un très bon pouvoir discriminant, de 84 %, avec un lambda de Wilks de 0,0449526 et une Pvalue < 0,0001. La matrice de confusion (*tableau 2*) nous donne un pourcentage de bons classements de 98 %.

Une seule parcelle expérimentale fortement contaminée en ETM est mal classée, il s'agit de AUBHCO qui du classement *a priori* « Fort » se retrouve en « Moyen ». D'après la *figure 1*, l'axe F1 discrimine les parcelles à modalités de faible niveau de contamination par rapport aux deux autres niveaux moyens et forts. L'axe F2 discrimine les deux niveaux de contamination Moyens et Fort. Ce résultat révèle une structure des données prévisible, dans la mesure où les sites à taux faibles de contamination sont aussi les sites à usages agricoles. Par contre, les sites de forte et moyenne contaminations en ETM sont constitués de bois, ou de friches industrielles et sont impliqués dans des programmes de restauration et de réhabilitation.

Nous constatons que les centroïdes des trois groupes sont bien séparés. Les variables les plus discriminantes par une sélection pas à pas descendante (ou ascendante) sont les PLFA_GTP (PLFA Gram+ Total), BIOmasse Totale Lombriciens (ou AB_AN – Abondance Anéciques), COLL_DIV (Indice de Diversité Collemboles), Arylsulfatase. Les variables AB_AN et BIO étant significativement corrélées ne peuvent être retenues en même temps. Avec un taux de bons classements de 80 %, le modèle prédictif issu de l'analyse factorielle est:

$$\text{Predict (Modalité i )}=i_{\alpha}$$
$$\text{avec } \alpha=\text{rang}(\max(i_{faible}; i_{moyen} ;i_{fort}))$$

avec

$i_{faible}= (366{,}533 PLFA_{GTP}+0{,}134\ ARYLS+0{,}102\ BIO-4{,}85 COLL_{div}-55{,}5)$

$i_{moyen}= (293{,}86 PLFA_{GTP}+0{,}48\ ARYLS+0{,}063\ BIO-3{,}031 COLL_{div}-41{,}45)$

$i_{fort}= (277{,}24 PLFA_{GTP}+0{,}137\ ARYLS+0{,}039\ BIO-0{,}922 COLL_{div}-33{,}698)$

**Tableau 2 -** Matrice de confusion avec toutes les variables de la sélection par RF du *tableau 1* pour le facteur de contamination ETM.

**Table 2 -** *Classification table taking into account all variables selected by RF of Table 1 for heavy metal contamination.*

| Actual Classe_ Contam_ETM | Group Size | Predicted Classe_Contam_ETM | | |
|---|---|---|---|---|
| | | Faible | Fort | Moyen |
| Faible | 23 | 23 (100,00 %) | 0 (0,00 %) | 0 (0,00 %) |
| Fort | 17 | 0 (0,00 %) | 16 (94,12 %) | 1 (5,88 %) |
| Moyen | 7 | 0 (0,00%) | 0 (0,00%) | 7 (100,00%) |

*Percent of cases correctly classified: 97,87 %*

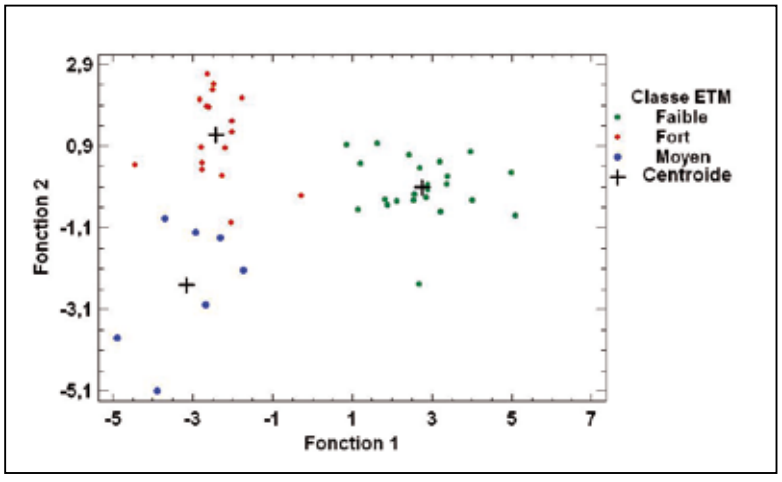
Ce modèle permet de prédire l'appartenance d'une nouvelle observation à l'une des modalités (faible, moyen ou fort) sur la seule connaissance des valeurs obtenues à partir des variables PLFA Gram+ Total, de la Biomasse Totale Lombriciens, de l'Indice de Diversité Collemboles et de l'Arylsulfatase.

## Variables sensibles aux niveaux de contamination organique

Le *tableau 3* décrit la sélection des variables biologiques pour les 47 parcelles expérimentales par la méthode des Random Forests. Plusieurs variables représentant des abondances et des indices de la mésofaune (en jaune), certaines variables de biomasse fongique ou microbienne (en vert), peu de biomasses et indice des Lombriciens (en rose), et trois activités enzymatiques (en gris) apparaissent sensibles à la contamination organique pour des modalités des sites expérimentaux considérés. Cette contamination est plurielle car elle implique des modalités contaminées en HAP principalement mais aussi en herbicides. La séparation selon les usages n'apparaît pas aussi clairement dans ce cas, car certains sites de types agricoles sont classés au niveau moyen de la contamination organique. Ce qui est clair, c'est la grande sensibilité des indicateurs de mésofaune (en jaune dans le *tableau 3*) qui prévalent dans cette analyse suivis de près par les indicateurs de biomasses bactériennes et fongiques (en vert).

**Figure 1 -** Plan factoriel discriminant pour la contamination en ETM.

***Figure 1 -*** *Factorial plan which discriminates heavy metal contamination.*




## Analyse discriminante pour le facteur de contamination organique

L'application de l'analyse par AFD (Analyse Factorielle Discriminante) aux variables sélectionnées par RF pour le facteur de contamination organique conduit aux résultats suivants (*tableau 4*).

La discrimination est parfaite entre les 3 niveaux de contamination organique Faible, Moyenne et Forte. La valeur du Lambda de Wilks est 0,0694469 (Pv < 0,0001). Les barycentres des 3 groupes sont bien séparés (*figure 2*). L'axe factoriel F1 discrimine les niveaux de contamination moyens et faibles, qui concerne des sites agricoles mais aussi en cours de restauration, et l'axe factorielle F2 discrimine bien le niveau de contamination Forte des deux autres. La matrice de confusion ci-dessous nous donne un taux de classement de 100 % pour les 47 parcelles expérimentales.

Les variables les plus discriminantes (stepwise) au nombre de 5 sont PLFA_GTP (PLFA Gram+ Total, GALACTOsidase, SI (Indice nématode SI), PHYTO (Abondance en nématodes phytophages) et PLFAFB (Ratio PLFA Fongique/PLFA Bactérien).

Nous remarquons qu'avec cette sélection nous avons un excellent modèle car près de 92 % des parcelles expérimentales sont bien classées.

Le modèle prédictif issu de l'analyse factorielle est dans ce cas:

$$\text{Predict (Modalité i)} = i_{\alpha}$$

avec $\alpha = \text{rang}(\max(i_{faible}; i_{moyen}; i_{fort}))$

et

$$i_{faible} = (0{,}175\,SI + 89{,}522\,PLFAFB + 1{,}28\,GALACTO - 0{,}00513\,PHYTO + 436{,}028\,PLFA_{GTP} - 67{,}23)$$

$$i_{moyen} = (0{,}006SI + 68{,}94PLFAFB + 1{,}843\,GALACTO - 0{,}0086\,PHYTO + 556{,}5PLFA_{GTP} - 95{,}74)$$

$$i_{fort} = (0{,}182SI + 135{,}665PLFAFB + 0{,}954\,GALACTO - 0{,}0054\,PHYTO + 401{,}659PLFA_{GTP} - 66{,}19)$$

L'appartenance d'une nouvelle observation à l'une des modalités (faible, moyen ou fort) pour le facteur de contamination organique

**Tableau 3 -** Variables obtenues par RF pour le facteur contamination organique, variables descriptives du groupe des biomasses microbiennes (vert), des fonctions microbiennes (gris), des lombriciens (rose), des nématodes et des collemboles (jaune).

***Table 3 -*** *Variables selected by RF for the organic contamination, descriptors of the group of microbial biomasses (green), microbial functions (grey), earthworms (pink), nematodes and mesofauna (yellow).*

| | |
|---|---|
| ADNR_18S | Biomasse ADN fongique |
| SI | Indice de Structure des nématodes |
| BIOMASSE | Biomasse microbienne totale |
| PLFAFB | PLFA ratio fongique/bacterien |
| AB_AR | Abondance des Arthropodes |
| PHYTO_FALC | Ab. des nématodes phytophages facultatifs |
| EI | Indice d'Enrichissement des nématodes |
| GALACTO | Galactosidase |
| OMNI | Abondance des nématodes omnivores |
| ERGO_TOT | Ergosterol Total |
| PHYTO | Abondance nématodes phytophages |
| AB_ACAR | Abondance des Acariens |
| RESP | Respiration microbienne |
| PLFA_GTP | PLFA bactéries à Gram+ Total |
| PLFA_TOT | PLFA Total |
| RICH_TAXO | Richesse Taxonomique Lombriciens |
| MICROARTHR_TOT | Abondance Microarthropodes Totale |
| BIO | Biomasse Totale Lombriciens |
| AB | Abondance des Lombriciens |
| AWCD | Diversité Bacterienne Fonctionnelle |
| P_AC | Phosphatase acide |
| RICH_SPE | Richesse Spécifique Lombriciens |

**Tableau 4 -** Tableau des inerties pour le facteur contamination organique.

***Table 4 -*** *Eigenvalues table for organic contamination.*

| Discriminant Function | Eigenvalue | Relative Percentage | Canonical Correlation |
|---|---|---|---|
| 1 | 6,46171 | 87,42 | 0,93058 |
| 2 | 0,929785 | 12,58 | 0,69412 |

**Tableau 5 -** Matrice de confusion avec toutes les variables de la sélection par RF pour le facteur contamination organique.

***Table 5 -*** *Classification table taking into account all variables selected by RF for the organic contamination.*

| Actual Classe_ Contam_ Orga | Group Size | Predicted Classe_Contam_ORGA | | |
|---|---|---|---|---|
| | | faible | fort | moyen |
| Faible | 39 | 39 (100,00 %) | 0 (0,00 %) | 0 (0,00 %) |
| Fort | 3 | 0 (0,00 %) | 3 (100,00 %) | 0 (0,00 %) |
| Moyen | 5 | 0 (0,00 %) | 0 (0,00 %) | 5 (100,00 %) |

*Percent of correctly classified: 100,00 %*

peut être déduite à partir des valeurs observées des variables SI, PLFAFB, la GALACTO, Phyto et les PLFA_GTP.

## Variables sensibles au facteur d'usage du sol

Le facteur usage du sol tel qu'il est défini dans l'étude est plus complexe car se découpe en 4 classes définissant les cultures (C), prairies (P) donc liés à un usage purement agricole, des bois et friches (B), et des forêts d'épicéa (F). Le *tableau 6* montre la sélection réalisée par RF selon des modalités décrites plus haut. Peu de variables de la mésofaune (en jaune) apparaissent sensibles à ces usages, contrairement aux variables de la macrofaune (en rose), en particulier l'IBQS, mais aussi de nombreuses variables des biomasses microbiennes (en vert), telles que les ADNs, de biomasse fongique telles que les ergostérols, la diversité bactérienne et peu d'activités enzymatiques (en gris) mais présentes. Par ailleurs, la diversité de la macrofaune et microbienne sont présentes dans la sélection.

L'application de l'analyse par AFD (Analyse Factorielle Discriminante) aux variables sélectionnées par RF pour le facteur d'usage du sol conduit à des résultats probants.

## Analyse discriminante pour le facteur usage des sols

En reprenant les 21 variables sélectionnées par RF pour le facteur usage du sol et en appliquant l'analyse discriminante, le *tableau 7* des inerties montre que deux axes factoriels restituent 83 % de l'inertie totale.

Toutes les modalités sont bien classées à l'exception de la parcelle Me117C classée a priori en C (Culture), mais se retrouvant dans le groupe B (Bois et Friches).

Deux axes factoriels regroupent 83 % de l'inertie totale. La *figure 3* montre la complexité du jeu de données. En effet, l'axe factoriel F1 discrimine parfaitement les forêts naturelles (F) des parcelles de modalités de cultures (C) mais aussi des bois et friches (B). L'axe F2 discrimine les usages agricoles de cultures et prairies des deux autres usages des sols les forêts (F) et les friches et bois (B). Une inertie résiduelle est faible mais existante, ce qui met en évidence une plus grande difficulté à sélectionner les variables les plus sensibles.

Par une sélection stepwise, les variables les plus discriminantes sont la respiration microbienne, le ratio Biomasse totale/Carbone Total (Biomasse_CT), les ergosterols total et libre, le nombre de Puits Positifs (PP) de la mesure de diversité fonctionnelle bactérienne, une abondance en nématodes Phytoparasitaires (PHYTOPARA), mais aussi la variable de diversité de la macrofaune telles que l'IBQS et l'abondance (AB) ou la biomasse totale des lombriciens (BIO) et l'activité enzymatique Arylsylfatase.

Les résultats mettent en évidence des ensembles d'indicateurs biologiques permettant de prédire l'état d'un sol. Certes, certains résultats sont prévisibles et la mise en œuvre selon deux étapes, par RF puis par AFD a permis de le montrer statistiquement. Pour le critère de l'usage du sol défini dans ce jeu de données, la démarche d'analyse statistique proposée permet de montrer l'importance des variables descriptives de la biodiversité, parmi l'ensemble de variables caractérisant les communautés vivantes et certaines fonctionnalités des sols (Thoisy-Dur *et al.,* 2011).

Dans la sélection par RF, les variables sensibles aux contaminations métalliques et organiques sont les suivantes: la respiration microbienne, les PLFA total, le PLFA Gram+ total, le ratio PLFA fongique/bactérien, la Phosphatase acide, la diversité bactérienne fonctionnelle, mais aussi les biomasses, abondance, richesses spécifique et taxonomique des lombriciens. Cependant, le fait de disposer dans le jeu des données de nombreuses variables pour les nématodes dont les indices et abondances des groupes trophiques, de deux ergostérols et de nombreuses activités enzymatiques conduit à des sélections qui semblent très diverses par les variables sélectionnées, mais qui le seraient beaucoup moins si nous avions un nombre plus réduit de variables pour le même indicateur communautaire ou fonctionnel. C'est ce qui se produit en particulier pour le groupe de nématodes, des collemboles et pour les activités enzymatiques. A l'intérieur de ces groupes, existent des variables clairement sensibles aux contaminations et usages des sols, mais ce ne sont pas toujours les mêmes variables descriptives sélectionnées selon le facteur considéré.

**Figure 2 -** Plan factoriel discriminant pour la contamination organique.
***Figure 2 -*** *Factorial plan which discriminates organic contamination.*

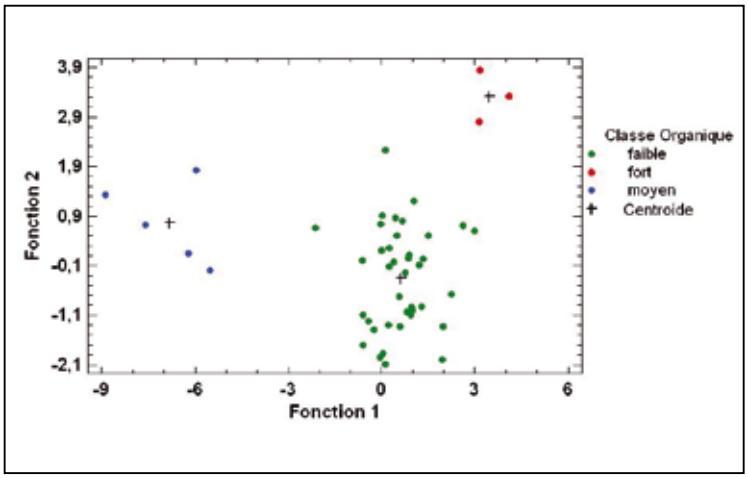
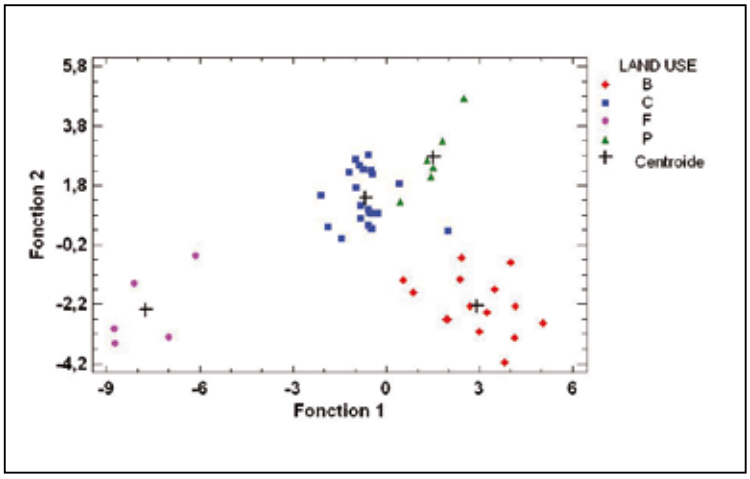



**Figure 3 -** Plan factoriel discriminant pour le facteur usage du sol.
***Figure 3 -*** *Factorial plan which discriminates land use.*



Le modèle prédictif pour la contamination métallique rassemble et concentre la grande capacité de prédiction des indicateurs mesurés dans ce jeu de données. Ces indicateurs liés au fonctionnement biologique des sols, par la biomasse totale des lombriciens, la diversité des collemboles, les PLFA de groupe bactérien à Gram+total, l'activité enzymatique arylsulfatase montrent la nécessité de prendre en compte un groupe d'indicateurs biologiques pour prédire un état, dans le contexte d'une forte contamination en ETM.

## CONCLUSION ET PERSPECTIVES

L'analyse globale nous a permis de montrer que certains bioindicateurs mis en relation permettent de prédire l'état d'un sol. Cet état est caractérisé par son usage et les niveaux de contamination métallique et organique qu'il supporte. Selon l'angle étudié, les indicateurs les plus discriminants et prédictifs ne sont pas systématiquement identiques: les contaminations organiques sont de loin bien décrites par les variables de la mésofaune et de la microfaune, alors que la contamination métallique forte est décrite pour tous les niveaux trophiques représentés dans ce jeu de données, tant bactérien et fongique, que la mésofaune et macrofaune. Par contre, les usages des sols très contrastés ont montré l'importance des variables de biodiversité, moins représentées que les biomasses et les abondances dans ce jeu de données.

La démarche que nous avons développée permet d'une part de sélectionner une batterie d'indicateurs discriminants et explicatifs et d'autre part d'élaborer un modèle prédictif d'une situation donnée (usage, contamination organique ou métallique) en ne faisant aucun *a priori* sur les variables en entrée (Taibi *et al.,* 2011, 2012). En effet la méthode des Random Forests a l'avantage de donner à toutes les variables le même statut. La dimension du tableau d'entrée s'est bien réduite alors, passant de près de 70 variables en entrée à seulement 22 (maximum) indicateurs discriminants en sortie et ce pour chaque situation (ETM, polluants organiques et occupation). La mise en œuvre de l'Analyse Discriminante devient alors plus aisée en confortant d'une part ce que nous avons obtenu comme résultats à partir des RF et d'autre part l'élaboration d'un modèle prévisionnel. Notons qu'en termes de discrimination sur des ensembles de variables de faible dimension, les RF et l'analyse discriminante sont toutes deux performantes (Laroutis et Taibi, 2011). Toutes les matrices de confusion par AD renvoient des taux de classement excellents, même après des sélections ascendantes ou descendantes.

Finalement, une démarche de sélection des bioindicateurs intégrant des critères de faisabilité pour la mise en pratique et le développement de ces outils (Ritz *et al.,* 2008) est en cours afin de moduler les résultats précédents par la prise en compte de leur aspect technique et socio-économique.

**Tableau 6 -** Variables obtenues par RF pour le facteur usage du sol, variables descriptives du groupe des biomasses microbiennes (vert), des fonctions microbiennes (gris), des lombriciens (rose), des nématodes et des collemboles (jaune).
***Table 6 -*** *Variables selected by RF for land use, descriptors of the group of microbial biomasses (green), microbial functions (grey), earthworms (pink), nematodes and mesofauna (yellow).*

| | |
|---|---|
| RESP | Respiration microbienne |
| BIO | Biomasse des Lombriciens |
| ADNR_18S | Biomasse ADN fongique |
| AB_AN | Abondanceb des Anéciques |
| ORD_BRUT | IBQS à l'ordre |
| BIO_AN | Biomasse des Anéciques |
| ARYLS | Arylsulfatase |
| PHYTOPARA | Nématodes Phytophages parasitaires |
| AWCD | Diversité bacterienne fonctionnelle |
| AB | Abondance totale Lombriciens |
| ORD_CLASSE | IBQS à la classe |
| Biomasse_CT | Ratio Biomasse microbienne/Carbone total |
| PP | Puits Positifs diversité bacterienne |
| AB_EN | Abondance des Endogés |
| BIO_EN | Biomasse des Endogés |
| ERGO_TOT | Ergosterol fongique total |
| NAG | N-acethyl-glucosidase |
| ERGO_LIBRE | Ergosterol fongique libre |
| ADN_RDT_ EXTRACT | Biomasse microbienne ADN extrait |
| PLFA_TOT | PLFA total |
| PLFAFB | PLFA ratio fongique/bacterien |

***Tableau 7 -*** *Tableau* des inerties pour le facteur usage du sol.
***Table 7 -*** Eigenvalues table for land use.

| Analyse discriminante (Usage du sol) | | | |
|---|---|---|---|
| Discriminant Function | Eigenvalue | Relative Percentage | Canonical Correlation |
| 1 | 10,4705 | 58,58 | 0,95542 |
| 2 | 4,37896 | 24,5 | 0,90227 |
| 3 | 3,02287 | 16,91 | 0,86685 |

**Tableau 8 -** *Matrice de confusion avec toutes les variables de la sélection par RF pour le facteur usage du sol.*

***Table 8 -*** *Classification table taking into account all variables selected by RF for land use.*

| Actual Land_ Use | Group Size | Predicted Land_Use | | | |
|---|---|---|---|---|---|
| | | B | C | F | P |
| B | 15 | 15 (100,00 %) | 0 (0,00 %) | 0 (0,00 %) | 0 (0,00 %) |
| C | 21 | 1 (4,76 %) | 20 (95,24 %) | 0 (0,00 %) | 0 (0,00 %) |
| F | 5 | 0 (0,00 %) | 0 (0,00 %) | 5 (100,00 %) | 0 (0,00 %) |
| P | 6 | 0 (0,00%) | 0 (0,00%) | 0 (0,00%) | 6 (100,00%) |

*Percent of cases correctly classified: 97,8 %*

## REMERCIEMENTS

… à l'Ademe pour le cofinancement du programme, aux gestionnaires des sites: Muriel Guernion, Francis Douay, Marc Legras, Aude Alaphilippe, Olivier Faure, Adnane Hitmi, Sébastien Conil, Thierry Beguiristain, Aurélia Michaud, Jean-François Vian, à Christian Mougin, Florence Dubs, Cécile Grand, Laurence Galsomies et Antoine Richard pour les nombreux échanges et à la Région Haute Normandie pour son soutien.

## BIBLIOGRAPHIE

Baize D., 2000 - Teneurs totales en « métaux lourds » dans les sols français. Résultats généraux du programme ASPITET. Le Courrier de l'Environnement de l'INRA, 39, pp. 39-54.

Bastida F., Zsolnay A., Hernandez T., Garcia C., 2008 - Past, present and future of soil quality indices: A biological perspective. Geoderma, 147, 3-4, pp. 159-171.

Breiman L., 2001 - Random Forests. Machine Learning, 45, pp. 5-32.

Breiman L., and Cutler A., 2005 - "Random Forests". http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (accessed 20 April 2008).

Brostaux, Y., 2005 - Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction, thesis, Gembloux University, 178 pages.

Cluzeau D., Péres G., Cannavacciuolo M., Bellido A., Guernion M., Ruiz N., Cortet J., Mateille T., Martin-Laurent F., Velasquez E., Mercier V., Bispo A., Villenave C., Ranjard L., Chaussod R., Rougé L., Jolivet C., Lermercier-Foucault B., Ponge J.F., 2008 - How to manage and analyse a large biodiversity data set: the case of the regional "RMQS BioDiv" experience - International Colloquium Soil Zoology Curritiba, Brésil, 25-29 Août 2008.

Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., and Lawler J., 2007 - Random Forests for classification in ecology. Ecology, 88, pp. 2783-2792.

Laroutis D. and Taibi S., 2011 - Discriminant analysis versus Random Forests on qualitative data: Contingent valuation method applied to the Seine estuary wetlands. International Journal of Ecological Economics & Statistics, 20, pp. 1-13.

Le Guédard M., J.-J. Bessoule J.J., 2013 - A self-evaluation method to attribute sound bio-values based on in situ ecotoxicological monitoring. Ecological Indicators, 30, 100-105

Pauget B., Gimbert F., Michaël Coeurdassier, Crini N., Pérès G., Faure O., Douay F., Hitmi A., Beguiristain T., Alaphilippe A., Guernion M., Houot S., Legras M., Vian J-F., Hedde M., Bispo A., Grand C., de Vaufleury A., 2013 - Ranking field site management priorities according to their metal transfer to snails, Ecological Indicators 29 445–454.

Pérès G., Vandenbulcke F., Guernion M., Hedde M., Beguiristain T., Douay F., Houot S., Piron D., Rougé L., Bispo A., Grand C., Galsomies L., Cluzeau D., 2011 -The use of earthworms as tool for soil monitoring, characterization and risk assessment. Example of a Bioindicator Programme developed at National scale (France). Pedobiologia 54, pp 77-87.

Pérès G., Bispo A., Grand C., Galsomies L., 2012 - Le programme de recherche ADEME "Bioindicateurs de l'état biologique des sols". Ses objectifs, sa mise en œuvre et son déroulement - Actes des Journées Techniques Ademe Bioindicateurs & Phytotechnologies: des outils biologiques pour des sols durables, Paris, France, 16-17 octobre 2012.

Remon E., Bouchardon J.L., Le Guédard M., Bessoule J.-J., Conord C., Faure O., 2013 - Are plants useful as accumulation indicators of metal bioavailability. Environmental Pollution, 175, pp 1-7.

Ritz K., Black Helaina I.J., Colin D., Campbell, Harris J A, Wood C., 2008 - Selecting biological indicators for monitoring soils: A framework for balancing scientific and technical opinion to assist policy development. Ecological Indicators, 9, pp. 1212-1221.

Taibi S., Lepelletier P., Pérès G., Rougé L., Dur J-C., Bispo A., 2011 -. Démarche en vue d'élaborer un indice d'état du sol - Journées de statistique de la Société Française de Statistique, Gammarth, Tunisie, 23-27 May 2011.

Taibi S., Thoisy-Dur J.C., Lepelletier P., Rougé L., Dantan J., Bispo A., Grand C., Pérès G., 2012 - Approche statistique de sélection d'Indicateurs et de Biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques - Journées d'Etude des Sols, Versailles, France, 13-23 mars 2012.

Thoisy-Dur J.C., Lepelletier P., Taibi S., Rougé L., Dantan J., Pérès G., Grand C., Bispo A., 2012) - Statistical approach to select soil bioindicators for soil monitoring, risk assessment and soil characterization. Results from the French national Programme "Bioindicators". - 6th SETAC World Congress, Berlin, Germany, 20-24 May 2012.

Villanneau E., Perry-Giraud C., Saby N., Jolivet C., Marot F., Maton D ., Floch-Barneaud A., Antoni V., et Arrouays D., 2008 - Détection de valeurs anomaliques d'éléments traces métalliques dans les sols à l'aide du Réseau de Mesure de la Qualité des Sols, Etude et Gestion des Sols, 15,3, pp.183-200.