L'arbre de régression multivariable: classification d'assemblages d'oiseaux fondée sur les caractéristiques de leur habitat

Marie-Hélène Ouellette[1], Jean-Luc DesGranges[2], Pierre Legendre[1], Daniel Borcard[1]

[1]Département de Sciences biologiques
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

[2]Service canadien de la faune,
Case Postale 10100
Sainte-Foy (Québec), Canada, G1V 4H5

RÉSUMÉ. Les problèmes écologiques liés aux fluctuations des niveaux d'eau (p.ex. dues aux barrages) sont nombreux et souvent mal connus. L'analyse présentée ici s'insère dans un programme de recherche de la Commission mixte internationale de gestion des eaux des Grands Lacs et du Saint-Laurent (CMI). Elle porte sur les assemblages d'oiseaux le long du fleuve Saint-Laurent en corrélation avec leur habitat. Le but est d'utiliser ces assemblages comme bioindicateurs de l'état écologique des milieux riverains. La technique appliquée ici consiste en un arbre de régression multivariable portant sur une sélection de 128 sites. Cet arbre permet de distinguer 6 groupes de sites caractérisés par des assemblages d'oiseaux et des propriétés environnementales spécifiques.

MOTS-CLÉS : arbre de régression multivariable, catégories, répartition spatiale, espèces indicatrices, oiseaux, caractéristiques environnementales.

1 Introduction

La gestion du débit des cours d'eau au moyen de barrages perturbe l'environnement, notamment en homogénéisant la végétation au détriment de la diversité de ses habitants [DES en prépar.2]. Dans le cadre d'un programme de recherche de la Commission mixte internationale de gestion des eaux des Grands Lacs et du Saint-Laurent (CMI), Environnement Canada a étudié les assemblages d'oiseaux le long du fleuve Saint-Laurent ainsi que leur habitat afin d'utiliser ces assemblages comme bioindicateurs de l'état écologique des milieux riverains. Nous appliquerons ici la méthode de l'arbre de régression multivariable de [DEA 02] afin d'identifier les caractéristiques environnementales auxquelles les assemblages d'oiseaux répondent le plus fortement.

2 Matériel

Les analyses présentées ici portent sur 128 sites où ont été répertoriées 73 espèces d'oiseaux. Les observations ont eu lieu en l'an 2003 du lac Ontario jusqu'au lac Saint-Pierre. 95 variables environnementales ont été mesurées ou observées à chaque site. 59 d'entre elles sont issues d'analyses d'images satellitaires; les autres sont essentiellement des descripteurs de contextes paysagers dérivés de l'analyse de la végétation.

3 Méthodes

3.1 Principe de l'analyse de l'arbre de régression multivariable

Cette analyse permet de regrouper les objets multivariables de la matrice réponse en se basant sur des variables explicatives externes. C'est une forme de groupement sous contrainte qui réalise une succession de divisions (partitions) binaires des objets. Chaque partition des objets en deux groupes est faite de façon à minimiser l'impureté (ou erreur, ou statistique TESS [LEG 98]) de la variable réponse (autrement dit, maximiser l'homogénéité intragroupe). Chaque partition est définie par une seule variable environnementale (dite primaire [BRE 84]). Le processus se poursuit jusqu'à l'atteinte d'une partition en petits groupes d'objets. On émonde ensuite l'arbre obtenu en remontant vers la racine, jusqu'à atteindre la taille désirée, en testant chaque partition par validation croisée. Un arbre se décrit par sa taille (nombre de groupes) et son erreur relative [quotient de la somme, pour tous les groupes, de l'impureté totale des objets au sein de chaque groupe (somme des carrés d'écarts à la moyenne multivariable du groupe) sur l'impureté du nœud racine (somme pour tous les objets des carrés des écarts à la moyenne multivariable)]. Parce que cette mesure fournit une estimation trop optimiste des capacité prédictives de l'arbre, on recourt généralement à une autre mesure, l'erreur relative de la validation croisée. Elle varie de 0 (prédictions impeccables) à 1 (prédictions complètement erronées).

Après ces calculs, l'utilisateur peut encore décider de remplacer certaines des variables explicatives définissant les nœuds (variables primaires) par d'autres pour faciliter l'interprétation de l'arbre. Pour guider ce choix, on calcule des indices de similarité entre les noeuds et les autres variables. Ces indices tiennent compte de la répartition des objets d'un nœud par rapport à celle d'une autre variable explicative; on calcule le nombre d'objets qui changent de groupe/le nombre total d'objets, ou encore le pourcentage ajusté quantifié par le nombre d'objets qui changent de groupe/nombre d'objets dans le plus grand groupe du noeud [BRE 84]. L'utilisation de ces variables permet de modifier la topologie de l'arbre et parfois d'augmenter le pourcentage d'explication de la matrice réponse. Pour choisir l'arbre final, on a recours à un réseau d'arbres construits à l'aide de la validation croisée. L'arbre finalement retenu a la capacité de prédire des assemblages en fonction des variables explicatives, ou, à l'inverse, de prédire des caractéristiques environnementales en fonction de la structure de l'assemblage.

La méthode définit aussi des espèces délimitantes pour chaque nœud. Une espèce délimitante a une importante contribution à la variance expliquée de l'arbre à un nœud donné. Ces espèces sont les mieux expliquées (plus petite somme du carré des erreurs) par ce nœud, qui est lui-même caractérisé par une certaine variable environnementale. Cela permet d'identifier les espèces qui répondent le mieux aux variables primaires de l'arbre.

Dans la présente étude, nous avons regroupé les sites en fonction de leurs abondances d'espèces d'oiseaux et caractérisé les habitats des groupes par les variables environnementales décrites à la section *Matériel*.

3.2 Espèces indicatrices

Pour identifier les espèces indicatrices de chaque groupe, nous avons utilisé une méthode statistique de recherche des espèces indicatrices [DUF 97]. Dans cette méthode, les espèces indicatrices sont identifiées à l'aide d'un test par permutation. La statistique du test (*IndVal*) combine la fidélité des espèces (proportion de sites d'un groupe où l'espèce est présente) et leur spécificité (à quel point une espèce ne se trouve que dans le groupe considéré).

3.3 Associations d'espèces

Nous avons utilisé une analyse de concordance de Kendall pour identifier les associations significatives d'espèces, soit les espèces qui ont des distributions géographiques semblables [LEG 05]. Les espèces ont d'abord été divisées en 4 grands groupes par la méthode des *K* centroïdes (*K-means*). La partition a été faite à partir des vecteurs propres, normés à la racine carrée de leur valeur propre, d'une ACP du tableau des abondances d'espèces centrées réduites. Nous avons ensuite identifié, au sein de chaque groupe, les espèces qui ont une concordance (*W* de Kendall) significative avec les autres membres du groupe.

4 Résultats

À partir du réseau d'arbres obtenu (Fig. 1), nous avons retenu l'arbre ayant la plus faible erreur relative. Cet arbre explique 34 %, soit environ 3 % de variation de plus que l'arbre initialement choisi par l'algorithme. Les calculs ont été réalisés à l'aide de la librairie MVPART du langage R [R 04]. Le modèle final retenu (Fig. 2) indique que les variables environnementales délimitantes des 6 groupes (codés de A à F de gauche à droite) sont, en ordre décroissant de contribution au coefficient de détermination multiple : la densité de lisière de marécages arborés (ED_Marbo) qui sépare les groupes A, B, et C des groupes D, E et F ; l'axe géographique (axe.total[,1]) qui sépare les groupes D et E du groupe F ; le pourcentage de prairies humides (CP_P_hum) qui sépare le groupe A des groupes B et C ; la densité de lisière de prairies humides (ED_P_hum) qui sépare le groupe D du E ; et la densité de lisière de friches dans la place-échantillon (ED_Frich) qui sépare le groupe B de C. De A à F, le nombre de sites par groupe est 42, 46, 3, 17, 18 et 2. Dans le groupe A, le Bruant des marais, le Carouge à épaulettes et la Paruline masquée sont les espèces les plus abondantes. Aucune espèce n'est indicatrice au sens du test IndVal, n'étant ni assez spécifique ni assez fidèle. Selon la partition de la variance pour chaque nœud par espèce, le Troglodyte des marais apparaît comme étant l'espèce délimitante pour le nœud qui sépare le groupe A des groupes B et C. Dans le groupe B, les espèces les plus abondantes sont le Bruant des marais, le Carouge à épaulettes et le Troglodyte des marais. Ce groupe n'a aucune espèce indicatrice au sens de IndVal. Les espèces les plus délimitantes sont le Carouge à épaulettes et le Troglodyte des marais. Le groupe D, pour sa part, est représenté par une grande abondance de Merle d'Amérique et de Parulines jaunes. Aucune espèce n'est significative au sens de IndVal. La Paruline jaune est l'espèce la plus délimitante de son nœud. Le groupe E est caractérisé par un grand nombre d'espèces abondantes : la Paruline jaune, le Bruant chanteur, le Carouge à épaulettes, la Paruline masquée, le Merle d'Amérique, le Bruant des marais, l'Hirondelle bicolore et le Quiscale bronzé. Aucune espèce n'est indicatrice au sens de IndVal ; l'espèce la plus délimitante est la Paruline jaune. Les groupes C et F, bien qu'écologiquement intéressants, ne sont pas assez représentés dans cet échantillon (4 et 2 sites) pour justifier une discussion détaillée.

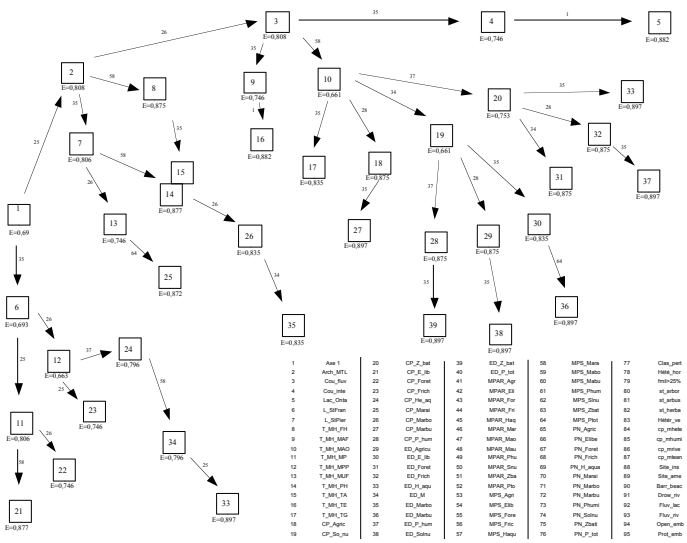

Figure 1. Réseau des arbres construits. Dans les cadres: numéros des arbres. Sur les flèches : variables primaires.

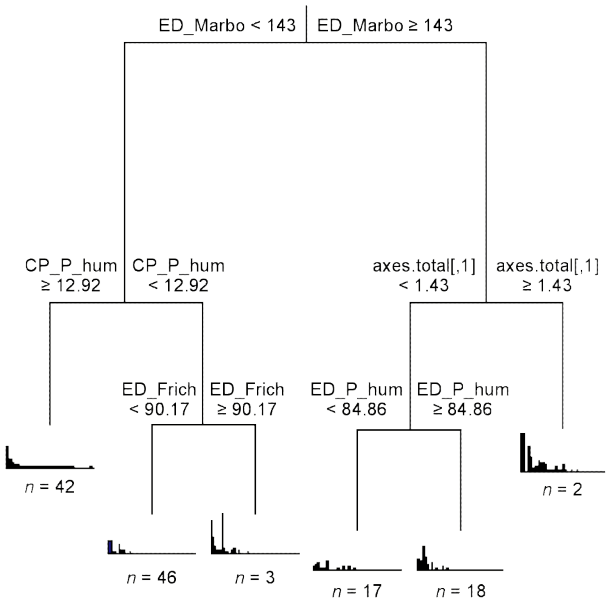

Figure 2. Arbre de régression multivariable final choisi à l'aide du réseau de la figure 1 (case 19).

5 Discussion

D'autres analyses ont été réalisées par DesGranges [DES en prépar.1], avec intégration explicite de descripteurs de l'hydrologie. Les résultats concordent avec ceux qui sont présentées ici dans la mesure où l'effet de l'hydrologie est en partie intégré par les descripteurs issus de l'analyse de la végétation. Ainsi, nos groupes A, B et C correspondent à des marais dépourvus d'arbres. Le groupe A se distingue des deux autres par l'absence de fluctuations de niveaux d'eau de longue durée au printemps. Les groupes B et C sont les plus inondés. L'identification du Troglodyte des marais comme espèce délimitante confirme les résultats obtenus d'autres analyses [DES en prépar.1]. Au pôle arboré du spectre, les groupes D, E et F représentent une toposéquence allant des sites les plus ouverts et humides (F) aux plus fermés (D).

6 Conclusion

L'arbre de régression multivariable est très intéressant lorsque l'objectif est de définir une typologie écologique qui soit non seulement explicative, mais aussi prédictive. Comparée à d'autres approches, cette méthode a le mérite de fournir un modèle d'apparence et d'interprétation simple, grâce à sa structure monothétique. Une telle caractéristique en fait un outil intéressant pour les praticiens de la conservation de l'environnement, pour lesquels l'efficacité repose sur des moyens de diagnostic et de décision simples et rapides. La comparaison des résultats présenté ici à ceux d'autres analyses réalisées sur les mêmes données montre que cette simplicité n'est pas obtenue au détriment de la qualité des résultats scientifiques.

7 Bibliographie

[BRE 84] BREIMAN L., FREIDMAN J., OLSHEN R., STONE C., *Classification and regression trees*, Chapman & Hall, 1984.

[CLA 92] CLARK L., PREGIBON D., "Tree-based models", Chambers J.M., Hastie T.J., editors, *Statistical models in S*, Wadsworth & Brooks, Pacific Grove, California, 1992, p. 377-420.

[DEA 92] DE'ATH G., FABRICIUS K., "Classification and regression trees; a powerful yet simple technique for the analysis of complex ecological data", *Ecology*, vol. 81, 2000, p. 3178-3192.

[DEA 02] DE'ATH G., "Multivariate regression trees: a new technique for modeling species-environment relationships", *Ecology*, vol.83, 2002, p. 1105-1117.

[DES en prépar.1] DESGRANGES J.-L., INGRAM J., DROLET B., SAVAGE C., BORCARD D., "Development of wetland bird assemblage predictive models and performance indicators for use in the environmental assessment of Lake Ontario and St. Lawrence River alternative water regulation plans", *Technical Report No 425*, Canadian Wildlife Service, Québec Region, en préparation.

[DES en prépar.2] DESGRANGES J.-L., LEHOUX D., DROLET B., DAUPHIN D., GIGUÈRE S. SAVAGE C., "Les oiseaux palustres : un groupe sensible aux conditions hydrologiques des zones humides du Saint-Laurent", Environnement Canada, *Série de documents d'évaluation de la science de la DGSAC*, en préparation.

[DIG 81] DIGBY P., GOWER J., "Ordination between and within groups applied to soil classification", *in: Down to earth statistics; solutions looking for geological problems*, Merriam D.F., editor, Syracuse University Geology Contributions, Syracuse, New York, 1981, p. 53-75.

[DUF 97] DUFRÊNE M., LEGENDRE P., "Species assemblages and indicator species: the need for a flexible asymmetrical approach", *Ecological Monographs*, vol.67, 1997, p.345-366.

[LEG 05] LEGENDRE P., "Species associations: the Kendall coefficient of concordance revisited", *Journal of Agricultural, Biological and Environmental Statistics*, 2005 (*sous presse*).

[LEG 98] LEGENDRE P., LEGENDRE L., *Numerical ecology. Second English edition*. Elsevier, Amsterdam, 1998.

[R 04] R DEVELOPMENT CORE TEAM, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2004.